EUROPEAN LAWYERS
EUROPEAN BARS
CCBE
BARREAUX EUROPÉENS
AVOCATS EUROPÉENS

Conseil des barreaux européens
La voix des avocats européens

Rue Joseph II, 40/8 - B-1000 Bruxelles
+32 (0)2 234 65 10 | ccbe@ccbe.eu | www.ccbe.eu

Guide technique sur l'utilisation des outils et modèles d'intelligence artificielle par les avocats

27 mars 2026

Sommaire

Introduction

Le présent guide vise à fournir aux avocats les connaissances techniques fondamentales leur permettant de sélectionner, d'évaluer et d'utiliser les outils d'intelligence artificielle disponibles dans le respect des lois applicables et des obligations professionnelles. À ce titre, il se concentre sur les fonctionnalités les plus courantes dans la pratique juridique et met en lumière les architectures sous-jacentes des différentes solutions d'intelligence artificielle. Un glossaire technique des termes clés accompagne ce guide à l'intention de son public juridique. Les termes expliqués dans le **glossaire** sont indiqués en **bleu** tout au long du texte. Le guide ne constitue pas une recommandation de produits ni un tutoriel étape par étape pour des produits spécifiques. Toutefois, étant donné que dans certains cas, il n'existe que quelques fournisseurs sur le marché[1], certaines références aux marques existantes sont inévitables.

Ce guide doit être lu conjointement avec le guide du CCBE sur l'utilisation de l'intelligence artificielle générative par les avocats (guide GenAI)[2] et les Lignes directrices du CCBE sur l'usage de l'informatique en nuage par les barreaux et les avocats[3]. Le guide GenAI avait pour objectif de présenter les principales caractéristiques et la définition de l'intelligence artificielle générative (GenAI) et d'exposer les avantages, les risques et les obligations professionnelles liés à l'utilisation de la GenAI par les avocats. Le guide sur l'informatique en nuage avait pour objectif de présenter les types de solutions disponibles et les obligations professionnelles liées à l'utilisation des services d'informatique en nuage.

Les avocats interagissent principalement avec les systèmes d'intelligence artificielle en tant qu'utilisateurs finaux et peuvent ne pas avoir une compréhension approfondie de leur fonctionnement technique interne. Toutefois, un avocat a l'obligation professionnelle de posséder une compréhension suffisante du fonctionnement essentiel des technologies qu'il utilise. Cette obligation est reflétée dans le principe de compétence professionnelle de la Charte des principes essentiels de l'avocat européen du CCBE (« Charte des principes essentiels »)[4]. Le commentaire de ce principe précise, entre autres : « *L'avocat doit être conscient des avantages et des risques relatifs à l'emploi des technologies dans sa pratique.* »

La compétence en matière de technologies est également une exigence de la législation de l'UE sur l'intelligence artificielle[5] (article 4), qui oblige les fournisseurs et les déployeurs de systèmes

[1] C'est le cas, par exemple, des fournisseurs de GPU, où NVIDIA occupe une position dominante sur le marché des GPU pour l'intelligence artificielle, avec une part de marché estimée entre 85 % et plus de 90 % pour les GPU optimisés pour l'intelligence artificielle.

[2] Guide du CCBE sur l'utilisation de l'intelligence artificielle générative par les avocats (2 octobre 2025) : https://www.ccbe.eu/fileadmin/speciality_distribution/public/documents/IT_LAW/ITL_Guides_recommendations/FR_ITL_20251002_CCBE-guide-on-the-use-of-the-use-of-generative-AI-for-lawyers.pdf

[3] Lignes directrices du CCBE sur l'usage de l'informatique en nuage par les barreaux et les avocats (27 février 2025) : https://www.ccbe.eu/fileadmin/speciality_distribution/public/documents/IT_LAW/ITL_Guides_recommendations/EN_ITL_20250227__CCBE-lignes directrices sur l'utilisation de l'informatique en nuage par les avocats.pdf

[4] Charte des principes essentiels de l'avocat européen et Code de déontologie des avocats européens : https://www.ccbe.eu/fileadmin/speciality_distribution/public/documents/DEONTOLOGY/DEON_CoC/FR_DEON_CoC.pdf. Voir également : partie 4.2 du guide du CCBE sur l'utilisation de l'intelligence artificielle générative par les avocats (voir note 2 ci-dessus).

[5] Règlement (UE) 2024/1689 du Parlement européen et du Conseil du 13 juin 2024 établissant des règles harmonisées concernant l'intelligence artificielle et modifiant les règlements (CE) n° 300/2008, (UE) n° 167/2013, (UE) n° 168/2013,

d'intelligence artificielle (y compris les avocats) à garantir « *un niveau suffisant de maîtrise de l'IA pour leur personnel et les autres personnes s'occupant du fonctionnement et de l'utilisation des systèmes d'IA pour leur compte [...].* »[6]

Il est essentiel que les avocats aient des connaissances de base sur le fonctionnement des outils d'intelligence artificielle, car cela leur permet d'identifier les solutions technologiques adaptées à leur pratique et d'évaluer les conséquences de ces choix sur le respect des obligations juridiques et professionnelles applicables. L'une de ces obligations est la protection de la confidentialité des communications entre l'avocat et son client. Les avocats sont les gardiens des données de leurs clients et ont l'obligation déontologique de protéger leurs clients et leurs données. Cette connaissance technologique permet aux avocats de faire la distinction entre les situations où ils ont le choix entre plusieurs outils, celles où il n'existe aucune solution viable et, surtout, celles où certaines technologies ne doivent pas être utilisées et pourquoi. Elle permet également aux avocats de reconnaître quand ils doivent consulter un spécialiste, tel qu'un professionnel de l'informatique, pour obtenir une assistance supplémentaire.

(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)

[6] Le texte complet de l'article 4 est le suivant : « Les fournisseurs et les déployeurs de systèmes d'IA prennent des mesures pour garantir, dans toute la mesure du possible, un niveau suffisant de maîtrise de l'IA pour leur personnel et les autres personnes s'occupant du fonctionnement et de l'utilisation des systèmes d'IA pour leur compte, en prenant en considération leurs connaissances techniques, leur expérience, leur éducation et leur formation, ainsi que le contexte dans lequel les systèmes d'IA sont destinés à être utilisés, et en tenant compte des personnes ou des groupes de personnes à l'égard desquels les systèmes d'IA sont destinés à être utilisés. »

Solutions actuelles d'intelligence artificielle générative

Le marché des outils d'intelligence artificielle évolue rapidement et offre de nombreuses options, même si celles-ci ne sont pas toujours facilement visibles. Par conséquent, les utilisateurs peuvent être enclins à adopter les outils les plus facilement accessibles, souvent ceux qui sont peu coûteux ou fortement promus, même lorsque ces outils ne sont pas spécifiquement conçus pour la pratique juridique.

Pour la plupart des professionnels du droit qui n'ont pas suivi de formation technique avancée, les systèmes d'intelligence artificielle générative actuellement disponibles auprès de divers fournisseurs peuvent sembler indiscernables. L'interface utilisateur comprend généralement une fenêtre de chat, principalement basée sur un navigateur, qui permet aux utilisateurs d'interagir avec le système d'intelligence artificielle par l'intermédiaire d'une saisie conversationnelle. Les utilisateurs peuvent poser des questions et envoyer des fichiers, tels que des textes, des images et des vidéos, et obtenir des réponses dont la qualité et la pertinence varient. Il est important de noter que très peu de ces systèmes permettent aux utilisateurs de comprendre leurs mécanismes sous-jacents ou fournissent des informations transparentes sur le traitement des données fournies par les utilisateurs.

