Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

26 avril 2024 V1.0[1]

Aide-mémoire de sensibilisation en matière de grands modèles de langage (*large language models,* LLM) au sein de l'administration fédérale

Les grands modèles de langage, qu'est-ce que c'est?

Les grands modèles de langage (*large language models,* LLM) sont des technologies recourant à l'intelligence artificielle (IA) qui sont à la base des systèmes d'IA[2] les plus courants dans le domaine du langage naturel. De nombreux outils d'IA génératifs utilisent ces LLM[3].
Les grands modèles de langage sont en mesure de traiter, de comprendre, d'interpréter et de générer du langage naturel. Ils sont en particulier capables d'accomplir une grande variété de tâches, et notamment de traduire, de comprendre ou de générer des textes. Entraînés avec de grandes quantités de données dans les formats les plus divers, les LLM les plus récents sont à même de produire des textes qu'il est souvent difficile de distinguer de premier abord d'un texte écrit par l'homme. On peut aussi envisager des inputs sous forme de sons ou d'images, puisque l'IA permet généralement de les convertir en texte avec une grande précision. De même, les voix générées par l'IA ressemblent de plus en plus à une voix humaine. Certains LLM constituent déjà des modèles multimodaux, qui sont par exemple en mesure de traiter et de générer non seulement du texte, mais aussi des images et des vidéos[4].
Quelques exemples de LLM: les séries de modèles GPT (utilisés dans ChatGPT d'OpenAI et Copilot de Microsoft), Gemini (utilisé dans Gemini de Google, anciennement Bard), les modèles LLaMA de Meta, les séries de modèles Grok de X ainsi que les modèles Claude d'Anthropics[5]. Pour acquérir la capacité d'interpréter et de générer du langage naturel, les LLM «apprennent» des relations statistiques à partir de textes qu'ils traitent lors d'un processus d'entraînement itératif exigeant une grande puissance de calcul. Ces modèles statistiques reposent sur des techniques et des méthodes de traitement du langage naturel (*natural language processing,* NLP[6]) qui permettent d'extraire la signification et les corrélations du langage humain.

Principes relatifs aux grands modèles de langage (LLM) au sein de l'administration fédérale[7]

En tant qu'utilisateur de technologies d'IA génératives (et notamment des LLM) au sein de l'administration fédérale, nous vous invitons à appliquer les principes suivants. Vous serez ainsi en mesure de faire vos propres expériences de manière responsable. Les technologies d'IA génératives (comme les LLM) peuvent vous aider dans

---

1 Cette fiche est mise à jour périodiquement afin d'intégrer les développements les plus récents et de rendre plus clairs les technologies d'IA générative (comme les grands modèles de langage) au sein de l'administration fédérale.

2 Un système IA (*AI system*) est un système automatique capable d'inférer, sur la base des «inputs» (entrées) qu'il reçoit et pour des objectifs explicites ou implicites, comment générer des «outputs» (résultats) tels que des prévisions, des contenus, des recommandations ou des décisions et qui, ce faisant, peut exercer une influence sur des environnements physiques ou virtuels. Les systèmes IA peuvent être dotés d'une autonomie plus ou moins grande. Une technologie d'IA désigne des fonctions isolées exécutables dans un ordinateur pour obtenir de l'intelligence artificielle («apprentissage automatique», p. ex.). Un système IA désigne ainsi une combinaison structurée et contextuelle de technologies d'IA en vue d'atteindre l'intelligence artificielle. *Source:* https://cnai.swiss/fr/products/terminologie/

3 L'«IA générative» est une notion large qui se réfère aux systèmes d'IA entraînés sur la base de grandes quantités de données provenant du monde physique et virtuel afin de générer des données de manière autonome (textes, images, enregistrements sonores, vidéos, simulations, codes, etc.). Ils sont souvent multimodaux, avec par exemple des inputs et/ou des outputs selon une ou plusieurs modalités (texte, image, vidéo, etc.). *Source:* https://cnai.swiss/fr/products/terminologie/

4 Aussi appelés grands modèles d'IA multimodaux («*large multimodal models*», LMM) à capacité générative (https://huyenchip.com/2023/10/10/multimodal.html).

