ACTES

DÉCODER L'IA

DES PROMESSES DES OUTILS AUX RÉALITÉS DES USAGES

Cycle d'ateliers

2025-2026

INSTITUT ROBERT BADINTER
Études et recherches sur le droit et la justice

La collection **Actes** propose un témoignage écrit des événements (colloques, séminaires, ateliers, etc.) organisés par l'Institut Robert Badinter anciennement IERDJ (Institut des études et de la recherche sur le droit et la justice).

Elle restitue la richesse des échanges entre intervenants et participants sur des thèmes où se croisent les perspectives des mondes professionnel et de la recherche.

Chaque publication de la collection constitue un document important de diffusion et de valorisation des savoirs pratiques et scientifiques sur le droit et la justice.

INSTITUT
ROBERT BADINTER
Études et recherches
sur le droit et la justice

Avant-propos

Le cycle d'ateliers sur l'intelligence artificielle ouvert par l'institut en 2025 et 2026 s'était donné comme objectif de confronter les promesses de l'intelligence artificielle, toutes plus séduisantes les unes que les autres, aux réalités techniques, institutionnelles et professionnelles dans lesquelles ses usages s'inscrivent désormais.

Conçu comme une invitation à « décoder l'IA », ce cycle a reposé sur une série d'ateliers mêlant interventions d'experts et échanges avec les professionnels du droit et de la justice. Il s'est structuré en deux temps complémentaires sur une année : après une conférence inaugurale ouverte au public au début de l'année 2025 à l'université de Strasbourg ayant permis d'aborder les grands débats que pouvaient susciter l'IA, le cycle s'est ouvert un mois plus tard sur trois ateliers de décryptage : l'un consacré à la notion d'agent, devenue depuis l'un des sujets les plus discutés en matière d'IA ; l'autre sur le développement d'outils d'IA sur mesure pour les professionnels, pressentant les critiques qui se seront depuis confirmées concernant les limites des grands modèles de langage ; le dernier enfin sur l'encadrement juridique de l'IA toujours en construction, laissant bien des utilisateurs sans réponse alors que les usages se multiplient, dans le champ du droit et de la justice aussi. Dans un second temps, des ateliers d'approfondissement dédiés à des applications spécifiques telles que la transcription automatique, la traduction ou encore la synthèse d'écritures, ont permis de rentrer concrètement dans le fonctionnement des IA et de les confronter aux réactions des praticiennes et praticiens, guidés par les chercheuses et chercheurs venus partager leurs travaux. L'ensemble a permis de croiser les approches techniques, juridiques, économiques et sociales, dans un format favorisant la discussion informée et le retour d'expérience.

La préparation, l'animation des ateliers et la rédaction des présents actes ont été assurées à mes côtés par Yannick Meneceur et Olivier Chevet, avec la collaboration de Mélanie Vay. Les ateliers ont bénéficié de la participation d'intervenants issus des mondes académique et professionnel, dont les contributions ont nourri des échanges riches avec les participants, en présentiel comme à distance. Nous tenons à les remercier très sincèrement, et s'agissant des intervenantes et intervenants, ils sont présentés dans la section « Méthodologie » en fin de document. Seuls les propos des intervenantes et intervenants sont référencés dans ces actes. La richesse des discussions avec les participants nourrit de manière complémentaire la synthèse qu'en ont réalisé les auteurs. Cette synthèse bénéficie aussi des connaissances et approfondissement qu'ils ont bien voulu donner à ces actes.

La publication des actes vise à conserver la mémoire de ces travaux et à en partager les principaux enseignements auprès d'un public élargi. Elle s'inscrit dans une démarche plus large de l'Institut consistant à accompagner les acteurs du droit et de la justice dans l'appropriation des transformations induites par l'intelligence artificielle, en identifiant à la fois ses potentialités et les défis qu'elle soulève. Elle constitue, à ce titre, une étape dans un travail appelé à se poursuivre, tant par le soutien de nouvelles recherches que par l'aménagement de futurs espaces de dialogue entre disciplines académiques et professions.

Les ateliers eux-mêmes étaient réservés aux membres et partenaires de l'Institut. Ils avaient été précédés le 23 janvier 2025 par une conférence publique « Décoder l'IA en 2025 : actualités d'une technologie en voie de banalisation » organisé avec l'université de Strasbourg. Cette conférence a donné lieu à la publication d'actes[1] et il est également possible de la revoir en vidéo sur le site de l'IRB[2].

Toujours dans un souci de partager avec un public plus large les enseignements des ateliers, le cycle s'est prolongé hors les murs par des interventions chez certains membres de notre groupement. Une première intervention a eu lieu le 11 décembre 2025, dans le cadre d'une conférence à l'ordre des avocats au conseil d'État et à la Cour de cassation, à l'invitation de son président Thomas Lyon-Caen. L'intervention a été menée à trois voix avec Claire Strugala, directrice adjointe du service de la documentation et la recherche (SDER) de la Cour de cassation, accompagnée de Yannick Meneceur et Olivier Chevet. Cet événement a donné lieu à la rédaction d'un article pour le numéro de 2026 de la revue *Justice et Cassation*[3], à paraitre aux éditions Dalloz.

Le 30 janvier 2026, le cycle a fait une halte à la cour d'appel de Montpellier, sur invitation des chefs de cour, en soutien au projet de juridiction. Après un propos introductif de Haffide Boulakras, directeur adjoint de l'ENM et auteur du rapport « L'IA au service de la justice : stratégie et solutions opérationnelles », ont suivi deux interventions de Yannick Meneceur et Olivier Chevet. Cet événement public a donné lieu à la publication d'actes spécifiques, également disponibles sur le site de l'IRB[4].

Le 23 mars 2026 enfin, dans le cadre de la journée de la relation magistrats-avocats qui portait sur ce thème, Olivier Chevet est intervenu à distance dans le cadre d'une demi-journée d'action sur le thème de la mise en œuvre concrète des outils d'intelligence artificielle par les professions judiciaires, à l'invitation de la cour d'appel de Fort de France.

En un peu plus d'une année, et avec l'ensemble de ces publications, nous espérons avoir contribué à la construction d'une culture commune autour du fonctionnement de l'IA : une culture complexe sans être compliquée, une culture qui ne peut totalement échapper à la technique, mais ne doit pas y rester enfermé non plus.

Harold ÉPINEUSE
Directeur adjoint de l'Institut Robert Badinter

1. Institut Robert Badinter. « Décoder l'IA en 2025. Décoder l'IA en 2025 », Jan. 2025, Paris, France. pp.70, 2025, Actes. ⟨hal-05497926⟩
2. Institut Robert Badinter, Université de Strasbourg « Conférence «Décoder l'IA en 2025 : actualités d'une technologie en voie de banalisation» », https ://institutrobertbadinter.fr/fr/evenements/conference-decoder-lia-en-2025-actualites-dune-technologie-en-voie-de-banalisation/
3. Olivier Chevet, Yannick Meneceur, Claire Strugala, « «Décoder l'IA à l'ordre des avocats au Conseil d'État et à la Cour de cassation», *Justice et cassation*, 2026, à paraître. Site de la revue, https ://www.ordre-avocats-cassation.fr/publications-scientifiques/revue-justice-cassation.
4. IRB, « Décoder l'IA à la cour d'appel de Montpellier », https ://institutrobertbadinter.fr/fr/evenements/decoder-lia-a-la-cour-dappel-de-montpellier/.

Introduction

L'irruption récente de l'intelligence artificielle dans le champ du droit et de la justice se caractérise moins par une innovation isolée que par un faisceau de transformations convergentes, touchant à la fois les outils, les pratiques et les cadres normatifs. Les systèmes dits « génératifs », dont la diffusion rapide alimente à la fois fascination et inquiétude, s'inscrivent dans un mouvement plus large de banalisation technologique qui tend à rendre leur usage presque ordinaire, tout en laissant subsister de fortes incertitudes quant à leurs effets réels. Comme le souligne le programme du cycle, cette tension entre promesse et réalité impose de dépasser les discours généraux pour revenir à une compréhension fine des mécanismes techniques et des conditions concrètes d'implémentation. A ce titre, il s'adresse plutôt à des lecteurs et lectrices déjà initiés au vocabulaire et aux problématiques de l'intelligence artificielle.

Dans ce contexte, le premier enjeu consiste à réintroduire de l'intelligibilité dans des objets techniques souvent appréhendés à travers des représentations simplificatrices. L'analyse des « grammaires » des IA génératives, en tant que systèmes probabilistes fondés sur des modèles statistiques du langage, met en lumière à la fois leur puissance opératoire et leurs limites structurelles. Loin de constituer des substituts autonomes à l'activité juridique, ces outils s'inscrivent dans des chaînes de traitement de l'information dont les paramètres, les données d'entraînement et les modalités d'usage conditionnent fortement les résultats. Il ne s'agit donc pas seulement de s'interroger sur ce que ces systèmes permettent de faire, mais de questionner la manière dont ils transforment les conditions de production du travail juridique.

Le second enjeu tient à l'identification des usages pertinents et des transformations professionnelles associées. Les applications explorées au cours du cycle – résumé de documents, reconnaissance vocale, traduction automatisé – révèlent une constante : l'intérêt de ces technologies réside moins dans une substitution que dans une reconfiguration des tâches, susceptible d'affecter les équilibres internes des professions juridiques. À cet égard, les arbitrages entre solutions génériques (« prêt-à-porter ») et développements spécifiques (« sur-mesure »), ainsi que les exigences de fiabilité, de traçabilité et de contrôle humain, apparaissent comme des déterminants majeurs des choix d'appropriation.

« L'intérêt de ces technologies réside moins dans une substitution que dans une reconfiguration des tâches, susceptible d'affecter les équilibres internes des professions juridiques.

Les ateliers d'approfondissement s'inscrivent dans un courant de pensée qui considère comme importante une compréhension interne minimale des outils mobilisés. Faute de quoi leur appréhension court le risque de rester soit abstraite, soit idéologique, et les choix opérés pour les insérer dans les processus professionnels de ne pas être optimaux, et parfois même délétères. La volonté a été de retenir des applications transversales des technologies génératives suffisamment mûres pour paraître presque déjà banales. Cette situation permet d'envisager leur déploiement rapide dans le monde du droit et de la justice. Pour autant, leur examen permet de constater qu'il reste

des difficultés à surmonter, que la confrontation aux cas concrets révèle des résistances et des points d'achoppement. Elle démontre aussi la dimension absolument essentielle des processus d'évaluation des outils génératifs.

Enfin, l'intelligence artificielle soumet le droit à une double exigence de régulation et d'adaptation. D'une part, l'émergence de cadres normatifs, tels que le règlement européen sur l'intelligence artificielle, traduit la volonté d'anticiper les risques, notamment dans des domaines qualifiés de « haut risque » comme la justice. D'autre part, ces évolutions légales demeurent en tension avec la rapidité des innovations et la diversité des usages, laissant aux acteurs une responsabilité accrue dans l'évaluation et la maîtrise des outils mobilisés. Les réflexions conduites dans le cadre de ce cycle invitent ainsi à envisager l'IA non comme une technologie exogène à encadrer, mais comme un élément désormais intégré aux environnements professionnels, appelant une vigilance continue et une capacité d'appropriation critique.

PREMIÈRE PARTIE
Ateliers de décryptage

Yannick MENECEUR
Expert associé à l'Institut Robert Badinter

Atelier n°1
De l'IA générative à l'IA agentique : une technologie toujours en évolution

- Qu'est-ce qu'un agent intelligent ?
- Le renouveau des agents face aux limites des IA génératives
- Architectures multi-agents : application à la recherche juridique

L'émergence récente de la notion d'« agent » dans le champ de l'intelligence artificielle ne relève pas d'un simple effet de vocabulaire, mais traduit une évolution plus profonde des systèmes numériques. À mesure que les modèles génératifs ont révélé leurs limites – en particulier en matière de fiabilité, d'actualisation des connaissances et de gestion de tâches complexes – s'est imposée l'idée qu'un modèle isolé ne suffisait plus. L'attention s'est alors déplacée vers des architectures capables d'organiser l'action, de structurer le raisonnement et d'articuler plusieurs outils ou sources d'information : c'est précisément le rôle dévolu aux agents.

Ce changement introduit une rupture importante. Là où l'IA générative produisait une réponse en réaction directe à une requête, l'agent inscrit cette réponse dans un processus : il planifie, décompose, vérifie et ajuste. Il ne se contente plus de générer du texte, mais orchestre une suite d'opérations orientées vers un objectif, dans une logique plus proche de l'activité humaine experte. Cette évolution ouvre la voie à des systèmes plus autonomes, mais aussi plus complexes, dont les performances dépendent autant de leur architecture que des modèles qu'ils mobilisent.

Dans ce contexte, il convient d'examiner ce que recouvre précisément la notion d'agent intelligent, les raisons de son renouveau face aux limites des IA génératives, ainsi que les formes concrètes que prennent aujourd'hui les architectures multi-agents, en particulier dans des domaines exigeants comme la recherche juridique.

I - Qu'est-ce qu'un agent intelligent ?

En informatique, et donc aussi dans le domaine de l'intelligence artificielle, le terme « agent » désigne un système logiciel autonome capable de percevoir son environnement, d'analyser des données et d'agir sur cet environnement en fonction d'objectifs définis. Contrairement à un programme classique exécutant une suite d'instructions figées, un agent se caractérise par sa capacité à apprendre, à s'adapter et à interagir de manière continue. Il intègre souvent des mécanismes de prise de décision et peut être doté d'effecteurs lui permettant d'exercer une action (par exemple envoyer une notification, effectuer une transaction, déplacer un robot physique, etc.). En résumé, un agent perçoit, décide, agit et s'adapte de façon relativement autonome.

Les agents peuvent être complètement autonomes, semi-autonomes (laissés en pilotage automatique avec supervision humaine) ou assistants (agissant uniquement sur sollicitation humaine). Historiquement, l'informatique a emprunté ce concept aux sciences économiques et sociales, où la notion d'agent rationnel était déjà utilisée pour modéliser le comportement d'un individu au sein d'une organisation ou d'un système. Les premiers travaux en IA dans les années 1950-60 (notamment ceux d'Herbert Simon, prix Nobel d'économie et pionnier de l'IA) ont posé les bases de l'agent intelligent en cherchant à doter les machines d'une forme de rationalité dans la prise de décision. Au fil des décennies, la recherche en systèmes multi-agents[5] s'est développée, avec l'idée de faire interagir plusieurs agents pour résoudre des problèmes complexes de manière distribuée (on peut trouver des analogies avec l'organisation d'une entreprise, d'une ruche ou d'un système biologique, où chaque agent joue un rôle spécifique au sein d'un tout coordonné).

> Le terme « agent » désigne un système logiciel autonome capable de percevoir son environnement, d'analyser des données et d'agir sur cet environnement en fonction d'objectifs définis.

Exemple concret d'un agent en informatique : l'assistant vocal dans un smartphone

L'assistant vocal inclus dans un *smartphone* se comporte comme un agent pour agir de manière relativement autonome une fois les instructions données et adapter ses décisions en fonction de l'environnement dans lequel se trouve l'utilisateur.

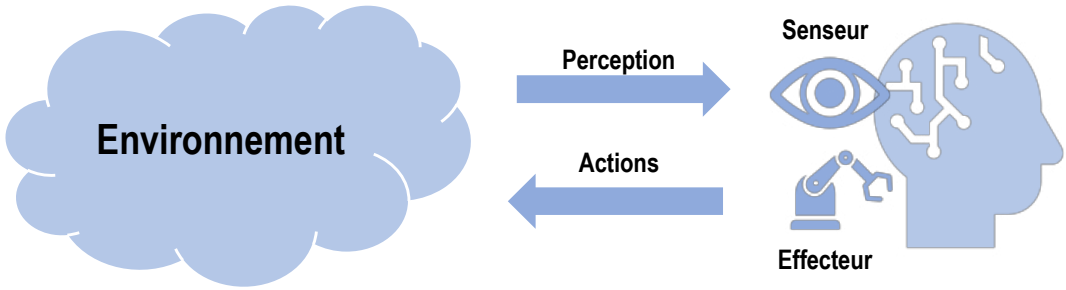
Ainsi, en réponse à l'instruction « Rappelle-moi d'acheter du lait quand je serai près d'un supermarché ouvert », l'assistant vocal fonctionne comme un agent « intelligent » :

L'assistant perçoit son environnement : il écoute votre voix, convertit vos mots en instructions pouvant être traitées par le système d'exploitation du téléphone et situe géographiquement l'utilisateur.
Il prend une décision : il émet de lui-même un rappel quand les conditions sont remplies.
Il agit automatiquement : il enregistre un rappel et surveille votre localisation en arrière-plan.
Il s'adapte : sauf instructions explicites et si les horaires sont bien disponibles en ligne, il est susceptible d'envoyer une notification à proximité de n'importe quel supermarché effectivement ouvert.

Autres exemples d'agents informatiques :

Les agents conversationnels qui répondent à des questions sur un site web et peuvent exécuter certaines tâches simples.
Les recommandations de plateformes vidéo ou musicales, qui analysent vos goûts pour vous suggérer des films ou musiques.
Les robots de *trading* en bourse qui achètent et vendent des actions selon des critères précis.

5. Sur la notion d'agents et de multi-agents, v. par exemple Jean-Pierre Briot, Yves Demazeau, « Introduction aux systèmes multi-agents », *in* Jean-Pierre Briot, Yves Demazeau, *Principes et architecture des systèmes multi-agents*, Paris, Hermès-Lavoisier, 2001, p.17-25.



Le concept d'agent
Source : Auteur

II - Le renouveau des agents face aux limites des IA génératives

L'essor des IA génératives (principalement les grands modèles de langage ou LLM pour *Large Language Models*) a mis en lumière des difficultés structurelles de ces systèmes[6]. D'une part, leur mécanisme de génération statistique (prédire mot après mot ce qui a le plus de probabilité d'apparaître) les amène à produire des réponses parfois factuellement erronées[7] bien qu'elles aient l'air convaincantes (phénomène des « hallucinations »). D'autre part, ces modèles entraînés avec un corpus vaste connaissent un gel de connaissances : leur base d'informations s'arrête à la date de fin d'entraînement et ils ne sont généralement pas en capacité d'actualiser leurs connaissances au fil du temps. Pour pallier cette lacune, une première solution a été l'approche dite RAG (*Retrieval-Augmented Generation* ou génération à enrichissement contextuel), qui consiste à coupler le modèle de langage avec une base de connaissances externe. Concrètement, avant de générer une réponse, le système effectue une recherche dans une base de documents (décisions de justice, textes de loi, articles, etc.) pour y puiser des informations à jour, puis intègre ces éléments factuels dans sa réponse générée. Cette méthode a démontré son efficacité pour améliorer la pertinence et l'actualité des réponses produites par l'IA.

Les agents intelligents proposent d'aller encore plus loin en s'attaquant à la fois au problème de fiabilité et à celui de la complexité des tâches à réaliser. Plutôt que d'avoir un modèle de langage opérant seul, un agent peut être vu comme un chef d'orchestre capable de décomposer une requête complexe en sous-tâches, d'enchaîner de manière autonome différentes étapes de traitement (recherche d'information, analyse critique, synthèse, etc.) et d'ajuster sa stratégie en cours de route en fonction des résultats obtenus. Dans un contexte juridique, pour répondre à une question très spécialisée, un agent pourrait d'abord interroger une base de données jurisprudentielle, puis utiliser un LLM pour formuler de manière claire les extraits pertinents trouvés, et enfin solliciter un module de synthèse pour rédiger une note de conclusion, le tout de façon transparente pour

> Un agent peut être vu comme un chef d'orchestre capable de décomposer une requête complexe en sous-tâches, d'enchaîner de manière autonome différentes étapes de traitement et d'ajuster sa stratégie en cours de route en fonction des résultats obtenus.

6. Sur les limites des grands modèles de langage appliqués au droit, v. Emmanuelle Legrand, Murielle Popa-Fabre, « IA générative & décision du juge », *in Impact du numérique sur la justice*, vol. 2, Paris, IERDJ, coll. « État des connaissances », octobre 2024, p. 26-28.
7. Pour une taxonomie des productions erronées des larges modèles de langage, v. par exemple Lei Huang, Weijiang Yu, Weitao Ma, *et al.* « A Survey on Hallucination in Large Language Models : Principles, Taxonomy, Challenges, and Open Questions », *Association for Computing Machinery*, 2023.

l'utilisateur final. Cette approche multi-étape vise à reproduire le raisonnement qu'aurait un humain expert réalisant ces opérations successivement[8].

Depuis la fin 2024, le concept d'agent autonome a fait un retour remarqué dans la communauté IA, popularisé notamment par des expérimentations comme Auto-GPT ou BabyAGI (des agents IA grand public capables de se donner des objectifs successifs). Appliqué aux modèles de langage, cela ouvre une nouvelle ère d'automatisation avancée. Un agent peut enclencher de lui-même une suite d'actions sans requête supplémentaire de l'utilisateur, comme détecter qu'une réponse initiale est incomplète ou peu fiable, décider de lancer une recherche documentaire complémentaire, puis réévaluer la réponse à la lumière de ces nouvelles informations. Cette proactivité vis-à-vis d'un LLM pur a pour ambition de fiabiliser la production (en évitant de se contenter de la première réponse « brute » du modèle) et de mieux gérer les tâches complexes nécessitant plusieurs étapes de réflexion ou de calcul. En somme, là où un modèle de langage répond à la requête telle quelle, un agent cognitif va chercher ce qu'il faut faire pour y répondre correctement, en mobilisant divers outils ou connaissances de manière autonome.

III - Architectures multi-agents : application à la recherche juridique

Pour illustrer concrètement l'apport des agents dans l'IA générative, il est intéressant d'examiner le retour d'expérience issu d'une expérimentation menée en 2024 dans le domaine de la recherche juridique. Zacharie Laïk, juriste et fondateur d'une legaltech (goodlegal.fr[9]), a développé plusieurs prototypes d'architectures multi-agents visant à améliorer la recherche de documents juridiques grâce à l'IA[10]. Son point de départ était le constat des limites des LLM génériques pour cet usage : bien qu'impressionnants, ces modèles se trompent fréquemment dans un contexte juridique, faute de compréhension fine de la hiérarchie des normes et de capacité à hiérarchiser les informations pertinentes dans une base de données volumineuse. Même l'ajout d'une couche RAG (pour interroger préalablement une base de décisions par exemple) ne résout pas tout : si l'IA « hallucine » moins, elle peine encore à filtrer et organiser correctement les résultats obtenus.

> Limites des LLM génériques bien qu'impressionnants, ces modèles se trompent fréquemment dans un contexte juridique, faute de compréhension fine de la hiérarchie des normes et de capacité à hiérarchiser les informations pertinentes.

1. Architecture multi-agents en série

La première solution testée fut une architecture multi-agents en série (*pipeline*). L'idée était d'enchaîner plusieurs agents spécialisés, chacun accomplissant une tâche précise à la suite du précédent. En pratique, le schéma mis en place était le suivant : un agent planificateur reçoit la question de l'utilisateur et découpe le travail en sous-tâches qu'il assigne à une série d'agents de recherche (par exemple, interroger successivement une base de jurisprudence, puis une base de lois, puis le web). Une fois ces recherches effectuées, un agent critique prend le relais pour évaluer la qualité et la pertinence des résultats obtenus, éventuellement suggérer des recherches additionnelles si des lacunes sont détectées. Enfin, un agent rédacteur compile et met en forme les informations vérifiées sous la forme d'un mémo ou d'un rapport synthétique répondant à la question initiale.

Cette architecture séquentielle (fig.1) a permis d'automatiser un cycle complet de recherche documentaire juridique, chaque agent se concentrant sur une étape du processus. Elle a montré des résultats encourageants en termes de structuration du travail : par rapport à une utilisation directe d'un LLM, l'approche multi-agents produit des réponses

8. Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, *et al.*, « Generative Agents : Interactive Simulacra of Human Behavior », *Association for Computing Machinery*, 2023.
9. Le site goodlegal.fr propose aux étudiants en droit plusieurs fonctionnalités s'appuyant sur de l'IA générative (génération de fiche d'arrêts, synthétiseur de documents, etc.).
10. Les éléments présentés par Zacharie Laïk ont été prépubliés sur LinkedIn : Zacharie. Laïk, *Multi-Agent Systems for Reliable Legal Research : Framework, Experiments, and Optimized Implementation*, LinkedIn, 15 décembre 2024, accessible sur : https ://www.linkedin.com/pulse/multi-agent-systems-reliable-legal-research-framework-zacharie-laïk-4lqfe/.

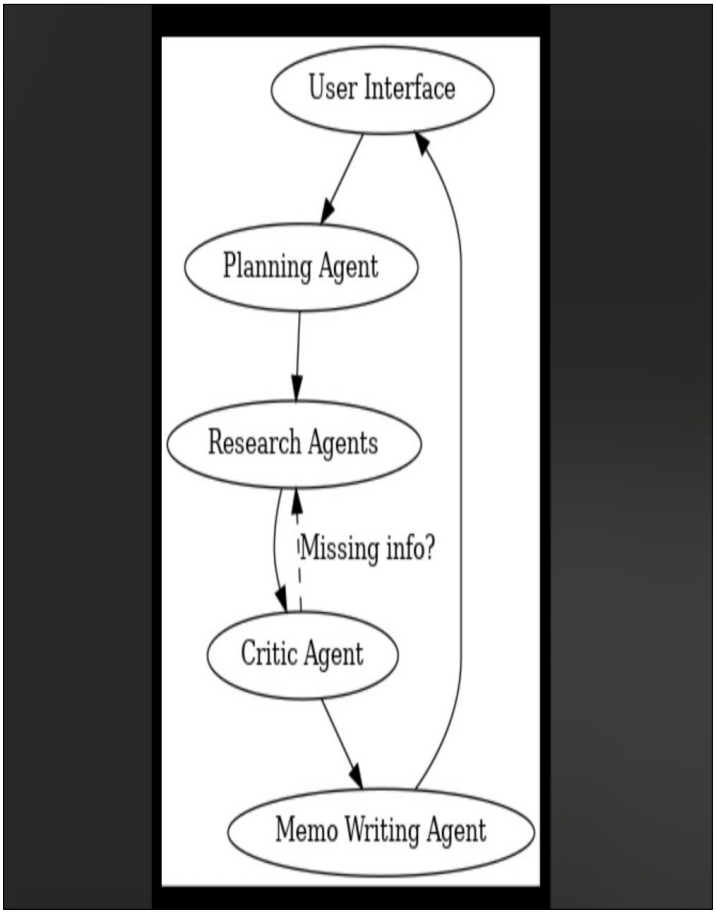


Figure 1 : Exemple d'une architecture multi-agents en pipeline pour la recherche juridique
Source : Zacharie Laïk

> Par rapport à une utilisation directe d'un LLM, l'approche multi-agents produit des réponses mieux étayées, car fondées sur des sources identifiées en amont et passées par un filtre critique humain-simulé.

mieux étayées, car fondées sur des sources identifiées en amont et passées par un filtre critique humain-simulé. Toutefois, ce premier prototype est resté simpliste de l'aveu même de son concepteur. En effet, tous les types de documents juridiques y étaient traités de la même manière, sans spécialisation fine : les décisions de jurisprudence, les textes de loi, les sources web étaient recherchés en série par les mêmes agents génériques, si bien que les résultats pertinents pouvaient être noyés dans un flot d'informations moins utiles, en plus de limites liées au nombre de jetons (*tokens*[11]) pouvant être traités en entrée. De plus, l'exécution strictement séquentielle limitait la réactivité et l'adaptabilité du système.

11. Un jeton est une représentation élémentaire traitée par un modèle de langage, allant du caractère ou ponctuation à une partie de mot ou un mot entier. Les modèles de langage sont limités en nombre de jetons pouvant être traités en entrée et en sortie : GPT-4o, Llama 3 et Mistral Large peuvent en traiter 128 000, Claude 3 200 000 et Gemini 2.0 Flash 1 000 000.

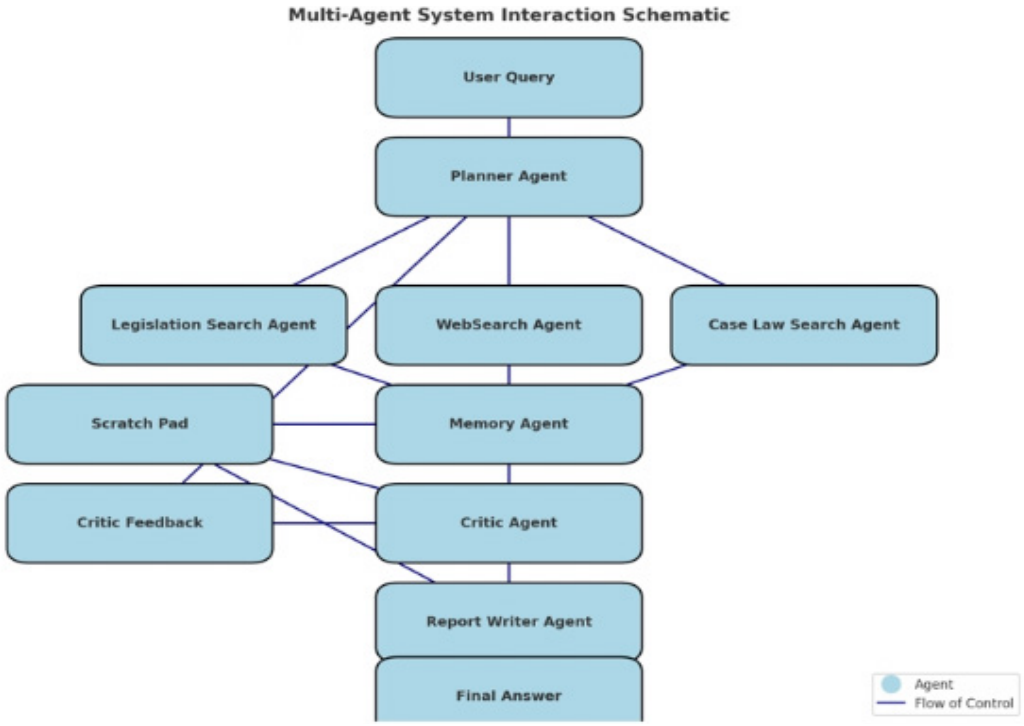


Figure 2 : Exemple d'une architecture d'agents en essaim pour la recherche juridique
Source : Zacharie Laïk

2. Architecture multi-agents en parallèle (ou en essaim)

Pour gagner en efficacité, l'architecture a évolué vers un schéma multi-agents en parallèle, organisé en essaim (*swarm*) d'agents. Concrètement, plusieurs agents de recherche spécialisés opèrent désormais simultanément sur différents types de sources : par exemple un agent dédié à la jurisprudence, un autre aux textes législatifs, un troisième à la recherche web. Chacun exploite de façon optimale sa catégorie de sources. Un agent de mémoire central maintient un bloc-notes partagé pour agréger les données récoltées par tous les agents en parallèle, tandis qu'un agent critique continue d'affiner itérativement les résultats en signalant d'éventuels doublons, incohérences ou manques. Enfin, l'agent rédacteur assemble les contributions de tout l'essaim pour produire le rapport final.

L'architecture en essaim (fig.2) a l'avantage d'une couverture plus complète et rapide des requêtes complexes : en sollicitant en parallèle des agents experts de chaque type de sources, on évite de passer à côté d'une information cruciale noyée dans la masse. La spécialisation par flux (lois, jurisprudence, web…) a nettement amélioré la pertinence du résultat global. Néanmoins, cette approche parallèle a introduit de nouvelles difficultés. La synchronisation des agents s'est révélée ardue : exécutant leurs tâches en même temps, ils produisaient parfois des redondances ou des conflits dans le bloc-notes partagé (données dupliquées, informations incohérentes). Il est même arrivé que certains agents se bloquent en attendant

> L'architecture en essaim a l'avantage d'une couverture plus complète et rapide des requêtes complexes : en sollicitant en parallèle des agents experts de chaque type de sources, on évite de passer à côté d'une information cruciale noyée dans la masse.

Figure 3 : Exemple d'une architecture d'agents en diamant pour la recherche juridique
Source : Zacharie Laïk

indéfiniment une ressource qu'un autre détenait, rappelant les problèmes classiques d'ordonnancements concurrents. En outre, la parallélisation accrue a fait exploser la consommation de ressources, notamment le nombre de *tokens* traités par les modèles de langage, ce qui augmente d'autant le coût financier et temporel de chaque requête. Dans certains tests, répondre à une question complexe pouvait coûter jusqu'à 0,8 $ en appels d'API (pour *Application Programming Interface* ou interface de programmation applicative)[12] de LLM, un montant non négligeable pour un usage répété. Enfin, offrir une interface utilisateur fluide capable de refléter en temps réel le travail simultané de multiples agents s'est avéré complexe.

3. Architecture multi-agents en « diamant »

Tirant leçon de ces difficultés, une troisième itération de l'architecture a été proposée : le modèle dit en « diamant ». Ce schéma cherche à combiner le meilleur des deux modèles précédents en centralisant une partie du processus pour éviter les comportements anarchiques de l'essaim, tout en conservant une exécution partiellement parallèle pour la spécialisation. Concrètement, l'utilisateur pose sa question, qui est d'abord optimisée automatiquement (réécriture du *prompt* pour le rendre plus précis). Puis cette requête optimisée se divise en trois branches spécialisées, exécutées en parallèle (jurisprudence, législation et sources ouvertes sur le web). Cependant, au lieu que chaque branche travaille indépendamment de bout en bout, elles sont coordonnées : chaque branche alimente un bloc-notes central géré par un agent mémoire commun, et c'est seulement quand ce bloc commun est complet qu'intervient la phase de synthèse par l'agent rédacteur. Ainsi, la gestion de la mémoire est recentralisée,

12. Une API est un ensemble de règles et de méthodes qui permet à deux logiciels de communiquer entre eux, sans que l'un ait besoin de connaître le fonctionnement interne de l'autre.

ce qui évite bien des incohérences, et le nombre de branches parallèles est limité pour réduire les problèmes de synchronisation.

L'architecture en diamant (fig.3) a apporté plusieurs améliorations notables. D'abord, elle a simplifié la coordination entre agents, réduisant significativement les conflits et doublons grâce à un flux de données plus clair et recentralisé. Ensuite, elle s'est révélée plus économe en ressources, la consommation totale de *tokens* ayant diminué d'environ 40 % par rapport à l'architecture en essaim. Cette baisse s'explique par une meilleure gestion du contexte partagé : au lieu que chaque agent transmette aux autres l'intégralité de ses résultats (comme c'était le cas dans l'essaim), ici le bloc-notes central agit comme un filtre qui limite les échanges redondants. Cette réduction de consommation de tokens aboutit aussi à un gain en temps de calcul et en coûts d'API. Par ailleurs, le processus plus linéaire (bien que multi-branches) facilite la conception d'une interface utilisateur cohérente grâce à laquelle l'utilisateur voit l'évolution progressive vers la réponse finale, plutôt qu'un foisonnement d'actions parallèles difficiles à suivre. Enfin, cette architecture demeure évolutive : elle permet d'ajouter facilement de nouvelles branches spécialisées en fonction des besoins de l'utilisateur (par exemple intégrer une recherche doctrinale ou une recherche dans une base de données privée) sans remettre en cause la structure générale.

> **L'architecture en diamant a apporté plusieurs améliorations notables. Elle a simplifié la coordination entre agents, elle s'est révélée plus économe en ressources le processus plus linéaire facilite la conception d'une interface utilisateur cohérente ette architecture demeure évolutive**

4. Des bénéfices à objectiver

En somme, l'emploi d'agents intelligents en complément des IA génératives ouvre des perspectives prometteuses pour améliorer la fiabilité et la sophistication des systèmes. L'exemple de la recherche juridique montre qu'en combinant plusieurs algorithmes spécialisés (LLM, outils de recherche, analyse critique) qui dialoguent entre eux, on peut approcher davantage le niveau d'analyse d'un humain expert tout en conservant la vitesse d'exécution de la machine.

Pour autant, il ne faut pas surestimer ces avancées : même avec la RAG et des agents multi-étapes, les modèles de langage conservent une propension structurelle à l'erreur dans les tâches complexes[13]. Dans un contexte juridique où l'exactitude est primordiale, il apparaît inacceptable d'introduire un système dont on sait qu'il produit toujours un pourcentage incompressible d'erreurs factuelles[14]. Cela explique le scepticisme prudent de nombreux praticiens du droit, malgré l'enthousiasme technologique ambiant. La recherche actuelle s'oriente donc vers la définition de standards d'évaluation (*benchmarks*) spécifiques à ces nouvelles applications, pour mesurer objectivement leurs performances et déterminer dans quelles conditions leurs erreurs éventuelles peuvent être jugées tolérables ou non.

13. Maurice Jakesh, Jeff Hancock, Mor Naaman, « Human Heuristics for AI-Generated Language are Flawed », *Proceedings of the National Academy of Sciences*, PNAS, 2023.
14. Sur les erreurs factuelles dans le domaine juridique, v. par exemple Matthew Dahl, Varun Magesh, Mirac Suzgun, Daniel E. Ho, « Large Legal Fictions : Profiling Legal Hallucinations in Large Language Models », *Journal of Legal Analysis*, 2024, vol. 16, n° 1, p. 64-93.

Atelier n°2
Usages et développement de l'IA générative dans le quotidien d'un cabinet d'avocats

- Panorama des usages actuels de l'IA générative en pratique juridique
- Limites et enjeux des IA génératives appliquées au droit
- Transformations du métier d'avocat et nouvelles compétences
- Choix technologiques : solutions du marché ou développement sur mesure ? L'exemple de FidalIA

À mesure que les technologies d'intelligence artificielle générative gagnent en maturité, leur mobilisation par les professions juridiques apparaît comme porteuse d'évolutions structurantes des pratiques professionnelles. En l'espace de quelques années, ces outils ont quitté le champ expérimental pour s'inscrire dans les usages quotidiens, portés par leur capacité à traiter de larges volumes d'informations, à produire des textes structurés et à assister le raisonnement dans ses dimensions les plus formalisées. Cette dynamique s'inscrit dans un double mouvement : d'une part, une recherche accrue d'efficience face à l'intensification des charges informationnelles pesant sur les juristes ; d'autre part, une voie de transformation plus profonde des modes de production du travail juridique, potentiellement médiés par des systèmes automatisés.

L'intérêt croissant pour ces technologies ne tient donc pas seulement à leurs performances techniques encore à éprouver, mais à leur potentiel de reconfiguration des équilibres entre analyse, rédaction et décision. Il invite à interroger à la fois les usages concrets qui se développent au sein des différentes spécialités juridiques, les conditions de leur appropriation, ainsi que les limites et risques qu'ils comportent. C'est dans cette perspective que seront examinés la diversité des cas d'usage actuellement observés, puis les enjeux que soulève cette mobilisation de l'IA pour les professions du droit et les transformations qu'elle induit dans leurs pratiques.

I - Panorama des usages actuels de l'IA générative en pratique juridique

1. Une vaste gamme de cas d'usage

Moins de trois ans après l'arrivée de ChatGPT, les professionnels du droit ont déjà commencé à explorer un large éventail d'usages pour les IA génératives. D'après une étude internationale de Wolters Kluwer (Future Ready Lawyer 2023), près de 73 % des avocats interrogés prévoyaient d'intégrer l'IA générative dans leur travail juridique dans l'année à venir[15]. Cette proportion, très significative, s'explique par la polyvalence des tâches que ces outils peuvent accomplir pour décharger les juristes de certaines besognes chronophages.

Parmi les cas d'usage les plus fréquents, on retrouve la rédaction assistée de documents juridiques (contrats, courriers, mémos, conclusions). Un assistant conversationnel couplé à un modèle de langage peut, par exemple, générer une première ébauche de contrat à partir de quelques clauses clés fournies, ou bien reformuler en langage courant des arguments juridiques pointus. De nombreux avocats utilisent déjà, parfois discrètement (« Shadow AI »), des outils comme ChatGPT pour obtenir un brouillon de lettre à un client résumant l'état d'un dossier, ou pour simplifier le phrasé d'une clause technique à destination d'un non-juriste. L'IA

15. Lyle Moran, « 73% of lawyers plan to use generative AI, report finds », *LegalDive*, 2023, https ://www.legaldive.com/news/generative-ai-legal-use-cases-wolters-kluwer-report/700342.

sert ainsi de « *rédacteur fantôme* » ou de pédagogue aidant à vulgariser le droit.

Un autre domaine d'usage majeur est la recherche et la synthèse d'informations juridiques. Un LLM entraîné spécifiquement peut exploiter le contenu des bases de données de jurisprudence ou de législation et en extraire les points saillants pour répondre à une question donnée. Il peut également résumer de longs documents juridiques (par exemple, un arrêt de cour de cassation très dense ou un rapport d'expertise) en quelques paragraphes synthétiques. Cette capacité de synthèse pourrait être précieuse pour gagner du temps lors de la lecture de décisions volumineuses ou de la préparation de dossiers : plutôt que d'éplucher page à page un jugement de 50 pages, le juriste peut demander à l'IA d'en dégager les faits, la procédure, les motifs et le dispositif principaux.

La revue de documents juridiques est également facilitée. En matière de mise en conformité ou de contentieux par exemple, il s'agit souvent de passer au crible des volumes importants de pièces (correspondances, contrats, pièces financières) : des outils d'IA peuvent aider à détecter des clauses spécifiques, à identifier des incohérences entre différentes versions d'un contrat, voire à anonymiser automatiquement des documents en retirant tous les noms et informations personnelles (nous reviendrons sur cet exemple d'anonymisation plus loin). De même, la traduction juridique bénéficie des progrès des modèles multilingues : on dispose désormais d'IA génératives capables de traduire un contrat du français vers l'anglais en respectant finement le ton et la terminologie juridique, avec une qualité souvent supérieure à celle des traducteurs automatiques classiques.

2. Une adoption en forte progression

Selon un sondage Ipsos réalisé début 2024, 15 % des Français avaient déjà utilisé une IA générative dans le cadre de leur travail, et pour une utilisation quotidienne pour les usages professionnels[16]. Les tâches plébiscitées par les utilisateurs incluaient justement la recherche d'informations (48 % des répondants), la rédaction de textes (38 %), la traduction (36 %) et la synthèse/le résumé (31 %). Le domaine du droit n'échappe pas à la tendance : un rapport du Sénat publié fin 2024 souligne que de nombreuses tâches juridiques, « *nonobstant une qualité encore inégale* », sont perméables à l'automatisation par ces modèles[17]. En particulier, les sénateurs citent la recherche jurisprudentielle, la synthèse de documents et la rédaction d'actes standardisés comme des applications évidentes de l'IA générative pour les juristes.

> **Un rapport du Sénat cite la recherche jurisprudentielle, la synthèse de documents et la rédaction d'actes standardisés comme des applications évidentes de l'IA générative pour les juristes.**

Il convient de noter que l'enthousiasme reste mesuré et variable selon les professions juridiques. Une enquête du Thomson Reuters Institute début 2024 montrait que les juristes d'entreprise voient comme cas d'usage numéro 1 la rédaction de contrats (88 % des juristes d'entreprise interrogés citent cette application)[18]. De leur côté, les juristes fiscalistes plébiscitent la génération de déclarations fiscales (69 % d'entre eux). Dans les cabinets d'avocats et chez les juristes de contentieux, c'est la recherche juridique qui domine les attentes envers l'IA. En somme, chaque spécialité identifie l'IA générative comme un levier de productivité pour ses tâches les plus consommatrices de temps : revue de contrats, veille jurisprudentielle, analyse de dossiers volumineux, etc. Un point commun important ressort de ces études : la relecture et l'analyse de documents figure presque systématiquement en tête des usages envisagés. Il s'agit là d'un besoin transversal du secteur juridique. La possibilité d'utiliser une IA pour passer au crible un dossier de plusieurs centaines de pages et en extraire les faits saillants ou pour comparer en un clin d'œil deux contrats afin de repérer les divergences de clauses représente un gain de productivité très concret aux yeux des premiers utilisateurs.

16. IPOS, « Etude IPSOS "l'usage de l'intelligence artificielle par les français" », 2025, https ://www.ipsos.com/sites/default/files/ct/news/documents/2025-02/ipsos-cesi-usage-intelligence-artificielle-rapport-complet.pdf.

17. Sénat, « L'intelligence artificielle générative et les métiers du droit : agir plutôt que subir », Rapport d'information n° 216, 2024, https ://www.senat.fr/rap/r24-216/r24-2162.html.

18. Thomson Reuters, « Generative AI for legal professionals : Top use cases », 2025, https ://legal.thomsonreuters.com/blog/generative-ai-for-legal-professionals-top-use-cases/.

II - Limites et enjeux des IA génératives appliquées au droit

Malgré ces promesses, les professionnels du droit gardent un œil critique sur les limites des IA génératives actuelles. La principale tient à l'incapacité de ces outils à modéliser fidèlement le raisonnement juridique humain[19]. Le droit n'est pas un simple ensemble de données textuelles : c'est un système complexe, semé d'exceptions, de subtilités de contexte, de principes hiérarchiques, que les LLM se limitant à une perception des régularités statistiques appréhendent mal. Comme l'exprime Herbert L. A. Hart, le droit est une « texture ouverte »[20] dont le sens dépend de nuances factuelles et téléologiques qu'un modèle purement statistique ne peut saisir entièrement. En d'autres termes, une phrase juridiquement correcte sur la forme peut être fausse ou inapplicable dès lors qu'on change légèrement les faits ou la finalité recherchée. Or l'IA, qui ne « comprend » pas réellement le sens, aura du mal à opérer ce type de discernement.

> Malgré ces promesses, les professionnels du droit gardent un œil critique sur les limites des IA génératives actuelles. La principale tient à l'incapacité de ces outils à modéliser fidèlement le raisonnement juridique humain.

La conséquence directe est de conduire l'IA générative à commettre des erreurs. Certaines peuvent être anecdotiques, comme une mauvaise citation d'un article de loi ou encore une confusion entre deux arrêts aux noms proches, d'autres peuvent être plus substantielles, comme l'inversion de responsabilités ou la citation d'une jurisprudence inexistante. L'une des principales difficultés, rencontrée tant par les concepteurs que par les utilisateurs, est le caractère aléatoire de la production de ces erreurs. À la différence d'un professionnel débutant dont on pourrait encadrer la progression, l'IA ne fournit pas de raisonnement explicatif de sa réponse, ce qui pose la question de l'efficacité des contrôles à mettre en œuvre. Les juristes doivent donc intégrer une étape supplémentaire de vérification systématique des productions de l'IA, au risque de propager des erreurs dans leur production. En pratique, tout contenu généré par une IA doit être validé par un professionnel du domaine avant d'en faire un plein usage. Loin de remplacer l'expertise, l'outil agit comme un assistant dont il faut sans cesse questionner le résultat.

Une autre limite est que ces modèles ne sont pas conçus pour respecter les spécificités du raisonnement juridique. Un LLM générique ignore la hiérarchie des normes, les fondements juridiques pertinents, la nécessité de motiver en droit chaque affirmation. Il peut toujours produire des argumentations juridiques incorrectes, d'autant plus si le modèle employé est généraliste. Par exemple, ChatGPT peut répondre de manière très littérale et non juridique à une question complexe, là où on pouvait s'attendre à ce qu'il applique un syllogisme juridique (majeure, mineure, conclusion). Des ajustements sont bien entendu possibles, comme la meilleure rédaction d'une requête (*prompt*) pour solliciter la mention des citations de textes de loi, ou l'entraînement d'un modèle avec des données juridiques afin qu'il dispose d'une représentation statistique du style juridique, mais aucune de ces mesures ne permet de garantir l'absence de production d'erreurs. L'IA, en réalité, ne « comprend » pas réellement les concepts juridiques, ce qui limite sa fiabilité pour des tâches demandant une interprétation fine et une mise en contexte dont elle ne dispose pas toujours.

En outre, dans un contexte où 30 % des professionnels utilisant l'IA y ont recours plusieurs fois par jour dans le cadre de leur travail[21], le recours aux IA génératives soulève d'importantes questions de confidentialité. Dès lors que les données traitées par des applications comme ChatGPT sont envoyées sur des serveurs localisés en dehors de l'Europe, la sécurité ou la réexploitation ne sont pas garanties, conduisant à de possibles fuites. Les avocats, tenus au secret professionnel, doivent donc être extrêmement prudents : anonymiser (au

19. V. par exemple Nicolas Regis, « L'intentionnalité du juge », *Archives de philosophie du droit*, 2022/1, tome 63, p. 463-476 ou encore Yannick Meneceur, *L'intelligence artificielle en procès*, Bruxelles, Bruylant, 2020, p. 97-98.
20. Herbert L. A. Hart, *Le concept de droit*, traduit par Michel Van De Kerchove, Bruxelles, Presses universitaires Saint-Louis Bruxelles, coll. « Droit », 2005.

21. IPSO op. cit.

> Le recours aux IA génératives soulève d'importantes questions de confidentialité. Cet enjeu de protection des données freine une adoption plus large dans les professions du droit, qui attendent des garanties techniques ou juridiques avant de s'engager.

sens strict du terme) les documents avant de les soumettre à une IA, vérifier les conditions d'utilisation des outils (certaines plateformes conservent par défaut les données soumises pour entraîner leurs modèles, comme OpenAI le faisait jusqu'à récemment, ce qui a conduit des entreprises à désactiver l'historique ou à opter pour des instances dédiées). Cet enjeu de protection des données freine une adoption plus large dans les professions du droit, qui attendent des garanties techniques ou juridiques avant de s'engager. L'Agence nationale de la sécurité des systèmes d'information (ANSSI) en France a émis en 2024 des recommandations de sécurité pour l'utilisation des IA génératives, insistant sur le chiffrement des données et le cloisonnement des accès pour les applications professionnelles sensibles[22]. De son côté, le Conseil national des barreaux (CNB) encourage la profession à se familiariser avec ces outils tout en respectant la déontologie, c'est-à-dire en choisissant des solutions garantissant la confidentialité ou le cas échéant en obtenant le consentement préalable éclairé du client[23].

Enfin, l'impact de l'IA sur la déontologie et la qualité du travail juridique fait débat. D'un côté, l'automatisation de certaines tâches répétitives est vue positivement par de nombreux professionnels, car elle libère du temps pour des missions à plus forte valeur ajoutée. D'un autre côté, certains craignent une forme de dévalorisation de l'expertise : si l'IA facilite trop certaines tâches, les clients ne vont-ils pas s'attendre à payer moins cher ces prestations, ou du moins à ce que l'avocat travaille plus vite ? Et surtout, le juriste ne risque-t-il pas de perdre ses réflexes à force de déléguer la réflexion à la machine ? Ce phénomène de baisse des capacités cognitives[24] par excès de confiance dans l'IA est relevé par 44 % des personnes dans l'étude d'Ipsos déjà citée. L'effet pervers, résultant d'un biais d'automatisation, serait de créer des professionnels hyper efficaces mais moins aguerris intellectuellement, enclins à accepter sans assez d'esprit critique les réponses de l'IA. Les formations insistent donc sur la nécessité pour les juristes de garder un regard critique sur les productions de l'IA et de ne jamais céder à la facilité au détriment de la rigueur.

