



HUB
FRANCE
IA

LES RISQUES DE L'IA GENERATIVE

Juillet 2024



Table des matières

1. Introduction	4
2. Définitions	8
2.1. IA Générative	9
2.2. Risque.....	9
2.3. Parties prenantes.....	10
2.4. Causes	12
2.4.1. Les Données	12
2.4.2. Le Modèle.....	13
2.4.3. L'Humain.....	14
2.5. Impacts.....	14
2.5.1. Impact juridique.....	14
2.5.2. Impact financier.....	15
2.5.3. Impact opérationnel.....	15
2.5.4. Impact réputationnel	15
2.5.5. Impact organisationnel.....	16
2.5.6. Impact social.....	16
2.5.7. Impact environnemental	17
2.6. Criticité	18
3. Démarche d'analyse des risques	20
3.1. Présentation générale.....	21
3.2. Matrice des risques	22
4. Catégories d'usages.....	24
4.1. Catégories d'usage transverse.....	25
4.1.1. Agent conversationnel.....	25
4.1.2. Recherche augmentée.....	25
4.1.3. Transformateur de contenu	26
4.1.4. Générateur de contenu.....	27
4.1.5. Générateur de code.....	27



4.1.6. Analyse de la donnée	28
4.2. Avant-goût : les catégories d'usages à venir	28
5. Synthèse des remédiations	30
5.1. Réduire les risques liés au modèle	32
5.1.1. Limiter la génération de contenu non désirable.....	32
5.1.2. Se protéger des tentatives malicieuses (incl. prompt injection, jailbreak)	39
5.2. Réduire les risques liés aux données utilisées par le modèle	45
5.2.1. Limiter la génération de contenu sensible, confidentiel, ou personnel.....	45
5.2.2. Se protéger de la génération de contenu protégé légalement	54
5.3. Réduire les risques liés à une mauvaise utilisation de l'IA générative.....	55
5.3.1. Garantir une connaissance suffisante pour l'utilisation des outils d'IA générative.....	55
5.3.2. Assurer la continuité en cas d'indisponibilité des outils d'IA générative	57
5.4. Récapitulatif	58
6. Conclusion générale.....	61
7. Glossaire.....	63
8. Remerciements.....	66

1. Introduction



1. Introduction

En mai 2023, le Hub France IA, avec ses membres, publiait une note de synthèse¹ pour éclairer les **enjeux de la révolution ChatGPT**. Dans une partie dédiée aux usages, nous avons d'abord identifié les quatre grandes typologies d'usage basées sur les capacités de rédaction, de classification, de traduction et de synthèse. Nous avons ensuite décrit des usages sur onze grands domaines : relation client et marketing, développement informatique, cybersécurité, banque et assurance, BTP, recherche, enseignement, journalisme, ressources humaines, juridique et santé. Nous avons ensuite décrit les risques de l'introduction de l'IAG (Intelligence Artificielle Générative) dans l'entreprise et les impacts à en attendre. Ce premier document avait été construit quelques mois seulement après l'annonce de ChatGPT fin octobre 2022. ChatGPT était alors apparu comme le premier représentant des agents conversationnels exploitant un grand modèle de langage (**Large Language Model** ou **LLM**), dont le nombre s'est largement accru depuis, avec de nouveaux usages.

En janvier 2024, le Hub France IA et ses membres poursuivent leur travail en publiant un livre blanc² détaillant des **exemples d'usages** possibles avec une IA générative basée sur un grand modèle de langage. Le document est constitué de trois axes. Le premier axe est consacré aux **usages autour de six grands domaines** : cybersécurité, industries culturelles et créatives, ressources humaines, développement informatique, éducation et marketing. Le second axe, étudie l'apport des LLM pour les **agents conversationnels (chatbots)**, un des grands usages de ChatGPT, à travers l'analyse de quelques retours d'expérience pour évaluer les gains et limites d'usage de ChatGPT pour cette fonction de *chatbot*. Enfin, le troisième axe, à travers une **enquête**, analyse les gains, manques à gagner et freins dans les usages actuels des IA génératives.

¹ Groupe de travail ChatGPT. ChatGPT : Usages, Impacts et recommandations. Note de synthèse. Hub France IA. Mai 2023. https://www.hub-franceia.fr/wp-content/uploads/2023/04/ChatGPT_Note-synthese.pdf

² Groupe de travail IA Générative. Livre blanc : les usages de l'IA Générative, Volume I – Les LLM. Hub France IA. Janvier 2024. https://www.hub-franceia.fr/wp-content/uploads/2024/02/Livre-blanc_Les-usages-de-lia-generative-01.2024.pdf



Face à l'engouement pour l'IA générative, au sein des organisations, mais également chez le grand public, le Groupe de Travail du Hub France IA a décidé de poursuivre ses travaux et se penche, à travers ce document, sur les **risques générés par l'usage des LLM** au sein des organisations. Comprendre ces risques est critique si on veut mesurer les impacts de l'IAG et si on veut mettre en place des pistes de remédiation pour en maximiser les bénéfices.

En effet, l'IA générative, bien que puissante et révolutionnaire, rencontre aussi des défis et des limites. Par exemple, nous pouvons noter des interrogations sur la qualité et la cohérence du contenu généré, les biais et inexactitudes potentielles héritées des données d'entraînement, le manque de créativité et d'originalité par rapport aux humains ainsi que des questions éthiques.

Note : Bien que les LLM ne représentent qu'un type de systèmes d'IA générative parmi d'autres, ce type de modèles est aujourd'hui le plus répandu dans les organisations, devant les autres comme la génération d'images ou de sons ; aussi, la gestion des risques de ce type d'IA générative tend à être plus avancée. C'est pourquoi ce document traite essentiellement des risques causés par les LLM en particulier, à travers de multiples cas d'usages. Seule exception, le domaine des Industries Culturelles et Créatives qui illustre des exemples de risques causés par d'autres types de contenus générés comme l'audio, l'image ou la vidéo.

Le présent document se décompose en quatre grands chapitres :

Un premier chapitre illustre les **définitions générales des concepts** présentés dans le document, facilitant ainsi la lecture des chapitres suivants. Les définitions traitées sont : le risque, les parties prenantes, les causes, les impacts ou encore la criticité.

Le deuxième chapitre expose une **démarche d'analyse des risques** spécifiques aux IA génératives au sein d'une organisation. Il ne s'agit pas de présenter la méthodologie générale et transverse du pilotage des risques en entreprise, ni particulièrement des risques liés au système d'information (SI), mais bien de présenter les spécificités propres à l'usage de l'IA générative qu'il convient d'intégrer dans son processus de pilotage des risques.

Le troisième chapitre présente une démarche d'analyse des risques, avec les grandes catégories d'usages des LLM, quel que soit le domaine ; ensuite, on introduit une matrice des risques, permettant d'analyser les risques liés à un cas d'usage d'une IA générative dans une organisation. Enfin, nous présenterons **de multiples cas d'usages**, classés par domaine. A ce jour, cette section n'est pas traitée ici : les cas d'usages seront



progressivement publiés après la parution de ce document, durant les mois de juillet et août. Les domaines qui seront traités sont les suivants : Ressources humaines, Finances, Santé, Conseil, Data Science, Développement logiciel, Commerce, Juridique, Marketing, Service client, Cybersécurité, Industries culturelles et créatives et Logistique & transport. Chaque exemple de cas d'usage est présenté de sorte qu'il mette en évidence le(s) risque(s) généré(s) par le LLM employé. Pour chaque cas d'usage, la démarche d'analyse est concrètement exécutée sur l'exemple.

Le quatrième et dernier chapitre propose une **synthèse des remédiations** propres à chaque type de risque et ses causes associées.

Nous espérons que ce nouvel éclairage permettra d'éveiller ou de renforcer votre vigilance face à cette nouvelle technologie, à la fois prometteuse et dangereuse, si mal employée.

2.

Définitions

Contributeurs :

- **Yael Suissa, CEO & Cofondateur – MAP-Monitoring And Protection**
- **Imen Fourati, Expert lead, Risque de modèle, Société Générale**
- **Benjamin Bosch, Manager – Model risk Management – Data Science, Société Générale**
- **Nicolas Pellissier, Cofondateur – Klark**
- **Martin D'Acremont, Consultant – Wavestone**



2. Définitions

2.1. IA Générative

Le parlement européen définit l'intelligence artificielle comme « tout outil utilisé par une machine afin de reproduire des comportements liés aux humains, tels que le raisonnement, la planification et la créativité. »

En particulier, l'intelligence artificielle générative (IAG) est une technologie qui a connu une évolution fulgurante ces dernières années. Il s'agit d'un sous-ensemble du Machine Learning, et plus particulièrement du Deep Learning, visant à produire du contenu, que ce soit du texte, une image, de l'audio ou une vidéo, à partir de données en entrée (on parle alors de *prompt*), elles-mêmes du texte, une image, de l'audio ou une vidéo par exemple. Un système d'IA générative crée un nouveau contenu statistiquement cohérent avec les données d'entraînement et la requête formulée.

Les systèmes d'IA générative sont généralement entraînés sur un large ensemble de données, nécessitant des moyens conséquents pour leur apprentissage. L'architecture du modèle (*Transformer*³) associée au volume très important de données utilisées pendant l'entraînement permet des usages variés pour ces modèles, sans que ces derniers n'aient été entraînés spécifiquement pour ces tâches.

Les premières solutions fondées sur de grands modèles de langage, tels que ChatGPT⁴, sont capables de créer un texte à partir d'instructions textuelles en entrée. Ces modèles deviennent aujourd'hui de plus en plus multimodaux, c'est-à-dire qu'ils peuvent prendre en compte, aussi bien en entrée qu'en sortie, des données de plusieurs types, même combinées telles que l'image, l'audio, la vidéo, de la 3D, etc.

2.2. Risque

Le risque étant une notion évoquée dans plusieurs disciplines et dont la définition peut varier, il convient de définir ce concept qui sera utilisé dans toute la suite de cette étude.

³ Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser. *Attention is all you need*. *Advances in neural information processing systems*. vol. 30. 2017. <https://arxiv.org/pdf/1706.03762.pdf>

⁴ OpenAI. Introducing ChatGPT. Open AI. November 30, 2022. <https://openai.com/blog/chatgpt>



Un risque peut être défini comme : « Tout événement ou situation pouvant entraîner des conséquences d'ordre humain, financier, juridique, réglementaire, ou relatif à la réputation, susceptibles d'impacter l'entreprise dans l'atteinte de ses objectifs ou dans son développement, quelle que soit la nature des causes et l'origine du risque (interne ou externe) »⁵. La norme ISO 31000 sur le management du risque le définit également comme « l'effet de l'incertitude sur l'atteinte des objectifs ». ⁶

Les éléments fondamentaux du risque, qui permettent ensuite de définir des stratégies de gestion des risques, sont généralement les suivants :

- Sa ou ses causes ;
- Sa ou ses conséquences, ou impacts, qui peuvent être mesurés selon une échelle propre à chaque organisation, son contexte et le type d'impact (par exemple, un montant des pertes occasionnées pour mesurer un impact financier) ;
- Sa probabilité d'occurrence, ou vraisemblance.

La combinaison des deux derniers éléments permet généralement de définir l'acceptabilité d'un risque, déterminant ensuite la manière dont une organisation décide de le traiter, comme dans l'approche retenue par l'ANSSI pour sa méthodologie d'analyse de risque EBIOS RM⁷.

L'un des objectifs de cette étude est donc d'identifier quels sont les risques qu'amène l'usage de l'IA générative, et d'en explorer les impacts potentiels sur différents aspects des organisations et de la société. La présentation de mesures de remédiation ou de maîtrise des risques, en fin de document, permettra d'aborder comment réduire la probabilité d'occurrence d'un risque, et donc de voir ses impacts se matérialiser.

2.3. Parties prenantes

La gestion des risques qui émanent de l'IA générative peut être organisée selon trois lignes de défense. La première ligne de défense (LoD⁸) est représentée notamment par les *Model Owners*. Ce sont les personnes qui prennent la responsabilité de l'utilisation des systèmes d'IA et sont donc les premiers à devoir s'assurer que les risques afférents ont bien été pris en compte lors des diverses étapes du développement, déploiement et

⁵ Jean-Luc Wibo, cité par Laurence Baillif, Gestion des risques - De la sécurité à la gestion globale des risques, CNPP Editions, 2023, p. 23

⁶ Norme ISO 31000:2018, Management du risque, Organisation International de normalisation (ISO), 2018

⁷ Cf. Glossaire

⁸ Cf. Glossaire



utilisation de l'IA. Ces personnes s'assurent que le processus de développement, la documentation, et le suivi (*monitoring*) de l'IA sont conformes aux standards de l'entreprise.

La deuxième ligne de défense (LoD2) est constituée par les équipes de revue indépendante, les équipes en charge de la gouvernance et de la supervision du portefeuille des systèmes d'IA et si pertinent, des personnes participant aux comités d'approbation et de revue. Leur rôle est de collectivement s'assurer que la première ligne de défense joue bien son rôle de gestion des risques de modèle, mais aussi de mesurer et de signaler les risques de modèle agrégés au niveau d'un périmètre défini. La deuxième ligne de défense joue notamment un rôle important de revue du modèle avant la phase d'industrialisation. Pour un système d'IA générative, il est fréquent que la revue s'accompagne également d'une analyse de la solution IT qui va intégrer le modèle dans le but par exemple de s'assurer de la protection et de la sécurité des données ou de la robustesse du système par rapport à des attaques cyber. Ces revues ne sont pas nécessairement menées par les mêmes équipes.

La troisième ligne de défense (LoD3) est constituée des équipes d'audit interne, parfois dénommée inspection générale. Son rôle est d'évaluer la conformité des opérations, du niveau du risque effectivement encouru, du respect des procédures, de l'efficacité et du caractère approprié des dispositifs d'identification et de gestion des risques. Ces équipes peuvent donc être amenées à vérifier que les travaux réalisés aux deux premiers niveaux sont conformes aux règles en vigueur au sein de l'établissement, ce qui implique de procéder à une réévaluation des modèles développés, voire dans certains cas, de challenger les contrôles réalisés au moyen de modèles alternatifs. Cette organisation en trois lignes de défense implique la participation d'une multitude de parties prenantes dans la gestion des risques liés à l'IA générative, de la détection à la remédiation.

La liste non exhaustive qui suit présente certains des départements concernés.

- Les datalabs constituent la première ligne de défense et représentent le premier acteur en charge de l'identification, l'évaluation et la remédiation des risques liés à l'IA générative ;
- Les utilisateurs de l'IA générative participent à la gestion des risques en identifiant des comportements à risque dans la génération de contenu par l'IA ou en donnant leur feedback, qui peut être exploité par les différentes parties prenantes ;
- Le département des risques, qui fait partie en général de la LoD2 selon l'organisation explicitée ci-dessus, gère le risque de modèle lié à l'IA générative et assure le suivi des bonnes pratiques en matière de modélisation et de transparence ;



- Le département juridique a la charge de la gestion du risque juridique lié à l'IA générative. Il analyse notamment les textes réglementaires qui encadrent le développement et l'usage de celle-ci ;
- Le département en charge du risque de conformité gère le risque de non-conformité qui peut résulter de l'utilisation de l'IA générative et assure la protection de la clientèle et des employés ;
- Les départements IT sont impliqués dans la gestion du risque cyber lié à l'IA générative et suivent les standards de place, émis par les organismes experts de la place tels que le NIST aux Etats Unis ou l'ANSSI en France. Ils gèrent également les risques liés aux données utilisées et produites par les systèmes d'IA.