Ce guide fournit des informations sur les différentes formes de systèmes d'intelligence artificielle générative et aide à déterminer la catégorie spécifique à laquelle un utilisateur (un avocat) a affaire. Tout d'abord, il est essentiel de comprendre la distinction entre un « **modèle d'intelligence artificielle** » et un « **système d'intelligence artificielle** » [7] . Un modèle d'intelligence artificielle n'est qu'un composant technique parmi d'autres au sein d'un système plus large : il s'agit du résultat final d'un processus complexe d'apprentissage automatique appelé « entraînement », qui constituera l'élément le plus précieux du système d'intelligence artificielle dans son ensemble. En revanche, un système d'intelligence artificielle désigne l'application opérationnelle complète qui intègre un ou plusieurs modèles d'intelligence artificielle avec des composants supplémentaires, tels que des données en pipelines, des interfaces utilisateur et une infrastructure sous-jacente, afin de fournir un service fonctionnel et accessible.

La plupart des modèles d'intelligence artificielle contemporains utilisés par les avocats sont des **grands modèles de langage (LLM)**. Ils s'appuient sur une architecture spécifique de modèles linguistiques de réseaux neuronaux qui suivent l'architecture des transformeurs. Au lieu de lire le texte mot par mot comme les systèmes précédents, les transformeurs « examinent » tous les mots d'un texte plus long en même temps. Cela leur permet, par exemple, de comprendre beaucoup plus efficacement le contexte, notamment les relations entre les mots, même dans des textes longs. Cette approche a connu un succès mondial après qu'OpenAI l'ait mise en œuvre dans une approche reposant exclusivement sur un décodeur via une architecture appelée GPT (Generative Pre-trained Transformers), qui utilise la partie décodeur de l'architecture du

[7] Voir également le glossaire technique.

transformeur (conçue pour prédire le mot suivant dans une séquence). Il s'agit de la partie spécialisée dans la génération de texte.

Plus récemment, les modèles multi- ou omni-modaux sont devenus plus courants et sont entraînés non seulement sur du texte, mais aussi sur des images et des vidéos, et ils ne sont pas toujours appelés LLM.

Les grands modèles requièrent énormément de mémoire, tant pendant leur entraînement, lorsqu'ils sont créés et perfectionnés, que pendant leur utilisation quotidienne, appelée **phase d'inférence**. Leurs besoins en calcul peuvent être si élevés que certains modèles ne peuvent pas fonctionner sur des ordinateurs portables, des téléphones ou des serveurs locaux classiques. Au lieu de cela, nos appareils envoient des requêtes à de puissants serveurs à distance qui traitent la demande et renvoient le résultat. Les modèles les plus grands et les plus avancés nécessitent une puissance de calcul, une énergie et un refroidissement tels qu'ils ne peuvent fonctionner que dans des centres de données spécialisés, et non sur du matériel grand public ni même sur du matériel professionnel standard.

Ces besoins élevés en puissance de calcul ont également un coût. En 2025, l'achat de matériel permettant d'exécuter les modèles les plus performants aussi rapidement que le font les fournisseurs commerciaux coûterait 6 000 000 € en frais ponctuels, avec des coûts mensuels d'environ 50 000 € pour la consommation d'électricité. La combinaison de frais élevés et de besoins importants en puissance de calcul signifie que ces modèles ne sont généralement accessibles qu'à distance en tant que service, tout en étant exécutés à partir des plus grands centres de données, dont très peu sont situés dans l'UE (**SaaS, ou « Software-as-a-Service »**).

Cependant, les LLM plus petits[8] sont récemment devenus de plus en plus puissants. De plus, les appareils finaux sont également de plus en plus équipés de puces qui réduisent les coûts d'inférence. Il est donc désormais possible pour un avocat de déployer et d'exploiter des systèmes d'intelligence artificielle avec des modèles de grande taille dans des environnements plus accessibles, plus abordables et plus contrôlables. Ces possibilités permettent à un avocat d'utiliser des modèles de grande taille de manière à contrôler leurs opérations et le traitement des données. Cela implique toutefois une plus grande responsabilité de la part de l'utilisateur en termes de sécurité informatique, tant physique que logique.

Les considérations ci-dessus, à savoir les coûts, la puissance de calcul, l'expertise requise et la responsabilité, peuvent être présentées sous la forme d'un éventail d'options disponibles. Ce guide identifie **quatre grandes catégories** qui sont primordiales pour comprendre comment les avocats peuvent utiliser les modèles d'intelligence artificielle :

1. **Appareils (auto-hébergés) sur site**
2. **Appareils auto-hébergés dans un centre de données privé/en colocation**

[8] Certains d'entre eux sont si petits (moins d'un milliard de paramètres) qu'ils ne sont même pas considérés comme des « modèles d'intelligence artificielle à usage général » au sens de la législation européenne sur l'intelligence artificielle. Ils ne devraient donc probablement pas être qualifiés de « grands » modèles linguistiques. Néanmoins, ces modèles peuvent encore être étonnamment performants, par exemple Qwen 2.5 0.5B, etc., même s'ils ne sont pas utilisables dans la pratique sans avoir été formés ou ajustés pour une tâche spécifique, telle que la classification.

3. **Modèle « propre » sur une infrastructure en tant que service (IaaS), des machines virtuelles ou une infrastructure en nuage similaire**
4. **SaaS/API entièrement géré (hébergé par le fournisseur)**

1. Appareils (auto-hébergés) sur site

Un modèle d'intelligence artificielle sur site fonctionne entièrement sur du matériel appartenant à l'organisation de l'utilisateur et contrôlé par celle-ci plutôt que dans le nuage ou sur les serveurs à distance d'un fournisseur. L'utilisateur bénéficie d'un contrôle maximal sur ses flux de données, la journalisation et la conservation des données et assume l'entière responsabilité de la sécurité et de la disponibilité des appareils. Toutes les données demeurent dans l'infrastructure de l'utilisateur, ce qui signifie que rien n'est envoyé à un fournisseur de nuage externe. L'organisation de l'utilisateur est responsable de la sécurité, des mises à jour, de la maintenance et de la conformité. Les utilisateurs ont un contrôle maximal sur le fonctionnement du modèle et le traitement des données.

1.1 Avantages

Le déploiement sur site est souvent préféré lorsque les utilisateurs traitent des données confidentielles ou sensibles. **Pour les avocats, c'est la solution la plus sûre en matière de respect de la confidentialité.** Les données sont stockées, traitées et consultées dans les locaux de l'avocat et restent sous son contrôle. Il n'y a aucun transfert de données à un tiers. Cette solution peut s'accompagner de protections supplémentaires (selon la législation locale) : le cabinet d'un avocat peut bénéficier d'une protection supplémentaire contre les mandats de perquisition de la police (par exemple, une autorisation spéciale du tribunal est nécessaire, un procureur ou un représentant du barreau doit être impliqué). Une telle solution permet également une personnalisation et une vérifiabilité complète.

1.2 Inconvénients

Le compromis réside dans le coût : l'exécution de grands modèles au niveau local nécessite du matériel spécialisé, une consommation d'énergie importante et une expertise technique. Si le modèle n'est pas connecté à Internet, le principal risque de cybersécurité se limite à la perte potentielle de données. S'il est connecté, le système peut être exposé à des menaces supplémentaires en matière de cybersécurité. Si l'utilisateur réduit ce risque supplémentaire, il peut toujours avoir l'assurance de conserver le contrôle total de ses données.