III - Transformations du métier d'avocat et nouvelles compétences

Le Conseil national des barreaux (CNB) considère que l'emploi de plus en plus courant par les avocats des IA génératives conduit à faire évoluer leurs pratiques et à développer de nouvelles compétences, afin de tirer parti des nouvelles opportunités, tout en contrôlant les risques[25]. Ainsi, le prompt engineering, c'est-à-dire la pratique d'une écriture efficace de dialogues avec des systèmes d'IA générative, est un savoir-faire en train de se développer[26]. Dans le même temps, le CNB encourage les avocats à développer un esprit critique vis-à-vis des résultats fournis par les IA génératives : il s'agit d'avoir la capacité de déceler les éventuels biais ou erreurs dans les réponses, de recouper avec des sources sûres et de procéder à une relecture approfondie des résultats de manière systématique. La formation initiale et continue des juristes commence à intégrer ces dimensions : certains programmes de master proposent déjà des modules sur l'utilisation responsable de

22. ANSSI, « Recommandations de sécurité pour un système d'IA générative » https ://cyber.gouv.fr/publications/recommandations-de-securite-pour-un-systeme-dia-generative.
23. CNB, « Une feuille de route pour que les avocats s'emparent de l'intelligence artificielle », https ://cnb.avocat.fr/actualite/une-feuille-de-route-pour-que-les-avocats-semparent-de-lintelligence-artificielle .
24. V. par exemple Nataliya Kosmyna, Eugene Hauptmann, *et al.* « Your Brain on ChatGPT : Accumulation of Cognitive Debt when Using an AI Assistant for Essay Writing Task », *arXiv*, 10 juin 2025, accessible sur : https ://arxiv.org/abs/2506.08872.
25. V. leur guide sur l'usage de l'IA générative par les avocats, accessible sur : https ://www.cnb.avocat.fr/fr/actualites/le-cnb-accompagne-toujours-plus-la-profession-dans-lutilisation-de-lia-generative.
26. V. annexe de l'ouvrage : Yannick Meneceur, *IA générative et professionnels du droit : comprendre et s'approprier la langue des probables*, Paris, LexisNexis, 2024.

l'IA[27], les écoles d'avocats abordent les enjeux de déontologie numérique[28], etc.

Pour les professionnels favorables à la banalisation des usages de cette technologie, le rôle des juristes tendrait alors à devenir celui d'un *superviseur* du travail des IA génératives. Ainsi, pour préparer toutes sortes d'analyses juridiques, il pourrait d'abord être demandé à l'IA une ébauche à partir de certaines données pertinentes qu'il s'agirait ensuite d'affiner progressivement en précisant le contenu de la requête. L'IA générative serait à employer là où elle est performante, c'est-à-dire pour le traitement massif d'informations et la génération rapide de texte, tout en reprenant la main là où l'expertise humaine est indispensable : le choix définitif des mots, l'analyse fine et les conclusions. Cette recherche d'une bonne complémentarité entre l'humain et la machine exige une formation approfondie et de la pratique : les professionnels du droit doivent bien comprendre ce qui peut être demandé (ou non) à des IA génératives et se montrer attentifs aux biais cognitifs (automatisation, ancrage). À cet égard, on voit apparaître des référentiels ou des bibliothèques internes de *prompts* pour permettre de produire des résultats de meilleure qualité. Des cabinets de taille importante, comme FIDAL, ont choisi de développer leurs propres systèmes, adossés à leur documentation interne, afin de garantir une production conforme à leurs standards d'écriture.

L'arrivée des IA génératives amène également la profession d'avocat à reconsidérer la tarification de ses différentes prestations. Certaines tâches facturées auparavant aux clients pour un coût important (comme des recherches juridiques longues ou complexes) sont accomplies plus rapidement par certains cabinets grâce à une bonne combinaison entre moteurs de recherche et IA. D'autres aspects du travail de l'avocat sont alors valorisés, comme le conseil stratégique, une expertise pointue ou la créativité dans les solutions, qui ne sont pas automatisables. L'IA générative est donc investie comme un nouveau levier d'efficience permettant de reconfigurer la distribution du travail. Cela peut se traduire par la délégation à des collaborateurs juniors de nouvelles missions, comme l'entraînement ou la supervision des IA internes à un cabinet, tandis que les tâches purement mécaniques, comme la recherche de preuves dans des masses importantes de documents, sont assumées par des machines.

IV - Choix technologiques : solutions du marché ou développement sur mesure ?

1. Usage de solutions standardisées ou développement spécifique ?

Devant l'abondance de systèmes d'IA générative disponibles, les cabinets d'avocats hésitent principalement entre deux stratégies d'adoption. La première consiste à utiliser des solutions standardisées du marché, par exemple les fonctionnalités d'IA intégrées dans les suites bureautiques (Microsoft 365 Copilot, etc.) ou des services en ligne génériques (ChatGPT, Perplexity, etc.). Ces outils présentent l'avantage d'être « clés en main » et peu coûteux à l'entrée : ils n'exigent pas de développement spécifique et un simple abonnement, sans engagement à long terme, suffit pour bénéficier des fonctionnalités. Cependant, ils sont généralistes et ne sont pas forcément adaptés aux spécificités du droit et des professions juridiques : outre le manque de personnalisation des connaissances métier, il existe de très importants problèmes de confidentialité.

> Devant l'abondance de systèmes d'IA générative disponibles, les cabinets d'avocats hésitent principalement entre deux stratégies d'adoption. La première consiste à utiliser des solutions standardisées du marché, la seconde stratégie est celle du développement spécifique d'une IA générative sur mesure

27. V. par exemple à l'université de Strasbourg au sein des masters Cyberjustice et Droit de l'économie numérique.
28. V. par exemple à l'École de formation professionnelle des barreaux (EFB), dans le cadre de la formation continue, le module « Déontologie de l'avocat augmenté ».

La seconde stratégie est celle du développement spécifique d'une IA générative sur mesure, calibrée pour les besoins d'une organisation. C'est l'option qui a été choisie par le cabinet français Fidal, qui est l'un des plus grands cabinets d'avocats d'affaires en Europe. Fidal a lancé en 2023 son projet interne nommé FidalIA, avec l'objectif de bâtir un assistant juridique réservé aux avocats du cabinet, disposant de leurs sources internes et répondant aux exigences de confidentialité et de qualité propres à la profession[29].

> La seconde stratégie est celle du développement spécifique d'une IA générative sur mesure.

2. Avantages d'un développement spécifique : l'exemple de FidalIA

Du côté des avantages, une telle IA générative peut être adaptée finement aux besoins métier et suivre au plus près les procédures de travail internes du cabinet. Fidal a commencé, en amont du développement, par un travail de concertation en interne pour identifier les cas d'usage les plus pertinents pour ses avocats dans les différents secteurs de leur activité.

Sur cette base, différentes fonctionnalités prioritaires ont été développées comme l'anonymisation de documents clients (pour pouvoir ensuite les exploiter sans risque dans d'autres modules d'IA), un robot conversationnel juridique sécurisé capable de répondre aux questions des avocats en interrogeant à la fois les sources publiques (codes, lois sur Légifrance) et les sources internes du cabinet (base de modèles d'actes, notes internes), la possibilité d'ajouter à la volée des documents d'un dossier pour poser des questions dessus, ou encore la traduction automatique intégrée. FidalIA permet aux avocats du cabinet de « glisser-déposer » un document dans l'interface (contrat reçu de la partie adverse, par exemple) et de poser des questions à l'IA à son sujet. Il doit être précisé que le document est utilisé uniquement pour le traitement demandé, puis il est supprimé immédiatement du serveur afin de répondre aux besoins de confidentialité : aucune donnée sensible n'est ainsi stockée durablement ou transmise à un tiers.

De même, FidalIA comporte une bibliothèque de *prompts* prérédigés, adaptés aux tâches juridiques courantes (comme vérifier la conformité au RGPD d'un texte ou analyser les faits d'un dossier contentieux), ce qui permet à ses utilisateurs de disposer de modèles déjà éprouvés, sur la base desquels ils peuvent ensuite développer leur propre pratique. L'autre atout d'une telle solution interne est l'intégration des données privées de l'organisation. Fidal a ainsi pu entraîner son IA avec des milliers de documents internes (actes juridiques, consultations, contrats types) préalablement anonymisés, et les croiser avec un large corpus de données publiques pertinentes. Un outil du marché n'aurait eu accès qu'aux données publiques ou aux bases intégrées par son éditeur, sans possibilité d'exploitation approfondie de l'expérience du cabinet.

3. Inconvénients d'un développement spécifique

Du côté des inconvénients, développer une IA générative en interne représente un investissement conséquent. Fidal a dû s'associer à des partenaires technologiques externes[30] et mobiliser une équipe dédiée pour bâtir son outil. Cela implique du temps (spécifiquement ici plusieurs mois de développement et de tests), un important investissement financier et une certaine prise de risque quant au résultat. De plus, maintenir en condition opérationnelle une solution sur mesure exige d'avoir en interne ou à disposition des compétences pointues en *data science*, en génie logiciel, en cybersécurité, etc., afin de faire évoluer le produit, l'adapter aux nouveaux modèles de langage et corriger les inévitables dysfonctionnements. Enfin, un outil propriétaire peut souffrir de lacunes temporaires en ne bénéficiant pas immédiatement de toutes les avancées disponibles sur le marché.

Par exemple, la version 1 de FidalIA ne comportait pas certaines fonctionnalités annexes du fait de contraintes techniques initiales, là où un outil commercial intégré les aurait

29. Ces développements s'appuient sur une présentation réalisée par Maître Aurélie Klein, lors d'un atelier à l'IRB le 14 février 2025.

30. Notamment Sopra Steria.

probablement incorporées par défaut. Il faut donc accepter d'améliorer par itérations une IA générative interne, en priorisant ce qui compte vraiment pour l'usage local.

4. Un choix à opérer en fonction du contexte

D'après un rapport de la RAND (pour *Research AND Development*, qui est un *think tank* états-unien), quatre projets d'IA sur cinq échoueraient dans les organisations[31]. Dans ce contexte, le retour d'expérience de FidalIA paraît très instructif : d'après Aurélie Klein, avocate responsable de l'innovation au sein du cabinet, ses confrères paraissent rassurés de disposer d'un outil maîtrisé en interne, plutôt que de recourir à un service grand public flou sur le traitement des données. Ils apprécient aussi que l'IA produise des textes dans un style parfaitement aligné sur celui du cabinet, grâce à un *prompt* initial, transparent pour les utilisateurs, configuré pour calibrer le ton et le niveau de langue.

S'il semble encore trop tôt pour porter une appréciation objectivée sur cette initiative, le cabinet Fidal semble avoir identifié l'essentiel des problématiques auxquelles les organisations sont aujourd'hui confrontées avec l'adoption de systèmes d'IA générative : exigences de confidentialité, politique de conduite du changement adaptée, recherche de limitation des biais et des « hallucinations ». Le choix entre une solution sur étagère ou un développement spécifique dépendra toutefois de la taille de la structure, de ses moyens et de ses objectifs : pour un petit cabinet, utiliser intelligemment des outils standards du commerce (éventuellement en payant pour des versions professionnelles garantissant la confidentialité et adaptées à la pratique juridique) sera sans doute plus réaliste que de vouloir créer son propre système maison ; pour de grandes structures ou des services publics, la voie d'un développement dédié pourra se justifier, afin de répondre aux exigences spécifiques (droit sectoriel pointu, secret professionnel, multilinguisme, etc.) mieux que ne le ferait une solution généraliste.

> « L'essentiel des problématiques auxquelles les organisations sont aujourd'hui confrontées avec l'adoption de systèmes d'IA générative : exigences de confidentialité, politique de conduite du changement adaptée, recherche de limitation des biais et des « hallucinations ».

31. James Ryseff, Brandon F. De Bruhl, Sydne J. Newberry, « The Root Causes of Failure for Artificial Intelligence Projects and How They Can Succeed – Avoiding the Anti-Patterns of AI », RAND, Research Report, 13 août 2024, accessible sur : https ://www.rand.org/pubs/research_reports/RRA2680-1.html.

Atelier n°3
L'encadrement juridique de l'emploi de l'IA dans le champ de la justice

- Présentation des nouveaux instruments juridiques contraignants (RIA et convention-cadre)
- Articulation des nouveaux instruments juridiques avec la protection des données à caractère personnel
- Contraintes pour le déploiement des IA génératives dans le domaine de la justice
- Évolution des cadres juridiques de responsabilité

L'essor rapide des systèmes d'intelligence artificielle s'accompagne d'une transformation profonde des cadres juridiques. De nouveaux instruments contraignants émergent. Ils structurent désormais l'encadrement des usages. Le règlement européen sur l'intelligence artificielle et la convention-cadre du Conseil de l'Europe en constituent les piliers. Leur articulation marque un changement d'échelle. Elle combine logique de marché, protection des droits fondamentaux et affirmation de principes démocratiques.

Ces nouveaux textes ne s'inscrivent pas en rupture. Ils prolongent et complètent des cadres existants, au premier rang desquels le RGPD. Ils introduisent toutefois des logiques inédites. L'approche par les risques, la régulation des cycles de vie des systèmes et la prise en compte de leur dimension évolutive redéfinissent les méthodes juridiques traditionnelles. Le droit s'adapte. Mais il se trouve aussi mis à l'épreuve par des technologies dont les effets sont incertains et parfois difficiles à anticiper.

Dans ce contexte, la question de la responsabilité devient centrale. Les régimes existants apparaissent partiellement inadaptés : difficulté d'établir un lien de causalité, multiplication des acteurs impliqués, opacité des systèmes, imprévisibilité des résultats. Autant d'éléments qui fragilisent les schémas classiques d'imputation. Le développement de l'intelligence artificielle impose ainsi de repenser en profondeur les équilibres entre innovation, sécurité juridique et protection des droits.

I - Présentation des nouveaux instruments juridiques contraignants (RIA et convention-cadre)

1. Le règlement européen sur l'intelligence artificielle (RIA)

Le règlement (UE) 2024/1689 du Parlement européen et du Conseil du 13 juin 2024 établissant des règles harmonisées concernant l'intelligence artificielle (RIA) est présentée par l'Union européenne comme la première législation adoptée au monde sur l'intelligence artificielle. Publié au Journal officiel de l'Union européenne le 12 juillet 2024, ce texte est entré en vigueur le 2 août 2024, avec une application graduelle de ses dispositions jusqu'en 2027[32].

L'objectif principal du RIA est d'améliorer le fonctionnement du marché intérieur en établissant un cadre juridique uniforme (art.114 TFUE[33]), tout en promouvant l'adoption d'une IA « axée sur l'humain et digne de confiance » et en garantissant un niveau élevé de protection des droits fondamentaux.

Bien que traitant également de la protection de droits fondamentaux (notamment la protection des données à caractère personnel, art.16

32. Règlement (UE) 2024/1689 du Parlement européen et du Conseil établissant des règles harmonisées concernant l'intelligence artificielle et modifiant les règlements (CE) n° 300/2008, (UE) n° 167/2013, (UE) n° 168/2013, (UE) 2018/858, (UE) 2018/1139 et (UE) 2019/2144 et les directives 2014/90/UE, (UE) 2016/797 et (UE) 2020/1828 (règlement sur l'intelligence artificielle)
33. Traité sur le fonctionnement de l'Union européenne.

> Le RIA introduit une classification des systèmes d'IA modulant l'intensité des contraintes réglementaires selon leur niveau de risque. Le RIA introduit une catégorie distincte pour les « modèles d'IA à usage général » (notamment les IA génératives), avec des obligations concernant la transparence, la documentation et l'atténuation des risques systémiques.

TFUE), l'approche du RIA est basée, comme tous les textes composant la législation sur les produits au sein de l'Union européenne, sur le nouveau cadre législatif de l'Union (*new legislative framework*) de 2008[34]. En ce sens, le RIA introduit une classification des systèmes d'IA modulant l'intensité des contraintes réglementaires selon leur niveau de risque.

- **Risque inacceptable** : le RIA interdit certaines pratiques d'IA jugées contraires aux valeurs européennes, comme la notation sociale, l'exploitation de vulnérabilités, ou encore la reconnaissance des émotions sur le lieu de travail. L'identification biométrique à distance à des fins répressives fait l'objet d'un régime strict d'encadrement.
- **Haut risque :** cette catégorie englobe les systèmes d'IA pouvant porter atteinte à la sécurité ou aux droits fondamentaux des personnes. Ces systèmes sont soumis à des exigences renforcées en matière d'évaluation de conformité et de documentation technique.
- **Risque spécifique en matière de transparence :** certains systèmes d'IA sont soumis à des obligations de transparence spécifiques, notamment les systèmes qui interagissent avec des humains.
- **Absence de risque ou risque minimal :** pour tous les autres systèmes d'IA, le règlement ne prévoit pas d'obligation spécifique. Les autres cadres réglementaires, comme le RGPD, demeurent toutefois pleinement applicables.

Le RIA introduit également une catégorie distincte pour les « modèles d'IA à usage général » (notamment les IA génératives), avec des obligations concernant la transparence, la documentation et l'atténuation des risques systémiques.

Sur le plan opérationnel, le RIA distingue plusieurs catégories d'acteurs, avec des obligations spécifiques pour chacun : fournisseurs, opérateurs de déploiement, importateurs, distributeurs et utilisateurs.

Le calendrier d'application est échelonné : les interdictions pour les systèmes d'IA présentant des risques inacceptables sont entrées en vigueur le 2 février 2025, les règles concernant les modèles d'IA à usage général sont devenues applicables le 2 août 2025, tandis que les dispositions relatives aux systèmes à haut risque s'appliqueront à partir du 2 août 2026 ou 2027 selon les cas.

Marina Teller, professeure de droit, rappelle qu'il s'agit avant tout « d'une régulation marchande, structurée autour de l'article 114 du traité sur le fonctionnement de l'Union européenne, dont l'objectif principal est l'harmonisation du marché intérieur », précisant que ce texte « reste peu armé pour prendre en charge les enjeux liés aux droits fondamentaux, à la démocratie et à l'éthique »[35]. Bertrand Cassar, responsable gouvernance des données du groupe La Poste, estime quant à lui que « ce maillage réglementaire, bien qu'ambitieux, constitue un véritable mur réglementaire qui rend son appropriation difficile pour les organisations publiques et privées ».

2. La convention-cadre du Conseil de l'Europe sur l'intelligence artificielle

Parallèlement au RIA, le Conseil de l'Europe a accueilli entre 2019 et 2024 les travaux et les négociations d'une convention-cadre sur l'intelligence artificielle et les droits de l'Homme, la démocratie et l'État de droit. Adoptée le 17 mai 2024 et ouverte à la signature le 5 septembre 2024 à Vilnius, cette convention est

34. Cet ensemble législatif, adopté en 2008, regroupe le règlement (CE) n° 765/2008 fixant les prescriptions relatives à l'accréditation et à la surveillance du marché des produits, la décision 768/2008 relative à un cadre commun pour la commercialisation des produits, qui comprend des dispositions de référence à incorporer dans les révisions de la législation sur les produits et le règlement (UE) 2019/1020 sur la surveillance du marché et la conformité des produits.

35. L'ensemble des citations du document de Marina Teller et Bertrand Cassar sont issues de l'atelier tenu à l'IRB le 11 avril 2025.

> Premier instrument international juridiquement contraignant dans le domaine de l'IA. la convention-cadre vise explicitement à garantir que les activités liées aux systèmes d'intelligence artificielle soient « pleinement compatibles avec les droits humains, la démocratie et l'État de droit, tout en étant propices au progrès et aux innovations technologiques ».

présentée par le Conseil de l'Europe comme le premier instrument international juridiquement contraignant dans le domaine de l'IA. Ce traité international a déjà été signé par l'Union européenne (avec des effets juridiques au sein de ses 27 États membres), plusieurs pays européens non-membres de l'Union mais membres du Conseil de l'Europe (Andorre, Géorgie, Islande, Liechtenstein, Monténégro, Norvège, République de Moldavie, Royaume-Uni, Saint-Marin, Suisse) ainsi que par des pays non-européens (Canada, États-Unis, Israël, Japon).

L'élaboration de la convention-cadre a impliqué un large éventail de parties prenantes. Les travaux ont débuté dès 2019 avec le Comité *ad hoc* sur l'intelligence artificielle (CAHAI), suivi en 2022 par le Comité sur l'intelligence artificielle (CAI), chargé de la rédaction et de la négociation. Ce processus a eu pour ambition d'intégrer des perspectives diverses, notamment celles de la société civile, pour renforcer la légitimité et la pertinence du texte final.

Conformément au mandat du Conseil de l'Europe et de manière complémentaire au RIA, qui adopte principalement une approche de régulation du marché, la convention-cadre vise explicitement à garantir que les activités liées aux systèmes d'intelligence artificielle soient « pleinement compatibles avec les droits humains, la démocratie et l'État de droit, tout en étant propices au progrès et aux innovations technologiques ». Elle revendique adopter une approche neutre sur le plan technologique afin de résister au temps.

Même si les deux textes sont distincts, la Commission européenne a assuré, par sa participation lors des négociations au Conseil de l'Europe entre 2022 et 2024, la cohérence juridique entre les initiatives. Ainsi, plusieurs concepts clés se retrouvent dans le RIA et la convention-cadre, tels que l'approche fondée sur les risques, les principes d'une IA digne de confiance (transparence, robustesse, sécurité) et le soutien à une innovation qualifiée de « sûre ». Cette convergence conceptuelle vise à faciliter une mise en œuvre cohérente des deux instruments et à ne pas fragmenter le cadre juridique relatif à l'IA au sein de l'Union.

3. L'innovation totale : un concept pour repenser la régulation de l'IA

Face à ces nouvelles régulations, Marina Teller propose une grille de lecture originale à travers le concept d'« innovation totale ». S'inspirant du concept de « fait social total » développé par l'anthropologue Marcel Mauss, elle suggère que certaines innovations technologiques constituent des innovations systémiques qui transforment « non seulement les pratiques professionnelles, mais aussi les fondements mêmes de la régulation juridique, du rapport à la décision, de la souveraineté numérique et de l'organisation sociale ».

Selon Marina Teller, une innovation totale se caractérise par plusieurs marqueurs : « son effet de masse mondial, sa rapidité d'adoption, sa dualité (civile et militaire), son irréversibilité et son opacité technologique ». Elle précise que « tout n'entre pas dans le champ des innovations totales », mais que certaines technologies comme le nucléaire, l'intelligence artificielle ou l'ordinateur quantique partagent ces caractéristiques.

Cette conceptualisation permet de mettre en lumière les limites de l'approche actuelle de la régulation. Marina Teller suggère qu'au lieu d'une approche par les risques, il aurait plutôt fallu développer « une approche par les bénéfices » obligeant les acteurs à démontrer, avant déploiement, les apports sociétaux, écologiques ou démocratiques de leurs systèmes d'IA.

Le concept d'innovation totale invite également à repenser le rapport au temps dans l'élaboration juridique. Comme le souligne Marina Teller, « Il y a un défi à savoir comment réguler aujourd'hui une innovation qui n'existe pas encore ». Cette tension entre la rapidité des innovations technologiques et le rythme plus lent de l'élaboration juridique constitue l'un des défis majeurs pour l'encadrement de l'IA.

II - Articulation des nouveaux instruments juridiques avec la protection des données à caractère personnel

1. Complémentarités entre le RIA et le RGPD

L'articulation entre le RIA et le règlement général sur la protection des données (RGPD) constitue un enjeu majeur pour les organisations développant ou déployant des systèmes d'IA. Ces deux textes s'appliquent conjointement lorsque des données à caractère personnel sont traitées par des systèmes d'IA, créant ainsi un cadre réglementaire complexe mais complémentaire.

Comme le précise la Commission nationale de l'informatique et des libertés (CNIL), « le RIA est très clair sur ce point : il ne remplace pas les exigences du RGPD, il les complète et prend le relais du RGPD sur certains points bien définis »[36]. Le RGPD s'applique à tous les traitements de données personnelles, tandis que le RIA s'adresse spécifiquement aux systèmes et modèles d'IA.

L'articulation entre ces deux textes peut être analysée à plusieurs niveaux.

- **Champ d'application :** le RGPD s'applique dès lors que des données personnelles sont traitées, indépendamment de la technologie utilisée. Le RIA, quant à lui, s'applique aux systèmes d'IA selon leur niveau de risque, qu'ils traitent ou non des données personnelles[37]. Certains systèmes d'IA peuvent donc être soumis uniquement au RIA, uniquement au RGPD, ou aux deux textes simultanément.
- **Gouvernance et contrôle :** le RIA prévoit la création d'autorités nationales de contrôle, qui pourront être les mêmes que celles désignées pour le RGPD. Par exemple, la Commission nationale pour la protection des données (CNPD) au Luxembourg a indiqué que le RIA « impactera sans doute [ses] compétences et nécessitera une coopération accrue entre [elle] et différents régulateurs, tant au niveau national qu'européen »[38].
- **Obligations de transparence et de documentation :** les deux règlements prévoient des obligations complémentaires de transparence et de documentation. Le RGPD exige des informations spécifiques sur les données à caractère personnel, tandis que le RIA impose des exigences spécifiques en matière de documentation technique pour les systèmes d'IA à haut risque.
- **Analyse d'impact :** le RGPD prévoit une analyse d'impact relative à la protection des données (AIPD) pour les traitements à risque élevé. Le RIA impose une évaluation de conformité pour les systèmes d'IA à haut risque. Ces analyses peuvent être complémentaires et potentiellement mutualisées.

Bertrand Cassar met en évidence une différence majeure entre ces deux cadres réglementaires : « On se retrouve à la fois avec cette démarche de mise en conformité et avec une approche juridique, mais on doit aussi avoir un moyen de contrôler ce système d'IA tout au long de son cycle de vie, parce qu'un système d'IA se dégrade au cours du temps ». Cette dimension temporelle et évolutive constitue une spécificité du RIA par rapport au RGPD, nécessitant le développement d'outils techniques de suivi et d'audit.

« On se retrouve à la fois avec cette démarche de mise en conformité et avec une approche juridique ». Cette dimension temporelle et évolutive constitue une spécificité du RIA par rapport au RGPD, nécessitant le développement d'outils techniques de suivi et d'audit.

2. Les données personnelles dans le contexte des IA génératives

Les IA génératives posent des défis particuliers en matière de protection des données personnelles. Ces systèmes sont généralement entraînés avec d'immenses corpus de données incluant potentiellement des informations personnelles et peuvent générer du contenu révélant des informations personnelles, parfois de manière indirecte ou inattendue.

36. CNIL, https ://www.cnil.fr/fr/entree-en-vigueur-du-reglement-europeen-sur-lia-les-premieres-questions-reponses-de-la-cnil.
37. https ://www.osborneclarke.com/insights/interplay-eu-ai-act-and-gdpr.
38. https ://cnpd.public.lu/fr/actualites/international/2024/07/ai-act.html.

Plusieurs problématiques spécifiques sont à souligner.

- **L'origine des données d'entraînement :** les questions relatives aux droits d'auteur et au consentement se posent avec acuité pour les IA génératives. Comme l'illustre la décision américaine *Thomson Reuters Enterprise c/ ROSS Intelligence, Inc.* du 2 février 2025, l'utilisation de données protégées pour l'entraînement d'IA peut être considérée comme déloyale lorsqu'elle vise à concurrencer directement le détenteur des droits. Le juge a notamment considéré qu'il existe « un marché potentiel des données d'entraînement d'IA », ouvrant la voie à une potentielle monétisation des données utilisées pour développer des systèmes d'IA39.
- **La réidentification des personnes :** les IA génératives peuvent parfois reproduire des informations personnelles présentes dans leurs données d'entraînement, posant ainsi des risques de divulgation non autorisée.
- **Le droit à l'effacement :** l'application du droit à l'effacement prévu par l'article 17 du RGPD s'avère particulièrement complexe dans le contexte des IA génératives. Comment garantir qu'une information personnelle soit effectivement « oubliée » par un modèle d'IA une fois celui-ci entraîné ?
- **La transparence algorithmique :** l'opacité inhérente aux modèles d'IA générative complique la mise en œuvre des principes de transparence et d'explicabilité exigés tant par le RGPD que par le RIA.
- **Le consentement et la finalité :** le principe de limitation des finalités du RGPD peut être difficile à appliquer aux IA génératives, dont les usages sont multiples et parfois imprévisibles au moment de la collecte des données.

III - Contraintes pour le déploiement des IA génératives dans le domaine de la justice

1. L'administration de la justice comme domaine à haut risque

L'annexe III (point 8) du RIA identifie explicitement les systèmes d'IA employés « pour la recherche et l'interprétation de faits ou de la loi » dans le domaine de l'administration de la justice comme étant à haut risque. Cette classification emporte des conséquences significatives, s'agissant de l'examen préalable de conformité et de la documentation (art.6 et s. du RIA).

> « Les systèmes d'IA employés « pour la recherche et l'interprétation de faits ou de la loi » dans le domaine de l'administration de la justice devront respecter l'ensemble des obligations prévues pour les systèmes à haut risque.

Ainsi ces systèmes devront respecter l'ensemble des obligations prévues pour les systèmes à haut risque, incluant notamment :

- la mise en place d'un système d'évaluation et de gestion des risques tout au long du cycle de vie du système,
- une gouvernance stricte des données d'entraînement, de validation et de test,
- l'élaboration d'une documentation technique complète et précise,
- la mise en œuvre de mécanismes de journalisation des événements pour assurer la traçabilité,
- une transparence renforcée de la part des acteurs déployant ces outils quant au fonctionnement et aux risques et limitations du système,
- la garantie d'un contrôle humain effectif,
- des mesures assurant l'exactitude, la robustesse et la cybersécurité du système.

Bertrand Cassar synthétise ainsi ces mesures : « tout système d'IA à haut risque doit être soumis à une procédure d'évaluation de conformité avant sa mise sur le marché, impliquant la constitution d'un dossier technique complet, la mise en place d'un système de gestion de la qualité et, dans certains cas, la certification par un organisme tiers ».

Ces exigences visent à garantir que les systèmes d'IA les plus sensibles utilisés dans le domaine judiciaire respectent les principes d'un procès équitable.

39. https ://www.uggc.com/affaire-thomson-reuters-c-ross-intelligence/.

2. Problématiques techniques et organisationnelles

L'emploi d'IA génératives dans le domaine de la justice soulève un certain nombre de problématiques techniques et organisationnelles, allant au-delà du respect du cadre juridique.

Une des premières problématiques majeures est relative à la qualité et à la mise à jour des données utilisées par ces systèmes. Les modèles d'IA générative peuvent avoir plusieurs mois, voire plusieurs années de retard sur le droit en vigueur, ce qui constitue un obstacle majeur dans un contexte juridique où la précision et l'actualité de l'information sont cruciales. Un système d'IA fournissant des références juridiques obsolètes peut conduire à des erreurs d'appréciation aux conséquences graves.

La deuxième problématique est relative à la production de corrélations fallacieuses (« hallucinations »), c'est-à-dire la génération d'informations factuelles ou juridiques fausses présentées de manière convaincante par le système. Ce risque est particulièrement problématique dans le domaine juridique, où la précision est essentielle. Un cas emblématique évoqué lors de l'atelier concerne des avocats américains sanctionnés pour avoir soumis à un tribunal des citations jurisprudentielles fictives générées par ChatGPT. Ce type d'incident illustre la nécessité d'un contrôle humain rigoureux des productions des IA génératives dans un contexte judiciaire.

Une troisième problématique, organisationnelle, a été évoquée lors de l'atelier à l'occasion de la séquence d'ouverture présentant des actualités : l'expérience australienne de déploiement de Copilot auprès de 5 000 agents publics a révélé en effet plusieurs points d'attention :

- une utilisation modérée des outils d'IA, malgré leur disponibilité,
- des gains de productivité réels mais limités par la nécessité de vérification,
- un décalage entre le style de langue généré par l'IA et celui attendu dans l'administration,
- des bénéfices inattendus en matière d'inclusion, notamment pour les personnes neurodiverses ou non-anglophones,
- des risques liés à l'accès non contrôlé à des documents sensibles.

Cette expérience souligne l'importance d'une évaluation rigoureuse avant, pendant et après le déploiement de systèmes d'IA dans l'administration publique et *a fortiori* dans le domaine judiciaire où les enjeux de confidentialité sont particulièrement sensibles.

3. Acceptabilité sociale des algorithmes dans la justice

Au-delà des aspects techniques et réglementaires, le déploiement des IA génératives dans le domaine judiciaire soulève la question de l'acceptabilité sociale. Marina Teller défend « l'idée que le droit ne peut se limiter à créer des obligations techniques de transparence, mais doit intégrer une véritable dimension participative »[40].

> « Au-delà des aspects techniques et réglementaires, le déploiement des IA génératives dans le domaine judiciaire soulève la question de l'acceptabilité sociale.

Selon elle, « le citoyen reste trop souvent exclu des processus de conception, de déploiement et d'évaluation des systèmes algorithmiques qui structurent pourtant des pans entiers de la vie publique ». Cette exclusion est particulièrement problématique dans le domaine judiciaire, où la légitimité des décisions repose en grande partie sur la confiance des justiciables dans le système.

Marina Teller appelle à « un principe de minimisation de l'usage des algorithmes dans l'administration, sur le modèle du principe de minimisation des données dans le RGPD », et plaide pour « des mécanismes de délibération démocratique autour de ces technologies ». Cette approche rejoint sa proposition d'une régulation par les bénéfices plutôt que par les risques, qui impliquerait de démontrer la plus-value sociétale des systèmes d'IA avant leur déploiement.

L'acceptabilité sociale des algorithmes dans la justice repose également sur des garanties de transparence et d'explicabilité. Le RIA prévoit des obligations spécifiques en ce sens pour les systèmes à haut risque, mais leur mise en œuvre pratique dans le contexte des

40. Atelier IRB, 11 avril 2025.

IA génératives, souvent qualifiées de « boîtes noires », reste un défi majeur. Comment expliquer de manière compréhensible le raisonnement d'un système d'IA générative qui a produit une analyse juridique ou une recommandation ?

> L'acceptabilité sociale des algorithmes dans la justice repose également sur des garanties de transparence et d'explicabilité. Le contrôle humain constitue un autre élément essentiel de l'acceptabilité sociale.

Le contrôle humain constitue un autre élément essentiel de l'acceptabilité sociale. L'article 14 du RIA exige que les systèmes d'IA à haut risque soient conçus de manière à pouvoir être efficacement supervisés par des personnes physiques. Dans le domaine judiciaire, cette exigence prend une dimension particulière : il s'agit de garantir que le juge ou le professionnel du droit conserve la maîtrise de la décision, l'IA n'intervenant qu'en support de son raisonnement.

IV - Évolution des cadres juridiques de responsabilité

1. Les lacunes actuelles en matière de responsabilité

L'une des questions juridiques à traiter par les organisations intégrant des IA génératives dans leurs processus est celle de la responsabilité en cas de dommages. Le RIA, s'il établit un cadre pour la mise sur le marché et l'utilisation des systèmes d'IA, n'a pas pour vocation de traiter des questions de responsabilité civile ou pénale en cas de préjudice.

Pour répondre spécifiquement à cette question en matière civile, l'Union européenne a tout d'abord procédé en 2024 à la réforme du régime de responsabilité du fait des produits défectueux pour l'adapter, au sens large, aux logiciels (dont l'IA)[41]. D'autre part, la Commission européenne a publié en septembre 2022 une proposition de directive adaptant le régime de responsabilité civile extra-contractuelle aux spécificités des systèmes d'IA[42]. Cette proposition visait à instaurer un régime de responsabilité civile pour faute spécifique, avec notamment un renversement de la charge de la preuve au profit des victimes, en partant du postulat que ces dernières ne disposent pas des moyens suffisants pour prouver la faute des fournisseurs d'IA.

Cette proposition de directive a toutefois été retirée par la Commission en février 2025, dans le cadre d'un effort de simplification réglementaire[43]. Ce retrait, sans laisser un vide juridique puisque le régime de droit commun continue de s'appliquer, va complexifier la recherche de responsabilité en cas de dommage causé par une IA générative. Les régimes de responsabilité de droit commun sont en effet mal adaptés aux spécificités des systèmes d'IA, pour plusieurs raisons.

- **L'établissement d'un lien de causalité :** comment démontrer qu'un préjudice résulte directement du fonctionnement d'un système d'IA, surtout lorsque le modèle au cœur de ce fonctionnement n'est pas explicable ?
- **L'identification du responsable :** dans la chaîne complexe des acteurs impliqués (développeurs, fournisseurs, opérateurs, utilisateurs), qui doit être tenu pour responsable d'un dommage ?
- **La prévisibilité du dommage :** les systèmes d'IA générative peuvent produire des résultats inattendus. Dans quelle mesure cette imprévisibilité peut-elle exonérer les développeurs de leur responsabilité ?
- **La responsabilité de l'utilisateur humain :** quelle est la responsabilité de l'utilisateur qui se fie à une IA générative sans vérifier ses productions ?

Bertrand Cassar a souligné lors de l'atelier cette complexité : « Si on doit prendre entre 3 ou 4 pays, quand on suit toute la chaîne contractuelle, en plus si l'un de ces acteurs est

41. Directive (UE) 2024/2853 du Parlement et du Conseil du 23 octobre 2024 relative à la responsabilité du fait des produits défectueux et abrogeant la directive 85/374/CEE du Conseil.
42. COM(2022) 496 final, Proposition de directive relative à l'adaptation des règles en matière de responsabilité civile extracontractuelle au domaine de l'intelligence artificielle, 28 septembre 2022.
43. V. le programme de travail 2025 de la Commission européenne : https ://commission.europa.eu/document/download/f80922dd-932d-4c4a-a18c-d800837fbb23_en ?filename=COM_2025_45_1_EN.pdf ; commentaire : https ://www.solutions-numeriques.com/avis-de-recherche-ou-est-passee-la-directive-sur-la-responsabilite-de-lia/.

hors-européen, avec l'idée du mandataire, de l'importateur, plus le fournisseur tiers, remonter cette chaîne est très compliqué ».

2. Impacts sur les professions juridiques et judiciaires

L'essor des IA génératives transforme profondément les métiers du droit et de la justice, entraînant l'émergence de nouvelles obligations déontologiques et la nécessité d'acquérir de nouvelles compétences.

Sur le plan déontologique, les professionnels du droit font face à plusieurs enjeux majeurs.

- **Le respect du secret professionnel :** l'utilisation d'IA génératives « grand public » soulève des questions de confidentialité, puisque les données saisies peuvent être stockées et réutilisées. Certains professionnels ont recours à des systèmes non validés par leur organisation (« Shadow AI »), au risque de compromettre le secret professionnel.
- **Le devoir de compétence :** les professionnels du droit ont l'obligation de maîtriser les outils qu'ils utilisent. Cela implique de comprendre les limites des IA génératives, notamment le risque de corrélations fallacieuses (« hallucinations »), et de vérifier systématiquement leurs productions.
- **Le devoir d'information :** une obligation de transparence envers les clients ou les justiciables sur l'utilisation d'IA génératives commence à émerger. Certains tribunaux exigent désormais que toute utilisation d'IA générative pour préparer des actes soit divulguée.
- **La responsabilité des productions :** les professionnels du droit qui utilisent des IA génératives restent pleinement responsables du contenu final qu'ils produisent, même si celui-ci a été partiellement généré par une IA.

Ces nouvelles exigences déontologiques s'accompagnent d'un besoin de formation et de renouvellement partiel de la réflexion sur les programmes d'enseignement dans les universités de droit, les écoles d'avocats, les écoles de formation initiale et continue de la justice (ENM, ENG, ENPJJ, ENAP). Parmi les compétences émergentes, figurent notamment la capacité à interagir efficacement avec les agents conversationnels (le « *prompt engineering* »), l'esprit critique appliqué aux productions des IA afin de limiter les biais d'automatisation, la compréhension des principes de fonctionnement des modèles d'IA générative et la connaissance du cadre réglementaire applicable aux IA.

Cette évolution s'accompagne également de l'émergence de nouveaux rôles au sein des organisations juridiques, comme des postes dédiés à l'innovation ou à la gestion de projets d'IA.

L'impact sur l'organisation du travail est également significatif. Si certaines tâches sont automatisées, comment redéployer le temps humain ainsi libéré ? Est-ce que le temps de vérification n'est pas supérieur au temps d'exécution d'une tâche sans IA ? Plusieurs scénarios sont à considérer : augmentation du volume de dossiers traités, montée en gamme des services, ou redéfinition des missions des collaborateurs, certains pouvant devenir des « référents IA » au sein de leur organisation.

3. Vers une approche équilibrée entre innovation et protection

Face aux nouveaux enjeux créés par les IA génératives dans le domaine juridique et judiciaire, l'enjeu est de trouver un équilibre entre l'encouragement à l'innovation et la protection des droits fondamentaux et des valeurs démocratiques.

> L'enjeu est de trouver un équilibre entre l'encouragement à l'innovation et la protection des droits fondamentaux et des valeurs démocratiques.

Le cadre réglementaire européen, composé du RIA et de la convention-cadre du Conseil de l'Europe, constitue une première tentative d'établir cet équilibre. Toutefois, ce cadre présente encore des lacunes, notamment en matière de responsabilité civile à la suite du retrait de la proposition de directive spécifique.

Marina Teller soutient une approche plus axée sur les bénéfices que sur les risques, proposant d'« avoir une sorte de grammaire intellectuelle pour dire très bien, ça c'est l'horizon de la technologie, faites-nous la démonstration *ex ante* qu'on a des bénéfices pour le consommateur, des bénéfices écologiques, des bénéfices

pour le bien-être social dans son ensemble ». Cette approche permettrait selon elle de mieux prendre en compte les dimensions éthiques et sociétales des IA génératives.

Le développement de « bacs à sable réglementaires », inclus dans le RIA, représente une piste intéressante pour concilier innovation et protection. Ces environnements contrôlés permettent d'expérimenter des solutions innovantes d'IA dans un cadre dérogatoire temporaire, tout en maintenant des garanties essentielles.

La question de la souveraineté numérique se pose également avec acuité. De nombreuses organisations juridiques françaises utilisent des solutions d'IA génératives développées par des entreprises américaines, ce qui soulève des enjeux de confidentialité et d'indépendance stratégique. Le cabinet Fidal a fait le choix de développer sa propre solution d'IA pour « conserver la maîtrise de ses données et garantir le respect du secret professionnel »[44].

Enfin, la dimension internationale de la régulation des IA génératives ne peut être négligée. Si l'Europe fait figure de pionnière avec son cadre réglementaire, d'autres régions du monde adoptent des approches différentes, parfois plus permissives. La convention-cadre du Conseil de l'Europe, ouverte à la signature d'États non-européens, représente une avancée importante vers une convergence réglementaire mondiale.

44. V. *supra* II - Articulation des nouveaux instruments juridiques avec la protection des données à caractère personnel.

DEUXIÈME PARTIE
Ateliers d'approfondissement

Olivier CHEVET
Responsable d'études et de recherches à l'Institut Robert Badinter

Atelier n°4
De la voix au texte : la reconnaissance vocale à l'épreuve des exigences juridiques et judiciaires

- Clés pour comprendre la transcription automatique : faire face à la rareté des corpus
- Le futur de la transcription automatique : dépasser des cas difficiles fréquents dans l'environnement judiciaire
- Chaînes de traitements et d'édition, leçons de mise en œuvre

La reconnaissance de la parole compte sans aucun doute parmi les concrétisations les plus fascinantes de l'intelligence artificielle. Magistralement mise en scène dans les échanges entre David Bowman et l'ordinateur HAL dans *2001 Odyssée de l'espace* dès 1968, puis 35 ans plus tard dans le film *Her*, ces deux œuvres illustrent la puissance de la conversation orale comme un nouveau paradigme d'interaction avec l'univers numérique.

Devenue réalité ces dernières années, ces capacités conversationnelles sont la réunion de deux champs de recherche longtemps restés relativement isolés : la synthèse et la reconnaissance de la parole. Cette dernière, omniprésente au point de devenir banale, semble sur le point de perdre son étrangeté, qui faisait qu'on la considérait comme une expression de l'intelligence machinique, pour se voir reléguer dans l'imaginaire collectif au rang de simple automatisme. Cet atelier sera toutefois l'occasion de mesurer tout le chemin restant à parcourir.

Pour André Leroi-Gourhan, l'enregistrement phonographique marque le moment où la parole s'extériorise[45], comme le langage avait quitté l'homme par l'écriture. Dans cette

45. Leroi-Gourhan conclut par ces mots son ouvrage majeur : « L'outil quitte précocement la main pour donner naissance à la machine : en dernière étape, parole et vision subissent, grâce au développement des techniques, un processus identique. Le langage qui avait quitté l'homme dans les œuvres de sa main par l'art et l'écriture marque son ultime séparation en confiant à la cire, à la pellicule, à la bande magnétique les fonctions intimes de la phonation et de la vision », *in* André Leroi-Gourhan, *Le geste et la parole I. Technique et langage*, Paris, Albin-Michel, 1964, p. 300.

perspective historique, la transcription automatique opère une jonction entre ces deux extériorisations. Elle marque l'étape où la main cesse d'être l'unique vecteur du langage écrit.

Mais de quelle forme de langage écrit parle-t-on ? Qui s'est déjà essayé à transcrire une conférence ou une audience a pu mesurer l'écart entre l'enregistrement et le texte écrit destiné à un lecteur absent de la scène captée. La distance naît de la langue elle-même, de formes linguistiques, d'hésitations, de répétitions, de vocabulaire propres à l'oralité. Mais elle naît aussi de la nécessité de transcrire les silences. De surcroît, le texte doit être appareillé, décoré de compléments nécessaires à la restitution de la situation d'énonciation. Une difficulté première est d'attribuer les propos à leurs locuteurs. Il est surprenant de constater combien face à la transcription brute, même parfaitement fidèle, reconnaître les tours de parole n'est pas immédiat, parfois même difficile. Et puis, en miroir des didascalies de l'auteur de théâtre, une transcription peut comporter des annotations paralinguistiques décrivant intonations, soupirs ou variations d'intensité sonore, tout comme des attitudes des locuteurs : des regards, des postures ou des gestes. Autant de mentions que portent les procès-verbaux d'auditions et que relèvent les greffiers dans leurs actes, procès-verbaux qui oscillent entre une transcription stricte et une reformulation négociée puis endossée par les personnes présentes.

Pour autant, même imparfait ou encore à distance du texte à produire, le résultat de la transcription automatique peut s'avérer particulièrement utile. Il en est ainsi pour une raison simple. En faisant entrer l'enregistrement sonore dans la sphère de l'écrit, en lui donnant une forme graphique, ce procédé affranchit le lecteur de la chronophagie de la consultation sonore, permettant d'aller et venir à la vitesse du regard. Mais surtout, par la transposition des signaux vocaux dans l'univers du texte numérisé, il permet d'y appliquer de très nombreux instruments de traitement automatique des textes, de l'indexation à la recherche, en passant par la classification ou l'analyse.

Pour faire de l'enregistrement sonore une source textuelle à part entière qui soit utile dans le corpus auquel elle est destinée, encore faut-il que la transcription soit fidèle. À défaut, si le contenu est d'importance dans le cadre d'un processus décisionnel, la possibilité d'un retour à la source pour vérification est essentielle. Comment alors s'en assurer ? La question est bien moins triviale qu'il n'y paraît.

« La transposition des signaux vocaux dans l'univers du texte numérisé permet d'y appliquer de très nombreux instruments de traitement automatique des textes, de l'indexation à la recherche, en passant par la classification ou l'analyse.

Notre objectif est de permettre au lecteur de percevoir la distance qui subsiste entre l'image de fidélité absolue d'un processus mécanique et la réalité des processus probabilistes inhérents aux approches connexionnistes sur lesquelles reposent les outils de transcription automatique. À cette fin, après avoir présenté l'état actuel de ces technologies, évoqué les situations dans lesquelles des progrès sont encore espérés – comme la parole superposée ou les formes dégradées d'énonciation –, seront abordés les retours d'expérience pour une mise en œuvre optimale.

I - Clés pour comprendre la transcription automatique : faire face à la rareté des corpus

Entre héritage scientifique, ruptures technologiques et enjeux industriels, la reconnaissance automatique de la parole s'est construite par strates successives, dont l'examen historique constitue aussi un révélateur des questions structurantes et des difficultés persistantes du domaine[46].

1. Origines et développement de la reconnaissance automatique de la parole

La reconnaissance automatique de la parole constitue un domaine ancien de la recherche en informatique et même l'un des premiers à avoir été exploré à l'issue de la Seconde Guerre mondiale. Dès 1945, le monde universitaire

46. Les développements de cette première partie reposent très largement sur le contenu de l'intervention de Christophe Servan lors de l'atelier tenu à l'IRB le 15 mai 2025.

prend conscience du rôle déterminant joué par l'informatique et l'électronique dans les avancées technologiques ayant contribué à l'effort de guerre, notamment avec les travaux d'Alan Turing. Cette prise de conscience entraîne, en particulier aux États-Unis, un essor significatif de la recherche dans ces disciplines.

> La reconnaissance automatique de la parole constitue un domaine ancien de la recherche en informatique et même l'un des premiers à avoir été explorés à l'issue de la Seconde Guerre mondiale.

Dans ce contexte, deux grands axes se structurent rapidement : d'une part, la traduction automatique, les premiers travaux remontant aux années 1950 ; d'autre part, la reconnaissance automatique de la parole (*automatic speech recognition* ou ASR), qui vise à transcrire automatiquement des conversations. En pleine guerre froide, ces recherches revêtent une portée stratégique évidente.

Au-delà de ce cadre historique, la reconnaissance vocale répond à un besoin fondamental d'interaction naturelle avec les machines. Elle est aujourd'hui devenue pleinement opérationnelle. Les utilisateurs attendent désormais des systèmes vocaux qu'ils soient capables de comprendre leurs demandes. Le domaine a donc été investi par tous les grands acteurs industriels : Amazon, Google et Apple figurent ainsi parmi ses principaux protagonistes. Siri développé par Apple a été l'un des premiers assistants vocaux à connaître une large diffusion, tandis qu'Alexa développé par Amazon s'est imposé comme l'un des assistants vocaux domestiques les plus répandus. Aujourd'hui, la recherche et l'industrie restent très actives, et la plupart des grands acteurs de l'IA continuent de mettre au point de nouvelles approches[47]. Une particularité de ce domaine de l'IA est que des acteurs plus spécialisés et moins puissants demeurent néanmoins très performants et compétitifs[48].

Un autre domaine connaît actuellement un développement très important : le sous-titrage et le doublage automatique, servant à quantité de contenus vidéo qui sont produits aujourd'hui dans des volumes gigantesques (plus de 500 heures de vidéos sont mises en ligne sur YouTube chaque minute[49]). On peut citer également d'autres applications spécialisées, comme les systèmes d'assistance vocale ou d'accessibilité[50], la transcription automatique de réunions ou la traduction orale en temps réel.

2. Une brève histoire de la transcription automatique

Les premières technologies de reconnaissance vocale reposaient non sur des logiciels, mais sur des dispositifs électroniques spécialisés. Avant les années 1960, en l'absence de systèmes programmables, ces dispositifs ne permettaient de traiter qu'un nombre très limité de signaux – par exemple, la reconnaissance de chiffres – ce qui constituait néanmoins une avancée significative. Ces travaux s'inscrivaient dans le prolongement des fondements théoriques posés par Claude Shannon[51], dont les apports continuent de structurer les approches contemporaines du traitement de l'information, du langage et du signal.

L'apparition des systèmes programmables dans les années 1960 marque un tournant décisif. Elle fournit les moyens de comparer des séquences acoustiques entre elles, en confrontant un signal entrant à des modèles préenregistrés, ouvrant ainsi la voie à des systèmes capables de reconnaître certains mots et structures sonores.

