HUB FRANCE IA

CHAT GPT : USAGES, IMPACTS ET RECOMMANDATIONS

NOTE DE SYNTHESE

Mai 2023

CHATGPT

USAGES, IMPACTS ET RECOMMANDATIONS

NOTE DE SYNTHESE

TABLE DES MATIERES

EDITO

Les 10 dernières années ont vu l'Intelligence Artificielle (IA) se déployer largement dans tous les pans de l'économie, après avoir envahi progressivement notre vie quotidienne. Mais le tempo change le **30 novembre 2022** quand OpenAI ouvre l'accès gratuit à ChatGPT. Cette ouverture déclenche un véritable tsunami : avec 1 million d'utilisateurs en 5 jours, puis 100 millions en 2 mois, ChatGPT devient l'application Web ayant connu la plus grande vitesse d'adoption par les internautes. On ne parle plus que de ChatGPT dans les entreprises et sur les réseaux sociaux. Et cet intérêt universel du grand public braque alors les projecteurs sur l'IA.

On peut ainsi d'ores et déjà anticiper l'arrivée d'une **deuxième vague de l'IA** exploitant l'**IA générative**, alors que la première vague était centrée autour des techniques d'apprentissage (*Machine Learning*) pour l'**IA prédictive**. L'**IA générative** (**IAG**), apparue il y a une moins d'une dizaine d'années[1], permet, par apprentissage sur des contenus existants, de générer de nouveaux contenus – texte, image, vidéo, etc. – pratiquement indiscernables de contenus réels. Si les briques technologiques utilisées pour produire ChatGPT ne sont pas nouvelles, l'adoption fulgurante de ce dernier constitue une révolution technologique et sociologique majeure comparable à l'arrivée d'Internet. Les **moteurs de recherche**, tels que nous les avons connus dans les 20 dernières années vont être transformés (et c'est le but fondamental de l'investissement de $10 Md de Microsoft dans Open AI), tout comme les **applications bureautiques** (dont notamment la suite Microsoft Office et son moteur de recherche Bing) ainsi que la **programmation**. Cette révolution technologique va avoir un profond impact sur nos sociétés, les emplois notamment.

Au-delà de la stupéfaction et de l'intérêt, mais aussi des critiques des utilisateurs, il est important de comprendre ce qui se joue sous nos yeux. Cette adoption pose en effet question tant dans les entreprises que dans la société civile, avec le risque que, faute de compréhension du fonctionnement de ChatGPT, cet outil soit employé à mauvais escient. Le Hub France IA a donc constitué un **groupe de travail** qui, pendant six semaines, a étudié comment ChatGPT fonctionne, quels sont ses usages, ses limites et ses impacts. Cette note de synthèse a pour ambition de restituer, de façon aussi accessible que possible, les travaux de ce groupe de travail et, enfin, de fournir des informations sur les alternatives existantes et des recommandations opérationnelles.

Depuis l'arrêt de nos travaux au 17 mars 2023, il y a chaque jour de nouvelles annonces, certaines très récentes (comme par exemple GPT-4, Google Bard, Adobe Firefly ou Microsoft Copilot 365) ne sont donc peu ou pas analysées ici. Nous espérons néanmoins que cette note de synthèse contribuera à éclairer les **enjeux de la révolution ChatGPT.**

[1] 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

INTRODUCTION

L'IAG vise à produire des contenus très réalistes. Nous nous intéresserons ici à la solution ChatGPT qui est une IAG **générant du texte à partir de texte**.

Comme pour un moteur de recherche, on interroge ChatGPT par une invite. Cette invite est plus longue qu'une requête pour un moteur de recherche : on parle de ***prompt.*** L'outil produit un texte (au lieu d'une liste de liens comme un moteur de recherche) plus ou moins long. On peut ensuite continuer la conversation en demandant des détails, en rebondissant sur une partie du texte, tout à fait comme on le ferait dans une conversation « normale » et pas du tout comme on en avait l'habitude dans une requête sur les moteurs de recherche, où les requêtes successives ne gardaient pas la mémoire des requêtes précédentes. C'est cette **mémoire de la succession des *prompts***, en particulier, qui rend la conversation aussi intéressante et surprenante. Cette aptitude à générer du texte, éventuellement selon un style spécifié, dans plus de cinquante langues et trois langages de programmation, fait potentiellement de ChatGPT un **outil utile** pour de nombreux usages. Cependant, ChatGPT **souffre de nombreuses limitations** : il peut donner des réponses plausibles mais fausses ou absurdes (on parle d'**hallucinations**), voire inappropriées ou biaisées ; et il est sensible à la façon d'exprimer la question (le *prompt*). De même, comme tous les systèmes d'IA par apprentissage à ce jour, ChatGPT n'a de connaissances que sur les données sur lesquelles il a été entraîné et n'a aucun **sens commun**. En fait, ChatGPT ne fait que fournir la réponse la plus probable au *prompt* qu'on lui propose : c'est simplement un système d'**auto-complétion** (comme on en a sur nos téléphones portables qui complètent nos débuts de mots par la suite la plus probable) sophistiqué, qui reste encore imparfait (sans logique propre ni capacité de déduction). OpenAI indique que ChatGPT est une version de recherche susceptible d'améliorations, notamment grâce aux utilisateurs dont les *prompts* et les avis servent à affiner le système.

Avant d'envisager l'utilisation de ChatGPT pour développer des cas d'usage, il convient de comprendre comment celui-ci fonctionne et quels sont les risques encourus. Les trois sections qui suivent abordent ces différents aspects.

COMMENT FONCTIONNE CHATGPT ?

Les méthodes de traitement du langage naturel (ou *Natural Language Processing*, NLP) ont largement évolué avec l'apparition des modèles de langage à base de *Transformer* basé sur le mécanisme d'attention[1]. Afin de sensibiliser les utilisateurs au bon usage à adopter tant dans la vie professionnelle que personnelle, il est essentiel de décrire le fonctionnement technique d'une IAG, telle que ChatGPT. C'est ce que propose cette section.

FONCTIONNEMENT DE CHATGPT

ChatGPT est un agent conversationnel (*chatbot*) basé sur une technologie d'IAG :

- Un **modèle de langage** (GPT-3 puis GPT-4 après le 14 mars 2023) entraîné sur un ensemble de données du Web, (extraites avant fin 2021 pour GPT-3) de près de 150 milliards de sources, soit environ 300 milliards de tokens (cf. plus bas), ou 600 GO de données. Wikipédia, qui fait partie des sources, représente seulement 0.6% de ces données d'apprentissage ;

- Un **entraînement à la conversation** et à la sécurité, réalisé par apprentissage par renforcement sur des données d'interactions entre utilisateurs, annotées manuellement. Divers modules pour garantir la sécurité (usages dangereux, malveillants, contenus haineux, etc.) sont également mis en œuvre.

Structure de ChatGPT

LE MODELE DE LANGAGE

GPT-3 est un grand modèle de langage (*Large Language Model* ou LLM) dont l'architecture repose sur un modèle appelé **Transformer**, permettant d'effectuer une analyse contextuelle approfondie de la séquence d'entrée, en utilisant une technique dite d'**attention** pour identifier les parties pertinentes de la séquence. Le Transformer se compose de deux parties principales : l'encodeur qui prend en entrée une séquence de mots (phrases ou paragraphes) et produit une représentation (***embedding***) de cette séquence et le décodeur qui réalise l'opération inverse pour générer la séquence de sortie la plus probable suivant la séquence d'entrée. La famille de modèles GPT-3 utilise la partie décodeur du *Transformer* pour générer des séquences de mots cohérentes et plausibles en se basant sur le contexte précédent. Les LLMs sont caractérisés par leur architecture complexe et leur nombre impressionnant de paramètres (175 milliards pour GPT-3).

