Bonnes pratiques

# Recommandations pour l'usage des IA génératives comme assistant personnel au sein d'INRAE

Depuis plus de 10 ans déjà, les systèmes d'intelligence artificielle (IA) nous assistent dans des tâches quotidiennes sans nécessairement que nous en ayons pris conscience (correction orthographique, dictées de messages, traduction de textes, traitement des photos de nos smartphones...). D'ici la fin de la décennie, il est vraisemblable que les systèmes d'intelligence artificielle nous accompagneront en continu et dans beaucoup d'autres tâches du quotidien. Cela pourrait prendre la forme de nombreux assistants personnels, qui accompliraient des tâches répétitives, appuieraient la réflexion et la prise de décision, et accéléreraient le travail de groupe. Les IA évoluent très rapidement et il parait aujourd'hui nécessaire d'accompagner ce changement dans le cadre du travail à INRAE. Cette note de recommandations présente les premières lignes directrices. Un plan d'action suivra dans un second temps.

## Que sont les IA génératives ?

L'intelligence artificielle est depuis longtemps un sujet d'étude pour la recherche, cependant, depuis la mise en ligne de ChatGPT par OpenAI en fin d'année 2022, son utilisation s'est largement étendue. Son récent succès a été rendu possible grâce à l'augmentation des puissances de calcul et par la disponibilité de grands volumes de données. Cette évolution a permis aux machines « d'apprendre » automatiquement des règles à partir de données de moins en moins structurées qui se sont diversifiées et affinées avec le temps.

Les systèmes d'IA générative sont entraînés sur d'importantes quantités de données et « apprennent » les relations entre celles-ci qui peuvent prendre la forme de texte, d'image, de son, de vidéo, de tableaux de valeurs, etc. Parmi ceux-ci, on trouve la catégorie des *Large Language Model* (ou LLM) qui produisent une relation statistique donnant les possibilités d'apparition d'entités (appelés *token*) les unes après les autres.

Lorsqu'ils sont entraînés sur de très grands volumes, ils forment des modèles à usage général qui peuvent ensuite être adaptés à beaucoup de tâches différentes. Une fois entraînés, ils peuvent être interrogés par un dialogue avec l'utilisateur, on parle alors d'IA conversationnelles.

L'entraînement d'un LLM est un processus long nécessitant un corpus de texte volumineux (l'intégralité de Wikipedia ne correspondrait qu'à quelques pourcents) et beaucoup de temps de calcul. Cela reste néanmoins un processus simple, sans intervention humaine. Le but en est de prédire un mot dans une phrase. A la fin de l'entraînement, le modèle dit de fondation est capable de produire du texte mais n'est pas encore vraiment utilisable. En effet, ce modèle pourra « répondre » de façon inappropriée, le modèle doit être alors contrôlé. Deux techniques sont utilisées à cette fin.

**Rédacteurs :** Micael Aliouat, Colette Cadiou, Jocelyn De-Goer-De-Herve, Remy Decoupes, Nathalie Gandon, Marjolaine Hamelin, Hadi Quesneville, Tristan Salord, Alban Thomas

INRAE - DipSO – Septembre 2024 – DOI : 10.17180/ztym-j930

### 1 - Fine tuning : spécialisation d'un LLM à apprentissage succinct

Le *fine-tuning* est une technique utilisant un corpus de textes défini qui permet de spécialiser sur une tâche spécifique un modèle de langage pré-entraîné. Cette méthode tire parti des connaissances générales acquises par le modèle lors de son entraînement initial sur un vaste corpus de textes et les affine pour répondre aux besoins d'un domaine ou d'une application particulière. Le *fine-tuning* permet de former de nouveaux LLM à partir de peu de données et de ressources de calcul par rapport à l'entraînement complet d'un modèle de fondation. Cette approche offre ainsi un moyen efficace d'exploiter la puissance des grands modèles de langage tout en les adaptant à des besoins spécifiques.