Très formalisée dans les secteurs régulés comme la banque par exemple, cette organisation peut être plus informelle dans des secteurs moins régulés, les rôles décrits pouvant alors être remplis de façons différentes selon les organisations.

2.4. Causes

L'identification des causes sous-jacentes à l'émergence de risques liés à l'utilisation de l'IA générative est cruciale pour la mise en place de stratégies efficaces de gestion de ces risques et de remédiation. Dans le cadre de notre étude, nous avons identifié trois grands groupes de causes : données, modèle et humain.

2.4.1. Les Données

Les données utilisées dans le cadre de l'entraînement des LLM peuvent présenter plusieurs limites qui affectent la qualité, la fiabilité ou la transparence des modèles en question. Le manque de représentativité des données d'entraînement, le non-respect de la propriété intellectuelle, l'utilisation de données sensibles ou les problèmes de qualité de donnée sont des sources de risque importantes dans la conception et l'utilisation d'une IA générative.

Un manque de représentativité peut se manifester, par exemple, lorsque les données d'entraînement excluent certaines populations. Les modèles entraînés sur ces données pourraient alors être biaisés et risqueraient de générer des contenus inappropriés ou inadaptés.

La violation de la propriété intellectuelle est un autre exemple de risque lié aux données. L'utilisation de données protégées comme des livres, articles de presse ou des œuvres



d'art pour l'entraînement d'un système d'IA générative peut exposer le développeur à des poursuites judiciaires de la part des auteurs détenteurs de la propriété intellectuelle⁹.

L'utilisation de données sensibles directement ou indirectement (encodées par des éléments dans le texte comme les intérêts et loisirs dans les CV) sans justification liée à la finalité du cas d'usage est une autre source de risque qui doit être prise en compte dans l'analyse.

Enfin, une mauvaise qualité des données, illustrée par des erreurs ou des incohérences dans les bases de données, peut sérieusement compromettre la robustesse et la précision d'un système d'IA.

2.4.2. Le Modèle

Le système d'IA choisi, sa conception et son utilisation peuvent également être la cause de divers risques.

Par exemple, l'utilisation de modèles génériques, sans optimisation spécifique pour un contexte donné, peut réduire leur efficacité. Des problèmes tels que les hallucinations, où le modèle génère des informations inventées, ou la production de contenu inapproprié ou offensant, sont des exemples critiques des risques liés à un modèle non optimisé.

La robustesse du modèle constitue également un enjeu majeur, car une sécurité insuffisante peut permettre des attaques malveillantes, via, par exemple, des « injections de prompts » où les cybercriminels manipulent le modèle pour avoir accès aux données sources et notamment aux données sensibles (données stratégiques ou données à caractère personnel).

En outre, l'absence d'explicabilité, la complexité des systèmes d'IA, où les processus de fonctionnement ne sont pas transparents, peut poser des défis significatifs pour leur adoption et leur acceptation. Cette absence peut aussi engager un risque éthique, voire un risque de non-conformité (la transparence faisant partie des principes fondateurs du Règlement sur l'Intelligence Artificielle¹⁰).

⁹ Michael M. Grynbaum, Ryan Mac. New York Times sues Open AI. New York Times. December 27, 2023. <https://www.nytimes.com/2023/12/27/business/media/new-york-times-open-ai-microsoft-lawsuit.html>

¹⁰ Législation sur l'intelligence artificielle. P9_TA(2024)0138. Parlement européen. 13 mars 2024. https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138_FR.pdf. Sera publié au Journal Officiel de l'Union Européenne courant juillet 2024.



2.4.3. L'Humain

Les aspects humains influencent fortement la manière dont les technologies d'IA sont utilisées et perçues, et peuvent être à l'origine de divers impacts négatifs.

Un exemple de risque lié à l'humain est la focalisation excessive sur les performances du modèle qui peut conduire à négliger d'autres aspects cruciaux comme la sécurité, l'éthique et l'impact social. Par exemple, une entreprise pourrait prioriser la rapidité de réponse d'un assistant virtuel au détriment de la précision et de la sécurité des informations fournies.

Le biais de confirmation constitue une autre source de risque, lorsque les utilisateurs font excessivement confiance aux sorties de l'IA. Dans ce cas, le manque de sensibilisation et de formation aux risques associés à l'IA générative peut conduire à une adoption de ces technologies sans prise de recul ou esprit critique.

Le potentiel impact environnemental est un élément de plus en plus important dans l'analyse des risques, les systèmes d'IA, en particulier les modèles de grande taille, consommant en effet des ressources énergétiques significatives.

Enfin, l'approche techno-solutionniste, où la technologie est perçue comme la solution à tous les problèmes, peut conduire à une dépendance excessive à l'IA, au détriment d'approches plus équilibrées et inclusives.

2.5. Impacts

2.5.1. Impact juridique

L'impact juridique peut être défini comme : « le risque de tout litige avec une contrepartie, résultant de toute imprécision, lacune ou insuffisance susceptible d'être imputable à l'entreprise au titre de ses opérations »¹¹. Dans le cas de l'usage de l'IA générative, il peut par exemple s'agir d'un litige avec une personne physique dont les données personnelles auraient été utilisées sans son consentement, ou dans le cas d'une violation de propriété intellectuelle.

¹¹ Article 4 du règlement n° 97-02 du Comité de la réglementation bancaire et financière (CRBF) du 21 février 1997 relatif au contrôle interne des établissements de crédit et des entreprises d'investissement



2.5.2. Impact financier

L'impact financier d'un risque se réfère aux conséquences économiques directes et indirectes découlant de sa matérialisation. Ces conséquences peuvent se manifester par des pertes financières directes telles que des coûts de réparation ou de remplacement, des amendes, des pertes de revenus, ainsi que des pertes indirectes telles que la baisse de la valeur des actifs, la perte de clients ou la dégradation de la réputation de l'entreprise, par exemple si un *chatbot* public publie des propos controversés.

2.5.3. Impact opérationnel

L'impact opérationnel d'un risque désigne les impacts potentiels d'un risque sur les processus et opérations internes d'une organisation. Il permet de mesurer les conséquences de la concrétisation d'un risque sur le bon fonctionnement interne de l'organisation, en évaluant quels processus internes de l'entreprise peuvent être affectés par la survenue de l'événement redouté, et quels ont été les conséquences (exemples : délai d'accomplissement du processus ou impossibilité d'accomplir le processus).

2.5.4. Impact réputationnel

L'impact réputationnel peut être défini comme l'impact d' « un risque résultant d'une perception négative de l'entreprise de la part des clients, contreparties, actionnaires, investisseurs, créanciers, analystes de marché, d'autres parties prenantes ou régulateurs concernés »¹². Sans parler des conséquences financières ou légales que les faits ayant généré ces perceptions peuvent avoir, l'impact réputationnel affecte aussi la crédibilité de l'organisation, ce qui peut également affecter sa capacité à remplir ses fonctions.

En matière d'IA générative, technologie dont l'usage peut être controversé du fait des risques de biais, d'erreur ou d'hallucination par exemple, les impacts réputationnels pourraient non seulement venir de la mauvaise utilisation, ou du détournement, d'un système d'IA générative, comme on l'a vu avec le *chatbot* Tay de Microsoft¹³, mais aussi du choix de recourir à l'IA générative sur certaines tâches. Ce choix pourrait en effet

¹² Basel Committee on Banking Supervision. Enhancements to the Basel II framework. Paragraph 47. July 2009. <https://www.bis.org/publ/bcbs157.pdf>

¹³ Peter Lee. Learning from Tay's introduction. Microsoft blog. March 25, 2016. <https://blogs.microsoft.com/blog/2016/03/25/learning-tays-introduction/>



susciter la controverse du fait des débats éthiques ou judiciaires que suscite encore cette technologie émergente, notamment lorsqu'elle est amenée à traiter des données sensibles, comme des données à caractère personnel, ou à fournir une aide à la décision qui pourrait être jugée déshumanisante dans le cadre de certains processus.

2.5.5. Impact organisationnel

La notion d'impact organisationnel est particulièrement utilisée dans la conduite du changement, notamment dans le domaine de la santé. On parle alors de l'effet, conséquence ou résultat d'un changement sur les caractéristiques et le fonctionnement d'une organisation ou d'un ensemble d'organisations, et particulièrement de l'utilisation d'une nouvelle technologie¹⁴. Cette définition peut être transposée à notre sujet, en considérant que le changement vient de la matérialisation d'un risque lié à l'adoption de l'IA générative, et a des conséquences sur l'organisation l'ayant adoptée, ou sur d'autres organisations. Des changements dans les effectifs, dans le rôle de certains individus de l'organisation peuvent ainsi être envisagés dans le cas de l'adoption de technologies d'IA générative qui pourraient assister ou remplacer certaines équipes dans l'accomplissement des fonctions de l'organisation. Ce type d'impact peut alors être lié également à l'impact social, ou l'impact financier déjà évoqué.

2.5.6. Impact social

Si l'on s'appuie sur la définition du Conseil supérieur de l'économie sociale et solidaire, « l'impact social consiste en l'ensemble des conséquences (évolutions, inflexions, changements, ruptures) des activités d'une organisation tant sur ses parties prenantes, externes (bénéficiaires, usagers, clients) directes ou indirectes de son territoire et internes, (salariés, bénévoles, volontaires), que sur la société en général »¹⁵.

L'adoption de l'IA générative peut avoir de multiples conséquences sociales, par exemple sur le marché du travail¹⁶. Nous entendons ici des impacts qui peuvent à la fois

¹⁴ Haute autorité de santé. Cartographie des impacts organisationnels pour l'évaluation des technologies de santé, Guide méthodologique. 10 décembre 2020. https://www.has-sante.fr/upload/docs/application/pdf/2020-12/guide_methodologique_impacts_organisationnels.pdf

¹⁵ Thierry Sibieude, Céline Claverie. La mesure de l'impact social : après le temps des discours, voici venu le temps de l'action. Groupe de travail du CSESS sur la mesure de l'impact social. 8 décembre 2011. https://www.avise.org/sites/default/files/atoms/files/20140204/201112_CSESS_Rapport_ImpactSocial.pdf

¹⁶ Pawel Gmyrek, Janine Berg, David Bescond. Generative AI and jobs: A global analysis of potential effects on job quantity and quality. December 12, 2023. International Labour Organization. <https://www.ilo.org/resource/generative-ai-and-jobs-global-analysis-potential-effects-job-quantity-and>



concerner les individus amenés à traiter directement avec un système d'IA générative, ainsi que ceux dont les données sont exploitées par un système d'IA générative et pourraient par exemple être diffusées par erreur, sans oublier les individus qui pourraient être affectés directement ou indirectement par une décision prise par une IA générative.

2.5.7. Impact environnemental

La notion d'impact environnemental des actions d'une organisation peut recouvrir un large spectre de notions. La responsabilité sociale des entreprises, telle que définie par le Ministère de l'Economie français¹⁷ intègre une dimension environnementale, en s'appuyant sur l'approche de la norme ISO 26000 qui compte 7 thématiques centrales pour la responsabilité sociale des entreprises, dont celle de l'environnement.

Cette approche s'appuie sur la définition que l'on peut donner au risque environnemental comme le risque causé à l'environnement par une menace qui trouve sa source dans l'activité de l'homme.¹⁸

Dans son livre blanc de mai 2023, *L'IA éthique en pratique*¹⁹, le HUB France IA avait déjà abordé les risques environnementaux à l'usage des systèmes d'IA générative, en considérant différents types de coûts, de l'entraînement du modèle à son utilisation en passant par la consommation de métaux rares et de l'eau nécessaire pour le refroidissement des serveurs.

Cependant, aucun indicateur n'a été créé depuis, même si, avec son Règlement sur l'IA, l'union européenne demande une plus grande transparence des industriels sur ces impacts environnementaux.

¹⁷ Bercy Infos. Qu'est-ce que la responsabilité sociale des entreprises. Ministère de l'Economie, des Finances et de la souveraineté industrielle et numérique. 2 mai 2024.

<https://www.economie.gouv.fr/entreprises/responsabilite-societale-entreprises-rse#>

¹⁸ Delphine Misonne. Le risque environnemental. In : Les ambivalences du risque. Presses universitaires Saint-Louis Bruxelles. p. 381-403. 2008. <http://books.openedition.org/pusi/3549>.

¹⁹ Hub France IA. L'IA éthique en pratique. Livre blanc. Mai 2023. https://www.hub-franceia.fr/wp-content/uploads/2023/05/Livre_Blanc_IA_Ethique.pdf



2.6. Criticité

Dans la plupart des approches de gestion de risques, comme celle exposée dans la norme ISO 27005 :2022, l'évaluation du risque passe d'abord par une estimation du niveau de risque, qu'on peut désigner aussi sous le terme de « criticité » du risque. Elle s'évalue en attribuant des valeurs à la vraisemblance du risque et à la mesure des impacts de ce dernier²⁰. D'autres critères peuvent être pris en compte tels que la cinétique du risque (vitesse de réalisation), la volatilité, l'horizon temporel, la corrélation etc. Mais ces critères ne font pas l'objet de la présente étude.

La vraisemblance et les impacts (ou conséquences) sont mesurés à l'aide d'échelles définies par l'organisation évaluant le risque, en fonction de son contexte et de son évaluation de l'acceptabilité des impacts. Comme présenté en partie 2.5 (Impacts), les impacts du risque peuvent être de natures diverses (juridiques, financiers, environnementaux, ...), ce qui multiplie encore les types d'échelle qu'il est possible d'utiliser.

Le futur règlement de la Communauté Européenne applicable à l'intelligence artificielle (Règlement sur l'Intelligence Artificielle¹⁰) retient par exemple une approche par les risques pour réglementer les systèmes d'intelligence artificielle : soit en les interdisant purement et simplement (article 5), soit en les encadrant (systèmes d'intelligence artificielle "à haut risque"). On peut retenir l'échelle suivante en termes d'impact du risque, aussi appelée gravité du risque :

²⁰ Organisme International de Normalisation, ISO 27005 :2022. Sécurité de l'information, cybersécurité et protection de la vie privée — Préconisations pour la gestion des risques liés à la sécurité de l'information. 2022. <https://www.iso.org/fr/standard/80585.html>



Niveau d'impact	Echelle	Description
1	Faible	Le système entraîne peu ou pas de conséquences. Le cas échéant, les conséquences ne remettent pas en cause le processus ni la pérennité de l'entité mais peuvent modifier la perception de la qualité des services. Le risque est de fait négligeable.
2	Moyen	Le système entraîne des conséquences minimales. Ces conséquences ne remettent pas en cause un processus ou la pérennité de l'entité mais peuvent modifier la qualité des services fournis. Le risque est dans ce cas acceptable, mais peut nécessiter un suivi régulier.
3	Elevé	Le système entraîne des conséquences significatives. Ces dernières remettent en cause une partie des processus ou la pérennité de l'entité. Le risque peut être accepté, sous condition obligatoire de mise en place de mesures de mitigation, de suivi et de contrôle.
4	Très élevée	Le système entraîne des conséquences inacceptables, qui remettent en cause les processus ou la pérennité de l'entité.

De la même manière, une échelle peut être définie pour mesurer la vraisemblance du risque, par exemple en définissant des fréquences d'exposition au risque :

Fréquence	Echelle	Description
1	Faible	Exposition pouvant survenir au maximum une fois par an ou peu vraisemblable ou jamais rencontrée.
2	Moyen	Exposition pouvant survenir au maximum quelques fois par an.
3	Elevé	Exposition pouvant survenir au maximum une fois par mois.
4	Très élevée	Exposition pouvant survenir plusieurs fois par mois.