2. Appareils auto-hébergés dans un centre de données privé/en colocation

Un modèle d'intelligence artificielle auto-hébergé dans un centre de données privé ou en colocation fonctionne sur des serveurs que l'utilisateur possède ou contrôle, mais qui sont physiquement hébergés en dehors de ses locaux (par exemple, dans un centre de données professionnel où les utilisateurs louent de l'espace physique, de l'électricité, le refroidissement

et la connectivité, ou dans le centre de données privé de l'utilisateur). Dans ce scénario, l'utilisateur conserve le contrôle total du modèle d'intelligence artificielle, des serveurs sur lesquels il fonctionne, des données qui transitent par le système, ainsi que de la sécurité, des politiques d'accès et de la vérifiabilité.

2.1 Avantages

Grâce à cette solution, l'avocat reste propriétaire de l'ensemble du matériel et dispose d'un profil de contrôle similaire à celui de la solution sur site.

2.2 Inconvénients

Le fournisseur de colocation garantirait la disponibilité des installations et la sécurité physique, parfois avec des « mains à distance » louables (pour effectuer les mises à jour nécessitant une intervention physique, etc.). Cela représente une dépense récurrente pour le cabinet d'avocats.

Du point de vue de l'avocat, cela signifie également que le fournisseur et l'emplacement doivent être fiables et choisis conformément aux règles déontologiques nationales régissant le recours à des fournisseurs informatiques externes. Une telle utilisation ouvre la voie à un nouveau vecteur d'attaque potentiel, à savoir la manipulation physique des appareils, qui peut également permettre d'accéder aux journaux ou aux données sensibles. En outre, les lois nationales qui offriraient une protection supplémentaire aux cabinets d'avocats en cas de mandats de perquisition pourraient ne pas s'appliquer dans ce cas en raison du fait que les données ne sont pas conservées physiquement dans les locaux de l'avocat.

3. Modèle « propre » sur une infrastructure en tant que service (IaaS), des machines virtuelles ou une infrastructure en nuage similaire

Dans ce scénario, l'utilisateur apporte son propre modèle d'intelligence artificielle et le fournisseur de nuage fournit l'infrastructure informatique. Cela diffère du SaaS étant donné que le fournisseur de nuage ne fournit pas le modèle, mais uniquement la couche matérielle ainsi que quelques fonctionnalités logicielles de base[9]. L'utilisateur utilise du matériel appartenant à des tiers, par exemple une machine virtuelle fonctionnant sur une machine physique appartenant à un tiers. Le fournisseur de nuage ou de machines virtuelles/louées gère le matériel et la sécurité de base dans le cadre d'un modèle de responsabilité partagée.

3.1 Avantages

Cette solution offre aux utilisateurs des avantages similaires à ceux de l'exploitation de leurs propres serveurs, mais sans le coût du matériel physique du centre de données. Dans ce modèle, l'utilisateur conserve le contrôle total du traitement des données et du fonctionnement du modèle. Dans la plupart des cas, l'utilisateur peut également décider de l'emplacement du centre de données où la machine virtuelle est exécutée (ou du moins, de la région géographique) et, surtout, du moment et du lieu où les données elles-mêmes peuvent être transférées, avec chiffrement et garde complète des clés de chiffrement, etc. Comme l'utilisateur exploite le

[9] Tels que le système d'exploitation, les couches de virtualisation, la surveillance, etc.

modèle d'intelligence artificielle sur des serveurs à distance, il garde le contrôle du modèle d'intelligence artificielle et de la version du modèle qu'il utilise.

Un avantage important de cette solution est l'indépendance étant donné que l'utilisateur évite d'être lié à l'écosystème d'un seul fournisseur d'intelligence artificielle. L'utilisateur peut également adapter le modèle à ses cas d'utilisation préférés (par exemple, les avocats peuvent ajuster le modèle en fonction de leurs documents internes et des domaines dans lesquels ils travaillent) et comprendre et contrôler la plupart des aspects importants de son système d'intelligence artificielle. De plus, les GPU en nuage peuvent être augmentés ou réduits selon les besoins. Enfin, le prix de ces environnements virtuels est généralement beaucoup plus attractif que celui des solutions sur site ou en nuage privé (les options 1 et 2 ci-dessus, respectivement).

3.2 Inconvénients

Cette solution ne permet généralement pas à l'utilisateur de disposer d'un droit de vérification complet. Ces droits, tels que spécifiés par les fournisseurs, signifient souvent que les utilisateurs se verront présenter les certificats pertinents sur demande. Du point de vue de l'avocat, il est nécessaire de tenir compte des règles professionnelles nationales relatives à l'utilisation de prestataires informatiques externes. Celles-ci peuvent comporter l'obligation pour les prestataires d'informer immédiatement l'avocat en cas de mesures de perquisition et de saisie. Il s'agit d'une obligation que les petits et moyens prestataires peuvent fournir ou être disposés à négocier, mais que les grands prestataires ne sont souvent pas disposés à faire. Il est important de préciser qu'il existe également des lois pertinentes régissant la notification des mesures d'enquête aux personnes qui en font l'objet[10].

4. SaaS/API entièrement géré (hébergé par le fournisseur)

Dans le cadre de cette solution, l'utilisateur peut accéder au modèle via une interface utilisateur web ou via une **interface de programmation d'application (API)**. Le fournisseur contrôle l'ensemble du système d'intelligence artificielle et gère le modèle d'intelligence artificielle, les serveurs et le stockage, toutes les mises à jour, les filtres de sécurité, les garde-fous, ainsi que le changement d'échelle, les performances et la disponibilité.

4.1 Avantages

Les avantages de cette solution sont la facilité d'utilisation étant donné que les utilisateurs peuvent accéder immédiatement au service via une interface utilisateur (UI) ou une API. Les utilisateurs ne sont pas responsables de l'infrastructure. Il s'agit de la solution la plus simple, et parfois la seule, pour que les utilisateurs accèdent aux meilleurs et aux plus puissants modèles de pointe. Cette solution est également rentable, en particulier par rapport à une solution sur site ou IaaS.

[10] Par exemple, l'article 13 du règlement (UE) 2023/1543 du Parlement européen et du Conseil du 12 juillet 2023 relatif aux injonctions européennes de production et aux injonctions européennes de conservation concernant les preuves électroniques dans le cadre des procédures pénales et aux fins de l'exécution de peines privatives de liberté prononcées à l'issue d'une procédure pénale, JO L 191, 28.7.2023, disponible ici : http://data.europa.eu/eli/reg/2023/1543/oj

4.2 Inconvénients

Cette solution présente des inconvénients, en particulier du point de vue des avocats. Les utilisateurs ne peuvent pas influencer la configuration interne du modèle, la fréquence à laquelle le fournisseur met à jour ou remplace le modèle, ni les filtres de sécurité ou les limitations appliqués. **Les données quittent l'environnement de l'utilisateur, ce qui, dans le cas des avocats, peut poser des problèmes de confidentialité.** En outre, les utilisateurs peuvent ne pas être en mesure de réaliser des vérifications complètes, étant donné que les fournisseurs ne montrent généralement que les certificats pertinents. Enfin, les utilisateurs sont confrontés à des risques de dépendance vis-à-vis des fournisseurs et à une transparence limitée. Cette architecture implique un contrôle minimal sur le fonctionnement interne étant donné qu'il s'agit d'un service entièrement partagé qui vise à servir le plus grand nombre de clients au coût le plus bas possible. Les obligations contractuelles qui accorderaient à l'avocat des droits de contrôle ou de vérification supplémentaires sont souvent non négociables parce que ces fournisseurs travaillent souvent à « hyper-échelle » et ne peuvent pas ou ne veulent pas accorder de traitement spécial à des clients individuels ou même à des groupes de clients.