Le **mécanisme d'attention**, introduit par Google en 2017 (note[1]), a révolutionné le traitement automatique du langage naturel. Il permet de capturer les dépendances entre les différentes parties du texte, en apprenant cette pertinence ou similarité entre les éléments. La compréhension d'un élément d'une phrase peut demander de le relier à une partie éloignée du texte. Le Transformer est capable de capturer les dépendances à longue portée des séquences d'entrée et de sortie grâce à des mécanismes d'attention. Ceux-ci permettent à ChatGPT de tenir compte d'un contexte de 4 096 *tokens* (le texte en entrée, le *prompt*, est décomposé en ***tokens*** : des unités de sens telles que des mots ou des symboles de ponctuation).

La force des LLMs réside dans leur capacité à prévoir et compléter une suite de mots en se basant sur un calcul de probabilité (le mot suivant le plus fréquent), tout en tenant compte de leur contexte. Un paramètre, la **température**, permet d'échantillonner dans les réponses possibles pour introduire de l'aléatoire (ce qui fait d'ailleurs que la même invite ne produit pas toujours la même réponse). ChatGPT génère donc du texte à partir d'un texte en entrée : c'est une IAG.

L'AGENT CONVERSATIONNEL

ChatGPT est basé sur un entraînement spécifique de GPT-3 en deux étapes :

- Un ensemble de paires « *prompt*-réponse » est produit et annoté manuellement (bonne / mauvaise réponse), puis utilisé pour un apprentissage supervisé de GPT-3 ;

- Un ensemble de quintuples (un *prompt* et quatre réponses possibles) est généré et classé manuellement pour ensuite alimenter une phase d'apprentissage par renforcement à rétroaction humaine (ou RLHF pour *Reinforcement Learning from Human Feedback*) pour apprendre à GPT-3 à générer la meilleure réponse.

Notons que, après la production de GPT-3 sur sa base d'apprentissage (**pré-2021 sur le web**), GPT-3 n'est plus connecté au web et n'est pas réentraîné régulièrement. En revanche, les interactions utilisateurs sont constamment **utilisées pour améliorer l'agent conversationnel** (ces informations ne sont donc pas protégées).

CE QUE CHATGPT PEUT ET NE PEUT PAS FAIRE POUR LE MOMENT

ChatGPT est un agent conversationnel recourant à une IA générative pour traiter du langage naturel et mener des conversations avec ses utilisateurs. L'un de ses principaux avantages est sa capacité à apprécier le contexte global de la requête et d'y répondre de façon cohérente.

Le principal apport de ChatGPT est de proposer une interface utilisateur conviviale, entraînée comme on l'a vu plus haut, pour spécialiser le modèle de langage pour un usage de type agent conversationnel.

ChatGPT ne dispose pas de **conscience** ni de faculté particulière de **raisonnement** et de **compréhension** de ce qu'on lui écrit ou de ce qu'il répond. Néanmoins, ChatGPT ne fournit généralement pas de réponses à caractère injurieux, diffamatoire ou encore pouvant porter atteinte à la morale. Pourtant, en insistant, l'utilisateur peut débloquer la génération de telles réponses, ce qui soulève la question de la capacité de ChatGPT à filtrer ses propos. Il ne se souviendra pas ultérieurement de la conversation que vous avez eue avec lui et n'apprendra rien de celle-ci, en tout cas pas automatiquement. Cependant, les conversations sont stockées et accessibles, à l'utilisateur **et à ChatGPT**.

Les réponses émises ne se fondent pas sur la vérité ou la logique mais sur la statistique. ChatGPT émet des réponses plausibles et rapides, mais **non vérifiées ni sourcées**, pouvant déboucher sur des **hallucinations**[2], c'est-à-dire simplement des éléments probables en fonction de la base d'apprentissage.

De plus, comme à ce jour, on ne connaît pas le contenu de la base d'apprentissage de GPT-3 (et encore moins de GPT-4), ni sa fréquence de mise à jour, ni les règles et valeurs (politiques, économiques, éthiques, philosophiques, etc.) utilisées par OpenAI lors de l'annotation pour l'apprentissage par renforcement, on ne peut pas considérer ChatGPT comme une source d'information **non biaisée ou neutre**.

Enfin, soulignons que ChatGPT n'est actuellement pas en mesure de fournir des **réponses "métiers"**, c'est-à-dire faisant référence à des connaissances internes à une organisation (par exemple le contenu d'une base de données). Les réponses générées sont uniquement adaptées au contexte fourni par l'utilisateur. La sortie par Open AI des API ChatGPT[3] vise à permettre

[2] https://en.wikipedia.org/wiki/Hallucination_(artificial_intelligence)
[3] https://openai.com/blog/introducing-chatgpt-and-whisper-apis

aux entreprises d'intégrer ChatGPT dans leurs applications, les utilisateurs restant propriétaires de leurs données d'entrée et de sortie des modèles. Une initiative à suivre.

LES AUTRES SYSTEMES D'IA GENERATIVE

Il existe de très nombreux LLMs et agents conversationnels associés. En particulier BARD (basé sur le LLM LaMDA) de Google est un concurrent direct de ChatGPT sur Bing (Prometheus[4]). Mais il en existe bien d'autres : LLaMA[5] (FAIR), Alpaca (Stanford), Chinchilla (DeepMind), etc. Ils diffèrent par la quantité de données d'entraînement (la conformité légale du *scrapping* (le fait d'aspirer un très grand nombre de données sur un site web) des données n'étant pas assurée) et par le nombre de paramètres des modèles, les deux pouvant varier séparément : une augmentation de la taille des données d'entraînement plutôt que du nombre de paramètres apportant semble-t-il de meilleures performances[6].

Une caractéristique importante de ChatGPT est la longueur des séquences de textes qu'il peut traiter (4 096 *tokens*, soit environ 3 000 mots anglais), lui permettant de se souvenir des échanges apparus plus tôt dans la conversation ou de générer des textes plus longs. Les modèles PaLM et LLaMa montrent de bonnes performances de raisonnement, numérique et logique, illustrées par leurs classements dans les benchmarks GSM8K, MATH et WinoGrande. Enfin, les *chatbots* LaMDA, PEER et Sparrow ou ChatGPT-Prometheus sont capables de citer les sources internet liées au texte généré. Lorsqu'il est indiqué que la source est donnée *a posteriori*, cela signifie que le chatbot génère sa réponse avant d'effectuer son travail de recherche documentaire (et non l'inverse !).

QUELLES EVOLUTIONS DES LLMS

Au cours des cinq dernières années, les LLMs ont connu un développement remarquable. Malgré leur succès, il reste un certain nombre de limitations à surmonter pour maximiser leur utilisation et minimiser les risques qu'ils peuvent présenter.

La **fiabilité** des informations produites par les LLMs et la **non protection des données utilisées** pour l'apprentissage sont des limitations majeures. Pour y remédier, des algorithmes d'explicabilité pourraient être intégrés aux interfaces pour aider l'utilisateur à évaluer la qualité des réponses. Pour renforcer la transparence, les textes générés commencent à être directement reliés aux **sources** d'informations principales par construction plutôt que via des recherches *a posteriori*. La mise à jour en temps réel de la base de connaissances et l'intégration de données expertes dans des domaines spécifiques permettraient d'obtenir des réponses justes, précises et à jour.

Les vastes quantités de données nécessaires pour entraîner ces modèles posent des problèmes de confidentialité, de biais ou de propriété intellectuelle. L'amélioration des processus de contrôle qualité sera cruciale pour l'évolution des LLMs, notamment avec l'arrivée de l'*AI Act*. Pour améliorer la modération des contenus générés, il sera nécessaire de développer des modèles

[4] https://blogs.microsoft.com/blog/2023/02/07/reinventing-search-with-a-new-ai-powered-microsoft-bing-and-edge-your-copilot-for-the-web/
[5] https://ai.facebook.com/blog/large-language-model-llama-meta-ai/, https://crfm.stanford.edu/2023/03/13/alpaca.html,
[6] https://arxiv.org/pdf/2203.15556.pdf

plus robustes à des modifications légères des *prompts* et d'améliorer la capacité à poser des questions plutôt que d'essayer de deviner la demande initiale de l'utilisateur. Les LLMs doivent également être en mesure de produire un langage plus naturel et moins verbeux pour améliorer l'expérience utilisateur

L'utilisation de LLMs en "**vertical**" avec du réapprentissage ou de la spécialisation des modèles sur des problématiques d'entreprises permettrait aux entreprises de gérer l'aspect confidentialité des données et de construire un outil spécialisé plus fiable sur leur périmètre que le généraliste ChatGPT. En particulier, le français LightOn se positionne sur ce type d'approche.

