L'utilisation de l'IA pour mieux tirer parti d'une base documentaire technique au bénéfice des industriels de la mécanique

Par Fanny LAMBERT
Responsable de la Veille sur les IA génératives de Cetim

Pour améliorer l'exploitation des documents techniques au bénéfice des industriels de la mécanique, le Cetim a fait développer un *chatbot* interne capable d'exploiter de manière sécurisée et contrôlée des milliers de documents produits en interne. Cette initiative vise à surmonter les limites des *chatbots* accessibles sur le *web*, notamment en matière de confidentialité et de capacité de traitement.

Le projet, aujourd'hui abouti, utilise les documents de veille du Cetim pour générer des résumés, des synthèses et des comparatifs, facilitant ainsi la collecte et l'analyse rapide d'informations scientifiques et techniques. Ce projet a aussi permis de comprendre cette technologie, d'évaluer les coûts, les infrastructures IT nécessaires, et les compétences requises pour l'intégration d'une IA en interne, nous permettant de le partager avec vous aujourd'hui.

Contexte

Les *chatbots* accessibles sur le *web* peuvent permettre d'exploiter des documents, mais ils comportent leurs limites. En effet, les fichiers que nous leur fournissons peuvent être sauvegardés et exploités pour l'entraînement du prochain modèle du *chatbot* utilisé, en fonction de ses conditions d'utilisation, ce qui peut être problématique d'un point de vue de la propriété intellectuelle et de la confidentialité des données. Autre point limitant : il n'est généralement possible de fournir que quelques documents (5-10 fichiers de taille moyenne), ce qui ne correspond pas forcément au besoin quand nous avons une grande quantité de fichiers à exploiter. Pour répondre à ces problématiques, il peut donc être intéressant de déployer une IA en interne, en mesure d'exploiter plusieurs milliers de documents de manière sécurisée et contrôlée, avec la possibilité de paramétrer l'IA en fonction du type de document exploité.

C'est cette réflexion que nous avons eue au Cetim (Centre technique des industries mécaniques) et qui nous a poussés à nous lancer dans un projet de développement d'un *chatbot* interne, exploitant des documents produits par le Cetim. Pour réaliser une première preuve de concept (POC), nous avons choisi d'exploiter les documents de veille que nous produisons et mettons à disposition de nos industriels sur notre extranet, la Mécathèque[1]. Ces quelques milliers de documents (plus de 6 500 à ce jour) contiennent des informations techniques et scientifiques sur de nombreux procédés et technologies de l'industrie mécanique. Les objectifs de ce projet étaient multiples :

- Permettre à nos cotisants et aux cétimiens d'exploiter plus finement et plus rapidement cette masse d'information conséquente pour de la collecte d'informations, l'éclairage de prise de décision ou encore l'obtention d'une synthèse de l'information. L'IA serait ainsi une sorte de moteur de recherche amélioré, complémentaire au moteur de recherche classique, capable de fournir le résumé d'un document, des synthèses de plusieurs documents avec les liens vers les sources, ou encore des tableaux comparatifs ou SWOT (pour *strengths, weaknesses, opportunities* et *threats*).
- Vivre un projet d'intégration d'IA en interne, voir quelle est la démarche pour le mettre en œuvre, et comprendre le fonctionnement de cette technologie. Nous voulions connaître les possibilités d'un tel outil, les infrastructures IT et les compétences à prévoir, évaluer les coûts et délais d'un tel projet, mais aussi être conscients des limites d'un *chatbot* en interne.
- Pouvoir partager ces expériences avec nos cotisants à travers des notes de veille, webinaires et interventions.

