AfIA

Association française
pour l'Intelligence Artificielle

Collège industriel

Dates	Noms des auteurs et contributeurs	Actions sur le document	Versions
04/12/2023	Valérie Reiner – Berger-Levrault	Auteur du draft	V0.1
	Bruno Carron - Airbus	Contributeur	
	Alain Berger - Ardans	Contributeur	
	Patrick Fabiani – Dassault Aviation	Contributeur	
	Mustapha Derras – Berger-Levrault	Contributeur	
02/02/2024	Christophe Bortolaso – Berger-Levrault	Contributeur	V0.2
	Pierre Feillet – IBM France	Contributeur	
	Elise Moris - Ardans	Contributeur	

Un petit texte collectif

Lors de la réunion du Collège Industriel du 20 novembre 2023, l'idée a été émise de rédiger collectivement un texte sur un sujet d'actualité dont pouvait se saisir tout particulièrement une société savante comme l'AFIA, sorte de point de vue collégial pris avec recul dans une approche raisonnable du sujet et non une vision plongeante.
Le texte n'a pas vocation aujourd'hui à être diffusé en dehors de la communauté de l'AFIA, en revanche ses auteurs ont clairement la volonté qu'il soit partagé avec les collèges académiques de l'association en une sorte de main tendue vers des échanges académiques-industriels. Il est également prévu d'en débattre à l'occasion du prochain FIIA.

Le sujet d'actualité

Le discours médiatique sur les modèles de langage à grande échelle (LLM) et l'intelligence artificielle générative, marqué par une profusion d'opinions, tend souvent à osciller entre des scénarios paranoïaques et des utopies irréalistes, notamment en ce qui concerne leur impact présumé sur notre vie quotidienne, la société, le travail et les métiers. Pour échapper à ces visions extrêmes, il convient de développer une compréhension du sujet qui soit à la fois réfléchie, fondée et orientée positivement.

De quoi ont besoin les industriels ?

Les industriels souhaitent profiter de la puissance et de la souplesse des moteurs de LLM qui embarquent des connaissances implicites inhérentes à tout travail de rédaction et en même temps, leur adjoindre des contraintes explicites qui encadrent les résultats générés avec une garantie mesurable. Les industriels veulent donc contraindre les réponses aux éléments contenus dans une base de connaissance mais aussi comprendre, mesurer, calculer la part d'invariant, somme toute, contrôler les comportements d'un LLM. Les LLM sont des architectures complexes de réseaux de neurones pré-entraînés. Pour acquérir une bonne capacité de génération de langage, les modèles exploitent l'apprentissage statistique sur de gigantesques quantités de données. Ils sont ensuite spécialisés par raffinement (Transformers fine tuning) ou intégration de tables d'informations de référence (Retrieval Augmented Generation) sur des corpus particuliers. Une spécificité des LLM est que leur prise en main par les usagers en tous domaines est très rapide et comme beaucoup de domaines de l'Intelligence Artificielle, les technologies évoluent très vite. Comment démêler, hinc et nunc, ce qui est recommandé ou ce qui est mauvais dans un contexte industriel ?

Les RAG : Un exemple d'hybridation d'IA à l'échelle industrielle

La simplicité d'emploi des LLM représente une opportunité et un écueil, et il est donc recommandé de faire appel à d'autres formes d'intelligence et d'expertise en complément des LLM. Cela peut être géré par une hybridation avec d'autres formes d'intelligence artificielle, symbolique ou structurée. Cela peut être aussi mis en œuvre par des processus et règles d'emploi organisés autour de l'humain pour permettre la bonne vérification, validation experte et décision dans les chaînes normales de responsabilité.

Les RAG (Retrieval-Augmented Generation), dans le contexte des LLM, se réfèrent à une architecture hybride qui combine un modèle de génération de langage, avec un système de récupération d'informations. Cette combinaison permet au modèle de puiser dans une base de données externe ou un ensemble de documents pour enrichir ses réponses. Le raffinement demande un réapprentissage qui peut être long là où la technologie RAG demande l'intégration d'informations bien tabulées, structurées et mises à jour. L'approche RAG étend donc les capacités des LLM traditionnels en leur permettant d'intégrer des informations spécifiques et contextuelles en temps réel, ce qui est particulièrement utile pour des questions nécessitant des connaissances à jour ou très spécifiques. Les RAG représentent une avancée importante car ils combinent la puissance de génération de langage des modèles avec la capacité de récupérer des informations spécifiques et actualisées à partir de sources externes. C'est typiquement le besoin des industriels !