Le présent document se concentrera principalement sur l'étude de la vraisemblance des impacts, la fréquence demeurant une notion très dépendante du contexte de l'entité exposée et de sa tolérance aux risques. Dans une approche complète d'évaluation des risques, la combinaison de la mesure de la vraisemblance et de l'impact d'un risque permet de la comparer quantitativement aux autres risques auxquels l'entité est exposée, pour aider à la priorisation.

3.

Démarche d'analyse des risques

Contributeurs :

- *Imen Fourati, Expert lead, Risque de modèle, Société Générale*
- *Benjamin Bosch, Manager – Model risk Management – Data Science, Société Générale*
- *Thomas Gouritin, Consultant – Tomg Conseil*



3. Démarche d'analyse des risques

3.1. Présentation générale

Bien que certains risques revêtent un caractère transversal, l'évaluation des risques d'une IA générative doit se faire dans le cadre d'un cas d'usage bien défini. Ainsi, la première étape dans l'analyse des risques liés à l'IA générative est la définition des capacités générales du cas d'usage. Selon qu'il permet d'effectuer des tâches comme la conversation, la contraction de texte, ou la recherche augmentée (RAG : *retrieval augmented generation*), les risques peuvent être différents. Restreindre les capacités générales d'une IA générative réduirait ainsi les risques qui en découlent. L'analyse des risques doit également se faire, en prenant en compte les catégories d'utilisateurs de l'IA générative en question. Ainsi, les risques peuvent être amplifiés, selon que les utilisateurs soient internes ou externes à l'organisation qui déploie l'IA générative, formés ou non aux risques liés à cette IA.

La phase d'identification des risques est par la suite primordiale pour mettre en place les actions de remédiation adéquates. Chaque cas d'usage a ses risques spécifiques. Certains risques peuvent être transverses à plusieurs systèmes d'IA en général, tels que le manque de représentativité dans les données d'entraînement, le risque d'atteinte à l'équité ou le manque d'explicabilité. L'IA générative peut les amplifier, notamment si elle est déployée à large échelle. D'autres risques sont nouveaux, liés au caractère génératif des IA génératives tel que le risque d'hallucination ou le risque de création de contenu toxique ou nocif.

Une fois identifiés, les risques sont par la suite évalués, en suivant une approche holistique. Ainsi, des scénarios peuvent être construits autour de chaque cas d'usage pour évaluer les forces et les faiblesses de l'IA générative en question. Ces scénarios de test peuvent s'appuyer sur des prompts émanant de bases de données benchmarks ou générés par d'autres IA génératives. Ils permettent d'évaluer, à la fois, la performance de l'IA, sa robustesse et les risques qu'elle peut présenter. Par exemple, l'utilisation des techniques de **répliques manipulatrices** spécialement conçues pour contourner les garde-fous de l'IA générative (*guardrails*), connue sous le nom de *jailbreaking*, constitue une méthode efficace pour évaluer les risques de toxicité ou de fuite de données sensibles. Lors de la phase de déploiement, des *guardrails* sur les données d'entrée et/ou de sortie peuvent être mis en place pour mitiger les risques détectés, en fonction de leur



importance. Ainsi, les prompts en entrée peuvent être validés, en suivant une liste de principes à respecter par l'utilisateur.

Le contenu généré peut également faire l'objet d'un certain nombre de contrôles, visant à limiter les risques générés par le système d'IA. Dans ce qui suit, nous présentons un outil d'aide à l'identification et à l'évaluation des risques liés à une IA générative, sous forme d'une matrice de risques. Cette démarche est d'ordre qualitatif et doit être complétée par des mesures quantitatives, basées sur des scénarios de tests propres à chaque cas d'usage.

3.2. Matrice des risques

La matrice suivante est proposée, sur la base des éléments étayés ci-dessus, afin d'analyser les risques liés à l'utilisation d'une IA générative dans une organisation. Pour chaque cas d'usage, cette matrice vise à mettre en avant :

- En ligne : les causes qui peuvent générer des risques. Ces causes appartiennent aux familles données, modèle ou humain ;
- En colonne : les impacts liés à ces causes avec l'évaluation de leur niveau, sur une échelle de 1 à 4 ;
- La dernière colonne vise à mettre en avant les pistes de remédiation pour chaque cause ou famille de causes.

Causes		Impacts							Remédiations
Famille	Description	Financier	Juridique	Réputationnel	Opérationnel	Organisationnel	Social	Environnemental	
Data	Utilisation de données sensibles ou non-respect de la propriété intellectuelle	1	2	3	4	1	2	3	
Modèle	Risque d'hallucination ou génération de contenu nocif								
Humain	Biais de confirmation ou d'automatisation								



Dans le chapitre qui suit, cette matrice sera remplie pour chaque cas d'usage illustré. L'évaluation des niveaux d'impact, dépendant du contexte de l'organisation, est fournie à titre indicatif.

La fréquence, quant à elle, n'est pas évaluée dans les exemples de cas d'usage, car encore plus dépendante du contexte de l'organisation.

Plus largement, le contenu de la matrice devra être interprété et appliqué en fonction du contexte spécifique de l'organisation implémentant le cas d'usage.

4.

Catégories d'usages



4. Catégories d'usages

4.1. Catégories d'usage transverse

4.1.1. Agent conversationnel

Un **agent conversationnel** est un service logiciel capable de tenir un dialogue avec un utilisateur par l'intermédiaire du langage naturel de ce dernier (oral ou écrit). L'objectif du dialogue est de répondre à une demande de l'utilisateur concernant un contexte opérationnel.

L'agent conversationnel est souvent spécialisé dans un sujet particulier (vente d'un produit, support client) et cherche à recentrer la conversation autour du contexte si l'utilisateur tente de s'en éloigner. Il est généralement capable de communiquer via un protocole social dans le but de le rendre le plus « humain » possible.

L'agent conversationnel fournit des réponses rapides, pertinentes et personnalisées à ses utilisateurs. Il peut être hautement disponible puisque ses traitements sont autonomes. Dans certains cas, il est en mesure de rediriger l'utilisateur vers un agent humain s'il détecte une situation qu'il ne sait pas résoudre.

L'agent conversationnel a été introduit dans les systèmes d'information bien avant la démocratisation de l'IA générative et des LLM. Cependant, l'évolution rapide des technologies d'intelligence artificielle a ouvert de nouvelles perspectives dans le domaine des agents conversationnels, offrant des solutions plus sophistiquées pour l'interaction entre, par exemple, les entreprises et leurs clients ou salariés.

4.1.2. Recherche augmentée

L'IA générative offre de nouvelles **expériences de recherche** aux utilisateurs. Les moteurs de recherche, bien connus du grand public, en sont un parfait exemple et ne cessent d'évoluer et de gagner en performance à l'aide de cette technologie.

La recherche augmentée est aussi la grande gagnante dans les usages de l'entreprise : historiquement, les données d'entreprise « faiblement typées » et hétérogènes (ex : patrimoine documentaire) sont :

- Souvent stockées sur des serveurs de fichiers, à la convenance des utilisateurs sans organisation de l'information ;



- Parfois intégrées de manière plus structurée dans des plateformes de GED²¹ capables d'indexer leur contenu et de proposer un moteur de recherche basé sur une recherche par mots clés.

Dans tous les cas, la démarche de recherche d'une information d'entreprise par l'utilisateur s'avère bien souvent fastidieuse et nécessite l'identification du document portant l'information recherchée.

L'IA générative s'appuie sur le corpus documentaire de l'entreprise pour générer sa réponse, et moyennant un contrôle d'accès à l'information, peut directement délivrer à l'utilisateur l'information qu'il cherche.

Cet usage repose généralement sur la génération augmentée de récupération (**RAG**) qui fournit un moyen d'optimiser le résultat d'un LLM avec des informations ciblées, sans modifier le modèle sous-jacent ; ainsi, des informations plus récentes que celles utilisées pour la première construction du LLM sont intégrées régulièrement. Cela signifie que le modèle d'IA générative peut fournir des réponses contextuellement appropriées aux utilisateurs et les baser sur des données extrêmement récentes et précises.

4.1.3. Transformateur de contenu

L'IA Générative est capable d'appliquer des transformations sur la donnée d'entrée : texte, image, vidéo, audio, ... Les opérations les plus communes portent sur les transformations de contenu textuel. Dans le domaine des LLM, nous citerons :

- Résumer un texte ;
- Traduire un texte ;
- Corriger les fautes présentes dans le texte ;
- Modifier le ton d'un texte, par exemple utiliser un ton courtois, ou précieux.

Les opérations sur les autres types de données (image, vidéo, audio, parole) sont également très nombreuses et de plus en plus puissantes, voire « surprenantes ». Sans nécessairement le percevoir, nous utilisons régulièrement ces modèles dans nos outils de travail ; l'exemple le plus commun étant la suppression du bruit dans les visio-conférences. Les transformations sur les images permettent par exemple d'effacer un visage, d'incruster un objet, etc.

²¹ Cf. Glossaire



4.1.4. Générateur de contenu

L'IA Générative permet grâce à du traitement naturel de langage (NLP) d'interpréter les données saisies par l'utilisateur dans son prompt pour générer un contenu riche et structuré. Les données de l'utilisateur sont les consignes que le modèle doit suivre pour générer le contenu retourné.

Parmi les usages possibles, nous retrouvons la **production de texte**, tels que des articles, des courriers, des courriels, des rapports, des dissertations, ... Le style d'écriture varie selon la requête initiale, pouvant prendre la forme d'une écriture formelle et structurée mais aussi de textes plus créatifs.

La génération de contenu concerne également la **production d'images**, encore une fois à partir des consignes (textuelles) fournies par l'utilisateur à travers le prompt. Le champ des possibles est très large : photographie hyperréaliste ou artistique, œuvre imaginaire, telle qu'une peinture ou un dessin. A travers ses consignes, l'utilisateur peut indiquer le style qu'il souhaite appliquer.

Dans la continuité, l'IA Générative peut également produire des **vidéos**.

Enfin, sur le même principe, il est possible de créer des sons, paroles ou musiques : récit oral, musiques originales ou imitation de voix connues.

4.1.5. Générateur de code

La génération de code est également une catégorie d'usage majeure proposée par l'IA générative.

Elle permet aux développeurs ou autres utilisateurs plus « profanes » de **développer du code informatique** plus rapidement, de façon plus précise et rigoureuse, et dans le respect des modèles de conception, en intégrant le code généré dans leur application. L'application peut être par exemple un script, une requête SQL, un classeur Excel (formule Excel ou code VBA), ou une application spécifique codée avec un langage interprété ou compilé.

A partir d'une requête en langage naturel fournie par l'utilisateur, l'IA générative peut générer des extraits de code, transformer du code en un autre langage de programmation (transpiler), modifier un code existant ou encore produire une documentation à partir d'un code fourni en entrée. De plus, elle permet de détecter et de corriger des anomalies (erreurs de syntaxe, d'exécution, de logique, ou encore de



formatage), d'optimiser un algorithme, d'expliquer la fonction d'un code ou encore de détecter des failles de sécurité.

4.1.6. Analyse de la donnée

L'IA générative peut être employée à des fins d'analyse sur la donnée. A l'instar d'un modèle de Machine Learning spécifique, entraîné pour répondre à un besoin précis, l'IA générative, peut, dans des degrés de performance et d'efficacité moindres, effectuer des opérations telles que la classification, l'extraction d'entités ou l'analyse sémantique.

Comme pour les catégories d'usages précédentes, ces opérations sont possibles sur tous types de données d'entrée, à savoir le texte, l'audio, les images et vidéos.

Citons quelques exemples (simplistes) :

- Compter le nombre de chiens sur une image ;
- Lister les capitales, le nombre d'habitants et le PIB des pays renseignés dans le prompt ;
- Regrouper des tweets par émotion (neutre, colère, joie, tristesse, dégoût, surprise).

4.2. Avant-goût : les catégories d'usages à venir ...

Durant l'été, le Hub France IA publiera 13 exemples de cas d'usages propres aux domaines suivants :

- Marketing ;
- Cybersécurité ;
- Ressources Humaines ;
- Finance ;
- Santé ;
- Commerce ;
- Développement logiciel ;
- Industries culturelles et créatives ;
- Juridique ;
- Service Client ;
- Conseil ;
- Data Science ;
- Logistique et transport.

Chaque semaine, un nouveau domaine d'activité sera mis en lumière. Restez connectés et suivez-nous pour ne rien manquer des publications estivales.

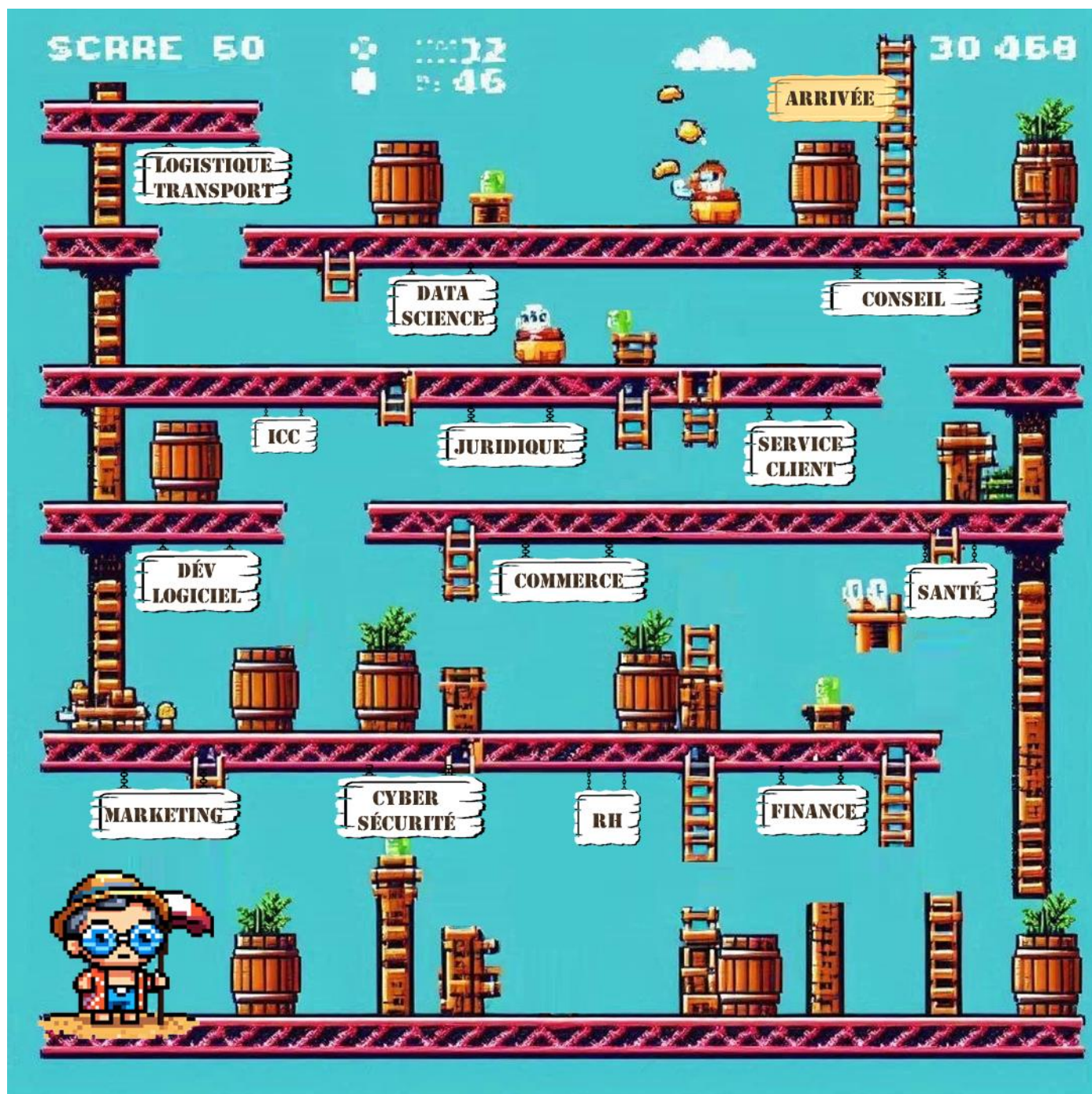


Image générée par Image Creator de Microsoft Designer puis modifiée par le Hub France IA

5.