Du point de vue de l'avocat, il est donc essentiel de comprendre les détails du contrat du fournisseur et leurs conséquences sur les activités de l'avocat. Dans ce contexte, les avocats doivent se poser de nombreuses questions :

- Les fournisseurs décrivent-ils dans leurs conditions contractuelles la manière dont les données des clients seront traitées ?
- Quel est le rôle du fournisseur dans le flux de traitement des données ? S'il assume le rôle de sous-traitant, fournit-il un accord de traitement des données et les dispositions de celui-ci sont-elles conformes au RGPD et aux autres réglementations applicables ?
- Les fournisseurs entraîneront-ils leurs modèles à partir des données téléchargées ? (c'est-à-dire les instructions et les réponses de l'avocat, via le contexte et l'historique des conversations)
- Les données seront-elles accessibles au personnel des fournisseurs et, si oui, dans quelles conditions ?
- Les données seront-elles accessibles à des tiers, y compris aux services répressifs (sur la base de mandats ou d'injonctions de production) ? Si oui, pendant combien de temps ?
- L'avocat sera-t-il informé par le fournisseur avant ou au moins immédiatement après la divulgation ou le transfert des données à des tiers, y compris aux services répressifs ?
- Les fournisseurs conservent-ils les données après leur suppression par l'utilisateur, par exemple dans leurs sauvegardes à des fins de reprise après sinistre ? Pendant combien de temps ?
- Que se passe-t-il si le fournisseur est victime d'une cyberattaque et n'est plus en mesure de respecter les engagements contractuels convenus avec l'utilisateur ?
- Où se trouve physiquement le centre de données du fournisseur ? Quelles lois régissent le traitement des données et la coopération avec les services répressifs ? Les données traitées seront-elles soumises à des juridictions différentes, voire concurrentes ?
- De plus, si le modèle est entraîné à partir des données ou des instructions de l'avocat, les avocats ne peuvent pas savoir quelles parties de leurs données seront réutilisées pour un autre utilisateur et comment ces données leur seront présentées.

- Quelles sont les restrictions de la politique d'utilisation acceptable (qui peuvent être cruciales pour les affaires pénales ou relatives aux droits humains) ?

Il est également essentiel d'expliquer comment les données saisies par l'utilisateur dans le système d'intelligence artificielle basé sur le LLM sont traitées. Lorsqu'un prompt est soumis à un grand modèle de langage, il est d'abord tokenisé et converti en représentations numériques, qui sont traitées dans la fenêtre contextuelle du modèle, puis les prédictions du token suivant sont générées pour former la réponse. Ce traitement est généralement transitoire, ce qui signifie que les tokens saisis ne sont conservés dans la mémoire de travail que pendant la durée techniquement nécessaire au calcul du résultat et ne sont pas automatiquement stockés dans une base de données récupérable ni constitués dans les paramètres du modèle.

Le fait qu'un prompt soit utilisé ultérieurement pour l'entraînement du modèle ne dépend pas de l'architecture du modèle elle-même, mais des conditions contractuelles, de la configuration technique et des politiques de gouvernance du fournisseur. Les services d'entreprise ou API excluent souvent contractuellement l'utilisation à des fins d'entraînement, tandis que les services destinés aux consommateurs peuvent se réserver le droit d'utiliser les entrées pour améliorer leurs services, sauf si l'utilisateur s'y oppose.

Même lorsque les prompts sont utilisés pour l'entraînement, cela ne signifie pas que le texte exact devient accessible à d'autres : l'entraînement implique des ajustements de pondération sur des milliards de paramètres plutôt que le simple stockage du corpus d'entraînement, de sorte que le modèle est censé apprendre des formules abstraites plutôt que de conserver une copie consultable du prompt.

Cependant, bien que la mémorisation mot pour mot (des prompts) ne soit pas l'objectif de la conception, des risques résiduels (par exemple, la journalisation, la révision humaine, des failles de sécurité ou des phénomènes rares de mémorisation) apparaissent au niveau du système d'intelligence artificielle.

Si une affaire est sensible, un avocat peut avoir besoin d'utiliser un modèle plus contrôlé étant donné que le secret professionnel, les obligations en matière de réglementation et la confidentialité des clients dépendent tous de qui peut accéder aux données et les copier. Moins nous contrôlons le **stack**, plus nous devons garantir le contrôle par des obligations contractuelles envers le fournisseur et nous fier aux promesses d'un fournisseur. L'avocat doit vérifier non seulement les dispositions du contrat, mais aussi la probabilité que le contrat soit exécutoire.

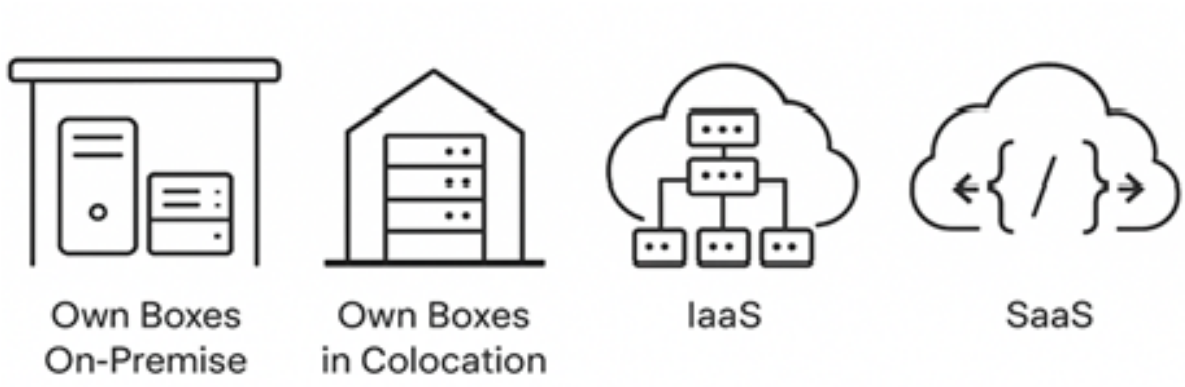
Cette infrastructure spécifique doit être prise en compte étant donné qu'elle implique que les données entrées dans l'interface utilisateur du système d'intelligence artificielle doivent être envoyées au modèle d'intelligence artificielle. Si le fournisseur du système d'intelligence artificielle et le fournisseur du modèle d'intelligence artificielle ne sont pas identiques, cela signifie également un transfert de données à un tiers, éventuellement situé dans un pays tiers.

Pour les données au repos, c'est-à-dire les données traitées par le modèle fonctionnant dans un environnement spécifique, le fournisseur peut modifier unilatéralement ses conditions générales, ou même simplement les paramètres par défaut. Un fournisseur peut être racheté par une partie peu fiable ou devenir insolvable. Un fournisseur peut également promettre la

suppression ou la non-divulgation des données, mais se trouver néanmoins contraint par des dispositions légales obligatoires de conserver ou de remettre les données, parfois sous le couvert de restrictions de confidentialité qui interdisent d'en informer l'utilisateur. Les flux transfrontaliers de données peuvent amplifier ce risque puisque les données relèveront de plusieurs juridictions. Malheureusement, même les plus grands fournisseurs peuvent être victimes de graves violations de sécurité, exposant ainsi les données à des tiers malveillants[11].

Pour les données en transit, le trafic vers un service géré peut traverser des réseaux sur lesquels l'avocat n'a aucun contrôle. Même si un fournisseur promet souvent un « chiffrement de bout en bout », aucune des parties ne sera en mesure de détecter ou de garantir à quel moment une communication donnée est réellement déchiffrée et qui peut avoir accès aux données au-delà de ce point. Le message ou le flux de données chiffrés peuvent toujours aboutir à des nœuds qui échappent au contrôle de cette partie, souvent sans aucun contrôle de la part de l'utilisateur, avec de nombreux sous-traitants tiers impliqués dans la journalisation et la surveillance. Dans de tels environnements, il est pratiquement impossible d'obtenir des informations sur les capacités d'interception légale et de contournement du chiffrement existantes.

