Strategie Geoinformation Schweiz
Stratégie suisse pour la géoinformation

État de l'art

Limites et contraintes des LLM

Risques et solutions, Version de Mai 2024

geospatial

Contexte

L'État de l'art *Limites et contraintes des LLM* a été mandaté dans le cadre du plan d'action 2024 (action 4-24-02) de la Stratégie suisse pour la géoinformation (SGS) en collaboration avec l'Office féderal de topographie (swisstopo). Ce rapport fait suite au rapport précédent, *État de l'art LLM et Géodonnées,* et contient les observations, les avis, les résultats et les conclusions de chercheurs et d'experts dans le domaine de la technologie géospatiale et de l'intelligence artificielle.

Remerciements

Ageospatial souhaite remercier l'Office féderal de topographie (swisstopo), COSIG (Coordination, Services et Informations Géographiques), l'Agence suisse pour l'encouragement de l'innovation (Innosuisse), EPFL Innovation Park (La Forge), l'Université de Genève (UNIGE), Geneva Responsable Enterpreneurship Center (GREC) et l'Institut des Sciences de l'Environnement (ISE).

TABLE DES MATIÈRES

PRÉAMBULE

La “Stratégie suisse pour la géoinformation”, adoptée le 11 décembre 2020 par le Conseil fédéral et la Conférence suisse des directeurs cantonaux des travaux publics, de l’aménagement du territoire et de l’environnement (DTAP), a pour objectif de fournir des informations géolocalisées fiables, détaillées, actualisées et interopérables pour tous les utilisateurs.

Cette stratégie vise à intégrer la géoinformation dans tous les aspects de la société, de l’économie et de l’environnement, pour faciliter la prise de décision, promouvoir le développement durable et favoriser l’innovation. Elle souligne l’importance d’une gestion durable et concertée de l’espace et des ressources limitées en Suisse, par la création et le maintien de représentations numériques des espaces géographiques.

Le mandat du Conseil fédéral est clair : exploiter la numérisation pour rendre le cadre de vie plus attrayant et durable, et pour maintenir la compétitivité de la Suisse en tant que site économique et de recherche. Les sept champs d’action définissent les grandes orientations pour les prochaines années : promouvoir l’écosystème de la géoinformation, relier les géodonnées, faciliter les processus, développer la science des géodonnées, encourager l’innovation, acquérir et renforcer les compétences, et développer les plateformes numériques. Ces actions visent à maximiser l’utilité des géoinformations et à encourager une approche agile et participative.

Une consolidation de l’infrastructure nationale de données géographiques (INDG) est actuellement en cours pour faciliter l’accès aux données géographiques via des interfaces standardisées (INDG 2.0) et pour préparer la mise en œuvre d’un réseau de plateformes collaboratives (INDG 3.0).

Les effets attendus de la stratégie comprennent un soutien accru à la prise de décision quotidienne, un accès à des géodonnées fiables et harmonisées, une meilleure connectivité entre les données non spatiales et les géodonnées, et une amélioration de la résilience et de la satisfaction des utilisateurs.

Pour garantir la mise en œuvre de cette nouvelle stratégie, des plans d’action annuels sont publiés afin d’établir une feuille de route pour les différentes parties prenantes. Ce rapport est réalisé dans le cadre de la Stratégie Suisse pour la Géoinformation (SGS) et de son plan d’action pour l’année 2024. Dans le champ d’action dédié à “Développer la science des géodonnées”, une initiative spécifique (4-24-02) porte sur l’utilisation de l’intelligence artificielle (IA) et des grands modèles de langage (LLM) dans les infrastructures de géodonnées. Cette initiative a pour objectif d’intégrer les capacités avancées des LLM au traitement et à l’analyse des données géospatiales.

INTRODUCTION

Comme nous l'avons vu dans l'état de l'art précédent, *LLM et Gédonnées, approches, outils et méthodologies*, les grands modèles de langage peuvent servir d'outils puissants pour de multiples applications réelles telles que l'analyse géospatiale.

Toutefois, lorsque nous travaillons avec de nouvelles technologies puissantes, nous devons prendre en compte les risques et les limites potentiels afin de pouvoir les utiliser de manière responsable et éthique. L'objectif de ce rapport est d'explorer les différentes limites des modèles LLM et MLLM actuels et de présenter des solutions potentielles pour surmonter ces limites.

Nous diviserons notre analyse en trois parties : la première partie tiendra compte des considérations éthiques telles que la confidentialité des données et la transparence des modèles, la deuxième partie se penchera sur les limites techniques telles que les hallucinations et les erreurs techniques, et la dernière partie traitera de l'impact sociétal et environnemental actuel et potentiel de l'intelligence artificielle générative.

CHAPITRE 1 :
CONSIDÉRATIONS ÉTHIQUES

1.1 BIAIS, STÉRÉOTYPES ET DISCRIMINATION

Comment les LLM peuvent renforcer les stéréotypes existants ou générer du contenu offensif et discriminatoire

Nous savons que la langue joue un rôle essentiel dans la formation de l'identité, de dynamiques sociales et de rapports de force. Elle reflète les normes sociétales et peut influencer la manière dont les individus sont perçus et traités. Les grands modèles de langage (LLM), dont l'objectif principal est de reproduire le langage humain, peuvent générer du contenu biaisé, inexact ou discriminatoire.

Ce type de contenu peut être le résultat de données d'apprentissage biaisées, d'un biais résultant d'un apprentissage continu sur des requêtes biaisées formulées par les utilisateurs, d'un biais algorithmique où le modèle est forcé de s'aligner sur une variable comme la précision ou l'engagement de l'utilisateur, et d'un biais basé sur le contexte géographique, tel que la langue ou les requêtes spécifiques à un lieu donné comme un pays. La taille du modèle a aussi une grande influence sur la génération de contenu biaisé. En effet, un modèle trop petit éprouvera des difficultés à s'adapter à de nouvelles situations et contextes et aura donc plus de chances de régurgiter de l'information biaisée, surtout si les données d'entraînement n'ont pas été vérifiées.

Toutefois, les modèles plus grands, dotés d'un plus grand nombre de paramètres réglables, ont une meilleure compréhension des relations complexes entre les mots et du contexte. Les plus gros modèles ont aussi la capacité d'ingérer une plus grande quantité de données réelles ou synthétiques qui peuvent venir en contrepoids et limiter les effets des données d'apprentissage biaisées [1].

Durant la génération de texte par un LLM, les biais peuvent se révéler de manière subtile, comme sur le choix des mots et le ton général. Par exemple, pour des tâches de traduction automatique, le LLM peut choisir par défaut des mots masculins en cas d'ambiguïté, comme pour la traduction de l'anglais "I am happy" à la forme masculine française "je suis heureux" plutôt qu'à la forme féminine "je suis heureuse" [2].

Les grands modèles sont aussi limités par d'importants biais géographiques et géopolitiques qui résultent de la représentation inégale des régions du monde dans les ensembles de données d'entraînement. Les biais géographiques des LLM se manifestent principalement par une surreprésentation des données provenant de pays géopolitiquement importants avec une forte présence sur internet (où la majorité des données d'entraînement sont récoltées), en particulier ceux du Nord, tels que les États-Unis et les pays d'Europe de l'Ouest. La conséquence est un modèle qui comprend et génère de manière disproportionnée des contenus pertinents pour ces régions, au détriment des autres régions sous-représentées.

Certains LLM sont également multilingues, c'est-à-dire qu'ils peuvent comprendre et générer du contenu dans différentes langues (jusqu'à 80 pour ChatGPT, par exemple [3]). Cela ajoute une nouvelle couche de complexité car, bien que ces modèles soient entraînés sur plusieurs langues, ils ne présentent pas nécessairement un niveau uniforme de compréhension ou de représentation du contexte géographique

de chaque langue. Cela entraîne une répartition inégale des connaissances géographiques entre les langues, où certaines langues peuvent être bien représentées en termes de diversité géographique, alors que d'autres ne le sont pas.

Nous pouvons quantifier et représenter ces biais au sein des LLM par l'extraction des *expert-units* ou des neurones spécifiques du modèle qui réagissent fortement à certaines notions géographiques. La figure ci-dessous montre une représentation géographique utilisant *les expert-units* (fig. 1.1) de BLOOM, indiquant comment différents pays sont connectés en fonction de leur proximité géographique ou culturelle au sein du modèle. Les clusters de couleurs illustrent des groupes de pays géographiquement ou culturellement proches, comme les pays d'Amérique du Sud en bleu clair, les pays africains en jaune, et les pays d'Asie du Sud-Est en bleu foncé. La taille des nœuds est proportionnelle à leur importance dans le graphe, représentant leur degré de connexion [4].

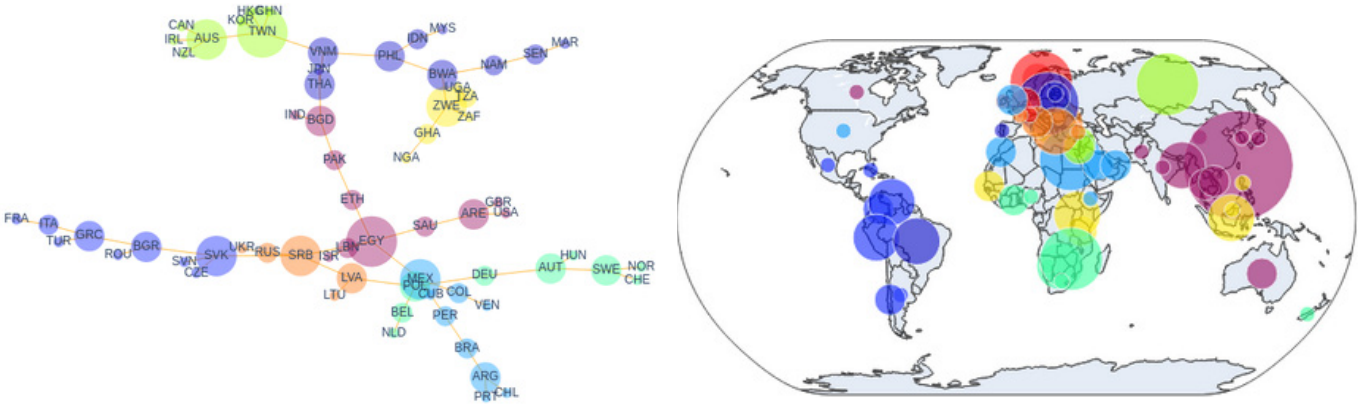

FIGURE 1.1 – Représentation géographique de BLOOM [4]

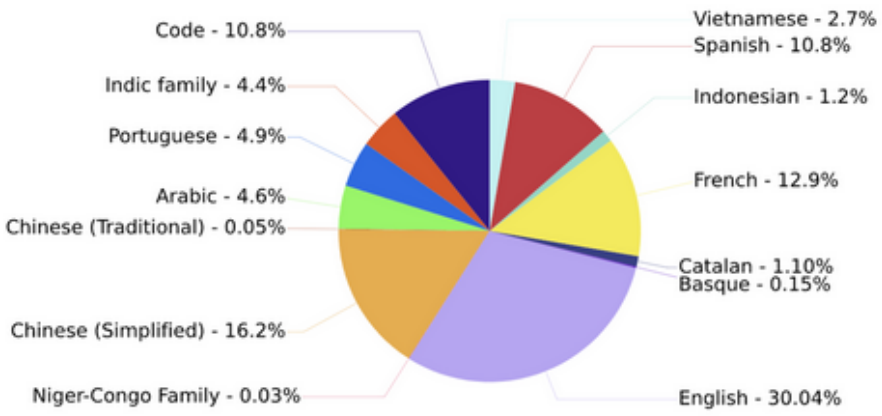

FIGURE 1.2 – Répartition des langues dans les données d'entraînement [5]

1.2 DONNÉES PERSONNELLES ET CYBERATTAQUES

Traitement et protection de données sensibles et privées

Alors que l'utilisation des grands modèles de langage se généralise, nous avons constaté certaines lacunes en matière de respect de la vie privée et de conformité réglementaire. Certains pays, comme l'Italie, ont même temporairement interdit ChatGPT en raison de préoccupations liées à la protection de la vie privée [6]. Les utilisateurs de ces LLM doivent donc être conscients des risques et prendre des mesures pour limiter les risques de fuites de données personnelles.

Les LLM sont formés, parfois en continu comme ChatGPT, sur de vastes ensembles de données qui comprennent souvent des informations sensibles ou personnelles, tirées de sites internet, de plateformes numériques ou de requêtes et documents téléchargés par les utilisateurs. Cette approche soulève deux problèmes principaux en lien avec les données personnelles: le risque de révéler des informations à caractère privé et la possibilité que ces modèles soient exploités pour générer des informations sensibles sur la base de leurs données d'entraînement. En effet, il est possible de manipuler le LLM pour qu'il génère mot pour mot ses données d'entraînement comme l'on fait des chercheurs de Google DeepMind en demandant à ChatGPT de répéter le mot "Book" ou "Poem" à l'infini [7]. Au bout de plusieurs répétitions, le LLM a révélé des données à caractère personnel telles que des noms, des adresses électroniques et des numéros de téléphone. On pourrait imaginer aller plus loin avec des LLM entraînés sur des données géospatiales sensibles, qui pourraient être liées à des informations personnelles comme l'adresse physique ou le numéro AVS. Il devient plus facile de commettre des usurpations d'identité, de créer de faux documents, ou de divulguer des informations personnelles en ligne, notamment de personnalités publiques.

La réglementation en vigueur relative à la confidentialité des données, telle que le Règlement général sur la protection des données (RGPD), accorde aux individus le droit d'accéder à leurs données personnelles, de les rectifier et de les effacer. Cependant, il est extrêmement difficile d'effacer les données ingérées par le LLM une fois qu'il a été entraîné. Cela signifie que nous devons nous assurer que les données de formation sont anonymisées à l'avance et que toutes les données personnelles identifiables sont supprimées. L'utilisateur doit ainsi vérifier que ses requêtes et les documents qu'il télécharge dans le LLM ne contiennent aucune donnée personnelle. Comme la méthode de manipulation utilisée par les chercheurs de Google, de nombreuses autres méthodes de cyberattaque sont envisageables pour l'extraction de données sensibles d'un LLM ou de ses données de traitement. Le hacker peut aussi "empoisonner" les données d'entraînement avec des données malveillantes ou compromettantes afin de perturber la sécurité, l'efficacité ou le code moral du modèle en question [8], mais aussi de créer des LLM malveillants capables de générer du texte ou du code pour des campagnes d'hameçonnage [9], par exemple.

La figure suivante (fig. 1.3), présente les différentes cyberattaques dont les LLM sont vulnérables ainsi que les différentes techniques de défense envisageables. Les lignes colorées réprésentent une technique de défense qui nous permet de nous défendre contre une ou un groupe d'attaques spécifiques. L'explication détaillée du graphique est disponible en annexe à la page 52.

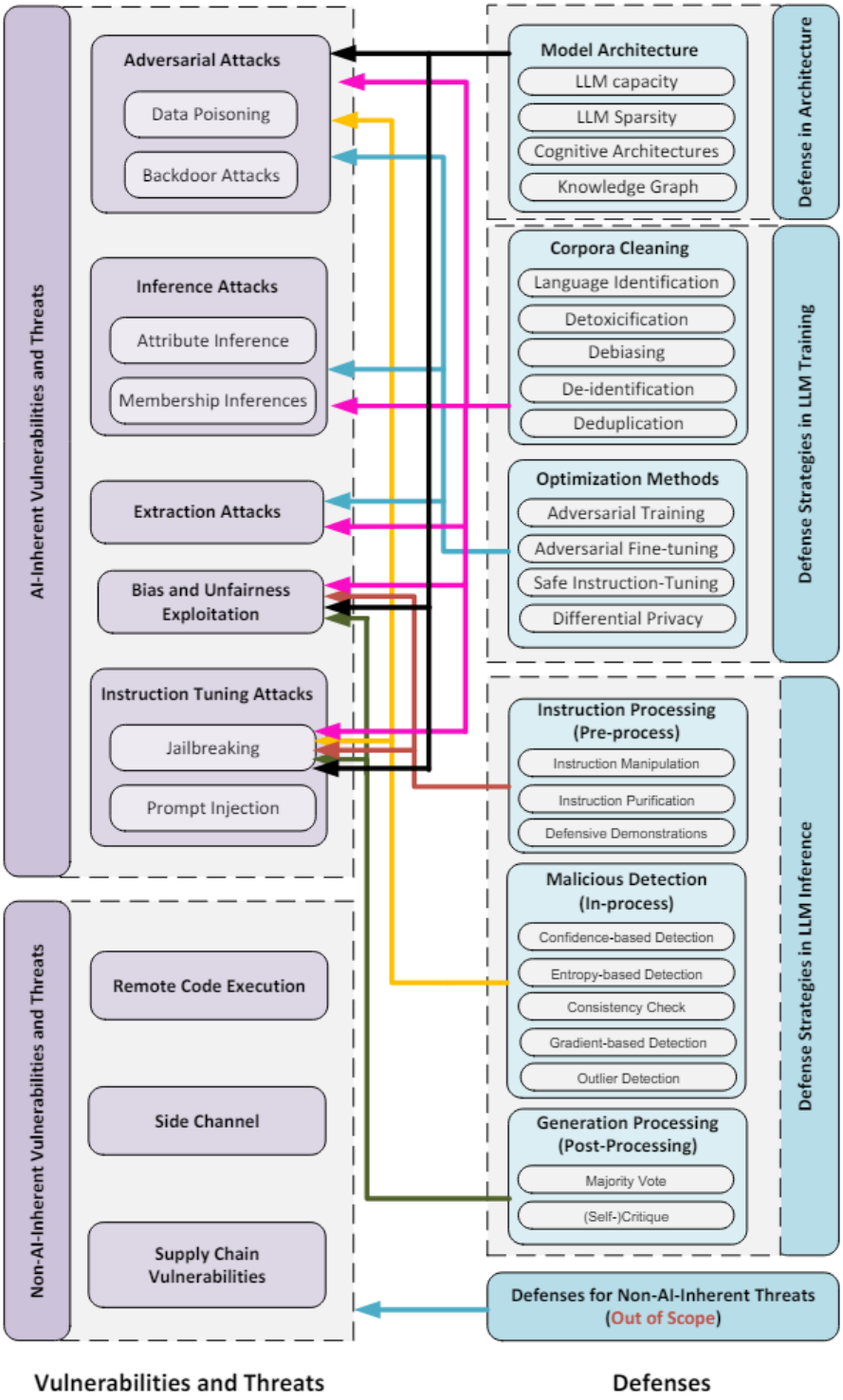

FIGURE 1.3 – Diagramme des vulnérabilités et des menaces liées aux cyberattaques et stratégies défensives envisageables [8]

1.3 DÉSINFORMATION

Capacité des LLMs d'aggraver ou de combattre la dissémination de la désinformation

La désinformation et la mésinformation constituent déjà un problème majeur, en particulier avec la croissance des réseaux sociaux et le développement de technologies de type "Deepfake", et autres technologies similaires. Avec l'avènement de LLM puissants tels que ChatGPT, de fausses informations peuvent être générées à une échelle sans précédent. Cependant, les LLM peuvent également être formés pour lutter contre la création de fausses informations, à condition que certaines mesures de prévention soient mises en place.