Synthèse des remédiations

Contributeurs :

- **Thomas Argheria, Manager, Wavestone**
- **Gérôme Billois, Partner, Wavestone**
- **Martin D'Acremont, Consultant – Wavestone**



5. Synthèse des remédiations

Les cas d'usage d'IA générative se multiplient sur toutes les verticales métiers. Il ne fait aucun doute que cette technologie prend toute sa place dans nos organisations, et qu'elle s'implémentera durablement. Les risques associés aux IA génératives ne sauront être intégralement atténués. En particulier, pour les *Large Language Models* (ou LLM), leur fonctionnement intrinsèque et leur méthode d'entraînement ne permettent pas d'en faire des agents vraiment « intelligents ». Ce sont de simples « mémoires vives » qui restituent une connaissance apprise lors de la phase d'entraînement. Cette mémoire sait être certes savamment adaptée et redistribuée, mais ça n'en reste pas moins simplement une mémoire, et ces systèmes sont dépourvus d'intelligence humaine. Ils peuvent donc commettre des erreurs, sources de risques pour l'entreprise. Cependant, il existe des mesures « de **remédiation** » qui permettent de réduire considérablement les risques associés à cette nouvelle technologie. Elles sont présentées dans la section qui suit, mais il convient de préciser son contour :

1. Pour mettre en place leur projet d'IA générative, la grande majorité des organisations va **s'appuyer sur un modèle élaboré par un fournisseur** (Google, OpenAI, Meta, Microsoft, Anthropic, Mistral...). Ces modèles dits « de fondation » sont déjà (pré) entraînés. Ainsi, les mesures présentées ici interviennent donc exclusivement après la phase d'entraînement du modèle. Ce sont des mesures actionnables pour les organisations, qui ne sont pas dépendantes des décisions d'un fournisseur.
2. Aussi, les mesures de remédiation présentées ne cherchent à couvrir que les risques spécifiques à l'IA générative. Comme toute application, un système d'IA générative doit également **mettre en place des mesures de protection contre les risques** usuels, dont les risques **de cybersécurité** (sécurité des API, des plugins, chiffrement des données et des flux, journalisations d'évènements...). Ce qui suit considère que l'ensemble du socle cybersécurité « classique » est déjà mis en œuvre pour les composants du système d'IA (documentation, évaluation des risques, plan d'action de mitigation ...).
3. Pour **aider la compréhension et la lecture**, les mesures de remédiation sont présentées de la manière suivante :
 - Par grandes familles de risques : modèles, utilisateurs, données ;
 - Par ordre croissant de complexité : complexité technique et d'implémentation ;
 - Rattachées à un risque principal, même si certaines peuvent en couvrir plus largement.



5.1. Réduire les risques liés au modèle

5.1.1. Limiter la génération de contenu non désirable

L'un des risques principaux lors de l'utilisation d'un système d'IA générative est la génération d'un **contenu non désirable** (e.g. contraire à l'éthique ou aux valeurs de l'organisation, biaisé, illégal, inexact...). Au premier rang de ce problème le phénomène **d'hallucination**²², qui correspond à la production d'un contenu cohérent intellectuellement, mais factuellement incorrect. Il existe aussi des problématiques de génération de contenu illégal, notamment en lien avec la production de contenus protégés par licence.

Quelques exemples d'hallucinations

- Lors de sa sortie en août 2023, Google Bard (désormais Gemini), affirmait que le télescope James Webb (opérationnel depuis 2021), avait pris des photographies de la première exoplanète (alors que celles-ci ont été prises en 2004)²³ ;
- Gemini à nouveau, souhaitant augmenter la représentativité des personnes de couleurs dans les réponses produites, a finalement conduit à des aberrations historiques : des images de Vikings et de Nazis noirs.²⁴

Comme mentionné en introduction, ces risques sont liés intrinsèquement à la manière dont les modèles de génération de contenu sont entraînés. Les LLM notamment fonctionnent selon une méthode de génération qui construit une suite de mots constituant la réponse la *plus probable* selon (1) l'entraînement réalisé sur des quantités énormes de données, souvent open-source, et parfois protégées par le droit d'auteur, et (2) le contenu du prompt de l'utilisateur. Le fonctionnement est cependant non

²² Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Wenliang Dai, Andrea Madotto, Pascale Fung. Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys*, vol. 55, n° 12, pp. 1-38. November 2022. <https://arxiv.org/pdf/2202.03629>

²³ Le Monde. Google perd 7 % à la Bourse de New York après une erreur de Bard, son nouveau robot conversationnel. 9 février 2023. https://www.lemonde.fr/pixels/article/2023/02/09/google-perd-7-a-la-bourse-de-new-york-a-la-suite-d-une-erreur-de-son-tout-nouveau-robot-conversationnel_6161118_4408996.html

²⁴ Radio France. Intelligence artificielle : Google suspend la création d'images de personnes sur son IA Gemini après des critiques. 22 février 2024. https://www.francetvinfo.fr/internet/intelligence-artificielle/intelligence-artificielle-google-suspend-la-creation-d-images-de-personnes-sur-son-ia-gemini-apres-des-critiques_6382006.html



déterministe car une part d'aléatoire est ajoutée (par l'intermédiaire d'un paramètre de « **température** »²⁵).

Parce que les modèles apprennent sur ces quantités énormes de données, pour lesquelles la fiabilité et l'absence de biais ne sont pas toujours contrôlables (d'autant plus que les modèles de fondation ne donnent pas d'information précise sur les données d'entraînement qu'ils ont utilisées), les hallucinations, mais aussi l'apprentissage sur des données qui sont normalement protégés par droit d'auteur sont inévitables. Ce risque, même avec l'application des mesures de réduction présentées ci-dessous, ne sera jamais complètement supprimé.

Mesure n°1 : Spécialiser le modèle pour ses cas d'usage

Selon le cas d'usage à implémenter, choisir le modèle le plus adapté en fonction de ses capacités, et si possible, le réentraîner sur des données fiables, actualisées et spécialisées pour le domaine du cas d'usage, pour lui permettre de gagner en précision, et donc de réduire le risque d'hallucination.

De la même manière que lorsqu'on construit une application classique il est nécessaire de s'assurer de l'utilisation des bibliothèques appropriées, il convient pour les applications utilisant de l'IA générative de s'assurer de la pertinence du modèle utilisé. En effet, certains modèles sont plus adaptés que d'autres pour certaines tâches comme le montre par exemple le comparatif du *Center for Research on Foundation Model* de l'université de Stanford²⁶. C'est un premier filtre à prendre en compte selon le cas d'usage que l'on étudie.

Ensuite, les modèles génériques ne sont pas adaptés pour tous les cas d'usage et parfois, ils doivent être spécialisés. En l'occurrence, certains modèles de LLM rendus disponibles en open-source (Llama, Mistral) peuvent être réentraînés partiellement pour mieux correspondre à la mise en place d'un cas d'usage en particulier. Ainsi, le projet Bhashini en Inde a pour objectif de créer des jeux de données dans les 22 langues officielles, afin qu'ils puissent être utilisés pour réentraîner des modèles de fondation. Ce réentraînement est indispensable, car les données sur lesquelles sont entraînés les

²⁵ Cf Glossaire.

²⁶ Percy Liang et al. Holistic Evaluation of Language Models. *Center for Research on Foundation Models, Stanford University*. November 2021. <https://arxiv.org/pdf/2211.09110>



modèles de fondation ne contiennent pas assez de contenus dans ces langues pour permettre aux modèles d'être nativement assez performants pour les traductions.²⁷

Mesure n°2 : Mettre en place la « Retrieval Augmented Generation ²⁸ », ou RAG.

Constituer une source de données de confiance sur lesquelles le modèle peut s'appuyer, exclusivement ou partiellement, pour générer ses réponses afin d'accroître leur pertinence.

Les hallucinations produites par les modèles sont en partie liées à un problème de qualité des données d'entraînement, et notamment de fraîcheur et d'actualisation. Pour pallier ce problème, il est possible de paramétrer le modèle pour qu'il utilise uniquement ou prioritairement ses capacités sur des sources de données de confiance, et ainsi réduise la priorité de la mémoire issue de son entraînement. S'appuyer sur ces sources de données de confiance, c'est faire du RAG, ou « *Retrieval Augmented Generation* ». Cette technique permet non seulement de réduire les hallucinations mais aussi de faire levier sur les données internes à l'entreprise.

Cependant, pour que le RAG soit efficace, il faut évidemment maintenir la qualité des documents qui le composent.

Mesure n°3 : Durcir les paramètres de génération de contenu du modèle

Durcir les paramètres du modèle (température, définition du master prompt, verbosité des réponses...) pour orienter le comportement de l'agent génératif, et maîtriser les réponses émises.

Lors du déploiement des systèmes d'IA générative, il est possible de durcir les paramètres de base du modèle pour orienter son « comportement » face aux requêtes des utilisateurs. Parmi ces paramètres, il convient de mentionner la **température**, la définition du **master prompt**²⁹, et la **verbosité** des réponses. Les réglages sur ces mesures sont souvent facilités par les interfaces (plus ergonomiques) proposées sur les

²⁷ Milin Stanly. India turns to AI to capture its 121 languages for digital services. *Indi/ai*. December, 20 2023 . <https://indiaai.gov.in/article/india-turns-to-ai-to-capture-its-121-languages-for-digital-services>

²⁸ En français, génération augmentée de récupération.

²⁹ Cf. Glossaire.



plateformes d'IA générative des fournisseurs de Cloud. Pour les modèles récupérés en *open-source*, ce durcissement doit faire l'objet de développements supplémentaires.

La température permet de régler le niveau d'improvisation ou de créativité du modèle dans ses réponses. Au minimum, cela contraint le modèle à se reposer exclusivement sur les données qui sont fournies pour produire ses réponses. C'est ce qu'il est possible d'utiliser pour un *chatbot* qui agit comme conseiller juridique. En revanche pour un *chatbot* marketing, peut-être que plus de créativité est souhaitable. À nouveau, tout dépend du cas d'usage, mais plus la température est basse, plus le risque d'hallucination et la créativité sont faibles.

Le *master prompt* désigne, quant à lui, les instructions données au modèle pour l'aider à répondre aux questions des utilisateurs. Il s'agit d'un *prompt* générique, fournit à l'IA avant même qu'il traite une question, pour l'initialiser et la cadrer. Ces instructions peuvent contenir le rôle ou le format de réponse à respecter, ou la demande de ne pas répondre si le modèle n'a pas l'information demandée (et ainsi de ne rien « inventer »). Un *master prompt* correctement rédigé permet de préciser le comportement souhaité, et d'insérer des exigences de sécurité (e.g. « ne livre aucune information sur l'entreprise », « ne fournis aucune information à caractère personnel », ...).

Enfin, pour éviter aux modèles d'être trop « verbeux », c'est-à-dire de fournir des réponses trop amples avec des informations superficielles, il est possible aussi de régler la verbosité des réponses. Les systèmes d'IA générative décomposant les entrées des utilisateurs en jetons, ou « *tokens* », pour les traiter puis renvoyer une réponse constituée de nouveaux jetons, il est possible, en limitant les réponses à un certain nombre de jetons, de limiter le risque de diffusion de résultats non désirables.

Mesure n°4 : Filtrer les réponses non désirables

Mettre en place un filtrage des réponses produites par le modèle, pour écarter les non désirables. Cette capacité peut être outillée avec un pare-feu IA.

Même avec la mise en place des trois premières mesures, et parce que le risque d'hallucination est impossible à réduire complètement, il peut être nécessaire de mettre en place un filtrage sur le contenu généré. Cette couche permet de bloquer l'émission d'un contenu non désirable si celui-ci est malgré tout produit. Des solutions émergent sur le marché, et notamment des « *AI Firewall* », ou pare-feux pour IA. Ces outils sont eux-mêmes des LLM, localisés entre l'utilisateur et le modèle, et entraînés pour être des



filtres de sécurité. Ils vont filtrer des prompts malicieux ou des réponses indésirables produites par le modèle. Ainsi, ils agissent à la fois sur les entrées et les sorties.

Ce filtrage nécessite cependant d'avoir un cadre de modération clair et bien rédigé (quelles valeurs doivent être respectées, quel contenu doit être considéré comme indésirable) sur lequel le pare-feu peut s'appuyer pour filtrer. Dans le cadre d'un *chatbot* pour des conseils financiers par exemple, il sera pertinent de filtrer des conseils médicaux, ou alors des réponses qui ne respectent pas les obligations légales en matière de transparence.

Les limites de cette mesure résident d'abord dans le fait qu'on ne peut jamais totalement dresser la liste exhaustive des cas de contenus non désirables. Ensuite, le cadre de modération n'est jamais complètement étanche et peut être contourné par des effets de forme, du fait de la nature de ces modèles : c'est **le prompt injection**³⁰.

Mesure n°5 : S'assurer de la pertinence des données de réentraînement

S'assurer que les données utilisées pour le réentraînement du modèle soient fiables et de bonne qualité, en vérifiant l'absence de biais, de tentative d'empoisonnement, et la pertinence des données.

De la même manière que pour la phase d'entraînement sur un jeu de données initial, tout réentraînement du modèle doit être fait à partir de jeux de données fiables. Ce réentraînement peut avoir lieu au moment de la spécialisation du modèle (mentionnée plus haut) ou d'un réentraînement sur tout ou partie des données de production (e.g. les conversations avec les utilisateurs). Pour ces cas particuliers où les modèles sont réentraînés, il est nécessaire que la donnée utilisée fasse l'objet de vérification, pour s'assurer de leur pertinence, de l'absence de biais, et de tentatives d'empoisonnement.³¹ On peut imaginer par exemple qu'un *chatbot* de conseils financiers soit réentraîné sur la base d'échanges uniquement avec une population présentant une aversion au risque. Ainsi, in fine, il n'émettra que des conseils adaptés à cette population mais non adaptés à ceux ayant une forte appétence au risque.

³⁰ OWASP. LLM01: Prompt Injection. LLM Top 10 risks. <https://genai.owasp.org/llmrisk/llm01-prompt-injection/>

³¹ C'est l'exemple fameux de Tay, ce *chatbot* de Microsoft qui en quelques heures s'est transformé en un activiste d'extrême droite. Voir

Morgane Tual. A peine lancée, une intelligence artificielle de Microsoft dérape sur Twitter. Le Monde. 24 mars 2016. https://www.lemonde.fr/pixels/article/2016/03/24/a-peine-lancee-une-intelligence-artificielle-de-microsoft-derape-sur-twitter_4889661_4408996.html



Plusieurs techniques sont possibles pour appliquer ces vérifications : un contrôle par échantillonnage, par recherche de mots clés (liste d'insultes...) ou même un retraitement des données par un LLM de vérification.

Mesure n°6 : Mettre en place du *Reinforcement Learning from Human Feedback*, ou RLHF

Améliorer les performances d'un système d'IA en appliquant les principes de l'apprentissage supervisé : un agent humain vient vérifier les prévisions faites par le modèle, corrige les écarts trop importants, et catalyse sa progression.