Un nouveau saut qualitatif intervient au début des années 1980 avec l'introduction des modèles de Markov cachés[52] qui s'imposent progressivement comme l'approche dominante. Ces modèles probabilistes remplacent les règles strictement déterministes et permettent

47. On peut mentionner la série Voxtral de Mistral, Amazon Transcribe, Microsoft, NVIDA Canary, IBM Granite Speech.

48. Par exemple Nova-3 de Deepgram, AssemblyAI ou Parakeet.

49. Michele Klawitter, « YouTube Statistics 2026 : Users, Revenue & Growth Data, 2026 », *Wytlabs blog*, 2026, https ://wytlabs.com/blog/youtube-statistics/.

50. Un système d'assistance vocale ou d'accessibilité est un logiciel informatique qui permet à un utilisateur d'interagir avec un appareil ou d'accéder à ses fonctionnalités par la voix ou via des aides alternatives (lecture à voix haute, commandes vocales, dictée, navigation simplifiée), notamment pour faciliter l'usage aux personnes en situation de handicap.

51. Claude E. Shannon, « A Mathematical Theory of Communication », *Bell System Technical Journal*, vol. 27, , juillet et octobre 1948, p. 379-423 et 623-656, http ://pespmc1.vub.ac.be/books/Shannon-TheoryComm.pdf.

52. Les *Hidden Markov Models* – HMM sont des modèles probabilistes utilisés pour décrire des phénomènes qu'on ne peut pas observer directement, mais dont on voit les effets.

l'apprentissage, à partir des données, de probabilités de transition entre sons, mots ou séquences. Ils rendent possible l'émergence de modèles de langue capables d'évaluer, voire de générer, des séquences linguistiques, ouvrant ainsi la voie non seulement à la reconnaissance, mais aussi aux premières formes de synthèse vocale. Ces approches resteront au cœur des systèmes jusqu'au début des années 2010.

Dans les premières architectures mises au point, la reconnaissance de la parole reposait sur une chaîne de traitement en plusieurs étapes. Le signal audio brut était d'abord transformé afin d'en extraire des caractéristiques acoustiques, puis celles-ci étaient exploitées par un modèle acoustique pour identifier des phonèmes, unités sonores élémentaires. Un modèle de décodage permettait ensuite de convertir ces phonèmes en mots, avant l'intervention d'un modèle de langue chargé de sélectionner les séquences les plus plausibles en fonction de probabilités d'enchaînement. Le processus suivait ainsi une progression du signal audio vers les phonèmes, puis les mots, jusqu'à la reconstruction d'une phrase cohérente. Ces approches, largement mobilisées dans les années 2000, ont conduit à l'émergence de plusieurs outils de référence dans la recherche.

3. L'arrivée des approches neuronales

Un tournant majeur intervient au milieu des années 2010 avec l'essor des approches neuronales profondes. S'inspirant notamment des progrès réalisés en traitement d'image, les chercheurs envisagent le signal audio comme une représentation séquentielle, analogue à une image dans le temps.

> Un tournant majeur intervient au milieu des années 2010 avec l'essor des approches neuronales profondes.

Des réseaux de neurones convolutionnels[53] sont alors mobilisés pour analyser ce signal à l'aide de fenêtres glissantes et en extraire automatiquement des caractéristiques pertinentes. Celles-ci sont ensuite traitées par d'autres architectures, souvent de type séquence-à-séquence, afin de convertir progressivement l'information acoustique en texte.

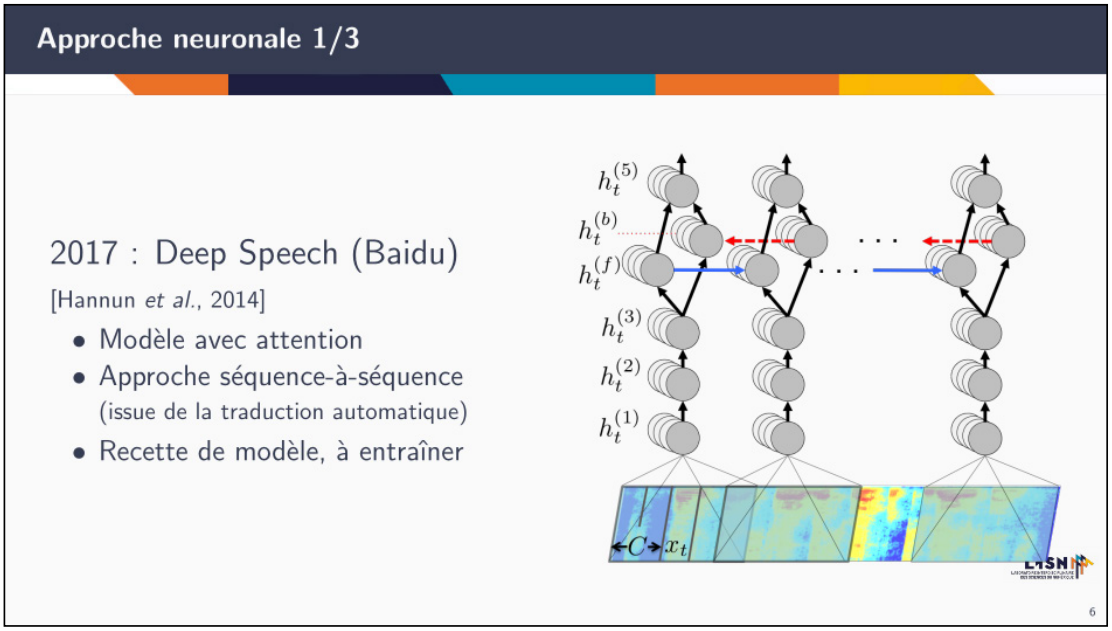
Parmi les premiers modèles marquants figure Deep Speech[54], développé par Baidu Research, qui intègre notamment des mécanismes d'attention[55] et présente des analogies avec certaines architectures de traduction automatique. Ces approches permettent d'améliorer significativement les performances des systèmes de transcription vocale tout en facilitant leur mise en œuvre. Elles nécessitent la disponibilité de très grands volumes de données d'apprentissage, sous forme de segments audio déjà transcrits, le texte attendu comportant des indications temporelles. À défaut, leur efficacité demeure limitée.

C'est dans ce contexte qu'ont émergé des initiatives visant à constituer des corpus ouverts. La Fondation Mozilla a ainsi lancé le projet Common Voice[56], destiné à collecter des enregistrements vocaux dans différentes langues.

Un nouveau tournant intervient à la fin des années 2010 avec l'essor des modèles auto-supervisés. Inspirées des avancées en traitement du langage, notamment des modèles fondés sur le masquage comme BERT[57], ces approches consistent à apprendre la structure du signal en reconstituant des portions manquantes. Appliquée à l'audio, cette méthode permet d'exploiter de vastes volumes de données non annotées, changeant ainsi l'échelle de l'apprentissage.

Les premiers résultats significatifs apparaissent en 2019 avec wav2vec, puis wav2vec 2.0, des modèles capables de couvrir un grand nombre de langues. Leur principe repose sur un

53. Les réseaux de neurones convolutionnels (CNN) sont des modèles d'apprentissage automatique conçus pour analyser des données structurées comme des images, en appliquant des filtres (convolutions) afin d'en extraire automatiquement des motifs importants (formes, textures, contours).
54. Awni Y. Hannun, Carl Case, Jared Casper *et al.*, « Deep Speech : Scaling up end-to-end speech recognition », *ArXiv*, 2014. https ://api.semanticscholar.org/CorpusID :16979536.
55. Il s'agit de techniques permettant au réseau de se concentrer automatiquement sur les parties les plus importantes d'une image ou d'une carte de caractéristiques, afin d'améliorer la qualité de l'apprentissage et des prédictions.
56. Common Voice est une plateforme libre et *open source* pour la création de données par des mécanismes participatifs. https ://commonvoice.mozilla.org/fr.
57. Il s'agit de modèles de langage apprenant à comprendre le sens d'un texte en masquant certains mots et en apprenant à les prédire à partir de leur contexte, ce qui leur permet de capter des relations fines entre les mots.



Deep Speech, premières approches neuronales pour la transcription
Source : Christophe Servan

encodeur entraîné par masquage[58], qui apprend des représentations du signal réutilisables pour diverses tâches, notamment la reconnaissance vocale, avec un besoin réduit en données annotées.

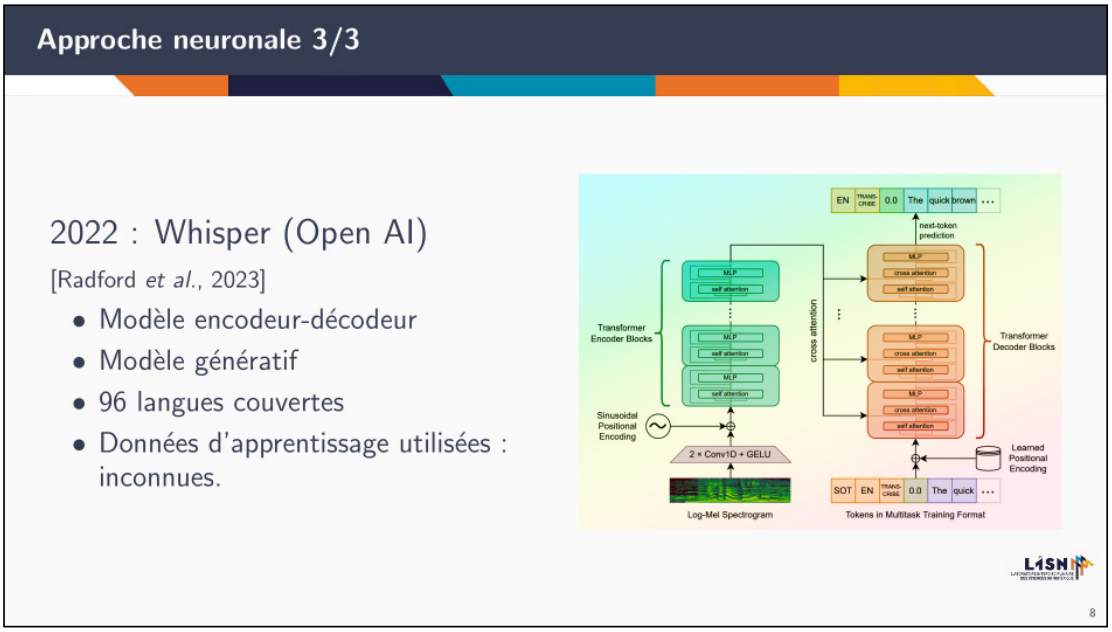
La question des données d'entraînement demeure toutefois centrale. Si certains acteurs communiquent partiellement sur leurs sources, celles-ci restent souvent peu détaillées. On sait néanmoins que, dans le cas de wav2vec, de larges corpus audio publics ont été mobilisés pour constituer les jeux de données d'apprentissage.

Avec Whisper, publié par OpenAI en 2022, une nouvelle étape est franchie. Ce modèle, fondé sur une architecture encodeur-décodeur, adopte une approche générative inspirée des modèles de type GPT : à partir d'un signal audio en entrée, il produit directement une séquence textuelle en sortie. Outre qu'il offre de bonnes performances, une très grande souplesse d'utilisation et la possibilité de fonctionner en local mais aussi avec des infrastructures de calcul internes très puissantes, il a permis de réduire considérablement le volume de données d'entraînement nécessaire pour chaque nouvelle langue[59].

La question des données d'apprentissage reste ouverte, d'autant que les informations disponibles sur les corpus utilisés demeurent comme on l'a dit souvent partielles, voire absentes. Il est probable que ces corpus combinent des données audio publiques, des enregistrements spécifiquement collectés et des annotations réalisées par des opérateurs humains. Or ces pratiques soulèvent des questionnements, tant sur les conditions de collecte que sur les modalités d'annotation, qui mériteraient à elles seules un examen approfondi.

58. Les modèles génératifs prédisent un élément à partir du contexte qui le précède : pendant l'apprentissage le modèle anticipe le mot suivant, puis ajuste ses paramètres en comparant sa prédiction à la valeur attendue. Les modèles à masquage mobilisent à la fois le contexte précédent et le contexte suivant pour reconstituer un élément manquant au sein d'une séquence. Concrètement, certains mots sont masqués, puis prédits à partir de leur environnement, selon un procédé répété sur l'ensemble de la séquence. Ils exploitent le contexte global, à la fois en amont et en aval, pour apprendre des représentations plus complètes. Chacune des familles a des domaines dans lesquels elle est plus performante. Les modèles à masquage sont généralement supérieurs pour la classification, l'extraction de données ou l'analyse syntaxique.

59. Par exemple pour le suédois, 300 heures d'audio ont suffi à dépasser les performances du meilleur modèle de l'époque, qui aurait nécessité des milliers d'heures d'audio pour des performances identiques.



Whisper d'Open AI, modèle *open-source*
Source : Christophe Servan

II - Le futur de la transcription automatique : surmonter des cas difficiles fréquents dans l'environnement judiciaire

L'influence du niveau de langue confirme les constats précédents : lorsque l'expression est structurée, avec une syntaxe maîtrisée, un vocabulaire stable et une diction claire – comme c'est souvent le cas chez les magistrats – les performances des modèles sont élevées et la transcription est fiable dans l'ensemble.

Un point plus fondamental doit par ailleurs être souligné : un modèle d'IA constitue toujours une approximation de la réalité, et non la réalité elle-même. Il en résulte nécessairement l'existence de situations non prévues, de cas rares ou de phénomènes insuffisamment représentés dans les données d'entraînement, susceptibles de générer des erreurs.

D'un point de vue technique, les modèles d'IA reposent sur des mécanismes probabilistes : ils apprennent à estimer ce qui est le plus probable dans le contexte audio donné. Or la probabilité ne saurait se confondre avec la certitude, ce qui rend les erreurs inévitables.

1. Des cas qui restent difficiles

Les situations les plus improbables, qui sont donc sources d'erreurs, sont soit celles qui n'ont pas été couvertes par le corpus d'entraînement, soit celles qui sont naturellement plus imprévisibles.

Il s'agit en particulier de la **parole spontanée**, qui reste l'un des principaux défis actuels et demeure un champ de recherche encore largement ouvert. Elle se caractérise par de nombreux phénomènes linguistiques difficiles à modéliser, tels que les disfluences, les hésitations, les répétitions, les autocorrections ou encore les faux départs.

Des énoncés du type « Je… enfin… ce que je voulais dire… » sont ainsi fréquents dans la conversation ordinaire, mais ils mettent en difficulté les systèmes de reconnaissance vocale, en perturbant les mécanismes de transcription et d'interprétation.

Dans le même sens, la **parole superposée** constitue une difficulté majeure. Dans les échanges naturels, les interlocuteurs se coupent, réagissent simultanément et parlent en chevauchement. Si ces phénomènes restent limités dans des contextes très formels, ils sont fréquents dans les situations réelles – réunions, interrogatoires ou conversations informelles. Pour les systèmes de reconnaissance vocale, le traitement de ces superpositions demeure particulièrement complexe.

Enfin, le multilinguisme introduit une difficulté supplémentaire, lorsque de nombreux locuteurs alternent entre plusieurs langues

au sein d'un même échange, par exemple en intégrant des termes techniques prononcés à l'anglaise dans un discours en français. Ces phénomènes de **mélange linguistique** complexifient la tâche des systèmes de reconnaissance automatique.

Plus en lien avec la structuration des données d'apprentissage, un autre facteur déterminant tient au **niveau de langue** présent dans les données d'entraînement. De nombreux corpus reposent sur un langage relativement formel ou standardisé, ce qui favorise de bonnes performances lorsque les locuteurs s'expriment dans un registre soutenu et structuré.

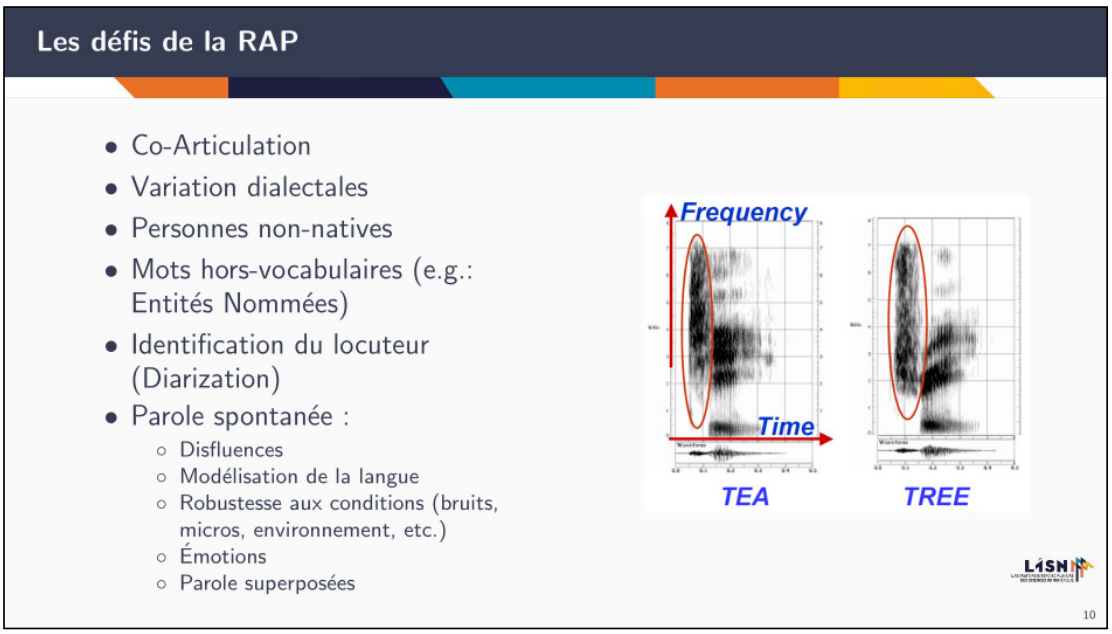
Ainsi, dans un contexte judiciaire, les interventions d'un juge sont généralement bien reconnues, sous réserve d'erreurs mineures. En revanche, les performances se dégradent lorsque les locuteurs présentent des difficultés d'élocution, mobilisent un registre différent ou s'expriment de manière moins structurée. On peut penser ici aux enfants, aux locuteurs non natifs du français, aux personnes atteintes de certains troubles psychiatriques.

En outre, la voix est fortement influencée par **l'état émotionnel et physiologique des locuteurs**. Dans des situations telles qu'un témoignage judiciaire, le stress, l'émotion ou l'agitation modifient la prosodie et les caractéristiques acoustiques, compliquant la reconnaissance automatique.

« Dans des situations telles qu'un témoignage judiciaire, le stress, l'émotion ou l'agitation modifient la prosodie et les caractéristiques acoustiques, compliquant la reconnaissance automatique.

Des facteurs plus ordinaires produisent des effets similaires, tels que la fatigue, un rhume ou les variations naturelles de la voix au cours de la journée. Ainsi, contrairement à une empreinte digitale, la voix ne constitue pas un identifiant biométrique parfaitement stable.

En dernier lieu, même dans les contextes où les performances sont élevées, certaines difficultés persistent, en particulier s'agissant des mots rares et des noms propres. Ce problème, appelé **reconnaissance des entités**



Les défis de la reconnaissance automatique de la parole
Source : Christophe Servan

nommées[60], constitue un point de fragilité classique : si des noms fréquents sont généralement bien reconnus, des noms plus rares ou d'origine étrangère, ou ceux qui n'ont jamais été entendus dans le corpus d'entraînement, donnent lieu à des transcriptions souvent approximatives, parfois surprenantes.

Un autre défi majeur concerne l'**identification des locuteurs**. Les systèmes actuels tendent davantage à reconnaître des caractéristiques liées à la source audio – comme le microphone ou l'environnement acoustique – qu'à identifier précisément les individus. Cette limite restreint certaines applications, notamment dans les réunions avec un unique micro par lieu de captation, et demeure un champ de recherche encore insuffisamment couvert. Les systèmes de transcription automatique de visioconférences bénéficient des particularités permettant d'attribuer le locuteur par les informations sur chaque participant et sa source spécifique, ce qui explique leurs performances supérieures dans ce contexte.

Ce phénomène peut conduire à des effets contre-intuitifs. Dans une configuration dans laquelle chaque participant est associé à une position autour de la table, un changement de place lors d'une réunion ultérieure peut induire le système en erreur : il attribue alors l'identité du locuteur à la position ou au microphone, plutôt qu'à la personne elle-même.

Ce comportement s'explique par la manière dont les modèles apprennent, en privilégiant les caractéristiques les plus saillantes dans les données, indépendamment de leur pertinence au regard de notre intuition. Ainsi, des différences imperceptibles pour l'oreille humaine peuvent être exploitées de manière déterminante par la machine.

De manière synthétique, il ressort que l'utilisation de la transcription automatique dans les contextes judiciaires pourrait être moins performante, à raison des situations d'énonciations souvent difficiles pour les outils actuels, qu'il s'agisse de l'attribution fiable des propos aux locuteurs, du caractère décisif de la détection des entités nommées, de la robustesse face à la parole spontanée, du traitement de la parole superposée, de la prise en compte de la diversité des niveaux de langue ou de l'impact des états émotionnels et physiologiques. Les systèmes existants fonctionneront, mais probablement avec des performances moindres et des taux d'erreurs plus élevés pouvant nécessiter des temps de correction et de reprise plus importants pour une transcription fidèle.

2. L'importance critique des corpus d'entraînement

Comme cela a été abordé, les performances des systèmes dépendent étroitement des données mobilisées lors de l'entraînement. Lorsque certaines variantes linguistiques sont peu représentées, les modèles peinent à les traiter correctement. Cette limite est particulièrement visible pour les accents régionaux, les dialectes ou les formes de langage local.

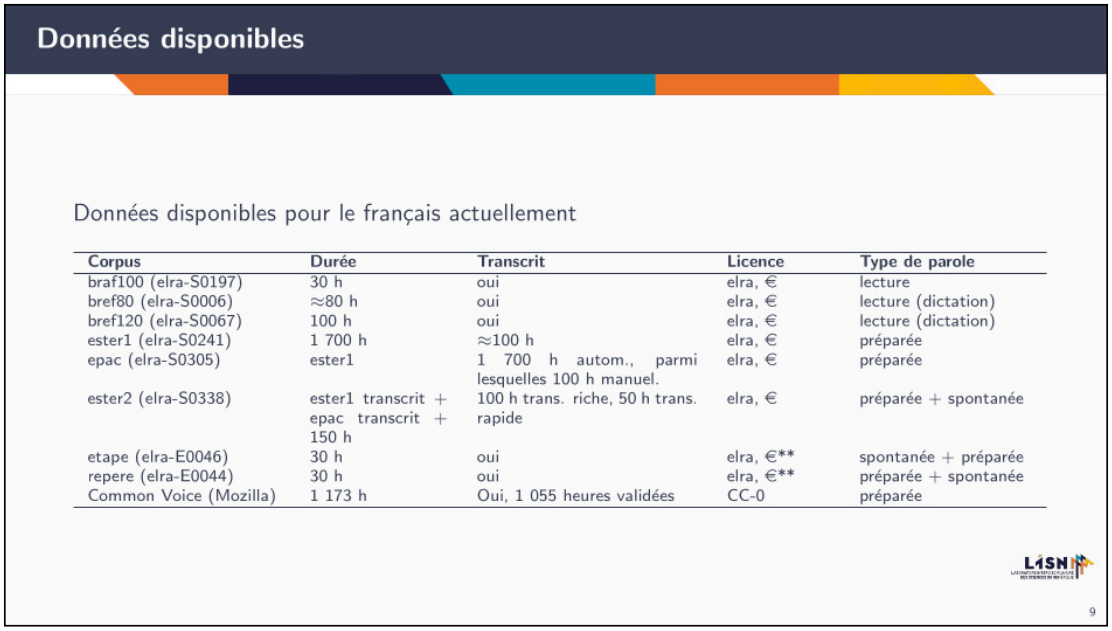
> « Les performances des systèmes dépendent étroitement des données mobilisées lors de l'entraînement.

Dans le cas d'accents très spécifiques, le manque d'enregistrements disponibles empêche un apprentissage satisfaisant. Même en cherchant à corriger ces lacunes, plusieurs obstacles subsistent : rareté des données, difficulté de collecte et contraintes juridiques pesant sur leur utilisation. Ces limites constituent ainsi des biais structurels des modèles actuels.

Ces difficultés montrent l'importance de disposer de sources plus proches de la parole ordinaire – telles que les films, séries ou contenus audiovisuels – afin de mieux refléter la diversité des locuteurs et des registres.

Si ces ressources offrent en théorie une grande richesse – parole informelle, accents variés, registres populaires –, leur exploitation se heurte à plusieurs obstacles : les contraintes juridiques liées aux droits d'usage, la qualité inégale des transcriptions disponibles, ainsi que la complexité intrinsèque de la parole spontanée. Pour ces corpus, les enjeux juridiques sont majeurs, en particulier au regard du droit

60. En traitement automatique des langues, une entité nommée est une expression linguistique qui désigne un référent singulier et identifiable (personne, lieu, organisation, date, etc.). À la différence du lexique commun, elle renvoie à une instance particulière du monde. Sa détection et sa catégorisation automatiques constituent une opération centrale, qui permet de structurer l'information textuelle tout en soulevant des enjeux de désambiguïsation et de référence. En anglais on parle de *Named Entity Recognition* ou NER.



Corpus	Durée	Transcrit	Licence	Type de parole
braf100 (elra-S0197)	30 h	oui	elra, €	lecture
bref80 (elra-S0006)	≈80 h	oui	elra, €	lecture (dictation)
bref120 (elra-S0067)	100 h	oui	elra, €	lecture (dictation)
ester1 (elra-S0241)	1 700 h	≈100 h	elra, €	préparée
epac (elra-S0305)	ester1	1 700 h autom., parmi lesquelles 100 h manuel.	elra, €	préparée
ester2 (elra-S0338)	ester1 transcrit + epac transcrit + 150 h	100 h trans. riche, 50 h trans. rapide	elra, €	préparée + spontanée
etape (elra-E0046)	30 h	oui	elra, €**	spontanée + préparée
repere (elra-E0044)	30 h	oui	elra, €**	préparée + spontanée
Common Voice (Mozilla)	1 173 h	Oui, 1 055 heures validées	CC-0	préparée

Les jeux de données en licence libre pour l'entraînement du français restent rares
Source : Christophe Servan

d'auteur. De nombreuses sources potentiellement foisonnantes – films, documentaires ou archives audiovisuelles – ne peuvent être librement exploitées pour l'entraînement des modèles.

À ces contraintes s'ajoutent les droits de diffusion et d'exploitation, ainsi que des exigences liées à l'anonymisation des voix et à la protection de l'identité des personnes enregistrées. Même des ressources particulièrement précieuses, telles que les archives de l'INA, demeurent fortement encadrées, ce qui en complique considérablement l'utilisation à des fins d'apprentissage.

Ces limites expliquent que de nombreux corpus d'entraînement demeurent fondés sur des données plus formelles et contrôlées, mais qui limitent la performance pour les formes d'expression moins rigoureuses.

Il existe donc un enjeu très important autour de la constitution de corpus d'enregistrements audio de situations réelles ou réalistes, avec leur transcription. Au fur et à mesure que des transcriptions seront opérées, la réalisation de transcriptions temporalisées sur des cas concrets d'audiences judiciaires constituera un axe important pour faire progresser les performances des générations futures de modèles de transcription.

> De nombreux corpus d'entraînement demeurent fondés sur des données plus formelles et contrôlées, mais qui limitent la performance pour les formes d'expression moins rigoureuses.

3. La question des données disponibles pour le français

Aujourd'hui, les systèmes de reconnaissance vocale atteignent des niveaux de performance très élevés, à la suite de progrès considérables. La qualité des résultats produits demeure toutefois dépendante de plusieurs facteurs, notamment la qualité des enregistrements, les conditions de captation et la disponibilité de données d'apprentissage adaptées.

S'agissant du français, la question des ressources disponibles reste particulièrement sensible. Les corpus, parfois libres ou partiellement accessibles, constituent des éléments essentiels pour le développement de ces technologies, mais leur volume demeure limité.

Comme cela a été souligné, atteindre des ordres de grandeur de plusieurs milliers d'heures de données reste difficile en pratique. Malgré l'existence de méthodes complémentaires permettant d'atténuer cette contrainte, les ressources disponibles pour le français demeurent sans commune mesure avec celles de l'anglais, pour lequel les volumes atteignent des centaines de milliers d'heures d'enregistrements.

III - Chaînes de traitements et d'édition, leçons de mise en œuvre

Après avoir dessiné le chemin restant à parcourir en matière de recherche sur cette thématique, la maturité de ces technologies permet d'aborder des situations de mise en œuvre concrète. Après la présentation d'un exemple tiré d'un cas proche d'une situation judiciaire réelle, la question des métriques pour la mesure des performances sera examinée.

1. Un exemple pratique tiré d'une séquence filmée

Afin d'illustrer concrètement ces enjeux, une démonstration a été réalisée à partir d'un extrait de documentaire filmant une situation d'audience judiciaire. L'accès à des enregistrements réels étant fortement contraint pour des raisons juridiques et éthiques, le choix s'est porté sur un extrait de la bande-annonce du documentaire *Bouche cousue*[61], consacré au travail des juges des enfants[62].

Cet extrait, filmé avec des moyens professionnels et en présence de participants conscients d'être enregistrés, présente des conditions favorables : qualité sonore élevée et diction relativement maîtrisée. Il ne reflète donc pas les situations les plus difficiles, mais permet d'observer le fonctionnement d'un système de transcription automatique dans un cadre réaliste.

Le système utilisé, récemment distingué par la National Association of Broadcasters, produit à première vue une transcription convaincante : identification des locuteurs, phrases globalement correctes, compréhension d'ensemble satisfaisante. Toutefois, une analyse détaillée fait apparaître plusieurs types d'erreurs.

Certaines relèvent de la ponctuation et peuvent altérer profondément le sens. Ainsi, la phrase prononcée par les parents suspectés de violence sur leur fille mineure « C'est pas nous on a frappé Rebeccas » est transcrite « C'est pas nous. On a frappé Rebecca », inversant l'interprétation des faits. D'autres concernent l'attribution des locuteurs : une intervention du père adressée à sa fille est, par exemple, attribuée au juge, ce qui peut affecter la compréhension des échanges dans un contexte formel.

On observe également des erreurs lexicales liées à une mauvaise interprétation du signal sonore, produisant des phrases proches de l'original mais dégradées dans leur intelligibilité. À cela s'ajoutent des phénomènes d'hallucination, lorsque le système génère temporairement des segments incohérents avant de se réaligner.

> On observe également des erreurs lexicales liées à une mauvaise interprétation du signal sonore, produisant des phrases proches de l'original mais dégradées dans leur intelligibilité.

Enfin, l'extrait met en évidence des écarts de performance selon les locuteurs : les interventions du juge, plus structurées, sont transcrites avec une grande précision, tandis que celles des parents, marquées par une parole plus spontanée, comportent davantage d'erreurs. Cette observation confirme que les systèmes actuels sont nettement plus fiables face à un langage formel que dans des situations d'expression naturelle ou émotionnelle.

Une seconde expérience, à première vue plus difficile, a été réalisée à partir d'un segment du documentaire *12 jours* de Raymond Depardon[63] consacré à des audiences relatives à l'hospitalisation psychiatrique sans consentement. L'extrait se caractérise par une parole plus hésitante, une syntaxe moins structurée et

61. Karine Dusfour, *Bouche cousue*, 416PROD, 2020, 71 mn, https ://416prod.fr/portfolio-item/bouche-cousue/ et bande-annonce sur https ://boutique.arte.tv/detail/bouche-cousue.
62. Cette section reprend une courte expérimentation réalisée par l'auteur, présentée lors de l'atelier du 15 mai 2025.
63. Raymond Depardon, *12 jours*, 2017, 88 min, https ://www.palmeraieetdesert.fr/films-1/12-jours.

Les objectifs du projet de recherche mené par le pôle judiciaire de la Gendarmerie nationale
Source : PJGN

un discours parfois difficile à suivre, mais dans un contexte de dialogue. Malgré ces conditions, la transcription automatique demeure globalement compréhensible, ce qui témoigne des progrès récents des systèmes.

Ces deux expériences conduisent à un constat nuancé. D'une part, les performances atteintes sont remarquables pour de nombreux fragments, y compris dans des conditions d'élocution difficiles. Ils permettent des gains de temps considérables. D'autre part, des erreurs parfois minimes peuvent altérer très sensiblement le sens des échanges, rendant indispensable une relecture humaine très attentive par une personne ayant entendu le contenu audio, en particulier dans des contextes sensibles tels que le domaine judiciaire.

2. Les outils de transcription développés par la Gendarmerie nationale

Le code de procédure pénale impose la rédaction de procès-verbaux d'audition dans le cadre des enquêtes. Cette activité essentielle est très chronophage. Quand la personne entendue est un adulte, le contenu n'est pas une transcription stricte des échanges, en ce qu'elle intègre des reformulations des propos tenus, dont la personne entendue valide la fidélité par sa signature. Les auditions de mineurs, dont l'enregistrement est prescrit par le législateur, donne lieu à des procès-verbaux beaucoup plus près des propos prononcés, sans reformulation. Ils sont d'autant plus sensibles quand ces auditions concernent des mineurs victimes, dont la parole a une valeur probatoire[64].

La Gendarmerie nationale a rapidement identifié la transcription automatique comme un gisement de productivité. Elle a donc réalisé un projet de recherche européen, Rosettai, pour réduire drastiquement ce temps de traitement, en visant un passage de huit à deux heures de travail par heure d'enregistrement sonore. Ce seuil correspond à un minimum incompressible : une heure d'écoute pour vérification, et environ une heure de correction. Le gain est significatif – de l'ordre de 75 % – et permet de réallouer le temps des enquêteurs à des tâches de plus forte valeur ajoutée, telles que l'analyse ou l'accompagnement des victimes. L'outil diffusé auprès de plusieurs services repose sur une interface volontairement simple. Il propose une visualisation du signal audio, facilitant l'ajustement des segments, ainsi qu'une segmentation des locuteurs – sans identification nominative – que l'utilisateur peut ensuite personnaliser. Une fonctionnalité centrale réside dans la correction des transcriptions : modification

64. Les développements de cette section s'appuient en grande partie sur l'intervention de Daniel Camara lors de l'atelier du 15 mai 2025.

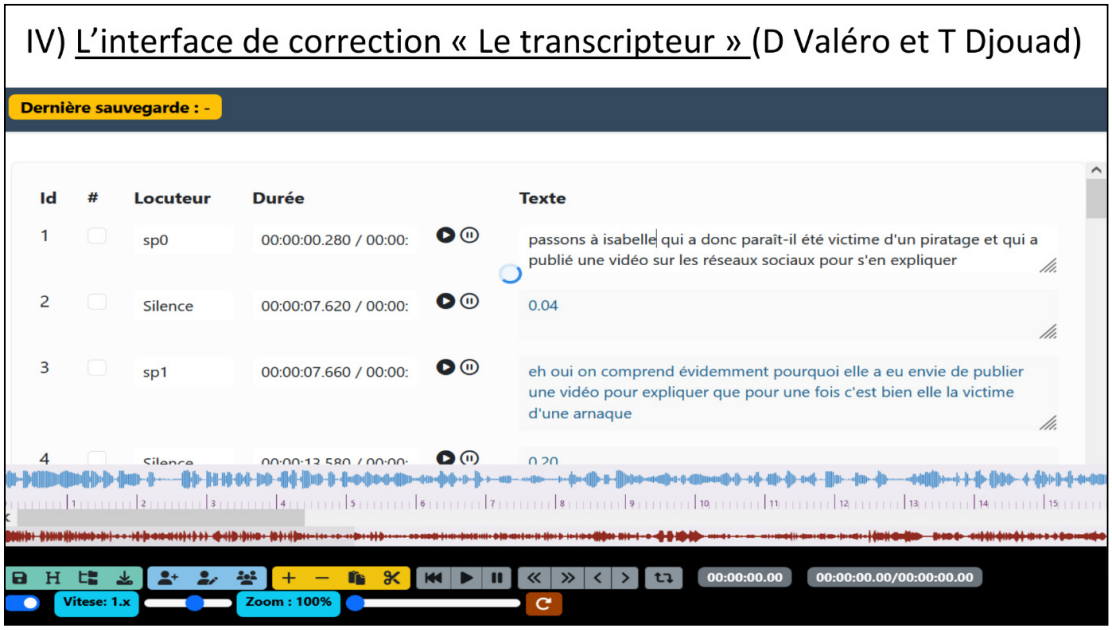


Figure : l'interface de correction des transcriptions de la plateforme PUD-GA
Remerciements à Max Béligné

du texte, ajustement des segments et mise en forme en vue d'une intégration directe dans les documents officiels, selon une logique suivant laquelle la machine assiste l'utilisateur.

Inscrit dans un contexte opérationnel, cet outil est utilisé par différents services, notamment pour les cas d'usage précités dont certains sont particulièrement exigeants comme les auditions de mineurs. Ces dernières situations cumulent plusieurs difficultés : spécificités du langage enfantin, diction parfois peu claire et emploi de formulations peu présentes dans les données d'entraînement, ce qui complique significativement la reconnaissance automatique.

Les outils déployés ne se limitent pas à une simple transcription automatique. Ils relèvent davantage d'une logique d'annotation de la parole, poursuivant un double objectif : d'une part, faciliter la transcription et l'annotation des enregistrements et, d'autre part, constituer des jeux de données destinés à l'entraînement et à l'amélioration des modèles. L'outil devient ainsi à la fois instrument de production des procès-verbaux et de perfectionnement des systèmes. Cette perspective, bien que techniquement prometteuse, soulève toutefois des enjeux juridiques importants, en particulier en matière de protection des données, ce qui en limite aujourd'hui le déploiement.

Elément important, le gain de temps initial apporté par la transcription automatique ne doit pas être perdu par la lourdeur du processus d'obtention de texte finalisé. Cela passe par la fourniture d'outils ergonomiques et performants pour assister ce travail de correction, notamment pour naviguer dans le document sonore et entre le texte et l'audio. En ce sens, l'exemple de la plateforme unifiée de données de l'université Grenoble Alpes est intéressant. Une large part de l'effort a été de mettre à disposition des chercheurs une interface de correction des transcriptions simple et efficace, en particulier pour permettre de réécouter très simplement chacun des segments audio de manière granulaire et, dans le contexte de la correction, de naviguer dans l'ensemble du document sonore, de visualiser les formes d'onde en particulier pour repérer les silences.

« Le gain de temps initial apporté par la transcription automatique ne doit pas être perdu par la lourdeur du processus d'obtention de texte finalisé.

Assurer la sécurité technique des plateformes mutualisées

Dans le cadre de l'outil expérimental, plusieurs mesures de sécurité ont été mises en place afin de limiter les risques. Les fichiers audio envoyés par les unités étaient conservés uniquement le temps du traitement, puis supprimés immédiatement après production et transmission de la transcription, sans conservation ni des enregistrements ni des résultats. Une approche similaire a été mise en place dans le cadre de la Plateforme universitaire de données Grenoble Alpes (PUD-GA)[65] : le stockage transitoire très bref limite considérablement le volume d'enregistrements susceptibles d'être subtilisés en cas de brèche de sécurité.

Par ailleurs, les fichiers étaient renommés automatiquement pour éviter toute conservation d'informations d'origine, puis stockés dans des répertoires chiffrés et non accessibles à d'autres utilisateurs. L'ensemble visait à garantir un niveau de sécurité jugé raisonnable, compte tenu du contexte.

L'outil de la Gendarmerie a également été conçu pour fonctionner en local, y compris avec un simple ordinateur portable. Cette capacité est essentielle dans le cadre d'enquêtes sensibles, pour lesquelles le transfert de données, même vers des infrastructures internes, peut être exclu. Dans ces conditions, l'ensemble du traitement est effectué directement sur la machine de l'utilisateur, sans connexion Internet, sans envoi de fichiers et sans circulation de données vers l'extérieur.

Un premier bilan : les difficultés d'expérimentation dans le cadre réglementaire

L'outil dans sa version expérimentale a connu une adoption particulièrement rapide au sein de la Gendarmerie. Initialement destiné aux pôles judiciaires, il a été diffusé assez largement à l'initiative d'utilisateurs, ce qui a entraîné une saturation rapide des serveurs face à l'augmentation des demandes. Il a notamment été mobilisé dans les maisons de protection des familles, unités spécialisées dans les affaires sensibles impliquant souvent des mineurs. À l'échelle nationale, ces structures génèrent plusieurs centaines d'heures d'auditions chaque semaine, révélant l'ampleur du besoin et expliquant la montée en charge rapide de l'outil.

Face au constat des attentes, des gains apportés et fort de ces résultats de l'expérience de recherche, un outil institutionnel a alors été déployé. Développé par l'École nationale supérieure de technologie et d'ingénierie (ENSTI) pour le compte du ministère de l'Intérieur, le système baptisé « Parole » se distingue par une volonté d'utilisation dans un plus grand nombre de services, son niveau de sécurisation et une conformité accrue[66].

S'agissant du cadre réglementaire, les enregistrements de la voix sont des données personnelles, indépendamment de toute finalité d'identification. Dès lors, le traitement d'enregistrements vocaux implique le respect des exigences légales, ce qui pose des questions juridiques et organisationnelles significatives.

> Le traitement d'enregistrements vocaux implique le respect des exigences légales, ce qui pose des questions juridiques et organisationnelles significatives.

La prise en compte des exigences juridiques, notamment dans le champ pénal, s'est révélée particulièrement complexe, en raison d'un décalage marqué entre les besoins d'expérimentation et le cadre réglementaire existant, ce qui a induit des délais importants, particulièrement lors de la mise en œuvre des procédures d'analyse d'impact sur la protection des données. Ces délais se comptent en années, alors même que les technologies évoluent très rapidement, ce qui rend plus difficile l'adaptation aux avancées récentes.

Cette situation a également conduit à des formes d'auto-limitation. Par souci de sécurité juridique, certains usages ont été volontairement restreints à des cas bien encadrés

65. La Plateforme universitaire de données Grenoble Alpes (PUD-GA) est une structure d'appui à la recherche, spécialisée dans l'usage des données quantitatives en sciences humaines et sociales (SHS). Elle a pour mission principale de faciliter l'accès, la compréhension et l'exploitation des données de recherche. Elle met notamment à disposition une chaîne de transcription automatique à destination des équipes de recherche. Voir https ://www.msh-alpes.fr/plateformes/pud-ga.

66. Voir l'article de Benjamin Polge, « De la recherche de disparus à la transcription des écoutes : l'IA révolutionne déjà la Gendarmerie », *Journal du Net*, 17 mars 2025, https ://www.journaldunet.com/intelligence-artificielle/1539917-de-la-transcription-des-ecoutes-a-la-recherche-de-disparus-comment-l-ia-revolutionne-la-gendarmerie/

– comme les auditions de mineurs, déjà soumises à une obligation d'enregistrement – alors même que les besoins opérationnels sont plus larges. Cette prudence institutionnelle, qui selon les acteurs paraît parfois renforcée à l'excès au regard des exigences strictes du cadre réglementaire, illustre les effets concrets de la complexité des procédures sur le déploiement des technologies.

Ces contraintes sont encore plus fortes pour la production de corpus d'entraînement comprenant des données vocales opérationnelles correspondant aux réalités de terrain, pourtant indispensables à l'affinage de modèles de reconnaissance vocale performants. La difficulté majeure tient aux nécessités de conservation à long terme des données d'entraînement, à l'impossibilité d'anonymiser des segments vocaux, à l'importance de disposer de segments d'une longueur suffisante pour intégrer des contextes réalistes et aux conditions de leur diffusion auprès des équipes de recherche et des fournisseurs de modèles.

« Ces contraintes sont encore plus fortes pour la production de corpus d'entraînement comprenant des données vocales opérationnelles correspondant aux réalités de terrain.

3. La problématique de la mesure de la performance

Les exemples précédents conduisent à s'interroger sur les modalités d'évaluation des systèmes de reconnaissance vocale. Jusqu'à présent, l'appréciation reposait principalement sur une évaluation intuitive, fondée sur le caractère plus ou moins compréhensible de la transcription[67].

L'étape suivante consiste alors à mobiliser des métriques objectives, permettant de mesurer de manière plus rigoureuse la performance de ces systèmes. Mais quels référentiels utiliser pour évaluer un système, et au regard de quelles attentes ? Cette interrogation constitue un point d'entrée déterminant. Il a donc fait l'objet de très nombreux travaux dans le cadre du traitement automatique de la langue, et donc de la transcription automatique.

Une métrique essentielle, mais insuffisante : le *Word Error Rate* (WER)

La métrique la plus couramment mobilisée pour cette tâche est le taux d'erreur sur les mots, ou *Word Error Rate* (WER). Son principe peut être illustré par un exemple simple. Considérons la phrase suivante : « La vitamine C, c'est bon pour la santé », qui comporte huit mots. Imaginons qu'un système de transcription automatique produise la version suivante : « La vie ta mine C bon pour la santé ». Le terme « vitamine » y est décomposé en trois mots (« vie », « ta », « mine ») et le « c'est » a disparu.

Intuitivement, on pourrait ne voir là que deux erreurs. Toutefois, le calcul du WER repose sur une autre logique : il s'agit de déterminer le nombre minimal d'opérations – substitutions, insertions ou suppressions – nécessaires pour transformer la transcription de référence en transcription automatique. Pour ce faire, on aligne d'abord les segments identiques, puis l'on identifie les écarts : une substitution (« vitamine » remplacé par « vie »), deux insertions (« ta » et « mine ») et une suppression (le « c'est »). On totalise ainsi quatre opérations rapportées aux huit mots de la phrase de référence, soit un WER de 50 %.

Cette métrique présente des avantages évidents : elle permet de comparer des systèmes entre eux et de suivre leurs progrès dans le temps. Néanmoins, ses limites apparaissent tout aussi clairement. Dans l'exemple retenu, l'omission du « c'est » affecte peu le sens global, tandis que l'altération du mot « vitamine » en modifie substantiellement la compréhension. Or le WER traite toutes les erreurs de manière équivalente, sans considération pour leur portée sémantique ni pour l'importance relative des mots.

Des indicateurs alternatifs existent, qui cherchent à pondérer les erreurs en fonction de leur contenu informationnel. Ils introduisent toutefois une part de subjectivité. En pratique, le WER demeure donc la référence dominante, largement utilisée dans les publications. Il importe, dès lors, d'en maîtriser les modalités de calcul tout en restant attentif à ses limites.

L'exemple qui vient d'être exposé est issu d'un rapport de recherche de l'Institut national de recherche en sciences et technologies du numérique (INRIA) consacré à ces questions,

67. Les développements de cette section s'appuient principalement sur l'intervention de Max Béligné lors de l'atelier du 15 mai 2025.

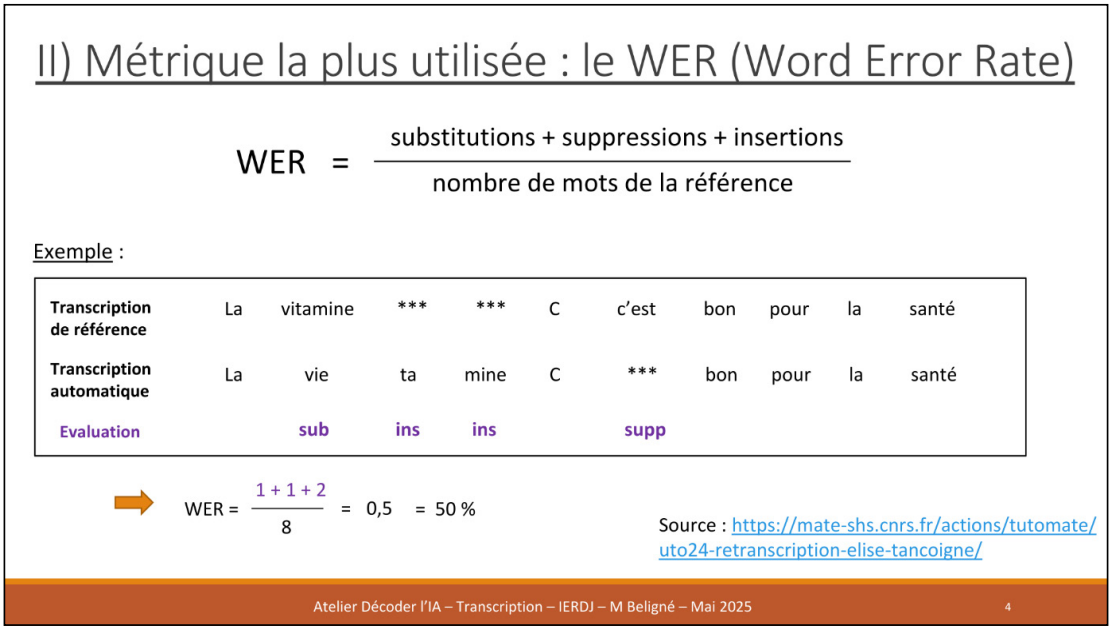


Figure : le *Word Error Rate* (WER)
Source : Max Béligné

publié en 2022[68]. Si les outils qui y sont analysés apparaissent aujourd'hui largement dépassés – tant l'évolution du domaine est rapide –, la méthodologie proposée conserve en revanche tout son intérêt et mérite une attention particulière. Autrement dit, en dépit de l'obsolescence des solutions techniques étudiées, la démarche adoptée demeure pleinement pertinente.

Pour comparer la performance de deux systèmes ou de versions successives d'un même système, il est essentiel de déterminer sur quel corpus d'enregistrements audio seront réalisées les mesures. Deux corpus d'évaluation font référence : Common Voice[69] et FLEURS[70]. Leur principal intérêt réside dans leur caractère multilingue, qui permet de tester des modèles sur un grand nombre de langues.

Ces corpus présentent toutefois des limites importantes. Ils reposent en effet, pour une large part, sur des lectures de phrases issues

> Deux corpus d'évaluation font référence : Common Voice et FLEURS. Leur principal intérêt réside dans leur caractère multilingue.

de Wikipédia, ce qui les éloigne des conditions réelles de parole spontanée. Cette différence de nature peut affecter significativement l'interprétation des performances mesurées.

Par ailleurs, il est essentiel de considérer le type de données avec lesquelles un modèle a été entraîné. Les évaluations sont généralement conduites en considérant plusieurs corpus distincts, car les résultats peuvent varier sensiblement entre les différents jeux de données, y compris lorsque ceux-ci semblent proches. Ainsi, pour certaines langues – comme l'allemand ou l'espagnol dans le corpus FLEURS –, on observe des taux d'erreur particulièrement faibles. Une telle performance ne signifie pas nécessairement que le modèle est intrinsèquement plus efficace dans ces langues mais peut s'expliquer par un phénomène de contamination entre les données d'entraînement et les

68. Élise Tancoigne, Jean Philippe Corbellini, Gaëlle Deletraz *et al.*, « Un mot pour un autre ? Analyse et comparaison de huit plateformes de transcription automatique », *Bulletin of Sociological Methodology/ Bulletin de méthodologie sociologique*, n° 155, 2022, p.45-81. https ://doi.org/10.1177/07591063221088322.
69. Pour rappel, Common Voice est un projet *open source* de Mozilla qui fournit une base de données libre d'enregistrements vocaux afin d'entraîner et d'améliorer les technologies de reconnaissance vocale.
70. FLEURS, mis au point par Google Research (aujourd'hui Google DeepMind), en collaboration avec des chercheurs de Meta AI, est un jeu de données informatique multilingue de parole, conçu pour évaluer et entraîner des systèmes de reconnaissance et de compréhension de la parole dans plus de 100 langues.

données de test : si des phrases similaires ont été rencontrées lors de l'apprentissage, le WER peut apparaître artificiellement bas.