5 Voir la «Fiche technique sur l'utilisation d'outils d'IA générative au sein de l'administration fédérale»; https://cnai.swiss/fr/products-autres-services-aide-memoire-pour-lia/

6 Le traitement automatique du langage naturel (NLP, de l'anglais *natural language processing*) est un domaine de l'IA qui traite des activités consistant à analyser, à comprendre et à générer des mots et des phrases orales ou écrites (langue naturelle). La plupart des techniques et des méthodes de NLP se basent sur l'«apprentissage automatique» pour extraire la signification et les corrélations du langage humain. Les applications de NLP comprennent par exemple la reconnaissance de textes («*text recognition*»), la reconnaissance de la parole («*speech recognition*»), les robots logiciels, les agents conversationnels («*chatbots*») et les assistants numériques. *Source:* https://cnai.swiss/fr/products/terminologie/

7 Inspiré du site «*Generative AI Framework for UK Government*» (consulté le 18 janvier 2024); voir https://www.gov.uk/government/publications/generative-ai-framework-for-hmg

vos tâches administratives quotidiennes. Ils sont en mesure de prendre en charge une multitude de tâches textuelles de manière partiellement ou entièrement automatisée. Osez faire le pas, mettez l'IA à l'épreuve! Avec un peu de créativité, vous contribuerez ainsi à rendre l'administration plus innovante. Mais prudence est mère de sûreté, et il vous faudra impérativement respecter les points ci-dessous.

1) Vous savez ce qu'est l'IA générative et ce que sont les LLM. Et vous connaissez leurs limites.
2) Vous utilisez les technologies d'IA génératives (et notamment les LLM) de manière responsable, en toute légalité et en respect des principes éthiques. Vous vérifiez s'il existe une base juridique suffisante pour l'utilisation prévue. Les exigences légales diffèrent selon que l'on recourt aux technologies d'IA génératives (les LLM, p. ex.) pour s'aider dans le but de traduire un texte, pour une aide plus substantielle lors de la rédaction d'un texte ou pour prendre une décision entièrement automatisée.
3) Vous savez comment utiliser les technologies d'IA génératives (comme les LLM) de manière sûre[8].
4) Vous assurez un contrôle humain approprié à chaque stade, du développement des technologies d'IA génératives (comme les LLM) à leur application.
5) Vous utilisez l'IA en général, et les LLM en particulier, en toute transparence. Un texte rédigé par un LLM peut donner l'impression d'avoir été écrit par l'homme. En travaillant au sein de l'administration fédérale, vous endossez une responsabilité particulière en matière de communication, que ce soit à l'interne ou à l'externe. Nous vous recommandons de signaler clairement si un texte a été généré à l'aide de l'IA (p. ex. «Ce texte a été généré à l'aide de l'intelligence artificielle, puis traité et vérifié par l'homme.»). Lors de décision individuelle automatisée au sens de l'art. 21 LPD[9], cette information est obligatoire.
6) Vous appliquez ces principes, mais aussi les lignes directrices en matière d'utilisation de l'IA au sein de l'administration fédérale et les principes de base d'une science des données (et d'une IA) centrée sur l'être humain et digne de confiance, tout en mettant en œuvre une assurance qualité appropriée.

Limites des grands modèles de langage (LLM)

Les LLM génèrent leurs résultats (outputs) en misant sur la probabilité statistique du mot suivant, et non sur la véracité du contenu. Ce procédé peut conduire à des énoncés factices, voire à des hallucinations. Autrement dit, il se peut que l'IA produise des résultats ne correspondant en rien à la réalité. Il est certes possible de limiter ce phénomène en fournissant au modèle un contexte (des documents, p. ex.). Reste que là aussi, il est possible d'obtenir des hallucinations dès lors que les outputs comprennent des contenus qui dépassent le contexte donné. Dans le cadre d'une utilisation appropriée, d'autres limites des LLM sont par exemple[10]:

- des résultats (outputs) non souhaités, des répétitions et des distorsions dans le modèle («*bias*») résultant d'une compilation déséquilibrée des contenus et des données utilisés pour l'apprentissage (données personnelles ou protégées par le droit d'auteur, textes au contenu douteux, erroné ou discriminatoire, etc.);
- des résultats obsolètes, les LLM générant du texte sur la base des données d'entraînement, qui se limitent forcément à des contenus disponibles au moment de l'entraînement du modèle;
- un manque de reproductibilité, car même si un LLM reçoit plusieurs fois un même input, l'output généré pourra différer de fois en fois tant sur le plan linguistique que sur le plan du contenu;
- un manque de sécurité du code généré (si un LLM a appris à créer du code, il peut aussi en générer qui contient potentiellement des failles de sécurité);

[8] Voir à ce sujet la «Fiche technique sur l'utilisation d'outils d'IA générative au sein de l'administration fédérale»; https://cnai.swiss/fr/products-autres-services-aide-memoire-pour-lia/