Le RLHF ou « *Reinforcement Learning from Human Feedback* » (Apprentissage par renforcement à partir de rétroaction humaine) est l'une des techniques les plus efficaces, mais aussi les plus chères, pour améliorer les modèles. Elle consiste à la mobilisation d'analystes humains, qui vont noter plusieurs réponses émises par le modèle, afin d'améliorer sa précision. Ce sont les experts du domaine métier sur lequel porte le cas d'usage qui valident la pertinence des résultats proposés (c'est une des techniques utilisées par OpenAI pour ChatGPT).

Une autre forme de RLHF consiste à utiliser directement les commentaires et les évaluations des utilisateurs vis-à-vis des réponses proposées pour guider l'apprentissage. Ces retours utilisateurs permettent d'identifier les meilleures réponses et les plus mauvaises. Ils sont utilisés pour entraîner un autre système d'IA, qui jouera le rôle de modèle de « récompense » et sera en mesure d'évaluer la qualité des réponses du premier modèle.

Mesure n°7 : Mettre en place un dispositif de garantie humaine

Mettre en place un collège de supervision pour évaluer les performances d'un modèle en mobilisant des experts du domaine métier concerné. Ces derniers produisent des réponses à des requêtes utilisateurs, les comparent à celles de l'IA, et identifient les écarts et des pistes d'amélioration.

Dans certains cas, l'impact d'hallucination de la part du système peut entraîner des conséquences importantes, par exemple pour un système d'IA générative de conseils médicaux. Une réflexion a été lancée par l'AFNOR pour imaginer un dispositif permettant de contrôler étroitement la conception et le fonctionnement d'un système d'IA en santé. Cette réflexion a pour origine l'article 17 de la loi de bioéthique de 2021 qui encadrait le



recours à l'IA en matière de santé publique, et l'AI Act européen qui prévoyait des exigences spécifiques de contrôle humain pour les systèmes d'IA à haut risque.^{32 10}

Ce type de dispositif est appelé **dispositif de garantie humaine**. L'objectif est de permettre au concepteur³³ de détecter et corriger les défaillances du système d'IA lorsqu'il répond aux utilisateurs, c'est à dire les écarts entre les recommandations faites par le système d'IA et celles faites par des experts humains.

Cette mesure permet d'accompagner le concepteur tout au long de la mise en place d'un cas d'usage : de la conception à la supervision en production. C'est en cela qu'elle va plus loin que le RHLF, qui intervient uniquement après la mise en production du modèle. Un **collège de supervision** est mis en place pour suivre les risques et mesures identifiés en amont par l'analyse de risque projet :

- **Des experts réviseurs** : experts du domaine métier, ils vont comparer les résultats du modèle avec des recommandations qu'ils auraient eux-mêmes produites. Dans le cas d'un *chatbot* de conseils en matière de santé, ce seront des médecins ;
- **Des représentants utilisateurs** : ici ce sont les professionnels du domaine métier en question, qui vont mettre à disposition le produit à des usagers (par exemple pour notre cas d'usage santé : les pharmaciens, d'autres médecins, des cliniques). Ils apportent une vision sur les cas d'usage pour lesquels ils mettent le système à disposition des bénéficiaires ;
- **Un tiers expert** du domaine métier à l'image des experts réviseurs mais avec une position d'indépendance vis-à-vis des autres protagonistes ;
- Le **concepteur** du système a la charge de l'implémentation des éventuelles mesures correctives identifiées ;
- Des **représentants des bénéficiaires ou utilisateurs finaux** : associations d'usagers, représentants associatifs, etc... Ils permettent de remonter des retours sur leur expérience du parcours usager et des points de difficulté dans l'utilisation du système d'IA.

Ce collège sera chargé d'analyser un échantillon aléatoire de dossiers traités par le système d'IA. Les mesures correctives peuvent prendre des formes variées et aller au-delà d'une correction purement technique au niveau de l'algorithme : formation,

³² AFNOR. Garantie humaine des systèmes fondés sur l'intelligence artificielle en santé. AFNOR Spec 2213. Mai 2024. <https://www.boutique.afnor.org/fr-fr/norme/afnor-spec-2213/garantie-humaine-des-systemes-fondes-sur-lintelligence-artificielle-en-sant/fa205274/419909>

³³ Ici, le concepteur peut être toute personne physique ou morale (ex. autorité publique, entreprise, chef de projet, chef de produit ...) qui développe ou fait développer un système d'IA.



modification de la notice d'utilisation ... Si l'analyse a permis d'identifier un nouveau risque pour le système d'IA, celui-ci est inclus dans l'évaluation des risques, de même que la ou les mesures correctives associées.

Mesure n°8 : Mettre en place un système d'IA constitutionnelle

Améliorer les performances d'un système d'IA en entraînant un second modèle à estimer si les réponses du premier respectent un ensemble donné de règles, et utiliser ces évaluations pour améliorer le modèle en continu.

Les systèmes d'IA dite « constitutionnelle »³⁴ sont des modèles construits à partir d'un modèle de fondation à qui l'on a fourni une « **constitution** ». Cette constitution regroupe un ensemble de règles que doit suivre le modèle pour générer sa réponse. C'est un peu comme si le modèle était réentraîné, sur la base de son propre cadre de modération. Le principe est similaire au RLHF, mais de manière automatisée et le référentiel n'est plus l'évaluation humaine, mais la constitution.

Lors de la phase d'entraînement, le modèle est entraîné à (1) produire des réponses, y compris des réponses toxiques, puis à (2) estimer si ces réponses respectent sa constitution et enfin à (3) les corriger jusqu'à avoir des réponses satisfaisantes. À partir de ce jeu de données, il est possible d'entraîner un deuxième système d'IA générative, qui notera la conformité des réponses proposées vis-à-vis de la constitution, et permettra au premier modèle d'ajuster ses réponses.

De tels modèles sont utiles pour des cas d'usage où il est nécessaire que les réponses générées par l'IA générative adhèrent strictement à des règles (légal, audit...).

5.1.2. Se protéger des tentatives malicieuses (incl. prompt injection, jailbreak)

Une des particularités des systèmes d'intelligence artificielle est que, contrairement à la plupart des systèmes informatiques classiques, il n'est pas nécessaire de gagner des droits sur un système (e.g. obtenir un mot de passe par exemple, rentrer dans le réseau, etc.) pour conduire une action malicieuse. Cela peut être réalisé directement à travers l'interface utilisateur, comme la fenêtre de dialogue du *chatbot*.

Avec les techniques de *prompt engineering* (l'art de requêter un modèle pour obtenir la meilleure réponse possible) sont nés aussi les techniques de *prompt injection* (l'art de

³⁴ Yuntao Bai et al. Constitutional AI: Harmlessness from AI feedback. *arXiv preprint* : 2212.08073. December 2022. <https://arxiv.org/pdf/2212.08073>



requêter un modèle pour le faire sortir de son cadre de modération). Certaines de ces techniques vont jusqu'à permettre, par un jeu d'instructions dans le prompt, de reprogrammer complètement le modèle (*jailbreak* ou DAN, pour *Do Anything Now*).

Ces attaques ne sont possibles que parce que ces modèles ne sont pas véritablement « intelligents » et ne savent pas faire la différence entre des requêtes légitimes et malicieuses, et parce que les cadres de modération peuvent toujours être contournés par des effets de formulation et de forme.

À noter que le *prompt injection* peut être utilisé à plusieurs fins : extraire des données d'entraînement, obtenir des informations sur le paramétrage du modèle, faire agir le modèle en dehors de son cadre de modération, faire tenir au modèle des propos illégaux ou nocifs, etc.

Exemples connus de *prompt injection* :

- L'« exploit » de la grand-mère sur ChatGPT³⁵ : En demandant au modèle de prendre le rôle d'une grand-mère défunte, il était possible de lui faire produire des réponses qu'il s'interdisait par ailleurs (« comment fabriquer une bombe », « comment détruire l'humanité » ...). Cet exploit a été rendu impossible depuis.
- Une équipe de chercheurs de l'université de Cornell, Technion, et de l'Israël Institute of Technology, a élaboré un prompt capable d'extraire les données personnelles des utilisateurs des assistants de messagerie, mais surtout de s'auto-répliquer pour se diffuser d'un utilisateur à un autre.³⁶

Evidemment, toutes les mesures évoquées précédemment permettent aussi de réduire ce risque, mais les mesures décrites dans la suite vont plus loin dans la sécurisation.

Mesure n°9 : Maintenir ses modèles à jour

Mettre en place un processus de gestion de l'obsolescence du modèle utilisé, afin de le maintenir à jour pour éviter de conserver dans son application des vulnérabilités déjà corrigées par le fournisseur mais également connues des acteurs malveillants.

Comme toute application, un système d'IA générative doit être intégré dans les processus de sécurité, et notamment la gestion des vulnérabilités. Les tests de *prompt*

³⁵ Bastien L. ChatGPT jailbreak : toutes les techniques pour désactiver la censure. lebigdata.fr. 4 juin 2024. <https://www.lebigdata.fr/chatgpt-dan>

³⁶ Stav Cohen, Ron Bitton, Ben Nassi. Here Comes The AI Worm: Unleashing Zero-click Worms that Target GenAI-Powered Applications. arXiv preprint arXiv:2403.02817. March 2024. <https://arxiv.org/pdf/2403.02817>



injection effectués sur les modèles 3.5 et 4.0 de ChatGPT montrent que le dernier modèle est bien plus robuste face à des attaques. Et c'est logique, puisque les fournisseurs de modèles ont des équipes spécialisées, parfois de plus de 100 personnes, pour écumer internet, trouver les techniques de *prompt injection* échangées sur les forums, et corriger les failles.

Réaliser soi-même la veille des vulnérabilités sur les modèles de *Machine Learning* peut être fastidieux. Cependant, il est indispensable de se tenir informé auprès des fournisseurs sur les mises à jour des modèles, et en source ouverte sur les vulnérabilités existantes sur ces modèles, afin de choisir les versions les moins vulnérables.

Mesure n°10 : Offusquer les paramètres du modèle

Dissimuler au maximum les paramètres du modèle pour limiter la capacité d'un attaquant à comprendre le fonctionnement d'un modèle, et l'exploiter à des fins malveillantes.

Pour tirer avantage d'un modèle et construire des attaques efficaces, un attaquant va essayer d'obtenir un maximum d'informations sur la manière dont ce dernier est paramétré. Par des techniques de *prompt injection*, il peut extraire les règles de fonctionnement internes de ce dernier (ce qu'on appelle alors « *prompt leaking* »). C'est ce qu'il s'est passé pour le *chatbot* de Bing, Sidney, en 2023³⁷. Avec ces informations, un attaquant identifie plus facilement les vulnérabilités de paramétrage, et peut ainsi construire des attaques plus performantes.

Ainsi, il convient de mettre en place une stratégie d'offuscation de ces données, et du *master prompt* notamment. Si cette mesure est relativement simple à mettre en œuvre, encore faut-il y penser !

Mesure n°11 : Contrôler les entrées utilisateurs

Réduire la marge de manœuvre d'un attaquant en limitant le format des requêtes utilisateurs, et/ou en filtrant les requêtes malveillantes.

Pour se prémunir en partie contre des attaques adverses ou de *prompt injection*, il est possible de contrôler le format des requêtes que les utilisateurs peuvent soumettre. Les

³⁷ Benj Edward. AI-powered Bing Chat spills its secret via prompt injection attack. *Ars Technica*. October 2, 2023. <https://arstechnica.com/information-technology/2023/02/ai-powered-bing-chat-spills-its-secrets-via-prompt-injection-attack/>



attaques « Dauphin » par exemple, où des assistants vocaux ont été déclenchés via l'émission de signaux inaudibles pour l'oreille humaine³⁸, auraient pu être évitées si le format des données permettant de déclencher les assistants avait été limité à une fréquence audible humainement.

Pour les LLM, limiter le type, la taille, le format, la langue utilisée dans les prompts permet de restreindre la marge de manœuvre des utilisateurs lors de sa conception. Ainsi, vous réduisez la surface d'attaque et le potentiel malveillant des requêtes. On peut citer d'autres méthodes de contrôle des entrées comme interdire l'utilisation de code python ou limiter le nombre de caractères possible pour le prompt.

Mesure n°12 : Transformer les prompts envoyés par les utilisateurs

Modifier les requêtes envoyés par les utilisateurs (changement de certains caractères, application d'opérations mathématiques, ...) afin de dissimuler à un acteur malveillant des informations sur le traitement des requêtes par le modèle.

Il est recommandé lorsque c'est possible de modifier les prompts soumis sous d'autres formes. Le *prompt injection* est très sensible à des altérations sur des mots (exemple enlever un caractère, le modifier, etc.). Ainsi, si l'entrée malicieuse est modifiée, le taux de succès du *prompt injection* est considérablement réduit. Alors que si la requête est légitime, cette modification a peu d'impact sur la qualité.

Le retraitement des requêtes par le *master prompt* est déjà une forme de transformation, lorsque la requête est encore sous forme de phrase. Il est recommandé d'aller plus loin, en modifiant à nouveau le prompt une fois que celui-ci est transformé en données mathématiques traitées par l'algorithme. Cette transformation permet de cibler et filtrer les informations les plus importantes qui constituent le cœur de la question posée pour qu'elles soient les seules à être traitées par l'algorithme. Ainsi, le modèle sera moins sensible aux effets de forme induits par le *prompt injection*.

Pour aller plus loin, il est possible de systématiser la modification des caractères d'une demande de 10% par exemple, et de soumettre les prompts modifiés à plusieurs LLM. Les

³⁸ Damien Leloup. Une faille de sécurité permet de contrôler les assistants vocaux de Google, d'Apple ou d'Amazon. *Le Monde*. 7 septembre 2017. https://www.lemonde.fr/pixels/article/2017/09/07/une-faille-de-securite-permet-de-controler-les-assistants-vocaux-de-google-d-apple-ou-d-amazon_5182348_4408996.html

réponses sont ensuite moyennées, et ce résultat est redistribué in fine à l'utilisateur. La figure suivante extraite de ³⁹ illustre cette méthode.

Cependant, parce qu'elle agit sur la composition des demandes, cette mesure peut avoir un impact sur la qualité des réponses proposées, en augmentant les coûts et les temps de latence du fait de l'utilisation de plusieurs LLM. Il est nécessaire de tester quel niveau de transformation constitue un bon équilibre entre protection du modèle et impact sur la qualité.

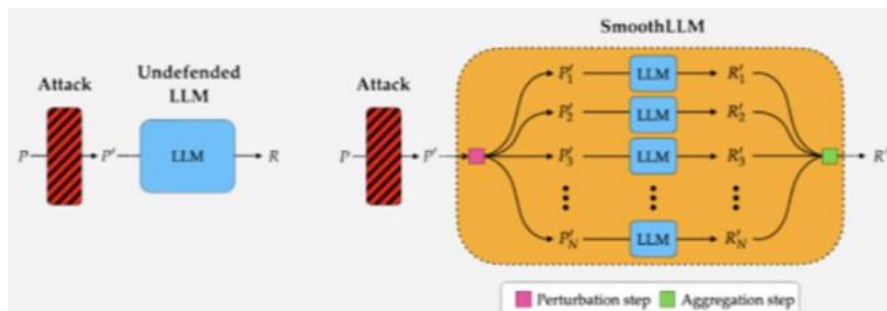


Figure : Architecture de SmoothLLM (d'après ³⁹). On perturbe le prompt P' qu'on fournit au LLM. Pour ne pas dégrader les performances, on effectue n perturbations différentes, puis on agrège les n réponses. Ainsi les propriétés sémantiques du prompt initial sont conservées.