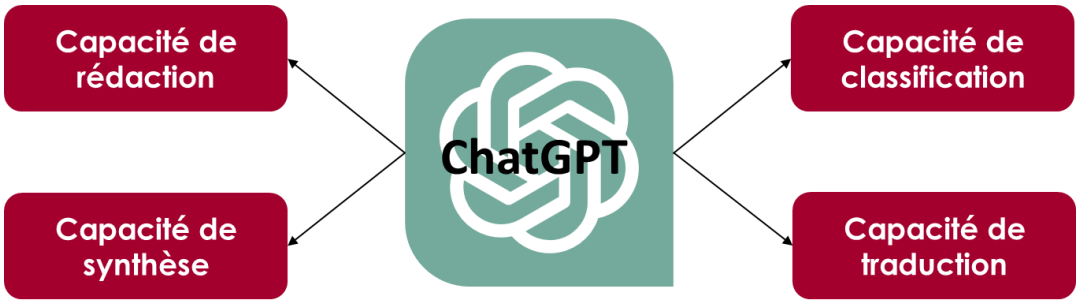
Les LLMs seront employés dans des tâches **multimodales**, impliquant l'utilisation de plusieurs types de données comme le texte, la voix, l'image ou la vidéo. Ainsi, GPT-4, dévoilé en mars 2023, est déjà capable de décrire par un texte une image qui lui est fournie dans le *prompt*.

Enfin, la consommation énergétique et les émissions de gaz à effet de serre associées à l'entraînement et à l'utilisation de ces modèles devront être optimisées. L'utilisation de LLMs pourrait paradoxalement accélérer la constitution de jeux de données spécifiques pour entraîner de plus petits modèles (*knowledge distillation*) moins énergivores.

USAGES

TYPOLOGIE DES USAGES

ChatGPT peut générer des réponses dans des styles et des tons variés et peut être utilisé pour accomplir plusieurs tâches classiques du NLP, comme la traduction, le résumé de texte, l'analyse d'opinions, la classification de texte, ou encore la génération de contenu.



Les usages de ChatGPT

Nous proposons une typologie (non exhaustive) des usages possibles de ChatGPT :

1. **Capacité de rédaction**
   - Générer de nouveaux contenus, tels que des articles de presse, une *newsletter* ou même des livres. Cela peut être utile pour les entreprises qui ont besoin de produire une grande quantité de contenus en peu de temps ou pour l'aide à la rédaction (documents marketing, réponses à des courriels, comptes-rendus, etc.). ChatGPT peut aussi être utilisé pour générer du code. Le storytelling est aussi un exemple d'usage de génération de contenu ;
   - Compléter des phrases, paragraphes, ou du code partiellement écrits ;
   - Hybrider l'expérience de recherche en proposant une expérience *Chat* aux consommateurs en complément des résultats du moteur de recherche.

2. **Capacité de classification**
   - Classer le texte en différentes catégories. Cela peut être utile pour les entreprises qui ont besoin de filtrer ou d'organiser de grandes quantités de données textuelles ;
   - Analyser les opinions d'un texte en déterminant si elles sont positives, négatives ou neutres. Cela peut être utile aux entreprises souhaitant analyser les opinions des clients à partir de leurs avis en ligne.
3. **Capacité de traduction**
   - Traduire du texte d'une langue à une autre. Cela peut être utile pour les entreprises qui ont besoin de communiquer avec des clients ou des partenaires dans plusieurs langues ;
   - Convertir du texte en parole, ce qui lui permet d'être utilisé dans une variété d'applications, telles que les assistants vocaux, le service client automatisé, etc.
4. **Capacité de synthèse**
   - Résumer de longs textes, pour accélérer leur compréhension. Cela peut être utile pour assurer des veilles informationnelles.

USAGES PAR DOMAINE

De nombreux professionnels issus de différents domaines d'activité utilisent déjà ChatGPT dans une grande variété de cas d'usage. Dans certains domaines, ces usages sont largement répandus (marketing, développement informatique, etc.). Dans d'autres, des tests et évaluations sont en cours (banque, BTP, etc.). Les paragraphes suivants synthétisent les cas d'usages identifiés par le Hub France IA au sein de 12 domaines d'activités.

De manière globale, l'utilisation de ChatGPT doit se faire avec vigilance et discernement, en vérifiant toujours les informations obtenues. En tout état de cause, il faut se souvenir que :

1. Toute information fournie dans le *prompt* est **recueillie par ChatGPT**. Il faut donc éviter de fournir dans le *prompt* des informations confidentielles et/ou à caractère personnel ;
2. ChatGPT est capable d'**halluciner**. Il faut donc toujours vérifier soigneusement l'exactitude des réponses (ce qui nécessite une connaissance du domaine) et éviter de l'utiliser pour des applications critiques.

RELATION CLIENT ET MARKETING

ChatGPT offre de nombreuses possibilités aux responsables relation client et marketeurs :

- **Génération de contenus accrocheurs** : ChatGPT peut être utilisé par les community managers pour générer du contenu marketing, comme des titres d'articles, ou identifier de nouveaux angles pour aborder des sujets récurrents ;
- **Gestion des publicités** : les marketeurs peuvent utiliser ChatGPT pour recommander des produits, accélérer les stratégies de *Search Engine Advertising*, générer des messages personnalisés pour les campagnes publicitaires ;
- **Résumé des communications** : ChatGPT génère une synthèse textuelle à partir des communications (conversations téléphoniques, courriels, etc.) entre le centre de contact et le client afin d'améliorer la qualité du service client ;
- **Chatbots** : ChatGPT peut être utilisé pour améliorer la capacité d'un *Chatbot* à comprendre et à générer des réponses plus naturelles et cohérentes en fonction du contexte fourni par l'utilisateur ;

- **Analyse des verbatims** : l'analyse des opinions de clients via ChatGPT à partir d'avis en ligne permet aux entreprises de mieux comprendre les préférences de leurs clients et d'adapter leurs produits ou services en conséquence.

DEVELOPPEMENT INFORMATIQUE

L'usage de ChatGPT le plus souvent décrit par les développeurs concerne le débogage de code. À partir d'un morceau de code fourni en *prompt*, ChatGPT peut fournir immédiatement une nouvelle version corrigée et commentée. Au contraire, le recours aux forums de développeurs nécessite du temps et peut s'avérer infructueux. ChatGPT apporte donc un gain de temps ainsi qu'un moyen de monter en compétences. Une première **limite** à cet usage est la quantité de code fournie à ChatGPT. L'extrait de code peut ne pas toujours contenir l'ensemble des informations requises pour comprendre et corriger le bug. Il est alors nécessaire d'ajouter des informations contextuelles de façon itérative pour obtenir une réponse pertinente : c'est ce qui permet à ChatGPT d'obtenir des résultats bien supérieurs aux méthodes de débogage standards. Une seconde **limite** est la **divulgation** du code dans le *prompt*, qui n'est du coup plus confidentiel.

D'autres usages concernent la génération de code à partir de langage naturel (comme une page web ou un script Python), la traduction de code informatique dans un autre langage informatique, la génération de la documentation d'un code informatique, ou encore l'automatisation des tests unitaires. Les retours d'expérience sur ces usages incluent également la nécessité de vérifier le résultat produit et d'adopter une démarche itérative pour affiner le résultat. Globalement, d'après une étude sur GitHub[7], 88% des développeurs indiquent une meilleure productivité en utilisant GitHub Copilot, une IAG basée sur Codex d'OpenAI et GPT-3.