Déroulement du projet

La première phase du projet a consisté à identifier et à sélectionner, parmi une douzaine de prestataires, quatre partenaires les plus adaptés aux besoins spécifiques du

[1] https://www.cetim.fr/mecatheque/Toute-la-richesse-des-etudes

projet. Cette sélection a donné lieu à un appel d'offres auprès de neuf sociétés, révélant une fourchette de budgets allant de 7 000 à 80 000 euros pour réaliser la phase 1. Les quatre prestataires retenus ont été mis en concurrence et se répartissent en deux catégories : deux intégrateurs de solutions « clé en main » (en l'occurrence Ekimetrics et The QA Company), et deux développeurs spécialisés en data science (Hurence et Cross Data). De nombreux tests rigoureux ont été effectués pour chacune des quatre solutions proposées. La seconde phase est quant à elle axée sur l'amélioration des performances du *chatbot* avec les deux prestataires que nous avons préférés (Hurence et QA Company), notamment en termes de pertinence des réponses, ainsi que sur l'optimisation de son ergonomie. Cette dernière phase visait également à préparer l'industrialisation du *chatbot* avec Hurence, que nous avons finalement retenu, afin de le déployer à grande échelle au sein de l'entreprise et de le mettre à disposition de nos cotisants.

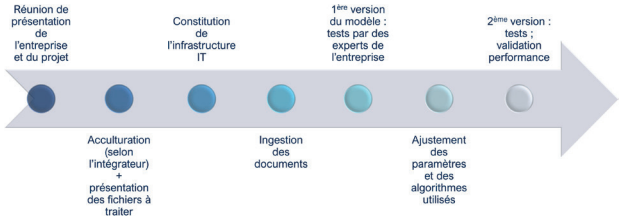

Figure1 : Les étapes d'un projet de développement d'un *chatbot* interne (Source : Auteure).

Fonctionnement d'un modèle de langage

Les IA conversationnelles comme ChatGPT, Mistral ou Gemini par exemple sont ce qu'on appelle des grands modèles de langage, ou LLM. Ce sont des réseaux de neurones entraînés en auto-supervision sur une très grande quantité de textes. Grossièrement, cela signifie que le modèle essaye de compléter des séquences de mots tronquées issues de son corpus d'entraînement, en calculant la probabilité que chaque mot de son vocabulaire soit celui attendu dans la phrase tronquée.

Par exemple, avec la phrase « le ciel est bleu », on va retirer le mot bleu et le modèle va essayer de retrouver la bonne proposition. Il essaye le mot qui lui semble le plus probable, il le compare ensuite avec le mot attendu, puis il met à jour ses probabilités en fonction du résultat. Il va répéter ainsi cette opération un très grand nombre de fois pour en déduire une régularité statistique, et c'est comme cela qu'il va développer ses « connaissances générales » (voir la Figure 2). Actuellement, un modèle de langage n'a donc pas de raisonnement à part entière, il fonctionne uniquement sur des statistiques pour construire ses réponses. La qualité du corpus fourni lors de l'entraînement va dès lors être déterminante dans la pertinence du modèle développé, puisque si vous lui fournissez des documents contenant des infox pour s'entraîner, cela perturbera ses statistiques et induira davantage de mauvaises réponses. De même, de nombreux biais risquent d'être transmis au modèle au cours de son entraînement. Pour vous donner un ordre d'idées, l'entraînement d'un modèle tel que GPT4 dure entre 28 et plus de 300 jours selon le nombre de machines réquisitionnées. Cette procédure est donc très coûteuse financièrement et énergétiquement.

Une notion clé dans le fonctionnement des modèles de langage est la vectorisation, également appelée *embedding*. Pour que le modèle comprenne le sens des mots et des phrases, chaque mot se voit attribuer différentes valeurs qui le représentent : s'agit-il d'un verbe, quel est son genre, est-il pluriel ou singulier, est-ce un humain, un animal, une plante, etc. Toutes ces valeurs sont représentées sous forme de vecteurs, ce qui permet au modèle de déterminer si deux mots sont similaires (comme « chat » et « chaton ») ou totalement différents (comme « chien » et « maison ») (voir la Figure 3). C'est ainsi que les relations sémantiques sont établies, permettant au modèle d'exploiter plus efficacement les données.