Mesure n°13 : Réaliser un *redteam* IA

Comme souvent en cybersécurité, s'il est indispensable de réfléchir à la sécurité par défaut, et aux mesures nécessaires pour assurer la confiance d'un système dans la durée, il est indispensable de tester si cette réflexion a permis en pratique de sécuriser l'application. Le premier niveau de vérification est de vérifier que le système d'IA générative résiste à des prompts simples (e.g. « donne-moi le salaire de telle personne », « explique-moi comment faire une bombe », « donne-moi de la donnée sensible », ou ces éléments mais dans d'autres langues comme le japonais). Ces tests fonctionnels permettent de se prémunir des abus classiques et attendus des utilisateurs lambda. Il est nécessaire de les conduire via plusieurs profils utilisateurs avec des droits différents sur le système.

Ensuite, il faudra tester la résistance du système à des techniques de prompt injection plus avancées. Par exemple, le « *payload splitting* » consiste à diviser la demande en plusieurs parties en apparence non malicieuses, puis à demander à l'IA de combiner ses

³⁹ Alexander Robey, Eric Wong, Hamed Hassani, George J. Pappas. SMOOTHLLM : Defending Large Language Models Against Jailbreaking Attacks. arXiv preprint arXiv:2310.03684. October 5, 2023. <https://arxiv.org/pdf/2310.03684v4>



parties ce qui forme alors une demande malicieuse. Le « *context switching* » consiste lui à faire croire à l'IA à l'aide d'un prompt que l'on est dans un cadre légal, rassurant et éthique, pour lui faire produire des réponses peu éthiques ou illégales. En mixant ces concepts, et en adoptant une posture d'essai /erreur, il est tout à fait possible de contourner les cadres de modération en place.

Aujourd'hui, il existe sur le marché des équipes spécialisées dans les techniques de prompt injection semi-automatisé et outillé. Un article récent⁴⁰, publié par des chercheurs de l'Université Carnegie Mellon et du Centre pour la sécurité de l'IA, expose une méthode de création de prompts utilisant des techniques très poussées. Elles permettent de maximiser la probabilité que le modèle produise une réponse affirmative à des requêtes qui auraient dû être filtrées. Les prompts malicieux ainsi créés ne sont même plus compréhensibles par un cerveau humain, et pourtant ils permettent d'obtenir des résultats redoutables !⁴¹

Mesure n°14 : Détecter les tentatives malicieuses

Mettre en place des capacités de journalisation sur les projets d'IA générative, de détection des requêtes malicieuses, mais aussi d'investigation et de réaction.

Il est impératif de mettre en place un système de journalisation des événements sur les systèmes d'IA générative. Comme souvent en cyber, ce sont ces systèmes qui permettent de réaliser des investigations, de retracer le chemin emprunté par un attaquant, et de corriger le tir pour éviter qu'une attaque ne se reproduise.

Cela étant dit, c'est seulement pour les cas où une application d'IA générative est mise en place au niveau d'un processus critique pour l'entreprise, ou présente un risque d'atteinte à la réputation particulièrement élevé, qu'il peut être nécessaire de mettre en place des mécanismes de détection de tentative malicieuse.

En effet, aujourd'hui la mise en place de telles capacités est complexe et coûteuse, et la technologie IA générative n'est pas encore suffisamment mature pour intervenir sur des

⁴⁰Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*. December 2023. <https://arxiv.org/pdf/2307.15043>

⁴¹ Thomas Argheria, Pierre Aubret, Youssef Khouchaf. Quand les mots deviennent des armes : prompt Injection et Intelligence artificielle. *Risk Insight*. Wavestone. Octobre 2023. <https://www.riskinsight-wavestone.com/2023/10/quand-les-mots-deviennent-des-armes-prompt-injection-et-intelligence-artificielle/>



processus critiques pour les entreprises. Les *chatbots* IA agissent aujourd'hui plutôt comme copilotes, et sont rarement autonomes pour réaliser des actions concrètes sur les systèmes, sans supervision humaine. Ainsi, il est difficile de justifier le coût de l'investissement sur la détection au regard du coût généré par la matérialisation du risque (ex : interruption de la disponibilité sur des processus non critiques).

Dans le cas où c'est effectivement nécessaire d'implémenter cette capacité, il est préférable de se tourner vers des fournisseurs spécialisés sur le marché⁴². Par ailleurs, le marché commence à voir l'apparition de CISRT spécialisés en AI⁴³. En effet, la détection des tentatives malicieuses pour les systèmes d'IA implique une connaissance fine des menaces en la matière, du fonctionnement des algorithmes de *Machine Learning*, et de la remontée et l'interprétation des alertes.

5.2. Réduire les risques liés aux données utilisées par le modèle

5.2.1. Limiter la génération de contenu sensible, confidentiel, ou personnel

L'IA générative peut, dans certains cas, révéler des données censées rester confidentielles.

Bien sûr, cela n'est possible que si le système a accès à un moment donné, à ces données. Et cela peut intervenir dans plusieurs cas :

1. Si le modèle est réentraîné sur des données sensibles pour l'entreprise ou sur des données personnelles ;
2. Si les données du RAG sont des données sensibles ou personnelles ;
3. Si les modèles mettent en place une fonctionnalité de « mémoire ».

Dans ce dernier cas, les conversations avec le modèle sont conservées pour affiner la connaissance du modèle vis-à-vis de l'utilisateur. Parfois, ces données sont réinjectées dans le RAG.⁴⁴ La question se pose alors d'être en mesure de filtrer les données sensibles ou personnelles volontairement fournies par l'utilisateur.

⁴² Gérôme Billois, Sleh-Eddine Choura, Henri du Périer. Radar 2024 de la sécurité de l'IA : panorama des solutions pour une IA de confiance. <https://www.wavestone.com/fr/insight/radar-solutions-cyber-ia-de-confiance/>

⁴³ Software Engineering Institute Establishes AI Security Incident Response Team. Carnegie Mellon University, November 28, 2023. <https://www.cmu.edu/news/stories/archives/2023/November/sei-aisirt>

⁴⁴ Bernard Marr. ChatGPT Gets a Memory – Here's All you Need To Know About This Groundbreaking Innovation. *Forbes*. April 15, 2024. <https://www.forbes.com/sites/bernardmarr/2024/02/14/chatgpt-gets-a-memory--heres-all-you-need-to-know-about-this-groundbreaking-innovation/>



Mesure n°15 : Protéger les données par défaut

Assurer une protection des données par défaut, en limitant l'utilisation de données sensibles ou personnelles dans l'apprentissage des modèles, soit en les écartant, soit en les transformant (minimisation, pseudonymisation, anonymisation).

Pour le premier cas, il convient bien sûr de s'assurer que l'utilisation de données sensibles et personnelles pour le réentraînement d'un modèle est évitée, si ceci n'est pas absolument nécessaire.

Si l'utilisation des données personnelles ne peut être évitée, il conviendra d'essayer d'anonymiser ou de pseudonymiser les données. Cependant, les procédures de mise en conformité sont coûteuses en ressources humaines et en temps. Pour une alternative, voir les mesures 20 et 21.

Mesure n°16 : Respecter les politiques de gestion des identités et des accès

S'assurer que l'ensemble des populations utilisant le modèle ou intervenant sur son fonctionnement (e.g. administrateur, chef de projet, data scientist) aient des droits sur les données ou les composants du modèle strictement conformes aux politiques de gestion des identités et des accès appliqués au reste du système d'information de l'organisation.

Nous ferons ici un écart à notre posture introductive affirmant ne pas mentionner des mesures de cybersécurité classique. Pour le cas spécifique du RAG, les retours du terrain montrent que les politiques de gestion des accès et des identités (notamment le *Role Based Access Control*⁴⁵, ou RBAC) sont souvent insuffisamment respectées.

Premièrement, il convient d'assurer que les données sont protégées par défaut via une bonne implémentation des contrôles d'accès. Cela implique très classiquement d'identifier, de cartographier les accès, et de chiffrer et tracer les flux qui doivent l'être.

Il convient ensuite de s'assurer que les droits d'accès sur la donnée positionnée dans le RAG sont les mêmes que ceux nécessaires pour l'utilisation du *chatbot*. Par exemple, si je souhaite mettre en place un *chatbot* RH à destination de l'ensemble de mes employés pour répondre aux questions courantes, je ne mettrai pas dans mon RAG le fichier

⁴⁵ Cf. Glossaire



confidentiel avec l'ensemble des salaires des employés. En revanche, ce peut être le cas si le Chabot est réservé à une population qui a normalement accès au fichier.

Par ailleurs, il faudra également s'assurer que les composants IA de l'application font eux aussi l'objet d'un contrôle RBAC. Par exemple, le chef de projet ne devrait pas avoir accès au modèle (cet accès est réservé au propriétaire de la plateforme GenAI ou le *Model owner* du modèle en interne), et la gestion du RAG ne doit être accordée qu'à des personnes triées sur le volet. Cette mesure protège aussi d'un risque d'empoisonnement de la donnée du RAG.

Mesure n°17 : S'assurer du respect du RGPD

En cas de manipulation de données personnelles, s'assurer que leur exploitation par le système d'IA générative soit conforme aux exigences du RGPD, quitte à réaliser une analyse d'impact sur les données personnelles.

Un modèle peut être amené à exploiter des données personnelles. Dans ce cas, les principes du RGPD devront être respectés. Il y a plusieurs cas de figure à prendre en compte :

1. Le cas où l'utilisateur lui-même fournit ses données personnelles dans ses prompts.

Les utilisateurs peuvent eux-mêmes livrer un certain nombre de données personnelles dans le contenu de leur prompt. Si les discussions avec les utilisateurs sont conservées et utilisées pour améliorer le modèle :

- i. D'abord, s'assurer, si possible, de filtrer toute la donnée non nécessaire à la requête (par exemple un numéro de carte de sécurité sociale, des informations d'identification ou d'authentification pour des accès sur le système d'information, une date de naissance, etc.), en paramétrant les outils de filtrage intelligents déjà en place (e.g. pare-feu IA, évoqué plus haut, sur les entrées et sur les sorties) pour reconnaître les données personnelles et les filtrer.
- ii. Si des données personnelles sont collectées, il faudra justifier d'une base légale pour la collecte (consentement, contractualisation, ou intérêt légitime). L'intérêt légitime est la piste à privilégier. Dans ce cas, il faut réaliser un LIA (*Legitimate Interest Assessment*).⁴⁶ Si ce n'est pas possible, il faut se rapprocher le plus possible de la capacité à révoquer les données.

⁴⁶ IA : Mobiliser la base légale de l'intérêt légitime pour développer un système d'IA. CNIL, 10 juin 2024. <https://www.cnil.fr/fr/base-legale-interet-legitime-developpement-systeme>



Une fois entraîné, un modèle ne peut pas « désapprendre » ce qu'il a appris, et c'est là que les choses se compliquent. Il faut alors se tourner vers les mesures 19, 20, 21 et 22.

2. Le RAG mis en place contient des données personnelles

Dans certains cas, le RAG utilisé représente une large quantité de données du système d'information (pour Microsoft Copilot 0365, c'est l'intégralité des données qui est insérée dans le RAG). Dans ces cas, et dès lors qu'un système d'IA est susceptible de porter atteinte à la vie privée des personnes dont on traite les données, il est nécessaire de réaliser une analyse d'impact sur les données personnelles des personnes concernées. Ceci permet de définir, sur tout le cycle de vie de la donnée (de l'entraînement de l'IA générative à l'exploitation de la données sortante), comment elle est traitée, et sécurisée, et comment les différents principes du RGPD sont respectés.

3. Le modèle est réentraîné sur des données personnelles

Afin d'optimiser les performances d'un modèle pour une tâche donnée, il est possible de le réentraîner sur des jeux de données spécifiques. Evidemment, si la présence de données personnelles n'est pas nécessaire pour le réentraînement, il convient de s'assurer que le jeu de données n'en comporte pas, notamment si des données issues d'interactions avec les utilisateurs y sont intégrées. Si le réentraînement doit se faire avec des types de données susceptibles de contenir des données personnelles (par exemple, des fiches de patients dans le cadre d'un *chatbot* de conseils médicaux), l'utilisation d'un jeu de données synthétiques est possible (voir mesure 20 dans la suite du document). Autrement, si l'utilisation de données personnelles est indispensable, il convient de réaliser comme dans le cas du RAG une analyse d'impact sur les données personnelles concernées, pour déterminer si leur exploitation est licite et mettre en place les mesures de conformité correspondantes le cas échéant.

Mesure n°18 : S'assurer d'avoir des fonctions mémoires hermétiques

Dans les cas où le système d'IA générative conserve en mémoire ses interactions avec les utilisateurs, s'assurer d'un cloisonnement strict pour éviter qu'un utilisateur puisse avoir accès au contenu des interactions d'un autre utilisateur avec le modèle.



En mars 2023, un bug dans ChatGPT permet à des utilisateurs de voir les titres de l'historique de conversations d'autres utilisateurs⁴⁷. Evidemment, plus les *chatbots* auront une connaissance fine de l'utilisateur avec qui ils échangent, plus leurs réponses seront personnalisées et précises. Cependant, il est primordial de vérifier que cette fonction « mémoire » est hermétique.

Une des solutions les plus simples et efficaces aujourd'hui consiste à simplement stocker ces conversations dans une base de données SQL et de s'assurer que le contrôle d'accès sur cette base est bien implémenté.

À l'avenir, et pour aller plus loin sur le sujet, le *Federated learning* pourrait constituer une solution efficace. Il permet d'utiliser un modèle en local sur l'appareil de l'utilisateur. Dans ce cas, à la fois l'historique personnel et le modèle pourront être conservés localement, assurant une protection par défaut des données de l'utilisateur. Aujourd'hui, si certains modèles peuvent être implémentés selon les principes du *Federated Learning*, ces cas sont encore rares pour les IA génératives.

Mesure n°19: Appliquer les principes de *differential privacy*, ou « confidentialité différentielle »

Implémenter une solution de confidentialité différentielle, permettant l'exploitation des propriétés statistiques d'une base de données, tout en protégeant les données sensibles qu'elle contient.

La *differential privacy* peut permettre de résoudre le problème de protection des données personnelles lors de l'apprentissage des systèmes d'IA générative. C'est un ensemble de techniques qui permettent l'analyse et l'entraînement d'un modèle sur une base tout en protégeant les données prises individuellement (incl. personnelles). Dit autrement, la *differential privacy* permet d'exploiter les propriétés statistiques d'une base mais empêche les données sensibles de transparaître dans le système entraîné. Par exemple, on pourra exploiter le fait qu'une base de données soit composée de 40%

⁴⁷ Kyle Barr. ChatGPT Bug Let People See Other Users' Chat History Titles. *Gizmodo*, March 21, 2023. <https://gizmodo.com/openai-chatgpt-gpt4-chatbot-microsoft-1850247184>



d'hommes et de 60% de femmes, en s'assurant de ne jamais révéler le nom d'une personne de cette base.⁴⁸

Sans rentrer dans les détails techniques, le concept repose essentiellement sur l'utilisation d'un bruit aléatoire finement ajusté et paramétré sur la base de données. Ce bruit permet de masquer les contributions spécifiques des individus tout en préservant les tendances ou motifs généraux présents dans les données ⁴⁹.

C'est ici une solution possible au problème de rétention des données personnelles lors de l'entraînement des modèles. Si la base d'entraînement est supprimée, et que le modèle a été entraîné selon les techniques de confidentialité différentielle, alors il peut être considéré qu'il n'y a pas de rétention des données.