CYBERSECURITE

ChatGPT peut aider les défenseurs à répondre aux menaces, mais il est également accessible aux attaquants. Ceux-ci peuvent utiliser les informations fournies par ChatGPT pour mieux comprendre les défenses actuelles et trouver des vulnérabilités à exploiter. Ils peuvent aussi l'utiliser pour générer du code malveillant ou pour automatiser ou optimiser certaines de leurs activités. Les risques principaux sont :

- Le renforcement des compétences des cybercriminels grâce à l'auto-formation individualisée liée à ChatGPT ;
- L'industrialisation et la meilleure qualité des activités malveillantes d'usurpation d'identité (phishing, fraude au président, etc.) grâce à la capacité de ChatGPT à rédiger des courriels convaincants, sans erreurs, dans plusieurs langues et adaptés aux différents contextes ;
- La détection des failles de sécurité dans un code en vue de les exploiter ainsi que la génération de logiciels malveillants tels que des *malwares*.

Même si OpenAI a mis en place des politiques et mesures de sécurité pour ne pas participer à des activités malveillantes, il est souvent malheureusement possible de contourner ces politiques avec un *prompt* dont la formulation est adaptée.

[7] https://github.blog/2022-09-07-research-quantifying-github-copilots-impact-on-developer-productivity-and-happiness/

BANQUE ET ASSURANCE

Le secteur bancaire est particulièrement régulé, et des interrogations subsistent quant à l'utilisation de ChatGPT et des données qui lui sont envoyées via les *prompts*. Pour assurer la protection des données, l'approche prudentielle adoptée jusqu'ici implique le blocage de l'accès à ChatGPT aux employés des établissements. Cependant, ChatGPT présente un intérêt certain pour améliorer les processus bancaires. Des démarches sont donc en cours pour tester les cas d'usages potentiels, tout en prêtant attention aux problématiques de sécurité et de protection des données des clients et des données internes. Parmi les expérimentations, nous pouvons citer :

- La synthèse d'un document afin d'en faire un résumé avec une longueur donnée ;
- La traduction d'un document ;
- La recherche d'informations sur un corpus documentaire de l'entreprise (sous forme de questions/réponses) ou corpus réglementaire ;
- La génération de code dans le cadre des développements internes d'applications (documentation et conversion de code, auto-complétion, génération de données, etc.) ;
- L'aide à la rédaction : documents marketing, documentation produit, réponses à des courriels, dossiers clients (crédit, etc.).

La réelle mise en application de ces cas d'usages dépendra de la manière dont ChatGPT sera mis à disposition des entreprises régulées.

BATIMENT ET TRAVAUX PUBLICS

Dans le domaine du BTP, un intérêt certain est porté à l'égard de ChatGPT, mais ce dernier est très peu exploité dans les entreprises sollicitées par le Hub France IA. Les utilisations pressenties porteraient principalement dans un premier temps sur l'assistance à la rédaction et la compilation des informations issues du terrain. ChatGPT est peu enclin à apporter des réponses techniques. La vigilance de son utilisation est à porter sur le contenu, notamment quant au respect des règles de l'art, des règles de sécurité et d'environnement.

RECHERCHE

Dans le domaine de la recherche, ChatGPT pourrait accélérer et augmenter la diffusion de l'information en apportant une aide bienvenue pour la rédaction d'articles scientifiques. Cependant, ChatGPT peut aussi être utilisé pour écrire de faux articles scientifiques. Ceci pourrait entraîner une hausse drastique du nombre de faux articles existants et "polluer" la littérature scientifique. C'est l'une des raisons pour lesquelles l'usage de modèles de langages pour rédiger des articles scientifiques est déjà très régulée par les éditeurs, qui, sans interdire totalement l'utilisation de ChatGPT, exigent notamment que les chercheurs fassent état précisément de l'usage qu'ils ont fait de ChatGPT.

ENSEIGNEMENT

Les cas d'usage de ChatGPT dans l'enseignement peuvent être rangés en deux catégories.

La première porte sur la capacité de ChatGPT à accélérer l'exécution de tâches que les enseignants réalisaient déjà auparavant : rédiger des courriels, réaliser des QCM, élaborer des

plans de cours détaillés avec des activités, objectifs, critères de réussite, une évaluation et une grille d'évaluation cohérentes, réaliser des cartes heuristiques (*mind-maps)* ou nuages de mots, concevoir des dictées avec des conditions spécifiques, construire des textes à trous, créer un glossaire, analyser ou résumer un texte, générer des exemples pour illustrer un concept, etc.

La seconde vise à tirer parti des spécificités de ChatGPT, tant ses forces que ses faiblesses, pour proposer de nouveaux types d'apprentissages. On trouve ainsi des cas visant à :

- Profiter de son accessibilité et de sa facilité d'utilisation pour sensibiliser les étudiants au numérique et vulgariser des concepts comme les biais algorithmiques ;
- Concevoir des exercices innovants visant à développer l'esprit critique, les savoirs des étudiants et les compétences en *prompt engineering* (vérifier les informations données par ChatGPT, générer un cours fourni par le professeur avec le moins de *prompts* possible, demander à un étudiant d'évaluer un texte de ChatGPT, etc.) ;
- Se servir de ChatGPT comme d'un tuteur fournissant des retours immédiats et personnalisés aux étudiants (corriger et reformuler leurs devoirs, améliorer leur pratique de l'anglais ou de la programmation, simuler des entretiens, etc.).

Pour encadrer ces pratiques, des initiatives de création de chartes ont été mises en œuvre.

ETHIQUE

L'utilisateur peut, avec précaution, soumettre ses dilemmes éthiques à ChatGPT pour obtenir un soutien à la réflexion en vue du "bien agir". De même, les organisations peuvent utiliser ChatGPT pour élaborer et réviser des politiques éthiques et des réglementations internes sachant que ni l'utilisateur ni l'organisation ne peuvent défausser la responsabilité de leurs actes sur ChatGPT. Attention cependant à ne pas délivrer des données confidentielles à l'IA et à rester vigilants quant aux biais de ChatGPT dans le cadre d'un tel usage. D'autre part, certaines utilisations de l'API GPT-3 posent de sérieux enjeux éthiques dont il est important de débattre avant que leur développement et leur déploiement n'augmentent. C'est par exemple le cas de l'initiative *Project December*[8], qui vise à entraîner un modèle de langage avec les données textuelles de personnes décédées pour concevoir un chatbot imitant la manière de s'exprimer de ces personnes (*Dead-bot*).

JOURNALISME

L'utilisation de l'IA et de ChatGPT dans le journalisme se développe de plus en plus. Des rédactions comme Buzzfeed l'utilisent pour cocréer des articles et des quizz afin de booster leur audience. Chez France Télévisions, ChatGPT a été utilisé pour trouver de nouveaux angles d'analyse, rédiger des courriels personnalisés pour des demandes d'entretiens auprès d'experts, ou encore synthétiser des articles scientifiques, offrant un gain de temps considérable aux journalistes. Malgré tout, l'incapacité de ChatGPT à citer ses sources et ses erreurs factuelles fréquentes limitent son déploiement dans les rédactions. Par ailleurs, ChatGPT ne peut pas remplacer le travail de recoupement des informations et d'investigation des journalistes, et il serait dangereux de se reposer **totalement** sur ChatGPT pour vérifier les sources et rédiger des articles.

[8] https://projectdecember.net/

RESSOURCES HUMAINES

Dans le domaine RH, les applications de l'IAG sont nombreuses : rédiger une offre d'emploi, préparer un entretien individuel, établir un plan d'action, répondre à des courriels, préparer des posts LinkedIn, rédiger un discours de fin d'année ou un article dans le journal d'entreprise. Dans tous ces cas, ChatGPT permet un gain de temps important, malgré des réponses standards qui nécessitent plusieurs itérations et modifications pour obtenir des résultats personnalisés qui contribuent à renforcer la culture d'entreprise. ChatGPT peut également être un outil d'aide à la rédaction des procédures internes (congés, règles salariales, représentation salariale) ainsi que des réponses circonstanciées aux demandes des employés sur ces procédures. Enfin, deux cas d'usage délicats qui demanderont un **encadrement strict** sont la gestion des conflits et l'aide au bien-être émotionnel au travail.

JURIDIQUE

Dans le domaine juridique, deux principaux cas d'usages ont été explorés.