Dans un RAG par exemple, une configuration typique est composée de trois éléments principaux :

- Composant de Récupération (Retrieval) : responsable de la recherche et de la sélection d'informations pertinentes à partir d'une base de données ou d'un ensemble de documents. Des algorithmes spécialisés en recherche d'informations eux-mêmes basés sur l'IA peuvent être utilisés. Ces algorithmes peuvent inclure des moteurs de recherche basés sur le texte, des systèmes de classement basés sur l'apprentissage automatique, ou des réseaux de neurones

conçus pour retrouver et réordonner des informations pertinentes. Les algorithmes et moteurs de recherche aident à automatiser la récupération, mais celle-ci doit être vérifiée d'une manière ou d'une autre (humain, filtres, ...).

- Composant de Génération (Generation) : en charge de l'utilisation d'un LLM pour générer des réponses ou du contenu en utilisant les informations récupérées. Ce composant est souvent un modèle de langage avancé capable de synthétiser les informations et de les présenter de manière cohérente et contextuelle. Le LLM de génération peut être conservé tel que sans réapprentissage tant que sa « performance » reste satisfaisante pour l'utilisation dans le cadre RAG en question. Cela est favorable à une utilisation en interne d'une entreprise ou d'une organisation sans lien avec l'extérieur. Sa ""mise à jour" éventuelle doit être considérée au bout d'un certain temps, mais avec prudence pour ne pas importer de biais ni diffuser d'information confidentielle de façon non désirée.
- Composant d'Augmentation (Augmented) : les choix sont nombreux et vont des systèmes de recommandation pour fournir des suggestions personnalisées exploitant le texte, en passant par des systèmes fondés sur des règles fournissant un cadre structuré pour certaines tâches spécifiques, à des IA plus complexes en charge de préparer les données. Les règles et cadres structurés doivent faire l'objet de validation experte attentive (tests de couverture des cas d'utilisation possible, ...) et les systèmes de recommandation doivent fournir des propositions ajustées au niveau de compétence des utilisateurs qui doivent être formés pour cela.

Nos exigences envers les machines

Les IA génératives donnent « l'illusion de l'intelligence » posant ainsi le problème du décalage entre les erreurs générées et la rigueur attendue de tout dispositif numérique dans un contexte industriel. L'acceptation d'erreurs renvoie à une dimension subjective inhabituelle plus proche des préoccupations des Sciences Humaines et Sociales que des Sciences dites dures. Quid de la rigueur, du déterminisme que l'on attend habituellement d'une machine ?

La question se pose dès lors que les cas d'usages d'exploitations industrielles intègrent cet aspect comme une dimension clé et non comme une difficulté ou une erreur à combattre ! Si cette particularité des IAG n'est pas acceptable dans une mise en œuvre il faut absolument éviter de les utiliser. Sinon, il faut s'attendre à devoir expliquer que :

- Lorsqu'une IAG répond à une question posée plusieurs fois de manières différentes ce n'est pas une erreur !
- Un humain ne sera pas capable de répéter 10 fois de suite une réponse avec les mêmes mots dans le même ordre au même rythme à une question, l'IA se comporte de la même manière, ce n'est pas une erreur !
- Chaque individu va synthétiser un même texte de manière différente et se focaliser sur des dimensions variées, une IA fera de même si une synthèse lui est demandée à répétition, ce n'est pas une erreur !
- Chaque individu va exposer une idée de manière différente exploitant des aspects nombreux de cette dernière, une IA fera de même, ce n'est pas une erreur !

Si vous disposez de moyens d'expertises, ayant une connaissance particulière des architectures, des langages et des univers de l'IA, en matière de réalisation de logiciels, le plus simple est d'essayer d'exploiter des données pour mesurer l'adéquation avec vos besoins.

Cela peut s'avérer parfois délicat s'agissant de données confidentielles ou couvertes par différents niveaux de secret. Attention toutefois à l'absolue nécessité de comprendre qu'il faut « vraiment » mettre en place des pratiques rigoureuses concernant la gestion des données et des compétences des utilisateurs. Il est notamment impératif que les utilisateurs soient formés ou informés des limitations d'usage (domaine opérationnel) de l'outil à base de LLM ou RAG qu'ils sont amenés à utiliser.