L'une des principales questions éthiques est la facilité avec laquelle les LLM peuvent être utilisés pour générer de la désinformation. Cette capacité peut être utilisée de manière malveillante pour produire un contenu qui semble factuel au premier regard, ce qui rend difficile la distinction entre les informations réelles et les informations erronées. Avec un peu de *prompt engineering*, les LLM peuvent aussi imiter le style d'écriture d'un individu, ce qui complique l'identification du texte généré pour les humains et les systèmes automatisés. Par ailleurs, avec l'aide de LLM multimodaux, de fausses images, faux sons répliquant la voix d'une personne et même de fausses vidéos peuvent être générés et diffusés sur les réseaux sociaux comme dans l'exemple ci-dessous (fig. 1.4). Un écosystème numérique où la désinformation est endémique peut conduire à une érosion de la confiance du grand public et peut avoir des effets dommageables à long terme, en particulier si l'opinion publique est influencée (lors des élections par exemple).

L'incorporation de la connaissance géospatiale dans la création de fausses informations permet d'augmenter considérablement leur crédibilité. En utilisant des données géospatiales précises, il est possible de contextualiser les informations fausses de manière locale, rendant ces informations plus convaincantes et difficiles à discréditer. Cette technique permet aux générateurs de désinformation d'intégrer des éléments spécifiques aux contextes locaux, tels que des repères géographiques (fig. 1.4), des événements régionaux, ou des aspects culturels propres à une communauté donnée.

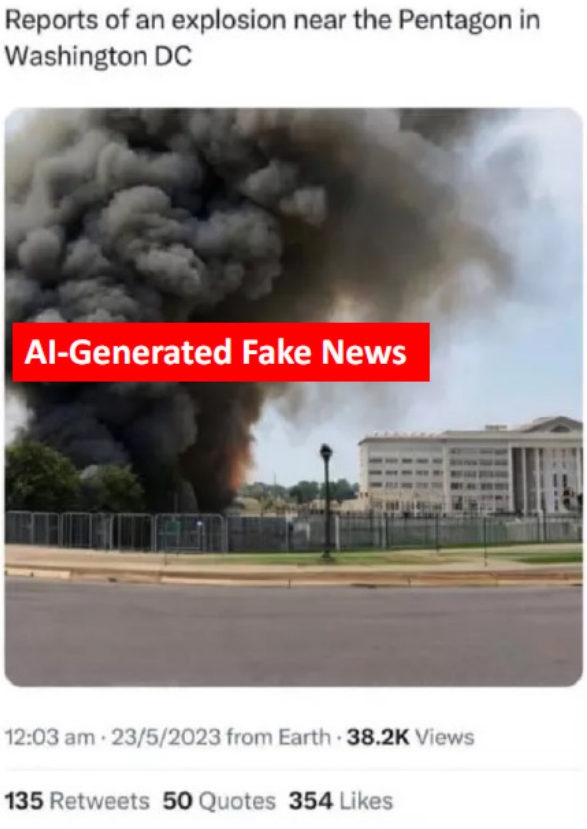

FIGURE 1.4 – Désinformation généré à partir d'un LLM multimodal [10]

Grâce à cette précision géospatiale, il est possible de produire des fausses informations extrêmement ciblées, conçues pour manipuler des groupes spécifiques ou des zones géographiques particulières. Par exemple, une fausse nouvelle concernant une catastrophe naturelle peut inclure des détails géospatiaux précis pour une région particulière, augmentant ainsi la crédibilité et la portée de l'information. De même, des informations erronées sur des politiques locaux ou des figures publiques régionales peuvent être adaptées pour maximiser leur impact et influencer l'opinion publique. Cette capacité de ciblage précis présente des défis importants pour la lutte contre la désinformation, d'où la nécessité de développer des méthodes permettant aux LLM de combattre la création et la propagation de fausses informations. Pour ce faire, il est essentiel de renforcer les capacités de raisonnement des modèles afin qu'ils puissent vérifier la validité des informations à partir de leurs données d'entraînement, par des recherches en ligne, ou en utilisant des outils de vérification externes comme PolitiFact.

De plus, on peut entraîner les modèles à détecter automatiquement le contenu généré par IA, notamment pour les contenus multimodaux, et appliquer une étiquette d'avertissement, comme illustré dans la figure 1.4 ci-dessus. D'autres solutions sont envisageables, notamment de mandater que tous les contenus générés par une IA soient étiquetés, soit explicitement, soit dans les métadonnées.

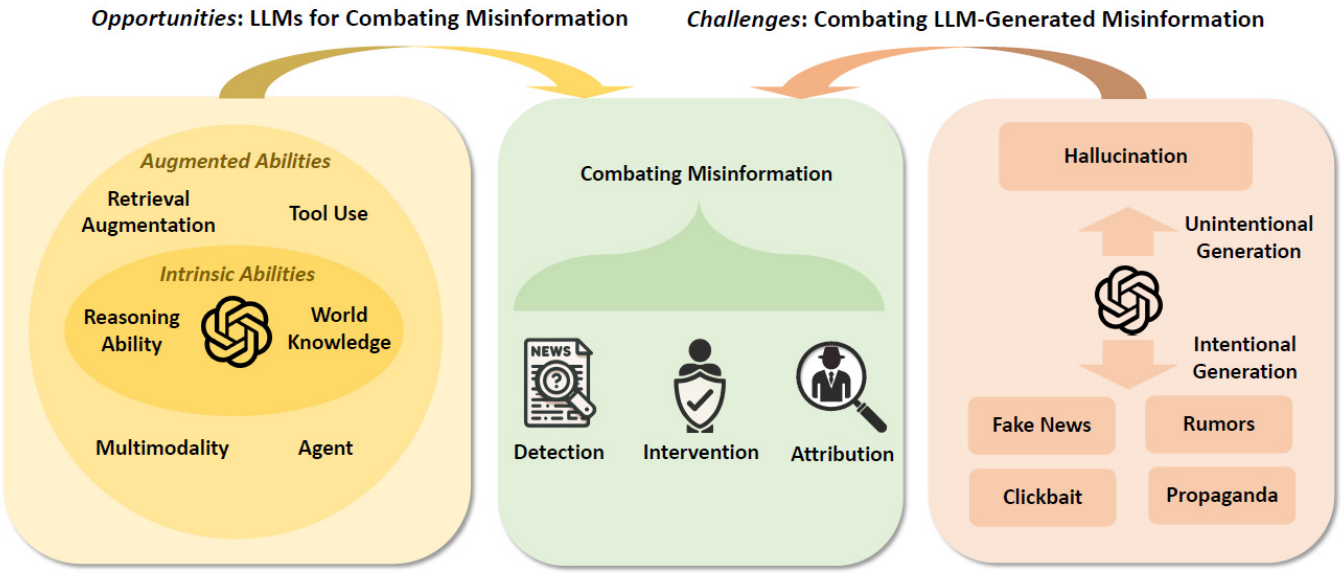

FIGURE 1.5 – Possibilités et défis de la lutte contre la désinformation dans l'ère des LLMs [10]

1.4 TRANSPARENCE

Clarifier le fonctionnement des LLMs pour un déploiement éthique de l'IA

En vue des risques comme la propagation des préjugés, la désinformation et la génération de contenus non fiables ou toxiques, nous devons évaluer le besoin impératif de transparence dans le déploiement des grands modèles de langage (LLM). Cela est particulièrement important dans les domaines exigeant un degré élevé de confiance de la part des utilisateurs, comme le droit ou la santé. L'article "*AI Transparency in the Age of LLMs: A Human-Centered Research Roadmap*" par Q. Vera Liao et Jennifer Wortman Vaughan souligne que, bien que les LLM offrent de remarquables possibilités d'innovation, ils comportent également des risques importants s'ils ne sont pas déployés de manière responsable. Pour atténuer ces risques, les chercheurs prônent l'élaboration de mécanismes de transparence tenant compte des divers besoins des différentes parties prenantes dans la gestion du cycle de vie des LLM.

Selon les auteurs, la transparence ne consiste pas seulement à rendre visible le fonctionnement interne du modèle, mais aussi à s'assurer que les parties prenantes comprennent les capacités et les limites des LLM. Les auteurs décrivent quatre stratégies principales qui ont été traditionnellement employées pour assurer une certaine transparence: la normalisation en matière de documentation telle que le "Model Card" (Figure 1.6), la publication des résultats d'évaluation suivant différentes méthodes (par exemple HELM [11]), la fourniture d'explications sur la manière dont le modèle interprète et traite les informations et l'incertitude du modèle. Ils notent toutefois que ces stratégies sont confrontées à de nouveaux défis dans le cadre des LLM en raison de leur nature complexe, opaque et souvent propriétaire. Par exemple, la normalisation de la documentation est compliquée par l'ampleur et la diversité des données d'apprentissage et des capacités des modèles, qui ne sont pas entièrement prévisibles, même par leurs développeurs.

Dans l'optique de fusionner les technologies géospatiales et les technologies d'intelligence générative, nous devons nous assurer de la qualité des données géospatiales utilisées, leurs métadonnées, ainsi que le degré de transparence des fournisseurs. L'absence d'uniformité dans les processus de collecte, de formatage et de partage des données est un problème qui complique actuellement la création de jeux de données spatiales normalisés pour l'entraînement des LLM. Les données géospatiales utilisées sont alors soit trop peu nombreuses, soit elles manquent de cohésion et de structure, affectant ainsi la fiabilité des informations géospatiales produites par les LLM. Nous observons donc une nécessité d'établir des normes pour la gestion et l'exploitation des données géospatiales à travers une collaboration plus étroite entre les différents acteurs à toutes les échelles. Sans cette normalisation des données, il sera difficile de garantir la qualité et la fiabilité des données géographiques traitées par le LLM.

Model Card

- **Model Details**. Basic information about the model.
  - Person or organization developing model
  - Model date
  - Model version
  - Model type
  - Information about training algorithms, parameters, fairness constraints or other applied approaches, and features
  - Paper or other resource for more information
  - Citation details
  - License
  - Where to send questions or comments about the model
- **Intended Use**. Use cases that were envisioned during development.
  - Primary intended uses
  - Primary intended users
  - Out-of-scope use cases
- **Factors**. Factors could include demographic or phenotypic groups, environmental conditions, technical attributes, or others listed in Section 4.3.
  - Relevant factors
  - Evaluation factors
- **Metrics**. Metrics should be chosen to reflect potential real-world impacts of the model.
  - Model performance measures
  - Decision thresholds
  - Variation approaches
- **Evaluation Data**. Details on the dataset(s) used for the quantitative analyses in the card.
  - Datasets
  - Motivation
  - Preprocessing
- **Training Data**. May not be possible to provide in practice. When possible, this section should mirror Evaluation Data. If such detail is not possible, minimal allowable information should be provided here, such as details of the distribution over various factors in the training datasets.
- **Quantitative Analyses**
  - Unitary results
  - Intersectional results
- **Ethical Considerations**
- **Caveats and Recommendations**

FIGURE 1.6 – Exemple de “Model Card” proposé par Mitchell et al. [12]

1.5 PRÉSERVER LA CONFIANCE

Dans quelle mesure les utilisateurs font-ils confiance aux solutions d'IA générative ?

D'après le rapport de KPMG, *Generative AI Consumer Trust Survey,* publié en janvier 2024 [13], nous pouvons voir que 74 % des personnes interrogées affirment qu'elles font confiance aux organisations qui utilisent l'IA générative dans leurs opérations au quotidien et 70 % pensent que les bienfaits surpassent les risques. Cependant, ce niveau élevé de confiance découle souvent d'un manque de sensibilisation du public aux risques, ce qui peut rendre les consommateurs particulièrement vulnérables à la désinformation, aux *Deepfakes* et aux cyberattaques [14].

"Si l'IA générative simplifie la vie des gens et apporte une réelle valeur ajoutée, elle franchira les barrières de l'âge. La technologie sera couramment adoptée une fois que ses limites auront été surmontées et qu'elle apportera une valeur ajoutée concrète aux consommateurs." Louis DiCesari, Responsable Mondial des Données, Analytics et de l'IA chez Levi Strauss. Le potentiel des technologies d'IA générative dans la simplification des tâches et l'amélioration de l'expérience des utilisateurs est indéniable.

Cependant, l'enthousiasme pour ces technologies entraîne également une série de défis éthiques et opérationnels qui doivent être abordés afin de maintenir la confiance des utilisateurs et d'assurer une intégration sécurisée de la technologie au sein de la société. Les pouvoirs publics doivent ainsi fournir un effort concerté pour réguler l'IA générative et normaliser les politiques divergentes en matière de protection des données personnelles, de sécurité des données et de droits de propriété intellectuelle et assurer que toutes les parties prenantes peuvent faire confiance à cette nouvelle technologie.

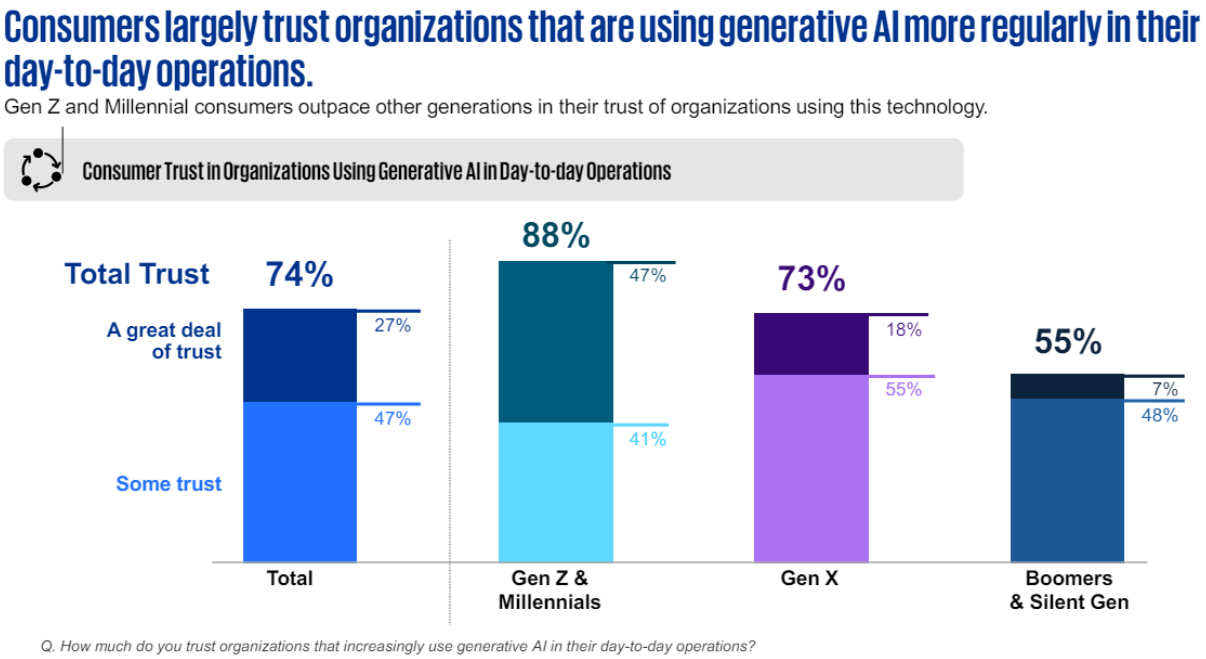

FIGURE 1.7 – Confiance des consommateurs envers les organisations qui utilisent l'IA générative dans leurs activités quotidiennes [13]