Le premier est relatif à du conseil juridique : donner des informations précises et des références. Si l'outil permet un grand gain de temps et d'argent comparé au recours à un expert juridique, reste que ses réponses sont souvent factuellement fausses, ce qui peut avoir des conséquences critiques dans le domaine juridique. Toutefois, l'imprécision du système sera probablement corrigée au fil des mises à jour. Dans tous les cas, le conseil juridique est interdit par la loi du 31 décembre 1971, frontière devenant particulièrement floue dans le cas de l'IAG.

Le second cas consiste à analyser des contrats juridiques. Ici, la nécessaire anonymisation des données confidentielles contenues dans les contrats réduit le gain de temps. Par ailleurs, ce masquage d'informations enlève des éléments contextuels importants pour la compréhension, ce qui altère la qualité des analyses réalisées par ChatGPT. Face au rapport négatif entre le temps d'élaboration du *prompt*, la qualité de la réponse et le risque de dévoilement d'informations confidentielles, l'usage de ChatGPT pour cette tâche n'est pas forcément intéressant.

SANTE

Différents cas d'usage ont été explorés dans le domaine de la santé.

Le premier concerne l'utilisation de ChatGPT pour l'autodiagnostic et l'automédication. Ici, les réponses de ChatGPT sont limitées car celui-ci manque de données contextuelles, contrairement au médecin qui dispose du carnet de santé du patient, peut échanger avec ce dernier, observer son comportement, l'ausculter avec des instruments spécialisés, etc. Pour un usage non dangereux de ChatGPT, l'utilisateur devrait avoir conscience des limites de ChatGPT (concernant par exemple les hallucinations), mais devrait aussi avoir des connaissances en médecine pour être à la fois capable de rendre compte de sa condition avec précision à l'IA, et pour être surtout capable de détecter les fausses informations qu'elle peut potentiellement délivrer.

Par ailleurs, ChatGPT peut être utilisé pour simplifier la tenue des rapports médicaux (rédaction de synthèses, traduction en langage de tous les jours, etc.), en s'assurant toujours du respect de la confidentialité des données entrées dans le *prompt* et de l'exactitude des résultats.

PROMPT ENGINEERING ET COMMENT BIEN UTILISER CHATGPT

PROMPT ENGINEERING

Le *prompt engineering* ou ingénierie d'invite est une technique qui consiste à formuler des requêtes adaptées pour amener des modèles de langage à produire des réponses précises et pertinentes. Il peut être composé de trois éléments : i) le contexte, ii) des exemples et iii) la question ou le problème à résoudre. Cette technique permet de guider le modèle afin d'améliorer la qualité des réponses fournies aux utilisateurs. Des formations dédiées à cette technique sont désormais disponibles.

Comme on peut rapidement le constater, les réponses de ChatGPT sont très dépendantes de la question posée et de la formulation du *prompt*. Voici quelques pistes de bonnes pratiques pour formuler des *prompts* pertinents au regard du besoin :

1. **Fournir un contexte clair et adéquat** : le contexte est le principal élément de ChatGPT pour focaliser sa recherche de réponses et augmenter la pertinence des résultats. Le *prompt* doit fournir le contexte le plus précisément possible ;
2. **Posez des questions claires** : évitez les questions ambiguës ou trop larges qui pourraient donner des réponses imprécises ou inutiles ;
3. **Évitez les questions trop complexes** : ChatGPT peut encore avoir du mal à répondre à des questions très complexes ou trop techniques ;
4. **Procédez par étapes :** demandez à ChatGPT d'expliciter les étapes qui mènent à la réponse. Le résultat d'une requête est meilleur si l'on fait travailler ChatGPT "pas à pas" ;
5. **Donnez des *feedbacks*** : guidez ChatGPT afin d'affiner sa réponse ;
6. **Utilisez ChatGPT avec vigilance :** vérifiez toujours les informations obtenues, ainsi que les sources (par exemple si vous utilisez ChatGPT dans Bing).

IMPACTS

IMPACTS EN ENTREPRISE

LES RISQUES A ADOPTER CHATGPT POUR UNE ENTREPRISE

Ces risques comprennent l'**adoption incontrôlée** par le personnel pouvant entraîner la diminution de la qualité des contenus produits dans l'entreprise (internes ou externes), avec les conséquences potentiellement négatives sur le business ou sur l'image, la **non-conformité** aux normes légales et aux enjeux éthiques, le développement de biais (liés à l'entraînement de ChatGPT) ainsi que les risques cyber. En outre, il peut y avoir des **risques juridiques** liés à l'ingestion de textes protégés par des droits d'auteur (comme en témoignent quelques procès en cours aux USA) bien que ceux-ci ne soient pas encore caractérisés, ou encore une utilisation hors contrat avec OpenAI. Ces conséquences seront particulièrement importantes dans des secteurs critiques.

De plus, ChatGPT renforce les problèmes de sécurité. Les *hackers*, en exploitant les vulnérabilités de ChatGPT, peuvent causer des dégâts considérables en influençant les décisions prises par l'IA, en volant des données privées ou en provoquant des incidents délibérément.

Les entreprises peuvent notamment, afin de minimiser les risques liés à l'utilisation de ChatGPT, publier une **charte d'usage** avec des règles d'utilisation claires, former les collaborateurs sur les limites et la précision de l'outil et mettre en place des outils de détection et de supervision pour prévenir les failles de cybersécurité, corriger les erreurs, les biais ou les incohérences.

Enfin, il est nécessaire de noter que le projet de règlement européen sur l'IA (*AI Act*) pourrait prendre directement en compte les risques spécifiques posés par les systèmes d'IAG.

LES RISQUES A NE PAS ADOPTER CHATGPT POUR UNE ENTREPRISE

Cependant, ne pas adopter ChatGPT pour une entreprise, c'est potentiellement ouvrir la porte à des concurrents qui vont l'utiliser pour : proposer un service client plus rapide, réduire les coûts d'opération et de main d'œuvre, baisser les prix des produits et services. Le gain de temps qu'apporte l'IA est évidement un gain financier et un **avantage compétitif**.

IMPACTS SUR L'EMPLOI

Les analystes indiquent que l'IAG menacerait 300 Millions d'emplois dans le monde, le plus souvent par remplacement de l'humain sur des tâches spécifiques[9]. Cette étude est à prendre avec précaution car, si l'on se réfère au processus de destruction créatrice de l'économiste Joseph Schumpeter, les nouvelles innovations engendrent un déclin dans un premier temps pour les activités s'appuyant sur d'anciennes innovations, puis une phase de croissance économique avec des emplois à la clé. Une étude de l'OCDE de 2005 estime qu'un tiers des gains de productivité du travail est dû à ce processus de destruction créatrice.

Plus que des destructions d'emplois, l'IAG va remodeler les métiers en automatisant certaines tâches et créera de nouveaux emplois. ChatGPT va améliorer la productivité, en réalisant certaines tâches de manière complémentaire. Une étude récente[10] estime qu'environ 19% des emplois ont au moins 50% de leurs tâches impactées par les LLMs en considérant leurs capacités actuelles et en anticipant l'intégration de ces modèles dans des logiciels.

Il est donc important pour les entreprises d'envisager des formations aux nouvelles compétences pour les employés dont les fonctions seront rendues en tout ou partie caduques par l'adoption de l'IAG : analystes produisant des rapports, graphistes d'illustrations marketing, etc. L'augmentation de l'humain ou l'évolution des métiers du fait de ces nouvelles technologies constitue sans nul doute un défi à relever dans les prochaines années.