Les défis à relever

Quels sont les moyens pour faire cohabiter les larges corpus hétérogènes et les corpus industriels souvent très organisés, ayant fait l'objet de stockage, d'indexation rigoureuse, d'une maintenance méthodique à grands frais sur de longues périodes, véritables patrimoines industriels ? On pense aux grandes bases documentaires ou autres systèmes d'information structurés porteurs de l'expertise voire du secret industriel. Evidemment, il est hors de question à ce jour de déverser ces patrimoines industriels sur les plates-formes de tests de telle ou telle plate-forme d'IA générative.

1. **Vérification des résultats :** De manière générale, de quels moyens disposons-nous pour vérifier les résultats ? Est-il possible de mettre en place des tests automatisés comme cela se pratique pour les logiciels classiques ? Un second besoin est lié à la mesure de la garantie des résultats générés, on pense ici à l'apport de démonstrations mathématiques qui aideraient à contrer les écueils de la variabilité des résultats voire du non-déterminisme si perturbant.
2. **Fiabiliser les LLM:** Concernant les erreurs et autres hallucinations, y a-t-il lieu d'interroger la possibilité de les réduire à défaut de les supprimer ? Si oui dans quelle proportion ?
3. **Vers une ingénierie des LLM :** A défaut de garantie ou vérification, avons-nous des méthodes systématiques qui indiqueraient les bienfaits : d'une ingénierie du prompt, de la taille d'un corpus à usage strict en fine-tuning, de l'hybridation avec une représentation des connaissances, d'injection pré ou post de règles ?
4. **Sortir du conversationnel :** Le dialogue ou le modèle conversationnel avec prompt sont-ils les plus à même de favoriser la convergence de données structurées et de LLM ?
5. **Impact énergétique :** Comment évaluer le coût écologique des LLM ? A partir de quel(s) critère(s) se fonder pour un choix éclairé et consenti de l'usage de cette technologie émergente ?
6. **Langue et impacts culturels :** La dominance de la langue anglaise a t'-elle des effets structurels sur les résultats générés ? Existe-t-il des comparaisons de taux d'erreurs par langues ? Qu'attendre d'un corpus large en anglais et d'un corpus strict en une autre langue ?
7. **LLM et facteurs de formes :** Faudrait-il éviter d'utiliser un vocabulaire trop anthropomorphique pour réduire les confusions, les déceptions ou inversement un niveau trop élevé de confiance et ainsi réduire les écueils d'enjeux culturels ?
8. **Stabilité et Qualité :** Comment entretenir le niveau de qualité des réponses (cf. les phénomènes de baisse de performance de GPT 4 en 2023) ? Comment garantir qu'un simple changement d'embedding ou de prompt n'entraine pas des baisses de qualité significatives ?
9. **Enrichissement permanent :** Comment s'assurer d'intégrer les experts dans la boucle d'amélioration continue des données sans appauvrir leur capacité d'apprentissage ? A contrario, comment gérer les données et connaissances obsolètes, présentes dans les modèles ?
10. **Apprentissage :** Comment assurer la prise en main par des utilisateurs néophytes des bots issus de la mise en œuvre de ces IAG ?
11. **Généricité et indépendance :** Comment se rendre indépendant d'un LLM (modèle) spécifique ? De la nécessité de mettre au point de nouvelles architectures et des plates-formes « LLM

indépendantes » ? Quelles autres IA sont nécessaires pour organiser, découper, curer, « accroitre/augmenter » les données ?

12. **Sécurité :** Comment assurer la sécurité (cyber) de ces plates-formes ? Avec quels tests s'en assurer ? Comment se prémunir des nouvelles techniques de prompt-injection que nous ne savons que mal maitriser à ce jour ?
13. **Le retour des documentalistes :** Comment réhabiliter certains métiers qui produisent les données indispensables à des mises en œuvre de ces LLM (documentaliste, rédacteur, tech publisher) ? Indispensable de rappeler l'importance de la qualité et de la quantité des données brutes !
14. **Gouvernance des données :** Comment renforcer/établir les principes de gouvernance des données et de disparition à terme des silos de données ? Ce sont des freins parfois définitifs à la mise au point d'assistants performants ! Cela amène à devoir rappeler quelques « anciens » principes qui eux-mêmes ont leurs outils : knowledge management, interopérabilité, gestion électronique des documents, etc. Ces outils deviennent indispensables pour produire de la donnée « consommable » par les LLM, mais aussi pour la faire vivre, la maintenir, l'adapter, l'enrichir.