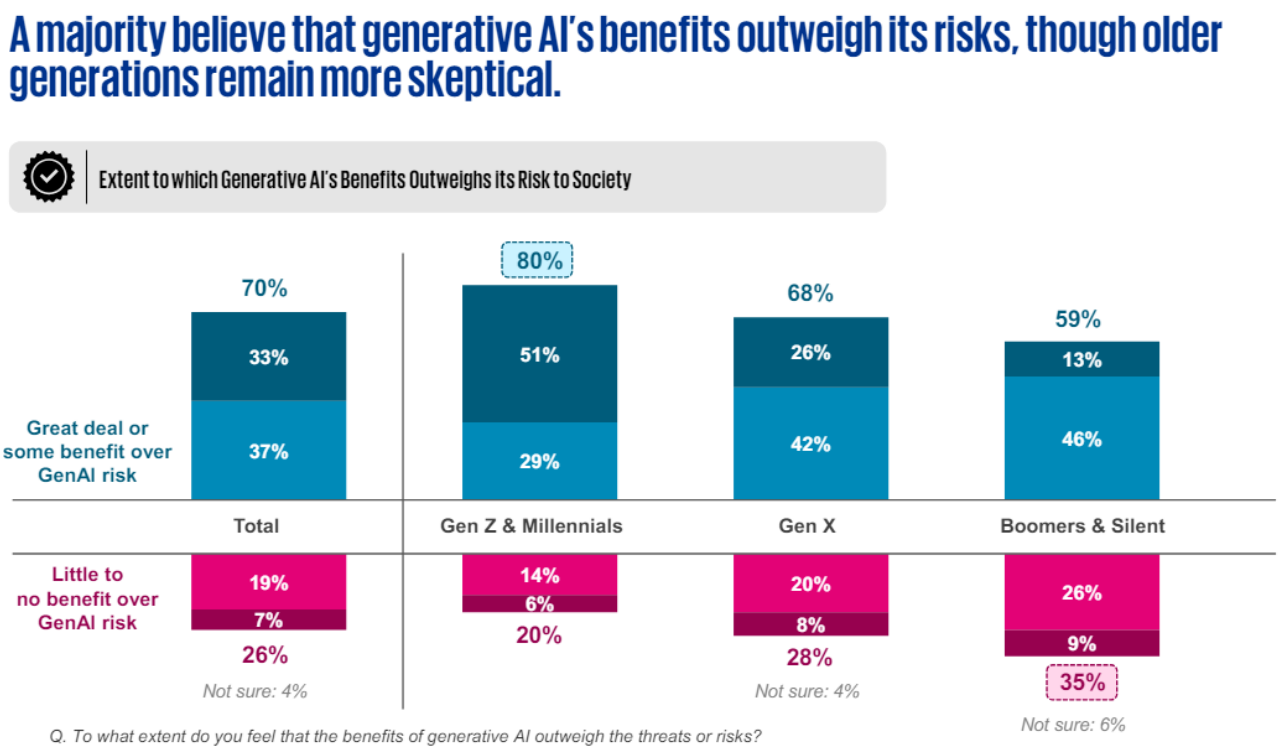

FIGURE 1.8 – Mesure dans laquelle les avantages de l'IA générative sont supérieurs aux risques qu'elle constitue pour la société [13]

CHAPITRE 2 :
CONSIDÉRATIONS TECHNIQUES

2.1 ERREURS TECHNIQUES

Erreurs qui affectent la fiabilité et les performances des LLM, telles que les erreurs de raisonnement

Au fil du temps, nous avons remarqué que les LLM manquent toujours de rigueur dans leur raisonnement logique pour des tâches complexes. Par exemple, les LLM peuvent commettre des erreurs dans la compréhension approfondie de problèmes complexes, ce qui se traduit par des erreurs dans le traitement des calculs ou dans le suivi des séquences logiques. Ces erreurs peuvent porter sur des étapes manquantes dans une chaîne de raisonnement ou sur des “hallucinations”, où le modèle génère des informations factuellement incorrectes ou inadaptées dans sa réponse.

Certaines techniques, comme l’élaboration d’une requête pour guider le modèle à raisonner en plusieurs étapes et articuler ses étapes de raisonnement de manière explicite avant de fournir une réponse finale, peuvent améliorer les capacités de raisonnement des LLM [15]. Cependant, ces améliorations peuvent être plus ou moins utiles en fonction de la nature de la tâche et du modèle utilisé.

Cependant, malgré leur capacité à extraire des informations à partir d’une quantité considérable de données, les LLM ne “comprennent” pas fondamentalement le contenu comme le ferait un être humain. Subbarao Kambhampati, chercheur à l’Arizona State University, a conclu que “rien de ce que j’ai lu, vérifié ou fait ne me donne une quelconque bonne raison de supposer que les LLM sont capables de raisonner et de planifier, comme on l’entend normalement” [16]. Pour lui, leur “raisonnement” est souvent le résultat d’une identification de tendances dans les données d’entraînement, plutôt qu’une véritable capacité de raisonnement.

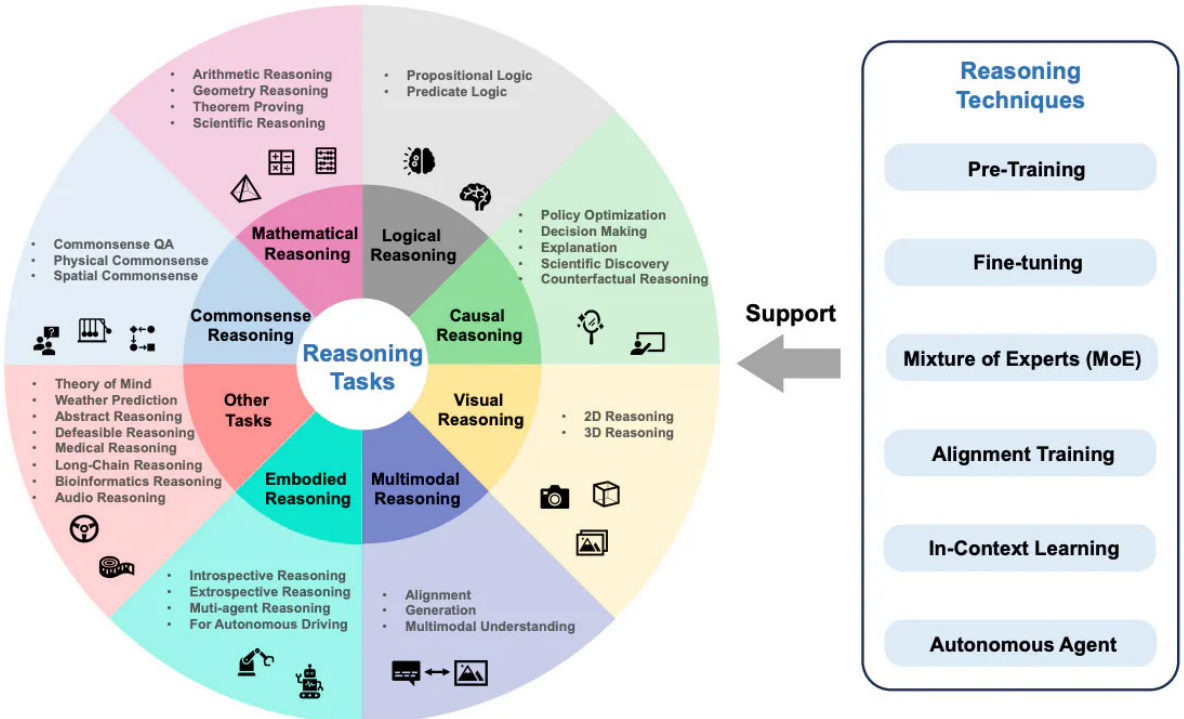

FIGURE 2.1 – Aperçu des techniques de raisonnement pour les LLM [17]

2.2 HALLUCINATIONS

Génération de contenu incohérent, factuellement incorrect ou absurde

Le phénomène d'hallucination dans les grands modèles de langage (LLM) représente un défi important et reste l'une des questions les plus complexes et les plus difficiles à résoudre. L'hallucination fait référence aux cas où ces modèles génèrent des réponses plausibles mais factuellement incorrectes ou dépourvues de sens. Il est essentiel de bien comprendre et d'atténuer ce problème car il peut avoir un impact sur la fiabilité et la crédibilité des applications d'intelligence artificielle. Les chercheurs dans l'article intitulé *A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions* [18] proposent de classer les hallucinations en deux catégories distinctes. La première catégorie, nommée *factuality hallucination*, fait référence aux hallucinations qui présentent des contradictions avec les faits réels (*factual inconsistency*) ou des informations inventées (*factual fabrication*). La deuxième catégorie, *faithfulness hallucination*, décrit comment les données de sortie du LLM peuvent dévier des instructions de l'utilisateur (*instruction inconsistency*), ignorer le contexte fourni par l'utilisateur (*context inconsistency*), ou présenter des incohérences logiques dans des tâches à plusieurs étapes (*logical inconsistency*).

Hallucination Type	Sub-Type	User Input	Model Output	Explanation
Factuality Hallucination	Factual Inconsistency	Tell me about the first person to land on the Moon.	Yuri Gagarin was the first person to land on the Moon.	The LLM's response is factual inconsistency as **Neil Armstrong was the first person to land on the Moon**, not Yuri Gagarin, who was the first human in space.
	Factual Fabrication	Tell me about the historical origins of unicorns.	Unicorns were documented to have roamed the plains of Atlantis around 10,000 BC, where they were considered sacred creatures and were often associated with royalty.	The LLM's response is a fabricated claim since **there's no verified, real-world evidence supporting the existence of unicorns in Atlantis**, especially in connection with royalty.
Faithfulness Hallucination	Instruction Inconsistency	Translate the following English question into Spanish: "What is the capital of France?"	The capital of France is Paris.	The user explicitly provided an instruction to translate the English question into Spanish. **Instead of performing the translation as instructed, the LLM chose to answer the question.**
	Context Inconsistency	Summarize the following paragraph about the Nile: The Nile is a major north-flowing river in northeastern Africa. It's traditionally considered the longest river in the world. Originating in the Great Lakes region of central Africa, it flows through several countries before emptying into the Mediterranean Sea.	The Nile originates from the mountain ranges of central Africa and flows into the Mediterranean.	The LLM's summary incorrectly states that the Nile originates from the "mountain ranges of central Africa," while **the paragraph clearly mentions that the Nile originates in the Great Lakes region of central Africa.**
	Logical Inconsistency	Solve the equation 2x + 3 = 11 step by step.	Step 1: Subtract 3 from both sides to get: 2x = 8. Step 2: Divide both sides by 2 to get: x = 3. So the result of the equation is x=3.	The first step correctly isolates the term with x to get 2x = 8. However, **the next step inaccurately divides 8 by 2 to yield a result of x = 3**, which is inconsistent with the earlier reasoning.

FIGURE 2.2 – Exemples de chaque catégorie d'hallucinations proposées par Huang et al. [18]

Dans le même article, les chercheurs estiment que les hallucinations sont dues à des données d'apprentissage défectueuses (fig. 2.3) ou des limitations des connaissances du modèle (fig. 2.4). Comme évoqué précédemment, les données d'apprentissage défectueuses peuvent perpétuer ou provoquer de la désinformation ou des biais qui se reflètent dans le texte généré par le modèle sous forme d'hallucinations.

Type	Premise	Input	Model Ouput	Explanation
Imitative Falsehood	Within the LLM's training data, a frequently referenced but incorrect fact appears multiple times: *Thomas Edison invented the light bulb.*	Who is credited with the invention of the light bulb?	Thomas Edison is credited with the invention of the light bulb.	**While Edison improved the light bulb's design, he was not its sole inventor.** The model's answer reflects a misconception prevalent in its training data, exemplifying an imitative falsehood.
Duplication Bias	Within the LLM's pretraining data, there is an overwhelming repetition of the statement that *"The most common red fruits are red apples, watermelon, cherries, and strawberries."*	Please list some red fruits, excluding apples.	Red fruits are red apples, watermelon, cherries, and strawberries.	**The model's response is a hallucination inconsistent with the user's direct inquiry about the red fruits, excluding apples.** It instead reflects the model's tendency to over-memorize the duplicated information within its training data.
Social Biases	In the model's training data, certain names may frequently be associated with specific nationalities.	Provide a summary of the following news: Dr. Kim recently published a breakthrough paper on quantum computing. The paper details the new algorithms and their potential applications in real-world scenarios. It has been hailed as revolutionary by experts in the field.	Dr. Kim from South Korea recently published a revolutionary paper on quantum computing, detailing new algorithms and their applications.	**The model erroneously added "South Korea" based on the name "Kim"** revealing its bias in linking certain names to specific nationalities, which can cause hallucinations in the summary.

FIGURE 2.3 – Exemples d'hallucinations résultant de la désinformation et des préjugés [18]

Type	Input	Model Ouput	Explanation
Domain Knowledge Deficiency	What causes phenylketonuria?	Phenylketonuria is caused by a mutation in the hemoglobin gene, which results in an unusual interaction with hemoglobin, affecting its ability to transport oxygen in the bloodstream.	Phenylketonuria (PKU) is actually caused by **mutations in the gene that helps make an enzyme called phenylalanine hydroxylase**. Due to the lack of specialized medical knowledge in LLM's training data, it may sometimes provide incorrect information on medical topics.
Outdated Knowledge	Which city hosted the most recent Winter Olympics?	The most recent city to host the Winter Olympics was Pyeongchang, South Korea, in 2018.	The most recent city to host the Winter Olympics was **Beijing, in 2022**. This is due to the fact that LLM stores outdated knowledge, and the answer to this question exhibits a time-shift phenomenon.

FIGURE 2.4 – Exemple d'hallucinations résultant des limites des connaissances dans les LLMs [18]

Huang et al. remarquent aussi que les hallucinations sont parfois provoquées par une incapacité du modèle à rapporter des informations présentes dans sa base de données. En effet, le modèle a parfois tendance à prendre des raccourcis et générer des réponses sur la base d'une forte corrélation entre les mots (Toronto et Canada dans l'exemple ci-dessous, fig. 2.5). Le modèle peut aussi échouer à se souvenir d'informations rares et peu fréquentes dans sa base de données et présente des difficultés à résoudre des problèmes complexes.

Type	Sub-Type	User Input	Model Output	Explanation
Knowledge Shortcut	Co-occurrence Bias	What is the capital of Canada?	Toronto is the capital of Canada.	The model leans heavily on the frequent co-occurrence of the terms *Toronto* and *Canada* in its training data, without truly capturing the factual knowledge about the capital of Canada
Knowledge Recall Failures	Long-tail Knowledge	Please generate a biography for George James Rankin.	George James Rankin is a dedicated educator known for his contributions to the field of education and his passion for fostering learning.	**George James Rankin is actually a politician**, but the LLM may have difficulty effectively utilizing knowledge about this long-tail entity despite being trained on comprehensive Wikipedia data during pre-training.
	Complex Reasoning	If Mount Everest were to descend by 500 meters, which mountain would become the world's highest peak?	If Mount Everest were to descend by 500 meters, it would still remain the world's highest peak.	The height of Mount Everest is 8844.43 meters, while K2's height is 8611 meters. If Mount Everest were to descend by 500 meters, K2 would become the world's highest peak. Facing complex multi-step reasoning questions like this, LLM may struggle to recall all the relevant knowledge associated with it.

FIGURE 2.5 – Exemple de sous-utilisation de la donnée d'entraînement [18]

Les chercheurs proposent et développent des solutions nouvelles ou existantes pour l'atténuation des hallucinations dans les LLM. Pour les hallucinations de type "*factuality hallucination*", les chercheurs proposent un système où le modèle peut consulter une base de données externe pour vérifier que sa réponse soit adaptée et factuelle (Principe de *retrieval-augmented generation* ou RAG, fig. 2.6). En parallèle, pour atténuer les hallucinations résultant d'une déviation des instructions de l'utilisateur, cinq systèmes différents peuvent être déployés pour assurer la fidélité et la cohérence de la réponse du LLM par rapport à la requête de l'utilisateur (fig. 2.7).

Finalement, pour combler les lacunes dans les connaissances du modèle, nous pouvons permettre au LLM d'accéder à des bases de données externes (fig. 2.8) soit pour une recherche unique (a), une recherche multiple (b), ou pour vérifier la fiabilité de sa réponse (c).