### 2 – Prompt-engineering : contrôle par un texte donné en entrée au niveau du prompt

**Les services d'IA génératives conversationnelles insèrent un prompt permettant aux utilisateurs de poser leur question à l'aide d'un texte.**

La méthode RAG (*Retrieval Augmented Generation*) associe à un LLM une base de connaissances pour le contextualiser. Cette famille de techniques utilisées pour spécialiser les réponses d'un LLM recherche des informations pertinentes dans une base de données externe et les utilise pour générer une réponse. Cette approche permet au modèle de fournir des réponses plus précises et plus informatives, en particulier lorsque les questions nécessitent des connaissances spécifiques qui ne sont pas intégrées dans le modèle de fondation. Les méthodes de RAG permettent notamment de pouvoir interroger les connaissances de documents de façon occasionnelle et qui ne nécessitent pas d'être intégrées au modèle via les techniques de *fine-tuning*. Contrairement au *fine-tuning*, les méthodes des RAG ne demandent pas de capacités de calcul accrues. Cependant, comme elles enrichissent les prompts, leurs tailles augmentent. Ceci induit un coût non négligeable puisque le modèle économique principal de l'IA impose une tarification au nombre de mots reçus et générés par l'agent conversationnel.

## Leur utilisation en pratique

Trois principales modalités d'utilisation des LLM sont possibles, chacune avec des avantages et des risques différents :

- via une application web hébergée chez un tiers (ChatGPT, LeChat, Gemini, HuggingFace, Groq, Cohere) ;
- via une ligne de commande ou un programme informatique accédant à une IA hébergée chez un tiers (ex : API de la plateforme openai) ;
- via une instance hébergée localement qui peut être accédée en version Web ou API.

Les usages types sont :

- **Obtenir des réponses à des questions ou rechercher des informations**.
  Cela peut être vu comme une alternative à l'utilisation des moteurs de recherche du web.
- **Aider à la rédaction de texte** en produisant des résumés, le plan d'un document, en améliorant le style, la grammaire, l'orthographe, en réalisant la traduction…
  Cela permet de gagner du temps sur la rédaction de documents.
- **Générer des codes informatiques** ou les vérifier.
  Cela facilite l'écriture de programmes, de requêtes à des bases de données, l'écriture de page web…

---

- IA : Intelligence artificielle
- LLM : Large language model – catégorie d'IA qui produit une relation statistique entre les données
- RAG : Retrieval augmented generation – méthode qui associe à un LLM une base de connaissances pour le contextualiser

## Des avantages, et de nouvelles opportunités

**Les LLM contribuent à l'augmentation de la productivité.** Ils permettent de gagner du temps pour la réalisation de tâches répétitives (résumés de documents, développement de code informatique…), ou longues et spécialisées (reformulation / vulgarisation de textes complexes). Beaucoup de tâches de bureau seront transformées, d'autres seront supprimées, et de nouvelles tâches apparaîtront. Ces transformations pourront concerner aussi bien des tâches à faible valeur ajoutée (ex : rédiger un courrier ou une note), que les tâches spécialisées qui composent le cœur d'un métier. Un accompagnement au changement sera nécessaire pour bien profiter de cette opportunité d'évolution.

**Les LLM pourraient faciliter et libérer du temps pour la production de nouvelles idées.** Le travail du chercheur pourrait s'en trouver facilité grâce à l'appui d'outils d'IA l'aidant à identifier de nouvelles hypothèses, créer des protocoles et réaliser des expériences. La production d'idées augmentera, entraînant la production de nouveaux résultats. L'exemple d'AlphaFold qui permet de prédire la structure 3D des protéines à partir de leur séquence d'acides aminés montre comment l'IA peut révolutionner un domaine. L'IA ne met pas ici en danger l'originalité de la création en elle-même et ses processus, elle n'offre qu'un outil nouveau qui enrichit la panoplie technique du chercheur.

**Ils pourraient offrir une plus grande précision par la prise en compte d'une quantité d'information plus importante.** L'augmentation du nombre de documents pouvant être considérés pour produire des méta-analyses, des articles de synthèse, des expertises, doit permettre d'obtenir des résultats plus précis et plus exhaustifs.