Voici donc quelques-unes des questions auxquelles les industriels doivent répondre avant de s'approprier les technologies des LLM et IA génératives. En effet, il reste à établir quelques certitudes nécessaires dès lors que l'on touche aux domaines du nucléaire, à la défense, à l'avionique, au droit, à la finance, à la pharmacie, au médical...

Y a-t-il une démarche à suivre ?

L'ensemble de ces questions recevra une réponse différente dans chaque entreprise selon son contexte de compétences internes, selon le niveau de confidentialité ou de secret de ses données et suivant les moyens de calcul et de stockage internes ou externes auquel elle peut avoir accès en satisfaisant ses exigences de confidentialité.

Un certain nombre de recommandations sont publiées d'ores et déjà concernant les risques de cybersécurité ou concernant les risques de fuites de données induits par l'usage des LLM. Un certain nombre de guide méthodologiques de bonnes pratiques en matière d'IA, de gestion des données et de formation des utilisateurs sont publiés : commission européenne, EASA, DGA, etc. Certaines de ces recommandations ne diffèrent pas de recommandations plus classiques en matière de cybersécurité, de fuites de données ou d'usage d'algorithmes complexes, à base d'IA symbolique, numérique ou neuronale ou pas, en connexion avec l'internet mondial ou non.

De manière générale, il parait essentiel d'insister sur la gestion des données. Au-delà d'un discours marketing insistant sur les algorithmes et le *prompting*, sachant que pour mettre au point les LLM, qui sont des modèles de langage, les données publiques ont été « absorbées » de façon globale et sans réel discernement via des stratégie « big data ». Mais pour les mises en œuvre spécialisées et utilisables dans le monde des entreprises nous aboutissons invariablement à des approches de type « small data » pour lesquelles la qualité, la gestion, la « vivacité » et l'instrumentation / enrichissement de la donnée brute doivent occuper une place prépondérante des projets exploitant des LLM. Dans ce registre rappeler l'importance d'outils de knowledge management ou les stratégies de gouvernance, par exemple, n'est-il pas indispensable ?

Il semble également important d'insister sur la nécessaire formation des utilisateurs afin qu'ils comprennent à quel genre d'outil ils ont affaire et sachent se poser la question d'en reconnaître les limitations d'usage et les risques. Il va de soi que les développeurs de solutions internes ou externes

doivent être des utilisateurs avertis ou a minima formés et informés des usages autorisés et ceux qui ne le sont pas concernant les bases de données qui seront exploités en développement ou en utilisation.

GLOSSAIRE

IAG : Intelligence Artificielle Générative
Domaine de l'intelligence Artificielle fondé sur une approche statistique et dont l'objectif est de fabriquer une réponse sous la forme d'un texte, d'une image, d'un code informatique, voire d'une combinaison de medias.

GPT : Generative Pre-trained Transformer ou Transformeur génératif pré-entraîné
C'est une famille demodèles de langage servant notamment au sein de moteur d'agent conversationnel (ChatBot) pour ChatGPT.

LLM : Large Language Models
C''est un langage basé sur des réseaux de neurones artificiel en capacité à apprendre des relations statistiques à partir de documents textuels (par exemple) lors d'un protocole d'entraînement supervisé (auto-supervisé voire semi-supervisé).
PROMPT : Prompting
Le prompt, en français *"invite"* ou "*ordre*", est le format de la requête qui est soumise à l'application d'IAG. Le "prompting" est l'art de fabriquer la bonne requête par rapport à ce que l'on attend et en fonction de l'application d'IAG sollicitée. La réponse à un "prompt" s'appelle "achèvement".

RAG : Retrieval Augmented Generation ou Génération Augmentée de Récupération
C'est une technique de traitement du langage naturel qui est considérée comme un sur-ensemble du LLM. L'objectif étant de prendre en compte "des règles ou des faits plus récents et plus fiables" afin de gommer le côté statistique des LLM.
Token :
Le Token ou jeton est une unité de texte que le programme de langage naturel utilise afin de gérer - c'est à dire comprendre puis de générer - du langage. Ainsi, il correspond approximativement à 4 caractères en anglais. Une centaine de tokens peuvent ainsi représenter 75 mots.