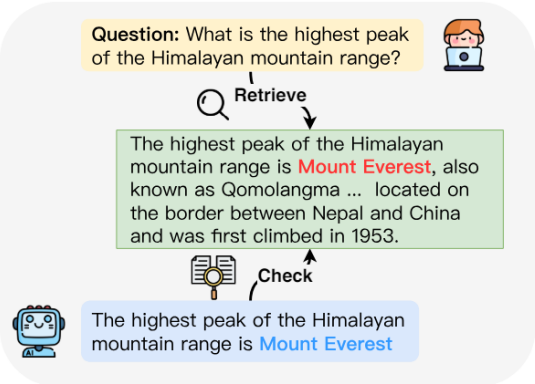

FIGURE 2.6 – Detection de *factuality hallucination* par la recherche de données externes [18]

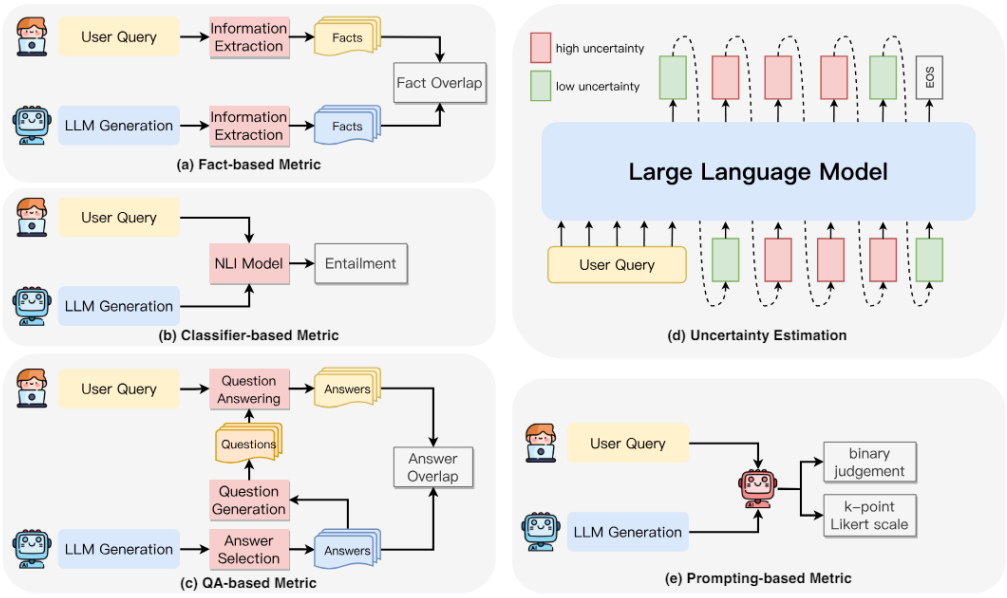

FIGURE 2.7 – Méthodes de détection d'hallucinations de type *faithfulness hallucination* [18]

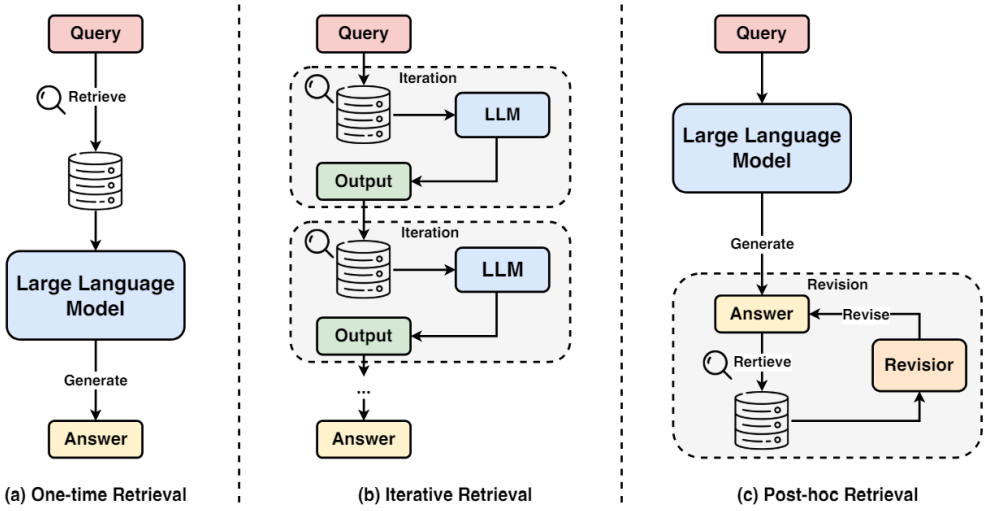

FIGURE 2.8 – Illustration de trois approches pour le *retrieval-augmented generation* [18]

Selon Xu, Jain et Kankanhalli [19], les hallucinations sont une limitation inévitable des LLM. Malgré les nombreux efforts déployés pour réduire les hallucinations, il est peut-être impossible de les éliminer. Les chercheurs affirment qu'à cause des limites en termes de ressources informatiques, de capacité d'apprentissage et de calcul, les LLMs n'auront jamais la capacité de comprendre toutes les nuances et subtilités de tous les sujets imaginables. Cette limitation inhérente suggère que les utilisateurs doivent être prudents, en particulier dans les applications dans lesquelles la fiabilité est primordiale, et doivent se fier aux solutions d'atténuation des hallucinations proposées ci-dessus.

2.3 REPRODUCTIBILITÉ

Pouvons-nous garantir que les données de sortie du LLM peuvent être systématiquement reproduits dans des conditions similaires?

Les LLM sont, par nature, de gigantesques calculateurs de probabilités, entraînés à prédire la suite d'une phrase ou d'un paragraphe au cours de leur formation. Le modèle génère donc une réponse en sélectionnant les mots ayant la probabilité la plus élevée de se suivre séquentiellement. Pour rendre la génération de texte plus flexible et naturelle, les développeurs de ces modèles ont intégré les paramètres suivants qui permettent d'influencer comment le modèle choisit ses mots.

Température : La température est un paramètre qui ajuste la distribution des probabilités des mots suivants. Une température basse (proche de 0) rend le modèle plus déterministe, sélectionnant les mots les plus probables. Une température plus élevée permet plus de diversité et de créativité dans les réponses, en permettant la sélection de mots moins probables.

Top-k : Cette méthode consiste à limiter le choix des mots suivants aux k mots les plus probables. Par exemple, si k=10, le modèle choisira le prochain mot parmi les 10 mots les plus probables, ce qui peut éviter des choix moins pertinents et améliorer la cohérence.

Top-p (ou *nucleus sampling*): Au lieu de fixer un nombre spécifique de mots comme dans le top-k, le top-p sélectionne les mots en fonction d'un seuil de probabilité cumulatif. Par exemple, si p=0.9, le modèle choisira parmi les mots dont les probabilités cumulées atteignent 90%, permettant une sélection dynamique basée sur la distribution des probabilités.

Largeur de recherche en faisceau (*Beam search width*) : La recherche en faisceau est une méthode qui explore plusieurs chemins possibles dans la génération de texte en parallèle. La largeur du faisceau (beam width) détermine le nombre de ces chemins explorés simultanément. Un faisceau plus large peut améliorer la qualité de la génération en considérant plus de possibilités, mais peut également augmenter le temps de calcul.

Seed Value: Il s'agit d'une valeur numérique spécifique utilisée pour initialiser le générateur de nombres aléatoires dans les processus de calcul. La définition d'une *seed value* garantit la reproductibilité des processus aléatoires.

Cependant, cette créativité et flexibilité inhérentes des LLM peuvent être désavantageuses, surtout lorsque nous avons besoin que le modèle produise la même réponse pour la même requête. Ainsi, en ajustant les différents paramètres, notamment la température de génération du modèle, nous pouvons garantir une certaine reproductibilité dans les réponses du LLM.

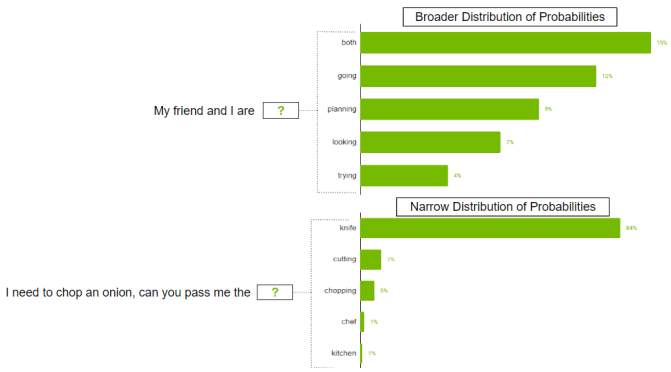

FIGURE 2.9 – Distribution des probabilités pour les résultats générés par le LLM [20]

2.4 LONGUEUR DES DONNÉES D'ENTRÉE ET FENÊTRE DE CONTEXTE

Limites opérationnelles relatives à la taille des données que les LLM peuvent traiter efficacement

Nous avons remarqué une tendance où les nouveaux modèles ne cessent d'élargir leur *context window*, ou nombre maximal de tokens que le modèle peut prendre en compte à un moment donné lors de la génération ou du traitement d'un texte. Cependant, la longueur des données d'entrée peut dégrader la performance du modèle à mesure que la longueur de ces données augmente. Ce phénomène ne se traduit pas seulement par un rendement décroissant, mais aussi par une baisse notable de la précision et de la capacité de raisonnement du modèle.

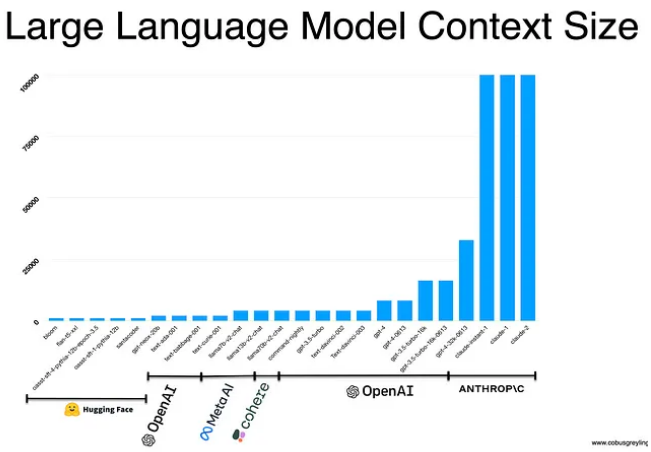

FIGURE 2.10 – Fenêtre de contexte des différents LLM [21]

Cela s'explique en partie par le fait que les modèles doivent naviguer parmi plus de "bruit" - des informations non pertinentes incluses dans l'entrée pour en augmenter la longueur. Ce bruit peut masquer les détails pertinents nécessaires à des réponses précises. Nous remarquons dans la figure ci-dessous (fig. 2.11) que GPT-4 est atypique par rapport aux autres modèles et répond bien à une technique appelée *chain-of-thought reasoning* (COT). Avec cette méthode, l'utilisateur demande au modèle de communiquer son raisonnement en plusieurs étapes. Par exemple, l'étude intitulée *Same Task, More Tokens : the Impact of Input Length on the Reasoning Performance of Large Language Models* [22] démontre que plus la longueur des données d'entrée augmente, plus les modèles peinent à maintenir le même niveau de performance observé avec des entrées plus courtes. La précision baisse en moyenne de 0,98 à 0,68 pour des données d'entrée de 3 000 tokens.

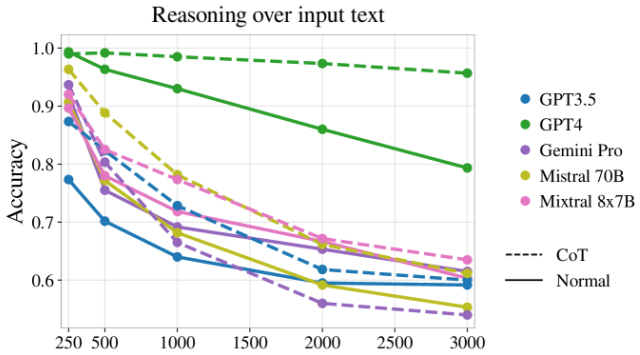

FIGURE 2.11 – Performance de cinq LLMs en fonction de la longueur des données d'entrée [22]

Les entrées plus longues nécessitent davantage de ressources informatiques, ce qui peut entraîner des inefficacités en termes d'énergie et de temps. Bien que la fenêtre de contexte du modèle le permette, il est préférable de limiter la longueur et la complexité des données d'entrée pour garantir la pertinence et la fiabilité du contenu généré par le LLM. Ces limitations sont particulièrement contraignantes pour les analyses géospatiales assistées par LLM, car elles requièrent souvent des opérations successives et une forte capacité de raisonnement à chaque étape. Or, si les capacités de raisonnement du modèle diminuent à chaque opération successive et avec chaque nouvelle entrée de données, nous risquons d'obtenir des réponses peu précises et fiables.

2.5 OPEN-SOURCE VS CLOSED-SOURCE

Comparaison des modèles LLM propriétaires et open-source et dans quel mesure sommes-nous dépendant de OpenAI?

Lorsqu'on évoque les LLM et l'intelligence artificielle générative, le nom d'OpenAI est l'un des premiers à venir à l'esprit. Fondée en 2015, cette start-up a connu une croissance exponentielle fin 2022, atteignant un million d'utilisateurs en seulement cinq jours grâce à son produit phare : ChatGPT. La popularité de ChatGPT s'explique par la facilité d'utilisation de son interface chat, sa performance inégalée et sa capacité multimodale. En plus d'un chatbot performant, OpenAI propose aussi une API facilitant l'intégration des LLM dans divers projets. En décembre 2023, OpenAI détenait plus de 39 % du marché et collaborait étroitement avec Microsoft, qui possédait 30 % des parts de marché [23]. Ensemble, ils contrôlent environ 70 % du marché, limitant ainsi le développement d'alternatives. Cette domination du marché leur permet aussi de dicter le rythme et oblige leurs concurrents à redoubler d'efforts pour rattraper leur retard. Cette 'course à l'armement' détourne surtout les ressources nécessaires au développement de LLM sécurisés, fiables et conformes aux normes en vigueur.

En mars 2024, on estime que 67 % des start-ups utilisent au moins un modèle d'OpenAI [24]. En plus de leurs propres produits, OpenAI a été très active dans le domaine *open-source*, contribuant au développement de plusieurs projets. Cette approche technique visant à favoriser les initiatives *open-source* a permis à OpenAI d'accélérer le développement de sa technologie et d'influencer considérablement le monde de l'IA, orientant divers projets vers une direction commune [25]. Cette situation soulève des questions sur la domination d'OpenAI sur le marché et les risques de standardisation des technologies d'IA générative, pouvant mener à une stagnation et à un manque de diversité dans les solutions proposées. On peut comparer cette stratégie à celle de Microsoft dans les années 1990, qui a imposé l'interface utilisateur basée sur des fenêtres, devenant ainsi la norme sur tous les systèmes d'exploitation actuels.

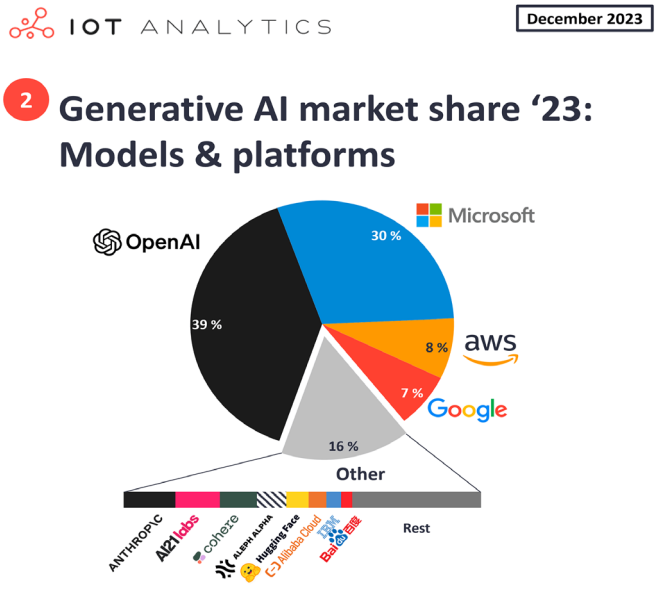

FIGURE 2.12 – Part de marché des fournisseurs de modèles et de plateformes d'IA générative [23]

La capacité d'OpenAI à diriger le marché lui permet d'innover rapidement dans un environnement peu régulé et oblige les législateurs, les décideurs politiques et leurs concurrents à adopter une position réactive. OpenAI conserve ainsi une avance systématique sur tous les acteurs publics et privés dans le domaine de l'intelligence artificielle générative. Cependant, cette approche peut susciter des réactions radicales, comme en Italie, où le pays a interdit ChatGPT pour des questions de sécurité des données personnelles [26]. Cette avance, bien qu'elle contribue au succès d'OpenAI,

est également une source d'instabilité pour l'entreprise et les utilisateurs finaux.

Dans un contexte où l'IA est de plus en plus présente dans notre quotidien, quelles alternatives s'offrent aux particuliers et aux professionnels? Nous observons une augmentation de la popularité des modèles *open-source*, notamment avec la création de Hugging Face, qui propose une plateforme centralisée pour la distribution de modèles *open-source*. Le format de distribution open-source permet non seulement de rendre l'IA accessible à tous, mais également d'impliquer la communauté dans sa création. L'utilisateur peut ainsi collaborer à son développement, mais aussi le modifier et le distribuer. L'utilisateur devient alors impliqué tout au long du cycle de vie du modèle et peut personnaliser son expérience selon ses besoins. Cette transparence dans les processus de création minimise les erreurs de développement et prévient certains abus, notamment en matière de respect de la vie privée.

Toutefois, nous devons être conscients des inconvénients liés à l'utilisation de modèles *open-source*. En effet, le déploiement de ces modèles requiert plus de temps que l'utilisation simple de l'API de modèles *closed-source* comme GPT-4. Il est aussi nécessaire de télécharger le modèle, qui peut atteindre plusieurs centaines de gigaoctets, de le déployer et de le configurer, ce qui nécessite des connaissances techniques spécifiques ainsi qu'un entretien régulier. Aujourd'hui, les modèles *open-source* commencent à combler l'écart avec OpenAI et deviennent de plus en plus accessibles grâce aux innovations dans le domaine des GPU, comme ceux développés par Nvidia, et aux optimisations telles que le nouveau modèle Phi-3 de Microsoft [27], qui offre des performances similaires à ChatGPT-3 tout en étant 97 % plus léger. Des plateformes comme ChatRTX facilitent également leur utilisation par le grand public. Les LLM *open-source*, par leur nature, nous poussent à réfléchir différemment et à innover malgré leurs limites contrairement aux LLM *closed-source* qui doivent répondre aux exigences du marché.

Dans le domaine géospatial, les modèles *open-source* se révèlent particulièrement pertinents. Ce domaine regroupe déjà une grande communauté développant des solutions *open-source* qui s'efforce constamment de rendre les données et l'analyse géospatiale plus accessibles. Leurs objectifs s'alignent avec le développement de LLM *open-source* qui peuvent automatiser la collecte, l'analyse et la visualisation des données géospatiales, réduisant ainsi le temps et les efforts nécessaires pour effectuer des analyses géospatiales complexes.

Bien qu'OpenAI domine actuellement le marché de l'IA générative, les alternatives *open-source* offrent une voie prometteuse pour diversifier et enrichir le paysage technologique, à condition que les utilisateurs soient prêts à investir les ressources nécessaires pour leur mise en œuvre.