Tableau 1 : Résumé des caractéristiques, avantages et inconvénients des différents modèles de déploiement



Modèle de déploiement	Qui fournit le modèle ?	Où le modèle est-il exécuté ?	Qui contrôle l'infrastructure ?	Lieu de stockage et confidentialité des données	Niveau de contrôle	Avantages	Inconvénients
SaaS/API entièrement géré	Fournisseur	Serveurs du fournisseur	Fournisseur	Données envoyées en externe	Faible	Accès facile et peu coûteux aux modèles de pointe	Contrôle limité, risques liés à la confidentialité
Modèle propre sur IaaS	Organisation de l'utilisateur	Machines virtuelles (VM) dans le nuage	Fournisseur de nuage	Sélectionnable par région, avec chiffrement	De moyen à élevé	Flexible et évolutif	Infrastructure externe, expertise requise
Appareils auto-hébergés dans un centre de données	Organisation de l'utilisateur	Colocation/centre de données privé	Partagé (infrastructure des installations	Données sur le matériel de	Élevé	Confidentialité et vérifiabilité	Maintenance, coût

[11] Telles que l'exploitation zero-day de Microsoft SharePoint ToolShell en 2025 (CVE-2025-53770) ou la violation de Snowflake Inc. en 2024.

privé/en colocation			et serveurs de l'organisation)	l'organisation de l'utilisateur			
Appareils (auto-hébergés) sur site	Organisation de l'utilisateur	Serveurs locaux/internes	Organisation de l'utilisateur	Données conservées en interne	Très élevé	Confidentialité maximale	Coût élevé, maintenance

Considérations techniques pour les options hors SaaS

Cette partie du guide explique quels modèles ouverts (gratuits) sont actuellement disponibles, ce qu'ils peuvent faire et quel type de matériel est nécessaire pour les faire fonctionner, y compris un coût indicatif pour l'achat de ce matériel. Elle s'adresse principalement aux lecteurs qui souhaitent explorer d'autres solutions que les solutions SaaS, où le matériel sous-jacent reste entièrement hors du contrôle de l'utilisateur. Pour ceux qui envisagent un modèle IaaS, les informations fournies sont principalement utiles pour déterminer le niveau d'infrastructure virtuelle (par exemple, vCPU, RAM et GPU) qui serait suffisant pour leurs besoins.

À titre informatif, le guide fournit ci-dessous quelques estimations de prix valables en septembre 2025 sans compter les frais d'installation, de déploiement ou de main-d'œuvre. Il serait plus difficile de comprendre et de comparer ces solutions si nous omettions tous ces chiffres. Il est important de souligner que les tendances du marché ont une incidence constante et considérable sur ces prix. Malheureusement, les prix du matériel informatique sont devenus très volatils ces derniers temps. Par exemple, les prix de la mémoire vive (RAM) ont doublé en 2025, ce qui a également entraîné une augmentation significative des coûts des capacités d'inférence en intelligence artificielle locales au cours des derniers mois de l'année[12].

Afin de guider le lecteur dans la deuxième partie, plus technique, le présent document donne une description générale de chacun des paramètres pertinents. Les avocats, en tant qu'utilisateurs finaux, utilisent généralement les modèles d'intelligence artificielle pendant la **phase d'inférence** (génération de résultats à partir de modèles entraînés) plutôt que de les entraîner davantage, ce qui nécessite beaucoup plus de ressources. L'inférence consiste à utiliser les **poids** entraînés du modèle, à traiter les données d'entrée (telles que les prompts, les images ou les textes) à l'aide de ces poids et à récupérer le résultat obtenu.

La plupart des modèles utilisés quotidiennement sont assez volumineux : alors que les plus petits peuvent ne nécessiter qu'1 Go d'espace de stockage, les plus grands (librement accessibles) peuvent occuper plus de 2 To.

Pour pouvoir exécuter un modèle, il faut disposer d'un ordinateur capable de stocker l'intégralité du modèle dans sa mémoire. De plus, pour pouvoir l'exécuter à des vitesses acceptables pour un usage professionnel, il faut également disposer d'un matériel spécialisé et relativement coûteux.

C'est pourquoi il n'est pas possible de se fier à l'unité centrale de traitement (**CPU**) normale de son ordinateur pour les modèles plus volumineux et qu'il est nécessaire de se procurer des unités de traitement graphique (**GPU**). La mémoire pour les CPU et les GPU est très coûteuse et

[12] Cette augmentation considérable des prix est due à une pénurie de mémoire et à la réservation, plusieurs années à l'avance, des capacités de production de mémoire des principaux fabricants mondiaux afin de construire ce que l'on appelle les « usines géantes d'intelligence artificielle ».

nécessite des investissements en capital importants, et les opérations associées sont très gourmandes en énergie, ce qui entraîne à son tour de nombreux projets de construction de centres de données dans le monde entier.

À l'inverse, les progrès réalisés dans les modèles commerciaux de pointe ont également eu des effets positifs sur les capacités des modèles plus petits, qui peuvent désormais être exploités par des petites entreprises à budget limité, voire par des personnes physiques, sur du matériel local ou sur un fournisseur d'infrastructure de leur choix, sous leur contrôle exclusif total. Il peut donc exister un « juste milieu » dans lequel un modèle plus petit effectue une tâche donnée de manière comparable aux systèmes commerciaux de pointe tout en restant relativement peu coûteux à exploiter. Les besoins en ressources d'un modèle d'intelligence artificielle dépendent toutefois de nombreux facteurs, notamment les suivants :

1. Les capacités du modèle

Il existe aujourd'hui divers modèles de chatbots conversationnels capables de répondre aux questions des utilisateurs, mais les performances prêtes à l'emploi des modèles plus petits restent nettement inférieures à celles des modèles LLM de pointe, tels que Claude Opus 4.6 ou GPT-5.4. Néanmoins, même ces systèmes plus petits peuvent être ajustés pour des tâches linguistiques spécifiques ou utilisés pour la classification de base de phrases et d'opérations routinières similaires. En revanche, les LLM plus grands sont à la base des modèles dits de raisonnement, qui sont conçus pour générer des résultats plus réfléchis et progressifs pour les utilisateurs finaux, généralement au prix d'un temps de réponse plus lent.

Un autre type de modèle, appelé modèle à plongement vectoriel (*embedding*), convertit les textes (tels que les documents, les rapports ou les textes législatifs) en une forme numérique pouvant faire l'objet de recherches à l'aide de la similarité sémantique plutôt qu'à l'aide de mots-clés exacts ou d'expressions régulières. Ces modèles ne sont généralement pas utilisés isolément, mais comme composants de systèmes de recherche d'informations plus complexes, tels que des moteurs de recherche comme ElasticSearch. Une autre utilisation possible consiste à transformer des enregistrements audios en transcriptions, ce qu'on appelle la **reconnaissance automatique de la parole** (ASR). Il s'agit de modèles à usage unique, mais souvent intégrés aux LLM.

Un autre type de modèle qui mérite d'être évoqué est celui des modèles vision-langage (VLM) qui, contrairement aux LLM traditionnels uniquement textuels, peuvent générer des descriptions textuelles à partir d'images afin de permettre la recherche et le filtrage d'images. Ils ont également considérablement amélioré les performances de **la reconnaissance optique de caractères (OCR)** classique, en extrayant de manière fiable les écritures manuscrites lisibles, les formules mathématiques et autres contenus similaires.