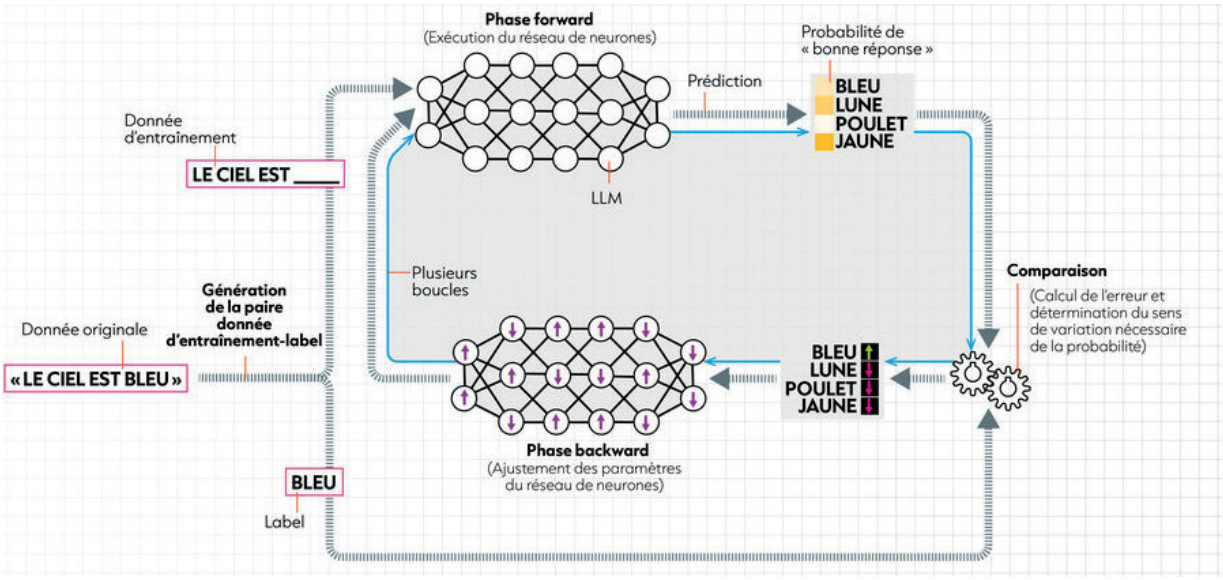

Figure 2 : Représentation schématique de l'entrainement d'un LLM par auto-supervision (Sources : Plongée dans les entrailles des grands modèles de langage qui font l'IA conversationnelle, *L'Usine Nouvelle* et Université de Genève).

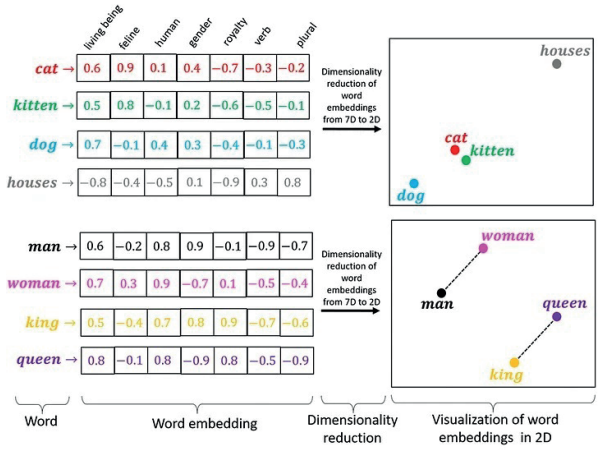

Figure 3 : Schématisation de la vectorisation des mots (Source : Medium).