IMPACTS ENVIRONNEMENTAUX

La question de l'impact écologique des IAG appelle une réponse complexe, notamment en raison de la difficulté d'établir et de stabiliser des indicateurs de mesure. Si la littérature consacrée à ce sujet est déjà riche, les résultats ne sont pas consensuels à l'heure actuelle. Pour

[9] https://www.key4biz.it/wp-content/uploads/2023/03/Global-Economics-Analyst_-The-Potentially-Large-Effects-of-Artificial-Intelligence-on-Economic-Growth-Briggs_Kodnani.pdf

[10] GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models. 2023. https://arxiv.org/pdf/2303.10130.pdf

calculer les projections d'impact environnemental des IAG, il faut tenir compte du coût énergétique de leur entraînement, souvent appréhendé par la consommation des serveurs, proportionnelle au nombre de processeurs utilisés, à la localisation des serveurs et à la durée de son exécution. Plus le nombre de paramètres utilisés dans le modèle est important, plus grande sera sa consommation. L'obsolescence des processeurs (désormais destinés à cet unique objectif) qui sont d'importants consommateurs de métaux rares (pollution extractive) est également à considérer. En effet, selon les études et les variables considérées, l'entraînement d'un modèle comme GPT-3, utilisé pour faire fonctionner ChatGPT, peut représenter cinq fois l'émission carbone du cycle de vie complet d'une voiture, soit entre 284 et 552 tonnes de CO2 émises[11], selon la méthode utilisée pour l'estimation. A titre de comparaison, un aller-retour San Francisco-New York émet approximativement 1,2 t de CO2 par passager.

Toutefois, le facteur le plus déterminant reste le volume d'utilisation des IAG, puisque calculer la réponse à une requête génère une dépense de ressources : une évaluation informelle[11] pour ChatGPT serait de 0,382 grammes de CO2 émis par requête (à comparer avec l'émission de 1400 grammes de CO2 émis pour la rédaction d'une page de 250 mots aux USA). Dans le cas de ChatGPT, le volume considérable d'utilisateurs entraîne mécaniquement un impact écologique de très grande ampleur.

IMPACTS JURIDIQUES ET ETHIQUES

IMPACTS JURIDIQUES

Il est nécessaire de bien distinguer entre (i) le **système d'intelligence artificielle** (SIA) correspondant à un traitement, et un algorithme de (ii) **gestion des données**. Si le premier fait actuellement l'objet d'un projet de règlement au sein de l'Union Européenne, à savoir l'*AI Act*, la gouvernance des données est dès à présent encadrée en droit français.

L'*AI Act* pourrait inclure des obligations spécifiques concernant les agents conversationnels. La classification de l'IAG en tant que SIA à haut risque devrait toutefois être discutée et de véritables enjeux subsistent dans les questions de SIA à usage général.

L'utilisateur de ChatGPT (qu'il soit une entreprise ou un particulier) doit donc veiller à ce que la **protection de ses données à caractère personnel** soit bien respectée (notamment concernant les règles du RGPD et en cohérence avec la politique de confidentialité). Dans le cadre d'une personne morale, une attention toute particulière doit également être faite concernant des données internes. Il existe en effet de nombreuses protections concernant les **données des entreprises**, telles que la confidentialité, la sensibilité des informations, le secret des affaires, la propriété intellectuelle, *etc*. Ces données sont des **données stratégiques** (essentielles pour l'entreprise et permettant d'identifier ses orientations), voire des **données souveraines** (lorsque ces données sont nécessaires ou reflètent les décisions d'un Etat). En l'état, le fonctionnement de ChatGPT permet à OpenAI de récupérer toutes les données qui lui sont proposées pour servir à l'entraînement et à la fourniture de la réponse. La réutilisation de ces données, pourtant protégées, est réalisée pour le bon fonctionnement de l'algorithme, faisant que des acteurs

[11] https://arxiv.org/abs/2303.06219

américains peuvent accéder aux données d'entreprises européennes. Cet enjeu de **souveraineté numérique** est d'autant plus accru avec l'application de textes, tels que le *Cloud Act* ou le *FISA*.

Il existe un enjeu fort, également, en ce qui concerne les **droits de propriété intellectuelle**. L'apprentissage de l'algorithme a été réalisé à la suite d'un large *scrapping* du web, posant la question du respect du droit de collecte, des droits personnels et des protections existantes empêchant l'accès à ces informations. De plus, l'usage de l'IAG soulève de nombreuses questions concernant la parenté d'une œuvre, le respect des droits extra-patrimoniaux d'une personne physique et, plus globalement, la paternité de l'œuvre issue des algorithmes. Il faudra suivre les premiers procès[12] aux USA sur ces sujets clés pour répondre à ces questions.

IMPACTS ETHIQUES

L'un des principaux problèmes liés à l'utilisation de ChatGPT est la possibilité que les résultats fournis par l'IA puissent être discriminatoires. En effet, les modèles d'IA peuvent être biaisés en fonction des données utilisées pour les entraîner, ce qui peut impliquer une amplification des biais sociaux existants. Ces biais peuvent être involontaires et difficiles à détecter, mais ils peuvent avoir des conséquences éthiques importantes, notamment en ce qui concerne la liberté d'expression lorsqu'ils sont utilisés pour la modération de contenu en ligne, ou le recrutement et la gestion des ressources humaines.

De plus, la possibilité d'une action malveillante est une préoccupation majeure. Les utilisateurs malveillants pourraient exploiter ChatGPT pour apprendre comment commettre des crimes, diffuser de la **désinformation** ou manipuler l'opinion publique. Cette menace peut avoir des conséquences dévastatrices pour la société, en particulier dans des domaines tels que la politique, la finance ou la sécurité nationale.

Les utilisateurs peuvent ne pas comprendre qu'ils communiquent avec un modèle d'IA et non pas avec un être humain. Cela peut être particulièrement préoccupant pour les personnes vulnérables, telles que les personnes âgées ou les enfants, qui peuvent être facilement trompées, renforçant les discriminations d'âge dans la société. Ces utilisateurs peuvent ainsi être amenés à prendre des décisions ou suivre des recommandations biaisées ou mal informées.

LES TRAVAILLEURS DU CLIC A NOUVEAU EN PREMIERE LIGNE

La « réussite » de ChatGPT repose en partie sur « des travailleurs du clic[13]», qui s'attachent à le rendre moins toxique en étiquetant pour apprendre à éliminer les contenus violents, sexistes et racistes. L'objectif est de fournir à l'IA des exemples étiquetés de violence, de discours haineux et d'abus sexuels, afin qu'elle puisse apprendre à détecter ces formes de toxicité sur le web et les filtrer avant qu'elles n'atteignent l'utilisateur.

Cependant, ces travailleurs du clic sont non-seulement sous-payés (ainsi, par exemple, OpenAI a utilisé des travailleurs kenyans externalisés gagnant moins de 2 dollars de l'heure), mais sont également exposés à des contenus extrêmement traumatisants sans soutien psychologique.

[12] https://www.theverge.com/2022/11/8/23446821/microsoft-openai-github-copilot-class-action-lawsuit-ai-copyright-violation-training-data
[13] https://www.casilli.fr/tag/travail-du-clic/

Face à ce constat, l'Organisation internationale du Travail préconise de s'appuyer sur le droit du travail en vigueur dans le pays et demande également aux entreprises « d'assurer la possibilité aux travailleurs de décliner des tâches ».

IMPACTS DANS LE MILIEU DE L'ENSEIGNEMENT

STRUCTURATION DE LA CONNAISSANCE ET DES INFORMATIONS

ChatGPT est dans une logique de complétion syntaxiquement et sémantiquement valide. Ainsi, si les réponses proposées par le modèle reflètent la société réelle (presque par "hasard statistique"), elles ne le font que partiellement du fait d'un jeu de données d'entraînement nécessairement limité. ChatGPT n'a pas de "modèle du monde" ni de capacité à "comprendre". Les textes produits ne s'inscrivent donc pas dans une structure "standard" de la connaissance. Cette absence de structure limite grandement les capacités de raisonnement de ChatGPT. De plus, les productions de ChatGPT n'étant pas structurées, elles ne peuvent pas être directement utilisées par des machines. S'il est d'usage de s'assurer de la fiabilité de ses sources pour construire une connaissance robuste, ChatGPT n'en cite aucune (sauf dans la version Prometheus incluse dans Bing).

Ainsi ChatGPT exprime des connaissances, mais ce n'est pas un système de connaissance et il lui reste encore beaucoup de progrès à faire. Toutefois, les modèles de langage ouvrent une nouvelle ère, propice à l'hybridation des compétences : il devient plus aisé que jamais de suppléer ses compétences par d'autres qu'on ne maîtrise pas forcément.

