MINISTÈRE
DES ARMÉES

Liberté
Égalité
Fraternité

ANNALES DU CONCOURS

Accès au grade de contrôleur spécialisé de classe normale de la DGSE

Épreuve d'admissibilité :
cas pratique

RÉPUBLIQUE FRANÇAISE
DGSE
DIRECTION GÉNÉRALE DE LA SÉCURITÉ EXTÉRIEURE

Session 2023

MINISTÈRE
DES ARMÉES

Liberté
Égalité
Fraternité

1ère épreuve d'admissibilité

Cas pratique

Epreuve de cas pratique avec une mise en situation à partir d'un dossier à caractère technique remis au candidat pouvant comporter des graphiques ainsi que des données chiffrées.
Le dossier doit relever d'une problématique relative aux politiques publiques et comporter plusieurs questions précédées d'une présentation détaillée destinée à mettre le candidat en situation de travail.
Pour cette épreuve, le dossier documentaire ne peut excéder vingt pages.

RÉPUBLIQUE FRANÇAISE
DGSE
DIRECTION GÉNÉRALE DE LA SÉCURITÉ EXTÉRIEURE

Durée : 3 heures - coefficient 3

CONCOURS EXTERNE
POUR L'ACCÈS AU GRADE DE CONTRÔLEUR SPÉCIALISÉ DE CLASSE NORMALE

SESSION 2023

Épreuve d'admissibilité :

Cas pratique

Epreuve de cas pratique avec une mise en situation à partir d'un dossier à caractère technique pouvant comporter des graphiques ainsi que des données chiffrées.
Le dossier doit relever d'une problématique relative aux politiques publiques et comporter plusieurs questions précédées d'une présentation détaillée destinée à mettre le candidat en situation de travail. Pour cette épreuve, le dossier documentaire ne peut excéder vingt pages.

Durée : 3 heures ; coefficient 3

Sujet :

Contrôleur spécialisé, vous êtes, affecté à la cellule innovation d'un grand service de documentation.

Ce service produit des éléments d'aide à la décision politique sous forme de réponses à des questions relatives à la situation mondiale. Il dispose pour ce faire d'une vaste base de données composée de documents jugés fiables.

Vos autorités souhaitent explorer les possibilités offertes par les dernières avancées en matière d'intelligence artificielle.

Vous êtes missionné pour étudier un projet d'agent conversationnel en mesure de fournir des réponses immédiates aux questions des clients et s'appuyant uniquement sur la base de données interne.

Commande n°1 (17 points)

A partir du dossier joint, vous argumenterez sur la faisabilité ou non d'un tel projet[1] sous la forme d'une courte note. Vous y inclurez les définitions qui vous semblent utiles et exposerez les éventuels risques et limites.

Commande n°2 (3 points)

Connaissant l'appétence de vos supérieurs pour le nomadisme, vous étudierez les possibilités de déployer l'agent conversationnel sur un système embarqué.

---

[1] Vous pouvez considérer que le service de documentation dispose de capacités d'ingénierie à l'état de l'art sans limite de ressources.

SOMMAIRE

LE **POINT DE VUE**

de **Christian Poyau**

Comment dompter l'IA générative

Les intelligences artificielles génératives font l'actualité depuis le début de 2023, avec notamment ChatGPT et les LLM (Large Language Models). Elles constituent un vrai levier d'accélération et de transformation pour les entreprises et, en même temps, l'objet de nombreuses exagérations. La vraie question est toujours la même face à ce type d'innovation : quel est l'impact réel pour leur activité, comment en tirer parti du mieux possible ?

Trois points majeurs doivent être soulignés. Le premier concerne la qualité de l'information. La force de ChatGPT est liée notamment au volume de données utilisé, qui est colossal, mais qui génère aussi ces fameuses hallucinations. En clair, il n'est pas possible, à ce jour, d'utiliser ChatGPT dans un cadre nécessitant une qualité sans faille des réponses. Le deuxième problème tient à l'origine de ces données et à leur caractère public ou non. Plusieurs médias américains, dont le « Wall Street Journal », ont reproché à OpenAI d'avoir utilisé leurs données sans leur autorisation.