Une fois ces bases posées, nous pouvons revenir à notre projet de *chatbot*. La technique retenue pour exploiter nos documents à l'aide d'une IA est appelée « RAG » (*Retrieval Augmented Generation*). En effet, cette méthode permet de connecter un LLM à une base de connaissances choisie, telle que des documents internes, une base de données ou autre, et confère ainsi des connaissances pointues au modèle sans qu'il soit nécessaire de le réentraîner. Cette méthode est assez économique et réduit les risques de réponses inventées ou fausses (que l'on appelle « hallucinations » dans le jargon). Un comparatif de cette technique par rapport à une IA utilisant uniquement ses connaissances liées à son entrainement est présenté dans le tableau ci-après.

La première étape du RAG consiste à découper vos documents en fragments (paragraphes, phrases…) appelés *chunks* par les développeurs. Ces *chunks* sont ensuite vectorisés et stockés. Lorsqu'une requête (*prompt*) est effectuée auprès de votre *chatbot*, celle-ci est également vectorisée pour être comprise par le LLM. Ensuite, les fragments de documents les plus pertinents sont recherchés dans la base de données. Enfin, ces *chunks* pré-sélectionnés sont analysés par le LLM pour générer une réponse (voir la Figure 4).

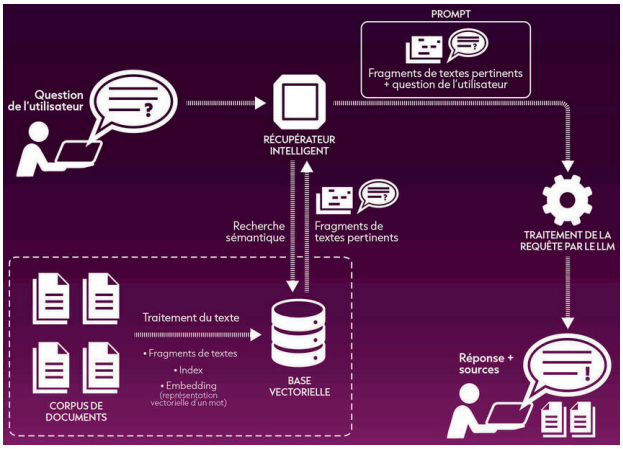

Figure 4 : Représentation simplifiée du fonctionnement du RAG (Source : L'intelligence artificielle générative se diffuse dans l'entreprise grâce au RAG).

Bien entendu, pour faire tourner ce type de technologie, vous aurez besoin d'infrastructures solides

Critère	LLM utilisant ses connaissances générales	LLM connecté à un RAG
Source des infos	Basée sur des données pré-entraînées	Accès en temps réel à des bases de données / API
Actualité des données	Fixe à la date d'entraînement	Infos à jour en temps réel
Questions spécifiques	Réponses parfois approximatives, générales	Recherche de documents précis
Contexte	Basé sur le modèle seul	Appui sur des documents externes
Données évolutives	Réentraînement nécessaire	Infos récentes sans réentraînement
Fiabilité des faits	Bonne pour les sujets courants	Plus fiable grâce aux sources récupérées
Ressources	Pas de mémoire externe requise	Nécessite des infrastructures supplémentaires
Vitesse de réponse	Très rapide	Plus lente mais acceptable (recherche externe)
Explications / Citations	Impossible de connaitre la source utilisée	Peut fournir des sources et citations
Utilisation typique	Assistants virtuels, réponses générales	FAQ dynamiques, expertises techniques nécessitant des informations actualisées

incluant des GPU (cartes graphiques) professionnels. Pour cela, deux options s'offrent à vous :

- Soit vous achetez votre matériel (*hardware*) pour l'installer dans vos locaux. Avec les prix actuels, vous en aurez pour environ 50 000 € pour 2 GPU. Vous pouvez également louer le matériel que vous installez chez vous pour 1 000-2 000 €/mois. Les besoins GPU varient en fonction du nombre d'utilisateurs, mais aussi en fonction du modèle IA choisi. En effet, un modèle comme GPT4 va consommer beaucoup plus de ressources qu'un modèle de chez Mistral par exemple.
- Soit vous installez votre IA sur un *cloud*, c'est-à-dire que vous allez réserver une petite partie d'un *data center* pour vos usages (abonnement classique ou à l'usage). Il vous faudra choisir entre un *cloud* souverain (français ou européen) comme OVH, Scaleway ou vast.ai par exemple, et un *cloud* américain comme Azure (Microsoft), AWS ou Google Cloud Platform, soumis quant à eux à l'US Patriot Act[2].