EDUCATION

ChatGPT induit des évolutions dans la sphère éducative, et plus largement dans notre rapport à la connaissance, l'information et la cognition. D'abord parce qu'il facilite l'accès au savoir, ensuite parce qu'il favorise "l'*adaptive learning*" :

- En favorisant l'inclusion des élèves quelles que soient leurs particularités : à la fois tuteur intelligent permettant d'améliorer sa compréhension via un dialogue interactif, et aide à la résolution de problème, etc. ;
- En allégeant la charge administrative qui incombe souvent aux professeurs pour leur permettre d'apporter une réelle plus-value aux apprentissages : aide à la conception de séquences et séances d'apprentissage, création de parcours d'apprentissage adaptés, parcours d'apprentissage sur mesure, etc.

Les modèles de langage impacteront à plus large échelle l'ensemble des processus liés à la cognition : le "travail cognitif" des apprenants sera déplacé. Un des enjeux de la construction des connaissances portera sur l'esprit critique et les capacités de vérification des contenus (pour pallier l'absence de précision factuelle de l'IAG), et à la compléter par des idées originales (idéation que l'IAG aujourd'hui ne permet guère).

Les organismes d'éducation, confrontés à la possibilité d'une **triche** massive, vont devoir entamer une réflexion afin d'apprendre à composer intelligemment avec cette nouvelle ressource. La réflexion en est à ses débuts, à l'image de Sciences Po qui a décidé très rapidement d'interdire l'utilisation de ChatGPT avant de revenir sur cette décision, la restreignant aux examens seulement.

COMMENT DETECTER DES TEXTES ISSUS DE CHATGPT

Outre les raisons sociétales telles que la triche aux examens ou le plagiat, il est important pour les acteurs de l'industrie des LLMs d'avoir un moyen de détecter le contenu produit par ChatGPT. En effet, depuis sa en novembre 2022, les contenus générés par ChatGPT rejoignent le contenu global du Web et peuvent donc à terme être utilisés pour entraîner les LLMs, malgré les erreurs qu'ils contiennent. Il est donc crucial que les futurs modèles de langage continuent à s'entraîner sur du contenu humain et non généré.

APPROCHE HUMAINE

Sans utiliser de techniques approfondies d'analyse, il est possible grâce à une lecture attentive de relever des indices qui pourraient indiquer une origine artificielle :

1. **Longueur des phrases :** le contenu généré par une IA contient souvent des phrases courtes. L'IA essaie de reproduire l'écriture humaine, mais n'a pas encore maîtrisé la complexité des phrases étendues ;
2. **Répétition de mots :** l'IA essaie d'enchaîner des mots clés pertinents. Cela se produit souvent lorsque l'on demande à ChatGPT une définition. Donc, si vous lisez un article et que vous avez l'impression que le même mot est utilisé encore et encore, il y a de fortes chances qu'il ait été écrit par une IA ;
3. **Manque d'analyse :** si vous lisez un article avec des redéfinitions de termes connus et que vous avez l'impression qu'il s'agit simplement d'une liste de faits sans réelle perspective ou analyse, il y a des chances qu'il ait été écrit par une IA ;
4. **Inexactitudes (hallucinations) :** ChatGPT n'as pas de connaissances intrinsèques et a besoin d'un contexte pour être précis dans ce qu'il dit. Un texte généré artificiellement peut très bien contenir une affirmation, une référence, une date qui serait fausse mais qui rend le texte crédible ;
5. **Absence de fautes d'orthographe** : le contenu généré par ChatGPT ne comporte pas de fautes d'orthographe, sauf si son utilisateur demande explicitement à ChatGPT d'introduire des fautes dans la réponse générée.

Cependant, l'évolution des LLMs pourra rendre ces indices assez rapidement obsolètes.

APPROCHES STATISTIQUES

Par construction, les LLMs produisent des textes en générant des tokens qui ont une forte probabilité d'apparaître en prenant en considération les *tokens* précédents. En analysant la distribution des probabilités des mots du texte on peut mettre en évidence que le langage humain produit plus de mots qui ont moins de probabilité statistique d'apparaître dans le texte. En d'autres termes, l'humain produit des textes avec un vocabulaire plus varié que ChatGPT.

Une analyse avec l'outil GLTR[14] (*Giant Langage model Test Room*) permet d'illustrer les probabilités d'apparition des mots dans le texte et ainsi d'identifier une probabilité que le texte soit artificiel.

APPROCHE WATERMARK

Open AI a produit un outil[15] de détection directement construit par apprentissage sur des paires de documents réels et artificiels : comme l'indique Open AI l'outil a encore beaucoup de limitations. Open AI travaille également sur une technique de *watermark*[16] (un filigrane numérique) utilisant des techniques de cryptographie, qui ne semble pas encore au point. On pourrait aussi utiliser une liste de mots "interdits" qui, s'ils sont trouvés dans le texte, signaleraient l'origine humaine, du fait que les LLMs vont éviter d'utiliser ces mots.

RECOMMANDATIONS

RECOMMANDATIONS ET CADRAGE JURIDIQUE POUR LES ENTREPRISES

Pour les experts interrogés à l'occasion de l'écriture de ce rapport, il faut penser au-delà des progrès techniques qui apporteraient des bénéfices incontestables pour les entreprises, tout en permettant l'avènement d'une "IA responsable", et considérer le caractère exceptionnellement massif des usages de ces IA, ainsi que les innombrables "effets rebonds" encore mal identifiés que ces progrès techniques pourraient encourager.

L'ensemble des composantes doivent être appréhendées, depuis la disponibilité du *hardware*, jusqu'à la pertinence des usages des IA, en lien avec leur impact social notamment sur les inégalités, et leur impact juridique (respect du RGPD et de la propriété intellectuelle). Étant donné la vitesse de développement des IAG, cette évaluation est urgente, certains pays ayant déjà pris la décision de bloquer l'usage de ChatGPT (annonce[17] de l'Italie le 31 mars 2023). Enfin, de par la multiplicité des cas d'usages possibles, et leur puissance servicielle et économique, les IAG font par ailleurs peser un danger écologique sur la société.

Dans l'état actuel de la technologie, nous recommandons aux entreprises d'encadrer soigneusement l'usage de ChatGPT et des IAG, par exemple par une charte dédiée.

RECOMMANDATIONS PEDAGOGIQUES POUR LES INSTITUTIONS

Les étudiants d'aujourd'hui sont les futurs citoyens d'un monde complexe où la capacité à composer avec les outils numériques sera primordiale. En s'inspirant d'initiatives lancées par des acteurs "pilotes" familiers de l'intégration des pratiques du numérique dans leurs enseignements depuis déjà plusieurs années nous incitons à une approche consistant à adopter progressivement et de façon contrôlée ChatGPT et les IAG en classe. Toutefois, l'intégration de ChatGPT dans tous les enseignements n'est pas non plus souhaitable et ne doit pas créer un rapport de

[14] http://gltr.io/
[15] https://openai.com/blog/new-ai-classifier-for-indicating-ai-written-text
[16] https://techcrunch.com/2022/12/10/openais-attempts-to-watermark-ai-text-hit-limits/, https://scottaaronson.blog/?p=6823
[17] https://www.repubblica.it/tecnologia/2023/03/31/news/privacy_garante_blocca_chatgpt-394368983/

dépendance des étudiants et enseignants à l'outil. L'enjeu est donc de convertir les débats techniques en débats sur des programmes d'enseignement cohérents, des formations à ces outils à disposition des enseignants, et une stratégie et des lignes directrices claires sur lesquels enseignants et étudiants pourront se reposer. Le tout déployé avec transparence et communication. Cette démarche a notamment été adoptée en France par Sciences Po[18], qui a fixé des règles sur l'usage de l'IA avec une charte anti-plagiat, tout en intégrant ChatGPT dans les activités pédagogiques. En attendant des actions institutionnelles plus généralisées, **nous recommandons aux enseignants d'expliciter très clairement auprès de leurs étudiants le cadre des utilisations autorisées ou non de ChatGPT et des IAG dans leurs enseignements**[19].