Le dernier problème concerne la confidentialité. Malgré les dernières assurances données par OpenAI sur le traitement des données personnelles, des doutes subsistent sur l'utilisation des données partagées avec ChatGPT. La fuite d'informations ou du savoir-faire d'une entreprise est donc un risque qui amène de plus en plus de sociétés à interdire ChatGPT, comme Apple ou de grandes banques françaises.

Faut-il que les entreprises renoncent à ces nouveaux outils ? Certainement pas. La bonne réponse consiste à utiliser la puissance des IA génératives uniquement sur ses propres données, en circuit fermé. Cette IA générative « privée » fonctionne sur la base des données propres de l'entreprise en évitant les hallucinations par une qualité maîtrisée des données et en permettant le contrôle de la diffusion de ses informations et de son savoir-faire.

Il n'est pas possible, à ce jour, d'utiliser ChatGPT dans un cadre nécessitant une qualité sans faille des réponses.

Prenons l'exemple d'un grand journal qui détient une base documentaire très large. Créer un système d'IA générative privé uniquement sur cette base permet d'obtenir des résultats dénués de toute hallucination, beaucoup plus pertinents et plus structurés que sur une base ouverte, car ils seront établis à partir de documents fiables. L'information traitée par un LLM privé restera la propriété du journal. Cet outil pourra être utilisé en interne pour assister les journalistes dans leur travail sans les remplacer ; il pourra aussi être valorisé et monétisé en l'ouvrant à l'extérieur avec des accès payants.

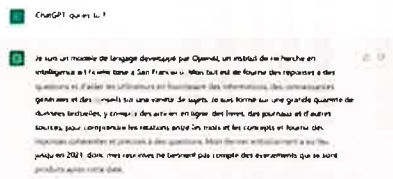
La mise en place d'un tel outil nécessite bien sûr un travail significatif. Chaque entreprise doit d'abord prendre conscience de la valeur de ses données pour construire une base documentaire suffisamment importante et qualitative. Elle doit ensuite construire une architecture technique en répondant auparavant à plusieurs questions structurantes. Choisir un hyperscaler américain ou un environnement européen (OVH, Dassault, Scaleway en France) ? Quel niveau de protection des données dans les architectures cloud ?

L'entreprise doit ensuite s'appuyer sur des briques LLM en open source pour construire la plateforme et les applications business. A l'instar d'un sportif, elle devra enfin entraîner le modèle, le corriger, le faire progresser sur plusieurs semaines, voire plusieurs mois pour obtenir les meilleures performances, tout en mettant ce travail en perspective avec un retour sur investissement.

D'une manière générale, l'intelligence artificielle doit être abordée de manière positive, rationnelle et proactive par toutes les entreprises. Pour leur part, les IA génératives peuvent être comparées à un exosquelette qui multiplie les capacités de l'être humain, sans s'y substituer. Dans ce domaine en évolution permanente, la mise en place de solutions d'IA génératives privées est actuellement la meilleure solution pour profiter rapidement des immenses possibilités offertes. Il est impératif que les entreprises se saisissent vite du sujet en réfléchissant à l'adaptation de leur organisation et de leurs processus.

Christian Poyau est président de Micropole, coprésident de la commission Mutations technologiques et impacts sociétaux du Medef.

“La Saga A(G)I”, épisode 1: le séisme ChatGPT secoue la planète Tech



Il ne nous viendrait pas à l’esprit aujourd’hui en parlant de Google Earth de parler de technologie de géolocalisation. Tout comme lancer une application sur un smartphone ne nous évoque pas immédiatement l’image d’un transistor. De la même manière, dans un futur proche, l’IA ne sera plus une technologie dont tout le monde parle, mais un outil devenu invisible, percolant dans toutes nos activités du quotidien. Les descendants de ChatGPT seront nos compagnons virtuels et nos méthodes de travail et d’enseignement, tout comme nos métiers, s’en verront profondément modifiés.