**Ils pourraient être un outil d'aide à l'apprentissage.** Elle permet une recherche d'information dans un vaste corpus de données et en fournit une synthèse. L'utilisateur pourra acquérir plus facilement de nouvelles connaissances ce qui favorisera l'interdisciplinarité. Le développement de tuteurs virtuels que nous voyons apparaître permet aujourd'hui d'envisager de nouvelles façons d'enseigner et d'apprendre via le dialogue avec ces IA.

**Ils permettent une nouvelle interactivité avec les systèmes informatiques et automatiques.** On connaît déjà les ChatBot qui fournissent une assistance aux utilisateurs pour l'accès à des informations sur des sites web. Certains LLM pourraient se développer pour conduire des entretiens automatisés. Enfin ces systèmes se basant sur des LLM sont capables de permettre à un utilisateur de dialoguer avec des machines (drones, robots, machines agricoles…), permettant ainsi de les piloter.

**Ils peuvent dialoguer entre eux.** Ces LLM ont la capacité de négocier entre eux et échanger des informations. Certains pensent les utiliser pour négocier automatiquement des contrats d'accès à des informations, rendant les systèmes plus interopérables (*Smart contract*).

- Le domaine est en rapide évolution, la présente note risque d'être obsolète au moment de la lecture. Il est clair que nous ne sommes pas encore en mesure d'identifier toutes les opportunités qui apparaîtront, ni leur ampleur.
- De plus, certaines expériences montreraient que certaines IA génératives ont des capacités émergentes, c'est-à-dire non prévues à la conception de celles-ci. On ne comprend pas encore bien les origines de ces nouvelles caractéristiques, mais elles semblent émerger au-dessus d'une taille critique de données ayant servi à l'entraînement.

- IA : Intelligence artificielle
- LLM : Large language model – catégorie d'IA qui produit une relation statistique entre les données
- RAG : Retrieval augmented generation – méthode qui associe à un LLM une base de connaissances pour le contextualiser

# Les limites et les risques

## Risques liés à l'exploitation

**L'IA générative a besoin de données en quantités considérables.** Les plus grands modèles de textes ont déjà utilisé la plupart des corpus existants. Leur évolution conduira probablement à économiser de la puissance de calcul et à mettre davantage l'accent sur la qualité des données que sur leur quantité. Le choix des données dites de qualité et spécifiques est déterminant pour obtenir de bons résultats.

**Impact environnemental de ces pratiques.** La consommation énergétique nécessaire à l'entraînement des grands modèles de langue a montré que les émissions de gaz à effet de serre liées à l'entraînement aux États-Unis d'un des premiers grands modèles de langue étaient du même ordre de grandeur que celles d'un vol entre New-York et San Francisco. L'utilisation après la phase d'apprentissage est quant à elle beaucoup moins consommatrice, mais mériterait aussi d'être évaluée. Cependant il convient de mettre en regard de cette consommation, ses potentiels bénéfices.

**Impacts sociétaux et éthique.** Des gains de productivité sont attendus avec l'usage d'assistants basés sur des IA génératives. Ils pourraient transformer rapidement plusieurs métiers, au risque d'un « décrochage » de certains au sein de la communauté. L'entraînement de certaines IA génératives ont par ailleurs conduit à l'exploitation de travailleurs sous-payés.

**Le degré d'ouverture du modèle.** Connaître les données d'entraînement des modèles, leur fonctionnement, leurs fragilités et leurs atouts est un prérequis pour la confiance dans les résultats d'une IA générative. Les sources rarement citées, en plus de leur complexité intrinsèque, vaut aux LLM une réputation de « boîte noire ». Il existe des biais dans les résultats liés à la nature et la qualité des données mises pour entraîner les modèles. Par exemple, le fait que les données d'internet utilisées pour entraîner les LLM soient principalement de langue anglaise introduit un biais sur la culture américaine.