Dans cette perspective, l'évaluation la plus probante consiste souvent à tester les systèmes à partir d'enregistrements spécifiques, produits dans des conditions réelles d'usage, dans l'esprit de l'exemple proposé précédemment.

La mesure des erreurs de diarisation

La diarisation vise à détecter et à identifier les locuteurs au sein d'un enregistrement. Elle est un processus déterminant pour la production de transcriptions fidèles et compréhensibles, et constitue une tâche chronophage lors de la correction d'une transcription initiale. Elle a pour particularité que, pour l'essentiel des changements de locuteur, la détection repose sur le signal audio, et non sur des indicateurs dans les mots prononcés. À l'instar de la transcription, la diarisation possède sa propre métrique d'évaluation : le *Diarization Error Rate* (DER).

Le principe est analogue à celui du *Word Error Rate* : il s'agit de comparer une annotation de référence à la prédiction du système. Cette comparaison fait apparaître trois types d'erreurs : la confusion de locuteur, lorsque la parole est attribuée à la mauvaise personne ; la fausse alarme, lorsque le système détecte à tort une prise de parole ; et la parole manquée, lorsqu'un segment effectivement prononcé n'est pas identifié.

> La diarisation vise à détecter et à identifier les locuteurs au sein d'un enregistrement. Elle est un processus déterminant pour la production de transcriptions fidèles et compréhensibles.

Le DER est ensuite calculé en rapportant la durée cumulée de ces erreurs à la durée totale de parole. Il constitue aujourd'hui l'indicateur de référence pour évaluer la performance des systèmes de diarisation.

4. L'importance et les limites des métriques standards

Il existe bien d'autres métriques que les deux précitées, qui occupent une place centrale dans les publications et les *benchmarks* et structurent les

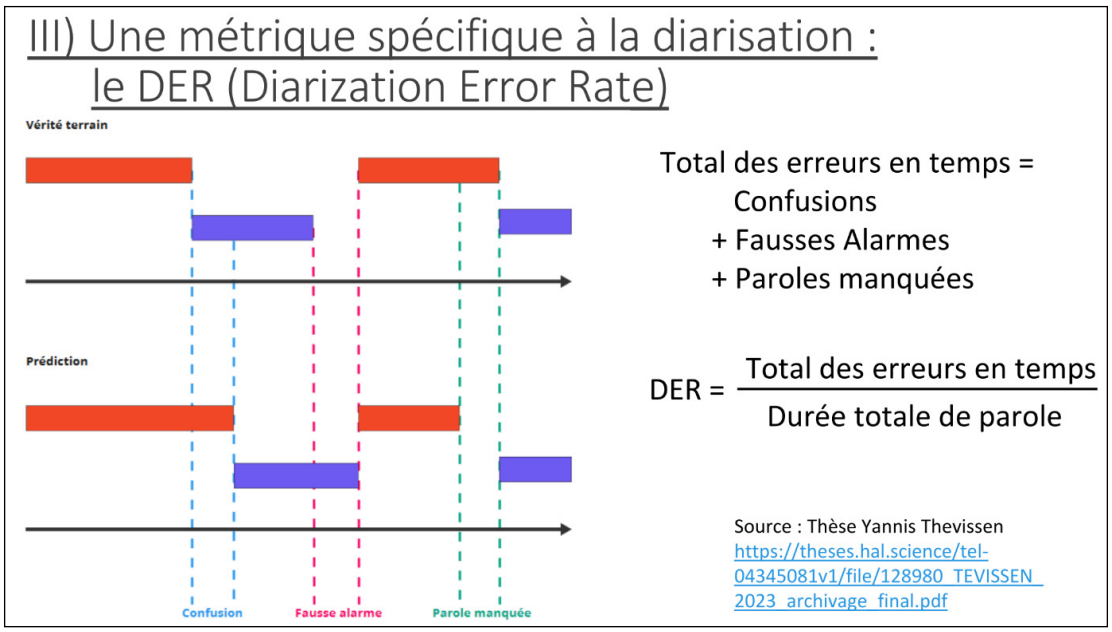


Figure : mesurer les erreurs de diarisation
Source : Yannis Thevissen

comparaisons entre systèmes. Elles fournissent des indicateurs quantifiés de performance, mais leur interprétation en situation réelle appelle à la prudence.

En principe, l'évaluation repose sur une séparation stricte entre données d'entraînement et données de test, distinctes mais comparables en termes de représentativité. À défaut, les mesures perdent en pertinence. L'analogie est simple : entraîner un système avec un corpus donné et l'évaluer avec ce même corpus conduit nécessairement à des performances biaisées, contrairement à une évaluation menée avec des données différentes.

« Entraîner un système avec un corpus donné et l'évaluer avec ce même corpus conduit nécessairement à des performances biaisées, contrairement à une évaluation menée avec des données différentes.

Or, avec les approches génératives récentes, l'origine des données d'entraînement n'est pas toujours identifiée. Si les données d'évaluation se retrouvent, même partiellement, dans les données d'apprentissage, les résultats peuvent être artificiellement élevés, le système ayant déjà été exposé aux exemples testés. On observe parfois des écarts de performance particulièrement marqués entre deux versions successives d'un même modèle, avec des gains spectaculaires pouvant dépasser largement les constats ressortant d'un état de l'art.

Ces résultats peuvent s'expliquer, dans certains cas, par une contamination des données d'évaluation par les données d'entraînement comme on l'a vu précédemment. Des comportements atypiques sont également susceptibles d'apparaître en fonction de la taille des modèles : ainsi, alors qu'une progression graduelle est attendue, certaines configurations montrent des sauts qualitatifs très importants entre différentes tailles de modèle sans justification évidente. À l'inverse, lorsque les corpus sont plus courants et donc mieux maîtrisés, les améliorations tendent à suivre une évolution plus régulière, mieux conforme aux attentes.

Cela renvoie à une question essentielle : celle de l'écart entre les erreurs mesurées par les métriques et les attentes réelles des utilisateurs. En pratique, l'impact d'une erreur dépend étroitement du contexte d'usage. Ainsi, dans le cadre de la transcription d'une réunion, la disparition d'un mot ou une légère approximation reste généralement acceptable, dès lors que l'objectif est d'obtenir un compte rendu ou un résumé global. Les utilisateurs font alors preuve d'une certaine tolérance. En revanche, dans d'autres contextes, les exigences peuvent être sensiblement plus élevées.

5. Le classement des modèles

S'agissant de la comparaison des modèles, il est utile de se référer à des classements publics, tels que le classement « Reconnaissance de la parole »[71] administré par la société Hugging Face, qui est une plateforme de distribution et de partage de modèles d'IA. Pour un utilisateur qui ne suit pas en continu l'actualité des modèles, ce type de classement constitue un outil précieux pour appréhender l'état de l'art. Il permet notamment d'observer certaines tendances récentes, comme la publication par la société Nvidia de nombreux modèles de petite taille.

L'analyse de ces classements suppose toutefois de prêter attention au contenu et au périmètre des modèles évalués. Ainsi, certains systèmes peuvent afficher des performances remarquables tout en étant limités à une seule langue. C'est par exemple le cas du modèle Parakeet, qui totalise environ 600 millions de paramètres et présente un WER très faible, mais qui fonctionne uniquement en anglais. D'autres modèles, légèrement plus volumineux, offrent également d'excellents résultats tout en couvrant plusieurs langues.

Le classement précité offre une vue synthétique des performances des différents systèmes disponibles. À l'occasion de sa dernière mise à jour en août 2025, on a pu observer que la valeur du WER pour les 10 meilleurs modèles est très proche, entre 5,42 et 6,02. Les performances moyennes pour la langue française sont globalement meilleures (entre 3,28 et 5,42). En revanche, les performances sur les formes longues en anglais sont sensiblement moins bonnes (entre 7,32 et 11,18), en particulier à cause des variations de prononciation, des

71. *Leaderboard Speech Recognition*, https ://huggingface.co/spaces/hf-audio/open_asr_leaderboard.

langues de spécialités et de la présence d'un grand nombre d'entités nommées.

Outre le *Word Error Rate* (WER), un indicateur souvent mobilisé est *Real-Time Factor inverse* (RTFx). Celui-ci met en relation la durée de l'audio et le temps de calcul nécessaire pour produire la transcription et permet ainsi d'apprécier la rapidité du modèle. Il indique concrètement si un système est en mesure de produire une réponse rapidement ou s'il nécessite un temps de traitement plus important.

Il existe par ailleurs des variantes de Whisper, telles que Crisper Whisper qui figure parmi les modèles les mieux classés. Cette variante vise à corriger un biais du modèle initial, entraîné avec des données souvent formelles et tendant, de ce fait, à « rehausser » le niveau de langue dans les transcriptions. Concrètement, des expressions familières peuvent être transformées en formulations plus normées. À l'inverse, Crisper Whisper cherche à restituer au plus près la parole réelle, en conservant notamment les disfluences – répétitions, hésitations –, ce qui contribue également à ses bons résultats au regard de certaines métriques.

6. Tolérance aux erreurs et contextes d'utilisation

Malgré les progrès réalisés, les métriques et les classements montrent que ces technologies ne sauraient être considérées comme infaillibles. Ces contraintes imposent d'adapter les systèmes aux usages. Dans un contexte judiciaire, la transcription doit nécessairement faire l'objet d'une relecture et d'une correction humaines. À l'inverse, pour des usages tels que le résumé de réunions, une tolérance plus grande aux erreurs peut être admise.

L'exemple de la transcription d'une audition de mineur illustre des exigences nettement plus strictes. Dans ce cadre, la fidélité doit être absolue : une erreur, même minime, peut altérer profondément le sens d'une déclaration, au point d'inverser les rôles entre victime et personne mise en cause. Pour la police judiciaire, les transcriptions des auditions constituent une obligation légale et une part significative de l'activité, tout en étant particulièrement coûteuses en termes de temps passé lorsqu'elles sont réalisées manuellement. Les difficultés sont accrues par la diction parfois incertaine des enfants, des conditions d'enregistrement dégradées – par exemple, en milieu hospitalier ou en cas d'utilisation de dispositifs de captation non spécialisés – et l'absence d'environnement acoustique maîtrisé.

> « Malgré les progrès réalisés, les métriques et les classements montrent que ces technologies ne sauraient être considérées comme infaillibles.

Sur la reprise des erreurs

L'expérience des plateformes traitant de grands volumes de transcriptions révèle que, dans la majorité des cas où le taux d'erreurs est anormalement élevé, l'origine des difficultés tient souvent à une qualité de la prise de son insuffisante. Un autre facteur tient à la distance excessive entre les corpus d'entraînement et les segments audio soumis. Un modèle entraîné avec des textes lus en studio aura des performances amoindries pour des conversations multi locuteurs très dynamiques.

Les retours d'expérience montrent également les difficultés liées aux environnements multilingues. Parfois un système de détection automatique de la langue est mis en place pour dispenser l'utilisateur de toute indication préalable. Une analyse des premières secondes de l'enregistrement – environ trente secondes – permet d'identifier la langue utilisée. Une erreur de détection initiale, qui peut se produire si l'enregistrement débute par du silence, du bruit, un propos liminaire dans une autre langue ou une activation prématurée du microphone, peut empêcher une identification fiable de la langue. Dans ce cas, l'ensemble de la transcription peut s'en trouver affectée au point de devenir totalement incohérente. Il peut en être de même en cas d'alternances de langues différentes parlées dans un même enregistrement.

Avec Whisper, un autre phénomène apparaît dès lors que le système se trouve en difficulté : celui des « hallucinations ». Dans ces situations, par exemple en cas d'un enregistrement bruité, la propension de ce type d'outil à générer à tout prix une réponse, y compris à partir de segments tout à fait inaudibles, favorise soit la génération de textes aléatoires, soit la génération du texte le plus probable sur ce segment. Le modèle peut alors générer, sur plusieurs lignes, un texte incohérent, avant de tenter de se réaligner avec

le contexte, ce qui rend l'ensemble de la transcription particulièrement instable.

Si l'on se place du point de vue du prestataire technique, il ne s'agit pas d'une défaillance intrinsèque du service, mais d'une limite intrinsèque du système, documenté comme n'étant pas parfait. Un risque déceptif émerge si les caractéristiques techniques du système mis en place sont mal documentées ou ne sont pas communiquées aux utilisateurs, auxquels revient la charge de la correction finale.

Sur les temps de reprise

Les taux d'erreurs, qui paraissent encore aujourd'hui largement irréductibles, impliquent des processus de reprise et de correction, dès lors que sont attendus des résultats exacts. D'après l'expérience convergente des intervenants, basée sur des centaines d'heure de transcriptions, le temps moyen de reprise en vue d'obtenir une transcription correcte de l'ordre de 2 heures pour 1 heure d'audio, contre des ratios de 8 heures de transcription pour 1 heure d'audio en mode purement manuel. Avec son premier outil de transcription automatique, la Gendarmerie nationale a donc économisé des milliers d'heures de travail.

> « D'après l'expérience convergente des intervenants, basée sur des centaines d'heure de transcriptions, le temps moyen de reprise en vue d'obtenir une transcription correcte de l'ordre de 2 heures pour 1 heure d'audio, contre des ratios de 8 heures de transcription pour 1 heure d'audio en mode purement manuel.

Les gains de temps généralement annoncés correspondent à des conditions d'usage optimales – enregistrement de qualité, parole claire, environnement maîtrisé. Dès que ces conditions se dégradent, les performances diminuent sensiblement et les bénéfices attendus peuvent être fortement réduits.

Dans un cas particulièrement difficile, avec des personnes âgées s'exprimant de manière informelle dans un patois local, la reprise est plutôt de l'ordre de 6 à 7 heures de transcription pour une heure d'audio, quasiment de l'ordre d'un travail manuel. Il apparaît ainsi que des gains de temps très significatifs sont possibles dans les situations dans lesquelles une transcription complète est d'ores et déjà requise, comme les procès-verbaux d'enquêteurs.

7. Les contraintes de déploiement

Les difficultés supplémentaires liée à transcription immédiate

Parmi les contraintes associées à la réalisation d'une transcription automatique, un paramètre opérationnel important est celui du délai de mise à disposition du résultat de la transcription. On distingue principalement deux situations : la première est la transcription dite « en différé » (*batch* en anglais) à partir d'un enregistrement audio terminé, qui est soumis à une chaîne de traitement en vue de produire la transcription. La durée de retour de la transcription automatique est alors égale à la somme du temps de transmission du fichier audio, puis de l'analyse automatique. Des facteurs comme la puissance de calcul disponible – en particulier sur un ordinateur personnel ou sur une infrastructure mutualisée, la mise en place de files d'attentes pour les traitements à opérer ou l'utilisation de capacités de calcul en heures creuses peuvent se traduire par des délais de l'ordre de plusieurs heures à plusieurs jours, étant entendu que des modèles plus performants sont aussi plus consommateurs de ressources.

À l'autre extrémité du spectre, la seconde situation correspond à certains modèles produisant la transcription dans des délais très brefs : on parle alors de mode « flux » (ou *streaming*). Ceux-ci produisent des segments de texte au fur et à mesure que les fragments audio leur parviennent. Ils nécessitent de disposer d'une puissance de calcul réservée pour pouvoir fonctionner, qu'elle soit locale, à l'instar d'un ordinateur disposant d'une carte graphique, ou distante, comme une infrastructure de calcul partagée. Cette contrainte du temps réel se traduit par une qualité de transcription moindre : la puissance de calcul disponible étant souvent plus limitée, il n'est pas possible de disposer des meilleurs algorithmes.

La différence de performance entre le mode *batch* et le mode *streaming* tient aussi à des

raisons algorithmiques, en lien avec la possibilité d'accès au contexte futur. Dans un scénario différé, le modèle possède une « connaissance parfaite » de l'avenir acoustique, ce qui lui permet de désambiguïser des segments de parole en fonction des mots prononcés plus tard dans la phrase. En *streaming*, l'attention doit être restreinte au passé. Or l'utilisation d'une attention purement causale entraîne souvent une dégradation significative des métriques. Les écarts peuvent être importants : la société Deepgram, qui fournit un service de transcription performant, affirme que le traitement en différé améliore le score WER de 10 à 17 points[72]. Pour compenser cela, certains chercheurs explorent des techniques mixtes[73]. Des efforts de recherche tendent à élaborer des modèles capables de fonctionner selon les deux approches[74].

Dans un scénario différé, le modèle possède une « connaissance parfaite » de l'avenir acoustique, ce qui lui permet de désambiguïser des segments de parole en fonction des mots prononcés plus tard dans la phrase. En *streaming*, l'attention doit être restreinte au passé.

Des modèles plus grands ou plus petits : le défi de l'efficacité

La question de la performance est étroitement liée à la taille des modèles mobilisés et donc à la puissance de calcul consommée. Cette contrainte devient particulièrement critique pour les usages en temps réel, tels que la transcription instantanée. Ainsi, si des modèles de grande taille – comme certaines versions complètes de Whisper – produisent des résultats de haute qualité, leur déploiement en temps réel nécessite des infrastructures de calcul importantes.

À l'inverse, des modèles plus compacts permettent des temps de réponse nettement plus rapides et facilitent le développement d'applications opérationnelles en temps réel. La question de la taille des modèles dépasse ainsi la seule dimension environnementale pour relever pleinement d'enjeux d'efficacité et d'usage.

De petits écarts de puissance nécessaire pour une transcription peuvent avoir une incidence dans l'hypothèse de grands volumes d'utilisation quand ces outils sont déployés à large échelle. L'exploitation de modèles de grande taille suppose des investissements matériels particulièrement élevés : une machine équipée de plusieurs GPU de dernière génération peut atteindre environ 200 000 euros, tandis que certaines configurations plus avancées s'élèvent jusqu'à 500 000 euros.

Un autre défi réside dans l'arbitrage entre précision et rapidité d'exécution. L'expérience acquise, notamment dans des contextes industriels comme la catégorisation de requêtes, illustre cette tension. Si les modèles de dernières générations offrent les meilleures performances, leur coût en temps de calcul peut devenir prohibitif : un temps de réponse de l'ordre de 200 millisecondes par requête, à raison de plusieurs milliers de requêtes par seconde, implique une infrastructure considérable, difficilement soutenable économiquement.

Dans ce contexte, le choix s'oriente souvent vers des modèles plus simples, légèrement moins précis, mais nettement plus rapides – de l'ordre de 100 à 200 fois. Une perte marginale de précision peut ainsi être jugée acceptable dès lors qu'elle permet un déploiement sur des infrastructures moins coûteuses et la gestion de volumes massifs de requêtes.

En environnement de production, ces coûts sont encore accrus par les exigences de redondance nécessaires à la fiabilité des systèmes, conduisant rapidement à des investissements de plusieurs millions d'euros pour faire fonctionner un modèle. Dans ce contexte, l'intérêt pour la miniaturisation des modèles et le développement d'architectures plus légères ne cesse de croître.

72. Selon le fournisseur Deepgram, « batch transcription usually improves WER by 10–17 points thanks to full context access ». Voir « Speech Recognition Accuracy : Production Metrics & Optimization 2025 », *Deepgram blog*, 2026, https ://deepgram.com/learn/speech-recognition-accuracy-production-metrics.
73. Pour produire la transcription avec un léger différé, avec des mécanismes dits de « *look ahead* » (contexte futur limité) ou de « *chunked attention* »» (attention par blocs).
74. Sharma, Bidisha, Karthik Pandia Durai, Shankar Venkatesan *et al.*, « Unifying Streaming and Non-streaming Zipformer-based ASR » *ArXiv* abs/2506.14434, 2025.

Améliorer les performances de transcription automatique en production

La bonne connaissance de l'ensemble de ces contraintes permet d'envisager des actions pour accroitre la qualité de la transcription et réduire en regard les temps de reprise.

Pour cela, une première démarche est de procéder à des évaluations des systèmes envisagés ou retenus à partir d'échantillons de données les plus proches possibles des conditions d'exploitation. Les *benchmarks* de référence sont composés de segments audio souvent peu représentatifs des conditions d'usage réel. L'objectif doit être de mesurer la performance opérationnelle et d'éviter de prendre des décisions sur la base de mesures théoriques. Selon la société Deepgram, un corpus d'évaluation de 1 à 5 heures d'audio le plus proche possible de la diversité des conditions opérationnelles permet déjà des évaluations fiables. Les évaluations ne doivent pas se limiter au WER comme indicateur.

Le principal facteur de dégradation des performances étant la qualité du signal enregistré, l'ensemble des efforts pour améliorer la qualité de la captation audio et du post-filtrage engendre des gains significatifs en qualité de la transcription. Dans les domaines spécialisés, comme la médecine ou le droit et la justice, les mécanismes de renforcement du vocabulaire métier autorisent également des gains rapides. Des procédures d'affinage, à partir de corpus pouvant aller de 10 à 100 heures d'enregistrements audio, produisent aussi des améliorations significatives, en particulier pour surmonter les faiblesses des modèles génériques en présence de vocabulaires métiers. En complément, des procédures de post-correction des transcriptions brutes correctement définies peuvent encore produire dans certains cas des améliorations significatives de la qualité des textes produits[75].

75. Article de Deepgram, *op. cit.*

Atelier n°5
D'une langue à l'autre : performance et enjeux de la traduction automatisée dans le champ du droit et de la justice

- Quels usages de la traduction automatique par les professionnels du droit ?
- Des clés pour comprendre la traduction automatique
- Limites et perspectives de la traduction automatique
- Enjeux juridiques du déploiement de solutions de traduction automatique dans la sphère judiciaire

Dans un pays comme la France, avec un État unitaire et centralisé et une langue nationale unique, les questions de la traduction sont souvent périphériques au fonctionnement des services publics. Pourtant les réalités contemporaines de circulation accrue des populations tendent à renforcer les situations de multilinguisme. Face à ces évolutions, l'examen des zones géographiques déjà multilingues peut être une source d'inspiration pour la recherche d'expériences, de solutions concrètes mais aussi de principes. Par exemple, émerge la notion de droit à la traduction, comme un défi majeur pour la citoyenneté participative et pour l'intégration[76].

Le champ judiciaire quant à lui fait un peu exception. La traduction et l'interprétation y sont des enjeux anciens[77]. Depuis l'Antiquité, la justice a été confrontée à la diversité linguistique et elle a dû recourir à des médiateurs du langage pour garantir l'équilibre du procès. Ce besoin élémentaire est aujourd'hui consacré comme un droit fondamental par de nombreux textes internationaux et européens, notamment l'article 6 de la Convention européenne des droits de l'Homme ou la directive 2010/64/UE relative au droit à l'interprétation et à la traduction dans le cadre des procédures pénales.

Malgré l'importance pour les ordres judiciaires de faire face à la diversité linguistique pour garantir l'effectivité de ce droit fondamental, la question spécifique de la traduction judicaire reste pourtant peu abordée dans les travaux de recherche. S'il en fallait un signe, parmi les 43 volumes de la collection de référence *Routledge Handbooks in Translation and Interpreting Studies*[78], aucun ne porte sur la question de la traduction judiciaire[79].

> Malgré l'importance pour les ordres judiciaires de faire face à la diversité linguistique pour garantir l'effectivité de ce droit fondamental, la question spécifique de la traduction judicaire reste pourtant peu abordée dans les travaux de recherche.

Derrière cet impératif juridique se cachent pourtant des difficultés concrètes et croissantes.

76. Reine Meylaerts, « Transnational Justice in a Multilingual World: An Overview of Transnational Regimes », *Meta*, 2012, vol. 56, n° 4, p. 743-757.
77. Dossier « Les langues du procès », *Les Cahiers de la Justice*, 2024, n° 2, https ://droit.cairn.info/revue-les-cahiers-de-la-justice-2024-2 ?lang=fr.
78. Manuels Routledge en études de traduction et d'interprétation.
79. Voir chez l'éditeur Routledge la collection *Handbooks in Translation and Interpreting Studies* (https ://www.routledge.com/Routledge-Handbooks-in-Translation-and-Interpreting-Studies/book-series/RHTI).

Le recours à des traducteurs et interprètes qualifiés se heurte à de nombreux obstacles : délais de recrutement, pénurie de compétences, niveaux de rémunération souvent inférieurs aux standards du marché. Ce déséquilibre s'inscrit dans un contexte où les besoins, loin de diminuer, sont appelés à croître.

C'est en partie pour ces raisons que, lors de la consultation qu'il a organisé autour des cas d'usages de l'intelligence artificielle, le ministère de la Justice français et les parties prenantes consultées ont identifié la traduction et l'interprétariat comme des applications des technologies d'intelligence artificielle susceptibles d'apporter des bénéfices concrets et immédiats.

« Le ministère de la Justice français et les parties prenantes consultées ont identifié la traduction et l'interprétariat comme des applications des technologies d'intelligence artificielle susceptibles d'apporter des bénéfices concrets et immédiats.

Similaires en apparence par leur objet visant le passage d'une langue à une autre, la traduction s'attache à produire un écrit à partir d'un premier écrit et l'interprétation à produire un discours oral à partir d'un discours oral initial, en plus d'être soumise à des contraintes d'immédiateté et d'intégration dans un espace d'interaction. Ces différences sont telles que les professionnels pratiquant ces deux activités ne sont pas les mêmes, que les instruments techniques pour leur numérisation reposent sur des principes distincts et que les outils n'ont pas la même maturité technologique. Ces écarts rendent difficile une analyse commune, ce qui nous amène à nous consacrer ici uniquement aux applications permettant la production d'écrits par traduction automatique.

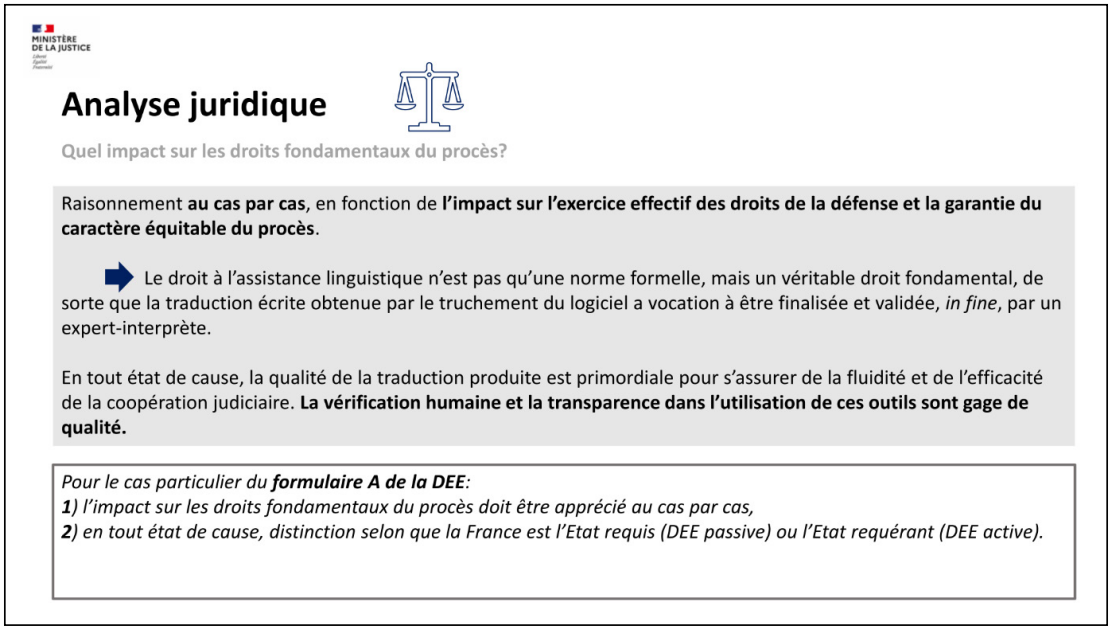
En cette matière, les technologies ont fait des progrès considérables au cours des deux dernières décennies, au point d'en devenir omniprésentes, presque banales, voire transparentes : les navigateurs intègrent gratuitement des fonctions de traduction automatique, les téléphones de dernière génération traduisent en temps réel les textes qui apparaissent sur une photographie, des systèmes de visioconférence proposent de traduire en simultané les sous-titres générés eux-mêmes automatiquement.

Plus discrètement, mais avec un impact considérable sur leur métier, les outils d'assistance à la traduction sont très mobilisés aujourd'hui par les professionnels de la traduction. Ils intègrent déjà très largement les techniques d'intelligence artificielle. C'est au point que, pour certains cas d'applications, le métier de traducteur se transforme à grand pas vers celui de la post-édition, c'est-à-dire de la correction d'écrits produits par des modèles d'IA. Ces évolutions se manifestent par une pression à la baisse sur les prix et sur les délais reposant sur l'idée, contestée par les traducteurs, qu'à qualité égale, ces nouveaux dispositifs de travail rendraient la traduction plus rapide et donc moins coûteuse.

Cette situation nouvelle de professionnels devenus correcteurs des propositions d'un système de traduction automatique est l'une des manifestations du principe cardinal de la supervision humaine des systèmes d'intelligence artificielle promu très largement comme la condition de leur déploiement. Que l'on parle de « *human control* » ou de manière plus floue de « *human in the loop* », les questions de contrôle effectif des productions des systèmes utilisant l'IA et de la responsabilité de celui qui en mobilise les résultats sont absolument centrales. Elles dépassent largement la problématique technologique pour aller vers l'interaction entre l'humain et les machines, ouvrant des questionnements d'organisation du travail, de compétences et de cognition.

Un contrôle effectif présuppose un niveau d'expertise suffisant du superviseur humain. En matière de traduction, le contrôle humain ne pourrait alors être opéré que par un traducteur correctement formé, et uniquement pour des productions vers sa langue maternelle. Apparaît ici de manière explicite la tension entre cet impératif de contrôle et la très large accessibilité de ces outils, qui sont utilisés de façon autonome par des personnes parfaitement incompétentes dans la langue source considérée.

Dans le cadre du procès, cette considération éthique rencontre des impératifs procéduraux : le devoir pour les parties de procéder à la traduction de certaines pièces en langue étrangère, l'obligation pour l'État de traduire certaines pièces pour les parties non francophones, mais aussi d'avoir recours à un traducteur expert pour


MINISTÈRE DE LA JUSTICE

Analyse juridique

Quel impact sur les droits fondamentaux du procès?

Raisonnement **au cas par cas**, en fonction de **l'impact sur l'exercice effectif des droits de la défense et la garantie du caractère équitable du procès**.

Le droit à l'assistance linguistique n'est pas qu'une norme formelle, mais un véritable droit fondamental, de sorte que la traduction écrite obtenue par le truchement du logiciel a vocation à être finalisée et validée, *in fine*, par un expert-interprète.

En tout état de cause, la qualité de la traduction produite est primordiale pour s'assurer de la fluidité et de l'efficacité de la coopération judiciaire. **La vérification humaine et la transparence dans l'utilisation de ces outils sont gage de qualité.**

*Pour le cas particulier du **formulaire A de la DEE**:*
***1)** l'impact sur les droits fondamentaux du procès doit être apprécié au cas par cas,*
***2)** en tout état de cause, distinction selon que la France est l'Etat requis (DEE passive) ou l'Etat requérant (DEE active).*

Le recours à la traduction automatique doit intégrer la préservation des droits fondamentaux du procès.
Source : secrétariat général, ministère de la Justice

certains actes. Cette obligation d'expertise est l'expression de l'importance d'une traduction fidèle des actes essentiels pour l'exercice de leurs droits par les parties et pour la qualité de la décision de justice. En l'état actuel du droit, elle entraîne l'impossibilité d'avoir recours à une traduction automatique pour traduire ces éléments.

Face aux problématiques multiples de traduction auxquelles ils font face, les acteurs judiciaires s'interrogent : des usages de la traduction automatique sans supervision sont-ils possibles, souhaitables ou faut-il redéfinir le périmètre de la traduction humaine obligatoire ? Cette dernière question ne peut être adressée sans s'interroger sur les performances effectives de ces outils, et surtout sur le niveau de qualité attendu selon chacune des situations d'usage. Dans certains cas non critiques, des erreurs ou des approximations sont potentiellement admissibles à la condition toutefois d'être perceptibles. Dans d'autres, elles seraient beaucoup trop lourdes de conséquences, à l'instar des traductions mobilisées par un juge pour régler un litige en fait ou en droit

Dans le cadre judiciaire, les outils de traduction automatique ont vocation à traiter des données nécessitant une protection : en premier lieu, les données personnelles souvent sensibles au sens du RGPD, comme des éléments de santé, d'appartenance politique ou syndicale, d'origine ethnique ; en deuxième lieu,

« Face aux problématiques multiples de traduction auxquelles ils font face, les acteurs judiciaires s'interrogent : des usages de la traduction automatique sans supervision sont-ils possibles, souhaitables ou faut-il redéfinir le périmètre de la traduction humaine obligatoire ?

des données porteuses d'enjeux de souveraineté ; en troisième lieu, des données relevant du secret des affaires, dont la révélation pourrait être de nature à obérer gravement les avantages concurrentiels des acteurs concernés .

Pour éclairer l'ensemble de ces questionnements, il est nous a semblé important, s'agissant de systèmes techniques, de fournir quelques clés pour comprendre la traduction automatique, afin d'appréhender, avec autant d'exactitude que possible pour un profane, les principes techniques sous-jacents, le niveau de performance, les limites et la dynamique de la

recherche scientifique et des développements industriels dans ce domaine.

Au-delà des aspect techniques et humains, les enjeux juridiques du déploiement de solutions de traduction automatique dans la sphère judiciaire sont également déterminants. Ils ont trait pour une large part à la nature et aux différents régimes de protection des données auxquels ils s'appliquent. Ces enjeux sont d'autant plus intéressants qu'ils présentent un certain niveau de généralité, la plupart des contraintes identifiées pouvant concerner aussi bien d'autres applications des technologies d'intelligence artificielle.

I - Des clés pour comprendre la traduction automatique

Dans cette première partie, nous soulèverons le voile des mécanismes de traduction automatique, qui sont extrêmement similaires à ceux mis en œuvre pour créer des robots conversationnels de type Mistral, ChatGPT ou Claude. Cette appréhension sommaire vise à donner quelques clés de compréhension des systèmes de traduction utilisant l'IA[80].

En premier lieu, ces systèmes portent une part d'imprévisibilité et de mystère, pour leurs utilisateurs mais aussi pour leurs créateurs. Certes les techniques mathématiques élémentaires utilisées pour les construire sont parfaitement définies, mais la représentation du sens qu'ils manipulent sous des formes numériques échappe à l'entendement humain, ce qui les rend difficiles à contrôler. Cela tient à leur double nature statistique et probabiliste, dans les phases d'apprentissage comme d'interaction avec l'utilisateur. L'effet d'opacité est amplifié par la taille considérable des modèles et des corpus d'entraînement qui renforce l'effet boîte noire. Sous certains aspects les chercheurs et ingénieurs en intelligence artificielle se rapprochent des biologistes et zoologues, tous observateurs de systèmes complexes qu'ils ne comprennent que partiellement.

En second lieu, ces systèmes sont construits et élaborés. Ils sont le résultat de choix opérés par les acteurs qui les conçoivent et des arbitrages rendus en fonction des objectifs que ces derniers se sont fixés. Ces processus concrétisent un ensemble de méthodes utilisées pour éduquer, instruire et former, selon une forme de pédagogie, qui manifeste le choix de données d'entraînement parmi une multitude de données disponibles, la préparation de ces données et leur mise en ordre, durant la phase d'apprentissage mais aussi durant les phases ultérieures d'alignement ou d'apprentissage par renforcement[81].

> Ces systèmes sont construits et élaborés. Ils sont le résultat de choix opérés par les acteurs qui les conçoivent et des arbitrages rendus en fonction des objectifs que ces derniers se sont fixés.

1. Vers la traduction

Les recherches sur la traduction automatique commencent dès les débuts de l'informatique avec les premiers écrits de Warren Weaver en 1947 et l'organisation de la première conférence sur ce thème en 1952 par Yehoshua Bar-Hillel.

Très tôt, ces travaux ont identifié le principe, toujours actuel, du recours à une représentation informatique intermédiaire des écrits, pour aborder la traduction par une approche en deux étapes : analyser la phrase source pour en extraire le sens sous la forme de cet intermédiaire informatique, puis générer à partir de cette dernière une nouvelle phrase dans la langue cible.

Des années 50 aux années 90, cette stratégie de complexification des niveaux d'analyse visait à construire des représentations toujours plus abstraites et à rendre la génération de

80. Le propos reprend très largement les développements de François Yvon, lors de son intervention dans le cadre de l'atelier tenu à l'IRB le 19 juin 2025, ainsi que dans son article « La traduction multilingue : analyse d'une prouesse technologique », complétés par d'autres articles scientifiques et les échanges oraux lors de cet atelier. Voir François Yvon, « La traduction multilingue : analyse d'une prouesse technologique ». *mediAzioni*, 27 décembre 2023, A17-A34 Pages. https ://doi.org/10.6092/ISSN.1974-4382/18785.

81. L'alignement désigne l'ensemble des techniques visant à faire en sorte que le comportement d'un système d'intelligence artificielle corresponde aux intentions humaines. Parmi ces techniques, l'apprentissage par renforcement (*reinforcement learning*) consiste à entraîner un modèle en lui attribuant des récompenses ou des pénalités en fonction de la qualité de ses réponses. Cette approche est souvent combinée à des retours humains (RLHF), permettant d'orienter le système vers des résultats jugés utiles, fiables et conformes aux attentes.

> Des années 50 aux années 90, cette stratégie de complexification des niveaux d'analyse visait à construire des représentations toujours plus abstraites et à rendre la génération de plus en plus longue et fluide.

plus en plus longue et fluide. Celle-ci s'opérait essentiellement avec des méthodes reposant sur des règles, des formalismes symboliques, des grammaires, partageant un principe commun de déterminisme.

L'approche statistique, qui a émergé dans les années 90, a été rendue possible par la disponibilité croissante de corpus parallèles de traductions bilingues réalisées par des humains, l'augmentation de la puissance de calcul des ordinateurs, le développement de méthodes d'évaluation automatique pour mesurer la performance des systèmes et surtout le développement des algorithmes d'apprentissage statistique, qui ont été appliqués à partir des années 2000 aux réseaux de neurones profonds et ont constitué une rupture majeure.

2. Créer des systèmes de traduction à réseau de neurones

Au cœur de la traduction neuronale, de manière similaire aux mécanismes que l'on retrouve dans les systèmes conversationnels de type GPT (*Generative Pretrained Transformer*[82]), se trouvent deux phases :

- la première, qui est la plus complexe, est celle de l'apprentissage : à partir de corpus parallèles, les concepteurs optimisent les paramètres du modèle pour réaliser la phase suivante de génération de manière efficace ;
- la seconde, qui est la phase de génération, part d'une phrase source et construit progressivement un équivalent en langue cible, en sélectionnant par itérations successives une traduction probable.

Que désigne le terme « transformer » ?

- L'algorithme actuellement dominant pour la génération s'appelle le « transformer ». À partir essentiellement d'une phrase source ou d'un début de phrase et à travers une série d'opérations mathématiques, le transformer construit une représentation numérique sous la forme d'un vecteur obtenu par combinaison de chacun des vecteurs des mots composant le contexte. Cette combinaison vise à représenter dans un espace numérique de grande dimension l'essentiel de l'information contenue dans la phrase source et la phrase cible.
- Pour opérer ces transformations, un modèle est caractérisé par un nombre considérable de valeurs numériques, appelées **paramètres** ou **poids**. Leurs valeurs définissent les calculs qui sont opérés sur la représentation numérique du texte.
- Un composant très important de ce calcul s'appelle l'**attention**. Il s'agit d'un mécanisme utilisé pour différencier et hiérarchiser les mots ou les symboles du contexte en entrée, afin de repérer, à chaque instant, quels éléments sont les plus importants pour faire la prédiction. Ces opérations sont très consommatrices de puissance computationnelle, qui croît de manière exponentielle avec la taille du texte fourni dans le *prompt*.
- Les différents calculs d'attention, paramétrés suivant les poids issus de l'apprentissage, permettent la prise en compte des multiples niveaux d'interdépendance entre les mots du contexte. Sans qu'il soit possible d'associer directement et simplement chaque niveau d'attention à une interdépendance particulière, ces mécanismes autorisent la prise en compte de règles linguistiques, de style, de niveau de langue mais aussi de domaines d'expression.

82. Système génératif pré-entraîné.

3. Comment la traduction bilingue est-elle construite ?

L'algorithme de génération ou de décodage repose sur un principe simple : la prédiction du mot suivant étant donné l'état courant du contexte. Le texte final est construit par l'ajout du nouveau mot au contexte et la répétition du processus de prédiction. Par exemple, si le seul contexte est le mot « bon », de nombreux mots peuvent statistiquement suivre pour produire un énoncé correct en français. Mais si le contexte est plus précis (« je suis de bonne… »), la distribution des mots possibles se réduit et il n'y en a plus que quelques-uns qui sont très probables (comme « humeur » ou « composition »).

Pour un système de traduction, le contexte fourni en entrée au système pour générer la traduction est légèrement différent de celui d'un modèle conversationnel : il se compose de la phrase à traduire dans la langue source, suivie du fragment de la phrase déjà traduite.

Pour prédire, le modèle sélectionne un mot suivant parmi les mots les plus probables. Pour ce choix, un léger aléa est introduit, pour éviter un comportement trop systématique. Cette dose de hasard peut être renforcée ou diminuée par un paramètre que l'on appelle la **température**. Ce mécanisme est la principale cause d'un phénomène que tous les utilisateurs observent : si l'on répète la même sollicitation plusieurs fois, la réponse sera à chaque fois dans une certaine mesure différente.

Qu'est-ce que la température dans un modèle ?

À chaque étape de la génération, après avoir appliqué les paramètres, le modèle dresse la liste ordonnée des mots suivants les plus probables. Il ne retient pas systématiquement celui dont la probabilité est la plus forte, mais un parmi les n premiers. Pourquoi ? Il s'agit d'améliorer la qualité des textes produits, réduire la production de textes très stéréotypés, limiter le risque de répétitions et augmenter la résilience du modèle après une erreur.
La température est un nombre qui influe sur le choix opéré par le modèle. Inférieure à 1, le modèle tendra à favoriser les mots les plus probables. Supérieure à 1, le modèle retiendra plus souvent des termes plus rares dans le corpus. Avec une température à 0, le modèle prendra systématiquement le terme le plus probable et apparaitra comme plus stable et plus prévisible.

La question déterminante est évidemment de savoir comment on détermine les paramètres d'un modèle. C'est le rôle de la phase d'entraînement qui va permettre d'en ajuster progressivement les valeurs.

> La question déterminante est évidemment de savoir comment on détermine les paramètres d'un modèle. C'est le rôle de la phase d'entraînement qui va permettre d'en ajuster progressivement les valeurs.

Pour la traduction, les fabricants de modèle de langage utilisent des corpus d'entraînement, composé d'un très large ensemble de paires de phrases déjà traduites et que le système doit reproduire. Après avoir masqué la fin de la traduction de référence, il est demandé au système de fournir le mot suivant. Si le résultat n'est pas correct, les paramètres du modèle sont légèrement ajustés par des calculs. Cette opération est répétée en confrontant à chaque fois les prédictions à la sollicitation, puis les paramètres du modèle sont modifiés afin de mieux aligner les prédictions avec la réalité du corpus. Les ajustements ont pour effet de rapprocher progressivement le vecteur du mot attendu du vecteur de la phrase source et de la traduction partielle.

On perçoit alors les raisons de l'immense capacité de calcul nécessaire pour l'apprentissage puisque, pour chaque mot de chacune des phrases du corpus d'entraînement, il faut multiplier des opérations de génération/ajustement.

L'évaluation automatique joue ici un rôle déterminant car elle permet de mesurer, au fur et à mesure de l'apprentissage, les progrès du système, jusqu'au moment où l'apprentissage plafonne. Il peut alors s'interrompre, le système étant suffisamment entraîné. Un principe majeur à observer est que les jeux de tests utilisés pour l'évaluation des progrès du modèle doivent être différents des données d'entraînement tout en étant suffisamment proches pour être pertinents.

L'entraînement doit maintenir un équilibre entre plusieurs risques parmi lesquels

l'instabilité et le surentraînement[83]. De nombreux facteurs peuvent conduire à ce que l'introduction de nouveaux exemples perturbe ce qui avait déjà été appris et rende ainsi impossible d'atteindre un état stable. À l'inverse, le surentraînement[84] aboutit à un modèle apprenant trop bien les données d'entraînement au point de mémoriser les exemples au lieu de généraliser à de nouveaux cas. Cela peut conduire à des réponses stéréotypées, peu créatives, ou à la reproduction mot à mot de passages vus pendant l'entraînement. Le surentraînement peut aussi poser des problèmes éthiques si le modèle mémorise des données sensibles ou privées.

Il convient de préciser que les explications précédentes concernent la réalisation de modèles capables de traduire d'une langue vers une autre à partir d'un corpus d'entraînement bilingue. Une fois un système français-anglais entraîné, on peut en inversant les corpus d'entrée et de sortie obtenir, pour le même coût, un système qui va de l'anglais au français.

4. La révolution neuronale : du bilinguisme au plurilinguisme

La traduction a connu une évolution majeure avec l'arrivée de modèles capables de traduire un grand nombre de langues dans toutes les combinaisons. Elle a été rendue possible par la segmentation des mots en unités sous-lexicales plus petites et indépendantes de la langue qu'on appelle « *tokens* » et la combinaison de données d'entraînement de plusieurs langues vers l'anglais et réciproquement, ce qui permet une traduction entre deux langues quelconques du corpus au moyen d'une double traduction passant par l'anglais utilisé comme langue pivot.

> « La traduction a connu une évolution majeure avec l'arrivée de modèles capables de traduire un grand nombre de langues dans toutes les combinaisons.

Dans un souci de simplicité, les grands modèles de langues ou les systèmes de traduction automatique ont été présentés jusqu'ici comme manipulant des mots. En réalité, comme on vient de le souligner, les modèles actuels manipulent des unités plus petites appelées unités sous-lexicales ou *tokens* en anglais. Par exemple, le mot « langue » se voit décomposé en deux unités : « lang » suivi de « ue ». Un premier avantage de ce procédé est la réduction de la complexité des outils, le nombre de *tokens* étant bien inférieur au nombre de mots. Ainsi tout se passe comme si on avait un grand vocabulaire englobant des unités plus petites, dont la qualité essentielle est d'être indépendant des langues.

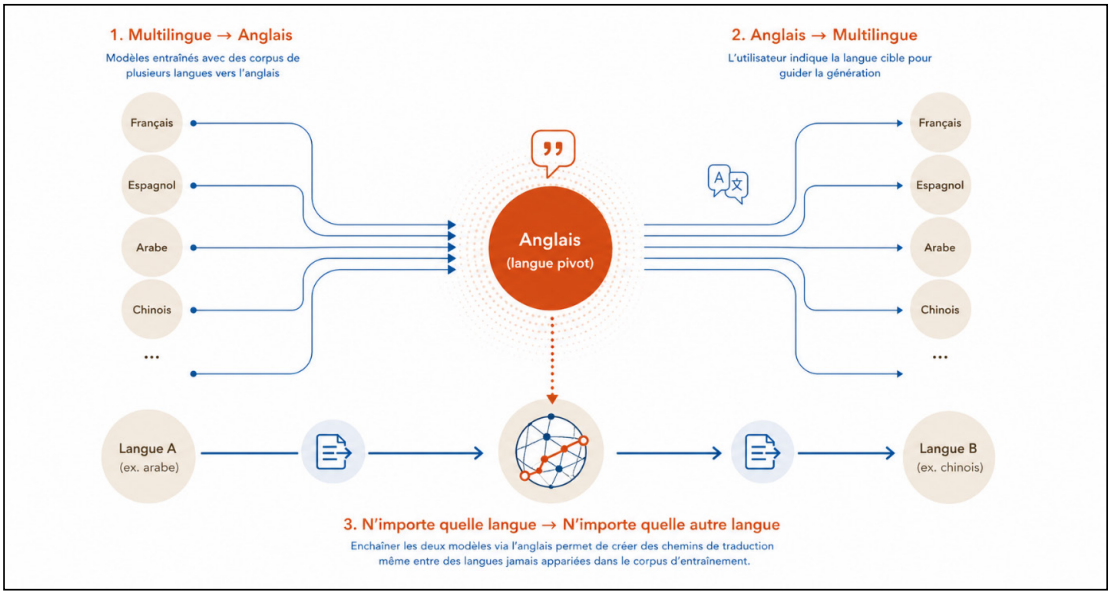
Le recours à une langue pivot

L'idée est donc de combiner les corpus de plusieurs langues vers l'anglais lors de la phase d'entraînement. Avec les mêmes méthodes que celles décrites précédemment, on obtient un modèle capable de produire des textes vers la langue pivot. Pour la traduction réciproque de la langue pivot vers une des autres langues, c'est à l'utilisateur de donner la consigne de la langue cible. Lors de l'entraînement, la langue cible est indiquée dans la consigne. À force de sanctions et de récompenses, les paramètres seront ajustés pour diriger les traductions selon cette consigne.

La capacité de traduire une langue quelconque vers l'anglais puis l'anglais vers une autre langue permet alors, en suivant ces deux étapes, de traduire n'importe quelle langue en n'importe quelle autre langue, y compris des langues n'ayant jamais été associées dans le corpus d'entraînement. Cette stratégie du pivot a donc représenté un grand pas vers la traduction multilingue. Elle a été utilisée historiquement dans des systèmes comme Google Translate.

Ce processus de double traduction n'est cependant pas sans incidence sur la qualité des traductions produites, du fait à la fois de l'accumulation voire l'amplification des erreurs produites à chaque étape, des pertes de nuance et de l'introduction de biais culturels ou conceptuels liés à la langue pivot elle-même. Comme a pu le constater, par exemple, Raoul Blin, les résultats se trouvent inférieurs à ceux obtenus avec des systèmes bilingues

83. Noëlie Debs, Sergio Peignier, Clément Douarre, *et al.*, « Apprendre l'apprentissage automatique : un retour d'expérience ». *JzeA*, n° 3, 2013, https ://doi.org/10.1051/j3ea/20222013.
84. *Overfitting*.



Le principe de la traduction automatique par l'anglais comme langue pivot
Source : auteur, généré par instructions à ChatGPT 5

spécifiquement entraînés pour un binôme de langues particulier[85].

La disparition de la langue pivot

À partir des années 2020, des acteurs comme Google, Facebook et d'autres ont fait le constat que tous les composants technologiques étaient réunis pour réaliser un système capable de traduire directement n'importe quelle langue en n'importe quelle autre langue. Il suffisait pour cela d'assembler des corpus parallèles mélangeant toutes les langues. Quand bien même ces corpus s'avéraient beaucoup plus volumineux, l'utilisation des mêmes techniques d'entraînement restait possible en ajoutant l'indication de la langue de génération.

Par cette approche, si le modèle possède suffisamment de paramètres, avec un processus d'apprentissage suffisamment long, le modèle obtenu peut recevoir et produire plusieurs dizaines voire centaines de langues. Le fait de disposer d'un modèle unique présente un triple avantage :

- proposer des traductions même pour des binômes jamais traités pendant l'apprentissage, mais avec une moindre qualité ;
- limiter le coût de développement d'un seul système ;
- simplifier le développement des systèmes que l'on peut faire évoluer aisément de manière très régulière.