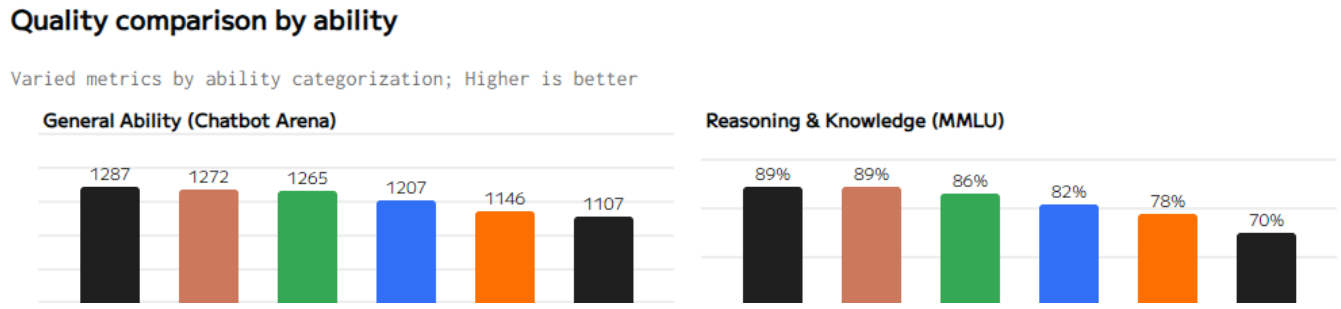

FIGURE 2.13 – Comparaison des différents modèles fournis par Azure AI [28]

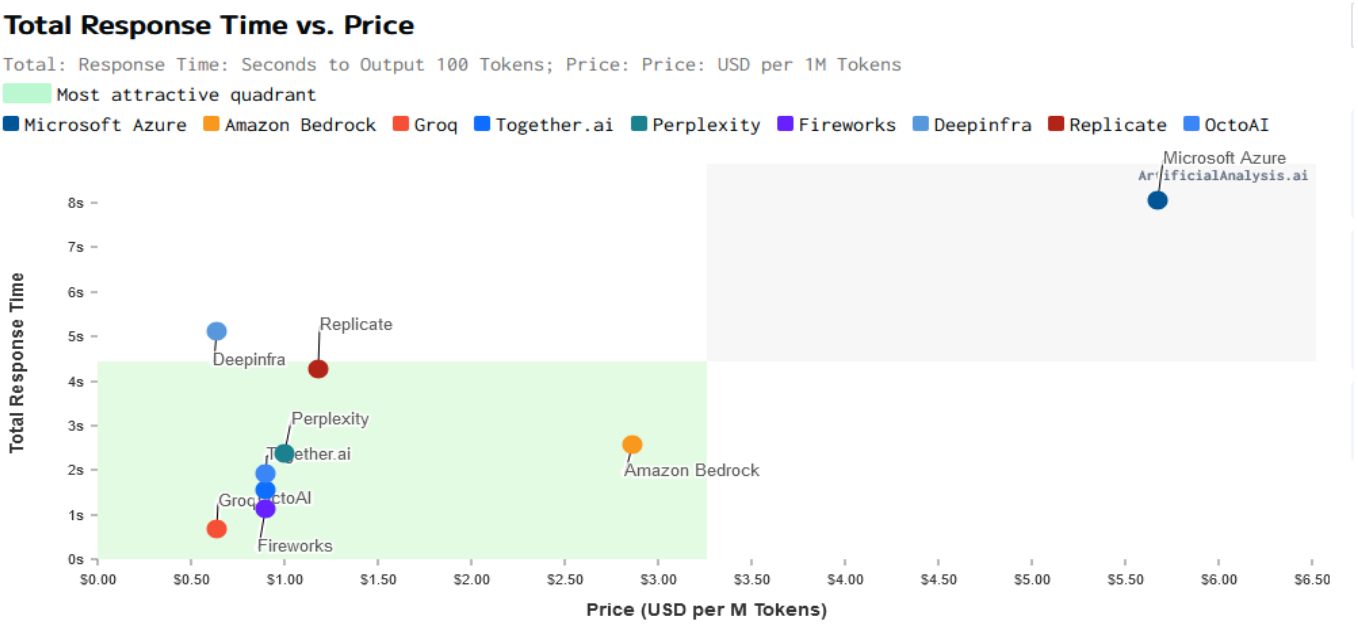

FIGURE 2.14 – Comparatif des temps de réponse et du prix de fournisseurs d'hébérgement de modèles d'intelligence générative. Le temps de réponse mesure le temps en secondes nécessaire pour recevoir uen réponse de 100 tokens et le prix est en USD/1 million de tokens. [28]

Open-source vs. closed-source LLMs

	Open-source LLMs	Closed-source LLMs
Availability	Freely available	Restricted: Paid customers, licence holders
Cost	Typically lower-cost or free	Higher cost, often with subscription fees
Integration	Can be integrated with a variety of applications	Limited to company-provided integrations
Implementation	Likely slower, especially if training is required	Ready-made APIs could make implementation as easy as plug and play
Customisation	Can be modified and adapted to specific needs	Limited customisation options
IP rights	No IP rights, free to use and modify	Company retains IP rights
Collaboration & innovation	Encourages collaboration and shared innovation	Limited to company's resources and vision
Model training, data access	Open data access, customisable training	Limited data access, pre-trained models
Transparency & explainability	Transparent, explainable model architecture	Often proprietary, less transparent
Security	Vulnerable to exploitation in open community	Company-driven security measures
Quality control	Varies by project and community	Company-driven quality control
Community support	Large developer community, support for popular models	Limited support, usually provided by the company
Updates & maintenance	Community-driven updates and maintenance	Company-driven updates and maintenance

Information compiled in 2023.

L'ATELIER BNP PARIBAS

FIGURE 2.15 – LLMs open-source versus closed-source [29]

CHAPITRE 3 :
CONSIDÉRATIONS SOCIALES ET ENVIRONNEMENTALES

3.1 EFFICACITÉ ÉNERGÉTIQUE

Consommation énérgétique nécessaires à la formation et au fonctionnement des LLM et les impacts environnementaux associés

“Les systèmes d’IA peuvent avoir une incidence importante sur l’environnement et une forte consommation d’énergie au cours de leur cycle de vie. Afin de mieux appréhender l’incidence des systèmes d’IA sur l’environnement, la documentation technique élaborée par les fournisseurs devrait inclure des informations sur la consommation d’énergie du système d’IA, y compris la consommation pendant le développement et la consommation prévue pendant l’utilisation” [30].

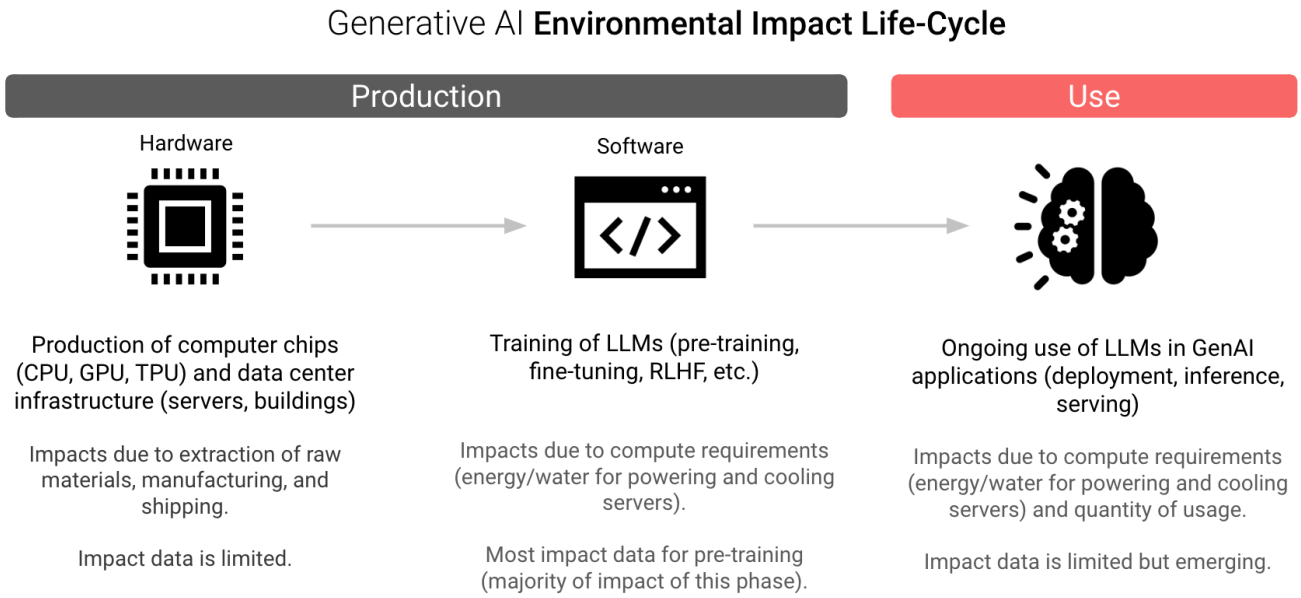

FIGURE 3.1 – Impact environnemental de l’IA générative [31]

Comme l’a mentionné le Parlement Européen ci-dessus, les LLM nécessitent des ressources énergétiques importantes liées à leur formation et à leur utilisation. Il est aussi important de se rappeler que le matériel informatique requis pour l’entraînement et l’hébergement du LLM nécessite une extraction de matières premières qui se traduit souvent par un impact écologique conséquent. L’ensemble des acteurs concernés par les technologies d’intelligence générative ignorent souvent cette dimension environnementale et seule l’Europe propose sérieusement d’imposer des réglementations sur l’efficacité énergétique des systèmes IA.

À mesure que ces modèles gagnent en complexité et en popularité, il devient de plus en plus difficile d'ignorer leur coût énergétique et leur impact environnemental. Sur Hugging Face, un référentiel de modèles d'IA, nous pouvons déjà observer que le nombre de paramètres d'un modèle d'IA génératif a une corrélation directe avec son efficacité énergétique. Comme nous pouvons voir sur le graphique ci-dessous, plus le modèle est grand, moins l'efficacité énergétique sera élevée.

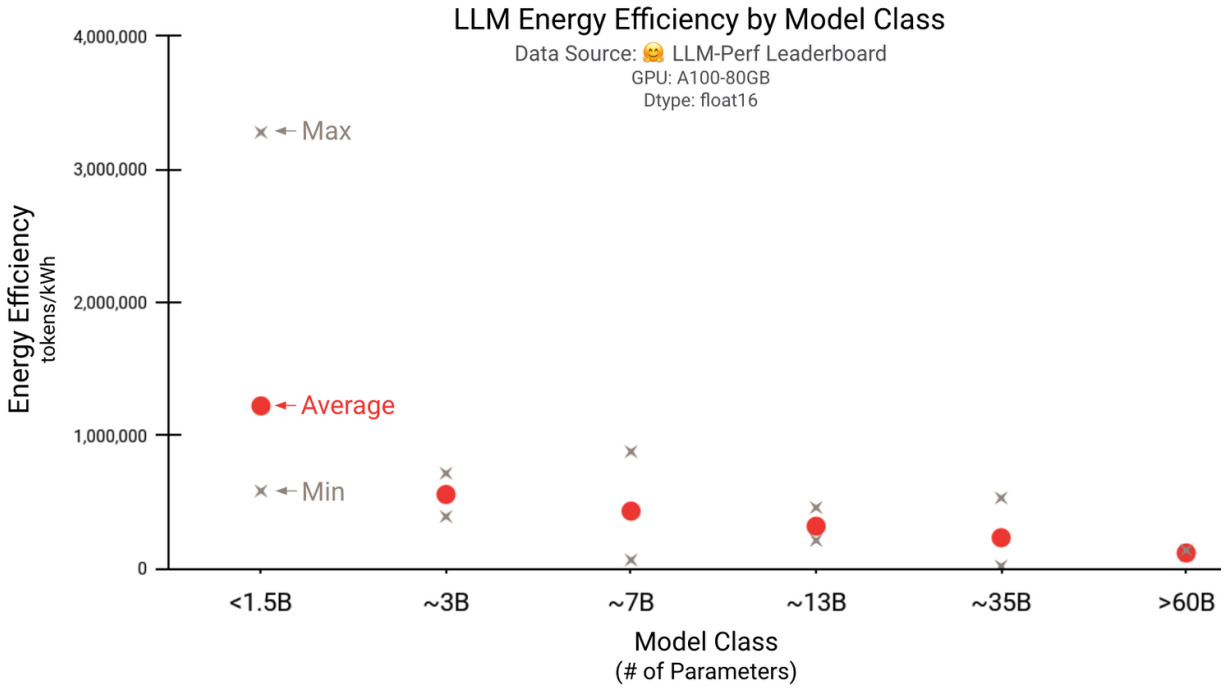

FIGURE 3.2 – Efficacité énergétique par nombre de paramètres [32]

La phase opérationnelle, en particulier l'inférence du modèle, consomme une quantité considérable d'énergie mais cet aspect reçoit moins d'attention que la phase d'entraînement. Des études telles que *From Words to Watts: Benchmarking the Energy Costs of Large Language Model Inference* [33], mettent en évidence les coûts énergétiques associés au fonctionnement de LLM comme LLaMA 65B sur différents GPU, montrant que même les tâches d'inférence peuvent être énergivores, en fonction de la complexité du modèle et du matériel utilisé. Sur la figure ci-dessous, les chercheurs ont estimé que le modèle LLaMA 65B consommait, pour une seconde d'inférence, entre 300 Watts et 1 Kilowatt, dépendant de la quantité de données d'entrées (batch size) et le nombre d'instances du modèle, actifs sur les deux processeurs graphiques utilisés lors de l'expérience (fig. 3.3).

Pour répondre à ces problèmes, les chercheurs et les développeurs des systèmes IA mettent en évidence certaines stratégies. Dans *A systematic review of Green AI*, un article publié par *Wiley Periodicals LLC*, les auteurs énumèrent les différentes études qui ont été menées dans l'optique de mesurer, optimiser et réduire la consommation énergétique des systèmes IA. Il s'agit notamment d'optimiser les méthodes d'entraînement du modèle, d'identifier et enlever les neurones redondantes, voir même de plafonner la puissance des processeurs graphiques, ce qui a été démontré dans l'article *From Words to Watts: Benchmarking the Energy Costs of Large Language Model Inference* (fig. 3.4). Cependant, la mise en œuvre de ces solutions demande aux développeurs des systèmes IA de faire des compromis en ce qui concerne la vitesse et la performance de leurs modèles tout en ajoutant des opérations supplémentaires potentiellement coûteuses telles que l'optimisation des méthodes d'entraînement et la mesure des émissions de carbone tout au long du cycle de vie du modèle. L'optimisation devient particulièrement essentielle lorsque nous travaillons avec des données géospatiales à l'aide d'un LLM. En effet, les données géospatiales sont beaucoup plus complexes et volumineuses que les données textuelles. Ainsi, pour traiter et visualiser ces données, une puissance de calcul accrue est nécessaire, entraînant par conséquent une consommation d'énergie plus élevée.

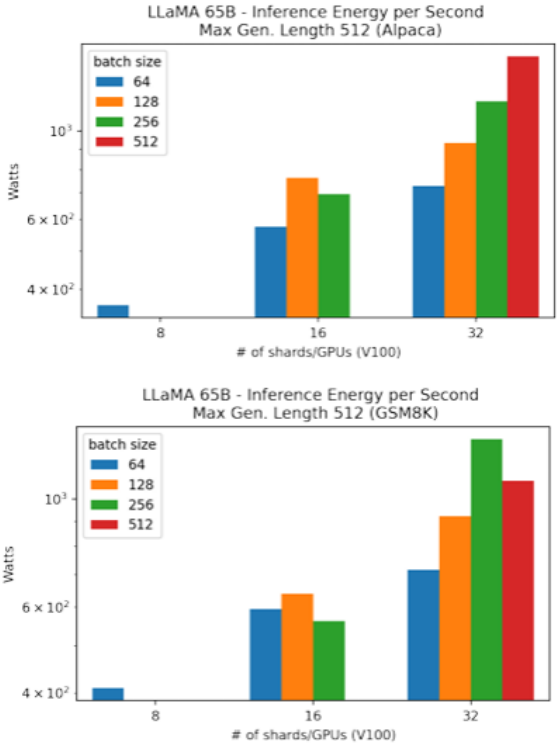

FIGURE 3.3 – Estimations de l'énergie par seconde (Watts) consommé par LLaMA 65B lors d'une opération de génération de texte [33]

Output length	Time % change		Energy % change		Token Rate % change	
	175W	150W	175W	150W	175W	150W
256	6.23	15.33	-21.82	-32.76	-5.87	-13.15
512	6.51	21.70	-23.95	-34.66	-6.11	-17.83
1024	7.40	21.65	-23.87	-34.59	-6.89	-17.80

FIGURE 3.4 – Effets du plafonnement de la puissance des processeurs graphiques sur les performances du modèle LLaMA 65B [33]

3.2 DÉPENDANCE ET CONSÉQUENCES SOCIO-ÉCONOMIQUES

Risques associés à une dépendance excessive aux technologies d'intelligence générative

Comme nous l'avons vu avec certaines innovations technologiques telles que Internet, les technologies d'intelligence artificielle pourraient transformer radicalement notre quotidien, à commencer par notre vie professionnelle. Le World Economic Forum [34], selon une enquête menée en 2020, estimait que 43% des entreprises étaient prêtes à réduire leurs effectifs en raison de l'intégration de nouvelles technologies et que d'ici 2025, le temps consacré par les humains et les machines sur des tâches quotidiennes au travail sera équivalent.

L'intégration croissante des LLM et autres systèmes d'intelligence artificielle dans les processus commerciaux quotidiens et les applications grand public vont sans doute entraîner d'importantes perturbations sur le marché du travail qui pourrait creuser les inégalités socio-économiques en augmentant de manière disproportionnée le potentiel de revenu et la productivité des personnes à revenu élevé par exemple [35].