Les capacités attendues d'un modèle déterminent en grande partie le type de matériel nécessaire pour le faire fonctionner[13].

Les avocats doivent être en mesure d'estimer les tâches qui peuvent être effectuées localement et à quel coût approximatif, ainsi que les fonctionnalités et les cas d'utilisation pour lesquels un service géré par des professionnels reste la meilleure solution. Mais pour pouvoir en faire l'estimation, les avocats doivent connaître les capacités et certains détails techniques des modèles locaux d'intelligence artificielle. Cela ressemble un peu à la situation des années 1980, où il était essentiel que les passionnés acquièrent un ordinateur personnel pour que les ordinateurs se diffusent dans tous les domaines de la vie. Les modèles d'intelligence artificielle locaux permettent aux avocats (et à la population en général) d'avoir cette expérience directe et de réduire leur dépendance vis-à-vis des grandes entreprises opérant dans diverses juridictions, même si l'utilisation de modèles locaux est moins confortable ou moins performante.

2. La taille du modèle

Le deuxième élément clé à prendre en compte est la taille des modèles compatibles avec une configuration matérielle donnée. La taille des modèles est généralement exprimée en nombre de paramètres (de millions à des milliards), un modèle 8B ayant 8 milliards de paramètres (c'est-à-dire 8 milliards de « poids et biais » distincts). Pour les modèles appartenant à la même famille ou provenant du même fournisseur, la taille est généralement corrélée à la fois aux capacités et aux exigences matérielles. Par exemple, DeepSeek-r1:14b (un modèle DeepSeek à 14 milliards de paramètres) est plus performant que DeepSeek-r1:8b (8 milliards de paramètres).

Cela n'est toutefois pas toujours vrai en raison de divers facteurs. Par exemple, le nouveau modèle Gemma 3 4B de Google surpasse l'ancien modèle Gemma 2 27B, plus grand, dans de nombreux comparatifs. De même, différents éditeurs de modèles peuvent offrir des performances supérieures dans la même gamme de paramètres grâce aux progrès réalisés en matière d'architecture, de corpus d'entraînement ou d'autres techniques. C'est pourquoi les praticiens s'appuient généralement sur des évaluations de capacités publiées sous forme de comparatifs dans des classements publics[14].

En termes de configuration matérielle requise, un VLM de 8 milliards de paramètres ou un modèle Stable Diffusion 8B de génération d'images ne fonctionnera pas à une vitesse acceptable sans une machine dotée d'un GPU suffisamment puissant, tandis qu'un LLM de 8 milliards peut fournir sa réponse à une vitesse acceptable dans des conversations utilisant uniquement un CPU. Cela signifie que la taille des paramètres n'est pas comparable entre les différents types de modèles[15].

---

[13] Il est possible de trouver une vaste bibliothèque de modèles téléchargeables et gratuits sur HuggingFace : https://huggingface.co/models

[14] https://huggingface.co/spaces/DontPlanToEnd/UGI-Leaderboard ou
https://artificialanalysis.ai/leaderboards/models
https://huggingface.co/spaces/hf-audio/open_asr_leaderboard
https://huggingface.co/spaces/TIGER-Lab/MMLU-Pro
https://huggingface.co/spaces/mteb/leaderboard

[15] Des instructions détaillées sont disponibles sur Internet pour savoir quels modèles peuvent généralement être exécutés sur quels types d'ordinateurs. Par exemple, en ce qui concerne les modèles de petite taille, consultez le site https://www.canirun.ai/ .

En outre, on trouve souvent le même modèle avec des **quantifications** différentes. Il est par exemple possible d'exécuter différentes versions du modèle GPT-OSS-20B gratuit d'OpenAI, telles que les versions FP16 et INT4, cette dernière nécessitant près de quatre fois moins de mémoire. Ces quantifications sont une simplification des poids du modèle entraîné afin d'accélérer l'inférence et de réduire la capacité de la mémoire, mais cela peut également entraîner une certaine perte de capacité.

Tableau 2 : Quantification – règle empirique concernant ses effets sur la taille et la qualité du modèle

Précision	Utilisation de la RAM	Effet sur la vitesse	Effet sur la qualité	Recommandé pour
FP16	1,0×	Référence	Fidélité totale	Précision maximale, résultats critiques pour la sécurité
INT8	~0,5×	+10 à 30	Assez proche de l'original pour la plupart des tâches	Chat général/RAG ; bon réglage par défaut pour les serveurs locaux
INT4	~0,25×	+20 à 50 %	Perte modérée	Économie de coûts et possibilité de contexte plus long avec la même mémoire

3. La vitesse et le nombre d'utilisateurs simultanés

La vitesse de production du modèle est généralement mesurée en « **tokens** par seconde », où un token est une conversion du texte original ou d'une image que le modèle d'intelligence artificielle peut comprendre de manière native. À titre d'illustration, la phrase précédente en anglais était de 173 caractères et correspondait à 35 tokens différents (en utilisant le tokeniseur GPT-4o d'OpenAI)[16]. Dans cette méthode de tokenisation en anglais, un token correspond en moyenne à quatre caractères. Il est important de noter que le nombre de tokens par caractère dépend de la langue : même si la traduction de la même phrase en hongrois ne comporte que 171 caractères, cela correspondra déjà à 53 tokens au lieu de 35 en anglais.

Il existe deux types de mesures de la vitesse des tokens, toutes deux exprimées en tokens par seconde (ou **tps**) :

- la vitesse de traitement des entrées (également appelée « préremplissage ») ; et
- la phase de génération (ou décodage).

Différentes tâches imposent différentes exigences à un modèle : les chatbots prennent généralement des entrées courtes et génèrent des sorties longues, tandis que les tâches de recherche d'informations reposent sur des entrées longues et des réponses courtes. À mesure

[16] Les tokens sont [976, 7733, 328, 2359, 4733, 382, 6971, 26489, 306, 392, 64329, 777, 3099, 672, 1919, 6602, 382, 261, 22165, 328, 290, 4756, 2201, 503, 448, 3621, 484, 290, 20837, 2359, 665, 297, 11594, 4218, 13]. Voir https://platform.openai.com/tokenizer

que les modèles deviennent plus grands, leur vitesse de sortie ralentit et ils nécessitent un matériel de plus en plus puissant et coûteux pour maintenir des performances acceptables[17].

En général, une vitesse de génération inférieure à 5 tps est trop lente pour une quelconque utilisation interactive, tandis qu'une vitesse supérieure à 20 tps dépasse la vitesse de lecture moyenne d'un avocat. Une vitesse d'environ 100 tps dépasse la vitesse de lecture rapide habituelle, même si elle peut encore sembler lente si la tâche implique, par exemple, la recherche dans une longue sortie de modèle. Il est important de noter que si plusieurs utilisateurs dépendent du même système local, son utilisation simultanée divisera et réduira encore davantage la vitesse de génération effective[18].

4. La longueur des entrées et des sorties

Le dernier terme technique à connaître est la « longueur du contexte », qui est également mesurée en tokens. La longueur typique pour la plupart des modèles locaux est maintenant de 16 384 tokens ou davantage, ce qui inclut à la fois le prompt et les fichiers joints (« tokens d'entrée ») et la longueur de la réponse (« tokens de sortie »). Comme décrit ci-dessus, une page imprimée moyenne en anglais peut contenir jusqu'à 500 tokens. Ainsi, si un utilisateur souhaite que le chatbot recherche toutes les lois et décisions judiciaires applicables dans un pays, il aurait besoin d'environ 3,5 milliards de tokens dans ce pays hypothétique. Cela est tout simplement impossible et, dans ces cas, une approche différente est suggérée, telle que la recherche d'informations (y compris les outils **RAG - génération augmentée par récupération**).