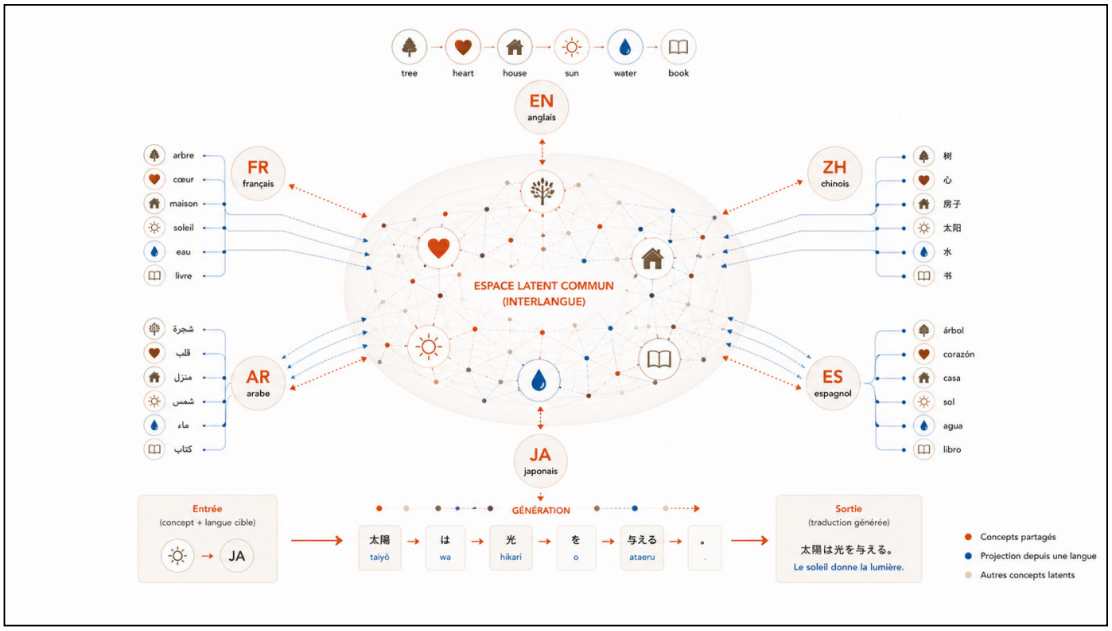
Ces méthodes offrent également un avantage majeur pour les opérateurs de ces systèmes : au lieu de maintenir des milliers de modèles traitant chacun un binôme de langues, il suffit d'en déployer un seul sur de nombreux serveurs à travers le réseau.

Cette simplicité apparente masque cependant une réalité concrète : pour la plupart des paires de langues, ces systèmes n'ont pas été soumis à des tests approfondis, voire n'ont pas été testés du tout.

L'émergence d'une interlangue

L'encodeur de ces systèmes, c'est-à-dire la partie qui prélève une phrase source et en calcule la représentation sous forme de nombres, ressemble à une forme d'interlangue numérique, dont la propriété est d'être spécifique et interne à chaque modèle. Pour un modèle donné, si on sélectionne deux phrases qui sont la traduction l'une de l'autre et qu'on les fait passer par l'encodeur, on obtient des représentations très proches, ce qui signifie que ces

85. Raoul Blin, « Traduire des corpus pour construire des modèles de traduction neuronaux : une solution pour toutes les langues peu dotées ? », *in Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheur, 2020s en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Volume 2 : Traitement Automatique des Langues Naturelles*, édité par Christophe Benzitoun, Chloé Braud, Laurine Huber *et al.* ATALA. https ://hal.science/hal-02784765.



Les modèles multilingues associent des concepts proches dans de multiples langues.
Source : auteur, généré par instructions à Chat GPT 5)

systèmes sont capables d'associer des phrases proches avec des représentations similaires.

Au-delà de son utilisation pour générer une traduction, cette propriété de l'encodeur est très utile pour développer des techniques de transfert interlinguistique[86], qui autorisent le traitement d'une phrase indépendamment de sa langue de production.

Ainsi, parmi les exemples d'application, on peut citer l'analyse de sentiment ou la détection de propos haineux. Quand bien même on ne disposerait pas de corpus d'entraînement dans la langue source mais uniquement en anglais, l'introduction du propos à contrôler dans un encodeur multilingue permet d'obtenir une représentation numérique proche de la traduction anglaise à laquelle il suffit ensuite d'appliquer le système de détection entraîné à partir de l'anglais.

> « Parmi les exemples d'application, on peut citer l'analyse de sentiment ou la détection de propos haineux.

5. La montée en puissance des capacités de traduction des LLM

Les grands modèles de langage[87] reposent sur une architecture différente, impliquant uniquement une phase de décodage à partir du *prompt* fourni par l'utilisateur et de leur capacité à suivre des instructions. En réponse à des *prompts* du type « Traduis le texte suivant en langue A » suivi du texte à traduire, ces modèles ont démontré de bonnes capacités de traduction.

Comme on le verra dans la suite des développements, les performances de traduction des LLM tendent aujourd'hui à égaler celles des modèles à encodeur/décodeur[88] et ils présentent d'autres avantages qui en font une des évolutions les plus prometteuses de la traduction automatique.

Leur meilleur atout est de ne pas nécessiter de corpus de traductions alignées toujours complexes et coûteux à constituer. Ils offrent

86. Traduction par l'auteur du terme *cross-lingual transfer*.

87. *Large Language Models* ou LLM.
88. Rajput, Ajeet Singh, Harsh Pratap Singh, Neha Raja Panwar, Neeraj Arya, Shuchi Mandhanya, et Khyati Bane. 2025. « Comparative study of large language models for machine translation ». In *Data science and applications*,. Springer Nature Singapore.

également à l'utilisateur la possibilité d'ajouter des instructions plus fines pour orienter la traduction, comme des indications de style ou de glossaires de traduction. Cela donne une très grande flexibilité dans la manipulation du contexte. Ils permettent ainsi d'envisager des traductions prenant en compte un contexte plus large que la phrase, qui reste l'entité de référence des architectures à encodeur/décodeur.

6. Quelques enseignements sur les LLM en général

Pour conclure, la nature, la qualité et le volume des données d'apprentissage sont des facteurs déterminants de la qualité du système une fois entraîné. Celle-ci dépend aussi de leur sélection, de leur préparation, de leur ordre de présentation, de la répétition de leur utilisation et du choix du mécanisme pour leur évaluation.

Plusieurs conséquences importantes en découlent pour les développeurs et les utilisateurs de systèmes d'IA.

- Le recours à des technologies d'apprentissage neuronal n'est pas en soi un gage de performance. L'étiquette « intelligence artificielle » n'est pas un label de qualité, mais une simple indication du type de technologies mobilisées. S'il fallait faire une analogie, la mention « moteur électrique » n'est pas le gage d'une voiture sûre, confortable et durable.
- Les performances d'un système d'IA ne sont pas forcément homogènes : des réponses de qualité lors de certaines interactions ne permettent pas de conclure que les performances seront aussi bonnes pour d'autres questionnements.
- Les systèmes d'IA sont des constructions et par la même la résultante de choix structurants opérés par leurs concepteurs en fonction de leurs objectifs. Ainsi, par construction, ils ne sont absolument pas exempts de biais ou d'inclinations, et, la compétence, le savoir-faire, la volonté de leurs créateurs et concepteurs ont nécessairement un impact.

II - Limites et perspectives de la traduction automatique

Du fait de l'ancienneté du recours à des techniques d'automatisation pour générer des traductions, la communauté des acteurs intéressés s'est confrontée de longue date aux aléas de ces technologies, qui ont rendu nécessaire de mettre en place des instruments et des cadres d'évaluation pour en mesurer les performances. Ces démarches présentent des aspects très méthodiques et s'adossent à une réflexion approfondie sur la mesure de la performance. À l'heure où les mondes du droit et de la justice s'ouvrent aux applications de l'intelligence artificielle, il y a là une source d'enseignements importante. L'évaluation ne sert pas ici une défiance vis-à-vis des technologies, mais constitue un moyen d'en appréhender les limites pour pouvoir améliorer leur efficacité et contenir en parallèle les risques inhérents.

> « Les systèmes d'IA ne sont absolument pas exempts de biais ou d'inclinations, et, la compétence, le savoir-faire, la volonté de leurs créateurs et concepteurs ont nécessairement un impact.

1. Mesurer la performance d'un système de traduction

L'évaluation des performances en traduction automatique repose historiquement sur des métriques automatiques dites « de surface », telles que BLEU, TER ou hLEPOR, qui comparent la sortie du système à une traduction de référence à partir de similarités lexicales ou syntaxiques. Ces approches, bien que robustes et peu coûteuses à déployer à grande échelle, présentent des limites structurelles : elles pénalisent fortement les paraphrases, les variations stylistiques légitimes et les divergences de longueur, sans pour autant refléter nécessairement une perte réelle de sens. Le rapport *The State of Machine Translation de 2024* précité souligne que ces métriques sont insuffisantes pour rendre compte de la qualité sémantique des traductions, en particulier dans les domaines spécialisés ou créatifs où les reformulations sont fréquentes[89].

89. *Ibid*em.

Pour dépasser ces limites, les évaluations contemporaines mobilisent de plus en plus des métriques sémantiques fondées sur des modèles neuronaux, telles que BERTScore ou COMET. Ces méthodes comparent les représentations vectorielles de la traduction automatique et de la référence humaine, et intègrent également parfois le texte source. COMET est présenté comme offrant l'un des plus hauts niveaux de corrélation avec le jugement humain lors des éditions du Workshop on Machine Translation (WMT). Toutefois, le rapport insiste sur un point méthodologique essentiel : un score COMET élevé ne suffit pas à caractériser finement les forces et faiblesses d'un système, notamment lorsque plusieurs modèles se situent dans un intervalle de confiance statistique proche (83 %).

La question de l'évaluation de la qualité des traductions réalisées par des traducteurs humains[90] a elle aussi une longue histoire, qui semble avoir convergé vers le cadre MQM[91]. Cette méthode repose sur l'identification, la catégorisation et la pondération des erreurs de traduction selon leur gravité (mineure, majeure, critique). Les erreurs sont classées en catégories telles que l'exactitude (omissions, ajouts, contresens), la fluidité, la terminologie, le style ou les conventions locales. Cette approche permet non seulement de produire un score agrégé, mais surtout d'analyser la nature des défauts de traduction, ce qui est décisif pour des usages professionnels impliquant des risques juridiques, médicaux ou financiers.

L'un des enseignements majeurs du champ de la traduction automatique est qu'il n'existe pas de méthode unique d'évaluation universellement valide. Les métriques automatiques sont indispensables pour comparer un grand nombre de systèmes sur des corpus étendus, tandis que les évaluations MQM – humaines ou assistées par LLM – sont nécessaires pour apprécier l'acceptabilité réelle des traductions dans un contexte donné. Le rapport montre également que les performances varient fortement selon les situations et que les évaluations ne sont pas toujours convergentes : une traduction jugée excellente selon une métrique globale peut présenter des erreurs critiques dans un contexte juridique ou réglementaire. L'évaluation des performances doit donc être conçue comme combinant plusieurs aspects, sélectionnés en fonction des exigences fonctionnelles et des risques associés à chaque usage de la traduction automatique.

« L'un des enseignements majeurs du champ de la traduction automatique est qu'il n'existe pas de méthode unique d'évaluation universellement valide.

2. Les performances actuelles des outils de traduction automatique

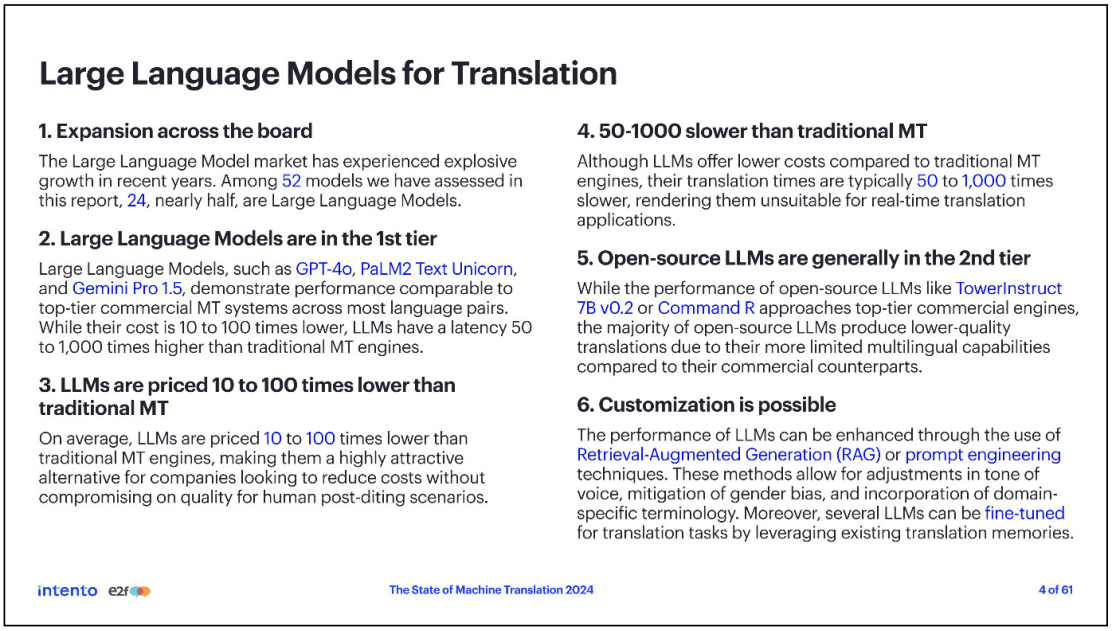
Un rapport de l'entreprise technologique américaine Intento sur l'évaluation des outils de traduction automatique de 2024[92] pointe les tendances suivantes.

L'un des constats majeurs du rapport est la **progression nette et généralisée de la qualité de la traduction automatique ces dernières années**, qui se traduit par une baisse relative des erreurs majeures et critiques et une proportion croissante des segments ne nécessitant aucune ou très peu de post-édition. Le rapport souligne que les performances atteintes rendent désormais utilisable de manière opérationnelle la traduction automatique dans un nombre croissant de contextes professionnels, sous réserve d'un cadrage approprié des usages.

La progression est particulièrement marquée pour les LLM des grands fournisseurs comme OpenAI, Google ou Anthropic, qui atteignent des niveaux de qualité proches de ceux des plateformes de traduction spécialisées.

L'analyse qualitative des erreurs montre que **les problèmes qui subsistent concernent principalement l'exactitude sémantique**. Les contresens, sous-traductions et sur-traductions représentent la grande majorité des erreurs classées comme majeures ou critiques, tandis que les erreurs de fluence (grammaire, ponctuation, orthographe) sont devenues relativement marginales. En d'autres termes, les systèmes produisent aujourd'hui des textes globalement

90. *Translation Quality Evaluation* ou TQE.
91. MQM (*Multidimensional Quality Metrics*).
92. Intento, « The State of Machine Translation 2024 ».



Large Language Models for Translation

1. Expansion across the board

The Large Language Model market has experienced explosive growth in recent years. Among 52 models we have assessed in this report, 24, nearly half, are Large Language Models.

2. Large Language Models are in the 1st tier

Large Language Models, such as GPT-4o, PaLM2 Text Unicorn, and Gemini Pro 1.5, demonstrate performance comparable to top-tier commercial MT systems across most language pairs. While their cost is 10 to 100 times lower, LLMs have a latency 50 to 1,000 times higher than traditional MT engines.

3. LLMs are priced 10 to 100 times lower than traditional MT

On average, LLMs are priced 10 to 100 times lower than traditional MT engines, making them a highly attractive alternative for companies looking to reduce costs without compromising on quality for human post-diting scenarios.

4. 50-1000 slower than traditional MT

Although LLMs offer lower costs compared to traditional MT engines, their translation times are typically 50 to 1,000 times slower, rendering them unsuitable for real-time translation applications.

5. Open-source LLMs are generally in the 2nd tier

While the performance of open-source LLMs like TowerInstruct 7B v0.2 or Command R approaches top-tier commercial engines, the majority of open-source LLMs produce lower-quality translations due to their more limited multilingual capabilities compared to their commercial counterparts.

6. Customization is possible

The performance of LLMs can be enhanced through the use of Retrieval-Augmented Generation (RAG) or prompt engineering techniques. These methods allow for adjustments in tone of voice, mitigation of gender bias, and incorporation of domain-specific terminology. Moreover, several LLMs can be fine-tuned for translation tasks by leveraging existing translation memories.

intento e2f The State of Machine Translation 2024 4 of 61

Principales conclusions du rapport « The State of Machine Translation 2024 » publié par la société Intento
(rapport disponible sur demande sur leur site : www.intento.io)

bien formés du point de vue linguistique, mais ils peuvent encore altérer le contenu informationnel ou pragmatique du message source. Ce constat fait ressortir l'importance d'une évaluation centrée sur le sens et non sur la seule qualité formelle du texte proposé en sortie.

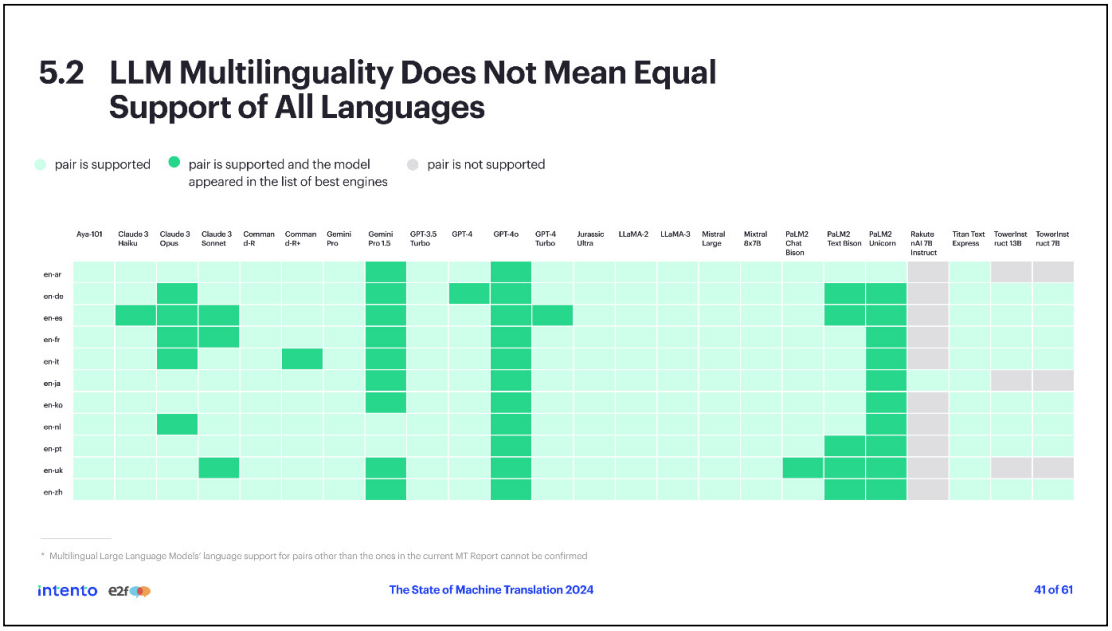
Les résultats confirment que la **qualité de la traduction automatique varie davantage selon le binôme de langues et la matière abordée que selon la technologie des modèles**. Les langues pour lesquelles les sources de données sont abondantes et les domaines généralistes ou techniques standardisés obtiennent globalement de meilleurs scores que d'autres qui concentrent une part plus importante d'erreurs significatives. **Le rapport met en évidence que les registres familiers (on parle ici de colloquialisme[93]) et certains usages spécialisés restent structurellement plus difficiles à traduire**, indépendamment des gains globaux observés.

Un résultat important concerne la distribution des erreurs par gravité : **si les erreurs critiques sont peu fréquentes désormais à l'échelle du corpus, leur impact potentiel demeure élevé**. L'application du cadre dit MQM pour *Multidimensional Quality Metrics*, (mesures de qualité multidimensionnelles) conçu pour mesurer, comparer et améliorer la qualité des traductions, montre qu'un nombre de contresens ou d'ajouts non justifiés même limité peut suffire à rendre une traduction inacceptable dans certains contextes d'usage. Cette asymétrie entre fréquence d'erreurs et niveau de gravité invite à dépasser les évaluations moyennes globales et à raisonner en termes de risques résiduels, en particulier pour les domaines sensibles (juridique, financier, santé).

> Un nombre de contresens ou d'ajouts non justifiés même limité peut suffire à rendre une traduction inacceptable dans certains contextes d'usage.

Enfin, le rapport met en évidence que **les meilleures performances globales sont atteintes lorsque les systèmes bénéficient de mécanismes d'adaptation** : glossaires,

93. Le colloquialisme désigne des expressions informelles utilisées à l'oral ou à l'écrit au quotidien. Ils peuvent inclure des contractions, des régionalismes, des tournures informelles et de l'argot.



Les performances de traduction des modèles multilingues varient très sensiblement selon les binômes de langues
Source : Intento, « The State of Machine Translation 2024 »

personnalisation terminologique ou ajustement à la matière. À l'inverse, les évaluations de modèles génériques montrent des plafonds de performance relativement stables pour certains types de contenus. Ce constat amène à considérer l'évaluation, non comme une mesure abstraite de la « qualité intrinsèque » d'un système, mais comme une appréciation située, dépendante de la proximité entre les données d'évaluation et les conditions réelles d'usage.

Même s'ils sont moins performants, les outils des générations précédentes présentent pour autant des avantages très significatifs au regard de leur consommation en puissance de calcul très inférieure, d'un ratio de 50 à 1 000 selon le rapport d'Intento[94]. Cette différence est lourde de conséquences, que ce soit en termes de vitesse de traduction bien supérieure, ce qui peut être décisif quand le volume de documents à traduire est important, ou de capacité à s'exécuter sur des infrastructures de calcul bien moins coûteuses et donc bien plus frugales en termes de consommation énergétique, ce qui permet d'envisager des traductions efficaces à partir d'un simple poste de travail.

« Même s'ils sont moins performants, les outils des générations précédentes présentent pour autant des avantages très significatifs au regard de leur consommation en puissance de calcul très inférieure.

3. Les obstacles à la performance de la traduction

La vérification des sorties constitue un enjeu central car, même si ces systèmes paraissent robustes et fiables à première vue, ils peuvent générer des erreurs significatives pour des raisons difficiles à anticiper. Ces fragilités sont bien sûr l'objet de l'attention et des efforts des travaux de recherche en cours.

Le premier ensemble de fragilités est dû au phénomène d'alignement lors de l'apprentissage et complexifie considérablement le contrôle. Le bon fonctionnement du système dépend de nombreux facteurs comme le contenu propre de chacun

94. Intento, *op. cit.*, p. 4.

des textes d'entraînement, la structure des grandes masses du corpus, la manière dont les exemples sont présentés et la configuration matérielle des machines. À raison de certains biais ou aléas dans les données, le système peut adopter des comportements qui ne sont pas ceux attendus.

Le deuxième type de fragilités résulte de l'incomplétude du contexte. Comme expliqué précédemment, le système détermine ses réponses à partir de la sollicitation qu'il reçoit, soit pour la plupart des systèmes la phrase en cours de traduction. Si le contexte n'est pas assez riche ou s'il présente des ambiguïtés, le système pourra ne pas fournir une réponse satisfaisante.

Cette incomplétude du contexte peut se manifester par exemple dans les situations dans lesquelles la langue cible opère des distinctions différentes de celles de la langue source. Ainsi, en hongrois, les pronoms n'ont pas de genre. Aussi, afin de déterminer le genre du pronom pour des traductions en français ou en anglais, il faut trouver cette information dans le contexte et utiliser les statistiques du corpus. Or on s'aperçoit que la phrase en hongrois « il/elle prépare un gâteau » sera traduite avec le pronom féminin « she » en anglais et « il » en français. Un traducteur humain pourrait lui se référer à des éléments du texte en amont ou en aval pour faire le bon choix.

Un troisième ensemble de limites tient au caractère probabiliste de ces systèmes, puisque tout ce qui échappe à la logique des grands nombres, en particulier les événements très rares, n'est pas appris, sauf s'agissant d'occurrences utilisées assez fréquemment pour pouvoir être mémorisées. Tout ce qui relève de l'idiosyncrasie, des expressions idiomatiques, des métaphores très rares risque ainsi d'être mal traduit.

Une quatrième catégorie de limites tient au caractère fragmenté des langues : elles sont composées en effet de nombreux sous-langages, mobilisés notamment dans les domaines spécialisés ou professionnels, mais aussi dans certaines communautés. Dans un rapport commun de 2022[95], les agences EU-Lisa[96] et Eurojust[97] soulignent les difficultés de traitement de la terminologie : « Ces systèmes ne distinguent pas les termes techniques des autres expressions, si bien que des traductions issues d'autres domaines – plus fréquentes dans les corpus parallèles – ont tendance à être privilégiées. Cela peut entraîner plusieurs problèmes : l'emploi d'équivalents erronés issus d'autres domaines, l'utilisation de variantes obsolètes ou encore l'incohérence dans le choix des synonymes terminologiques ».

> Une quatrième catégorie de limites tient au caractère fragmenté des langues : elles sont composées en effet de nombreux sous-langages, mobilisés notamment dans les domaines spécialisés ou professionnels, mais aussi dans certaines communautés.

Enfin, une dernière source d'erreurs est propre à l'algorithme de génération, qui construit les phrases mot à mot en complétant au fur et à mesure le contexte de génération. Une fois qu'une erreur a été introduite, celle-ci peut facilement se propager et plonger le système dans des boucles d'erreurs incontrôlées.

4. Les langues à faibles ressources

Comme il ressort du rapport précité de la société Intento, la qualité de traduction reste très inégale selon les binômes de langues : pour les langues d'utilisation très courante, la traduction est souvent satisfaisante, mais pour les langues plus rarement mobilisées, la génération reste difficile. Par ailleurs, certaines langues s'avèrent plus difficiles à traduire que d'autres, comme les langues germaniques. Ce constat a son importance pour les usages judiciaires, car l'un des apports potentiels de ces technologies est notamment de pallier la difficulté à trouver des traducteurs vers et depuis le français pour certaines langues.

95. EU-LISA, Eurojust, « Artificial Intelligence supporting cross-border cooperation in criminal cases », 2022, p. 17, https ://www.eurojust.europa.eu/publication/artificial-intelligence-supporting-cross-border-cooperation-criminal-justice.
96. EU-Lisa : Agence européenne pour la gestion opérationnelle des systèmes d'information à grande échelle au sein de l'espace de liberté, de sécurité et de justice.
97. Eurojust : Agence de l'Union européenne pour la coopération judiciaire en matière pénale.

Cela tient pour beaucoup à la disponibilité très variable de corpus d'entraînement, qui demeure très dépendante encore de l'histoire et des rapports de domination entre les différentes langues. La faible disponibilité de données alignées pour de très nombreuses langues introduit une asymétrie dans la mise à disposition d'outils de traduction automatique. Des efforts importants sont réalisés actuellement pour constituer des corpus d'entraînement couvrant des centaines de langues, comme Language in the Wild[98] ou Madlad[99]. L'initiative No Language Left Behind (NLLB)[100] vise à développer des corpus d'entraînement pour 150 langues à faibles ressources s'intégrant dans le modèle *open source* éponyme NLLB-200, qui supporte 200 langues. Parallèlement, Google annonçait en 2022 son objectif de supporter 1 000 langues dans Google Translate[101].

Chaque langue étant porteuse de cultures, de conceptions et de prismes différents, la prédominance de l'anglais pour l'entraînement des outils amène à s'interroger sur les effets qu'elle pourrait produire sur les traductions. L'enjeu de la constitution et du maintien de corpus d'entraînement dans une langue apparaît donc possiblement comme un élément essentiel de son existence numérique future et de défense des cultures dont elle est porteuse.

Chaque langue étant porteuse de cultures, de conceptions et de prismes différents, la prédominance de l'anglais pour l'entraînement des outils amène à s'interroger sur les effets qu'elle pourrait produire sur les traductions.

98. Isaac Caswell, Theresa Breiner, Daan van Esch, et Ankur Bapna, « Language ID in the Wild : Unexpected Challenges on the Path to a Thousand-Language Web Text Corpus », 2020. *arXiv* :2010.14571. Prépublication, https ://doi.org/10.48550/arXiv.2010.14571.
99. Sneha Kudugunta, Isaac Caswell, Biao Zhang *et al.* 2023. « MADLAD-400 : A Multilingual And Document-Level Large Audited Dataset ». Version 1. Prépublication, arXiv. https ://doi.org/10.48550/ARXIV.2309.04662.
100. NLLB Team, Marta R. Costa-Jussà, James Cross *et al.* 2022. « No Language Left Behind : Scaling Human-Centered Machine Translation ». Version 3. Prépublication, arXiv. https ://doi.org/10.48550/ARXIV.2207.04672.
101. Google Inc, « 3 ways AI is scaling helpful technologies worldwide", *Blog.google*, novembre 2022, https ://blog.google/technology/ai/ways-ai-is-scaling-helpful/ (consulté le 18 juillet 2024).

5. L'émergence de petits modèles de langage spécialisés pour la traduction

Les systèmes de traduction bénéficient des progrès de la recherche générale sur l'IA. Ainsi à partir du grand modèle NLLB-200 précité, par l'application de techniques dites de « distillation »[102], il a été possible de créer une version moins performante (NLLB-1,3B), mais 40 fois moins énergivore et capable de fonctionner sur un ordinateur personnel puissant.

Une part très importante des efforts de la recherche sur l'IA porte sur la réalisation de modèles de langages plus petits, les *Small Language Models* (SLM), qui sont plus efficaces et moins gourmands en puissance de calcul, dont les résultats très concrets ouvrent des perspectives prometteuses. Les bénéfices écologiques et économiques évidents pourraient cependant être largement absorbés par une explosion des usages de la traduction automatique que, par un effet rebond, ces progrès eux-mêmes pourraient alimenter.

Cette dynamique de réduction des besoins en puissance de calcul converge avec la tendance historique des progrès de l'électronique dans le sens d'une démultiplication de la puissance des ordinateurs personnels, qui se poursuivra dans les années à venir, certains acteurs prédisant une multiplication de la capacité de calcul de l'IA utilisée en local entre 5 et 10 d'ici à 2030. Selon une étude de marché de l'Institut Gartner, dès 2026, une large part des ordinateurs personnels vendus disposeront de telles capacités de calcul IA en local. L'étude précise que « les SLM rendent possible l'exécution de fonctions IA sophistiquées avec une consommation énergétique maîtrisée et une protection accrue des données utilisateur » et que le « passage vers l'IA embarquée traduit la volonté du marché de s'affranchir des dépendances au cloud et d'amener l'IA au plus près de l'utilisateur »[103].

Google propose déjà des modèles dont les poids sont accessibles sous la dénomination Gemma (collection de modèles d'IA ouverts

102. La distillation d'un modèle de langage consiste à entraîner un modèle plus petit (« élève ») à reproduire les comportements d'un modèle plus grand et plus performant (« professeur »), non seulement en imitant ses réponses finales, mais surtout en apprenant à partir des probabilités qu'il attribue aux différentes solutions possibles ; ce transfert permet d'obtenir un système beaucoup plus léger et rapide, tout en conservant l'essentiel des compétences du modèle initial, au prix d'une légère perte de précision.
103. Serge Lebial, « Vers une prépondérance des PC IA en 2026 selon Gartner », *Le Monde Informatique*, 2026 https ://www.lemondeinformatique.fr/actualites/lire-vers-une-preponderance-des-pc-ia-en-2026%C2%A0selon-gartner-97706.html (consulté le 3 janvier 2026).

et compacts), dont les versions les plus petites peuvent fonctionner à des niveaux de performance satisfaisants sur des ordinateurs personnels puissants. TranslateGemma[104], propose la traduction croisée entre 55 langues. Le niveau de qualité de la langue générée est présenté comme très supérieur à ce qu'il était possible d'obtenir jusqu'alors pour des traductions en exécution locale. Ces modèles représentent une évolution majeure vers la possibilité de produire des traductions de qualité sans transférer des documents sensibles et avec des niveaux de puissance de calcul raisonnables.

6. Vers l'entraînement de modèles affinés pour la traduction judiciaire en français

Comme l'a fait ressortir une analyse du déploiement massif d'outils d'intelligence artificielle dans l'institution judiciaire au Brésil, celui-ci repose avant tout sur la constitution d'un très vaste corpus de données disponibles pour l'entraînement. À partir de celles-ci, il est possible d'entraîner et d'affiner des modèles spécifiques à partir de données anonymisées. Ces modèles ont alors pu être partagés et utilisés par les juridictions via une plateforme dédiée appelée SINAPSES[105]. Si le Brésil a commencé à y déployer des outils d'intelligence artificielle générative, leur utilisation était encore limitée en 2024[106]. Dans ce pays, le large déploiement de l'IA s'appuie sur des modèles métiers parfaitement adaptés aux besoins de l'institution judiciaire[107], en matière de traduction comme dans d'autres domaines.

Comme nous l'avons déjà abordé, les performances de traduction peuvent être sensiblement améliorées par un entraînement complémentaire de petits modèles de traduction à l'aide de jeux de données spécifiques à une sous-langue particulière. C'est le cas de la langue juridique, et plus particulièrement encore de la langue judiciaire, qui possède un vocabulaire et des formulations qui lui sont propres.

> **Les performances de traduction peuvent être sensiblement améliorées par un entraînement complémentaire de petits modèles de traduction à l'aide de jeux de données spécifiques à la langue juridique, et plus particulièrement encore la langue judiciaire.**

Or, si les petits modèles de langage présentent l'avantage de la frugalité pendant les phases d'inférence lorsque le modèle en quelque sorte met en pratique ce qu'il a appris, c'est également le cas au préalable lors de l'entraînement. Avec l'abaissement régulier du coût de la puissance de calcul, la réalisation de modèles de traduction aux performances renforcées pour la traduction judiciaire paraît aujourd'hui tout à fait accessible, tant pour ce qui est des petits modèles de langage que des modèles encodeur/décodeurs bilingues du type de ceux mobilisés dans le projet Bergamot[108].

La plus grande difficulté pour les concepteurs de tels modèles affinés est de disposer de corpus d'entraînement spécifiques, qu'il s'agisse de lexiques bilingues ou d'ensembles de segments de textes traduits et alignés. Deux pistes se dessinent : le recours à des outils d'IA pour faciliter la production de corpus d'entraînement et la valorisation des corpus traduits existants au sein de l'institution judiciaire.

La constitution de corpus d'entraînement pouvant être longue et complexe, la recherche en informatique explore actuellement les possibilités de recourir à l'assistance de grands modèles de langage pour la rendre plus productive. Il peut s'agir d'aide à l'extraction d'une terminologie dans des domaines spécifiques et cela peut aller jusqu'à la génération d'un corpus synthétique d'entraînement. L'ensemble de ces orientations sont explorées notamment par Yasmin Moslem[109].

104. Mara Finkelstein, Isaac Caswell, Tobias Domhan *et al.*, « TranslateGemma Technical Report ». 2026. arXiv :2601.09012. Prépublication, *arXiv*, janvier 19. https ://doi.org/10.48550/arXiv.2601.09012.
105. Conselho Nacional de Justiça, et Programa das Nações Unidas para o Desenvolvimento.. *Pesquisa inteligência artificial no Judiciário 2024 : resumo executivo*. 2025
106. Conselho Nacional de Justiça,. *O uso da inteligência artificial generativa no Poder Judiciário brasileiro : relatório de pesquisa*. 2024 https ://www.cnj.jus.br/wp-content/uploads/2024/09/cnj-relatorio-de-pesquisa-iag-pj.pdf .
107. Résolution du groupe de travail sur l'intelligence artificielle du pouvoir judiciaire..
108. Le projet Bergamot est une initiative européenne menée avec Mozilla et plusieurs universités pour développer une traduction automatique neuronale qui fonctionne entièrement en local, directement dans le navigateur de l'utilisateur, sans envoyer de données vers le cloud.
109. Yasmin Moslem, « Language Modelling Approaches to Adaptive Machine Translation ». Version 1. 2024. Prépublication, *arXiv*. https ://doi.org/10.48550/ARXIV.2401.14559.

Une autre possibilité de développement repose sur le constat suivant : compte tenu du volume de traductions réalisées chaque année et depuis plusieurs décennies par l'institution judiciaire, celle-ci dispose d'un fonds très large de traductions de très bonne qualité réalisées par des traducteurs professionnels. L'entraînement devant impérativement être effectué à partir de données anonymisées, le fait que l'ensemble de ces documents contiennent des données personnelles pourrait toutefois apparaître comme une limite insurmontable. Mais, dans le cas particulier de la traduction, les corpus d'entraînement sont composés de paires de phrases dans plusieurs langues. Cette structure très précise pourrait sans doute permettre d'isoler dans les corpus existants des segments sans données personnelles, ou, des phrases qui en contiendraient tout en étant susceptibles d'être facilement anonymisées à raison de leur brièveté.

« L'institution judiciaire dispose d'un fonds très large de traductions de très bonne qualité réalisées par des traducteurs professionnels. Le fait que l'ensemble de ces documents contiennent des données personnelles pourrait toutefois apparaître comme une limite insurmontable.

L'une des évolutions récentes des LLM est l'élargissement considérable du spectre du contexte, passé de quelques milliers à des centaines de milliers de mots pour certains très grands modèles de langage de dernière génération.

Pour la traduction, si le fait de pouvoir générer des textes plus longs apparaît porteur d'un bénéfice immédiat pour l'utilisateur, la recherche de nouvelles possibilités s'appuie surtout sur les capacités d'apprentissage en contexte (*in-context learning*[110]). Cette faculté permet d'envisager d'utiliser une partie du contexte pour fournir des éléments spécifiques et diriger la traduction, tels que des lexiques spécialisés, des exemples de formulation ou de mots à privilégier dans la langue cible, voire des exemples de phrases déjà traduites et validées du document en cours ou d'autres documents similaires. Cette perspective est également explorée par Yasmin Moslem dans la publication précitée. Ses résultats montrent qu'un LLM utilisé pour la traduction avec apprentissage en contexte permet d'atteindre voire de dépasser les performances des outils dédiés à la traduction.

De telles perspectives pourraient permettre de faire l'économie d'une phase d'affinage des modèles, mais pas de se dispenser de celles de création préalable d'un corpus spécialisé pour nourrir le contexte et d'identification de la terminologie et des exemples de traductions dont la qualité déterminera l'ampleur des bénéfices escomptés. Cependant, le volume de calculs à opérer par un LLM évoluant de manière exponentielle avec la taille du contexte, cette approche porte le risque d'une consommation encore accrue de calcul et d'énergie, alors même que, nous l'avons vu, ces modèles sont déjà 10 à 100 fois plus gourmands que les outils traditionnels de traduction.

* * *

À l'issue de ce tour d'horizon, il apparaît que la question de la traduction judiciaire est loin d'être une question réglée. De nombreux travaux sont en cours ou restent à mener, avec des arbitrages importants à faire entre complexité d'entraînement, vitesse de traduction, qualité des textes produits, coûts d'exploitation et consommation énergétique. Suivant les solutions, les écarts peuvent en effet être extrêmement significatifs.

110. L'*in-context learning* (apprentissage par le contexte) désigne la capacité d'un modèle de langage de grande taille à adapter son comportement à partir des seuls exemples fournis dans la requête, sans modification de ses paramètres internes. Concrètement, on insère dans la séquence de texte soumise au modèle quelques exemples et le modèle infère, par analogie statistique, la règle implicite à appliquer aux nouveaux cas. Il ne s'agit donc pas d'un apprentissage au sens classique (aucune phase d'entraînement supplémentaire), mais d'une généralisation conditionnée par le contexte textuel immédiat, rendue possible par l'ampleur des données et des corrélations internalisées lors de l'entraînement initial.

III - Enjeux juridiques du déploiement de solutions de traduction automatique dans la sphère judiciaire

Pour envisager les perspectives de mise en œuvre de solutions de traduction automatique par les acteurs judiciaires, les problématiques juridiques liées aux régimes de protection applicables aux données de procédure doivent être prises en compte. Un très rapide tour d'horizon des impératifs que pose la protection des secrets dont sont porteuses les pièces de procédure est donc proposé[111].

1. Les impératifs des régimes de protection de données issues de pièces judiciaires

La traduction automatique est un traitement automatique de données et ces données peuvent impliquer le respect de règles de protection afférentes. Pour les usages judiciaires, le traitement informatique de données de procédure pose des problèmes épineux du fait de la superposition de multiples régimes de protection dont nous ne pourrons donner ici qu'un aperçu extrêmement superficiel. Dans tous les cas, parce que les données de procédure sont considérées comme particulièrement sensibles[112], elles sont soumises aux règles de protection des données sensibles et souveraines et le recours à des sous-traitants par les acteurs publics est soumis à des exigences particulières de certification. S'agissant des données en matière civile et administrative, dès lors qu'elles contiennent de manière presque systématique des données personnelles, elles sont soumises aux dispositions du règlement général sur la protection des données (RGPD). Ce règlement impose des précautions particulières en cas de recours, directement ou indirectement, à un prestataire situé dans une zone ne faisant pas l'objet d'une décision d'adéquation[113], comme c'est le cas de la Chine à l'heure actuelle.

> Parce que les données de procédure sont considérées comme particulièrement sensibles, elles sont soumises aux règles de protection des données sensibles et souveraines et le recours à des sous-traitants par les acteurs publics est soumis à des exigences particulières de certification.

En matière pénale, les règles de la directive dite « police justice »[114] transposées aux articles 87 à 114 de la loi Informatique et Libertés[115] s'appliquent aux autorités publiques et imposent un régime spécifique et des précautions accrues par rapport au RGPD. Concernant les avocats et les experts qui se voient remettre des informations relevant de cette directive, le régime applicable prévu est celui du « destinataire »[116] défini à l'article 3 §10 et au point (36) : ces professionnels sont soumis au RGPD, mais également aux restrictions d'usages que peut leur imposer le gestionnaire du traitement à l'origine de la transmission des données personnelles.

Les données pénales échangées sont également soumises aux règles de protection des secrets professionnels et parfois du secret des affaires. Pour les avocats par exemple, le Conseil national des barreaux (CNB) rappelle que « le secret professionnel s'applique […] quels qu'en soient les supports, matériels ou immatériels, notamment électronique » et qu'il est « d'ordre public, général, absolu et illimité dans le temps »[117].

Il convient de souligner une particularité du RGPD et de la directive « police justice ». Le

111. Les développements de cette partie s'appuient en grande partie sur l'intervention de Marie Jonca lors de l'atelier du 19 juin 2025.
112. Article 31 II de la loi n° 2024-449 du 21 mai 2024 visant à sécuriser et à réguler l'espace numérique.
113. Acte juridique adopté par la Commission européenne qui reconnaît qu'un pays tiers (hors UE) ou une organisation internationale offre un niveau de protection des données équivalent à celui garanti par le RGPD.
114. Directive (UE) 2016/680 du 27 avril 2016 relative à la protection des personnes physiques à l'égard du traitement des données à caractère personnel par les autorités compétentes à des fins de prévention et de détection des infractions pénales, d'enquêtes et de poursuites en la matière ou d'exécution de sanctions pénales, et à la libre circulation de ces données, et abrogeant la décision-cadre 2008/977/JAI du Conseil (2016).
115. Loi n° 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés.
116. « 10. la personne physique ou morale, l'autorité publique, le service ou tout autre organisme qui reçoit communication des données à caractère personnel, qu'il s'agisse ou non d'un tiers. »
117. §19. « RGPD : une foire aux questions pour aller plus loin », https ://cnb.avocat.fr/actualite/rgpd-une-foire-aux-questions-pour-aller-plus-loin consulté le 2 mars 2026

RGPD prévoit en son article 55 que « les autorités de contrôle ne sont pas compétentes pour contrôler les opérations de traitement effectuées par les juridictions dans l'exercice de leur fonction juridictionnelle ». Parallèlement, la directive « police justice » exclut de même cette possibilité en son article 45. Cette disposition est reprise dans la loi Informatique et Libertés qui prévoit que, pour le contrôle du RGPD, la CNIL « n'est pas compétente pour contrôler les opérations de traitement effectuées, dans l'exercice de leur fonction juridictionnelle, par les juridictions et leur ministère public »[118]. Pour l'ensemble de ces textes, l'interprétation de la notion de fonction juridictionnelle apparaît très restrictive : elle ne concerne pas l'outil en tant que tel mais spécifiquement les « opérations de traitement ». Elle semble donc plutôt une limitation de l'exercice des pouvoirs de contrôle destinée à protéger l'indépendance juridictionnelle et concerne les usages spécifiques des acteurs juridictionnels dans le cadre strict de leurs fonctions juridictionnelles.

S'il n'existe pas à notre connaissance de dispositions spécifiques au recours à la traduction automatique, les avis rendus par la CNIL et différents ordres professionnels, tant sur la mise en œuvre d'outils numériques en conformité avec le RGPD que sur la mobilisation spécifique de LLM, fixent un cadre tout à fait approprié à la traduction automatique.

Dès lors, la question du recours à des outils de traduction automatique pour des pièces de procédure est particulièrement sensible pour les acteurs judiciaires comme pour l'ensemble des professions du droit, y compris les experts traducteurs appelés à traduire de telles pièces dans la mesure où leurs outils d'aide à la traduction mobilisent cette technologie.

> La question du recours à des outils de traduction automatique pour des pièces de procédure est particulièrement sensible pour les acteurs judiciaires comme pour l'ensemble des professions du droit, y compris les experts traducteurs.

Aborder la question de la protection des secrets dont sont dépositaires les intervenants à la procédure judiciaire revient à s'interroger sur les possibilités que la chaîne de traitement concrètement mise en place permette l'accès de tiers non autorisés aux données protégées. Par exemple, une affaire récente a révélé à cet égard que les prestataires de cloud opéraient des analyses des contenus stockés, en particulier pour détecter des contenus dont la détention est illégale.

2. Première perspective – le recours à des services hébergés

Aujourd'hui et de manière assez durable sans doute, les meilleurs résultats de traduction automatique sont obtenus avec les grands modèles de langage proposés par les principaux opérateurs sous la forme de services exécutés exclusivement sur leurs propres infrastructures.

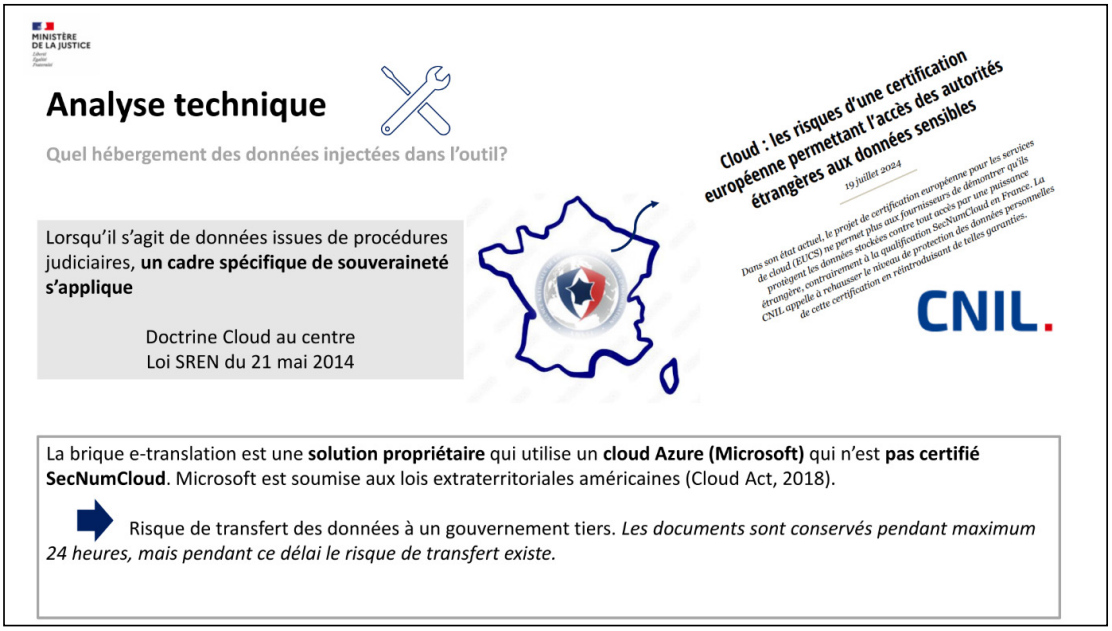
Les services de traduction hébergés accessibles actuellement sont de deux sortes. Il s'agit, d'une part, des services fournis par les grands modèles de langages (OpenAI, Claude, Mistral, DeepSeek…), qui prennent la forme d'agents conversationnels accessibles depuis un navigateur via des applications installées sur un ordinateur ou un smartphone, ou des agents mobilisés par des outils d'orchestration (n8n par exemple)[119], ou plus directement par des interfaces de programmation applicative (API)[120]. Il s'agit, d'autre part, des services dédiés de traduction en ligne, comme DeepL, Lara, eTranslation ou Google Cloud. Quel que soit le mode d'accès, la remise d'un texte pour traduction sur ces plateformes comporte la transmission du contenu vers les serveurs des prestataires et le stockage temporaire des données sur les plateformes à partir desquelles sera effectuée l'opération de traduction.

Les règles relatives à la protection des données souveraines s'imposent aux acteurs publics. Elles obligent à prévenir tout accès extraterritorial. C'est le cas des sociétés

118. Art 19 V de la Loi n° 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés

119. Les outils d'orchestration de type n8n sont des plateformes d'automatisation de flux de traitement (souvent sans programmation ou avec peu de programmation) permettant de connecter des applications, services, API et bases de données afin d'automatiser des tâches sans développement lourd.

120. Une API (*Application Programming Interface* ou interface de programmation pour applications) désigne un ensemble de règles et de points d'accès qui permet à deux programmes informatiques de communiquer entre eux, de s'échanger des données ou d'exécuter des actions automatiquement.



Analyse du risque de recours au service eTranslation de l'UE pour le traitement de pièces de procédure pénale
Source : secrétariat général, ministère de la Justice

américaines ou chinoises que leurs services soient utilisés de manière directe ou indirecte. Il n'est donc pas possible d'avoir recours aux services d'OpenAI, Anthropic ou Microsoft pour traduire des pièces de procédure. Les pratiques de sous-traitance étant très fréquentes dans l'offre de services IA, un examen attentif des conditions générales des prestataires est impératif. À titre d'illustration, le service italien Lara Translate utilise les services d'Amazon Web Services et le service de traduction eTranslation mis en place par la Commission européenne repose sur les services de Microsoft Azure.

Le recours à des services européens de traduction ne faisant pas appel à des prestataires américains et dont les centres de calcul sont localisés en Europe assure en principe une protection contre ces législations extraterritoriales. Pour autant, leur conformité aux impératifs du cloud souverain doivent être vérifiés. L'ANSSI publie un catalogue des applications certifiées, en particulier le standard SecNumCloud qui nous intéresse ici[121].

La problématique se complexifie encore avec la prise en compte des règles de protection du secret professionnel, qu'il s'agisse des rapports entre avocats, du secret de l'enquête et de l'instruction ou du secret des affaires. En effet, le débat semble ouvert sur le fait de savoir si transmettre à un prestataire souverain pour traduction automatique des documents de procédure pourrait s'analyser comme la remise à un tiers à la procédure, et donc comme une violation du secret.