La dépendance croissante et excessive à l'égard des LLM et l'ignorance des limites et risques que peuvent poser ces technologies pourrait conduire à la dissémination en masse de la désinformation, provoquer des problèmes juridiques conséquents et exposer ses utilisateurs à toute une série de vulnérabilités en matière de protection des données personnelles [36]. Il est également conseillé de prendre des précautions lorsqu'on autorise l'IA à prendre des décisions importantes ou automatiser des tâches sensibles à l'erreur, sous peine de conséquences désastreuses, particulièrement dans le cas où aucune vérification manuelle par un humain est prévue.

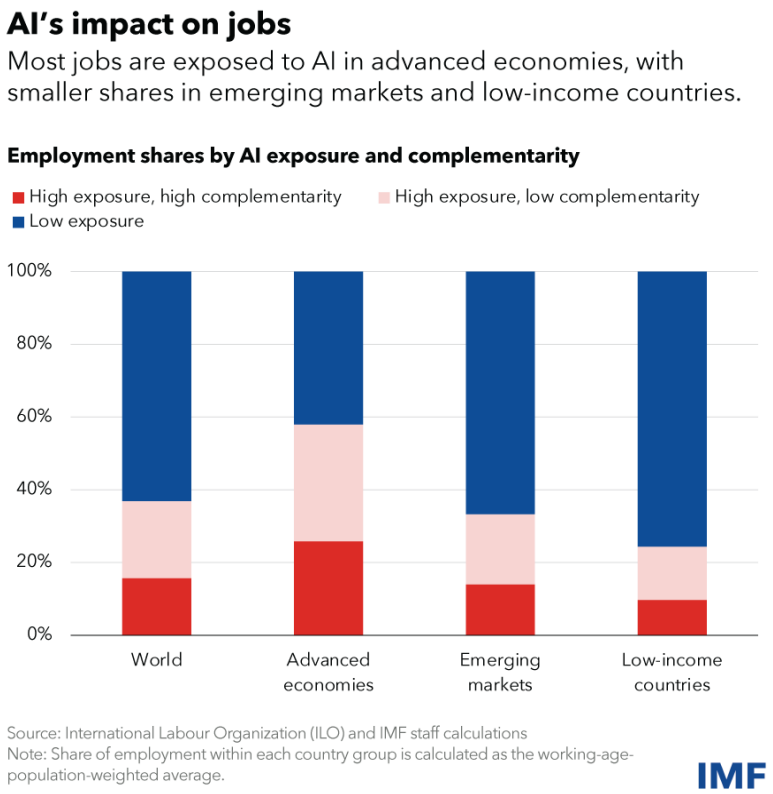

FIGURE 3.5 – L'impact de l'IA sur l'emploi selon le FMI [35]

CHAPITRE 4 :
CONSIDÉRATIONS GÉOSPATIALES

4.1 CONNAISSANCES LIMITÉES

En quoi le manque de connaissances spécifiques pourrait avoir un impact sur les compétences et les performances du LLM

La connaissance est l'ensemble des informations et des faits que nous possédons et que nous utilisons pour accomplir des tâches. Elle comprend les sujets que nous comprenons et les informations que nous connaissons. Posséder des connaissances implique que nous sachions comment réaliser une tâche, mais pas que nous puissions réellement le faire. C'est ce qui distingue les connaissances des compétences. À première vue, les LLM semblent être quasi-omniscients et extrêmement compétents dans la ou les tâches qui leur sont affectées par leurs développeurs. Or, les LLM ont globalement une connaissance générale sur un grand nombre de sujets et ne sont souvent que très compétents pour la génération de texte en langage naturel et, dans certains cas, relativement compétents sur d'autres tâches comme la génération de code par exemple.

Certaines tâches complexes, telles que l'analyse géospatiale, nécessitent à la fois une connaissance approfondie et spécifique du domaine géospatial et de bonnes compétences sur une variété de tâches qui peuvent se succéder et être interdépendantes. La complexité de tous les éléments liés au domaine géospatial, notamment les données, les outils et les connaissances associés, rend l'analyse géospatiale souvent trop complexe pour être déléguée à un LLM. Dans cette partie, nous allons d'abord nous focaliser sur les limites des connaissances géospatiales des modèles LLM et les solutions possibles pour combler ces lacunes et améliorer leur compétence sur des tâches géospatiales.

Même si certains LLM actuels, comme GPT-4, possèdent une connaissance plutôt approfondie du domaine géospatial, ils présentent encore des lacunes évidentes, restent enclins aux hallucinations et sont incapables d'effectuer des tâches géospatiales complexes sans aide externe. Dans le premier rapport, LLM et géodonnées [37], nous avons vu que les LLM sont capables d'exécuter des tâches géospatiales dites "simples", telles que la génération des coordonnées d'une ville à partir d'une requête [38], avec un niveau de précision qui est relativement bon mais pas idéal pour un cas d'utilisation concret.

Les chercheurs théorisent que les connaissances géospatiales limitées des LLM découlent d'un manque d'échantillons spatiaux dans les données de formation [39]. On peut également supposer que les connaissances géospatiales des LLM sont dérivées de données principalement ou entièrement textuelles qui leur ont été fournies au cours de la phase de formation. Ces données purement textuelles donnent certainement une bonne vue d'ensemble du domaine, mais ne permettent pas d'en avoir une compréhension holistique. L'une des solutions pour fournir au LLM une compréhension holistique des concepts géospatiaux a été présentée dans le premier rapport avec la technique du Geospatial Location Embedding ou GLE [40]. Pour rappel, cette technique consiste, durant la phase d'entraînement, à encoder des attributs spatiaux, tels que les coordonnées géographiques, les distances et les relations spatiales entre les lieux, en parallèle des données textuelles et dans une forme

compréhensible pour le LLM. Ces lacunes dans les connaissances du LLM pourraient être comblées en améliorant sa formation à des tâches spécifiques.

Le processus de collecte de données et de formation d'un LLM pouvant être extrêmement coûteux et complexe, une solution plus simple consisterait à faire appel à des données externes, comme nous l'avons montré dans le deuxième chapitre du premier rapport. Cependant, l'appel à des données externes présente ses propres inconvénients et limites. En effet, la qualité des données géospatiales disponibles est très variable, et chaque fournisseur a ses propres méthodes et normes pour la collecte, le traitement et la distribution des données. Chaque fournisseur établit également des licences spécifiques pour la distribution et l'utilisation de leurs données. Par exemple, certaines données sont ouvertes mais non commercialisables. Nous devons également tenir compte du fait que la nomenclature que nous utilisons pour les données et les métadonnées est conçue pour être lisible par un humain, ce qui signifie que peu ou pas de considération est accordée à la façon dont un LLM pourrait "lire" ces données. Sans effort de la part des fournisseurs pour normaliser leurs bases de données et appliquer des normes de stockage et de nomenclature adaptées à la requête et au traitement par les LLM, la capacité des LLM à effectuer une analyse géospatiale sur ces données externes restera limitée.

De plus, récupérer ces données externes requiert l'utilisation d'outils qui introduisent une couche de complexité supplémentaire. Jusqu'à récemment, seuls certains modèles propriétaires, tels que ChatGPT-4, ont été conçus et sont compétents dans l'utilisation d'outils grâce à l'appel de fonctions (function calling) [41]. Ce n'est que récemment, avec la sortie de LLama 3.1 [42] fin juillet 2024, que les modèles open source ont commencé à égaler ChatGPT dans l'utilisation d'outils. Nous ne devons cependant pas oublier qu'il est nécessaire de fournir au LLM des informations sur la manière d'utiliser les outils fournis, ce qui soulève la question de savoir comment harmoniser cette connaissance spécifique avec ses connaissances générales [43].

Les outils utilisés pour appeler des données externes nécessitent parfois que le LLM génère du code en SQL ou SPARQL, comme cela a été démontré dans l'article GeoQAMap - Geographic Question Answering with Maps Leveraging LLM and Open Knowledge Base [44]. Comme nous le savons, tous les modèles n'ont pas la même connaissance et compréhension des différents langages de programmation. Cela entraîne de grandes différences entre les modèles dans leur capacité à écrire du code cohérent pour exécuter une tâche spécifique telle que la requête de données géospatiales.

Une dernière solution que nous pouvons imaginer serait de lier le LLM à d'autres modèles de fondation spécialisés dans le traitement des données géospatiales tels que les *Geospatial Foundation Models* (GFM) [45]. L'idée serait que le GFM dispose des connaissances spécifiques et des compétences nécessaires pour exécuter certaines tâches géospatiales et que le LLM soit responsable du traitement

du langage naturel et de l'interaction avec l'utilisateur. Par exemple, Prithvi-100M [46] est un modèle de fondation entraîné sur un jeu de données constitué d'imageries satellite harmonisées provenant des satellites Landsat et Sentinel (Harmonized Landsat Sentinel-2). Ce modèle a été créé dans le cadre d'une collaboration entre IBM et la NASA et est capable d'effectuer des tâches d'imputation multitemporelle des nuages, de cartographie des inondations, de segmentation des cicatrices d'incendie et de segmentation multitemporelle des cultures.

Prithvi-100M répond à plusieurs limitations associées aux GFM grâce à un développement en deux étapes comprenant une phase de pré-entraînement et plusieurs étapes d'affinage pour chaque tâche (fig. 4.1). Dans la phase de pré-entraînement, l'apprentissage auto-supervisé (SSL) est utilisé à l'aide d'un autoencodeur (*masked autoencoder* ou MAE) qui masque jusqu'à 75 % des pixels d'une image satellite. Depuis cette image masquée, le modèle doit reconstituer l'image pixel par pixel. Cette méthode d'entraînement donne au modèle une bonne compréhension des caractéristiques et des relations spatiales au sein des données géospatiales.

Pour que le modèle puisse saisir la temporalité de certains phénomènes géospatiaux comme les inondations, les chercheurs ont inclus des images satellites de la même région prises à des moments différents. Cette technique permet au modèle d'apprendre comment les paysages évoluent dans le temps et de reconnaître les caractéristiques de certains phénomènes géographiques.

Suite à cette phase de pré-entraînement, le modèle est affiné pour des tâches spécifiques à l'aide d'un mélange de données étiquetées et non étiquetées. Comme le modèle a déjà été pré-entraîné, le processus d'affinage peut se focaliser sur l'apprentissage de la tâche finale que l'on souhaite que le modèle puisse accomplir. Par exemple, dans la tâche de détection des inondations, le modèle doit classer chaque pixel comme étant "sans données", "sans eau" ou "avec eau". Cette division du développement du modèle en deux étapes, et l'alignement étroit entre les tâches d'apprentissage auto-supervisé et la tâche finale, améliore les performances du modèle et la précision des résultats. En effet, les GFM sont généralement formés en une seule étape et le processus d'apprentissage implique parfois la répétition d'une tâche qui n'est pas nécessairement en lien avec la tâche qu'il devra effectuer en aval.

En termes de performances, Prithvi a obtenu des résultats supérieurs dans plusieurs tâches géospatiales. Par exemple, dans la détection des inondations, le modèle a atteint une précision globale d'environ 95%. Ces résultats sont comparables, voire supérieurs, à ceux obtenus par les modèles d'apprentissage profond traditionnels comme U-Net, qui affichent généralement des précisions globales de l'ordre de 85 à 96% pour des tâches similaires. Cependant, ces modèles d'apprentissage profond n'ont pas la même conscience spatio-temporelle des phénomènes géographiques et auraient du mal à généraliser leurs compétences aux régions du monde qui n'ont pas été abordées dans leurs données

d'apprentissage. Ces performances résultent, en partie, de la sélection méticuleuse des données et de leur qualité. Par exemple, l'ensemble de données sur les inondations comprend des images provenant de 14 biomes, 357 écorégions et 6 continents, couvrant 11 événements d'inondation. Cet ensemble de données diversifié garantit que Prithvi puisse s'adapter à différents contextes géographiques tout en préservant des performances élevées. Cependant, il est difficile d'obtenir des données géospatiales ouvertes de cette qualité et la sélection et l'étiquetage manuels des données sont particulièrement chronophages. C'est pourquoi Prithvi ne dispose que de 100 millions de paramètres, alors que les grands modèles de langage en possèdent des milliards.

Les modèles GFM, tels que Prithvi et Clay [47], nous montrent comment les chercheurs tentent de surmonter les limites et améliorer les modèles existants. Si nous parvenons à combiner de grands modèles linguistiques avec des modèles de fondation géospatiaux tels que Prithvi dans la même solution, nous pourrons créer une relation symbiotique dans laquelle le LLM détient la majorité des connaissances géospatiales, interagit avec l'utilisateur et est responsable de la résolution des problèmes, tandis que le GFM se focalise sur la maîtrise et l'exécution d'un certain nombre de tâches géospatiales sous la directive du LLM. Nous pouvons ainsi établir une passerelle entre la connaissance et la compétence pour surmonter certaines des limites exprimées dans cette partie.

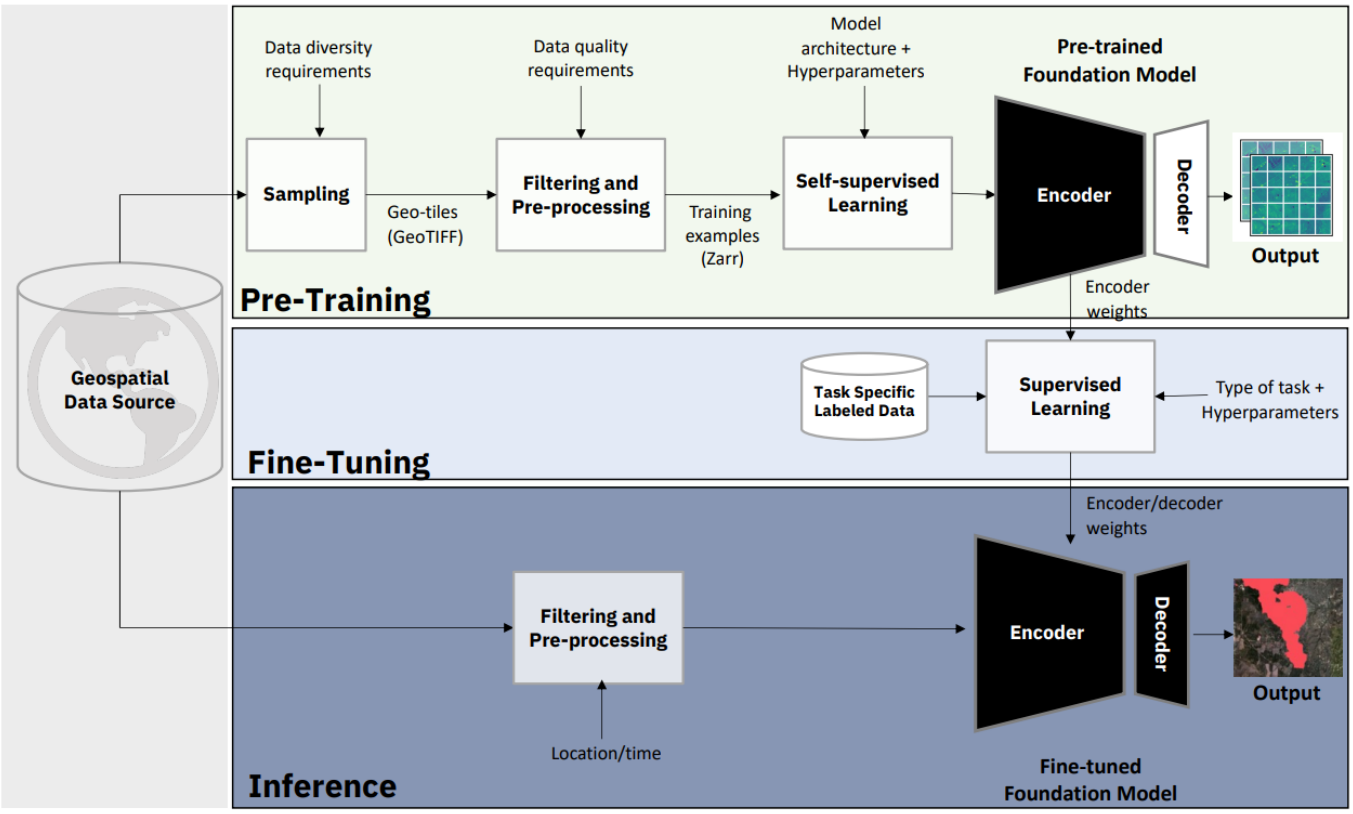

FIGURE 4.1 – Framework proposé pour l'entraînement, l'affinage et l'inference de Prithvi-100M [46]

4.2 LIMITATIONS TECHNIQUES ET TECHNOLOGIQUES

Quelles sont les limitations technologiques qui affectent la capacité des LLM à effectuer une analyse géospatiale ?

Comme nous l'avons vu tout au long de ce rapport, la majorité des limitations des LLM sont liées à leur architecture, à leur nature probabiliste inhérente et à la quantité et qualité de données de pré-entraînement. La performance des LLM dans l'exécution de tâches géospatiales varie de manière significative en fonction de ces facteurs, même lorsque les modèles ont le même nombre de paramètres. Ceci a été confirmé dans le rapport précédent lorsque nous avons évoqué une expérience où des chercheurs ont demandé à plusieurs LLM de générer les coordonnées d'une ville [38]. Grâce à cet essai, nous avons appris que différents modèles pouvaient produire des résultats très différents même si le nombre de paramètres était identique. En effet, les nuances dans l'architecture, les données de formation et les méthodes d'entraînement du modèle peuvent conduire à des différences conséquentes dans la façon dont la connaissance géospatiale est assimilée et mobilisée. La taille du modèle est également un facteur important, car les modèles qui détiennent un plus grand nombre de paramètres affichent généralement de meilleures performances. Cela s'explique par le fait que les modèles plus grands présentent généralement de meilleures capacités de raisonnement (voir partie 2.1) et disposent d'une base de connaissances beaucoup plus vaste.