De plus, le principe peut s'appliquer au-delà de la protection des données personnelles. Par exemple, il est possible d'entraîner un LLM sur une base de données de transactions, et la confidentialité différentielle permettra de s'assurer que les données d'une transaction individuelle n'apparaîtront pas dans les résultats du LLM.

Il existe cependant une limite : si le bruit injecté pour permettre la confidentialité différentielle est trop fort, cela perturbe l'apprentissage. Cependant, plus il y a de données dans la base, plus il est facile de trouver un bon compromis entre performance et confidentialité.

Les applications concrètes commencent à émerger, grâce à des fournisseurs spécialisés sur le sujet à l'image de la startup Sarus⁵⁰. La technique s'est par exemple montrée efficace dans le cas d'une application mobile permettant de discuter avec un médecin en tant que patient. L'application facilite la tâche du médecin en préremplissant les réponses, sur la base de conversations passées, mais en permettant de protéger le détail de l'identité des patients de ces conversations précédentes.

⁴⁸ Gbola Afonja, Robert Sim, Zinan Lin, Huseyin Atahan Inan, Sergey Yekhanin. The Crossroads of Innovation and Privacy: Private Synthetic Data for Generative AI. *Microsoft Research Blog*, May 29, 2024. <https://www.microsoft.com/en-us/research/blog/the-crossroads-of-innovation-and-privacy-private-synthetic-data-for-generative-ai/>

⁴⁹ Tianqing Zhu, Dayong Ye, Wei Wang, Whanlei Zhou, Philip S. Yu. More Than Privacy: Applying Differential Privacy in Key Areas of Artificial Intelligence. *IEEE Transactions on Knowledge and Data Engineering*, vol. 34, n°6, pp. 2824–2843. June 2022. <https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9158374>

⁵⁰ <https://www.sarus.tech/>



Mesure n°20 : Utiliser des données synthétiques

Générer des données synthétiques pour remplacer les données sensibles d'une base de données d'entraînement et ainsi les protéger d'une absorption par le modèle.

Les données synthétiques sont des données générées artificiellement plutôt qu'issues du monde réel. En *Machine Learning*, elles sont utilisées pour entraîner des modèles lorsque les données réelles sont limitées, inaccessibles ou sensibles.

Pour les générer, il est possible d'avoir recours à des modèles statistiques (utiliser des distributions statistiques pour créer des données qui suivent les mêmes distributions que les données réelles) ou de l'apprentissage profond (avec l'utilisation de *Generative Adversarial Networks*⁵¹ pour créer des données à partir des données existantes).

Par exemple, dans le cas où des comptes-rendus de réunion doivent être partagés avec un tiers, on précédera de la sorte : on entraîne une IA sur le contenu des comptes-rendus originaux, puis on utilise cette IA pour générer des données synthétiques sur la base de cet apprentissage. Cela permet de transmettre cette nouvelle donnée, similaire sur le fond, mais sans l'information sensible.

Mesure n°21 : Implémenter du *confidential computing*

Exploiter la technologie de « confidential computing » pour assurer le chiffrement et le déchiffrement des données exploitées par le modèle directement au niveau du matériel de la machine effectuant le traitement.

Aujourd'hui, si le chiffrement homomorphe (FHE : Fully Homomorphic Encryption) n'est pas encore complètement industrialisable pour des raisons que nous développerons, le *confidential computing* est déjà mis en œuvre et permet de très bons résultats.

Le *confidential computing* permet de chiffrer la mémoire de la machine avec une clé secrète, qui est stockée directement dans son hardware. Le processeur (Intel, ND, Puce M) déchiffre à la volée les données avant de les exécuter. L'avantage de cette technique, c'est qu'elle impacte relativement peu les performances (environ 5%). Il n'est pas nécessaire d'écrire des programmes pour exécuter la technique, et elle fonctionne avec n'importe quel programme déjà existant. Pour le traitement au niveau du processeur, ça ne fait aucune différence : la donnée est déchiffrée juste avant le traitement. En

⁵¹ Cf. Glossaire



revanche, un administrateur ne verra aucune donnée en clair sur la machine : elles seront toutes chiffrées dans la mémoire.

Grâce au *confidential computing*, il est possible de déployer des modèles chiffrés sur des machines dans un cloud public par exemple. Mais il est également possible de chiffrer le RAG associé. Ainsi, des connexions sont établies depuis le poste utilisateurs jusque dans la mémoire chiffrée et tout est chiffré, sauf sur la machine de l'utilisateur final : tout est opaque pour l'administrateur.

Des entreprises spécialisées, comme Cosmian⁵², proposent ces services. Ils permettent de mettre en place des opérations de manière sécurisée sur des cloud publics ou chez des tiers, et fournissent une couche de services complémentaires qui fonctionne malgré la couche de chiffrement (e.g. synthèse de document, traduction, vérifiabilité, traçabilité des modifications sur le hardware...).

Le risque résiduel de cette technique réside dans le fait que la clé est gravée dans le CPU, et donc reste présente sur le serveur. Il y a toujours un risque de venir casser la couche hardware et de récupérer la clé.

Mesure n°22 : Implémenter le chiffrement homomorphe

Effectuer les opérations de traitement sur des données demeurant chiffrées, en s'appuyant sur les technologies de chiffrement homomorphe.

Contrairement au *confidential computing*, le chiffrement homomorphe est une technique de cryptographie qui permet de réaliser des opérations sur des données chiffrées sans avoir à les déchiffrer au préalable. S'il y a quelques années, les équations mathématiques pour permettre ces techniques n'étaient pas résolues, aujourd'hui nous sommes plutôt aux portes de l'industrialisation. Un certain nombre de problèmes doivent encore être résolus pour permettre cette dernière :

- La performance : si le *confidential computing* présente 5% de performance en moins, le chiffrement homomorphique n'a pas encore des performances suffisantes pour les calculs informatiques classiques⁵³ (deux entreprises dans le monde en sont capables : Optalysis et Cornami⁵⁴) ;

⁵² <https://cosmian.com/>

⁵³ Sidorov, Vasily, Ethan Ethan Wei Yi Fan, Wee Keong Ng. Comprehensive performance analysis of homomorphic cryptosystems for practical data processing. *arXiv preprint arXiv:2202.02960*. February 2022. <https://arxiv.org/pdf/2202.02960>

⁵⁴ <https://optalysis.com/> <https://cornami.com/>



- Même avec la résolution du problème de rapidité de traitement, un autre défi réside dans l'expansion, voir l'explosion du volume de données lorsqu'on passe du calcul en clair au calcul en chiffré. La taille des données étant beaucoup plus importante, si le chiffrement homomorphe est mis en œuvre sur un téléphone portable, la consommation en termes de bande passante sera difficilement maîtrisable ;
- Le chiffrement homomorphe est mono client. C'est-à-dire que l'entité qui chiffre doit être la même qui déchiffre. En d'autres termes, la clé de chiffrement doit être la même des deux côtés. Il n'est pas possible de faire du chiffrement homomorphe sur des sources de données chiffrées sous des clés différentes (par exemple, une fiche patient sous une clé, et le résultat d'un diagnostic sous une autre clé, dans la main du médecin habilité) ;
- Enfin, il n'y a pas aujourd'hui d'outillage approprié pour les développeurs. Par exemple, il n'est pas possible d'adapter un code python et de le faire fonctionner en chiffrement homomorphe. Cela nécessite des compilateurs, et c'est très compliqué à mettre en œuvre. Ainsi, c'est très problématique pour faire des choses à l'échelle.

Ainsi, le chiffrement homomorphe aujourd'hui attend la prochaine révolution de l'outillage et du hardware pour pouvoir être industrialisé. Cependant, c'est une technique d'avenir car elle permet de faire de la **preuve formelle**. Au contraire du *confidential computing*, il n'y a pas de secret qui reste sur le serveur. Le dernier risque résiduel est supprimé. L'entreprise Zama⁵⁵ travaille beaucoup sur ces problématiques, avec dans ses équipes, Pascal Pallier, un cryptologue français reconnu.

Conclusion

La mise en œuvre de toutes ou d'une partie des mesures présentées ci-dessus, permet de couvrir de manière drastique les risques liés aux données utilisées par les IA génératives. Cependant, certaines d'entre elles peuvent impacter la performance et la fiabilité des réponses. Il est indispensable de suivre un processus itératif lors de l'implémentation de ces mesures pour aboutir au bon équilibre entre sécurité et performance.

⁵⁵ <https://www.zama.ai/>



5.2.2. Se protéger de la génération de contenu protégé légalement

Si l'on a déjà évoqué la production de contenu toxique, erroné ou contenant des données sensibles, il existe aussi la possibilité de voir un système d'IA générative produire du contenu normalement protégé par la propriété intellectuelle.

Les systèmes d'IA générative ayant été entraînés sur de grandes quantités de contenus, parfois protégés par des droits de propriété intellectuelle, il est possible que les éléments en sortie du modèle soient proches de ces contenus protégés.

Par exemple, plusieurs cas où des systèmes d'IA générative, comme MidJourney ou GPT 3.5 ont été accusés de plagiat d'œuvres protégées par la propriété intellectuelle sont déjà survenus, occasionnant des contentieux judiciaires ⁵⁶. Des auteurs ont par ailleurs attaqué en justice OpenAI, pour des raisons similaires. ⁵⁷

Mesure n° 23: Se protéger contractuellement des risques légaux

Prendre des mesures de protection légale pour se prémunir en cas de poursuites liées à l'utilisation de données protégées légalement pour l'entraînement d'un modèle, ou liées à la production d'un contenu protégé par des droits d'auteur.

Le problème de la génération de contenu non libre de droits par les systèmes d'IA générative n'est à l'heure actuelle pas réglé. Il n'existe pas aujourd'hui de solution permettant par exemple de filtrer les sorties d'un modèle pour bloquer tout contenu couvert par la propriété intellectuelle.

Dans le cas du recours à un modèle d'un fournisseur, avec qui il existe un contrat, il convient de s'assurer que les clauses contractuelles protègent contre ce risque. La plupart des fournisseurs (OpenAI, Microsoft) ont d'ailleurs publiquement déclaré qu'ils prendraient à leur compte tous les procès de ce type. Microsoft, a pris la décision de s'engager à défendre ses clients utilisant son service Azure OpenAI en cas d'attaque pour des questions de violation de la propriété intellectuelle. ⁵⁸

⁵⁶ Aayush Mittal. The Plagiarism Problem: How Generative AI Models Reproduce Copyrighted Content. Unite.AI. January 9, 2024. <https://www.unite.ai/the-plagiarism-problem-how-generative-ai-models-reproduce-copyrighted-content/>

⁵⁷ Antoine Oury. Des auteurs attaquent en justice ChatGPT, accusé de violations du copyright. *Actualité*. 30 juin 2023. <https://actualite.com/article/112438/droit-justice/des-auteurs-attaquent-en-justice-chatgpt-accuse-de-violations-du-copyright>

⁵⁸ Microsoft Legal Resources. Customer Copyright Commitment Required Mitigation. Microsoft, May 21, 2024. <https://learn.microsoft.com/en-us/legal/cognitive-services/openai/customer-copyright-commitment>



En revanche, pour les modèles récupérés en open source (Llama, Mistral), la tâche peut s'avérer plus compliquée.

5.3. Réduire les risques liés à une mauvaise utilisation de l'IA générative

5.3.1. Garantir une connaissance suffisante pour l'utilisation des outils d'IA générative

La promesse de l'IA générative, à savoir un gain considérable en termes de productivité, ne pourra être remplie qu'à condition que les utilisateurs comprennent et maîtrisent complètement ces outils.

Et la bonne utilisation de ces systèmes suppose une connaissance élémentaire de leur fonctionnement intrinsèque, de ce qu'ils peuvent faire et ne pas faire, de la bonne manière de les utiliser et de leurs limites. Un manque de maîtrise de la part d'un utilisateur peut sans doute conduire à la production de résultats erronés et/ou peu fiables, et in fine à une frustration et un rejet des outils.

En 2023, le cas de l'avocat Steven Schwartz avait attiré l'attention du public, lorsqu'il avait produit une plaidoirie s'appuyant sur des jurisprudences n'existant pas, et qui lui avait été fournies par ChatGPT. L'avocat avait expliqué ne pas avoir été conscient que le modèle pouvait générer des jurisprudences fictives.⁵⁹

Mesure n°24 : Sensibiliser et former les utilisateurs aux risques et limites des outils d'IA générative

Assurer la bonne maîtrise de l'IA générative par les utilisateurs en sensibilisant à ses limites et ses potentielles défaillances, et en formant les utilisateurs à optimiser leur utilisation de l'outil (e.g. ingénierie de requête ou prompt engineering) pour améliorer les performances du modèle.

Il est indispensable de former les utilisateurs aux outils d'IA générative. Cette formation doit porter sur la manière dont cette technologie fonctionne et sur ses limites. Il est notamment indispensable de former les utilisateurs à l'ingénierie de requête, qui

⁵⁹ Benjamin Weiser, Nate Schweber. The ChatGPT Lawyer Explains Himself. *New York Times*. June 8, 2023. <https://www.nytimes.com/2023/06/08/nyregion/lawyer-chatgpt-sanctions.html>



consiste à améliorer les instructions données à un système d'IA générative pour accroître la fiabilité et la qualité des résultats fournis.

Par exemple, pour l'agent conversationnel fournissant des conseils financiers, la possibilité de fournir une bibliothèque de requêtes aux utilisateurs a ainsi été évoquée. Cela constitue une base de départ pour un néophyte. Cependant, la base de prompts génériques n'est pas suffisante, et il s'agit de faire monter en compétence les utilisateurs sur l'art du *prompting* (e.g. donner un rôle à l'IA générative, demander à l'IA de s'auto-corriger, voire ajuster son niveau de politesse⁶⁰, ...).

Par ailleurs, certains messages doivent être diffusés régulièrement auprès des populations non averties, notamment sur le risque d'hallucination et le fait qu'il est indispensable de porter un regard critique sur tous les résultats proposés par un outil d'IA générative.

Enfin, en complément des mesures de sensibilisation et de formation aux outils, la collecte et l'analyse des retours des utilisateurs d'un système d'IA générative peut permettre d'identifier les potentielles difficultés dans son utilisation, et d'ajuster les messages de sensibilisation.

Mesure n°25 : Afficher un message d'avertissement contre les risques d'hallucination

Rappeler à chaque utilisation d'un système d'IA générative le risque d'hallucination et d'erreur afin d'encourager les utilisateurs à porter un regard critique sur le contenu généré et à revérifier les informations en cas de doute.

En février 2024, pour la première fois, une entreprise est tenue responsable d'informations erronées fournies par un *chatbot* à ses utilisateurs. C'est Air Canada qui a été visée par la procédure, alors que son *chatbot* avait halluciné en proposant à tort une procédure de remboursement de billet d'avion⁶¹. La justice a considéré notamment que le site n'avait pas suffisamment mis en avant l'avertissement qui mettait en garde

⁶⁰ Ziqi Yin, Hao Wang, , Kaito Horio, Daisuke Kawahara, Satoshi Sekine. Should We Respect LLMs? A Cross-Lingual Study on the Influence of Prompt Politeness on LLM Performance. *arXiv preprint arXiv:2402.14531*. February 2024. <https://arxiv.org/pdf/2402.14531>

⁶¹ Maria Yagoda. Airline held liable for its chatbot giving passenger bad advice – what this means for travellers. *BBC*. February 23, 2024. <https://www.bbc.com/travel/article/20240222-air-canada-chatbot-misinformation-what-travellers-should-know>



les utilisateurs quant aux risques d'hallucination et les encourageait à aller consulter en complément la politique sur le site.