> « Le recours à des services européens de traduction ne faisant pas appel à des prestataires américains et dont les centres de calcul sont localisés en Europe assure en principe une protection contre ces législations extraterritoriales.

121. p. 67. ANSSI, *Catalogue produits · services · profils de protection · sites, certifiés · qualifiés · agréés*, février 2026, https ://cyber.gouv.fr/offre-de-service/solutions-certifiees-et-qualifiees/services-de-securite-evalue/decouvrir-les-solutions-certifiees-qualifiees/.

3. Deuxième perspective – une infrastructure internalisée dédiée de traduction

Plusieurs grands modèles de langage sont disponibles avec des licences en source libre (ou plus exactement avec publication de leurs poids), ce qui permet, comme le prévoient leurs conditions d'utilisation, de les faire fonctionner sur une infrastructure de calcul spécifique. Celle-ci peut être soit internalisée soit externalisée totalement, mais elle s'intègre en tout état de cause à l'infrastructure numérique de l'administration ou de l'organisation et elle est entièrement administrée par ses services. Cette architecture permet en particulier de renforcer les garanties des politiques de sécurité spécifiques, notamment en matière de droits d'accès et de contrôle des personnes susceptibles d'accéder aux données transitant sur ces plateformes.

> Cette architecture permet en particulier de renforcer les garanties des politiques de sécurité spécifiques, notamment en matière de droits d'accès et de contrôle des personnes susceptibles d'accéder aux données transitant sur ces plateformes.

C'est l'approche retenue par le ministère de la Justice français[122], à savoir un assistant reposant sur un grand modèle de langage (LLM) en source libre et hébergé par une infrastructure cloud certifiée SecNumCloud. L'immense majorité des grands modèles de langage possédant aujourd'hui des facultés de traduction, cela devrait permettre d'offrir des capacités de traduction automatique sécurisées aux agents du ministère autorisant la soumission de pièces de procédures civiles ou pénales.

De manière complémentaire et dans un souci de frugalité et de minimisation de la consommation énergétique, une telle architecture devrait permettre également de déployer des grands ou des petits modèles de langage spécialisés pour la traduction, comme TranslateGemma de Google[123] ou les modèles spécialisés de la famille NLLB conçus par Meta[124] déjà mentionnés.

Cependant, sur le plan de la protection des secrets, la question de la sécurisation des accès à la plateforme dédiée elle-même pourrait se poser. Par exemple, la Plateforme nationale des interceptions judiciaires (PNIJ), qui a pour vocation de centraliser des données judiciaires sur des enquêtes et instructions en cours, mais également des données du renseignement, possède un régime de protection particulièrement fort, qui prévoit le recours systématique à des techniques d'encryptions pour le transfert et le stockage[125]. Ces mécanismes permettent de donner le contrôle à l'autorité judiciaire titulaire d'une affaire sur l'octroi de droits d'accès temporaires aux prestataires, soit pour des raisons techniques, soit pour des raisons d'interprétariat ou de traduction.

4. Troisième perspective – des capacités de traduction sur le poste de travail

Pour la traduction automatique, la possibilité de disposer d'outils d'un niveau de performance de l'ordre du niveau actuel fonctionnant localement, donc sans surcoût économique, avec une consommation énergétique modeste et sans risque supplémentaire lors de la transmission de pièces de procédure à un prestataire externe, apparaît comme une perspective réaliste pour les prochaines années.

D'une part, les technologies de traduction à base d'encodeur/décodeur, bien que produisant un texte de moindre qualité, présentent plusieurs avantages qui, dans certains contextes, peuvent être intéressants. La compacité des modèles a pour corollaire leur grande efficacité en termes de puissance de calcul, et donc de consommation énergétique. C'est cette compacité qui permet une utilisation en local, sans aucun transfert de données vers une machine tierce.

Le projet européen Bergamot, intégré désormais au navigateur Firefox, en est

122. Haffide Boulakras, « Rapport sur l'IA au service de la justice : stratégie et solutions opérationnelles », 2025, p. 36, consulté le 16 janvier 2025, https ://www.vie-publique.fr/rapport/299818-lia-au-service-de-la-justice-strategie-et-solutions-operationnelles.

123. Voir *supra* L'émergence des petits modèles de langage spécialisés pour la traduction.

124. Voir *supra* Les langues à faibles ressources.

125. Pour un aperçu de l'infrastructure de sécurité mis en place, voir par exemple la délibération n° 2020-103 du 15 octobre 2020 portant avis sur un projet de décret modifiant le décret n° 2014-1162 du 9 octobre 2014 relatif à la création de la « Plate-forme nationale des interceptions judiciaires » (PNIJ) de la CNIL, https ://www.legifrance.gouv.fr/jorf/id/JORFTEXT000044315821.

l'illustration : il permet une traduction d'une qualité satisfaisante entièrement en local sur un ordinateur standard. La particularité de cette approche est la modularité. Ce modèle repose sur une série de modèles bilingues (actuellement pour 45 dont 13 langues européennes) vers l'anglais utilisé comme langue pivot, à télécharger en fonction des besoins. S'agissant d'un logiciel libre, il peut être inclus à des applicatifs métier. En s'intégrant au navigateur Firefox, il offre une solution pérenne et crédible et une bonne dynamique de développement par l'intégration de nouvelles langues.

De manière très concrète, il est donc possible à l'heure actuelle de traduire en local des documents HTML simplement en installant Firefox. Bergamot est également utilisable pour des traductions depuis et vers l'anglais avec l'application téléchargeable *open source* Translate Locally[126]. Elle permet des traductions sur un simple PC à des vitesses de l'ordre de 2 pages par seconde, bien supérieures à celle des solutions reposant sur des LLM[127].

D'autre part, il est possible d'ores et déjà de faire fonctionner en local sur un PC équipé d'une accélération graphique[128] d'entrée de gamme des LLM dédiés à la traduction comme TranslateGemma, ce qui paraissait inenvisageable il y a seulement un an. Les progrès à venir tant en termes de performance des ordinateurs, que de performance et d'optimisation des modèles et des techniques d'entraînement ne pourront que renforcer cette technologie, en offrant l'accès à des textes traduits de meilleure qualité rédactionnelle.

Conclusion

Quels enseignements dans la perspective d'applications juridiques et judiciaires des technologies de traduction automatique ?

Un premier enseignement est que, pour la nature des tâches de traduction intéressant les juristes, à savoir la traduction de formes longues de textes présentant à la fois une sophistication et le recours à des sous-langages spécialisés, les technologies actuelles ne produisent pas encore de textes cibles du même niveau qu'un traducteur professionnel, même si cet écart tend à se réduire.

> **Pour la nature des tâches de traduction intéressant les juristes, les technologies actuelles ne produisent pas encore de textes cibles du même niveau qu'un traducteur professionnel,**

Pour les applications les plus critiques, la traduction pleinement opérée ou réalisée, sinon *a minima* validée, par un traducteur humain qualifié apparaît indispensable. Pour d'autres, elle n'est peut-être pas nécessaire. Pour décider des situations concrètes dans lesquelles il est possible d'avoir recours à la traduction automatique sans supervision d'un traducteur humain, le déterminant principal n'est pas l'outil technologique, mais bien le niveau de qualité nécessaire dans chaque situation d'usage particulière. Comme le souligne l'agence Eurojust, « même si les documents traduits automatiquement ne pourront pas être utilisés comme preuves, ils fourniront au moins un aperçu rapide et des indications sur les parties des documents qui pourraient être les plus importantes à faire traduire officiellement »[129].

Le recours à la traduction automatique depuis une langue que l'utilisateur ne maîtrise pas percute les principes de la supervision humaine : sa situation d'incompétence absolue le place dans une dépendance totale vis-à-vis de l'outil de traduction et nécessite donc un degré de confiance extrêmement élevé, qui ne peut sans doute venir que d'un processus de certification pour l'usage qu'il en fait. Il est important de rappeler ici cette difficulté évoquée précédemment : la qualité de traduction d'un outil multilingue est variable selon chaque binôme de langues. Par ailleurs un bon niveau de performance générale ne présage que partiellement des performances dans une sous-langue technique telle que le droit.

126. https ://translatelocally.com/downloads/.
127. Voir « L'émergence des petits modèles de langage spécialisés pour la traduction » et « Les performances actuelles des outils de traduction automatique ».
128. L'accélération graphique permet de confier certaines tâches d'affichage ou de calcul visuel à la carte graphique plutôt qu'au processeur central et d'obtenir ainsi de meilleures performances, une image plus fluide et une utilisation plus efficace des ressources.
129. EU-LISA, Eurojust, 2022 op.cit.

Cependant la situation du procès n'est pas celle de l'analyse d'une pièce isolée et autonome. Au-delà des documents pour lesquels le législateur impose indirectement une traduction par un expert donc humaine, les éléments traduits s'insèrent dans une histoire que les parties connaissent partiellement et dans un dossier qui donne aux magistrats un contexte d'appréciation. Par ailleurs, la présence au dossier de documents n'implique pas leur caractère décisif – soit parce qu'ils sont redondants, soit parce qu'ils ne sont pas nécessaires à la décision à intervenir. De surcroît, la procédure contradictoire a pour conséquence que les documents sont mis en débat, les parties pouvant dans certains cas discuter du contenu d'une traduction et/ou solliciter une contre-traduction, soit de droit, soit éventuellement après avoir produit elles-mêmes une autre traduction automatique. Les instruments du procès ouvrent peut-être un autre regard sur les moyens d'opérer le contrôle humain des traductions automatiques : plutôt que de l'exercer *a priori* sur tous les éléments versés en langue étrangère, une piste de réflexion pourrait être d'en faire un objet du débat judiciaire et de le réaliser *a posteriori*, sous le contrôle du juge, selon la sensibilité des éléments de preuve concernés.

Un deuxième enseignement est que les améliorations considérables de performance de traduction reposent pour une part significative sur la croissance de la taille des modèles, qui a pour corollaire la nécessité à l'heure actuelle de faire réaliser les traductions au sein de centres de calcul. Le transfert de documents que cela implique rend nécessaire le respect de différents régimes de protection des données et des secrets de l'enquête et de l'instruction, professionnel et des affaires. Il est par exemple difficile d'envisager la traduction de pièces se trouvant dans les actes d'un dossier d'instruction en cours, alors même qu'elles ne se trouvent aujourd'hui dans aucun système informatique centralisé jusqu'à leur cotation. Le besoin croissant en calcul est en partie contrebalancé, tant par les progrès extrêmement rapides de l'électronique en puissance de calcul, que par les progrès des logiciels qui permettent d'envisager l'exécution à distance de modèles de traduction.

Un troisième enseignement souligne l'importance d'explorer les possibilités et la pertinence de la spécialisation des modèles de traduction pour les sous-langues juridiques et judiciaires. L'ensemble des travaux de recherche montrent une amélioration significative de la qualité des traductions lorsque les systèmes sont affinés à partir de données d'entraînement spécifiques au domaine de spécialité, même si cela ne va pas sans d'importants défis à relever. Outre les questions économiques que pose le coût très important de l'entraînement de grands modèles au regard des bénéfices qui peuvent être attendus en termes de qualité, cette perspective dans le cas de la traduction pose l'épineuse question de la constitution des corpus d'entraînement. Le volume des traductions de pièces judiciaires opérées chaque année fournit un potentiel important de données d'entraînement en provenance de procédure closes de bonne qualité car traduites par des experts humains. La question de leur réemploi à cette fin soulève de multiples questions, notamment celle du risque, qui reste à qualifier, de voir le système révéler une partie de ses données d'entraînement non anonymisées, ou celle de la difficulté particulière de procéder à l'anonymisation nécessaire dans une langue étrangère, qu'elle soit la langue source ou la langue de destination. Ces difficultés semblent dépassables à terme. Le lancement d'une initiative de recensement ou de constitution d'un corpus de traductions judiciaires alignées à partir des traductions opérées dans le cadre de procédures judicaires, voire l'ajout aux missions des traducteurs d'une tâche d'anonymisation des textes sources et cibles, pourrait constituer un investissement intéressant pour le développement de glossaires à destination des outils de traduction assistée.

« Le besoin croissant en calcul est en partie contrebalancé, tant par les progrès extrêmement rapides de l'électronique en puissance de calcul, que par les progrès des logiciels qui permettent d'envisager l'exécution à distance de modèles de traduction.

Atelier n°6
La synthèse d'écritures et de dossiers : promesses, risques et usages maîtrisés du résumé automatique pour la justice

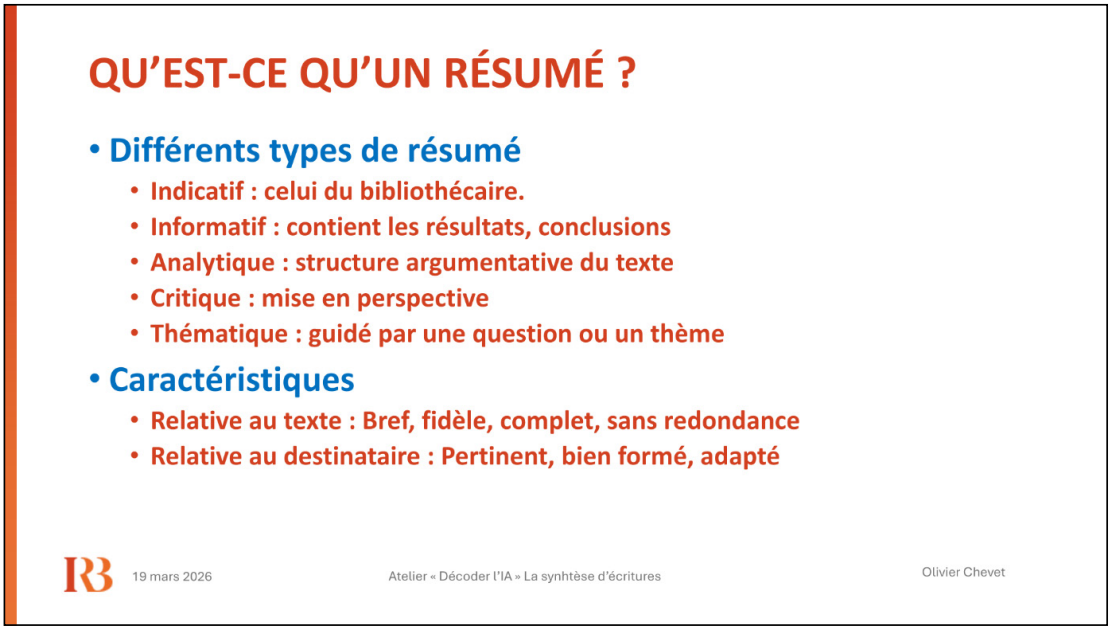
I - Le résumé, un objet protéiforme

Parmi la grande diversité des potentialités offertes par les technologies d'intelligence artificielle, les techniques génératives, qui permettent la production de textes bien rédigés et structurés, ont apporté un important renouveau. Elles se sont assez naturellement déclinées dans la production automatique de résumés et cela dans de très nombreux domaines de la connaissance. Le droit et la justice n'échappent pas à cette dynamique. Cet essor impose de clarifier ce que recouvre, en pratique, la notion même de « résumé ». L'intitulé de l'atelier – « La synthèse d'écritures et de dossiers : promesses, risques et usages maîtrisés du résumé automatique pour la justice » – témoigne d'ailleurs de l'ambivalence de cet objet à la fois technique, fonctionnel et normatif. Dans le contexte de l'institution judiciaire, l'objectif auquel doivent répondre les résumés ne consiste pas seulement en la production de textes plus courts que les textes originaux, mais en la transformation d'un corpus documentaire en une représentation exploitable dans un cadre décisionnel. Cette perspective justifie d'interroger simultanément la définition du résumé, les attentes institutionnelles et les limites structurelles de l'automatisation.

Le résumé ne constitue pas une catégorie homogène, mais un ensemble de pratiques différenciées selon les finalités poursuivies. On distingue classiquement les résumés indicatifs, visant à orienter le lecteur potentiel dans un ensemble d'ouvrages, les résumés informatifs, restituant résultats et conclusions, et les résumés analytiques, centrés sur la structure argumentative. À ces formes s'ajoutent encore des résumés critiques ou thématiques, qui introduisent un point de vue ou un prisme de sélection. Cette typologie succincte révèle que le résumé n'est jamais une simple compression neutre du texte : il procède toujours d'un point de vue et d'une hiérarchisation implicite de l'information. En matière judiciaire, cette pluralité est déterminante, car elle conditionne la manière dont les faits, les demandes, les moyens et les qualifications sont reconstruits et représentés.

> Dans le contexte de l'institution judiciaire, l'objectif auquel doivent répondre les résumés ne consiste pas seulement en la production de textes plus courts mais en la transformation d'un corpus documentaire en une représentation exploitable dans un cadre décisionnel.

Suivant une vision plus large de la représentation abrégée du contenu d'un document, les activités d'extraction de données sous formes codifiées apparaissent relever d'une logique proche, qu'il s'agisse de produire des notices bibliographiques ou documentaires ou de rattacher un document à des classifications ou à



des typologies. Pour ce type d'application, le destinataire est le plus souvent un autre système informatique. Il s'agit alors de produire une représentation codifiée dans un langage informatique, en vue d'isoler des entités factuelles, comme des références à des personnes, des lieux ou des biens, ou des références juridiques, comme des textes ou des concepts légaux. Parfois cela peut s'opérer au moyen de représentations informatiques plus sophistiquées, comme des graphes relationnels entre les entités détectées.

Les qualités attendues d'un résumé se répartissent en deux registres distincts. D'une part, il doit répondre à des exigences relatives au texte source : fidélité, complétude et absence de redondance. D'autre part, il doit satisfaire à des exigences relatives aux besoins et attentes des destinataires : brièveté, pertinence, lisibilité, conformité à des conventions formelles et adéquation au niveau de langue attendu. Cette seconde dimension est décisive dans le contexte judiciaire, où le résumé s'inscrit dans des chaînes d'usage spécifiques (préparation de l'audience, rédaction de décisions, diffusion de la jurisprudence). Ainsi, un même document pourra donner lieu à des résumés différents selon qu'il est destiné à un magistrat, à un greffier ou à un justiciable, voire selon le moment procédural. Le résumé apparaît alors aussi comme une opération de transposition

« Les qualités attendues d'un résumé se répartissent en deux registres distincts. Répondre à des exigences relatives au texte source : fidélité, complétude et absence de redondance. Satisfaire à des exigences relatives aux besoins et attentes des destinataires : brièveté, pertinence, lisibilité, conformité à des conventions formelles et adéquation au niveau de langue attendu.

dans un certain contexte, et non de simple réduction du volume de caractères.

Pour approcher au plus près la notion de résumé telle qu'elle s'appréhende dans le milieu du droit et de la justice, l'analyse des travaux du ministère de la Justice est une entrée féconde. Les orientations stratégiques du ministère de la Justice[130] confirment ainsi l'importance fonc-

130. Voir Haffide Boulakras, « Rapport sur l'IA au service de la justice : stratégie et solutions opérationnelles », 2025.

tionnelle de la synthèse documentaire parmi les applications potentielles de l'IA.

Dans le cadre d'une première phase de son programme de déploiement d'outils d'intelligence artificielle, le ministère de la Justice travaille à mettre à la disposition de ses agents, sur une plateforme souveraine, un grand modèle de langage généraliste adapté aux réalités judiciaires, qui permet alors d'envisager son utilisation avec des documents judiciaires. Parmi les services attendus de ce type d'outil, la synthèse d'écritures figure en bonne place, le rapport précité évoquant l'« extraction rapide des informations clés », la « synthèse de contenus », la « comparaison de plusieurs documents pour repérer les similitudes » et très explicitement la « création de résumés ou de tableaux comparatifs ». On retiendra que la plateforme de l'AMIAD (Agence ministérielle pour l'IA de défense) du ministère de la Défense, qui a inspiré en partie les réflexions en ce sens au sein du ministère de la Justice, propose une fonctionnalité appelée « Synthétiseur », qui permet l'extraction des « idées et informations importantes […] dans un format synthétique et personnalisable ». À ce stade, la synthèse de document apparaît donc comme une capacité transversale de l'outil d'IA, immédiatement mobilisable, mais encore peu spécialisée.

Dans une seconde phase de ce programme seront déployés des outils mieux adaptés pour répondre à 12 cas d'usages prioritaires identifiés par les personnels judiciaires. Parmi ceux-ci figurent notamment l'aide à la rédaction et la synthèse au civil (cas n° 6) et au pénal (cas n° 9)[131]. Les projets opérationnels tels que « Mon assistant civil » et « Mon assistant pénal » illustrent cette évolution vers des usages de l'IA intégrés à des applicatifs métier. Le premier doit servir à résumer les conclusions des parties, identifier les convergences et divergences et prérédiger l'exposé du litige ; le second à détecter la nature des infractions et fournir des vues synthétiques du dossier.

La présentation de ces outils révèle des attentes qui se structurent autour de catégories juridiques : faits constants, procédure, prétentions, moyens, ou encore concordances et contradictions entre pièces. Le résumé attendu n'est alors plus de type générique, mais sa finalité est orientée par des schémas de lecture propres au raisonnement judiciaire. Il s'agit moins de condenser un texte que de reconfigurer un dossier selon les besoins de l'analyse juridictionnelle. Le résumé devient ainsi une composante d'un processus plus large de production de la décision, situé à l'interface entre traitement documentaire et raisonnement juridique.

« Le résumé attendu n'est alors plus de type générique, mais sa finalité est orientée par des schémas de lecture propres au raisonnement judiciaire.

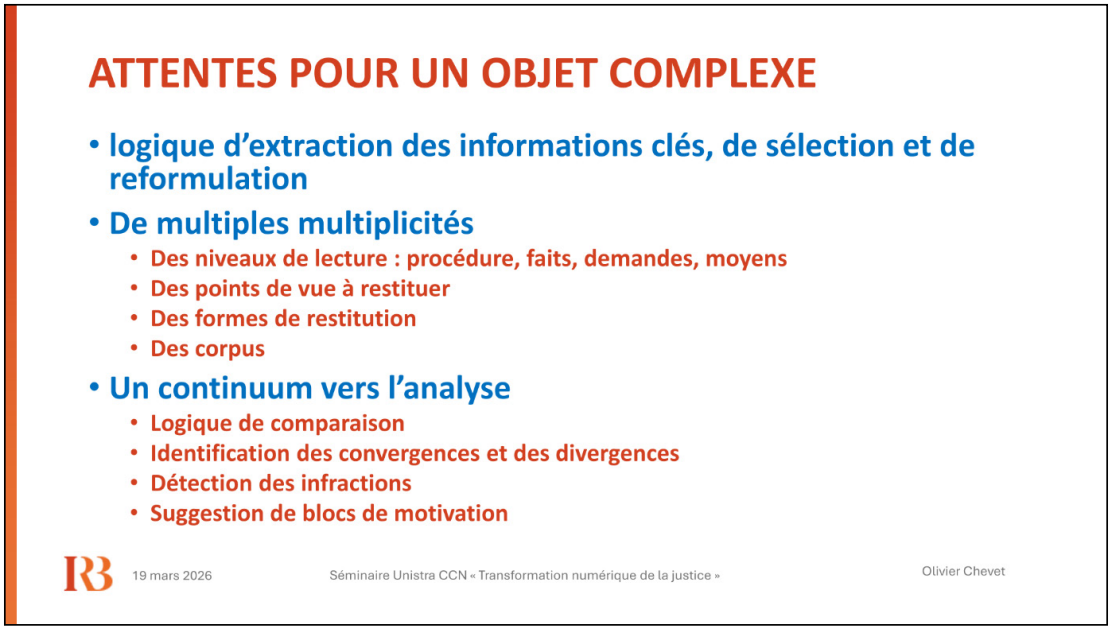
Cette compréhension large des attentes explique la prudence exprimée par la Cour de cassation quant à l'usage du résumé automatique pour ses besoins propres[132]. Celle-ci souligne les risques d'analyses parcellaires, de perte de contexte et de contresens, qui sont susceptibles d'alourdir le travail de vérification plutôt que d'alléger la tâche des personnels. En conséquence, la Cour privilégie l'exploration d'approches alternatives mais procédant de la même dynamique, telles que la mise en évidence de passages pertinents, qui autorisent le maintien d'un contrôle humain élevé. De manière significative, elle exclut également le recours au résumé automatique pour la diffusion de la jurisprudence, en raison des exigences de précision s'attachant à cette activité. Cette position met en lumière une tension centrale entre productivité et fiabilité.

L'analyse comparée des usages internationaux, telle qu'elle ressort des annexes du rapport précité[133], renforce ce constat. Les termes mêmes de « résumé » et de « synthèse » y apparaissent peu, au profit de notions comme l'aide à la rédaction ou la personnalisation des contenus. Cette absence interroge : la problématique de la synthèse, particulièrement saillante en France, pourrait ainsi recouvrir des enjeux plus larges liés à la structuration de l'information juridique.

131. *Ibid.*, p. 32.

132. Sandrine Zentiara. 2025. *IA : Préparer la Cour de cassation de demain*. Cour de cassation. https ://www.courdecassation.fr/publications/autre-publication-de-la-cour/ia-preparer-la-cour-de-cassation-de-demain-cour-de.

133. *Ibid.*, p.125 « Annexe 2 : études du service des relations internationales ».



En définitive, le résumé judiciaire automatisé constitue un objet intrinsèquement complexe, situé à l'intersection de plusieurs logiques : extraction, sélection et reformulation. Cette complexité est accentuée par la multiplicité des niveaux de lecture (faits, procédure, moyens), des points de vue à restituer et des formes de restitution attendues. Elle explique également la porosité entre résumé et analyse, dès lors que l'on introduit des opérations de comparaison, de détection ou de qualification. Comprendre cette nature hybride apparaît comme un préalable indispensable, en particulier au regard du possible décalage avec les capacités actuelles des grands modèles de langage à produire des résumés dans les formes et suivant les mécanismes d'évaluation avec lesquels ils ont été entraînés.

> Le résumé judiciaire automatisé constitue un objet intrinsèquement complexe, situé à l'intersection de plusieurs logiques : extraction, sélection et reformulation. Complexité accentuée par la multiplicité des niveaux de lecture, des points de vue à restituer et des formes de restitution attendues.

Face à ces attentes de l'institution judiciaire, il est utile d'avoir en tête ce double constat : d'une part, la problématique du résumé automatique est déjà ancienne dans la communauté de l'intelligence artificielle et, d'autre part, les modèles de langage, et particulièrement les versions les plus récentes des grands modèles de langage, ont constitué un progrès important pour la génération automatique de résumé. Aussi, une compréhension plus intime des mécanismes par lesquels ces outils sont en mesure de parvenir à de telles productions peut aider à comprendre les atouts et les limites de ces outils, et ainsi à déterminer leur meilleur emploi et à circonscrire les risques associés à leur intronisation.

L'approche contemporaine du résumé automatique repose sur des processus relativement standardisés, qui s'appliquent de manière comparable à différents domaines, y compris juridique : on trouve en entrée un ensemble de documents, que l'on soumet à un algorithme de résumé, qui produit en sortie un texte synthétique, idéalement accompagné d'une évaluation de sa qualité.

Au cours de cet atelier, l'accent a été mis sur les modèles génératifs, qui évoquent des résumés de type *abstractif par opposition à extractif*.

Cette distinction est essentielle car, contrairement aux approches extractives qui consistent à sélectionner et réassembler des segments du texte d'origine, les modèles abstractifs génèrent un texte nouveau, reformulé et censé capturer l'essentiel du contenu initial.

> Pour comprendre les implications de ces outils de résumé automatique dans le contexte judiciaire, il est nécessaire de distinguer plusieurs types d'usage des modèles de langage, qui reposent sur des logiques différentes.

Pour comprendre les implications de ces outils de résumé automatique dans le contexte judiciaire, il est nécessaire de distinguer plusieurs types d'usage des modèles de langage, qui reposent sur des logiques différentes.

Un premier ensemble correspond à des tâches de reformulation. Dans ce cas, le modèle n'est pas sollicité pour apporter de nouvelles informations, mais pour réorganiser, traduire ou synthétiser un contenu fourni. Ce cas d'usage recouvre celui du résumé automatique, mais aussi de la rédaction assistée (lettres, rapports), de la traduction, ou encore du compte rendu de réunion. Dans ces situations, l'utilisateur fournit l'ensemble des éléments nécessaires et attend du modèle une transformation formelle : clarification, condensation, mise en forme. Le modèle agit alors comme un outil linguistique avancé, sans mobilisation significative de connaissances externes.

À l'opposé, d'autres usages relèvent d'une logique exploratoire. Il s'agit notamment de générer des idées (*brainstorming*), de formuler ou de tester des hypothèses, ou de répondre à des questions ouvertes. Dans ces cas d'usage, l'utilisateur attend explicitement du modèle qu'il mobilise les connaissances acquises lors de son entraînement. Le modèle est sollicité comme un réservoir de savoirs capable de produire des informations qui ne figurent pas dans les documents en entrée. On passe alors d'une logique de transformation à une logique de génération enrichie, par laquelle le modèle complète, extrapole, voire infère.

Lorsque les outils d'IA sont mobilisés pour produire des résumés, l'objectif général des systèmes est clair : produire des textes aussi proches que possible de ceux rédigés par des experts humains dans la même situation, tant du point de vue du contenu que de la formulation.

Pour atteindre cet objectif, trois aspects complémentaires sont abordés :

1. **le fonctionnement des modèles de résumé automatique**, avec un focus sur les approches génératives ;
2. **les méthodes d'évaluation**, qui constituent un enjeu central et particulièrement complexe dans ce domaine ;
3. **les adaptations nécessaires selon les domaines**, en particulier le domaine juridique pour lequel des stratégies spécifiques – notamment autour du contrôle des hallucinations – doivent être envisagées.

II - Comment apprend-on à un LLM à résumer ?

Les modèles de langage contemporains reposent sur un principe fondamental, souvent simplifié à l'extrême : ils sont entraînés à prédire le mot suivant à partir d'un contexte donné. Ce contexte peut être une invite (un « *prompt* » en anglais), une instruction ou plus généralement n'importe quelle séquence textuelle[134].

Cette caractéristique, bien que connue, est fréquemment sous-estimée dans ses implications. En réalité, elle emporte une contrainte structurante : au sortir de son pré-entraînement, un modèle de langage ne sait rien faire d'autre que compléter une séquence. Il ne dispose pas, en tant que tel, de capacités intrinsèques de raisonnement, de synthèse ou de compréhension au sens fort. Toutes les compétences que l'on observe – répondre à une question, résumer un texte, traduire – sont apprises *a posteriori*, à travers des phases d'entraînement spécifiques.

Il en résulte une dimension essentielle à avoir à l'esprit : les comportements attendus des modèles ne sont pas véritablement émergents au sens où ils n'apparaissent pas spontanément. Ils sont induits par les données et les procédures d'apprentissage.

134. Les propos de cette partie s'appuient en grande partie sur l'intervention de Vincent Guigue lors de l'atelier du mars 16 mars 2026.

La disponibilité des données d'entraînement est un déterminant majeur de la performance des modèles dédiés à la fonction de résumé. De ce fait, la capacité à résumer est une fonction beaucoup plus travaillée dans certains domaines particuliers. C'est par exemple le cas du résumé de news (actualités), à raison d'une forte disponibilité de données, sous forme d'articles et de résumés rédigés par des journalistes. Il est donc naturellement facile de travailler sur ces corpus et d'obtenir de meilleurs résultats à partir de ces données.

Il en est de même pour les articles scientifiques : chaque article contient le plus souvent un *abstract*, ce qui permet d'entraîner des modèles à très grande échelle. Même si la forme de ces résumés varie selon les disciplines, les modèles s'avèrent plutôt performants dans ce domaine.

Des performances satisfaisantes de production de résumé dans les domaines proches des corpus d'entraînement ne permettent pas de présupposer des performances du même ordre pour d'autres corpus ou champs de connaissance. À ce jour, il semble difficile de considérer que les modèles les plus performants disposeraient d'une capacité générale à résumer.

> Des performances satisfaisantes de production de résumé dans les domaines proches des corpus d'entraînement ne permettent pas de présupposer des performances du même ordre pour d'autres corpus ou champs de connaissance.

Une difficulté majeure réside dans la sélection des informations qui méritent d'être retenues. Pour la surmonter, il faut pouvoir identifier des régularités dans les choix humains, qui se lient par exemple à la structure du document, à la position de l'information dans le texte ou au type de contenu. Or ces récurrences signifiantes sont spécifiques à chaque domaine de connaissance, voire à chaque objectif de résumé. Mais dès lors qu'elles sont suffisamment fortes et observables, les modèles peuvent les apprendre. Chacun des cas particuliers permet d'espérer raisonnablement des niveaux de performances du même ordre dans d'autres domaines, dès lors que l'on dispose de suffisamment de données d'entraînement.

1. Le rôle déterminant des données supervisées

L'entraînement d'un modèle de langage se décompose en deux grandes phases, dont l'importance respective est souvent mal appréciée.

La première, qui nécessite le plus grand volume de calcul, consiste à exposer le modèle à de vastes corpus textuels – typiquement issus du web – afin qu'il apprenne à prédire le mot suivant dans une grande diversité de contextes. Cette phase représente environ 80 à 85 % de l'effort computationnel.

La seconde phase, plus discrète en apparence, n'en est pas moins déterminante pour les usages concrets. Elle consiste à enseigner explicitement des comportements au modèle, à partir de données supervisées : des paires de question-réponse, des instructions suivies de sorties attendues ou encore des exemples annotés dans des domaines spécifiques.

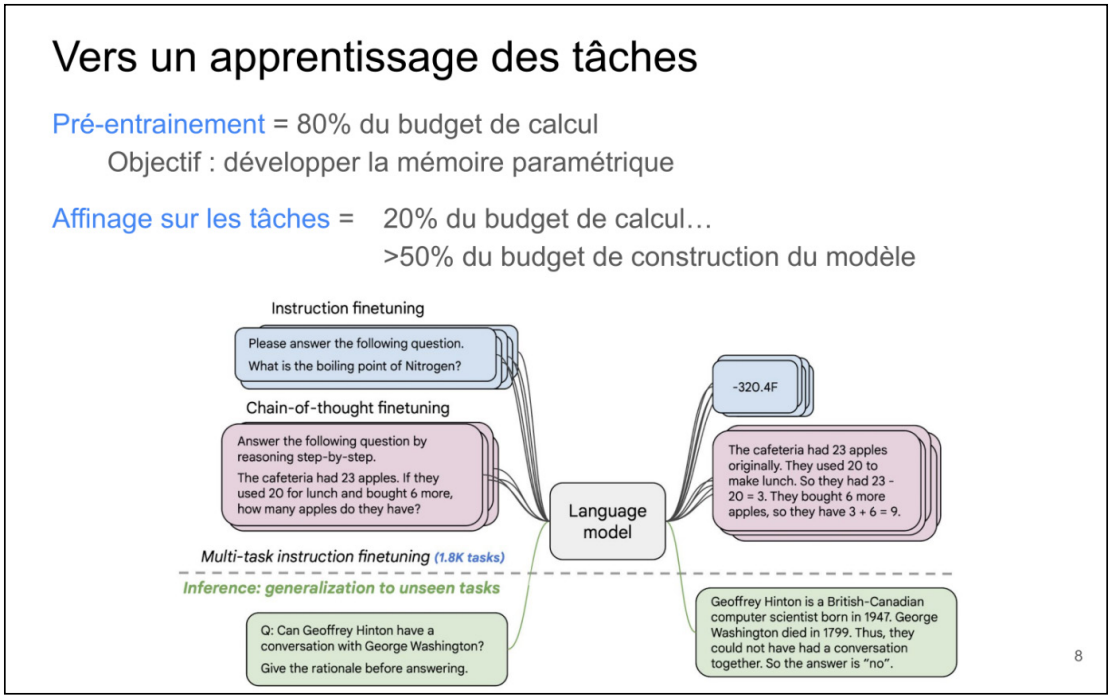
Sans cette étape, le modèle n'est pas en mesure de reconnaître qu'une entrée constitue une question, ni qu'elle appelle une réponse structurée. Autrement dit, il ne dispose pas des repères nécessaires pour produire une sortie pertinente dans un cadre applicatif donné.

Cette phase supervisée implique la constitution de bases de données massives et spécialisées, couvrant une grande diversité de domaines – scientifique, technique, juridique. Elle représente un coût considérable, souvent supérieur à celui du calcul lui-même. Le facteur limitant n'est donc pas uniquement l'infrastructure, mais bien la production de données de qualité.

L'alignement : une construction normative

À cette logique d'apprentissage s'ajoute une dimension essentielle : l'alignement. Au sortir de son entraînement, un modèle de langage n'a pas la faculté de refuser de répondre à une question, quelle qu'elle soit. Si on lui demande, par exemple, des instructions dangereuses ou illicites, il produira une réponse dès lors qu'il a appris des séquences similaires dans ses données.

Le fait qu'il s'abstienne de répondre ou qu'il reformule sa réponse de manière prudente est le résultat d'un apprentissage supplémentaire, fondé sur des interventions humaines. On



L'apprentissage à réaliser des tâches représente une part importantet du cout d'entrainement
Source : Karen Pinel-Sauvagnat et Vincent Guigue

lui apprend explicitement à ne pas répondre à certaines requêtes ou à adopter une certaine forme de discours.

Cela implique une conséquence directe : ces modèles ne sont pas neutres. Ils intègrent nécessairement une forme de ligne éditoriale, inscrite dans les choix d'annotation, de filtrage et de supervision.

L'apprentissage des modèles ne se limite pas à une opposition simple entre non supervisé et supervisé. Il existe des approches intermédiaires qui relèvent de l'auto-supervision structurée. Par exemple, on peut entraîner un modèle à déterminer si deux phrases se suivent logiquement dans un document ou si elles sont indépendantes. Si ce type de tâche ne nécessite pas d'annotation humaine explicite, il permet au modèle de développer une capacité à structurer et agréger le sens à l'échelle de la phrase ou du paragraphe.

Ces mécanismes sont particulièrement importants dans le cadre du résumé automatique, où il ne s'agit pas seulement de prédire localement des mots, mais de saisir la cohérence globale d'un texte.

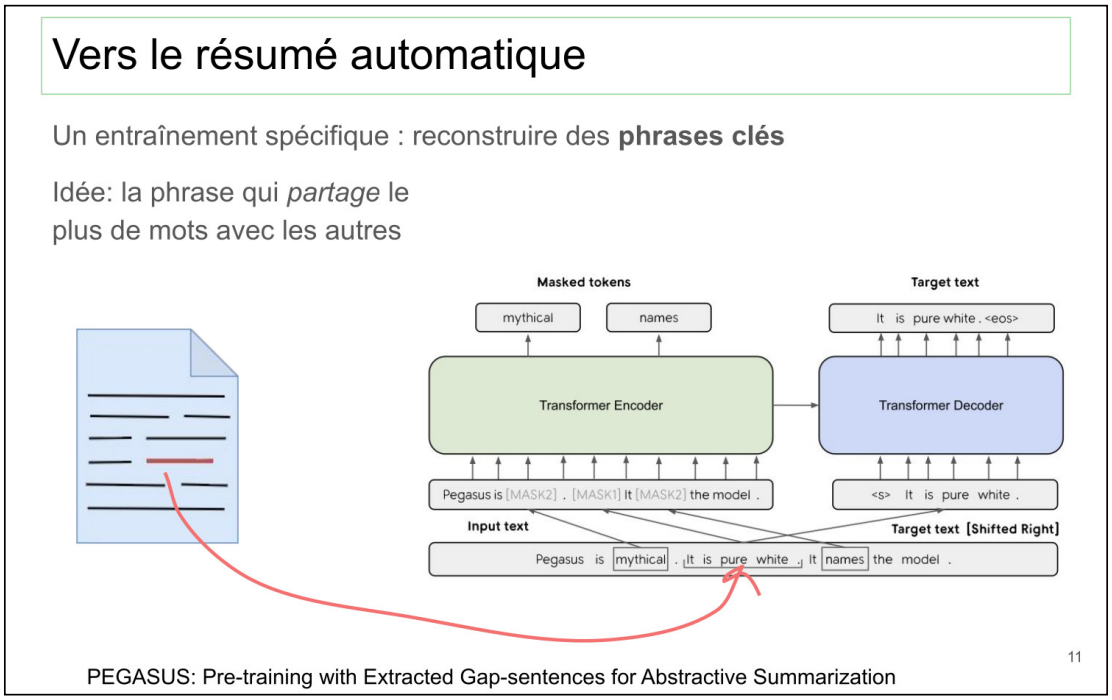
Résumé automatique : de l'extraction à la génération

Historiquement, le résumé automatique a d'abord été abordé sous un angle extractif, l'idée étant d'identifier, dans un document, les phrases les plus représentatives, selon des critères relativement simples.

Une méthode classique consiste à mesurer la similarité lexicale entre les phrases du texte : on construit un graphe dans lequel chaque phrase est reliée aux autres en fonction du nombre de mots qu'elles partagent. Les phrases les plus « centrales », c'est-à-dire les plus connectées au reste du document, sont alors sélectionnées comme candidates au résumé. Ces approches, bien que rudimentaires, présentent un avantage majeur : elles sont facilement automatisables à grande échelle.

Le résumé extractif n'a pas disparu : il a continué à être utilisé pour certaines applications. Son principal avantage reste en effet sa forte fidélité au texte source. Puisqu'il repose sur la sélection de phrases existantes, il limite fortement le risque d'introduire des erreurs factuelles ou des informations « hallucinées ».

Certaines approches hybrides ont d'ailleurs émergé : elles consistent à produire d'abord un résumé extractif, puis à utiliser un modèle de langage pour le reformuler ou le lisser. L'objectif



Apprendre à un modèle à résumer repose sur des tâches spécifiques
Source : Karen Pinel-Sauvagnat et Vincent Guigue

est alors de combiner la fiabilité de l'extraction avec une amélioration de la lisibilité.

Cependant, cette approche révèle aussi les limites structurelles du résumé extractif. Pour le destinataire, les résumés sont souvent peu fluides, la structure peut paraître fragmentée ou heurtée et leur lecture est moins naturelle qu'un texte rédigé de manière cohérente. Or les attentes ont évolué. Les utilisateurs s'attendent désormais à des résumés fluides, cohérents et proches d'un texte rédigé humainement.

Dans ce contexte, les modèles de type LLM (*Large Language Models*) ont profondément modifié les pratiques. Leur capacité à produire directement des résumés abstractifs (c'est-à-dire reformulés) de bonne qualité rend souvent inutile le passage par une étape extractive intermédiaire.

En pratique, on observe donc un basculement : on assiste, non pas à une disparition totale du résumé extractif, mais à une perte de centralité, au profit d'approches plus directes et aux résultats plus lisibles. Le compromis fondamental néanmoins reste inchangé : entre fidélité stricte (extractif) et qualité rédactionnelle/ synthèse (abstractif).

Les méthodes extractives peuvent d'ailleurs être détournées pour servir de base à l'apprentissage de modèles génératifs. Le principe est le suivant : à partir d'un document, on identifie automatiquement une phrase jugée importante ; on retire ensuite cette phrase et on demande au modèle de la reconstruire à partir du reste du texte.

Ce processus constitue un signal d'apprentissage : le modèle est incité à produire une information centrale à partir d'un contexte élargi. Il s'agit d'une forme d'auto-supervision, qui permet d'éviter le recours exclusif à des données annotées manuellement.

L'intérêt est double : réduire le coût de production des données, tout en permettant l'entraînement du modèle à partir d'une tâche proche du résumé.

> On assiste, non pas à une disparition totale du résumé extractif, mais à une perte de centralité, au profit d'approches plus directes et aux résultats plus lisibles.

Le rôle persistant des données supervisées

Ces approches ne remplacent toutefois pas les données supervisées, qui restent indispensables.

De nombreux corpus contiennent déjà des paires document-résumé, notamment dans les articles scientifiques ou les bases juridiques. Ces ressources permettent d'entraîner directement les modèles à produire des résumés conformes à des standards humains.

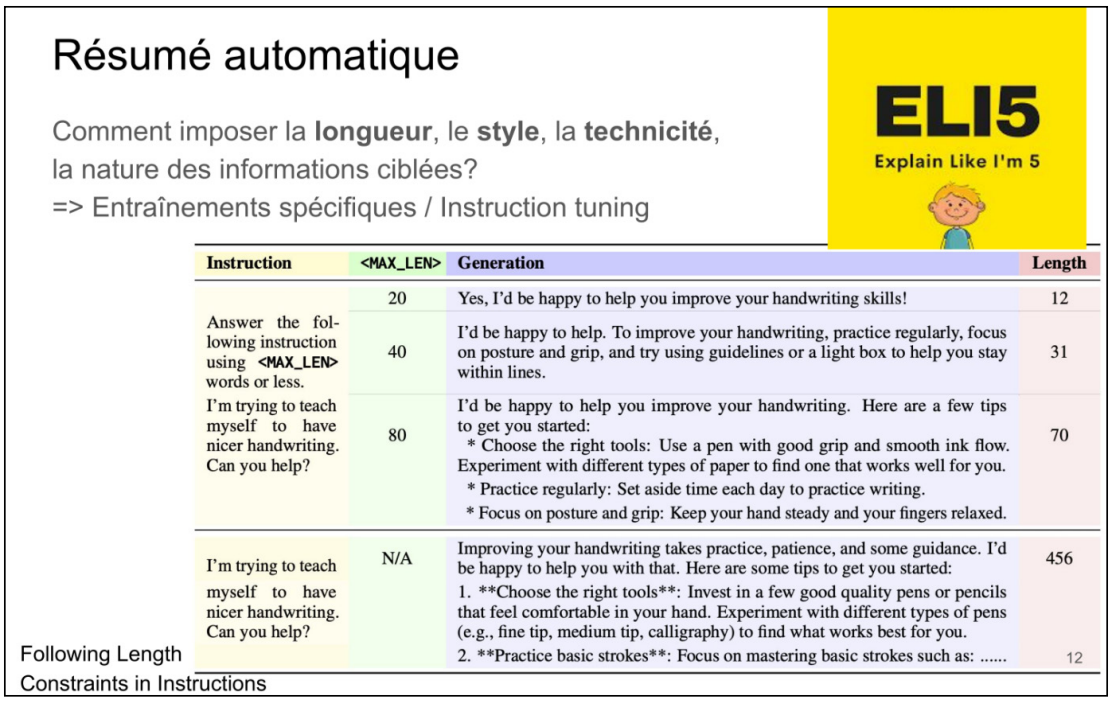
« De nombreux corpus contiennent déjà des paires document-résumé, notamment dans les articles scientifiques ou les bases juridiques. Ces ressources permettent d'entraîner directement les modèles à produire des résumés conformes à des standards humains.

Le processus est alors classique : le modèle génère un résumé mot par mot, et ses prédictions sont comparées à la référence. À chaque divergence, il est corrigé. Progressivement, il apprend à reproduire les structures et les choix rédactionnels attendus.

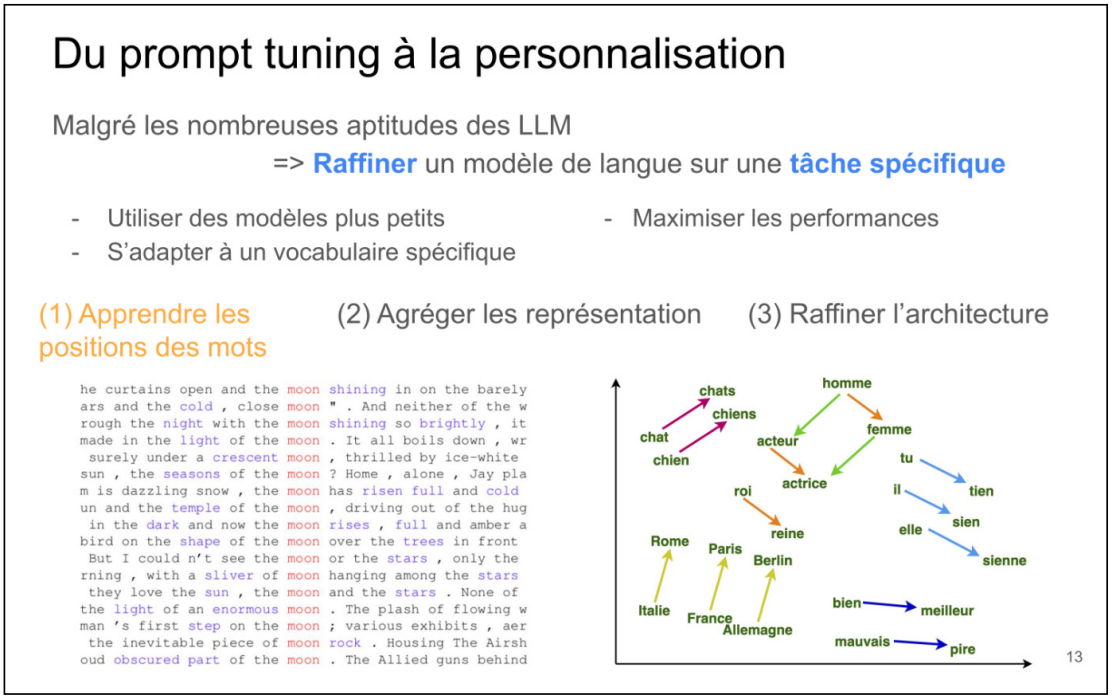
La production d'un résumé automatique ne répond pas à une définition unique. Elle dépend étroitement de l'objectif poursuivi et du destinataire final. Un résumé peut être extrêmement concis ou, au contraire, relativement développé ; il peut privilégier l'exhaustivité ou la lisibilité ; il peut adopter un vocabulaire technique ou, à l'inverse, chercher à simplifier le propos. Ces variations ne relèvent pas d'ajustements marginaux : elles traduisent des contraintes fonctionnelles distinctes, que le modèle doit apprendre à intégrer.

Or ces capacités d'adaptation ne sont pas spontanées. Elles reposent, là encore, sur des stratégies d'apprentissage supervisé. Concrètement, on expose le modèle à des exemples répondant à des consignes explicites : produire un résumé de 20 mots, de 40 mots ou de 80 mots ; reformuler un texte dans un registre simplifié ; adapter le niveau de langage à un public donné.

À travers la répétition de ces exemples, le modèle apprend progressivement à conditionner sa génération de texte selon des



ELI5 : un exemple de corpus d'apprentissage pour le résumé
Source : Karen Pinel-Sauvagnat et Vincent Guigue



Raffiner un modèle pour accroître ses performances sur des tâches spécifiques
Source : Karen Pinel-Sauvagnat et Vincent Guigue

instructions qui lui sont fournies. Un exemple emblématique est le jeu de données ELI5[135], construit autour de la tâche dite *« explique comme si j'avais 5 ans »* (*Explain Like I'm Five*). Elle consiste à reformuler des contenus complexes dans un langage accessible à un enfant. Ce type de corpus illustre bien un point fondamental : la capacité de reformulation – souvent perçue comme naturelle pour les modèles – est en réalité le produit d'un entraînement ciblé.

Sans données appropriées, un modèle ne développe pas spontanément un style, un registre ou un niveau de simplification donné.