COMMENT BIEN INTEGRER CES EVOLUTIONS DANS LA SOCIETE CIVILE ET LES ENTREPRISES

Certains demandent, dans une lettre ouverte, un moratoire[20] sur l'IAG, car ces outils, d'après eux, présentent des risques profonds pour nos sociétés et l'humanité. L'Italie (note[17]) a en effet bloqué l'utilisation de ChatGPT : il est toujours possible d'interdire un outil. Cependant, cela n'empêchera pas les laboratoires de recherche de continuer leurs travaux. Nous recommandons plutôt de prendre les mesures nécessaires pour intégrer ces IAG dans nos sociétés.

Nous recommandons aux pouvoirs publics de se saisir du sujet de l'implémentation de solutions fondées sur des IAG, en privilégiant les enjeux de **souveraineté** et de création de dispositifs de **financement** adaptés pour renforcer les écosystèmes français et européen. Il faut noter l'importance de **solutions open-source** (ou *a minima* ne nécessitant pas la sortie des données du territoire), accessibles aux startups et PME pour développer des **solutions verticales** (par affinage d'un modèle préentraîné sur les données d'entreprise), sans que les données des *prompts* ne sortent de l'entreprise.

Face aux évolutions apportées par ChatGPT et l'IAG, il devient urgent d'accélérer la **formation** de tous les citoyens à l'IA et aux IAG. Les impacts massifs attendus sur l'emploi, sur l'éducation, sur le monde de l'entreprise accompagneront une forte augmentation de la productivité des entreprises (+7% d'après une étude récente de Goldman Sachs[21]), mais avec une menace sur 300 millions d'emplois dans le monde. Il faut donc d'urgence préparer la société à ces évolutions.

Les LLMs sont quasi intégralement dans les mains de quelques entreprises américaines (ou chinoises). Il faut donc que l'Europe s'empare rapidement du problème et mette en place les financements, les moyens de calcul et les soutiens pour que des **solutions souveraines** soient développées **au service de l'écosystème européen**. Cela devra passer par le développement à très court terme de LLMs open source, ou *a minima* souverains, sans crainte de voir partir les données fournies dans les *prompts* vers les USA. Des solutions du type de celles annoncées

[18] https://www.sciencespo.fr/fr/actualites/sciences-po-fixe-des-regles-claires-sur-lutilisation-de-chat-gpt-par-les-etudiants
[19] Ethan et Lilach Mollick https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4391243
[20] https://www.nytimes.com/2023/03/29/technology/ai-artificial-intelligence-musk-risks.html
[21] https://www.lefigaro.fr/conjoncture/les-intelligences-artificielles-comme-chatgpt-menacent-300-millions-d-emplois-dans-le-monde-selon-goldman-sachs-20230330

récemment par Stanford [22]pourraient être mises en œuvre pour développer, à partir de solutions open source, des clones de ChatGPT à prix réduit.

L'accélération du déploiement de l'IA Générative met une forte **contrainte sur les délais** : les actions en faveur d'IAG françaises doivent se mettre en place à très court terme pour donner des résultats sous 12 mois au plus. Faute de quoi, l'économie française ne réussira pas à récolter les bénéfices de cette révolution.

CONCLUSION

La vague des IAG est lancée, il est illusoire de penser qu'elle peut être stoppée. Les IAG vont se développer et les entreprises, doivent s'y préparer pour en tirer les bénéfices, sans courir les risques que nous avons décrits au cours de ce rapport.

Les IAG ont propulsé l'intelligence artificielle hors du seul champ des spécialistes ; citoyens, journalistes, enseignants, politiques ou encore philosophes, tous sont concernés par l'essor fulgurant de cette technologie. De ce fait, il est essentiel d'acculturer à tous les niveaux pour un usage raisonné de ces outils, loin des discours alarmistes mais de façon non naïve en tenant compte de recommandations telles que celles émises dans ce rapport.

- **Mettre en place des actions de formation massive** des salariés ;
- **Ne pas interdire l'usage de l'IAG, mais mettre en place des chartes d'usage claires et opérationnelles** : interdire exclusivement les actions inacceptables (fuite de données confidentielles, triche, etc.), mais permettre les activités à valeur ajoutée en les encadrant soigneusement ;
- **Protéger les données confidentielles** et ne pas les introduire dans les *prompts* ;
- **Toujours vérifier les résultats produits par ChatGPT**, pour éviter les hallucinations ;
- **Contribuer au développement de solutions souveraines**, open-source si possible, accessibles aux petites entreprises ;
- **Travailler à ce que ces solutions permettent de développer des solutions verticales par métier**, avec, pour l'entreprise, la maîtrise de ses données (confidentielles).

[22] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, Tatsunori B. Hashimoto https://crfm.stanford.edu/2023/03/13/alpaca.html

CONTRIBUTEURS

Voici la liste des contributeurs et cosignataires, qui ont œuvré à la rédaction de cette note de synthèse :

Coordinateurs

- **Pierre MONGET** – Hub France IA
- **Françoise SOULIE** – Hub France IA
- **Marie-Aude AUFAURE** – Hub France IA
- **Camille SOUILLART** – Hub France IA

Contributeurs

- **David GRIMBERT** – Advestis
- **Nicolas MORIZET** – Advestis
- **Thomas CZERNICHOW** – Aleia
- **Marina BOJARSKI** – Atos / Paris Saclay
- **Jérôme LEBECQ** – BNP Paribas
- **Gregory COUSTOU** – ClicNwork.io
- **Julien ARTIGUE** – ClicNwork.io
- **Yixuan ZHAO** – ClicNwork.io
- **Jean-Patrice GLAFKIDES** – Datavaloris
- **Alexandre BARANOV** – DecisionBrain
- **Kajetan WOJTACKI** – DecisionBrain
- **Belkacem LAIMOUCHE** – DGAC
- **Youssef LAAROUCHI** – EDF
- **Cédric LOPEZ** – Emvista
- **Benoît BERGERET** – ESSEC Business School
- **Léo NEBEL** – EvidenceB
- **Sacha Martini** – FFB Occitanie
- **Kati BREMME** – France Télévisions
- **Yannick SANCHEZ** – France Télévisions
- **Stéphane REQUENA** – GENCI
- **Alex COMBESSIE** – Giskard.ai
- **Jean-Marie JOHN-MATHEWS** – Giskard.ai
- **David RODRIGUEZ** – ICAM / CERTOP-CNRS
- **Yann FERGUSON** – ICAM
- **Pierre PLEVEN** – Indépendant
- **Goulven PERSONNIC** – IOD Solutions
- **Bertrand LAFFORGUE** – Konverso
- **Emmanuelle BLONS** – L'atelier
- **Bertrand CASSAR** – La Poste
- **Hélène IMBERT** – Magic Lemp
- **Baptiste AVRIL** – Matrice
- **Jean CONDE** – Matrice
- **Yves COLLINET** – Micropole
- **Justine LUCAS** – Neovision
- **Miguel SOLINAS** – Neovision
- **Vincent FOUCTEAU** – Neovision
- **Malika BENAMAR** – OPPBTP
- **Loïc RAKOTOSON** – Opscidia
- **Sylvain MASSIP** – Opscidia
- **Thibaut SOUBRIE** – Preste
- **Laëtitia FAUCONNET-VIEGAS** – Probayes
- **Ronan LE HY** – Probayes
- **François-Alexandre GLAUDET** – Quavitra
- **Ivan MONNIER** – QWAM
- **Alexandre LAVALLEE** – Selas Studio
- **Hélène FLAMEIN** – SNCF
- **Audrey AGESILAS** – Société Générale
- **Benjamin BOSCH** – Société Générale
- **Nadir OUADA** – Société Générale
- **Emmanuel VIENNET** – Université Sorbonne Paris Nord
- **Karel BOURGOIS** – Voxist

CHAT GPT : USAGES, IMPACTS ET RECOMMANDATIONS

NOTE DE SYNTHESE

Mai 2023

HUB FRANCE IA