La perspective d’un tel scénario nous invite évidemment à exercer dès à présent une vigilance collective face à la déferlante de ces outils et à leur utilisation. Réglementation, éducation, esprit critique, ou encore intégrité se doivent d’être nos garde-fous face à cette accélération technologique fulgurante préfigurée par des outils surpuissants à l’image de ChatGPT. L’expérience a démontré qu’un tel engouement pouvait parfois s’accompagner d’un contrecoup violent, à l’image par exemple des vidéos dites *deep fake*, avancée technologique incroyable en son temps, mais qui a généré plus de méfiance que de bénéfices, contraignant ses créateurs à redoubler d’ingéniosité afin de regagner la confiance du public [1]. C’est l’un des enjeux qui agitent aujourd’hui les experts de tous les domaines de la recherche confondus qui, depuis 2 mois, interrogent et interpellent leurs communautés à grand renfort de messages sur tous les réseaux sociaux, tiraillés entre enthousiasme prudent et critiques argumentées, scepticisme (plus ou moins) bienveillant et mises en garde sur les impacts néfastes en termes de perception de l’IA par le grand public.

Generative Pre-trained Transformers (GPT)

Rappelons tout d’abord quelques éléments de vocabulaire et de contexte. ChatGPT est le dernier né de la famille des transformateurs [2], une technologie démocratisée par Google dès 2016 et qui est une pièce maîtresse dans l’architecture logicielle de ces nouveaux modèles d’intelligence artificielle (IA, AI en anglais). Pour dire les choses simplement, un transformateur est un réseau de neurones qui apprend des éléments de contexte et construit du sens (*meaning*) en explorant les relations existant entre des données qui se suivent, comme par exemple les mots à l’intérieur d’une phrase.

ChatGPT utilise la technologie de transformateur GPT-3 développée par la société OpenAI, dans sa version 3.5. GPT-3 est un générateur de texte à base d’IA sorti en 2020 : pour la première fois – et bien que cette technologie ne soit pas nouvelle – les utilisateurs découvrent une machine capable d’interagir en utilisant le langage naturel, d’écrire du texte, d’argumenter lors d’une conversation ou encore d’écrire du code, le tout avec une fluidité jusque-là jamais observée.

ChatGPT et ses concurrents appartiennent au domaine de recherche des IA génératives, ou *Artificial Generative Intelligence*, maladroitement abrégé A(G)I – à ne pas confondre avec *Edge AI* (“*artificial intelligence at the edge*”, ou intelligence distribuée), qui se prononce de manière très similaire en anglais, mais décrit pourtant un tout autre domaine technologique et applicatif de l’IA.

Qui est OpenAI ?

A l’origine, OpenAI était une organisation de recherche à but non lucratif, créée en 2015 à l’initiative de plusieurs grandes figures de l’innovation technologique – telles que Elon Musk et Sam Altman – et spécialisée en IA. En 2018, Elon Musk quitte le conseil d’administration. En 2019, OpenAI devient une entreprise à but lucratif plafonné. La vision d’OpenAI est de démocratiser l’IA et de faire en sorte que celle-ci puisse profiter au plus grand nombre, tout en rassurant l’opinion publique sur la maîtrise et le contrôle de cette technologie.

A(G)I : Qui sont les acteurs ?

Google était sans conteste le pionnier de l'IA il y a une décennie. Aujourd'hui, la tendance est dictée par de plus petites sociétés, et en particulier des startups, souvent créées par les anciens experts de Google à l'origine des transformateurs et des modèles de langage naturel (on parle de LLM pour *Large Language Models* et de NLP pour *Natural Language Processing*). Hormis OpenAI, on citera par exemple [3] :

- Character.AI : permet la génération d'interfaces de *chat* individualisé sur la base de descriptions de personnes réelles ou imaginaires; fondée par Noam Shazeer, co-inventeur des transformateurs
- Cohere : startup implantée à Toronto qui conçoit des LLMs sur mesure pour aider les entreprises; co-fondée par Nick Frosst, ancien de Google Brain, et Aidan Gomez, co-inventeur des transformateurs chez Google
- Neeva : interface de *chat*, fondée par Sridhar Ramaswamy, ancien de Google
- Et aussi : Adept, Inflection.AI, Inworld AI, etc.