Malgré leur potentiel, dans la quasi-totalité des essais présentés dans le premier rapport, un humain devait intervenir à un moment ou à un autre. Cette intervention est d'autant plus nécessaire pour les opérations géospatiales complexes où le modèle doit faire appel à des données externes ou générer du code. En effet, l'humain doit souvent guider le LLM en utilisant des techniques de *prompt engineering* ou d'affinage et vérifier que le LLM ne fasse pas d'erreur surtout si la solution à la question de l'utilisateur nécessite un raisonnement en plusieurs étapes. Il y a également un travail à faire en amont, comme l'intégration de documents d'orientation et des instructions détaillées, comme par exemple la liste et la description des outils qu'il peut utiliser, pour aider le modèle à effectuer les tâches demandées. L'analyse géospatiale implique souvent des sources de données, des algorithmes et des hypothèses complexes, ce qui nécessite une approche de résolution des problèmes par essais et erreurs. Cette stratégie reflète les méthodes humaines de résolution de problèmes, où les analystes explorent diverses solutions possibles et éliminent les pistes infructueuses. Ces tâches représentent un défi pour les humains donc nous ne pouvons pas, à l'heure actuelle, attendre à ce que le LLM fournisse une solution complète et précise à un problème complexe sans intervention par un être humain (concept du *Human-in-the-Loop* [48]).

Comme nous l'avons mentionné précédemment, l'incapacité de générer un code cohérent ou de le vérifier constitue également une limitation majeure, en particulier lorsqu'il s'agit d'extraire ou de traiter des données stockées dans une base de données. Les chercheurs qui ont développé GeoQAMap ont mentionné qu'ils devaient

souvent revoir le code généré par le LLM et vérifier qu'il avait bien appelé les bonnes données. En effet, le code généré pour appeler les données doit être précis et ne pas contenir d'erreurs. Cela peut s'avérer difficile avec les données géospatiales, car certains lieux géographiques peuvent avoir des noms qui prêtent à confusion pour le LLM. Par exemple, dans l'un des essais présentés dans Are Large Language Models Geospatially Knowledgeable [38], les chercheurs ont essayé de déterminer si le modèle pouvait correctement identifier des villes géographiquement proches lorsqu'il est sollicité avec des prépositions géospatiales (fig. 4.2). Pour ce faire, ils ont conçu une série de tests où le modèle doit compléter des phrases du type "<Ville-A> est proche de <Ville-B>", en utilisant diverses prépositions géospatiales. Les phrases de départ sont construites avec trois prépositions géospatiales : "proche de", "près de" et "loin de". Une expérience de contrôle est également menée en remplaçant la préposition géospatiale par la conjonction "et". Les chercheurs ont également créé des requêtes qui mentionnent l'état où se situe chaque ville, par exemple, "Albany, New York est proche de...". Nous remarquons qu'avec l'exemple de la ville d'Albany, la mention de l'état dans les requêtes permet d'augmenter considérablement la précision des résultats. Ceci est dû au fait qu'il existe une autre ville nommée Albany en Californie, ce qui a induit le LLM en erreur. Nous pouvons donc comprendre le problème lié à la façon dont les lieux géographiques peuvent avoir des noms similaires ou identiques, ce qui pourrait pousser le LLM à générer des données erronées.

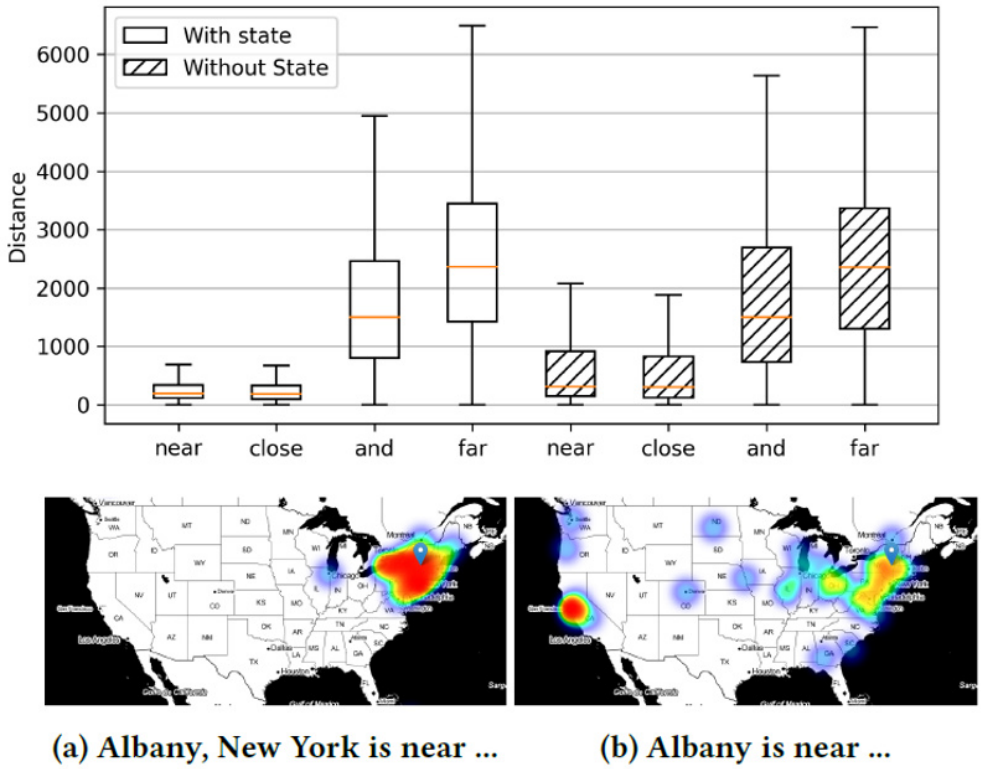

FIGURE 4.2 – Prédiction des distances entre les villes avec différentes prépositions et cartes thermiques des lieux générés avec "near". Les informations sur les États améliorent la désambiguïsation des villes [38]

Ce problème de noms ne se limite pas aux noms de villes et pourrait également s'appliquer aux autoroutes et aux rues portant le même nom et/ou le même numéro, mais situées à des endroits différents (l'Autoroute "A1" existe en Suisse et en France, pour donner un exemple). Ce facteur pourrait être davantage amplifié si le LLM a accès à un ensemble de données global et si plusieurs langues sont impliquées, en particulier celles qui n'utilisent pas l'alphabet latin. Par conséquent, compte tenu de son stade de développement actuel, l'intelligence artificielle générative a toujours besoin de conseils et d'aide humaine lorsqu'elle accomplit des tâches difficiles et à plusieurs étapes. Certaines solutions existent, comme nous l'avons vu dans la partie sur les hallucinations, où un deuxième LLM (ou un agent intelligent) peut être employé pour vérifier la réponse générée par un autre LLM. Cependant, comme avec les humains, des erreurs se produiront inévitablement, et nous devons définir notre niveau de tolérance vis-à-vis de ces erreurs et établir des stratégies pour les minimiser.

Enfin, l'analyse géospatiale ne se limite pas à répondre aux questions "quoi ?" et "où ?". Les LLM actuels se concentrent sur la manière d'exécuter les tâches, mais disposent d'une capacité très limitée pour répondre aux questions "pourquoi ?", car elles nécessitent des connaissances géospatiales approfondies et la faculté de formuler et tester des hypothèses [39]. La réponse aux questions "pourquoi ?", souvent nécessaire pour comprendre et expliquer des phénomènes géographiques complexes comme les flux (migratoires, démographiques, économiques, ...) ou les événements climatiques, implique la formulation d'un plan pour la recherche, d'une ou plusieurs hypothèses basées sur la question posée, une mobilisation autonome des données disponibles, et une compréhension du contexte (géographique, géopolitique, socio-économique, historique, etc...). Pour aller plus loin, les questions les plus difficiles auxquelles un LLM ou une personne pourrait répondre seraient des questions prédictives de type "Que se passera-t-il à cet endroit si telle chose se produit à cet autre endroit ?". Ce type de question combine tous les défis des questions de type "pourquoi" et ajoute l'incertitude associée à la prévision d'événements futurs.

Dans l'ensemble, les LLM sont remarquablement puissants et nous ne pouvons pas négliger leur potentiel pour l'analyse géospatiale. Nous devons néanmoins garder à l'esprit que la technologie de l'intelligence artificielle générative n'en est qu'à ses débuts et que les limites techniques et technologiques actuelles que nous avons observées sont inévitables lorsque nous travaillons avec des technologies de pointe. Il ne nous reste plus qu'à réaliser progressivement le potentiel de ces modèles et à développer de nouvelles stratégies et solutions pour surmonter leurs limites.

4.3 LIMITES DU SECTEUR GÉOSPATIAL

Évaluation du niveau de préparation du secteur géospatial pour les solutions d'intelligence artificielle générative

Selon la Commission Géospatiale (CG), un comité d'experts responsable de définir la stratégie géospatiale du Royaume-Uni, une statistique souvent citée indique que « les scientifiques des données dépensent 80 % de leur temps à rechercher et à préparer des données et seulement 20 % à effectuer des analyses et à générer des informations pertinentes » [49]. Si c'est le cas, pourquoi ne pas utiliser un LLM pour automatiser la recherche des données ? Les LLM sont, en effet, déjà capables de rechercher la donnée géospatiale de manière autonome mais se heurtent aux mêmes problèmes que les scientifiques des données mentionnés ci-dessous. D'après la CG, ce processus d'acquisition de données peut être comparé à la pyramide des besoins de Maslow (fig. 4.3), où un individu ne peut atteindre son potentiel (l'auto-actualisation) à moins que tous ses besoins de niveau inférieur ne soient d'abord satisfaits (par exemple, la nourriture et l'eau, la sécurité, les relations et le soutien émotionnel). De la même manière, pour atteindre le haut de la pyramide où nous pouvons effectuer une analyse de la donnée, nous devons: Savoir où chercher la donnée et où la trouver, déterminer si la donnée est disponible et accessible (licences, libre accès ou payant, etc.), s'assurer de la qualité et fiabilité de la donnée et s'assurer que le format ne pose pas de problème de compatibilité et Avoir la possibilité de joindre plusieurs jeux de données ensemble tout en maintenant la cohérence. Par exemple, si nous avons un jeu de données qui utilise les unités métriques pour mesurer les distances et un autre jeu de données qui utilise les unités impériales, nous devons pouvoir convertir les unités de mesure pour maintenir la cohérence de notre analyse. Cela permet de s'assurer que toutes les données sont comparables et peuvent être correctement intégrées pour des analyses précises et fiables.

L'efficacité de toute solution d'intelligence artificielle générative pour les analyses géospatiales dépendra donc de la qualité et de l'accessibilité des données qu'elle doit acquérir et traiter. Deux grandes organisations définissent actuellement les normes de l'industrie géospatiale : l'Open Geospatial Consortium et l'Organisation internationale de normalisation. Leurs normes sont généralement adoptées par les grandes organisations telles que les Nations Unies, mais les entités plus petites, telles que les gouvernements locaux, qui fournissent généralement des données à l'échelle locale, sont plus enclines à suivre leur propre système de normes, probablement en raison d'un manque de capacité financière et technique. Cela est particulièrement vrai pour les gouvernements des pays du Sud, qui ne possèdent pas l'infrastructure ou la capacité technique nécessaire pour recueillir, stocker et maintenir des données géospatiales normalisées, accessibles et qualitatives.

On pourrait évoquer qu'il manque parfois une incitation financière pour recueillir et maintenir certaines données, ce qui est souvent dû à une compréhension limitée de la plus-value que les données géospatiales peuvent apporter [50]. Ce manque de sensibilisation sur les avantages que peuvent apporter les données géospatiales pourrait

potentiellement restreindre la demande pour ces données et limiter l'innovation, car les entreprises risquent d'hésiter à investir dans la commercialisation d'une nouvelle donnée dont le coût et la rentabilité sont difficiles à évaluer. Pour revenir à la limitation de l'innovation, elle s'applique également au développement de solutions d'intelligence artificielle générative pour l'analyse géospatiale. En effet, si la demande pour ce type de solution est limitée en raison d'un manque de compréhension du domaine géospatial par le marché dans son ensemble, les entités privées cherchant à développer de telles solutions auront du mal à trouver le financement nécessaire.

Il existe un segment de l'industrie géospatiale qui a particulièrement réussi à susciter de l'intérêt et à attirer des financements importants de la part des investisseurs. Ce segment comprend des entreprises spécialisées dans l'observation de la Terre dont certaines lancent leurs propres satellites dans l'espace. Cela est possible grâce à la commercialisation de l'orbite terrestre basse, avec des avancées telles que les nanosatellites (*CubeSat*) et des coûts de lancement spatiaux plus abordables. Ce développement a permis de réduire le coût des images satellites de haute résolution et donc de les rendre plus accessibles à une plus large palette de consommateurs [51].

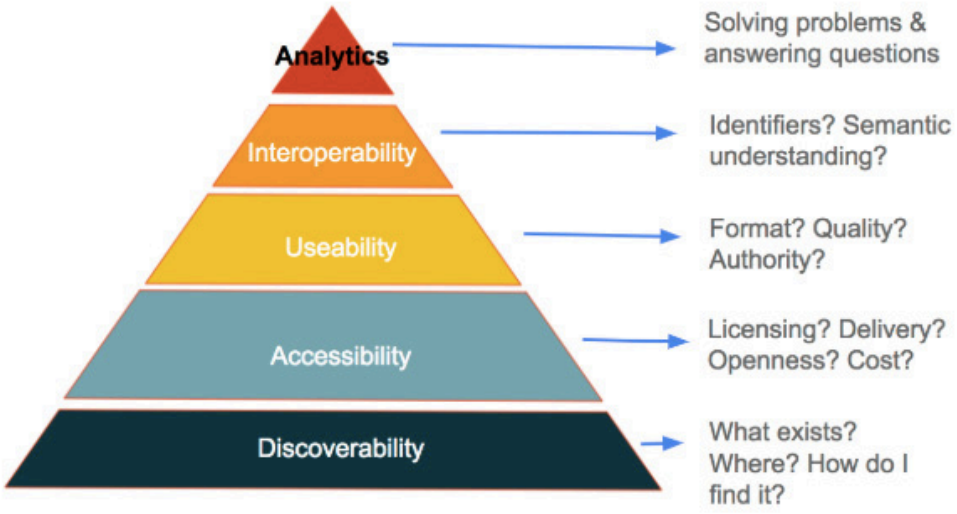

FIGURE 4.3 – Pyramide de Maslow pour l'analyse des données [49]

CONCLUSION

L'intégration de grands modèles de langage (LLM) dans le traitement des données géospatiales constitue une opportunité prometteuse pour améliorer l'analyse et l'interprétation des informations géospatiales. Ces données, qui englobent divers formats tels que l'imagerie satellitaire, les données des systèmes d'information géographique (SIG) et les données géoréférencées, nécessitent des techniques d'analyse sophistiquées pour en tirer des informations pertinentes et utiles à l'utilisateur final. Les LLM, avec leurs capacités avancées de traitement du langage naturel, offrent des avantages potentiels comme l'automatisation et la démocratisation de la donnée et l'analyse géospatiale.

L'une des principales contraintes techniques liées à l'utilisation des LLM pour le traitement des données géospatiales est la capacité de raisonnement des LLM et leur capacité à exécuter des opérations de manière séquentielle, deux éléments nécessaires pour effectuer de l'analyse géospatiale complexe. Les LLM sont généralement conçus pour traiter des entrées textuelles et leurs performances peuvent se dégrader avec l'augmentation de la complexité et du volume des données. La nécessité de gérer de vastes ensembles de données peut entraîner une diminution de la précision et une augmentation des cas d'hallucinations, où le modèle génère des informations incorrectes ou absurdes. Ces imprécisions peuvent être particulièrement problématiques dans les contextes géospatiaux où la précision est primordiale.

De plus, le traitement des données géospatiales à l'aide de LLM nécessite des ressources informatiques conséquentes. Les phases d'apprentissage et d'inférence exigent une puissance de calcul et des ressources énergétiques considérables, ce qui peut avoir une incidence sur l'efficacité des analyses géospatiales complexes à grande échelle. Cette contrainte souligne la nécessité d'optimiser les architectures des LLM et d'explorer des méthodes plus efficaces sur le plan énergétique pour équilibrer les performances et la consommation de ressources.

Les implications éthiques et sociales du déploiement des LLM dans le traitement des données géospatiales ne peuvent être négligées. Les questions de confidentialité des données et de partialité sont primordiales, étant donné que les données géospatiales peuvent contenir des informations sensibles et personnelles ou créées à partir d'informations confidentielles. Il est essentiel de veiller au respect des réglementations en matière de protection des données et de mettre en œuvre des mécanismes pour atténuer les biais, notamment géographiques. Le risque d'utilisation abusive des LLM pour produire des informations géospatiales trompeuses ou nuisibles est une autre préoccupation majeure. La transparence dans la formation des modèles, les sources de données et les processus décisionnels est essentielle pour maintenir la confiance du public et assurer un usage responsable de la technologie.

L'image ci-dessous (fig. 3.6) illustre la conformité des principaux fournisseurs de modèles de base avec le projet de loi sur l'IA de l'Union Européenne. Elle montre la conformité de différents modèles de langage (GPT-4, Cohere Command, Stable Diffusion v2, Claude 1, PaLM 2, BLOOM, LLaMA, Jurassic-2, Luminous, GPT-NeoX) par rapport aux exigences du projet de loi. Les résultats révèlent des différences significatives dans la conformité de chaque modèle aux exigences réglementaires et aucun modèle pour l'instant, même les LLM open-source, ne respecte tous les critères. Ceci met en évidence les défis et les progrès en matière de régulation de l'intelligence artificielle.