**L'utilisation de l'IA générative pour la recherche pose une question de souveraineté.** De plus, lorsqu'on utilise des modèles d'IA générative développés par des tiers, nous devons envisager la possibilité que le fournisseur puisse un jour interrompre le modèle. Cela pourrait avoir un impact important sur la reproductibilité de nos travaux et impacter notre productivité.

**Augmentation massive des productions.** L'augmentation de la production de texte par des IA risque d'amener un surcroît d'activité de lecture. Par exemple, le *peer-reviewing* pourrait s'en trouver encore plus engorgé. Cela aura un impact sur les délais de *reviewing* (les éditeurs peinent à trouver des *reviewers*) et donc la publication des articles scientifiques.

## Risques liés à l'usage

**Des risques d'erreurs factuelles, appelés hallucinations.** Les LLM fonctionnent sur la base de probabilités, ce qui leur donne leur flexibilité, mais aussi leur capacité de se « tromper ». En effet, un LLM ne comprend ni ce qu'a écrit un utilisateur, ni la réponse qu'elle vient de générer. Les résultats peuvent être vraisemblables d'un point de vue des probabilités, mais peuvent être faux. Le résultat est souvent affirmé sans aucune nuance de doute s'il y en avait. Il devient alors difficile d'évaluer la fiabilité des résultats. Ce constat n'empêche pas d'utiliser un LLM comme assistant, mais impose de contrôler le contenu généré.

- IA : Intelligence artificielle
- LLM : Large language model – catégorie d'IA qui produit une relation statistique entre les données
- RAG : Retrieval augmented generation – méthode qui associe à un LLM une base de connaissances pour le contextualiser

**Des risques d'information obsolète (date de l'apprentissage du modèle).** Les LLM sont entraînés sur des données jusqu'à une date spécifique et sont donc ignorants de tout événement ou information produite au-delà de cette date. Il est nécessaire de la connaître afin d'évaluer quelles questions sont appropriées pour son utilisation.

**Des risques juridiques.** Elles exposent à un risque de perdre la nouveauté ou l'antériorité nécessaire pour l'obtention ou la reconnaissance d'un droit de propriété intellectuelle s'agissant des données injectées dans une IA. Il y a aussi un risque accru de contrefaçon, à l'insu de l'utilisateur, si les modèles sont entraînés par exemple sur des publications ou des données soumises sous licences restrictives ou sans tenir compte du droit d'auteur.

De la même manière, elles peuvent engendrer une perte de confidentialité sur des données sensibles lorsque celles-ci sont utilisées pour l'entraînement de ces modèles. En effet, une partie de la masse des données déjà utilisées comme corpus d'entraînement sont des données à caractère personnel. Elles sont donc déjà présentes et peuvent en plus être injectées par l'utilisateur de l'IA. **Les risques pour la vie privée** vont donc s'accroître. Les premiers contentieux sur les IA génératives sont portés au niveau des autorités de contrôle des données à caractère personnel et ces dernières sont pointées comme étant les régulateurs compétents. Par ailleurs, ces premiers contentieux montrent que la fiabilité des informations personnelles utilisées dans les corpus de ces IA génératives peut être remise en cause et pointent la grande difficulté pour les concepteurs de supprimer les fausses informations.

**Risques de reproduction de normes sociales discriminantes.** Les IA génératives peuvent être un vecteur de reproduction des normes sociales et culturelles présentes dans les données d'entraînement, ce qui peut influencer les résultats et conduire à des conclusions erronées si ces normes ne sont pas correctement identifiées et corrigées.

**Des risques déontologiques, intégrité scientifique.** Les IA génératives peuvent favoriser et renforcer les phénomènes de bulles informationnelles déjà présentes dans les réseaux sociaux. Les fausses informations (hallucinations ou *fake news*), l'augmentation des fraudes scientifiques et du plagiat sont autant de problèmes qui peuvent s'intensifier. Il est actuellement difficile de détecter les contenus générés par LLM (en particulier du texte), mais des développements technologiques sont en cours pour leur apposer un « tatouage numérique ». Ce marquage qui pourrait être rendu obligatoire par la future législation européenne pour encadrer l'IA (*AI Act*) devrait donc limiter le risque de fraude sans toutefois l'éliminer. Enfin, il y a un risque de délégation totale de tâches à l'IA générative entraînant une perte de compétences et des risques d'erreurs accrues.