Même si l'on souhaite simplement rechercher une affaire spécifique dans des documents judiciaires, examiner les preuves ou les déclarations des témoins, on peut facilement atteindre un volume de 100 pages, ce qui pourrait représenter environ 50 000 tokens.

Tableau 3 : Quelle est la priorité pour un cas d'utilisation ? Longueur et vitesse

Tâche	Longueur de l'entrée	Longueur de la sortie	Vitesse d'entrée (priorité de préremplissage)	Vitesse de sortie (priorité de décodage)
Questions-réponses sur le chat	Court	De courte à moyenne	●	●●
RAG juridique (prompt long)	Long	De courte à moyenne	●●●	●
OCR en masse + sous-titrage VLM	Long (images/PDF)	Courte	●●	●
Résumé par lots (documents courts)	Court	Courte	●	●
Transcription (ASR)	Long (audio)	Longue	●	●●●

[17] Vous pouvez tester la vitesse acceptable pour votre cas d'utilisation sur ce simulateur : https://kamilstanuch.github.io/LLM-token-generation-simulator/

[18] La liste des vitesses disponibles en fonction du matériel est disponible ici : https://llm.aidatatools.com/results-windows.php.

Quels modèles peuvent être exécutés localement ?

Compte tenu de ces explications, cet article examine maintenant ce qu'un avocat peut raisonnablement exploiter localement. Il présente les différentes solutions possibles en fonction du budget disponible : de l'utilisation d'ordinateurs existants à des dépenses qui ne sont pas vraiment envisageables pour les petits cabinets.

La solution la plus économique consisterait à **utiliser un petit modèle sur un ordinateur existant** (même s'il date de quelques années). Il peut s'agir de chatbots conversationnels, tels que deepseek-r1:1.5b, ou de modèles à plongement vectoriel (*embedding*) et de recherche d'informations sur un ordinateur Windows standard doté d'au moins 8 Go de RAM. Une machine de 16 Go peut déjà faire fonctionner des modèles plus performants comme deepseek-r1:14b à une vitesse modérée de 2,5 tps ou être utilisée, bien que lentement, à des fins de reconnaissance automatique de la parole (ASR). Les coûts sont minimes, à l'exception du temps nécessaire pour configurer les outils concernés. Il convient de noter qu'il existe déjà des « moteurs d'inférence » accessibles, tels que Ollama, LMStudio ou AnythingLLM, qui peuvent aider à télécharger un modèle approprié.

Le niveau supérieur et la solution la plus économique consistent en une **machine dédiée à l'utilisation de modèles d'intelligence artificielle locaux**. En se basant sur les prix de septembre 2025, il convient de prévoir une dépense unique d'environ 2 000 € (hors TVA). Ce prix comprend tous les composants nécessaires à un ordinateur, une machine d'inférence, y compris une carte mère avec 128 Go de RAM[19] et un processeur rapide, deux GPU peu coûteux et 24 Go de VRAM (par exemple, deux à quatre GPU lorsqu'on utilise une carte mère adéquate). Cela permet d'exécuter des modèles de 20 à 40 milliards de paramètres en mode texte à une vitesse confortable[20]. Dans cette gamme de prix, il n'est plus souhaitable d'exécuter des modèles sans GPU[21].

[19] Veuillez noter les remarques préliminaires concernant l'extrême volatilité des prix de la RAM.

[20] La plupart des cartes mères grand public ne peuvent accueillir qu'un seul GPU à pleine vitesse, mais certaines peuvent en recevoir jusqu'à quatre à la fois (grâce à quatre emplacements PCI Express x16 mécaniques, comme la Gigabyte B650 EAGLE AX), même si les trois derniers GPU ne pourront pas utiliser la pleine vitesse du processeur (fonctionnant plutôt en mode x1). Néanmoins, cela permet tout de même d'améliorer la vitesse pour un budget modeste, ce qui est très apprécié des passionnés. Cependant, cela n'est pas vraiment utilisable pour les cas d'utilisation VLM.

[21] Veuillez noter qu'il s'agit d'une simplification importante. Une fois que l'on dispose d'une mémoire RAM (VRAM) suffisante pour héberger un modèle, la question cruciale suivante est celle de la bande passante mémoire. Si les cartes mères de type serveur et station de travail peuvent offrir une bande passante rapide, elles ont tendance à être beaucoup plus coûteuses que les composants grand public. Les GPU offrent donc un bien meilleur rapport qualité-prix et une vitesse supérieure jusqu'à une certaine capacité de mémoire. Cependant, seules quelques cartes mères très coûteuses permettent d'insérer quatre GPU ou davantage, ce qui nécessite également des solutions de onéreuses pour y procéder. De plus, seuls les GPU très coûteux contiennent plus de 32 Go de VRAM (à l'instar de la version 96 Go à 8 000 € pièce). Étant donné que les modèles à poids ouverts les plus volumineux nécessitent 1 à 2 To de mémoire, ces modèles ne pourraient pas s'inscrire dans des budgets inférieurs à 150 000 € lorsqu'on utilise uniquement l'inférence GPU. Parallèlement, il est possible de les faire fonctionner sur des machines à 20 000 € équipées uniquement d'un processeur ou en partie sur des GPU relativement peu coûteux, mais à une vitesse réduite (5 à 8 tps).

La solution suivante, plus coûteuse, consiste à acheter un GPU plus cher, tel que le RTX Pro 6000 de Nvidia avec 96 Go de VRAM (environ 8 000 €). Il peut être utilisé avec les ordinateurs existants à condition qu'ils disposent de suffisamment d'espace et de puissance, ou avec une machine d'inférence locale dédiée. Avec la machine d'inférence locale, il est possible d'utiliser le meilleur modèle ouvert actuel d'OpenAI, appelé GPT-OSS-120B. Le deepseek-r1:14b précité peut atteindre une vitesse de 114 tps en utilisant ce GPU (avec une longueur de contexte limitée).

Au-delà de ces solutions, il faudrait envisager l'achat d'un serveur ou d'une configuration spécifique à une station de travail. En effet, les cartes mères grand public disposent rarement d'une bande passante suffisante pour plus d'un GPU (à l'exception de certains cas d'utilisation très limités, comme l'exemple de la **carte mère** à 2 000 € ci-dessus). Ces cartes mères permettent d'utiliser jusqu'à 1 ou 2 To de mémoire et peuvent être équipées de jusqu'à quatre GPU très coûteux.

Un budget de 20 000 € permettrait d'utiliser certains des modèles à poids ouverts les plus performants (tel qu'un 671B DeepSeek V3 quantifié en 8 bits ou un Qwen3-235B-A22B, même lentement)[22] ou de partager un GPT-OSS-120B bien équilibré simultanément avec plusieurs utilisateurs, en utilisant de grandes fenêtres de contexte. Le point important est que les capacités de compréhension du langage de ces modèles dépasseraient même celles du GPT-4o de 2024 et qu'ils peuvent être exécutés facilement[23].

Cependant, les machines dont le prix dépasse les 20 000 euros sont généralement des appareils spécifiques et dédiés, tels que le DGX H100 de Nvidia (environ 350 000 euros chacun) ou le GB300 NVL72 (coûtant jusqu'à 3 millions d'euros). Ce dernier nécessite une alimentation électrique et des capacités de refroidissement spécialisées qui rendraient difficile son installation dans un centre de données voisin (même si l'on pouvait se permettre le matériel et les frais d'électricité, qui peuvent atteindre des dizaines de milliers d'euros par mois). Si l'on a véritablement besoin d'un tel niveau de capacité, l'installer dans ses propres bureaux ne serait plus une solution pratique.