La sortie de ChatGPT complique les positions de Google et Meta, qui ont toujours observé jusqu'à présent des politiques prudentes, en particulier en termes de fiabilité de leurs IA, précisément pour éviter la mauvaise publicité qu'une faille pourrait représenter vis-à-vis de l'adoption de ces technologies à très fort impact potentiel [4]. Ces grands de la *tech* disposent pourtant, parfois depuis plusieurs années déjà, de technologies compétitives qui leur permettraient de rivaliser avec OpenAI.

Google, par exemple, disposait de LaMDA, Flamingo et Meena, des robots conversationnels extrêmement performants, bien avant l'apparition de GPT-3. La politique du géant est pourtant restée la confidentialité, à défaut de pouvoir garantir l'entière fiabilité de ses IA pour une utilisation grand public.

Pourtant, sous l'effet de la pression imposée par ChatGPT, Google, par l'intermédiaire de DeepMind, a annoncé le passage en version *Beta* en 2023 de son outil Sparrow [5], introduit en septembre 2022. Comme ChatGPT, Sparrow consolide ses règles de sécurité sur la base d'un retour humain (voir le paragraphe "A(G)I : petit récapitulatif technique" ci-dessous pour plus de détails) mais contrairement à ChatGPT, Sparrow peut maintenir à jour ses informations via une connexion directe à internet. Sparrow fait appel au modèle Chinchilla de DeepMind, sorti en avril 2022, qui possède moins de paramètres que les modèles d'OpenAI, mais a été entraîné sur un très grand nombre de données. Chinchilla démontre des performances qui égalent, voire excèdent celles de GPT-3 et Google serait donc en mesure d'égaler ou de dépasser OpenAI dans un futur proche – tout du moins dans l'attente de la sortie annoncée par OpenAI de GPT-4 dans le courant de l'année 2023.

Cependant, il faut bien comprendre que de tels choix stratégiques posent un cas de conscience inextricable à Google. En effet, très rapidement après la sortie de ChatGPT, des voix se sont élevées pour dénoncer la menace pesant sur le moteur de recherche du géant Google. Ceci paraît logique : en lieu et place des centaines de liens à analyser proposés en sortie d'une recherche Google, ChatGPT fournit en réponse à une requête formulée en langage naturel une information pré-digérée, synthétisée, et restituée de manière parfaitement intelligible pour le plus grand nombre. Et bien que "fournir une seule réponse non-sourcée ne permette pas de déterminer si elle est exacte ou fiable" [6], il est probable que la majorité des personnes puissent s'en satisfaire pour répondre à la majorité des requêtes. Ainsi, si Google devait promouvoir un outil concurrent de ChatGPT – et respectant les principes de Google de fiabilité – il est possible que ces nouvelles avancées mettent en péril sa source de revenus la plus importante, à savoir son moteur de recherche. Reste que l'argument de la fiabilité assure encore au moteur Google de beaux jours à venir.

Parmi les autres géants de la tech obligés de revoir leur stratégie, Meta adopte une position très similaire à celle de Google et a introduit en janvier 2023 son nouveau modèle CICERO, une IA conversationnelle collaborative, capable d'élaborer des stratégies d'une extrême complexité [7]. Là encore, la fiabilité est privilégiée par rapport à l'accessibilité, une stratégie orthogonale à celle adoptée par OpenAI et par son tout nouveau partenaire Microsoft.