## Recommandations

**Continuer la veille sur les LLM,** et leur usage en tant qu'assistant personnel. Nous vivons actuellement une période de foisonnements de modèles et de services basés sur des LLM. Il est difficile de prédire aujourd'hui quel(s) modèle(s)/service(s) seront les plus pertinents pour assister des chercheurs. Cette veille pourra se traduire par une mise à jour des présentes recommandations et de la formation.

**Développer une expertise sur l'évaluation des LLM.** Pour connaître la confiance que l'on peut accorder à ces IA, il est nécessaire de les tester. Les chercheurs doivent être en mesure d'évaluer les limites de performance de ces outils. Cette précaution permet de comprendre leurs limites pour en réduire les risques et s'assurer que leurs systèmes remplissent efficacement l'usage pour lequel ils sont prévus.

L'institut doit suivre dans la durée l'évolution des performances et des biais des systèmes d'IA afin d'anticiper de nouveaux risques. INRAE dispose de chercheurs dans ce domaine qui peuvent éclairer sa direction sur ce sujet. Ce besoin d'évaluation concerne aussi bien les biais (culturels, sociétaux, informationnels…) que l'impact

- IA : Intelligence artificielle
- LLM : Large language model – catégorie d'IA qui produit une relation statistique entre les données
- RAG : Retrieval augmented generation – méthode qui associe à un LLM une base de connaissances pour le contextualiser

environnemental. Tout ceci implique de disposer de capacités d'évaluation des systèmes d'IA dont la fiabilité soit reconnue par tous.

**Maîtriser la spécialisation des modèles.** Il est très difficile d'améliorer la qualité d'un LLM de fondation. Cela nécessiterait des ressources de calculs gigantesques que seules quelques entreprises ont aujourd'hui. Mais un tel modèle peut être affiné, spécialisé pour des usages particuliers (technique de RAG et *fine tuning*). Cette maîtrise est clef pour le bon usage de ces IA.

**Maîtriser le flux de données.** L'utilisation des IA implique l'utilisation de données pour l'apprentissage (ou la spécialisation), mais aussi lors de leur utilisation. Celles-ci n'échappent pas au cadre régi par la gouvernance des données. Les données à caractère personnel et les données sensibles ne doivent pas être accessibles pour d'autres. Il convient donc d'avoir des systèmes d'IA hébergés dans des infrastructures de confiance suivant la réglementation en vigueur sur le partage des données. Il sera nécessaire alors d'encourager l'usage de LLM installés localement, de vérifier la sécurité de ces LLM, de privilégier des contenus délimités et vérifiés pour l'apprentissage, de s'abstenir d'injecter des données à caractère personnel, confidentielles ou sensibles, de s'abstenir d'injecter des informations qui pourraient faire l'objet d'une protection future, et de privilégier des modèles de langue « open-source ».

**Sensibiliser, former et accompagner les utilisateurs à l'usage des IA génératives.** La qualité des réponses de l'IA générative est fortement influencée par l'entrée utilisateur ou le prompt. Il est nécessaire de proposer des formations pour aider les utilisateurs à maîtriser cet outil. Cependant, l'expérimentation est indispensable pour élaborer des prompts clairs, spécifiques et structurés de manière appropriée afin que l'IA générative génère la sortie avec le style, la qualité et le but souhaités. Des personnes ressources pourraient être identifiées pour accompagner les prises en main individuelles. Il faut également responsabiliser l'agent sur ce qui a été généré. L'utilisateur doit savoir qu'il doit garantir le respect de la réglementation.