Tableau 4 : Exemples de modèles adaptés à certains cas d'utilisation locale de l'intelligence artificielle

Tâches d'intelligence artificielle locales	Type de modèle	Caractéristiques minimales pour être pratique
Rédaction et révision (adapté aux modifications simples dans les langues prises en charge)	LLM 7B–8B	CPU suffisant
RAG à contexte long (200-500 pages)	LLM 13B–34B + embeddings	GPU ≥16–24 Go ou CPU rapide avec INT4
OCR PDF + légendes	VLM 8B–14B + OCR	GPU ≥12 Go

[22] Cependant, le plus grand modèle à poids ouvert libre actuellement disponible, Kimi K2 Thinking of 1000B parameters, ne rentrera pas dans un tel budget.
[23] https://huggingface.co/spaces/TIGER-Lab/MMLU-Pro

Tâches d'intelligence artificielle locales	Type de modèle	Caractéristiques minimales pour être pratique
Transcription (ASR)	Base ASR grande (<1,5 Go)	CPU suffisant
Résumé par lots (100 documents)	LLM 7 milliards à 13 milliards	GPU ≥ 12 Go
Extraction de clauses/Assurance qualité des contrats	LLM 7B–13B + embeddings	CPU suffisant (GPU recommandé)

Évolutions à venir

Les technologies de l'intelligence artificielle progressent à un rythme rapide et de nouvelles évolutions apparaissent continuellement. L'un de ces domaines est celui de l'intelligence artificielle agentique, qui désigne des systèmes autonomes capables d'exécuter des tâches en plusieurs étapes, d'accéder à des sources de données externes et d'agir au nom des utilisateurs avec une intervention humaine limitée, voire inexistante. Bien que ces technologies ne soient pas encore suffisamment matures pour que le CCBE puisse évaluer leurs effets spécifiques sur les services juridiques et les obligations professionnelles, leur accessibilité croissante signifie que les avocats y seront inévitablement confrontés, que ce soit dans leur propre pratique ou dans les affaires qu'ils traitent. Les principes généraux énoncés dans le présent guide, tels que les considérations relatives à la protection de la confidentialité, au contrôle des données et aux risques inhérents aux services dans le nuage et aux services de tiers, offrent un cadre pertinent pour évaluer ces technologies et d'autres technologies émergentes en matière d'intelligence artificielle. Le CCBE continuera à suivre ces évolutions et a l'intention de les aborder dans les futures mises à jour du présent guide.

Glossaire technique

Modèle d'IA	Un modèle d'intelligence artificielle est un programme entraîné sur un ensemble de données pour reconnaître certains schémas ou prendre des décisions sans intervention humaine supplémentaire. Les modèles d'IA appliquent différents algorithmes aux données pertinentes afin d'accomplir les tâches ou de produire les résultats pour lesquels ils ont été conçus.
Système d'IA	Un système d'IA est un système basé sur une machine capable, pour des objectifs explicites ou implicites, de déduire à partir de données d'entrée comment générer des sorties telles que des prédictions, du contenu, des recommandations ou des décisions.
Interface de programmation d'application (API)	Un ensemble de commandes programmables permettant à un logiciel externe (par exemple une plateforme LegalTech) d'interagir avec un modèle d'IA hébergé ailleurs.
Reconnaissance automatique de la parole (ASR)	La reconnaissance automatique de la parole est une technologie convertissant la langue parlée en texte, permettant ainsi aux ordinateurs de comprendre et de traiter la parole humaine.
Mécanisme d'attention	Le mécanisme d'attention permet au modèle de déterminer quels mots d'entrée sont importants et comment chaque mot se relie à tous les autres. Cela explique pourquoi ces modèles peuvent générer du texte cohérent et suivre des instructions complexes.
Unité centrale de traitement (CPU)	Le processeur central d'un ordinateur, qui exécute le système d'exploitation et définit la logique séquentielle des programmes. Il gère les ressources, coordonne les données entre le stockage, la mémoire cache, la RAM et les GPU.
Plongement vectoriel (*embedding*)	Les plongements vectoriels (*embeddings*) sont des représentations numériques (vecteurs) de données telles que du texte, des images ou de l'audio. Ils capturent le sens (la sémantique) et les relations, transformant une information complexe en vecteurs continus plus petits et exploitables.
Fine-tuning	L'ajustement fin d'un modèle pré-entraîné sur un ensemble de données plus restreint et spécifique (par exemple des contrats juridiques) afin qu'il produise des résultats plus pertinents.
Unité de traitement graphique (GPU)	Un matériel spécialisé dans le traitement parallèle. À l'origine destiné au rendu graphique rapide, il est aujourd'hui utilisé pour accélérer des tâches intensives comme l'entraînement et l'inférence de grands réseaux neuronaux.
Infrastructure en tant que service (IaaS)	Un modèle d'informatique en nuage fournissant des ressources informatiques virtualisées (serveurs, stockage, réseau) à la demande, sans que les organisations n'aient à gérer des centres de données physiques.
Inférence	La phase durant laquelle un modèle d'IA déjà entraîné utilise les connaissances acquises pour analyser de nouvelles données et produire des résultats (prédictions, classifications, décisions).
Grand modèle de langage (LLM)	Un type de modèle d'IA entraîné sur d'immenses ensembles de données à l'aide de réseaux neuronaux profonds (transformateurs), capable de comprendre et générer du texte de manière similaire à un humain.
Carte mère	Le circuit imprimé principal d'un ordinateur, constituant son ossature physique et électrique.
Traitement du langage naturel (NLP)	Un domaine de l'IA permettant aux ordinateurs de comprendre, interpréter, générer et manipuler la langue humaine, écrite ou orale.
Modèles à poids ouverts	Des modèles dont les poids (paramètres entraînés) sont accessibles publiquement, permettant à d'autres de les utiliser ou de les affiner.

Reconnaissance optique de caractères (OCR)	Une technologie permettant d'identifier et de convertir du texte présent dans des images ou des documents scannés en texte éditable et exploitable.
Paramètre	Un poids numérique interne au modèle influençant la manière dont il transforme les entrées en sorties. Les LLM modernes comptent des milliards de paramètres.
Prompt	Le texte ou l'instruction fournie à un LLM pour orienter sa réponse.
Ingénierie de prompt	L'art de formuler des prompts pour obtenir des réponses fiables et conformes, par exemple en ajoutant des contraintes ou des exemples.
Quantification	Une technique visant à réduire la taille d'un modèle et sa consommation mémoire en stockant ses poids avec une précision moindre (par exemple INT8, INT4 au lieu de FP16/FP32), au prix éventuel d'une faible perte de précision.
Mémoire vive (RAM)	Mémoire rapide et volatile stockant les programmes et données en cours d'utilisation par le CPU.
Génération augmentée par récupération (RAG)	Une technique combinant un modèle de langage avec une base documentaire pour permettre au modèle de s'appuyer sur des sources réelles.
SaaS	Un modèle de distribution logicielle dans lequel les applications sont hébergées par un fournisseur et accessibles en ligne via abonnement.
Stack	L'ensemble des technologies, frameworks et outils constituant l'architecture d'une application d'IA.
Token	L'unité fondamentale de texte traitée par les modèles de langage (mots, sous-mots, caractères).
Tokenisation	Le processus consistant à découper du texte brut en tokens.
tps	Tokens par seconde.
Transformeurs / Modèles transformeurs	Des architectures de réseaux neuronaux analysant simultanément tous les mots d'un texte, permettant une compréhension avancée du contexte.
VRAM	La mémoire vidéo utilisée par les GPU.
Poids	Les valeurs numériques apprises par un modèle d'IA lors de son entraînement et définissant son fonctionnement.