Fidal IA, évolutions & usages après un an.

Compte rendu rédigé par ANDSI & Pierre Delort

En bref…

Stéphane MAROTTO, DSI de Fidal, a proposé aux membres un point sur les évolutions et usages de Fidal IA, après plus d'un an d'utilisation.

Approche générale

Le projet Fidal IA a été initié au moment du lancement de ChatGPT, dans le but d'exploiter le potentiel de traitement de texte via les modèles d'apprentissage offert par l'IA. Cet outil est destiné à servir une population extrêmement exigeante de 2 000 utilisateurs, principalement des avocats. Le choix a été fait d'une interface la plus simple et la plus ergonomique possible pour limiter le temps de formation et de déploiement de l'outil.

L'IA dans le domaine du juridique

L'IA représente un changement majeur pour des avocats qui travaillent essentiellement sur une base documentaire. La généralisation de son usage a rendu indispensable la mise en place rapide d'un outil interne performant. Avec la première version de Fidal IA, les avocats ont commencé par réaliser des tâches très simples. Un certain nombre de prompts ont été mis à disposition dans l'outil, les plus utilisés d'entre eux étant celui destiné à la rédaction de mails commerciaux. Fidal n'a pas opté pour une solution Copilot généralisée à l'origine, pour des raisons de coûts, bien que cet outil soit utilisé sur Outlook ou PowerPoint par certaines équipes, ou pour des tâches simples s'il permet d'éviter le recours à des ChatGPT gratuits.

L'IA générative est aujourd'hui très utilisée chez Fidal pour la partie marketing et communication, notamment pour la génération d'images. Le marché de l'IA dans le domaine du juridique est très animé, avec un grand nombre de start-ups proposant différents outils, notamment de justice *prédictive*, dont les résultats sont assez probants. Fidal a testé l'outil, mais ne l'a pas adopté, préférant se fonder sur son expérience. Certaines applications sont utilisées, par exemple pour la génération de contrats, et pèsent très fortement sur un budget DSI. La mise en place de Fidal IA s'avère moins coûteuse que les abonnements aux outils d'IA juridiques classiques qui font référence sur la place. En outre, les start-ups du domaine juridique proposaient jusqu'ici différents services assurés par in fine par des avocats. L'émergence de l'IA générative dans ce secteur va profondément changer la donne et menace directement certaines activités.

L'autre clientèle de ces outils à base d'IA est celle des directions juridiques des entreprises. Même si l'intervention d'un avocat est un passage obligé *in fine,* 71 % d'entre elles utilisent déjà l'IA générative pour des recherches juridiques et diverses analyses, et 33 % pour la gestion de contrats. Ce recours challenge fortement les équipes de Fidal qui doivent maintenir leur haut niveau. Il implique aussi un autre changement profond : les avocats seniors les plus ouverts à la technologie utilisent l'IA générative avec une certaine

efficacité. Cette évolution pose question en terme de formation des jeunes, formation à laquelle Fidal est attaché. Une future version de Fidal IA devrait d'ailleurs être dotée d'une dimension coaching, afin d'aider les jeunes avocats à monter en compétences sans dépendre d'avocats seniors.

Transition vers l'IA agentique

La première version de Fidal IA a été entraînée sur des données internes et sur certaines spécialités, avec quelques centaines de milliers de documents. Avec l'ajout de documents open data, cette base est montée à plus d'un million et demi de documents, mais bien que davantage de domaines soient couverts, la qualité des réponses a diminué, en raison d'un espace de choix bien trop vaste. Dans la version 2, le RAG (Retrieval Augmented Generation) s'appuie sur 8 millions de documents qu'un agent IA est capable de contextualiser, pour définir la zone de recherche, les spécialités concernées et ne sélectionner que les documents pertinents.

Int : Au-delà de 50 000 documents de référence, il faut contextualiser au maximum les documents et joindre de la métadonnée utile au moment du chargement.
SM : Les métadonnées sont un domaine que nous explorons également. Sur certains sites d'open data juridiques, il est possible de récupérer des données et de les vectoriser efficacement grâce à un agent. Des solutions comme Pappers peuvent s'avérer pertinentes pour accéder à ce type de documents tagués.
Int : Nous utilisons l'IA pour analyser les documents et générer des métadonnées, avant chargement.
SM : Nous *chunkons* les documents standards, l'enjeu étant ensuite de conserver la cohérence des documents découpés.
Int : Le coût de la puissance de calcul nécessaire au traitement de millions de documents est-il un frein ?
SM : Le coût du recours à nos modèles est une bonne surprise, car il reste raisonnable.