**Structurer une gouvernance de l'IA.** L'évolution du domaine est rapide et continue, il est indispensable de suivre son évolution et de l'accompagner pour adapter au mieux notre usage. En effet, l'utilisation de l'IA emporte des risques réglementaires, éthiques, organisationnels pour les collectifs, et environnementaux pour lesquels un accompagnement des communautés est indispensable. De façon similaire à l'établissement d'une gouvernance des données, algorithmes et codes, il est souhaitable d'établir une gouvernance de l'IA. Le lien fort existant avec les données suggère d'établir une gouvernance commune. Cette gouvernance aura en charge de suivre les usages de l'IA au sein d'INRAE, mais aussi les initiatives nationales et internationales qui se mettent en place pour l'accompagnement à leur utilisation, leur contrôle et leur développement. Elle permettra de participer à l'organisation du partage des connaissances et des expériences sur ces nouveaux outils au sein de l'Enseignement supérieur et de la recherche. Enfin, elle pourra fournir des recommandations pour leur bon usage au sein de l'Institut.

## Références


- *A quick guide of using GAI for scientific research*. MIDAS. Web. 24 Apr. 2024. https://midas.umich.edu/generative-ai-user-guide
- Balaguer A et al. *RAG vs fine-tuning: pipelines, tradeoffs, and a case study on agriculture*. 30 Jan. 2024. arXiv.org. Web. 24 Apr. 2024. http://arxiv.org/abs/2401.08406
- Bagenal J (2024*). Generative artificial intelligence and scientific publishing: urgent questions, difficult answers*. The Lancet, 403, 1118–1120. https://doi.org/10.1016/S0140-6736(24)00416-1
- Devillard A (2024) *Loi sur l'IA : "Le tatouage numérique permet de repérer les deepfakes."* Sciences et Avenir.


- IA : Intelligence artificielle
- LLM : Large language model – catégorie d'IA qui produit une relation statistique entre les données
- RAG : Retrieval augmented generation – méthode qui associe à un LLM une base de connaissances pour le contextualiser

- *Generative AI Usage Guidance for the Research Community*. Provost & Chief Academic Officer - UNC Chapel Hill. Web. 24 Apr. 2024. https://provost.unc.edu/generative-ai-usage-guidance-for-the-research-community
- *IA : Notre ambition pour la France.* 15 Mar. 2024. vie-publique.fr. Web. 24 Apr. 2024. https://www.vie-publique.fr/rapport/293444-ia-notre-ambition-pour-la-france
- *Rapport d'information n°2207*. Web. 24 Apr. 2024. https://www.assemblee-nationale.fr/dyn/16/rapports/cion_lois/l16b2207_rapport-information
- Veissier L (2024) *Quand ChatGPT tient la plume*. TheMetaNews. https://themeta.news/quand-chatgpt-tient-la-plume/

**Pour citer ce document** : Micael Aliouat, Colette Cadiou, Jocelyn De-Goer-De-Herve, Remy Decoupes, Nathalie Gandon, Marjolaine Hamelin, Hadi Quesneville, Tristan Salord, Alban Thomas, 2024. Recommandations pour l'usage des IA génératives comme assistant personnel au sein d'INRAE, INRAE (France), 7 p. DOI : 10.17180/ztym-j930


## Les rédacteurs

Micael Aliouat
Direction des Affaires juridiques - INRAE, DAJ, 75338, Paris, France.

Colette Cadiou, Marjolaine Hamelin, Hadi Quesneville, Tristan Salord, Alban Thomas
Direction pour la Science ouverte - INRAE, DipSO, 75338, Paris, France.

Jocelyn De-Goer-De-Herve
Université Clermont Auvergne, INRAE, VetAgro Sup, UMR EPIA, 63122, Saint-Genès-Champanelle, France.

Remy Decoupes
Univ. Montpellier, AgroParisTech, CIRAD, CNRS, INRAE, UMR TETIS, 34090, Montpellier, France.

Nathalie Gandon
Déléguée à la protection des données d'INRAE - INRAE, CODIR, 75338, Paris, France.

- IA : Intelligence artificielle
- LLM : Large language model – catégorie d'IA qui produit une relation statistique entre les données
- RAG : Retrieval augmented generation – méthode qui associe à un LLM une base de connaissances pour le contextualiser