[9] https://www.fedlex.admin.ch/eli/oc/2022/491/fr#art_21

[10] Nous nous basons sur le chapitre «2.3 Risques des LLM» du rapport (version 1.1 du 27 mars 2024) du Bundesamt allemand de la sécurité dans la technologie de l'information («*Bundesamt für Sicherheit in der Informationstechnik*») sur les opportunités et les risques des LLM; voir https://www.bsi.bund.de/SharedDocs/Downloads/DE/BSI/KI/Generative_KI-Modelle.pdf

- une réaction erronée à des inputs spécifiques, car un écart aussi insignifiant soit-il dans les inputs peut entraîner de grandes différences dans les outputs. Si par exemple les inputs fournis à un LLM diffèrent des textes utilisés pour l'apprentissage, le modèle ne sera souvent plus en mesure de les traiter correctement et génèrera des outputs erronés;
- en règle générale, les LLM génèrent des textes irréprochables du point de vue de la langue, convaincants sur le plan du contenu et pertinents dans les domaines les plus variés. Cela peut donner l'impression que les capacités de la machine sont semblables à celles de l'homme, et donc inciter l'utilisateur à accorder une trop grande confiance aux outputs du modèle et à ses aptitudes («*automation bias*»);
- une réutilisation fréquente des données saisies à titre d'entraînement: des données erronées sont donc susceptibles d'entrer dans le processus d'apprentissage;
- une vulnérabilité à l'interprétation d'un texte comme instruction, les LLM interprétant tous les inputs de la même manière, sans faire de distinction entre les instructions et les autres textes;
- un manque de confidentialité des données saisies, les LLM étant souvent proposés en tant que service sur Internet au moyen d'interfaces appropriées (via un navigateur Web, p. ex.). La saisie de données personnelles dans un tel service via Internet peut constituer une transmission de données à des tiers. La loi sur la protection des données s'applique;
- le développement et l'exploitation de LLM par un fournisseur sur son infrastructure peut entraîner une forte dépendance (p. ex. manque de souveraineté des données, faible contrôlabilité du modèle);
- les LLM peuvent générer des références inexactes, voire inexistantes, et s'y tenir coûte que coûte.

Les technologies d'IA génératives (et notamment les LLM) offrent de nombreuses opportunités et des applications intéressantes. Leur développement actuel est très dynamique. Cette évolution impose de nouveaux impératifs en matière de sécurité, que ce soit pour le développement, l'exploitation ou l'utilisation de ces modèles. Pour utiliser ces technologies d'IA en toute sécurité, il est essentiel d'analyser les risques de manière systématique.

Dans ce sens, les personnes travaillant pour l'administration fédérale sont tenues de procéder à une analyse de risque pour chaque technologie d'IA générative (LLM) qu'elles souhaitent intégrer dans leurs processus de travail. Pour ce faire, elles peuvent se référer aux principes de cet aide-mémoire, et au final faire leurs propres expériences de manière responsable.

En cas de questions:

- sur l'utilisation de l'IA au sein de l'administration fédérale: groupe de travail IA et pôles de compétence du Réseau de compétences en intelligence artificielle – CNAI;
- sur la sécurité informatique et la protection des données: les ISBO et les PPDO de votre unité administrative;
- sur des services concrets dans le domaine de la science des données et de l'IA: DSCC.

Autres recommandations:

Lignes directrices en matière d'utilisation de l'intelligence artificielle au sein de l'administration fédérale

Les sept lignes directrices en lien avec l'IA à la Confédération restent en vigueur: placer l'être humain au cœur des préoccupations; créer des conditions propices au développement et à l'utilisation de l'IA; assurer la transparence, la traçabilité et l'explicabilité; déterminer la responsabilité; assurer la sécurité; promouvoir la participation active à la gouvernance en matière d'IA en impliquant tous les acteurs pertinents aux plans national et international.

Code de bonnes pratiques de la Confédération pour une science des données (et pour une IA) centrée sur l'être humain et digne de confiance

Le code de bonnes pratiques sensibilise les unités administratives de la Confédération aux principes de base d'une science des données (et d'une IA) centrée sur l'être humain et digne de confiance. Il fournit des exemples pratiques et contribue à une application dans le travail quotidien. Ses principes fondamentaux sont: la protection des données et de l'information; la sécurité de l'information; la sécurité des données; la gouvernance des données; la non-discrimination; l'explicabilité; la traçabilité; la transparence; la reproductibilité; la neutralité; l'objectivité ainsi que le traitement éthique des données et des résultats.

Le présent aide-mémoire a été élaboré au sein du groupe de travail «IA au sein de l'administration fédérale» du CNAI avec la participation de représentants de tous les départements et de la ChF.