Caractéristiques de la nouvelle version de Fidal IA

La nouvelle version de Fidal IA s'appuie sur GPT 5-Mini, tout en conservant GPT 4-1 et Mistral Large, fonctionnant sur Azure, qui ne permet pas d'accéder aux pleines capacités de ces modèles en termes de nombre de tokens traités.

Fidal IA V.2 propose un espace Sharepoint sur lequel les avocats déposent leurs documents pour alimenter le système, ce qui suppose anonymiser préalablement ces derniers à l'aide d'un module d'IA. La nouvelle version permet de vérifier les références fournies par l'outil par un simple clic. L'ergonomie générale a été améliorée. Il offre la possibilité de traiter des documents directement sous Word, Excel ou PowerPoint.

Fidal IA propose une large base de prompts. Les avocats qui le souhaitent peuvent partager leurs meilleurs prompts afin qu'ils bénéficient à tous après vérification par la DSI. L'interface permet de choisir les sources, le modèle (ChatGPT ou Mistral), d'accéder à un guide d'utilisation, à des bibliothèques de prompts, de créer des collections de documents pour les travaux en cours. Des explications sont fournies par des pop-up au passage de la souris sur chaque bouton.

Les prompts sont classés par types de droit. Une option permet l'analyse de transcriptions de réunions. L'option « traitement » permet de traduire, reformuler et anonymiser les documents, pour une importation ensuite dans le format souhaité. L'outil propose des traductions en 7 langues, différents types de résumés, des revues comparatives.

Différents cas d'usage ainsi qu'un certain nombre de prompts sont présentés.

Int : Les prompts sont-ils toujours aussi longs ? Quelle formation ont reçue les avocats pour les rédiger ?
SM : Ce sont des prompts souvent « costauds » émanant des avocats eux-mêmes. On a pris du temps pour les former.
Int : Combien de temps prend la génération de la réponse à un prompt de plus d'un page ?
SM : La réponse est générée en une minute, d'un bloc.
Int : De notre côté, nous sommes passés d'une minute à quelques secondes avec affichage progressif de la réponse.
SM : Initialement, le temps de réponse était de 10 minutes sur la V2. Il a été nettement réduit, au prix d'une très légère diminution en qualité.
Int : Le RAG adresse-t-il l'analyse vectorielle au LLM de manière confidentielle ?
SM : L'information reste cantonnée à notre *tenant* et les données sont anonymisées. Les instances hébergées par Microsoft sont privées et leur contexte n'est pas partagé. Toutefois, les CGU des fournisseurs d'IA révèlent qu'ils se réservent le droit d'enregistrer les instances en cas d'utilisation de certaines formulations ou certains mots. Fidal a contesté ces conditions, les documents liés aux affaires sur lesquelles les avocats interviennent étant truffés de mots de ce type !
Int : Meta interdit l'utilisation de ses modèles dans des domaines comme le nucléaire, ce qui démontre qu'ils sont capables de vérifier le contenu des requêtes.
Int : Comment sont stockés les documents ?
SM : Nos documents confidentiels restent sur nos propres data centers et ne sont pas traités par Fidal IA. Les documents générés disparaissent après chaque session, seules les collections sont conservées le temps nécessaire.
Int : Qui se charge du *chunking* et de la vectorisation documentaire ?
SM : Nous avons sous-traité cette partie à Sopra Steria. Le *chunking* est automatisé.

Mise en place d'un monitoring

La nouvelle version de Fidal IA permet d'en monitorer l'utilisation. Environ 1 100 personnes différentes l'utilisent chaque mois en interne, et 300 à 400 chaque jour. Les assistantes juridiques l'utilisent au même titre que les avocats. Les outils mis en œuvre pour ce monitoring sont divers : Matomo (analyse de trafic), Grafana, Open Telemetry, PostGre Vector. Avec cette version, le RAG dit modulaire a laissé place à un système agentique, possédant une véritable capacité d'analyse. L'écart qualitatif de cette nouvelle version étant considérable, les usages se sont fortement développés, ce dont les KPIs attestent. Les réponses fournies sont systématiquement évaluées par les avocats, ce qui permet d'apporter des améliorations au système.

Int : La prédictibilité de l'outil est-elle bonne ? Observez-vous des divergences de résultats pour des requêtes identiques ?
SM : Les requêtes ne sont jamais exactement identiques. Aucun avocat mécontent d'une réponse n'a signalé avoir formulé la même requête et obtenu une réponse différente à notre connaissance.

Perspectives pour 2026

Aujourd'hui, Copilot de Microsoft s'est nettement amélioré et peut effectuer des recherches tout à fait pertinentes sur Internet. Par ailleurs, nos équipes de développement utilisent Claude pour coder sur les applications historiques, et les résultats obtenus en termes de génération de code s'avèrent spectaculaires.