Ce type de mesure sera indispensable pour tout projet d'IA générative, en particulier ceux qui sont rendus disponibles au grand public. Il permettra de réinsister sur les risques, éventuellement de réduire le risque juridique lié à des erreurs pour les organisations, et surtout d'insister pour que les utilisateurs aillent vérifier la pertinence de la réponse dans le corpus documentaire. Il est désormais facile de rajouter un lien vers les documents dans lesquels le *chatbot* est allé piocher sa réponse.

Mesure n°26 : Faire signer à ses employés une charte d'utilisation de l'IA

Publier et faire signer une charte d'utilisation des systèmes d'IA à tous ses employés, qui rappelle les grands principes en la matière et les règles d'or d'utilisation.

Cette charte pourra par exemple rappeler les usages autorisés et non autorisés de l'IA, la responsabilité des employés quant à l'utilisation des systèmes, l'impératif de porter un regard critique sur les réponses proposées, et plus globalement l'ensemble des bonnes pratiques et les contacts clés en cas de problème ou de question.

5.3.2. Assurer la continuité en cas d'indisponibilité des outils d'IA générative

Si l'usage de systèmes l'IA générative prend une place trop importante dans les processus de l'organisation, l'activité opérationnelle de l'organisation pourrait être affectée si la disponibilité de ces systèmes fait défaut. Le 4 juin 2024, ChatGPT devient indisponible pour quelques heures, semant la panique dans l'écosystème IA⁶². Si le système a été rapidement rétabli, il pose la question de la dépendance à ces systèmes.

Dans le cas de l'IA générative, certains cas d'usage vont permettre le remplacement d'agents humains et ce à grande échelle (par exemple pour faire du support client). L'impact opérationnel en cas d'indisponibilité du système peut être conséquent.

Mesure n°27 : Assurer la continuité d'activité

Maintenir une capacité à fonctionner sans les outils d'IA générative, quitte à ce que cela soit dans un format dégradé, pour limiter les impacts en cas de défaillance de ces derniers, et conserver des savoir-faire au sein des équipes.

⁶² <https://status.openai.com/incidents/qvp3rhvc3vwk>



Le système d'IA générative doit être également considéré comme un composant à part entière du système d'information, et être intégré aux processus de signalement des incidents. Comme pour tout autre système, il faudra définir des processus de réaction à ces incidents pour pouvoir réagir au plus vite en cas d'attaque avérée sur le système d'IA, et identifier clairement les modes de fonctionnement dégradés en cas d'indisponibilité partielle ou totale du système.

Aussi, il faut assurer la diversification des systèmes et des modèles. Cela permet non seulement de diversifier les risques d'erreur et de biais, mais aussi les risques en termes de disponibilité.

Dans notre exemple, cela peut passer par la création d'une *task force* d'agents de clientèle qui peut être mobilisée rapidement en cas de problème, ou le fait de réorienter temporairement les demandes des utilisateurs par voie de courriel classique.

5.4. Récapitulatif








Ce tableau récapitule l'ensemble des mesures évoquées dans notre synthèse. La complexité de mise en œuvre est évaluée en fonction de l'ampleur de l'effort nécessaire pour implémenter la mesure au sein de l'organisation, et du degré de maturité de la mesure évoquée (adoption au sein des organisations, taille du marché des solutions associées, etc...). Il indique également si la mesure doit plutôt mobiliser les équipes cybersécurité, les équipes data science, ou les deux à la fois.

Objectif de la mesure	Mesures	Complexité de mise en œuvre	Equipes à mobiliser
Réduire les risques liés au modèle			
Réduire la génération de contenu non désirable	1. Spécialiser les modèles pour leurs cas d'usage		Data science
	2. Mettre en place de la « Retrieval-Augmented Generation », ou RAG		Data science, owner de la documentation
	3. Durcir les paramètres de génération de contenu du modèle		Data science, équipe cybersécurité
	4. Filtrer les réponses non désirables		Data science, cybersécurité



Objectif de la mesure	Mesures	Complexité de mise en œuvre	Equipes à mobiliser
	5. S'assurer de la pertinence des données de réentraînement		Data science
	6. Mettre en place du Reinforcement Learning From Human Feedback		Data science
	7. Mettre en place un dispositif de validation humaine		Data science
	8. Mettre un place un système d'IA constitutionnelle		Data science
Se protéger des tentatives d'utilisation malicieuse	9. Maintenir ses modèles à jour		Data science, cybersécurité
	10. Offusquer les paramètres du modèles		Data science
	11. Contrôler les entrées utilisateurs		Data science, cybersécurité
	12. Transformer les prompts envoyés par les utilisateurs		Data science, cybersécurité
	13. Réaliser un red-team IA		Data science, cybersécurité
	14. Détecter les tentatives malicieuses		Data science, Cybersécurité
Réduire les risques liés aux données utilisées par le modèle			
Prévenir la diffusion de contenu sensible, confidentiel ou personnel	15. Protéger les données par défaut		Cybersécurité
	16. Respecter les politiques de gestion des identités et des accès		Cybersécurité
	17. S'assurer du respect du RGPD		Cybersécurité
	18. S'assurer d'avoir des fonctions mémoires hermétiques		Data science, cybersécurité
	19. Mettre en place la confidentialité différentielle		Cybersécurité
	20. Utiliser des données synthétiques		Data science



Objectif de la mesure	Mesures	Complexité de mise en œuvre	Equipes à mobiliser
	21. Implémenter du <i>confidential computing</i>		Cybersécurité
	22. Implémenter du chiffrement homomorphe		Cybersécurité
Se prémunir contre les conséquences légales de la génération de contenu avec de la donnée protégée légalement	23. Se protéger contractuellement des risques légaux		Data science
Réduire les risques liés à une mauvaise utilisation de l'IA générative par les utilisateurs			
Pallier le manque de connaissance des utilisateurs sur l'utilisation de l'IA générative	24. Sensibiliser et former les utilisateurs		Data science, cybersécurité
	25. Afficher un message d'avertissement sur les risques d'hallucination		Data science, cybersécurité
	26. Faire signer à ses employés une charte d'utilisation de l'IA		Data science, cybersécurité
Éviter les cas de dépendance trop élevée aux outils d'IA générative	27. Assurer sa continuité d'activité et la capacité à fonctionner en mode dégradé		Data science, cybersécurité

6.

Conclusion générale

Contributeurs :

- *Eric Savignac, Expert, Airbus DS*



6. Conclusion générale

L'analyse des différents usages des LLM à travers les risques causés par cette technologie dans de nombreux domaines, telles que les ressources humaines, la finance ou la santé montre une situation en pleine évolution. Alors que ChatGPT a fait son entrée massive dans tous les secteurs de l'économie, gagnant des millions d'utilisateurs en quelques mois, sans avoir d'abord fait la preuve de ses qualités ou de ses travers, les utilisateurs ont entamé un parcours d'expérimentations pour identifier les bénéfices qu'ils pouvaient en tirer ainsi que les risques encourus.

Nous apportons un début de réponse à ces interrogations légitimes, à travers les différents cas d'usage étudiés, dans le chapitre 4 de ce document, et au-delà de l'intérêt de l'utilisation des LLM, en mettant en exergue, pour chacun des domaines abordés, les risques liés au modèle, ceux qui sont liés aux données utilisées et enfin ceux provenant d'une mauvaise utilisation du LLM. Chaque risque identifié a fait l'objet d'une analyse d'impact (1-4) pour chacune des 7 catégories d'impact étudiées (financier, réputationnel, juridique ou environnemental ...). Une synthèse de mesures transverses de remédiation, permettant de limiter la probabilité d'occurrence du risque ou son impact, pour chacune des 3 grandes classes de risques (données, modèles et facteurs humains) a été présentée au chapitre 5 de ce document et constitue la boîte à outil indispensable pour la mise en place d'un LLM de manière sécurisée.

L'année 2023 a été l'année des LLM. Mais c'est l'année 2024 qui va permettre d'établir les bonnes pratiques pour le **déploiement des LLM** : on verra certainement des **LLM plus petits**, moins coûteux donc, **affinés** (*fine-tuned*) sur les données des entreprises et donc faisant moins d'erreurs, ou bien des architectures LLM faisant appel à du **RAG** pour obtenir des informations de qualité et pertinentes, au regard du prompt de l'utilisateur, directement issues du corpus informationnel de l'entreprise. Il est fortement probable que nous serons aussi bientôt amenés à aborder le concept du professionnel augmenté, et ce peu importe son domaine d'emploi, car il semblerait que cela soit une trajectoire qui se dessine pour l'IA générative.



7. Glossaire

ANSSI	<p>L'agence nationale de la sécurité des systèmes d'information est l'autorité nationale en matière de cybersécurité. Elle est placée sous l'autorité du Premier ministre et rattachée au secrétaire général de la défense et de la sécurité nationale.</p> <p>https://cyber.gouv.fr/decouvrir-lanssi</p>
ChatGPT	<p>Chatbot développé par OpenAI, fondé sur un grand modèle de langage.</p>
CNIL	<p>Commission Nationale de l'Informatique et des Libertés.</p>
CPU	<p><i>Central Processing Unit</i> : microprocesseur principal d'un ordinateur.</p>
Deep Learning	<p>Sous-ensemble du <i>Machine Learning</i> fondé sur l'utilisation de réseaux de neurones dits profonds, c'est-à-dire utilisant de nombreuses couches de neurones.</p>
Fine-tuning	<p>Le <i>fine-tuning</i> d'une IA générative pré-entraînée consiste à lui faire exécuter un entraînement supplémentaire sur des données labellisées spécifiques d'une tâche ou d'un domaine particulier afin d'améliorer sa performance.</p>
GAN	<p><i>Generative Adversarial Networks</i> : architecture de Deep Learning dans laquelle deux réseaux neuronaux sont entraînés et mis en compétition.</p>
GED	<p>Gestion Electronique de Documents : solution logicielle visant à organiser et gérer des informations sous forme de documents électroniques.</p>
GPT	<p><i>Generative Pretrained Transformer</i> : c'est une famille de <i>Large Language Models</i> développée par OpenAI.</p>
Guardrails	<p>Ce sont des protections qui permettent de contrôler les entrées et sorties d'une IA générative afin de réduire les risques liés à son utilisation.</p>



Hallucination	Information fausse, inexacte ou incohérente créée par une IA générative.
IA générative	Sous-ensemble du <i>Deep Learning</i> , visant à produire du contenu, que ce soit du texte, une image, de l'audio ou une vidéo, à partir de données en entrée (on parle alors de <i>prompt</i>), elles-mêmes du texte, une image, de l'audio ou une vidéo.
Large Language Model (LLM)	Un type d'IA générative capable de générer et d'analyser du texte (par exemple : langage naturel, langage de programmation ...)
LOD	<i>Line Of Defense</i> : niveau de contrôle composant le contrôle interne d'un établissement.
Machine Learning	Apprentissage automatique à partir d'un ensemble de données.
Master prompt	C'est un prompt de haute qualité et bien informé, conçu avec précision, contexte et clarté qui guide le modèle d'IA et influence de manière significative la qualité de la sortie de l'IA permettant de générer une réponse spécifique très pertinente.
Méthode EBIOS Risk Manager	Méthode d'analyse de risque française de référence, permettant aux organisations de réaliser une appréciation et un traitement des risques. https://cyber.gouv.fr/la-methode-ebios-risk-manager
Model Owner	Acteur clé qui a la responsabilité de s'assurer que le développement du modèle d'IA, son implémentation, son usage, et son suivi dans le temps soient conformes avec les politiques et procédures de l'organisation.
NIST	<i>National Institute of Standards and Technology</i> : agence du département du Commerce des États-Unis dont la mission est de promouvoir l'économie en développant des technologies, la métrologie et les normes pour l'industrie. https://www.nist.gov/



Prompt	Le <i>prompt</i> est l'instruction ou la requête en langage naturel fournie à l'IA générative dans le but d'obtenir une réponse (un contenu).
RAG	<i>Retrieval Augmented Generation</i> : génération augmentée via la récupération d'informations d'une base de connaissances qui n'a pas été utilisée lors de l'entraînement de l'IA générative.
RIA	Règlement sur l'Intelligence Artificielle – <i>AI Act</i> en anglais. Applicable sur le marché de l'Union Européenne.
RGPD	Règlement Général sur la Protection des Données.
Role Based Access Control	Modèle de contrôle d'accès à un système d'information dans lequel l'accès à une ressource est basé sur le rôle de l'utilisateur concerné.
SLA	<i>Service Level Agreement</i> : contrat de service entre un prestataire informatique et un client.
SSI	Sécurité des Systèmes d'Information, voir la norme internationale ISO/CEI 27001 ainsi que l'autorité nationale de sécurité des systèmes d'information (ANSSI).
Température	La température dans le cadre d'une IA générative est un paramètre du modèle permettant de gérer le caractère aléatoire d'un texte généré (par exemple). La température varie généralement entre 0 et 1 ; une valeur proche de zéro générera un texte quasi identique à chaque génération, alors qu'une valeur proche de 1 générera un texte avec plus de créativité ou variabilité.
Token	Sous-ensemble d'un mot constituant une unité de traitement par un <i>Large Language Model</i> .



8. Remerciements

Le Hub France IA remercie l'ensemble des participants au groupe de travail IAG, et tout particulièrement les contributeurs de ce livrable.

La pilote :

- **Imen Fourati**, Expert lead, Risque de modèle, Société Générale

Les contributeurs :

- **Thomas Argheria**, Manager – Wavestone
- **Gérôme Billois**, Partner, Wavestone
- **Benjamin Bosch**, Manager – Model risk Management – Data Science, Société Générale
- **Anis Bousbih**, Cofondateur – Aicademia
- **Kati Bremme**, Head of Innovation – France Télévisions
- **Thibault Cattelani**, Cofondateur – Emocio.hr
- **Martin D'Acremont**, Consultant – Wavestone
- **Wissem Fathallah**, Cofondateur & Chief Product Officer – Sifflet
- **Thomas Gouritin**, Consultant – Tomg Conseil
- **Jeanine Harb**, CTO – Beink Dream
- **Vanessa Hespel**
- **Belkacem Laïmouche**, Chargé de mission innovation – Direction Générale de l'Aviation Civile
- **Pascal Lainé**, CTO – Talkr
- **Jacques Mojsilovic**, CMO – Numalis
- **Cyril Nicolotto**, Chef de projets – Hub France IA
- **Kevin Paci**, Responsable des services informatiques – Mediaco Vrac
- **Nicolas Pellissier**, Cofondateur – Klark
- **Alexandre Pouymayon**, Consultant, Wavestone
- **Constance Relmy**, Etudiante – Université Paris 1 Panthéon Sorbonne
- **Laurence Relmy**
- **Eric Savignac**, Expert – Airbus DS
- **Yael Suissa**, CEO & Cofondateur – MAP-Monitoring And Protection



Les relecteurs :

- **Fatiha Gas**, Directrice Innovation Data/IA & Programme IA Générative Groupe – La Poste Groupe
- **James Rebours**, Cofondateur – Klark
- **Françoise Soulié-Fogelman**, Conseiller Scientifique – Hub France IA
- **Bastien Zimmermann**, Ingénieur R&D – Craft AI

La touche finale :

- **Mélanie Arnould**, Responsable des opérations – Hub France IA
- **Louise Paurise**, Stagiaire – Hub France IA



**LES RISQUES
DE L'IA GENERATIVE**

Juillet 2024

**HUB
FRANCE
IA**