Représentation du langage : des vecteurs au sens

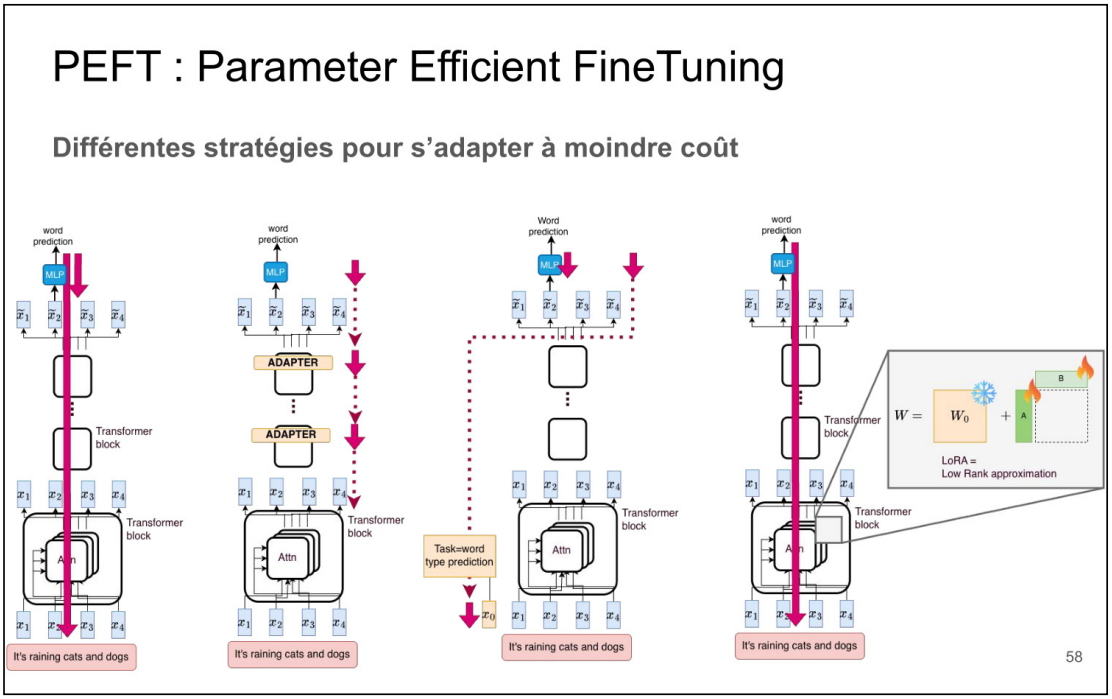
Sur le plan technique, les modèles de langage manipulent des représentations abstraites : les mots sont encodés sous forme de vecteurs dans un espace de grande dimension. Dans cet espace, la proximité entre vecteurs correspond à des similarités sémantiques ou contextuelles. Des mots proches auront tendance à apparaître dans des contextes similaires, ce qui permet au modèle de capter des relations de synonymie ou d'usage.

L'architecture de type transformer[136] a précisément pour fonction d'agréger ces représentations afin de produire une compréhension contextuelle des séquences. Le modèle ne se contente pas d'aligner des mots : il construit

« Sans données appropriées, un modèle ne développe pas spontanément un style, un registre ou un niveau de simplification donné.

135. Angela Fan, Yacine Jernite, Ethan Perez, *et al.*, « ELI5: Long Form Question Answering », *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, p. 3558-3567. https ://doi.org/10.18653/v1/P19-1346 ; consultable https ://facebookresearch.github.io/ELI5/explore.html.

136. L'architecture transformer est un modèle introduit par Google Research dans un article devenu célèbre « Attention Is All You Need » qui repose sur un mécanisme d'attention : il traite une séquence en évaluant directement les relations entre tous ses éléments, plutôt que de les parcourir un à un. Cette approche permet à la fois une meilleure efficacité de calcul et une compréhension plus globale du contexte. Elle est le fondement des modèles de la famille *Generative Pretrained Transformer* (GPT).



Le raffinage : spécialiser les modèles sans les réentraîner complètement
Source : Karen Pinel-Sauvagnat et Vincent Guigue

progressivement une représentation intégrée du sens à l'échelle de la phrase, voire du texte.

Le critère opérationnel de cette « compréhension » reste néanmoins pragmatique : si, à partir du contexte, le modèle est capable de prédire correctement le mot suivant, on considère qu'il a capturé une partie pertinente du sens.

La question devient alors la suivante : comment orienter un modèle vers un usage spécifique, comme le résumé dans un domaine particulier ? La réponse tient dans l'ajustement de ses paramètres internes. Un modèle de grande taille peut comporter des dizaines, voire des centaines de milliards de paramètres – chacun représentant une variable ajustable dans le processus de génération. L'apprentissage consiste à modifier ces paramètres de sorte que, pour une entrée donnée, le modèle produise la sortie attendue. Ce processus est au cœur de toutes les techniques d'adaptation. Pour répondre à cette contrainte, différentes stratégies ont été développées.

La première consiste à modifier l'intégralité des paramètres du modèle. Cette approche offre une grande flexibilité, mais son coût la rend difficilement déployable à grande échelle dans des environnements opérationnels. Ce *fine-tuning* complet est en effet extrêmement coûteux, tant en calcul qu'en ressources.

La seconde, plus pragmatique, consiste à n'ajuster qu'une partie des paramètres, typiquement ceux situés dans les couches finales du modèle, là où s'opère la prise de décision. Cette approche permet de spécialiser le modèle à moindre coût, tout en conservant l'essentiel des connaissances acquises lors de l'entraînement initial.

Ces méthodes de *fine-tuning* partiel constituent aujourd'hui un levier central pour adapter les modèles à des tâches spécifiques – comme le résumé juridique – sans mobiliser des ressources prohibitives.

Il faut concéder très clairement que cette activité relève fondamentalement de la recherche expérimentale. On ne sait pas, avant

Ces méthodes de *fine-tuning* partiel constituent aujourd'hui un levier central pour adapter les modèles à des tâches spécifiques – comme le résumé juridique – sans mobiliser des ressources prohibitives.

d'essayer, quelle est la meilleure manière d'adapter un modèle, de choisir les bons paramètres ou de définir la stratégie optimale pour un domaine donné.

Le processus repose donc sur des intuitions initiales, suivies de phases d'expérimentation et d'itérations successives. Si une méthode universelle existait, la question serait déjà résolue — ce qui n'est pas le cas.

En pratique, les chercheurs formulent des hypothèses liées aux caractéristiques du domaine. Par exemple, dans le domaine juridique, un axe de travail pertinent consiste à se focaliser sur les entités (personnes, dates, institutions, etc.), car elles structurent fortement les textes. Mais ce type de choix est contextuel, c'est-à-dire non transférable tel quel à d'autres domaines, et doit être validé empiriquement.

Les modèles de langage modernes sont extrêmement complexes. De par cette complexité, les approches statistiques trouvent leur limite dans la difficile compréhension de leur fonctionnement. Dans ce contexte, le terme de « bricolage » n'est pas totalement inapproprié. Cette approche s'oppose à une tradition plus classique (notamment en statistique) qui consiste à chercher à construire des modèles interprétables. Aujourd'hui, de telles ambitions restent largement hors de portée pour les modèles de langage.

S'agissant de l'adaptation des modèles au domaine du droit, les progrès pourraient être plus rapides car les signaux y seraient plus forts en raison de sa structuration logique. Mais c'est là une hypothèse de recherche, non encore pleinement validée.

Prompting et personnalisation : une adaptation par le contexte

Au-delà des mécanismes classiques d'apprentissage et de *fine-tuning*, il existe une modalité d'adaptation plus légère, mais particulièrement structurante dans les usages actuels : la rédaction d'invite ou *prompting*, et plus spécifiquement ce que l'on désigne parfois comme « pré-invite » (*pre-prompt*) ou « invite système » (*system prompt*).

Dans les interfaces usuelles, l'utilisateur a le sentiment d'interagir directement avec le modèle en formulant une invite contenant une instruction. En réalité, cette interaction est précédée – de manière invisible – par des instructions initiales, qui encadrent le comportement du modèle. Ces instructions peuvent définir un rôle, un ton, des contraintes ou des objectifs généraux.

« Une modalité d'adaptation plus légère, mais particulièrement structurante dans les usages actuels : la rédaction d'invite ou *prompting*, et plus spécifiquement ce que l'on désigne parfois comme « pré-invite » (*pre-prompt*) ou « invite système » (*system prompt*).

Cette pré-invite constitue un levier d'adaptation puissant. Elle permet d'introduire des informations contextuelles persistantes, notamment liées au profil de l'utilisateur ou à la nature de la tâche. Le modèle ne génère pas sa réponse uniquement à partir de la question posée, mais à partir de l'ensemble du contexte, incluant ces éléments implicites.

Un exemple simple permet d'en saisir les implications. Face à un terme polysémique comme « cellule », la réponse attendue varie radicalement selon le domaine d'appartenance de l'utilisateur – biologie, informatique ou droit. En intégrant cette information en amont, dans la pré-invite, il devient possible d'orienter la génération vers une interprétation pertinente.

Ainsi, la rédaction d'invite ne se limite pas à une interaction ponctuelle : elle participe d'une logique plus large de personnalisation et de spécialisation dynamique, par laquelle le modèle ajuste ses réponses en fonction d'un contexte enrichi.

Cette idée peut être poussée plus loin avec des techniques dites d'ajustement des invites[137]. Plutôt que de modifier les paramètres internes du modèle, on agit sur les éléments d'entrée, en apprenant au modèle à construire des séquences initiales optimales. Ces séquences, parfois invisibles pour l'utilisateur, fonctionnent comme des « vecteurs d'instruction » qui orientent la génération de texte.

Dans certains cas, ces pré-invites peuvent elles-mêmes être apprises automatiquement, en fonction des préférences ou des usages observés. On entre alors dans une logique d'adaptation continue : le système affine

137. *Prompt tuning* en anglais.

progressivement la manière dont il doit interpréter les requêtes et formuler les réponses.

Cette approche présente un avantage majeur : elle permet de spécialiser un modèle sans en modifier l'architecture ni les paramètres internes, ce qui en réduit considérablement le coût.

2. Une limite structurelle : l'absence de garantie de véracité

Le processus d'apprentissage inscrit dans les paramètres une représentation condensée et statistique du corpus d'entraînement. Pour la génération, le modèle combine cette mémoire probabiliste avec le contenu du *prompt*. Le texte produit sera un mélange du contenu reçu et des connaissances probabilistes du modèle. Il dépend du contenu et des instructions de l'invite (ex. : « produire un résumé avec telles caractéristiques »), mais il est également fonction des capacités apprises du modèle à respecter plus ou moins fidèlement les instructions reçues. Il en résulte une limite structurelle : le modèle peut générer des informations plausibles, bien formulées, mais factuellement incorrectes. Ce phénomène, souvent qualifié d'hallucination, découle directement de la nature probabiliste du processus de génération.

Dans les situations où l'utilisateur fournit peu de contenu, le modèle aura davantage recours à sa mémoire probabiliste, et il est plus susceptible de produire des hallucinations. Cela correspond au comportement attendu dans des usages créatifs. Un *prompt* comme « Parle-moi de la chanson "Madame" de Barbara », titre un peu confidentiel de l'artiste, donne selon les modèles et leurs versions des textes souvent sans rapport avec le propos de l'œuvre.

Dans les cas de reformulation et analyse de documents, comme le résumé, le cadre est beaucoup plus contraint par le contexte. L'utilisateur fournit un ou plusieurs documents, pose une question et attend que la réponse soit sans contradiction avec le contenu communiqué. Cette contrainte est centrale dans des domaines comme le droit, où la traçabilité et la fidélité aux sources conditionnent la validité de l'analyse.

Pour répondre à cet objectif de fidélité, des architectures spécifiques ont été développées, regroupées sous le terme de « génération augmentée par récupération (RAG) »[138]. Ces architectures reposent sur une séparation fonctionnelle en deux composantes : un module de recherche d'information, chargé d'identifier, au sein d'un corpus, les segments pertinents pour répondre à une question ; un module de génération, qui produit une réponse à partir des éléments extraits.

En pratique, le système ne travaille pas directement à l'échelle du document entier, mais à celle de fragments plus fins, généralement des paragraphes. Cette granularité permet d'améliorer la précision de la recherche et de limiter le bruit informationnel.

L'intérêt de l'approche RAG pour le résumé est double : elle permet de restreindre l'espace de génération aux seules informations jugées pertinentes et elle introduit une forme d'ancrage de la réponse dans des contenus identifiables.

Apprendre la fidélité : ne pas ajouter d'information

Cependant, ce type d'architecture ne suffit pas en lui-même à garantir la fidélité des réponses. Le modèle génératif conserve sa tendance naturelle à compléter de manière plausible, y compris au-delà des informations fournies. Il est donc nécessaire de l'entraîner explicitement à ne pas extrapoler.

Des protocoles d'apprentissage ont été développés en ce sens. Par exemple, par la construction de bases de données contenant des affirmations volontairement incorrectes – telles que « la tour Eiffel est située à Rome » – et l'association de tests simples.

« Le modèle génératif conserve sa tendance naturelle à compléter de manière plausible, y compris au-delà des informations fournies. Il est donc nécessaire de l'entraîner explicitement à ne pas extrapoler.

138. *Retrieval-Augmented Generation* (RAG).

L'objectif ici n'est pas de transmettre une connaissance correcte, mais de tester la capacité du modèle à se limiter strictement au contenu du corpus, même lorsque celui-ci est erroné. Ce type de dispositif permet d'évaluer et de renforcer la discipline du modèle face à ses données d'entrée.

Une difficulté majeure : apprendre à ne pas répondre

Un enjeu particulièrement délicat consiste à apprendre au modèle à reconnaître ses limites.

Dans un cadre idéal, si une question porte sur une information absente du corpus, le système devrait répondre explicitement qu'il ne dispose pas des éléments nécessaires. En pratique, cette capacité est difficile à obtenir.

Le comportement par défaut d'un modèle de langage est de produire une réponse plausible, même en l'absence de fondement. L'apprentissage du « non-savoir » – c'est-à-dire la capacité à répondre je ne sais pas – nécessite donc un entraînement spécifique, fondé sur de nombreux exemples.

> L'apprentissage du « non-savoir » – c'est-à-dire la capacité à répondre je ne sais pas – nécessite un entraînement spécifique, fondé sur de nombreux exemples.

Des jeux de données ont été constitués à cette fin, permettant d'enseigner au modèle à distinguer les situations dans lesquelles une réponse est attendue et justifiée de celles dans lesquelles elle ne l'est pas. Ce travail est essentiel pour réduire les erreurs et améliorer la fiabilité des modèles.

Les architectures agentiques : dépasser les limites des modèles de langage par une hybridation fonctionnelle

Les développements récents autour des architectures agentiques visent à répondre à une limite bien identifiée des modèles de langage : leur incapacité à exécuter de manière fiable certaines tâches élémentaires, malgré leur performance globale impressionnante.

Un exemple simple l'illustre immédiatement. Un modèle de langage est notoirement peu fiable pour effectuer un calcul et, de surcroît, l'exécution de ce type d'opération par génération de texte est inefficiente en termes de coût computationnel. En revanche, il est parfaitement capable de formuler une instruction dans un langage de programmation permettant d'appeler un outil externe – par exemple, une calculatrice – et d'en exploiter le résultat.

Cette capacité constitue le point de départ des architectures agentiques : le modèle n'est plus envisagé comme un système autonome, mais comme un orchestrateur capable d'interagir avec un ensemble d'outils spécialisés.

Dans ce cadre, le modèle de langage agit comme une interface de pilotage. Il peut interroger une base de données, appeler un service externe, exécuter un calcul via un module dédié ou encore récupérer des documents pertinents.

Cette hybridation permet de compenser les faiblesses intrinsèques des modèles, notamment en matière de fiabilité factuelle. En s'appuyant sur des sources externes identifiées, le système peut produire des réponses ancrées dans des données vérifiables et éventuellement accompagner ses réponses de références explicites.

On retrouve ici une convergence avec les architectures de type RAG[139], mais avec un degré de sophistication supplémentaire : le modèle ne se contente plus de générer à partir d'un contexte fourni, il décide dynamiquement des ressources à mobiliser.

Un autre apport structurant des approches agentiques réside dans la dissociation entre génération et vérification. Lorsqu'un modèle est sollicité pour répondre à une question ouverte, il opère dans un régime génératif : il produit une séquence de mots en fonction des probabilités apprises. En revanche, lorsqu'on lui demande d'évaluer la véracité d'une affirmation, le processus mobilisé est différent : il s'agit alors d'estimer la cohérence ou la plausibilité d'un énoncé au regard des connaissances internalisées.

Fait notable, les modèles se révèlent souvent plus fiables dans des tâches de vérification que dans des tâches de génération libre. Ce constat ouvre la voie à des architectures dans lesquelles plusieurs agents interagissent : un premier agent produit une réponse, un second agent en évalue la validité. Cette logique de boucle de contrôle permet d'améliorer la

139. Cette notion est expliquée au 2.1.2.

robustesse globale du système, en exploitant différentes facettes du comportement du modèle.

En étendant ce principe, on peut concevoir des systèmes composés de plusieurs agents spécialisés, chacun chargé d'une fonction spécifique : génération, vérification, recherche d'information, reformulation, etc.

Ces systèmes multi-agents permettent de structurer le traitement de l'information en chaînes de décisions explicites, plutôt qu'en une génération monolithique. Ils offrent ainsi une meilleure maîtrise des erreurs et une plus grande transparence dans le processus de production des réponses.

3. Une capacité générale à résumer émergente mais encore imparfaite

On observe effectivement une forme d'émergence d'une capacité générale des systèmes à produire des résumés au-delà du corpus d'entraînement. Des modèles qui n'ont pas été explicitement spécialisés pour cette tâche sont capables de produire des résumés cohérents, souvent pertinents et parfois même impressionnants au regard de l'absence d'entraînement dédié.

Cette capacité s'explique par leur apprentissage général de la langue et des structures discursives. Les modèles internalisent implicitement des mécanismes de sélection de l'information, des logiques de compression sémantique et des formes de reformulation.

Cependant, cette compétence reste imparfaite et dépendante du contexte. Plusieurs limites apparaissent, la première d'entre elles étant une variabilité selon le domaine, car les performances peuvent chuter pour des textes très techniques, spécialisés ou éloignés des données et distributions vues à l'entraînement.

Par ailleurs, si le modèle peut produire un résumé « acceptable » dans des domaines qui sont nouveaux pour lui, il peut cependant omettre des éléments importants, introduire des imprécisions ou mal hiérarchiser l'information.

4. Le stockage de la connaissance : des bases structurées aux modèles

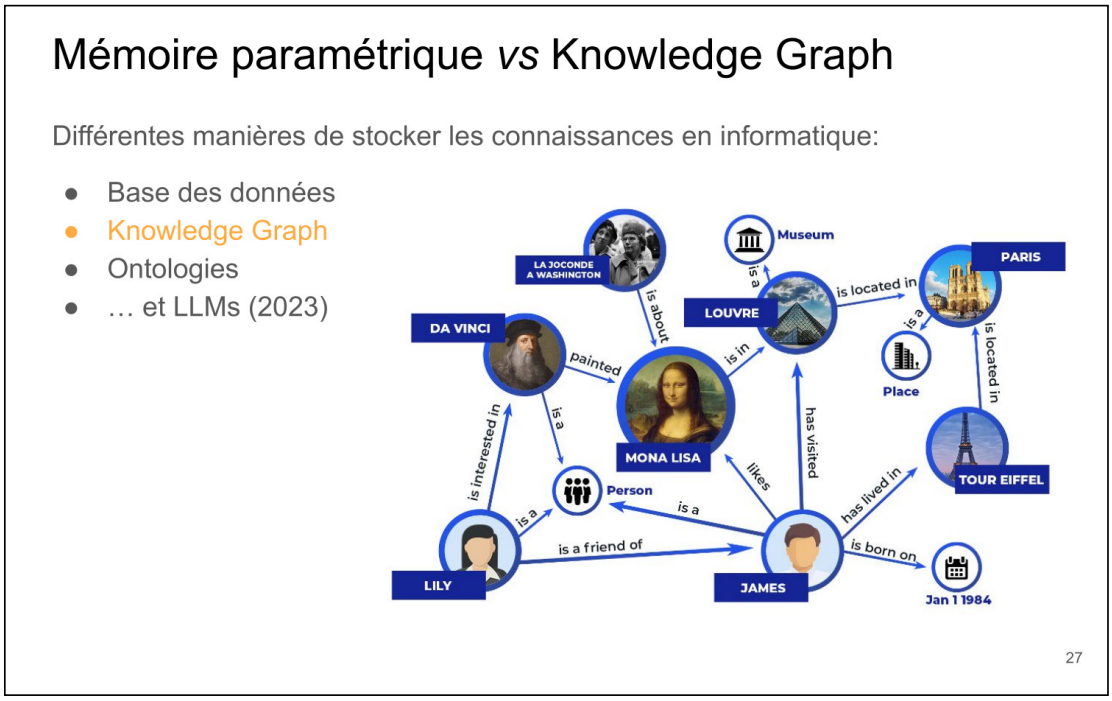
Si l'utilisation des modèles de langage la plus visible est la production de formes rédigées de résumés à destination de lecteurs humains, il est important pour appréhender les perspectives de la recherche en ce domaine de replacer ces modèles de langage dans la perspective plus large des modes de stockage de la connaissance en informatique.

De longue date, cette fonction est assurée par les bases de données, notamment relationnelles. Celles-ci reposent sur une structuration explicite de l'information : tables, relations, contraintes. Les données y sont généralement saisies ou validées par des humains, ce qui confère à ces systèmes un haut niveau de fiabilité.

Dans cette configuration, la connaissance est explicite, structurée et contrôlée. À l'inverse, les modèles de langage stockent une forme de connaissance implicite et distribuée suivant leurs paramètres. Cette connaissance n'est pas directement accessible ni vérifiable, ce qui pose des défis spécifiques en matière de traçabilité et de confiance.

L'intérêt des architectures agentiques apparaît alors clairement : elles permettent d'articuler ces deux paradigmes. Les bases de données offrent des informations fiables, structurées et auditables. Les modèles de langage apportent des capacités de traitement linguistique, de synthèse et d'interaction. En combinant ces approches, il devient possible de concevoir des systèmes capables de raisonner sur des données fiables tout en produisant des réponses adaptées au langage naturel.

« Les bases de données offrent des informations fiables, structurées et auditables. Les modèles de langage apportent des capacités de traitement linguistique, de synthèse et d'interaction. En combinant ces approches, il devient possible de concevoir des systèmes capables de raisonner sur des données fiables tout en produisant des réponses adaptées au langage naturel.



Compléter la représentation du monde du modèle par des graphes de connaissance
Source : Karen Pinel-Sauvagnat et Vincent Guigue

Graphes de connaissances : structurer l'information par les relations

À partir des années 2000, une nouvelle approche de la représentation des connaissances a été développée : celle des graphes de connaissances (*knowledge graphs*).

Contrairement aux bases de données relationnelles, qui reposent sur des schémas rigides, les graphes de connaissances proposent une structuration plus souple et plus expressive. Ils consistent à représenter des entités – personnes, lieux, œuvres, événements – et à modéliser explicitement les relations qu'ils entretiennent.

Cette approche permet de capturer une forme de connaissance plus riche, car elle ne se limite pas à stocker des attributs, mais intègre directement la dimension relationnelle du savoir.

> Cette approche permet de capturer une forme de connaissance plus riche, car elle intègre directement la dimension relationnelle du savoir.

En pratique, ces graphes sont généralement construits à partir de sources structurées et fiables. Un exemple classique est celui des « infoboxes » de Wikipédia, qui synthétisent des informations essentielles sur une entité donnée : dates, lieux, fonctions, événements marquants.

Ces informations peuvent être extraites, normalisées et intégrées dans un graphe de connaissances. Une fois structurée, cette base permet de répondre directement à certaines requêtes, sans passer par une consultation documentaire complète.

Ce mécanisme est aujourd'hui largement visible dans les moteurs de recherche. Là où l'utilisateur était auparavant renvoyé vers une page – par exemple, celle de Barack Obama – les systèmes actuels sont capables de fournir une réponse immédiate, comme son âge, accompagnée d'éléments contextuels.

Cette évolution repose sur la constitution de bases de connaissance validées, souvent enrichies par des contributions humaines et des processus de vérification.

Ontologies : intégrer la logique dans la représentation

Dans le prolongement de ces approches, l'ingénierie des connaissances a développé des structures plus formelles : les ontologies. Une ontologie ne se limite pas à stocker de

l'information ; elle intègre également un mécanisme d'inférence logique. Autrement dit, elle permet de déduire automatiquement de nouvelles connaissances à partir de faits existants.

Par exemple, si un système sait que « une personne est née dans une ville », que « cette ville appartient à un État » et que « cet État appartient à un pays », il peut en déduire, par raisonnement, la nationalité ou l'origine géographique de cette personne.

Ce type de système repose sur des règles explicites et garantit un haut niveau de cohérence logique. Il constitue historiquement l'une des formes les plus abouties de représentation structurée de la connaissance.

Les modèles de langage : une mémoire implicite et probabiliste

L'émergence récente des grands modèles de langage introduit un paradigme radicalement différent. Ces modèles ne stockent pas la connaissance sous forme explicite, mais dans leurs paramètres, selon une logique probabiliste. On parle désormais de mémoire paramétrique : une forme de connaissance distribuée résultant de l'apprentissage des probabilités de cooccurrence des mots.

Cette mémoire présente plusieurs caractéristiques :

- elle est **implicite et opaque** : elle ne peut pas être examinée, ou seulement de manière très limitée ;
- elle est **probabiliste** : elle repose sur des corrélations statistiques, non sur des règles logiques explicites ;
- elle est **imparfaite** : elle peut produire des approximations ou des erreurs, tout en restant globalement cohérente.

Malgré ces limites, cette forme de connaissance est aujourd'hui massivement utilisée, en raison de sa couverture extrêmement large et de sa facilité d'accès via le langage naturel.

Vers des architectures hybrides : extraction et reconstruction

Face aux limites respectives des approches symboliques et statistiques, des stratégies hybrides se développent. Les modèles de langage peuvent être mobilisés pour effectuer des tâches d'extraction d'information : identifier dans un texte les entités pertinentes (personnes, lieux, dates) et les relations entre elles.

Ces informations peuvent ensuite être structurées, validées (éventuellement par des experts humains) et intégrées dans un graphe de connaissances.

On obtient ainsi une chaîne complète de l'extraction automatique, structuration, validation, réutilisation pour la réponse à des requêtes. Cette approche permet de combiner la puissance de traitement linguistique des modèles et la fiabilité des structures explicites.

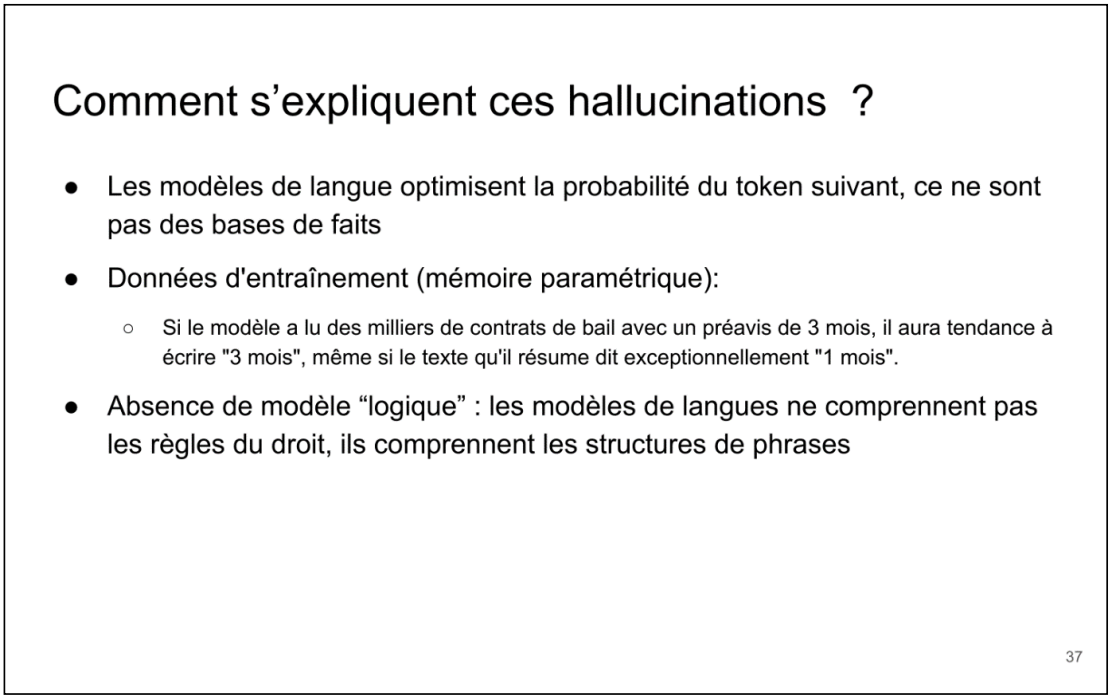
Aucune de ces approches ne constitue une solution universelle. Les bases de données et les ontologies offrent une forte fiabilité, mais une couverture limitée et un coût de construction élevé ; les graphes de connaissances enrichissent la représentation, mais restent dépendants de processus de validation ; les modèles de langage couvrent un spectre très large, mais au prix d'une perte de contrôle sur la véracité. Les systèmes contemporains s'inscrivent donc dans une logique de compromis, cherchant à articuler ces différents paradigmes pour tirer parti de leurs complémentarités.

> Les systèmes contemporains s'inscrivent donc dans une logique de compromis, cherchant à articuler ces différents paradigmes pour tirer parti de leurs complémentarités.

L'enjeu n'est donc plus tant de concevoir des modèles performants, que de les intégrer dans des systèmes complexes, capables de contrôler la provenance de l'information, de limiter les erreurs et d'adapter la réponse au contexte d'usage.

Deux écosystèmes encore largement séparés

Au-delà de la question de la qualité, il existe aujourd'hui une distinction assez nette entre deux paradigmes : les modèles de langage (LLM), qui fonctionnent principalement avec du texte brut, sont entraînés avec des entrées/sorties textuelles et produisent des réponses flexibles, mais sans garantie formelle de véracité ; les graphes de connaissances, qui relèvent d'un univers structuré et formel, s'interrogent via des langages de requête et offrent des réponses fiables mais uniquement sur un périmètre couvert et peuvent produire des non-réponses si l'information n'est pas présente.



L'origine des hallucinations
Source : Karen Pinel-Sauvagnat et Vincent Guigue

La passerelle entre ces deux mondes est précisément le domaine de l'extraction d'information. C'est elle qui doit permettre de transformer du texte en structures exploitables (graphes) et potentiellement de réinjecter ces structures dans des systèmes génératifs.

Mais cette étape reste aujourd'hui le maillon faible : malgré des progrès rapides, les performances ne sont pas encore suffisantes pour une automatisation fiable à grande échelle. Cela en fait un champ de recherche actif, mais encore loin d'être stabilisé.

Dans cette perspective, le modèle de langage apparaît moins comme une solution complète que comme un composant central, dont la valeur dépend de son articulation avec d'autres sources et d'autres mécanismes.

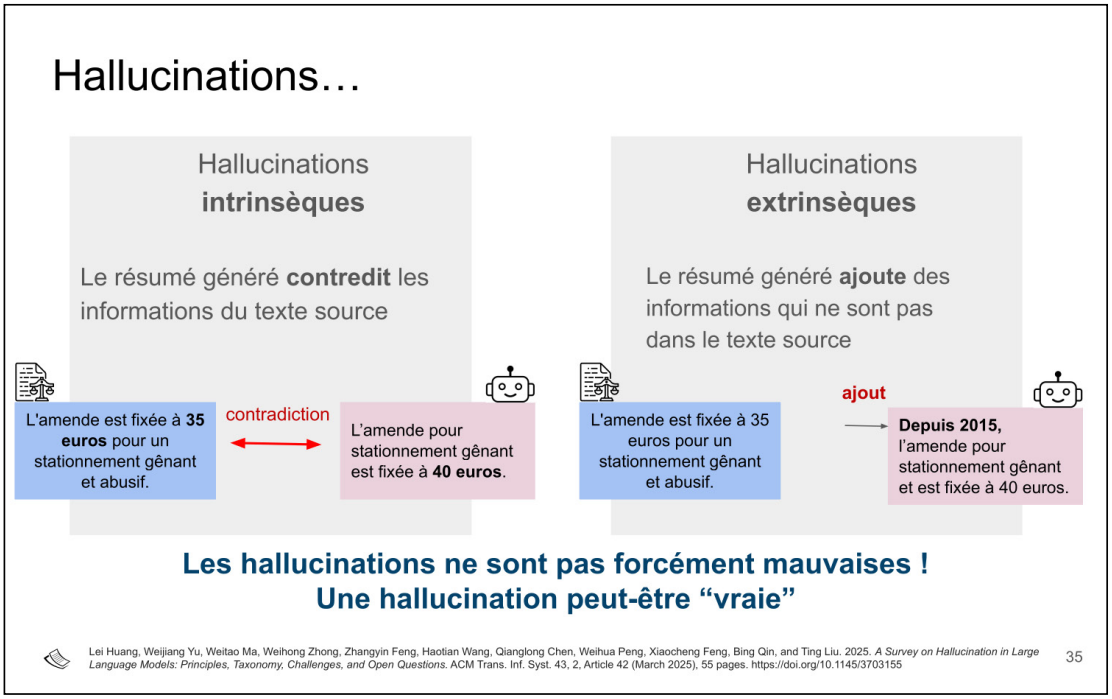
Un champ de recherche très fertile est celui du chemin inverse, à savoir la réinjection de fragments de graphes de connaissances comme source d'information pour la génération de résumé. Cette approche se heurte toutefois à une difficulté majeure : la fiabilité de l'information extraite.

Les progrès en extraction d'information sont néanmoins réels et significatifs. De manière schématique, le niveau de performance est passé de chiffres de l'ordre de 60 % à des scores proches de 85 %. Cependant, ces chiffres restent trompeurs. En pratique, cela signifie que des entités importantes (noms propres, dates) sont encore manquées, que des erreurs subsistent dans les relations entre les éléments (confusions entre événements, inversions, incohérences) et donc que le graphe produit contient encore de nombreuses erreurs structurelles.

Or un graphe de connaissances est généralement perçu comme une source hautement fiable, presque « certifiée ». Or les performances des outils d'extraction automatiques de graphes de connaissances à partir de texte sont insuffisantes. En conséquence, une utilisation fiable de ces graphes suppose encore, dans la plupart des cas, une reprise et validation humaine.

III - Comment évalue-t-on la qualité d'un résumé ?

L'évaluation de la qualité des résumés automatiques constitue un préalable à toute utilisation opérationnelle des modèles de langage. Cette question prend une dimension particulière avec les grands modèles généralistes, faute de savoir comment ils ont été entraînés à produire des résumés, avec quels corpus, pour quelles typologies et dans quels champs de connaissance. Il en résulte une tension fondamentale entre la nature du processus d'apprentissage et les exigences attendues d'un bon résumé, qui suppose



Les hallucinations peuvent contredire ou compléter le texte source
Source : Karen Pinel-Sauvagnat et Vincent Guigue

> Une tension fondamentale entre la nature du processus d'apprentissage et les exigences attendues d'un bon résumé, qui suppose à la fois sélection, condensation et cohérence sémantique.

à la fois sélection, condensation et cohérence sémantique.[140]

Dans ce contexte, la notion d'« hallucination » occupe une place centrale dans l'analyse de la qualité. Elle désigne les cas dans lesquels le modèle produit des informations problématiques par rapport au document source, mais recouvre en réalité des phénomènes distincts. Les hallucinations dites intrinsèques correspondent à des contradictions internes : le résumé altère une information présente dans le texte d'origine, par exemple en modifiant un montant ou un fait explicitement mentionné. À l'inverse, les hallucinations extrinsèques renvoient à l'ajout d'informations absentes du document initial, issues des connaissances générales apprises par le modèle. Cette distinction est essentielle, car elle révèle que toutes les hallucinations ne relèvent pas nécessairement d'une erreur brute. Certaines peuvent, dans certaines situations, améliorer la qualité informative du résumé, soit en corrigeant une inexactitude du texte source, soit en apportant un contexte pertinent. L'enjeu n'est donc pas uniquement d'éliminer ces phénomènes, mais de les qualifier et de les contrôler.

Cette problématique conduit à distinguer deux critères d'évaluation souvent confondus : la fidélité et la factualité. La fidélité renvoie à la conformité du résumé au document source ; elle mesure la capacité du modèle à restituer correctement les informations présentes dans le texte d'origine. La factualité, en revanche, évalue la véracité des informations produites au regard du monde réel ou d'un corpus de connaissances externes.

Ces deux dimensions peuvent diverger. Un résumé peut être parfaitement fidèle tout en étant factuellement faux si le document source contient lui-même une erreur. À l'inverse, un résumé peut être factuellement correct tout en

140. Les développements de cette partie s'appuient en grande partie sur l'intervention de Karen Pinel-Sauvagnat lors de l'atlier du 17 mars 2026

manquant de fidélité s'il introduit des informations absentes du texte initial, même si celles-ci sont exactes. Cette dissociation met en évidence une difficulté méthodologique majeure : définir ce qu'est un « bon » résumé dépend du cadre d'usage, notamment du degré de tolérance à l'enrichissement externe.

L'objectif recherché dans la plupart des travaux actuels consiste néanmoins à atteindre un équilibre pour produire des résumés qui soient à la fois fidèles et factuels. Cette double exigence structure les principales orientations de recherche visant à améliorer la qualité des productions.

Une première approche consiste à contraindre le modèle à se concentrer sur les entités et les faits saillants du document, afin de limiter les dérives interprétatives et d'ancrer la génération dans le contenu source. Cette stratégie est particulièrement pertinente dans des domaines spécialisés, comme le droit, dans lesquels l'identification précise des entités (parties, dates, qualifications juridiques) conditionne la validité du résumé.

Une deuxième ligne de travail repose sur des techniques de post-entraînement, qui consistent à affiner le comportement du modèle à partir d'exemples annotés. Ces méthodes impliquent généralement une intervention humaine importante, sous forme de classements ou d'évaluations comparatives des résumés générés. Elles permettent d'orienter le modèle vers des réponses jugées de meilleure qualité, mais au prix d'un coût élevé, lié à la mobilisation d'expertises pour produire des annotations fiables.

Enfin, une troisième approche intervient directement au moment de la génération, en introduisant des mécanismes de pénalisation pour les éléments qui s'éloignent du texte source. L'idée est de contraindre dynamiquement le modèle à privilégier des formulations ancrées dans le document d'origine, réduisant ainsi la probabilité d'hallucinations. Cette régulation en temps réel illustre une tendance plus générale consistant à encadrer la génération plutôt qu'à modifier uniquement l'apprentissage en amont.

1. L'importance du jugement humain

L'évaluation des systèmes de résumé automatique repose, de manière structurelle, sur l'intervention humaine. Celle-ci constitue un élément central, à la fois indispensable et pourtant difficilement mobilisable à grande échelle. Le jugement humain intervient d'abord dans la construction de ce que l'on appelle des résumés de référence, ou « vérités terrain ». Ces résumés sont conçus comme des cibles idéales, vers lesquelles les modèles doivent tendre. Ils servent également de base comparative pour mesurer la qualité des productions générées : l'évaluation consiste alors à quantifier l'écart entre un résumé produit par le modèle et ce ou ces résumés de référence.

Cette approche se heurte toutefois à une difficulté conceptuelle majeure : l'absence d'unicité du « bon » résumé. Contrairement à d'autres tâches pour lesquelles une réponse correcte peut être objectivement définie, le résumé relève d'une activité interprétative. Plusieurs résumés différents peuvent être également valides, chacun mettant en avant des aspects distincts du document source. Dès lors, la construction d'un référentiel exhaustif apparaît hors de portée. Même dans un cadre restreint, la diversité des formulations et des choix de contenu rend illusoire l'idée de couvrir l'ensemble des résumés acceptables.

Cette indétermination est renforcée par les désaccords entre experts eux-mêmes. Les études empiriques montrent que, lorsqu'ils évaluent des résumés produits par leurs pairs, les experts attribuent des scores souvent hétérogènes, y compris sur des critères fondamentaux tels que la cohérence ou la pertinence. Autrement dit, l'évaluation humaine, loin d'être parfaitement stable, est elle-même sujette à variabilité. Cette instabilité fragilise la notion même de référence et, par extension, les méthodes d'apprentissage et d'évaluation qui en dépendent.

« L'évaluation humaine, loin d'être parfaitement stable, est elle-même sujette à variabilité. Cette instabilité fragilise la notion même de référence et, par extension, les méthodes d'apprentissage et d'évaluation qui en dépendent.

À cette difficulté qualitative s'ajoute une contrainte pratique : le coût du recours à

l'expertise. Dans les conditions réelles de la recherche, les experts sont rares et difficiles à mobiliser. Les jeux de données sont donc souvent constitués par des annotateurs non spécialistes, voire des travailleurs recrutés sur des plateformes de micro-travail. Or la littérature montre que la corrélation entre les annotations produites par ces différents profils est faible, parfois proche de zéro. Il en résulte un compromis à rechercher entre la qualité et la quantité des données : les annotations les plus fiables sont aussi les plus coûteuses, tandis que les plus abondantes sont les moins précises. Cette tension affecte directement la robustesse des évaluations, qui nécessitent pourtant des volumes importants de données pour être statistiquement significatives.

Dans ce contexte, l'évaluation d'un résumé ne peut se limiter à une mesure unique. Elle s'organise autour de plusieurs dimensions complémentaires. La première consiste à mesurer la similarité entre le résumé généré et le résumé de référence, c'est-à-dire à évaluer dans quelle mesure les contenus se recoupent. Cette approche, largement utilisée, repose implicitement sur l'idée que la proximité formelle ou lexicale est un indicateur de qualité. Elle ne suffit cependant pas à capturer l'ensemble des propriétés attendues d'un bon résumé, ce qui explique le développement de critères plus fins, intégrant notamment les notions de fidélité, de cohérence et de pertinence.

L'ensemble de ces éléments met en évidence une difficulté fondamentale : l'évaluation du résumé automatique ne dépend pas uniquement des performances des modèles, mais aussi de la manière dont on définit, construit et mesure la qualité elle-même.

2. La mesure de la proximité à un résumé idéal

L'évaluation de la qualité des résumés automatiques repose sur une combinaison de critères complémentaires, qui visent à appréhender différentes dimensions du texte généré. À partir des résumés de référence, construits par des experts ou des annotateurs, il s'agit d'abord de mesurer la fidélité au document source, définie comme la capacité du modèle à restituer correctement les informations essentielles. Cette évaluation ne se limite toutefois pas à une simple correspondance informationnelle : elle intègre également des propriétés linguistiques et discursives telles que la cohérence interne, l'absence de redondance ou encore la fluidité

« L'évaluation du résumé automatique ne dépend pas uniquement des performances des modèles, mais aussi de la manière dont on définit, construit et mesure la qualité elle-même.

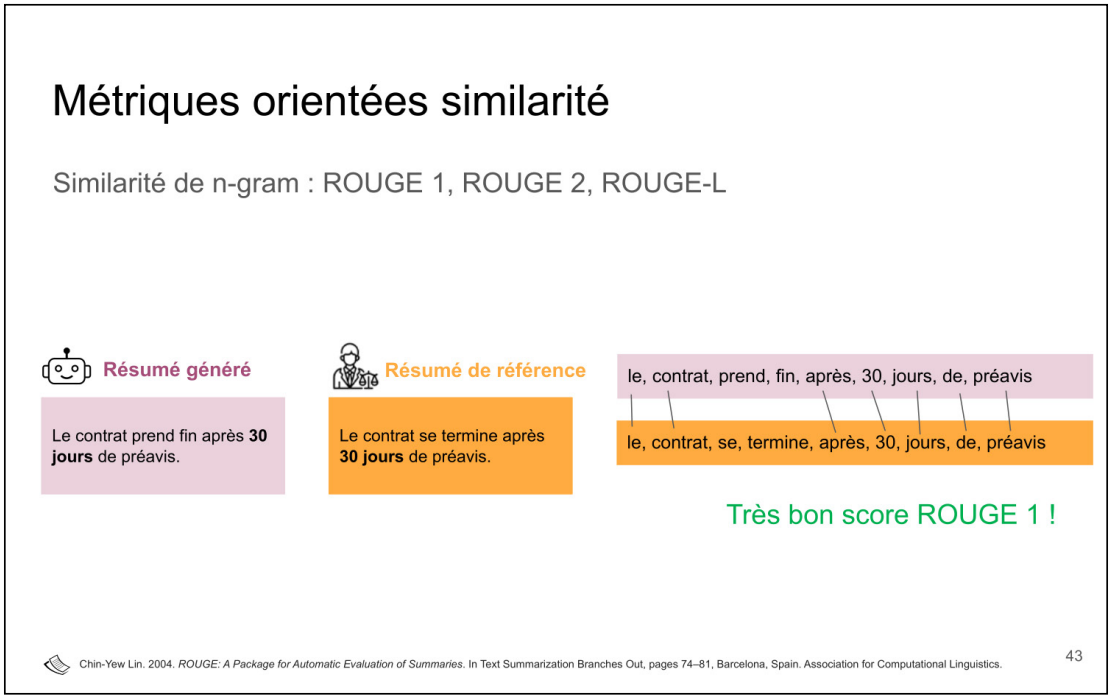
du texte. En pratique, ce champ de recherche a élaboré un grand nombre de métriques – plusieurs dizaines, voire davantage – sans qu'aucune ne s'impose comme solution pleinement satisfaisante.

ROUGE : distance lexicale au résumé idéal

Parmi ces métriques, la famille des scores ROUGE est la plus ancienne et demeure la plus largement utilisée. Leur principe est relativement simple : il consiste à comparer un résumé généré avec un résumé de référence en mesurant le chevauchement lexical. Selon les variantes, la comparaison porte sur des unités différentes – mots isolés, paires de mots ou plus longues séquences communes. L'idée sous-jacente est que, plus les deux textes partagent d'éléments en commun, plus le résumé généré est considéré comme proche du résumé attendu, et donc de meilleure qualité.

Cette approche présente l'avantage d'être facile à mettre en œuvre et à interpréter, ce qui explique son statut de métrique de référence dans la littérature scientifique. Elle permet notamment de produire des évaluations standardisées et comparables entre elles, ce qui en fait un outil incontournable, malgré ses limites. Toutefois, une analyse plus fine révèle des biais importants. En premier lieu, cette métrique ne prend pas correctement en compte l'ordre des mots, ce qui peut conduire à attribuer un score élevé à des phrases dont le sens est pourtant inversé. Une permutation des rôles syntaxiques peut ainsi passer inaperçue, alors même qu'elle modifie radicalement la signification.

En outre, les scores de type ROUGE sont peu sensibles à certaines distinctions sémantiques cruciales. Une variation minimale, comme l'ajout ou la suppression d'une négation, peut produire un énoncé diamétralement opposé sans affecter significativement le score. Cette insensibilité



Mesurer la fidélité par la similarité lexicale
Source : Karen Pinel-Sauvagnat et Vincent Guigue

aux nuances sémantiques constitue une limite majeure, en particulier dans des domaines où la précision du sens est déterminante, comme le droit. Enfin, ce type de métrique repose sur une correspondance lexicale stricte et ignore les relations de synonymie ou de reformulation. Un résumé utilisant des termes différents mais équivalents sur le plan sémantique peut ainsi être injustement pénalisé.

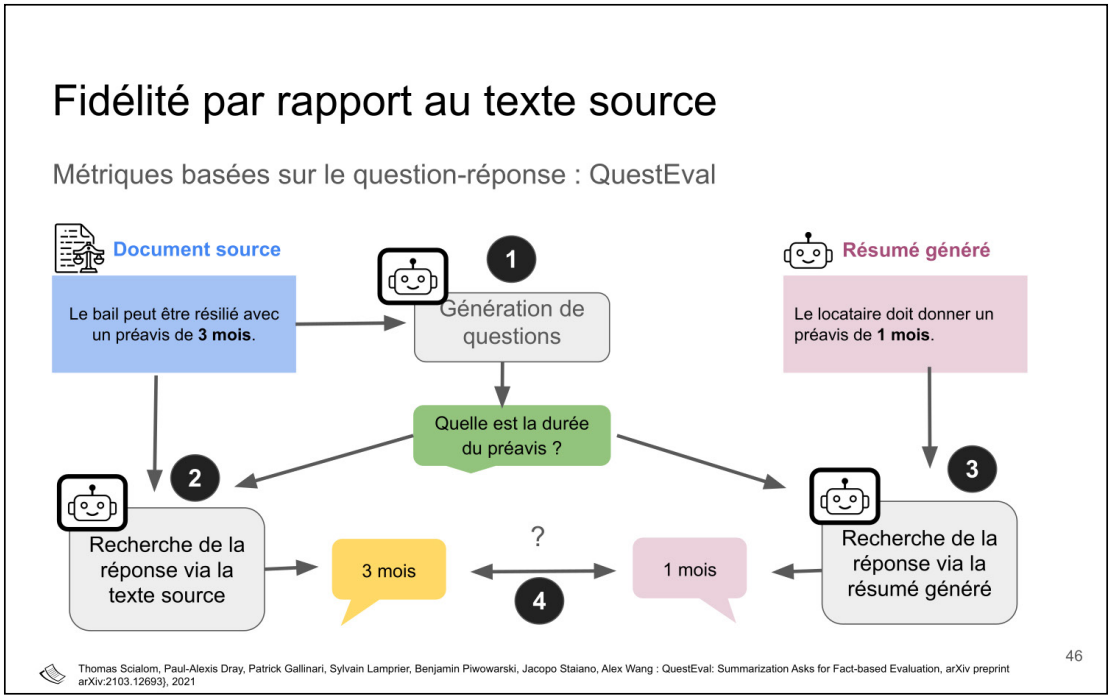
> « Cette insensibilité aux nuances sémantiques constitue une limite majeure, en particulier dans des domaines où la précision du sens est déterminante, comme le droit.

Ces limites expliquent que, malgré son usage systématique, la métrique ROUGE ne puisse être considérée comme une mesure complète de la qualité des résumés. Elle capture une forme de similarité de surface, mais reste largement aveugle à la structure du sens et aux propriétés discursives. Son maintien comme standard tient moins à sa pertinence intrinsèque qu'à son rôle de point de référence commun, permettant la comparaison entre approches. L'évaluation du résumé automatique se trouve ainsi dans une situation caractéristique : elle repose sur des outils indispensables mais imparfaits, dont l'interprétation nécessite une analyse critique.

BERTScore, la mesure de la distance sémantique

D'autres métriques ont été proposées pour dépasser les limites des approches fondées uniquement sur le recouvrement lexical. Parmi elles, le BERTScore vise à traiter plus directement la question de la similarité sémantique. L'idée consiste à ne plus comparer les mots entre eux, mais leurs représentations vectorielles issues de modèles de langage. Dans ce cadre, des phrases comme « l'avocat défend son client » et « le juriste assiste son mandaté » peuvent obtenir un bon score, dans la mesure où les termes utilisés sont proches du point de vue sémantique et donc également proches dans l'espace vectoriel.

Cette approche permet de corriger certaines limites des métriques comme ROUGE, notamment leur incapacité à prendre en compte les synonymes. Pour autant, elle ne constitue pas une solution parfaite. S'appuyer sur des



QuestEval, mesurer la fidélité par le contenu sémantique
Source : Karen Pinel-Sauvagnat et Vincent Guigue

proximités sémantiques peut conduire à lisser des distinctions importantes, car en pratique la proximité des représentations vectorielles ne correspond pas systématiquement à une équivalence de sens. Ainsi, des énoncés comme « le tribunal ordonne la saisie » et « le tribunal suggère la saisie » peuvent être considérés comme proches par le modèle, alors même que leur signification est sensiblement différente.