En encourageant la collaboration interdisciplinaire, l'innovation et la recherche, l'ensemble des parties prenantes peuvent exploiter les avantages des LLM dans le domaine de l'analyse géospatiale tout en atténuant leurs risques. Ces parties prenantes, présentes tout au long de la durée de vie du modèle, doivent assurer un déploiement et une utilisation éthique, sécurisée et durable des technologies

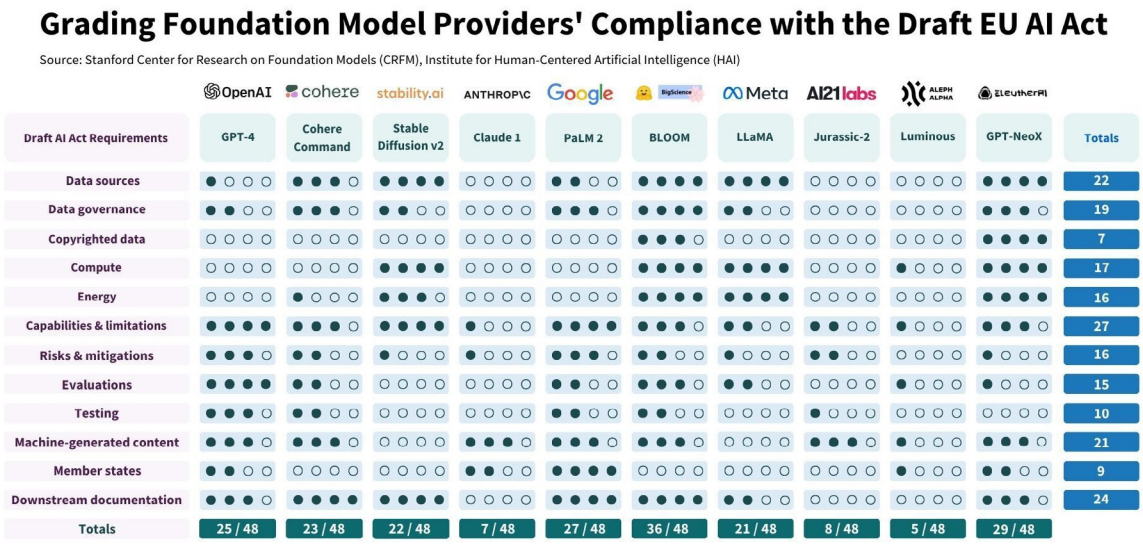

FIGURE 5 – Conformité des modèles LLM avec avec la législation européenne sur l'intelligence artificielle [52]

LISTE DES FIGURES

BIBLIOGRAPHIE

[1] Muhammad Usman Hadi, asem al ashi, Rizwan Qureshi, et al. (2023, July 10). A survey on large language models: Applications, challenges, limitations, and practical usage. TechRxiv.

[2] Gallegos, I. O., Rossi, R. A., Barrow, J., Tanjim, M. M., Kim, S., Dernoncourt, F., Yu, T., Zhang, R., & Ahmed, N. K. (2024). Bias and fairness in large language models: A survey.

[3] Funelas, R., 2024. ChatGPT Language Capabilities: The Breakdown. Tomedes. https://www.tomedes.com/translator-hub/chatgpt-language-capabilities

[4] Faisal, F., & Anastasopoulos, A. (2022). Geographic and geopolitical biases of language models. arXiv.

[5] BigScience Large Open Science Open-access Multilingual Language Model. (2022). https://huggingface.co/bigscience/bloom

[6] Saran, C. (2023). ChatGPT returns to Italy after OpenAI tweaks privacy disclosures, controls. CSO Online. https://www.csoonline.com/article/575219/chatgpt-returns-to-italy-after-openai-tweaks-privacy-disclosures-controls.html

[7] Condliffe, J. (2023). ChatGPT wrote a new poem, forever. Wired. https://www.wired.com/story/chatgpt-poem-forever-security-roundup/

[8] Yao, Y., Duan, J., Xu, K., Cai, Y., Sun, Z., & Zhang, Y. (2024). A survey on large language model (LLM) security and privacy: The good, the bad, and the ugly. High-Confidence Computing, 4(2), 100211.

[9] Amos, Z. (2023). What is FraudGPT? HackerNoon. https://hackernoon.com/what-is-fraudgpt

[10] Chen, C., & Shu, K. (2023). Combating misinformation in the age of LLMs: Opportunities and challenges.

[11] Liang, P., et al. (2023). Holistic evaluation of language models.

[12] Mitchell, M., et al. (2019). Model cards for model reporting. In Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* '19. ACM.

[13] KPMG. (2024). Generative AI consumer trust survey. https://kpmg.com/kpmg-us/content/dam/kpmg/corporate-communications/pdf/2024/kpmg-generative-ai-consumer-trust-survey.pdf

[14] Capgemini Research Institute. (2024). Creative and generative AI. https://www.capgemini.com/insights/research-library/creative-and-generative-ai

[15] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2023). Chain-of-thought prompting elicits reasoning in large language models.

[16] Kambhampati, S. (2024). Can large language models reason and plan? Annals of the New York Academy of Sciences, 1534(1), 15–18.

[17] Sun, J., et al. (2024). A survey of reasoning with foundation models.

[18] Huang, L., et al. (2023). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions.

[19] Xu, Z., Jain, S., & Kankanhalli, M. (2024). Hallucination is inevitable: An innate limitation of large language models.

[20] Surla, A. (2023). How to get better outputs from your large language model. NVIDIA Technical Blog. https://developer.nvidia.com/blog/how-to-get-better-outputs-from-your-large-language-model/

[21] Greyling, C. (2023). RAG + LLM = Context size. https://cobusgreyling.medium.com/rag-llm-context-size-6728a2f44beb

[22] Levy, M., Jacoby, A., & Goldberg, Y. (2024). Same task, more tokens: The impact of input length on the reasoning performance of large language models.

[23] Wegner, P. (2023). The leading generative AI companies. IoT Analytics. https://iot-analytics.com/leading-generative-ai-companies

[24] Jones, H. (2024). How startups are using AI. Kruze Consulting. https://kruzeconsulting.com/blog/how-startups-using-ai

[25] Ma, H. (2024). Marketing strategy of Open AI. Advances in Economics, Management and Political Sciences, 73. https://doi.org/10.54254/2754-1169/73/20231500

[26] RTS. (2023). L'Italie bloque le robot ChatGPT pour protéger les données personnelles. https://www.rts.ch/info/monde/13908803-litalie-bloque-le-robot-chatgpt-pour-proteger-les-donnees-personnelles.html

[27] Bilenko, M. (2024). Introducing Phi-3: Redefining what's possible with SLMs. Microsoft Azure Blog. https://azure.microsoft.com/en-us/blog/introducing-phi-3-redefining-whats-possible-with-slms

[28] Artificial Analysis. (2024). Llama 3 Instruct (70B): API provider benchmarking & analysis. https://artificialanalysis.ai/models/llama-3-instruct-70b/providers

[29] Tarraf, G. (2023). Everything you need to evaluate open-source (vs. closed-source) LLMs. L'Atelier BNP Paribas. https://atelier.net/insights/evaluating-open-source-large-language-models

[30] Parlement européen. (2023). Résolution du Parlement européen sur l'intelligence artificielle et le droit d'auteur, Amendement 81 Proposition de règlement Considérant 46. https://www.europarl.europa.eu/doceo/document/TA-9-2023-0236_FR.html

[31] Gamazaychikov, D. (2024). MPG + LLMs: Exploring the energy efficiency of generative AI. https://www.linkedin.com/pulse/mpg-llms-exploring-energy-efficiency-generative-ai-gamazaychikov/

[32] Moutawwakil, I., & Pierrard, R. (2023). LLM-Perf Leaderboard. Hugging Face. https://huggingface.co/spaces/optimum/llm-perf-leaderboard

[33] Samsi, S., et al. (2023). From words to watts: Benchmarking the energy costs of large language model inference.

[34] World Economic Forum. (2020). The future of jobs report 2020. https://www3.weforum.org/docs/WEF_Future_of_Jobs_2020.pdf

[35] Georgieva, K. (2024). AI will transform the global economy. Let's make sure it benefits humanity. IMF Blog. https://www.imf.org/en/Blogs/Articles/2024/01/14/ai-will-transform-the-global-economy-lets-make-sure-it-benefits-humanity

[36] Kundu, R. (2024). AI risks: Exploring the critical challenges of artificial intelligence. Lakera. https://www.lakera.ai/blog/risks-of-ai

[37] Ageospatial. (2024). LLM et Géodonnées: Approches, outils et méthodologies (Version de Mai 2024).

[38] Bhandari, P., Anastasopoulos, A., & Pfoser, D. (2023). Are large language models geospatially knowledgeable? In Proceedings of the 31st ACM International Conference on Advances in Geographic Information Systems (SIGSPATIAL '23). Association for Computing Machinery. https://doi.org/10.1145/3589132.3625625

[39] Li, Z., & Ning, H. (2023). Autonomous GIS: The next-generation AI-powered GIS. arXiv preprint arXiv:2305.06453. https://doi.org/10.48550/arXiv.2305.06453

[40] Tucker, S. (2024). A systematic review of geospatial location embedding approaches in large language models: A path to spatial AI systems. arXiv. https://doi.org/10.48550/arXiv.2401.10279

[41] OpenAI. Function calling. https://platform.openai.com/docs/guides/function-calling

[42] Meta AI. (2024). Introducing LLaMA 3.1. Meta AI. https://ai.meta.com/blog/meta-llama-3-1/

[43] Zhang, Y., Wei, C., Wu, S., He, Z., & Yu, W. (2023). GeoGPT: Understanding and processing geospatial tasks through an autonomous GPT. arXiv preprint arXiv:2310.14992.

[44] Feng, Y., Ding, L., & Xiao, G. GeoQAMap - Geographic Question Answering with Maps Leveraging LLM and Open Knowledge Base.

[45] Mai, G. (2024). Geo-Foundation Models. https://www.researchgate.net/publication/377981578_Geo-Foundation_Models

[46] Jakubik, J., Roy, S., Phillips, C. E., Fraccaro, P., Godwin, D., Zadrozny, B., Szwarcman, D., Gomes, C., Nyirjesy, G., Edwards, B., Kimura, D., Simumba, N., Chu, L., Mukkavilli, S. K., Lambhate, D., Das, K., Bangalore, R., Oliveira, D., Muszynski, M., Ankur, K., Ramasubramanian, M., Gurung, I., Khallaghi, S., Li, H. S., Cecil, M., Ahmadi, M., Kordi, F., Alemohammad, H., Maskey, M., Ganti, R., Weldemariam, K., & Ramachandran, R. (2024). Foundation models for generalist geospatial artificial intelligence. arXiv. https://doi.org/10.48550/arXiv.2310.18660

[47] Clay Foundation. Clay Foundation Model. https://clay-foundation.github.io/model/

[48] Google Cloud. Human-in-the-loop AI. https://cloud.g oogle.com/discover/human-in-the-loop

[49] Phillips, H. (2020). Getting the most from our national location data. Geospatial Commission. https://geospatialcommission.blog.gov.uk/2020/01/27/getting-the-most-from-our-national-location-data/

[50] Frontier Economics. (2020). Geospatial data market study. https://assets.publishing.service.gov.uk/media/5fb7aa488fa8f559e2153975/Frontier_Economics_-_Geospatial_Data_Market_Study.pdf

[51] Government Office for Science. (2019). Future technologies review. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/827507/Final_Version_-_Future_Technologies_Review.pdf

[52] Bommasani, R., Klyman, K., Zhang, D., & Liang, P. (2023). Do foundation model providers comply with the EU AI Act? Stanford Center for Research on Foundation Models. https://crfm.stanford.edu/2023/06/15/eu-ai-act.html

ANNEXE

AI Inherent Vulnerabilities and Threats fait référence aux vulnérabilités et risques inhérents aux modèles d'intelligence artificielle (architecture, fine-tuning, méthodes d'entraînement, etc.).

<table>
<tr><td rowspan="2">*Adversial attacks*: Ensemble de techniques et de stratégies utilisées pour manipuler ou tromper intentionnellement les modèles d'intelligence artificielle.</td><td>*Data poisoning:* Injection de données nuisibles dans les données de formation du LLM pour compromettre sa sécurité et son code éthique.</td></tr>
<tr><td>*Backdoor attacks:* Insertion d'une porte dérobée pour introduire des déclencheurs cachés qui pourront manipuler le comportement et les réponses du LLM.</td></tr>
<tr><td rowspan="2">Inference attacks: Utilisation de requêtes spécifiques pour provoquer des fuites involontaires de données personnelles dans les réponses du LLM.</td><td>*Attribute inference attacks:* Manipulation du LLM à l'aide de requêtes précises et de contenu en ligne publié par la victime pour en déduire des informations potentiellement sensibles (localisation, revenu, sexe, âge, etc.).</td></tr>
<tr><td>*Membership inference:* Type d'attaque qui cherche à déterminer si une partie d'ensemble de données fait partie des données d'apprentissage. Ceci permet donc de dévoiler la provenance des données d'entraînement.</td></tr>
<tr><td colspan="2">*Extraction attacks:* Types d'attaques plus directes et focalisées dans l'extraction de données personnelles et sensibles spécifiques.</td></tr>
<tr><td colspan="2">*Bias and unfairness exploitation:* Exploitation des biais existants dans les LLM pour produire du contenu discriminatoire et préjudiciable.</td></tr>
</table>

Instruction tuning attacks: Exploitation des vulnérabilités et limites des LLM affinés. Ces attaques ciblent les instructions ou les exemples de tâches utilisés pour affiner le modèle.	*Jailbreaking:* Désigne les méthodes utilisées pour détourner les protocoles de sécurité d'un LLM pour qu'il réponde à des questions potentiellement dangereuses.
	Prompt injection: Création et injection de requêtes pour manipuler le comportement d'un LLM.
	Denial of Service: Attaque par déni de service (ou DoS attack) est un type de cyberattaque dont l'objectif est de surcharger un système informatique pour le rendre inaccessible aux utilisateurs. Ce type d'attaque vise à exploiter la puissance de calcul élevée nécessaire à l'exploitation du LLM.

Non-AI Inherent Vulnerabilities and Threats: Regroupe les menaces externes que les LLM pourraient rencontrer notamment au niveau de toute l'infrastructure informatique nécessaire pour l'entraînement et l'exploitation du LLM, y compris les serveurs, les réseaux, les bases de données, les systèmes de stockage et les interfaces d'accès.

Remote code execution: L'exécution de code à distance (RCE) permet d'attaquer et de compromettre les applications, services web ou serveurs qui hébergent un LLM.
Supply chain vulnerabilities: L'utilisation de données, de services et de composants tiers comme les plugins peuvent compromettre la sécurité du LLM en introduisant de nouvelles failles et vulnérabilités. Par exemple, les plugins peuvent être utilisés comme point d'accès pour lancer des cyberattaques.

Les attaques par canaux auxiliaires sont utilisées pour extraire des informations sensibles en exploitant des tendances observées dans le comportement d'un système informatique, permettant ainsi d'identifier des failles.

Dans le contexte de l'entraînement des grands modèles de langage (LLM), on pratique souvent la déduplication des données pour éliminer les doublons. Cette opération vise à améliorer l'efficacité et à réduire le surapprentissage du modèle. Toutefois, si un attaquant connaît la fréquence typique d'occurrences d'une donnée avant la déduplication, il peut déduire que la réduction significative de cette fréquence indique la présence de multiples exemplaires de cette donnée dans l'ensemble des données d'entraînement.

Par exemple, en observant qu'il y a de nombreuses mentions de certaines maladies spécifiques après la déduplication, un attaquant pourrait en déduire que des dossiers médicaux ont été utilisés pour entraîner le modèle. En exploitant les vulnérabilités créées par la déduplication, l'attaquant peut tenter d'extraire ou de deviner ces documents, compromettant ainsi la confidentialité des informations médicales, telles que les diagnostics, les traitements, ou les informations d'identification personnelle des patients.

Les méthodes de défense pour les LLM se basent sur trois axes principaux : adapter l'architecture des modèles pour qu'ils soient plus résilients face aux attaques, optimiser le processus d'entraînement pour atténuer les attaques visant les données d'entraînement, et développer des systèmes de protection actifs.

<table>
<tr><td colspan="2">*Defense in model architecture:* Nous pouvons améliorer la résistance des LLM aux cyberattaques en adaptant l'architecture du modèle pour augmenter ses capacités de raisonnement et en adoptant de nouvelles méthodes d'entraînement.</td></tr>
<tr><td rowspan="2">*Defense strategies in LLM training:* Méthodes d'optimisation dans la collecte, le nettoyage et la génération de la donnée.</td><td>*Corpora cleaning:* Certaines données, notamment celles récupérées des sites internet, peuvent contenir du langage offensif, biaisé et parfois mensonger. Pour assurer la qualité des données d'entraînement, ces données de "mauvaise qualité" doivent être méticuleusement contrôlées et nettoyées par des processus de détoxification, débiaisement et d'anonymisation.</td></tr>
<tr><td>*Optimization methods:* Optimisation du processus d'entraînement pour améliorer le comportement du LLM, la sécurité du modèle et l'alignement éthique.</td></tr>
</table>

Defense strategies in LLM inference: Stratégies de défense active et dynamique implémentées lors du déploiement du modèle.	*Instruction processing:* Traitement et purification des instructions envoyées par les utilisateurs pour éviter que le LLM ne reçoive des instructions suspectes et potentiellement malveillantes.
	Malicious detection: Renforcer la capacité des LLM à détecter des instructions suspectes et potentiellement malveillantes à travers la détection de mots ou phrases atypiques.
	Generation processing: Permettre au LLM de vérifier le contenu généré et le modifier avant de l'envoyer à l'utilisateur.

Ageospatial Sàrl