Résultats

Le développement du *chatbot* destiné à exploiter les ressources de la Mécathèque a permis de dégager plusieurs enseignements, tant sur la qualité des réponses fournies que sur les ressources nécessaires à un tel projet.

Les performances de celui-ci ont été évaluées par les dix membres de l'équipe de veille technologique, auteurs des documents utilisés. Cette évaluation portait sur un échantillon varié de questions techniques posées au *chatbot*, couvrant les thématiques majeures de la Mécathèque. Une cinquantaine de tests ont été réalisés pour chaque version proposée par les prestataires.

Les résultats montrent qu'à la fin des tests, 55 % des réponses étaient jugées bonnes, c'est-à-dire complètes et techniquement exactes, 33 % des réponses étaient moyennes, souvent correctes mais incomplètes, nécessitant des recherches supplémentaires par l'utilisateur, et 12 % des réponses étaient mauvaises, comprenant des erreurs ou des hors-sujets.

Les erreurs étaient principalement dues à une mauvaise interprétation des termes techniques ou à une confusion dans le traitement de documents complexes (mélange d'informations par exemple), mais ne correspondaient pas à des hallucinations.

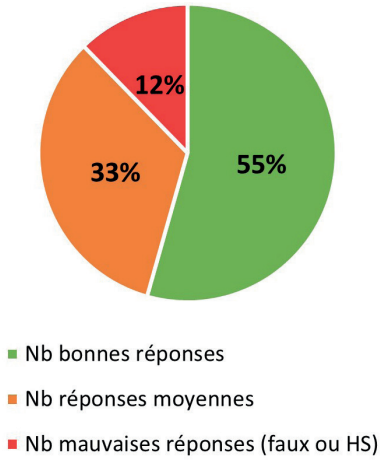

Figure 5 : Répartition de la qualité des réponses dans la dernière version du *chatbot*, à ce jour (Source : Auteure).

Pour atteindre ces résultats, le projet a mobilisé :

- une équipe de testeurs avec une bonne connaissance des documents, permettant d'évaluer plus aisément la pertinence des réponses ;
- une implication importante des équipes SI, pour le développement et l'intégration du *chatbot* ;
- plusieurs mois de travail, pour l'entraînement de l'IA générative, la structuration des données et l'amélioration continue des réponses ;
- un budget estimé entre 40 et 70 k€ pour financer un seul prestataire jusqu'à la fin de la phase 2, sans compter le temps de travail des équipes internes et les coûts d'infrastructure.

Cette expérience nous a permis d'acquérir une meilleure compréhension du fonctionnement des IA génératives, notamment concernant l'entraînement, la vectorisation et le RAG détaillés précédemment. Nous avons également réalisé à travers nos tests l'importance de la qualité et la bonne structuration (titres, tableaux, légendes, images) des documents fournis dans le RAG pour faciliter leur exploitation par l'IA. Il faut d'ailleurs prendre en compte que les modèles de base ne sont pas capables de lire et analyser les images. Ce type de fonctionnalité peut être ajouté, mais représente un coût supplémentaire.

Cependant, le langage technique spécifique à l'industrie mécanique reste un défi, nécessitant une amélioration continue de l'IA pour mieux gérer le jargon et les subtilités des documents de veille.

[2] https://shs.cairn.info/dictionnaire-du-renseignement--9782262070564-page-592?lang=fr