> S'appuyer sur des proximités sémantiques peut conduire à lisser des distinctions importantes, car en pratique la proximité des représentations vectorielles ne correspond pas systématiquement à une équivalence de sens.

Ces deux exemples de métriques – ROUGE et BERTScore – illustrent les approches qui consistent à comparer le texte généré à un résumé de référence produit par des experts. D'autres métriques cherchent à évaluer directement la fidélité du résumé par rapport au texte source. Il ne s'agit plus de mesurer un écart avec une référence humaine, mais de vérifier si le contenu généré correspond bien aux informations présentes dans le document d'entrée.

QuestEval,

Parmi ces méthodes, certaines reposent sur des mécanismes de génération de questions et de réponses. L'une d'entre elles, représentative de cette approche, est QuestEval[141]. Le principe est le suivant. À partir du document source, un premier modèle génère automatiquement des questions pertinentes. Par exemple, si le texte indique que le bail peut être résilié avec un préavis de trois mois, une question générée pourrait être : « quelle est la durée du préavis ? ». Un second modèle, appelé évaluateur, est ensuite mobilisé pour répondre à cette question à partir du document source, puis à partir du résumé à évaluer.

L'évaluation repose alors sur la comparaison entre les deux réponses. Si les réponses

141. Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier *et al.*, « QuestEval: Summarization Asks for Fact-Based Evaluation », *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021, p. 6594-6604, https ://doi.org/10.18653/v1/2021.emnlp-main.529.

divergent, la correspondance entre elles est faible, ce qui traduit un défaut de fidélité du résumé par rapport au document d'origine.

Les techniques d'inférence en langage naturel

Un autre type de métrique, également centré sur la fidélité au texte source, repose sur les techniques d'inférence en langage naturel (*Natural Language Inference*, NLI[142]). L'objectif est de déterminer si les informations produites dans le résumé, considérées comme des propositions ou « *claims* », sont corroborées par le document d'origine.

Dans ce cadre, chaque énoncé du résumé est évalué au regard du texte source. Le modèle chargé de cette tâche rend une décision binaire : soit le contenu est conforme à la source, soit il ne l'est pas. Dans l'exemple évoqué, le modèle conclurait que l'énoncé n'est pas conforme.

Ce type de modèle repose sur un apprentissage préalable réalisé à partir de données annotées dans lesquelles de nombreuses variations et transformations ont été introduites. Cet entraînement lui permet d'apprendre à distinguer les cas dans lesquels une information est corroborée par un texte de ceux dans lesquels elle ne l'est pas. Une fois entraîné, le modèle peut être réutilisé pour évaluer automatiquement des résumés, en produisant systématiquement une décision de type oui/non sur le caractère conforme ou non des informations générées.

Cette approche constitue ainsi une deuxième grande catégorie de métriques centrées sur la vérification de la validité des contenus produits par rapport au texte source

Utiliser un modèle comme évaluateur : « *LLM as a judge* »

Une évolution récente des méthodes d'évaluation repose sur l'utilisation directe des modèles de langage eux-mêmes comme outils d'évaluation, selon une approche désormais désignée sous le terme de « *LLM as a judge* » (« modèle comme évaluateur »). Cette pratique tend à se généraliser et vient s'ajouter aux métriques plus classiques. En conséquence, les évaluations se présentent souvent sous la forme de tableaux combinant un grand nombre d'indicateurs, chacun produisant ses propres scores.

Une évolution récente des méthodes d'évaluation repose sur l'utilisation directe des modèles de langage eux-mêmes comme outils d'évaluation.

Le principe est de mobiliser un modèle distinct, qualifié de modèle juge, auquel sont fournis le document source ainsi que le résumé à évaluer. On lui demande ensuite de produire des scores sur plusieurs dimensions, telles que la factualité, la fluidité ou la cohérence, et généralement sur des échelles numériques prédéfinies, par exemple de 0 à 5. Cette approche peut sembler circulaire, dans la mesure où un modèle évalue la production d'un autre modèle. Toutefois, les travaux empiriques montrent que les scores ainsi obtenus présentent une corrélation significative avec les jugements humains, ce qui justifie leur usage croissant.

Dans certaines implémentations, l'évaluation est structurée de manière plus élaborée. Le modèle juge peut être invité à expliciter son raisonnement avant de produire une note, dans une logique de chaîne de raisonnements. Des travaux, notamment autour de ce que l'on appelle le « GPTScore »[143], consistent ainsi à demander au modèle de détailler les étapes de son analyse avant de fournir une évaluation. Le score final peut ensuite être calculé à partir des probabilités associées aux différentes réponses générées, par exemple sous la forme d'une moyenne.

Cette approche soulève néanmoins des difficultés, en particulier en matière d'explicabilité. Si certains travaux tentent de contraindre le modèle à justifier les scores qu'il attribue, ces justifications restent imparfaites et difficiles à interpréter de manière rigoureuse. Le caractère largement opaque du processus décisionnel

142. Lili Mou, Rui Men, Ge Li *et al.*, « Natural Language Inference by Tree-Based Convolution and Heuristic Matching », *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, volume 2 : Short Papers, 2016, p. 130-136, https ://doi.org/10.18653/v1/P16-2022.

143. Fu, Jinlan, See-Kiong Ng, Zhengbao Jiang, et Pengfei Liu, « GPTScore: Evaluate as You Desire », *Proceedings of the 2024 Conference of the North American. Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1: Long Papers, 2024, p. 6556-6576, https ://doi.org/10.18653/v1/2024.naacl-long.365.

limite la compréhension fine des raisons qui conduisent à une évaluation donnée.

Parallèlement, d'autres pistes de recherche visent à rendre l'évaluation plus transparente. Certaines méthodes proposent de décomposer le document source et le résumé en unités élémentaires, sous forme d'énoncés simples. L'évaluation consiste alors à identifier, parmi ces énoncés, ceux qui sont corrects, incorrects ou absents, et à expliciter la manière dont ces éléments contribuent au score final. Cette granularité permet de mieux comprendre l'origine des erreurs et d'affiner l'analyse de la qualité du résumé.

De manière générale, aucun indicateur ne permet à lui seul de rendre compte de l'ensemble des dimensions pertinentes. Il est donc nécessaire de mobiliser plusieurs métriques de manière complémentaire. Celles-ci permettent notamment d'identifier les résumés manifestement défaillants, mais elles ne suffisent pas à garantir la production d'un résumé optimal. En définitive, l'évaluation humaine demeure indispensable, en particulier pour valider les résultats dans des contextes exigeants.

« Aucun indicateur ne permet à lui seul de rendre compte de l'ensemble des dimensions pertinentes. Il est donc nécessaire de mobiliser plusieurs métriques de manière complémentaire.

La difficulté spécifique de mesurer les éléments passés sous silence

Le résumé implique nécessairement une réduction de l'information. Mais cela soulève une difficulté majeure : en cherchant à être plus concis, ne risque-t-on pas de supprimer des éléments essentiels ? Dans certains contextes – par exemple, en matière judiciaire –, un élément apparemment secondaire (un témoignage isolé, un détail factuel précis) peut avoir une importance décisive. S'il disparaît du résumé, il peut ne jamais être reconsidéré, ce qui pose un problème critique d'usage.

Cette problématique est bien connue en recherche et elle est en partie abordée à travers la notion de rappel (*recall*). Le rappel mesure la capacité du résumé à conserver les informations importantes présentes dans le texte source.

Par exemple, dans des métriques classiques comme ROUGE**,** on compare les éléments (mots, segments) du résumé généré avec ceux d'un résumé de référence. Si certains éléments présents dans la référence sont absents du résumé produit, alors le rappel est faible. Ces métriques permettent donc, dans une certaine mesure, de détecter des omissions.

Mais il y a une limite fondamentale : comment identifier ce qui est « important » ? Le point critique se situe ailleurs. Ce qui compte réellement n'est pas seulement ce qui est omis, mais ce qui n'aurait pas dû être omis. Cette dimension correspond à ce que l'on appelle généralement la pertinence : a-t-on retenu les bonnes informations ? a-t-on éliminé de bonnes informations ?

Il s'agit de la dimension la plus importante mais aussi la plus difficile à évaluer et à modéliser. Contrairement à d'autres critères comme la fluidité (qualité linguistique) ou la fidélité locale (correspondance avec le texte), la pertinence dépend fortement du contexte, varie selon le domaine et repose souvent sur une interprétation experte.

Dans les évaluations humaines, c'est d'ailleurs le critère le plus exigeant, celui pour lequel les modèles sont les moins performants et pour lequel la corrélation avec le jugement humain est la plus faible.

3. L'enjeu majeur de la disponibilité des données d'entraînement

La disponibilité des données constitue un enjeu central, en particulier lorsqu'il s'agit de disposer de données supervisées. Dans le cas du résumé automatique, il ne suffit pas d'avoir accès aux documents : il est également nécessaire de disposer des résumés correspondants pour entraîner les modèles à produire ce type de réponse. Or ces résumés représentent un coût élevé puisqu'ils doivent être produits manuellement. Ils constituent précisément ce que l'on désigne comme la supervision, c'est-à-dire les exemples de réponses attendues permettant d'orienter l'apprentissage du modèle.

La difficulté de la tâche joue un rôle déterminant dans les besoins en données. Certaines tâches peuvent être considérées comme relativement simples : elles sont exécutées rapidement par un humain, reposent sur des règles

explicites et mobilisent principalement des transformations lexicales. Dans ce cas, les systèmes peuvent apprendre efficacement avec des volumes de données plus limités.

À l'inverse, lorsque la tâche est complexe, qu'elle mobilise des jugements subjectifs et que les règles qui la gouvernent sont difficiles à formaliser, les besoins en supervision augmentent fortement. Le résumé automatique s'inscrit largement dans cette seconde catégorie. Plus la tâche est difficile à expliciter et à stabiliser, plus il devient nécessaire de disposer de grandes quantités de données annotées pour guider l'apprentissage des modèles.

4. Les contraintes spécifiques au traitement des données sensibles

La question de la sensibilité des données constitue une autre contrainte majeure, qui ne concerne plus seulement les réponses attendues, mais les données d'entrée elles-mêmes, dans les phases d'apprentissage comme dans les phases de génération. Dans certains domaines, les documents utilisés pour entraîner ou interroger les modèles peuvent contenir des informations personnelles, confidentielles ou sensibles. Or l'usage des modèles de langage soulève ici des enjeux spécifiques, dans la mesure où ceux-ci sont généralement déployés sur des infrastructures distantes. Les données sont alors transmises à des serveurs externes, qui peuvent non seulement les traiter, mais aussi les stocker, avec des risques de réutilisation ou de fuite.

La sensibilité des données ne se réduit toutefois pas à une opposition binaire entre données sensibles et non sensibles. Elle s'inscrit plutôt dans un continuum, qui appelle des réponses techniques différenciées. À un premier niveau, on trouve des documents sans contrainte particulière. Viennent ensuite les données à caractère personnel, puis des informations confidentielles, par exemple liées à des projets en cours. À un niveau encore plus contraint se situent des données particulièrement sensibles, comme les données médicales, dont la circulation est strictement encadrée.

Dans ce dernier cas, les contraintes sont telles que les données ne peuvent pas être extraites de leur environnement d'origine. Le traitement doit alors être réalisé localement, au sein même des infrastructures dans lesquelles les données sont conservées, par exemple au sein même d'un établissement hospitalier. Cela implique de déplacer les modèles ou les capacités de calcul vers les données, et non l'inverse. Des dispositifs spécifiques ont été mis en place pour faciliter ce type d'accès, mais ils restent fortement encadrés.

À chaque niveau de sensibilité correspondent ainsi des choix techniques distincts. Lorsque les contraintes de protection des données sont fortes, l'exécution locale des modèles devient nécessaire afin de garantir la confidentialité des informations. Cette contrainte illustre une limite importante du déploiement standard des modèles de langage et souligne l'importance d'adapter les architectures et les pratiques aux exigences juridiques et opérationnelles des domaines concernés.

> « Lorsque les contraintes de protection des données sont fortes, l'exécution locale des modèles devient nécessaire afin de garantir la confidentialité des informations.

Ainsi le déploiement des modèles d'intelligence artificielle est étroitement lié à celle de la sensibilité des données et des garanties de contrôle que l'on souhaite conserver. Il n'existe pas de solution unique, mais plutôt un éventail de configurations, qui correspondent à des niveaux de contrainte et de maîtrise différents.

À un premier niveau, il est possible de faire fonctionner des modèles localement. Certains modèles, plus compacts, peuvent être exécutés sur des infrastructures internes. Cela suppose néanmoins de disposer de ressources de calcul suffisantes, mais cette option permet de conserver un contrôle complet sur les données, puisqu'elles ne quittent pas l'environnement local.

Lorsque les contraintes sont légèrement relâchées, une autre option consiste à déployer des modèles dans un cadre institutionnel maîtrisé. Dans ce cas, l'organisation contrôle non seulement l'infrastructure matérielle et le réseau, mais également les caractéristiques du modèle lui-même : ses poids, les données sur lesquelles il a été entraîné et donc, dans une certaine mesure, sa « ligne éditoriale ». Le modèle n'est pas nécessairement exécuté en

local strict, mais il reste dans un périmètre fortement sécurisé et contrôlé.

Un niveau intermédiaire consiste à recourir à des modèles proposés par des acteurs externes, tels que Mistral AI. Dans cette configuration, l'organisation ne maîtrise pas entièrement les données d'entraînement ni les choix de conception du modèle, mais peut néanmoins en contrôler le déploiement. Ces modèles peuvent être installés sur des infrastructures internes, ce qui permet de conserver la maîtrise du matériel, du réseau et des conditions de sécurité, tout en bénéficiant de solutions développées par des tiers.

Au-delà, on trouve les modèles commerciaux accessibles sous licence. La protection repose toutefois sur un cadre juridique et contractuel, et non sur un contrôle technique direct. Elle suppose donc un niveau de confiance dans le fournisseur et dans les engagements associés à la licence.

Enfin, à l'extrémité de ce spectre, se situent les outils librement accessibles, souvent gratuits, dont les conditions d'utilisation peuvent évoluer rapidement. Dans ce cas, les garanties offertes sont les plus faibles, tant en termes de stabilité contractuelle que de protection des données. Cette dernière catégorie suscite logiquement davantage de méfiance, en particulier lorsque les données manipulées présentent un caractère sensible.

L'ensemble de ces configurations illustre un continuum entre maîtrise et externalisation. Le choix d'une solution de déploiement dépend directement du niveau de sensibilité des données traitées et du degré de contrôle que l'on souhaite conserver sur l'infrastructure, le modèle et les conditions d'utilisation.

IV - L'automatisation des résumés dans le domaine juridique

Comme on l'a vu précédemment, l'automatisation performante de la production des multiples formes de résumé dans le domaine du droit et des professions judiciaires implique une adaptation au domaine. Elle est nécessaire parce que la langue du droit possède des spécificités, que les documents présentent des structures spécifiques, mais aussi que le raisonnement juridique et judicaire emporte des règles de sélection des informations pertinentes qui lui sont propres.

1. Utiliser des données juridiques pour l'entraînement

Pour les langues spécialisées, l'adaptation à un domaine passe en grande partie par une adaptation du vocabulaire. Les modèles ne travaillent pas directement au niveau des mots, mais au niveau de *tokens* de groupes de lettres, souvent proches de radicaux ou de fragments fréquents dans la langue.

Ce mécanisme fonctionne globalement très bien. Mais il montre ses limites dès qu'on entre dans des domaines spécialisés : certains mots rares ou techniques sont mal découpés, parce qu'ils apparaissent peu dans les données d'entraînement, et ne bénéficient donc pas de statistiques fiables. Dans ces cas-là, le modèle segmente ces mots en de nombreux petits fragments, ce qui peut dégrader la compréhension.

Cependant, même si le découpage est imparfait, les modèles peuvent compenser par le phénomène d'agrégation. Ils reconstruisent une représentation globale à partir des *tokens*, et peuvent ainsi récupérer une partie de l'information perdue au moment du découpage.

Se pose alors une importante question pratique : faut-il redécouper le vocabulaire pour un domaine spécialisé ou au contraire conserver le vocabulaire général et seulement adapter le modèle ?

Les conclusions ont évolué avec le temps. Aujourd'hui, sans qu'il y ait encore de consensus, la tendance est plutôt de ne pas redéfinir le vocabulaire, afin de conserver les connaissances générales apprises sur de grands corpus, et de travailler plutôt au niveau de l'agrégation.

Comme évoqué auparavant, le raffinement des modèles repose sur de nombreuses stratégies possibles, qui se distinguent par leur coût et leur efficacité plus ou moins grande. Il n'existe pas de méthode unique et il est nécessaire de tester, comparer, ajuster.

2. Faire face au risque de restitution de fragment des données d'entraînement au moment de l'inférence

L'institution judiciaire dispose de grandes quantités de données au format numérique : des procédures, des décisions, des synthèses déjà produites. Au-delà des données relatives au droit lui-même, par nature publiques et impersonnelles, de nombreuses applications potentielles des résumés automatiques pour les professions juridiques concernent des documents contenant

des données sensibles, entendues comme des données soumises à des régimes de protection. La proportion de ces données est parfois très importante, et elles sont d'une grande diversité, comme un numéro de police d'assurance, des coordonnées bancaires, un profil ADN, des coordonnées GPS ou un numéro IMEI (identité internationale d'équipement mobile). Or, parce que le travail des professions juridiques s'articule autour de l'application de la loi générale à des cas particuliers, ces données participent de l'analyse et de la compréhension des faits d'une situation

Comme nous l'avons vu, les performances de synthèse des LLM peuvent être améliorées par un entraînement à partir de données représentatives du champ de connaissance et des attentes des utilisateurs pour différents types de résumé. Pour obtenir des modèles performants pour les résumés de dossiers judiciaires, intégrant en particulier le fond des dossiers et les faits, il serait indéniablement intéressant de constituer des corpus d'entraînement avec des données de dossiers véritables.

La question se pose : quels sont les risques à entraîner des modèles avec ce type de contenus, notamment lorsqu'ils incluent des données personnelles comme c'est le cas de tous les dossiers judiciaires ? Ces risques s'associent à deux phénomènes complémentaires : la mémorisation par le modèle de fragments de ses données d'entraînement, d'une part, et les conditions dans lesquelles le modèle pourrait se livrer à leur restitution lors de l'inférence, d'autre part.

Dans ce champ de recherche très actif, les travaux de Nicholas Carlini ont montré que le phénomène de mémorisation n'était pas théorique, même si la restitution passait par la mobilisation de techniques relevant plus de la cyberattaque que de l'usage courant des LLM[144] [145]. Il montre que le risque est d'autant plus important qu'une information est répétée, et qu'il s'accroît avec la taille des modèles. Des expériences ont été menées pour tester précisément ce risque avec des corpus de données personnelles synthétiques[146].

> Le phénomène de mémorisation n'était pas théorique, même si la restitution passait par la mobilisation de techniques relevant plus de la cyberattaque que de l'usage courant des LLM.

Un premier constat important a été tiré de ces expériences : la répétition est déterminante. Ainsi il est en pratique étonnamment difficile de faire restituer des données personnelles. Pour que cela se produise, il faut que les données apparaissent très fréquemment dans le corpus ou qu'elles soient issues de sources très fortement pondérées.

Il faut cependant distinguer deux situations : les contenus fréquents ou fortement présents dans les données plus susceptibles d'être reproduits et les données personnelles isolées beaucoup plus difficiles à extraire. Cela fait que le risque est réel mais moins immédiat qu'on pourrait le craindre dans des usages standards.

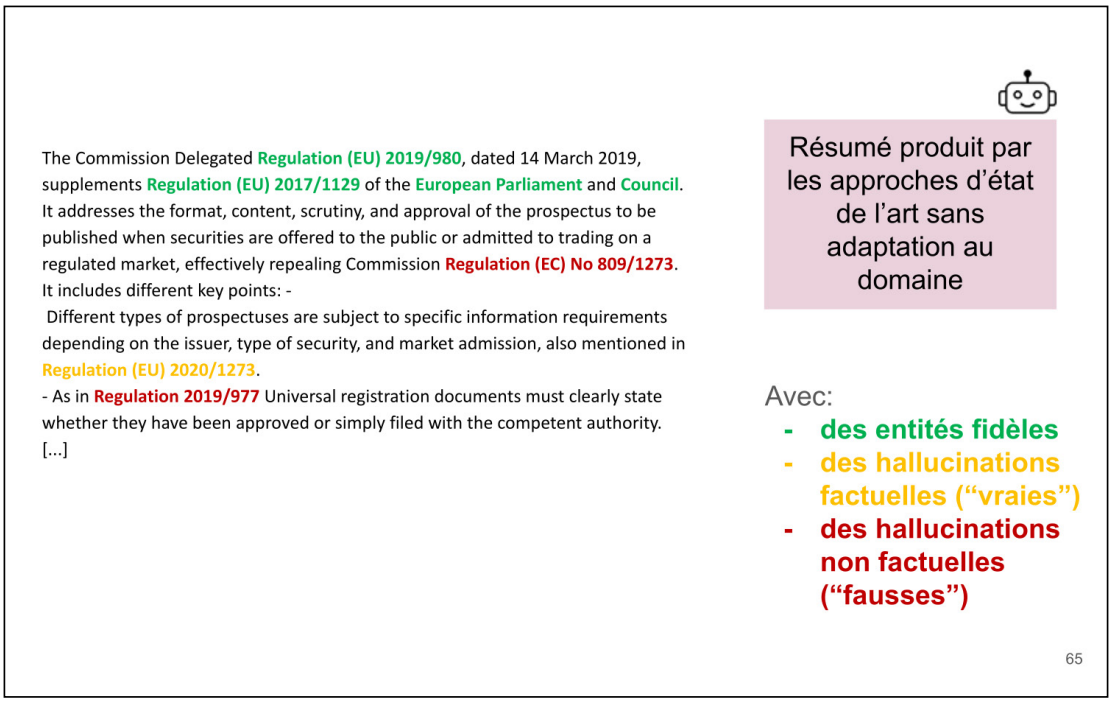
Mais ce risque existe également lors des phases de raffinement (*fine-tuning*) et, de manière paradoxale, il peut être même plus important à ce stade. Pourquoi ? Parce que le raffinage intervient en fin d'entraînement, souvent avec des jeux de données plus petits et plus ciblés, ce qui augmente la probabilité que certaines informations soient mémorisées de manière plus directe. Autrement dit, ce qui est injecté tard dans le modèle peut avoir un impact disproportionné.

À ce jour, il n'existe pas de garantie forte contre ce risque, ni de solution totalement satisfaisante. Le problème reste donc identifié, étudié, mais non complètement maîtrisé.

Pour contourner les difficultés d'accès aux données, notamment celles liées au respect de la vie privée, une des évolutions importantes

144. Nicholas Carlini, Florian Tramèr, Eric Wallace *et al.*, « Extracting Training Data from Large Language Models ». USENIX Security Symposium, 2020, https ://export.arxiv.org/pdf/2012.07805v2.
145. Pour une recherche sur la reconstitution de textes de romans, Ahmed Ahmed, A. Feder Cooper, Sanmi Koyejo, et Percy Liang, « Extracting books from production language models », arXiv :2601.02671. Prépublication, arXiv, 2026. https ://doi.org/10.48550/arXiv.2601.02671.

146. Sriram Selvam et Anneswa Ghosh, « PANORAMA: A synthetic PII-laced dataset for studying sensitive data memorization in LLMs », arXiv :2505.12238. Prépublication, *arXiv*, 2025. https ://doi.org/10.48550/arXiv.2505.12238.



Les ajouts du modèle peuvent être ou non factuellement exacts
Source : Karen Pinel-Sauvagnat et Vincent Guigue

actuellement est d'explorer le recours à la génération de données synthétiques pour nourrir les corpus d'entraînement.

> Pour contourner les difficultés d'accès aux données, notamment celles liées au respect de la vie privée, une des évolutions importantes actuellement est d'explorer le recours à la génération de données synthétiques pour nourrir les corpus d'entraînement.

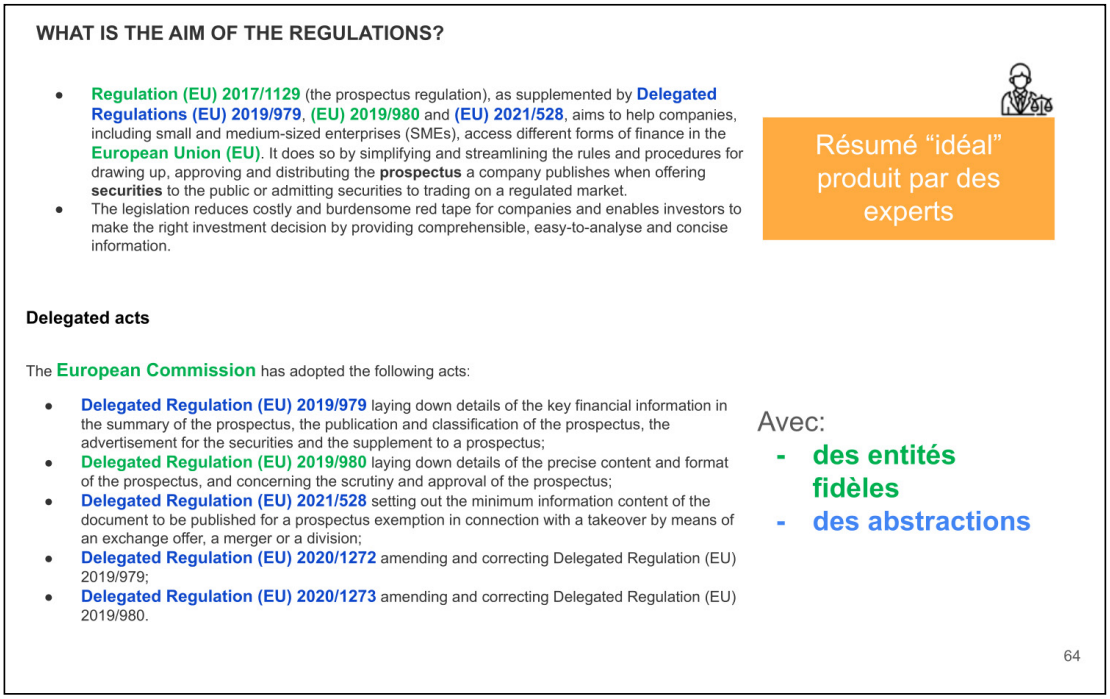
Historiquement, l'idée de générer des données synthétiques, pour entraîner des modèles ne fonctionnait pas. C'était globalement voué à l'échec. De premiers résultats probants ont finis par être obtenus. Un exemple souvent cité est celui d'AlphaZero initialement entraîné avec des parties d'échecs, puis uniquement à partir des règles du jeu, en laissant le système générer lui-même ses propres données. Cela a très bien fonctionné, mais dans un cadre très structuré qui est celui du jeu d'échecs.

Dans le domaine des images, on utilise depuis longtemps la *data augmentation*, un mécanisme permettant d'augmenter artificiellement le jeu de données en créant de nouvelles entrées à partir de données existantes (rotation, bruit, etc.). Mais cela suppose toujours de disposer de données réelles de départ. On ne part pas uniquement de données synthétiques.

Pour le texte, la situation est différente. On en est encore aux balbutiements. Les approches restent limitées. Par exemple, dans certains travaux sur le résumé, les chercheurs génèrent des variantes de données ou créent des exemples contrastifs (en modifiant le texte source) pour entraîner les modèles à ne pas se tromper. Mais cela relève davantage de la *data augmentation* que de la véritable donnée synthétique.

On observe néanmoins une dynamique dans plusieurs domaines : en séries temporelles, par exemple, on commence à entraîner des modèles sur des signaux synthétiques. Cela suggère que la tendance est générale, même si elle n'est pas encore aboutie pour le texte.

Le principal problème est celui des biais. Si un modèle génère des données qui servent ensuite à entraîner un autre modèle, qui est lui-même évalué par un système similaire, alors on crée une boucle où les biais se reproduisent et s'amplifient.



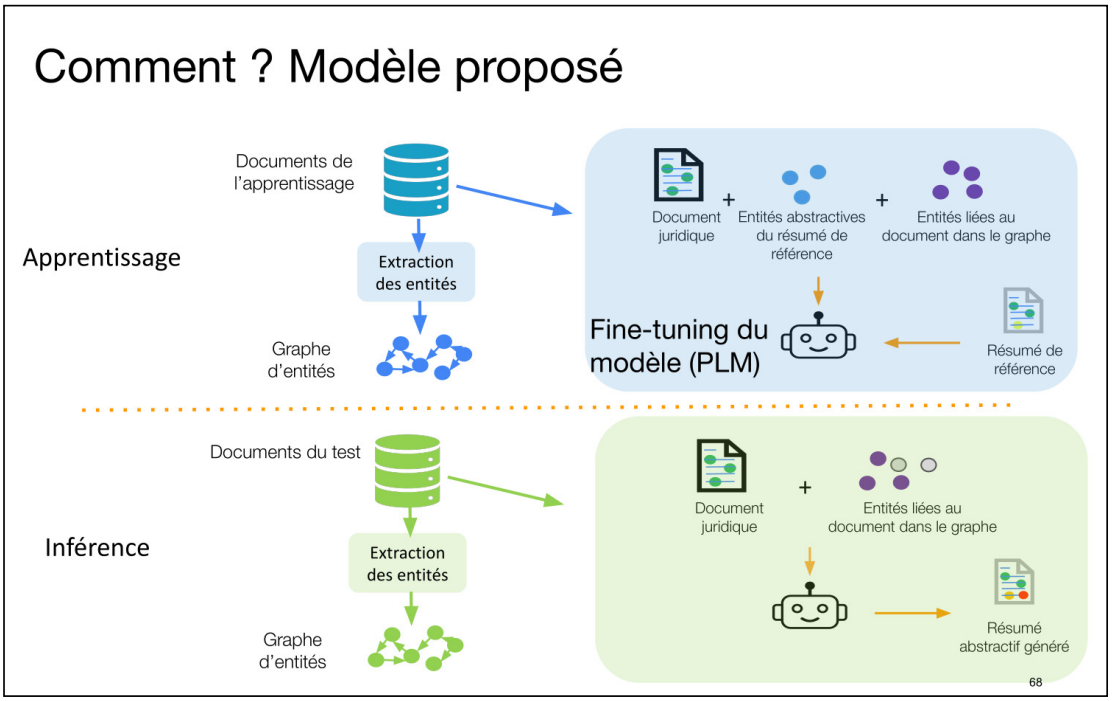
EUR-Lex-Sum : un résumé de référence contient des éléments non présents dans le document d'origine.
Source : document fourni par Karen Pinel-Sauvagnat et Vincent Guigue

3. Cas d'application : le résumé de la législation de l'Union européenne pour un éditeur juridique

Pour donner une dimension concrète à l'ensemble des considérations précédentes, les intervenants ont présenté un travail de thèse visant à mobiliser les approches génératives pour produire des résumés de textes législatifs de l'Union européenne pour un éditeur juridique[147]. L'activité de production de résumés était préexistante et faisait partie des services de l'éditeur. L'ensemble des résumés déjà produits a donc pu constituer une base d'entraînement. Il a été complété par le corpus Eur-Lex-Sum[148]. Ils ont été réalisés dans le cadre d'une thèse de Nihed Bendaham[149]. Elle a été menée dans le cadre d'un contrat CIFRE au sein de la société Berger-Levrault[150], confrontée à des besoins opérationnels de résumés de documents juridiques, notamment pour faciliter le travail de veille.

L'approche s'appuie sur des corpus existants, en particulier le corpus Eur-Lex-Sum[151]. L'analyse des résumés de ce type de données produits par des experts met en évidence une caractéristique importante : la forte présence d'entités juridiques. Plus précisément, ces entités se répartissent en deux catégories. Il s'agit, d'une part, des entités fidèles, directement issues du document source et, d'autre part, des éléments ajoutés par les experts qualifiés d'« abstractions », qui permettent de contextualiser ou d'enrichir le résumé. Ces ajouts ne sont pas marginaux : ils représentent une proportion significative des

147. Nihed Bendahman, Karen Pinel-Sauvagnat, Gilles Hubert, et Mokhtar Boumedyen Billami, « Not All Hallucinations Are Good to Throw Away When It Comes to Legal Abstractive Summarization », *Proceedings of the 2025 Conference of the Nations of the Americas. Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1 : Long Papers, 2025, p. 5331-44. https ://doi.org/10.18653/v1/2025.naacl-long.275.
148. Dennis Aumiller, Ashish Chouhan, and Michael Gertz, « Eur-lex-sum: A multi-and cross-lingual dataset for long-form summarization in the legal domain », *in Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, p. 7626-7639.
149. Nihed Bendahman. *Évaluation et réduction des hallucinations dans la génération automatique de résumés dans le contexte spécifique des documents juridiques*. Intelligence artificielle [cs.AI]. Université de Toulouse, 2025. Français. ⟨NNT : 2025TLSES196⟩. ⟨tel-05532695⟩
150. La société française Berger-Levrault exerce une activité d'édition de logiciels principalement à destination des acteurs publics. Elle possède des activités d'édition juridique, en particulier avec les services *Légibase Justice* et *Légibase Collectivités* (https ://www.berger-levrault.com/fr/).
151. Dennis Aumiller, *op.cit.*



Le corpus d'apprentissage est également utilisé comme source ce construction d'une représentation du corpus sous forme de graphe de connaissances
Source : travaux de Nihed Bendahman

contenus, ce qu'on observe à la fois sur Eur-Lex et pour d'autres corpus juridiques en français.

La plateforme EUR-Lex-Sum offre un matériau précieux pour la recherche : elle met à disposition à la fois les textes de loi originaux et les résumés rédigés par des experts. Ce type de corpus permet de comparer directement la production des modèles à une référence humaine de haute qualité.

L'analyse de ces résumés rédigés par des experts met en évidence plusieurs caractéristiques structurantes. En premier lieu, l'utilisation d'un vocabulaire hautement spécialisé pose des défis classiques de compréhension et de reformulation pour les modèles.

Mais surtout, et plus subtilement, les résumés ne se limitent pas à condenser l'information du texte source. Ils intègrent souvent une contextualisation externe en mobilisant des éléments absents du document initial.

Concrètement, un résumé peut faire référence à d'autres textes de loi ou cadres normatifs qui ne sont pas explicitement mentionnés dans le document d'origine. Cette pratique correspond à une démarche experte : elle vise à replacer le texte dans son écosystème juridique.

Ce constat introduit une difficulté majeure pour le résumé automatique. Les modèles abstractifs sont traditionnellement conçus pour produire une synthèse fidèle au contenu d'entrée. Or, dans ce cas, un résumé pertinent nécessite d'introduire des informations complémentaires, issues d'un savoir externe.

> « Dans ce cas, un résumé pertinent nécessite d'introduire des informations complémentaires, issues d'un savoir externe.

Cela soulève une difficulté. D'un côté, la génération de contenu non présent dans la source est généralement considérée comme un risque – celui des hallucinations. De l'autre, dans certains contextes comme le droit, cette capacité peut être nécessaire et même attendue, à condition qu'elle soit maîtrisée.

Les travaux de Nihed Bendahman s'inscrivent précisément dans cette problématique : il s'agit d'identifier et de reconstruire ces

informations contextuelles absentes du document source, afin d'approcher le comportement des experts humains.

Cette observation fait toucher du doigt une difficulté majeure pour les approches classiques de résumé automatique. Les modèles actuels produisent à la fois des entités fidèles et des hallucinations, elles-mêmes de deux types : factuelles, lorsqu'elles ajoutent des informations correctes mais absentes du texte source, et non factuelles, lorsqu'elles introduisent des informations erronées. Or les abstractions produites par les experts – bien qu'extérieures au document – correspondent, du point de vue du modèle, à des hallucinations. Dès lors, les approches visant uniquement à réduire les hallucinations ne permettent pas de reproduire le comportement expert.

L'analyse quantitative confirme ce point : environ deux tiers des entités présentes dans les résumés de référence relèvent de ces abstractions. L'objectif ne peut donc pas être de supprimer toute forme d'ajout, mais de distinguer entre enrichissements pertinents et erreurs. Le problème est alors reformulé en termes d'alignement : il s'agit de faire correspondre, d'une part, les entités fidèles du résumé généré avec celles du résumé de référence et, d'autre part, les abstractions expertes avec des hallucinations factuelles produites par le modèle tout en éliminant les hallucinations non factuelles.

Pour répondre à cette difficulté, une méthode de raffinement a été proposée. Elle consiste à exploiter la structure des entités présentes dans les données d'apprentissage. À partir du corpus, un graphe d'entités est construit, reliant différents types d'informations : documents, thématiques juridiques, organisations, localisations, mots-clés. Ce graphe représente une forme de connaissance structurée du corpus. Lors de l'apprentissage, le modèle reçoit en entrée à la fois le document, les entités issues du graphe et les abstractions présentes dans les résumés de référence, et il apprend à produire ces derniers. À l'inférence, il peut alors mobiliser ce graphe pour sélectionner des entités pertinentes à intégrer dans le résumé, y compris au-delà du contenu strict du document.

Les expérimentations montrent que cette approche permet d'améliorer les performances selon plusieurs métriques, notamment en augmentant les hallucinations factuelles – c'est-à-dire les ajouts pertinents – et en réduisant les hallucinations non factuelles. Les gains ne sont pas parfaits, mais ils indiquent une amélioration par rapport aux approches de l'état de l'art.

> Une méthode de raffinement a été proposée. Elle consiste à exploiter la structure des entités présentes dans les données d'apprentissage.

Enfin, des expérimentations préliminaires ont été menées avec des modèles de grande taille, en leur fournissant directement ces informations (entités ou triplets) sous forme de consignes. Les résultats ne sont toutefois pas concluants à ce stade. Dans les conditions testées, un modèle plus petit, correctement affiné, obtient de meilleurs résultats qu'un modèle de grande taille utilisé sans adaptation spécifique. Plusieurs explications sont envisagées, notamment la sensibilité aux instructions (*prompting*) ou la difficulté pour ces modèles d'exploiter efficacement des structures externes sans entraînement dédié.

Conclusion

L'un des points saillants de la discussion tient à un paradoxe désormais bien identifié : les outils de traitement automatique du langage atteignent aujourd'hui un niveau de performance spectaculaire, parfois déroutant. Cette efficacité apparente constitue précisément leur principal piège.

La tentation est forte, face à des résultats jugés convaincants, d'envisager une intégration immédiate dans les pratiques professionnelles. Une telle démarche est pourtant problématique. L'adoption de ces systèmes ne peut être guidée par leur seule performance perçue ; elle doit reposer sur une analyse préalable des besoins. En d'autres termes, il ne s'agit pas d'adapter les usages aux capacités de l'outil, mais bien de configurer l'outil en fonction d'objectifs clairement définis.

Cette inversion de perspective est essentielle. Un système de résumé automatique, aussi performant soit-il, n'a de pertinence que s'il répond à un besoin effectif. L'identification de ce besoin relève du terrain, des pratiques réelles, et non d'une fascination technologique. Les modèles doivent ensuite être ajustés pour répondre à ces exigences, et non l'inverse.

Au-delà des contraintes techniques, les enjeux liés à l'usage sont déterminants. L'automatisation de tâches cognitives, notamment dans des domaines sensibles comme la médecine ou le droit, introduit des risques bien documentés.

Parmi ceux-ci figure la tendance à accorder une confiance excessive aux systèmes automatisés. Cette surestimation de leur fiabilité repose sur un mécanisme simple mais puissant : lorsqu'un outil produit des résultats corrects dans une large majorité des cas – par exemple, neuf fois sur dix – il devient difficile pour l'utilisateur d'anticiper ses erreurs.

> Lorsqu'un outil produit des résultats corrects dans une large majorité des cas – par exemple, neuf fois sur dix – il devient difficile pour l'utilisateur d'anticiper ses erreurs.

Ce phénomène, souvent désigné comme un biais d'automatisation, affecte différemment les publics. Les utilisateurs non spécialisés et les professionnels ne commettent pas les mêmes erreurs, mais restent exposés à une difficulté commune : maintenir une vigilance critique face à des résultats généralement pertinents.

La performance moyenne élevée masque ainsi la variabilité des résultats. C'est précisément cette asymétrie – entre fréquence de réussite et impact potentiel de l'erreur – qui rend ces systèmes à la fois puissants et dangereux.

En définitive, les systèmes de résumé automatique ne doivent pas être appréhendés comme des solutions génériques, mais comme des outils spécialisés, dont l'efficacité dépend étroitement de leur contexte d'usage. Leur déploiement suppose une définition rigoureuse des besoins, une évaluation réaliste de leurs performances et une prise en compte explicite des risques d'usage.

L'enjeu n'est donc pas uniquement technologique. Il est aussi méthodologique et épistémologique : il s'agit de comprendre ce que ces outils font réellement, dans quelles conditions ils le font et à quelles limites ils se heurtent.

> Les systèmes de résumé automatique doivent être appréhendés comme des outils spécialisés, dont l'efficacité dépend étroitement de leur contexte d'usage. Leur déploiement suppose une définition rigoureuse des besoins, une évaluation réaliste de leurs performances et une prise en compte explicite des risques d'usage.

C'est à cette condition que leur potentiel pourra être exploité sans céder aux illusions qu'ils suscitent.

Conclusion du cycle

L'Institut Robert Badinter tient à remercier nos intervenantes et intervenants, grâce auxquels ce cycle d'ateliers a, nous l'espérons, tenu toutes ses promesses. Grâce à elles nous avons pu aborder quelques points fondamentaux concernant la mobilisation des technologies de l'intelligence artificielle et en discuter avec l'ensemble des participantes et participants, séance après séance. Nous avons bien perçu dans ces discussions l'importance de repartir des besoins spécifiques du terrain. C'est à cette condition que l'expertise scientifique peut être mobilisée efficacement, afin de concevoir et de déployer des outils adaptés.

Un point commun de l'ensemble des ateliers est le constat que les avancées reposent sur une croissance exponentielle de la taille des modèles et des besoins en calculs, qui pose des défis majeurs. Le premier d'entre eux est environnemental, à raison de la consommation électrique et des besoins en ressources pour le refroidissement. Par ailleurs cette croissance complexifie l'accès à l'état de l'art pour les acteurs académiques eux-mêmes, pourtant dotés de ressources spécifiques. Le contrôle et la supervision des modèles deviennent également plus difficiles pour les régulateurs. Cette croissance implique des investissements matériels considérables, un renouvellement accéléré des équipements et questionne enfin la soutenabilité du modèle actuel de développement. Parvenir à une IA plus frugale apparaît comme absolument indispensable.

Un enseignement central se dégage des échanges ensuite : la question du rapport entre l'utilisateur et l'automate constitue un enjeu structurant pour le bon développement de l'IA. Elle renvoie à la place que chacun choisit d'occuper face aux outils, à la place qu'il entend donner à leurs productions, aux tâches qu'il décide de déléguer, ainsi qu'à celles qu'il entend conserver.

Ce positionnement relève en grande partie d'un arbitrage individuel. Il implique des choix propres à chaque utilisateur, en fonction de ses pratiques, de ses exigences et de sa perception des outils. Des choix en partie aussi contraints par un contexte de travail et de relations entre professionnels soumis à de grandes évolutions. Cette double dimension personnelle et organisationnelle ouvre un champ de réflexion particulièrement large, appelé à se développer dans les travaux à venir de l'institut.

Méthodologie et intervenants

Pour conduire ces ateliers exploratoires, selon les principes habituels de l'IRB, permettant de confronter les travaux et hypothèses de la recherche à l'expérience et aux questionnements des praticiens du droit et de la justice, l'Institut s'est adjoint le concours de Yannick MENECEUR, expert associé, magistrat, inspecteur de la justice et auteur de plusieurs ouvrages sur les usages judiciaires de l'intelligence artificielle et d'Olivier CHEVET, responsable d'études et de recherches, magistrat, par ailleurs ingénieur en génie logiciel et titulaire d'un DEA en informatique documentaire. Ils ont organisé les ateliers, les ont animés et ont rédigé le présent rapport. Conçu et réalisé sous la direction d'Harold Épineuse, directeur adjoint, ce cycle d'ateliers comme la rédaction de ces actes ont bénéficié de la collaboration de Mélanie VAY, responsable d'études et de recherches, docteure en science politique.

Chacun des ateliers a été introduit et conclu par Harold ÉPINEUSE. Une séquence initiale de brèves d'actualité était proposée par Mélanie VAY, Yannick MENECEUR et Olivier CHEVET, chacun d'eux présentant une brève en lien avec l'actualité de l'intelligence artificielle, sans forcément de lien direct avec le thème de l'atelier.

Les séances ont été organisées autour de l'intervention d'un ou plusieurs invités, sélectionnés pour leur pertinence sur les thèmes correspondants. Les échanges avec l'assistance, en présentiel et en distanciel, ont eu lieu tout au long des présentations.

Le 1er atelier de décryptage s'est tenu le vendredi 14 février 2025. Consacré aux agents, il a été l'occasion d'écouter les interventions de :

- **Yannick MENECEUR**, magistrat, expert associé auprès de l'IRB
- **Zacharie LAÏK**, avocat, fondateur et dirigeant de GoodLegal.fr.

Le 2ème atelier de décryptage s'est tenu le vendredi 21 mars 2025. Consacré aux usages professionnels de l'IA, il a été l'occasion d'écouter l'intervention de :

- **Aurélie KLEIN**, avocate experte Data & Digital – directrice de l'innovation digitale au cabinet FIDAL, maîtresse de conférences associée à la faculté de droit, de science politique et de gestion de l'université de Strasbourg.

Le 3ème atelier de décryptage s'est tenu le vendredi 11 avril 2025. Consacré aux questions juridiques pour la mise en œuvre d'outils d'intelligence artificielle, il a été l'occasion d'écouter les interventions de :

- **Marina TELLER**, professeure de droit privé à la faculté de droit et de science politique de l'université de Nice Côte d'Azur, directrice de la Chaire 3IA «Droit économique et intelligence artificielle », directrice du programme DL4T (Deep Law for Technologies)
- **Bertrand CASSAR**, directeur Valorisation stratégique des données au sein du groupe La Poste – docteur en droit et professeur associé à l'université Paris 1 Panthéon-Sorbonne.

Le 1er atelier d'approfondissement s'est tenu le jeudi 15 mai 2025. Consacré à la transcription automatique, il a été l'occasion d'écouter les interventions de :

- **Christophe SERVAN**, chercheur contractuel CNRS au sein de l'équipe SEME du Laboratoire interdisciplinaire des sciences du numérique (LISN) à Orsay, et désormais chercheur à l'Agence ministérielle de l'intelligence artificielle pour la défense (AMIAD)
- **Daniel CAMARA**, responsable technique du Centre forensique d'intelligence artificielle du pôle judiciaire de la Gendarmerie nationale, qui s'est exprimé en particulier à propos du premier outil de transcription automatique déployé de manière expérimentale au sein de la Gendarmerie nationale pour l'assistance à la rédaction de procès-verbaux d'audition
- **Max BELIGNÉ**, ingénieur de recherche auprès de la Plateforme universitaire de données Grenoble Alpes.

Le 2ème atelier d'approfondissement s'est tenu le jeudi 19 juin 2025. Consacré à la traduction automatique, il a été l'occasion d'écouter les interventions de :

- **François YVON**, professeur en informatique, chargé de mission IA & Traitement automatique des langues à l'ISIR (Institut des systèmes intelligents et de robotique), Sorbonne Université
- **Marie JONCA**, adjointe au chef du département de l'évaluation et des projets de modernisation au service de l'expertise et de la modernisation (SEM) du secrétariat général du ministère de la Justice.

Enfin, **le 3ème atelier d'approfondissement** s'est tenu le jeudi 17 mars 2026. Consacré à la synthèse d'écritures et au résumé, il a été l'occasion d'écouter les interventions de :

- **Karen PINEL-SAUVAGNAT**, maîtresse de conférences à l'université de Toulouse, HDR, membre du laboratoire Institut de recherche en informatique de Toulouse (IRIT), dont les sujets de recherche portent sur la recherche d'information, le résumé et la synthèse et qui est également très impliquée dans l'ARIA (Association de recherche d'information et application)
- **Vincent GUIGUE**, professeur à AgroParisTech, membre du laboratoire Modélisation Mathématique, informatique et Physique (MMIP), dont le domaine de recherche est le « *machine learning* ».

Table des matières

Directeur de la publication : Valérie SAGANT

Coordination : Harold ÉPINEUSE, Yannick MENECEUR et Olivier CHEVET

Rédaction : Olivier CHEVET, Yannick MENECEUR, avec la collaboration de Mélanie VAY

Remerciements : Zacharie LAÏK, Marina TELLER, Bertrand CASSAR, Aurélie KLEIN, Christophe SERVAN, Maxime BELIGNE, Daniel CAMARA, François YVON, Vincent GUIGUE et Karen PINEL-SAUVAGNAT

Suivi d'édition : Xavier CAYON et Léa DELION

Réalisation : Justine DE SPIEGELAERE

Secrétariat de rédaction : Olivier CHEVET et Mélanie VAY

Impression : CIN – Juin 2026

DÉCODER L'IA
DES PROMESSES DES OUTILS AUX RÉALITÉS DES USAGES

À la suite de la conférence inaugurale tenue à Strasbourg en janvier 2025, le cycle d'ateliers « Décoder l'IA » avait pour ambition de confronter les promesses de l'intelligence artificielle aux réalités complexes de cette technologie replacée dans son environnement technique, professionnel et institutionnel.

Autour d'invités issus de différents horizons dont plusieurs chercheurs en informatique, six séances ont nourri une réflexion à la fois pratique et interdisciplinaire. Après des ateliers consacrés à la notion d'agent, aux usages au sein des cabinets d'avocats et à l'encadrement juridique de l'IA, le cycle s'est approfondi autour de cas d'usage spécifiques : transcription, traduction et synthèse.

Ces actes mettent en lumière les exigences croissantes de contrôle humain, d'évaluation, de traçabilité et de fiabilité dans un secteur où l'erreur peut produire des effets majeurs. À rebours des discours simplificateurs, ils proposent une lecture critique et documentée de l'IA générative, entre promesses d'efficacité, dépendances techniques et défis de régulation.

Courrier postal
Institut Robert Badinter
Ministère de la Justice
13, place Vendôme
75042 Paris Cedex 01

Venir dans nos locaux
47 bis, rue des Vinaigriers
75010 Paris

contact@institutrobertbadinter.fr
institutrobertbadinter.fr

INSTITUT ROBERT BADINTER
Études et recherches sur le droit et la justice

L'Institut Robert Badinter (anciennement Institut des études et de la recherche sur le droit et la justice - IERDJ) a pour mission de nourrir la connaissance et les échanges sur le droit et la justice en lançant des réflexions originales et prospectives, en finançant et en accompagnant des travaux de recherche et en diffusant largement les connaissances sur les normes, la régulation et le fonctionnement de la justice, toutes disciplines scientifiques confondues.

Groupement d'intérêt public créé en 2022, l'Institut Robert Badinter est issu de la fusion de l'Institut des hautes études sur la justice et de la Mission de recherche droit et justice, deux entités internationalement reconnues.