PD : Le DSI de Hachette Livre nous a présenté récemment un cas d'utilisation de l'IA pour refondre son système informatique historique (10 millions de lignes de code en plus de 10 langages de programmation).
SM : Je n'en suis pas étonné. Il est amusant de constater que ce sont les développeurs seniors qui ont adopté plus rapidement l'IA, au même titre que nos avocats seniors, sûrement parce qu'ils sont plus à même de juger immédiatement de la pertinence de ce qui est produit.

Dans des domaines comme la santé ou de la finance, l'apport de l'IA est considérable. Il en est de même dans le domaine juridique. Plusieurs fiscalistes de Fidal jugent les calculs produits par Fidal IA très pertinents, tout comme des avocats spécialisés dans le social pour le calcul par exemple des indemnités de licenciement. L'IA est aussi appelée à jouer un rôle de plus en plus important dans l'industrie de fabrication.

Int : On parle ici d'IA générative
SM : Dans l'industrie ou la santé, il s'agit de modèles de machine learning. Pour le reste, les modèles de LLM ont atteint certaines limites et l'IA générative et *a fortiori* agentique ouvre de nouvelles perspectives.

Il est important d'évoquer également le coût énergétique du recours à l'IA, en rappelant que les nouvelles NPU embarquées sur les ordinateurs ou les mobiles peuvent souvent suffire à faire tourner certains modèles locaux sans passer systématiquement par un serveur. Il est probable que le recours aux très grands modèles ne soit plus aussi incontournable à l'avenir.

Int : Apple Intelligence tourne déjà en local sur les Macs.
SM : Il existe aussi des systèmes de traduction embarqués utilisant les NPU.

Débat

Int : L'Europe entend privilégier la frugalité énergétique, grâce au recours à des modèles de base re-spécialisés pour consommer le moins de GPU possible.
Int : Le recours à l'algorithmie permet de contourner l'*overkill* des LLM pour des tâches simples.
SM : Il existe des outils d'anonymisation hors LLM à base d'algorithme depuis des années. Mais ces solutions sont coûteuses par rapport à ce que permet une brique IA.

Int : Monitorez-vous le coût de l'ensemble du système ? Le fait de passer par un agent peut générer plus d'allers-retours avec une maîtrise budgétaire moindre.
SM : Nous le monitorons de façon permanente, ce qui donne lieu à un copil mensuel. Certes les allers-retours sont plus nombreux, mais le coût par *token* a baissé. L'inférence nous coûte aujourd'hui 3 à 4000 euros par mois.

Int : Les prompts sont-ils envoyés indépendamment du modèle ?
SM : Tout à fait. Les avocats ont effectué des comparaisons, les différences selon les modèles sont faibles.

Int : Quid de l'usage de Copilot ?
SM : Copilot Chat est inclus dans notre *tenant* sécurisé. Le but est de ne pas avoir à utiliser la version grand public de ChatGPT. Nous utilisons trois modèles, OpenAI, Mistral et Copilot et tous attaquent le même RAG. C'est un choix qui est dicté par une précaution en cas d'envol des tarifs d'OpenAI, et par la volonté de conserver une solution française.

Int : Maintenir l'inférence sur trois modèles doit avoir un coût…

SM : Non, car les moteurs interrogent à tour de rôle une *tokenisation* et un *embedding* effectués sur la base de Fidal.
Int : La base de données est donc hébergée en interne.

Int : Comment vous assurez-vous de l'anonymisation des données soumises au RAG ?
SM : Nous faisons confiance aux avocats, qui sont extrêmement stricts sur ce point.

Int : Comment avez-vous traité la question de la souveraineté et de la confidentialité de vos propres données avant de choisir Azure ?
SM : En 2023, les solutions étaient rares. Nous avons testé celle-ci qui a satisfait les conditions de confidentialité, moyennant l'anonymisation des documents. LE risque de fuite est pris très au sérieux dans nos métiers.

Int : Etes-vous capable de générer un data set à partir des retours utilisateurs lorsqu'une réponse n'est pas satisfaisante pour le réinjecter dans le RAG ?
SM : Nous n'avons pas pu gérer cet aspect.
Int : Vous n'entraînez donc pas de modèle, puisqu'il s'agit d'inférences via le RAG.
SM : Un tel système serait trop coûteux. Nous avons privilégié une solution moins onéreuse. Nous pourrons tester d'autres solutions à l'avenir si notre budget est étendu, ce que le succès de la V2 de Fidal IA pourrait favoriser

Présentation des orateurs

Stéphane MARIOTTO, ingénieur des Mines & AI Cogmaster ENS, a débuté en IA au CNES puis dirigé plusieurs DSI (assurance, industrie) avant Fidal.