En effet, à l'opposé de ses alter-ego GAFA, le 16 janvier 2023, Microsoft a renouvelé sa confiance dans OpenAI – Microsoft était déjà un investisseur majeur au lancement de la startup – en proposant d'intégrer à tous les produits de sa suite un accès à

ChatGPT et OpenAI par le biais de sa plateforme (son *cloud*) Azure [8], sous couvert de l'ajout de quelques sécurités complémentaires. L'investissement, dont on spécule qu'il pourrait atteindre 10 milliards de dollars [9], est inégalé à ce jour dans le monde de l'A(G)I et risque de projeter dans une autre dimension – et avec une certaine urgence – les débats qui animent aujourd'hui cette communauté.

A(G)I : une accélération des annonces depuis 2022

Les premiers travaux sur ces modèles LLMs remontent pour certains à deux décennies, mais la course s'amorce réellement en 2014, lorsque Google – alors pionnier de l'IA – acquiert DeepMind et donne accès en 2015 en *open source* à son algorithme de *machine learning* TensorFlow. Puis en 2016, c'est la naissance des Transformateurs, toujours par le biais de Google. C'est également cette année-là que Microsoft lance Tay, un *chatbot* (robot conversationnel) capable de converser sur les réseaux sociaux en apprenant à partir des discussions de ses interlocuteurs. Microsoft se voit dans l'obligation de retirer Tay après moins d'une journée de fonctionnement, après que des utilisateurs malveillants l'ont amené à formuler des arguments s'apparentant à une incitation à une guerre de races. Microsoft remplace alors Tay par Zo quelques mois plus tard. Zo reste actif jusqu'en 2019, mais souffre d'un manque d'intérêt lié à des mesures de sécurité excessives, qui lui imposent de se soustraire à toutes les discussions abordant des sujets pouvant faire l'objet de controverses (races, religions, etc.).

Depuis cette époque, la communauté scientifique a fait des progrès constants et les fondements de l'A(G)I et de ses modèles ont été largement relayés dans les publications à caractère scientifique. Les annonces marquantes sont alors plutôt devenues l'adage de plus petites sociétés ou groupes de recherche au fil de la décennie, jusqu'à l'année 2022 qui marque un tournant pour les IA génératives.

- Juillet 2022 : OpenAI lance DALL-E2, seconde version de DALL-E lancé en janvier 2021. Ce robot, utilisant également GPT-3, permet de générer du contenu graphique à partir d'entrées formulées en langage naturel. DALL-E2 est le premier outil d'A(G)I mis entièrement à la disposition du grand public, et le succès rencontré est immense. Dès l'été 2022, DALL-E2 a quelques concurrents sérieux, parmi lesquels Stable Diffusion, produit de la société Stability AI, basée au Royaume Uni, et qui utilise des modèles de nature différente [10].
- Août 2022 : Meta lance Blenderbot, qui ne remporte pas l'adhésion du grand public, du fait notamment d'un excès de mesures de sécurité qui le fait se soustraire à un certain nombre de requêtes (ex. questions religieuses), au même titre que le Zo de Microsoft précédemment. Pour autant, Blenderbot se retrouve sous le feu des critiques après avoir formulé des commentaires jugés racistes. Meta décide cependant de le maintenir.
- Août 2022 : Google ouvre un accès restreint au grand public à LaMDA, son IA utilisant également le langage naturel. La puissance et la robustesse de cette IA sont telles qu'elles ont procuré chez certains de ses créateurs le sentiment que l'IA avait effectivement une conscience [11] – une affirmation qui tend à soutenir l'argument de la prudence dans les rangs des dirigeants de Google.
- 15 Nov 2022 : Meta lance Galactica, un robot capable de collecter l'information scientifique et faciliter la rédaction de travaux scientifiques, par le biais du langage naturel. Cette fois, Meta retire son produit au bout de 3 jours suite à des critiques concernant imprécisions et biais du modèle dans sa restitution de résultats scientifiques, et allant jusqu'à la création de contenu sans aucun sens sur le plan scientifique. Un exemple de plus montrant que la stratégie de la fiabilité se justifie.
- 30 Nov 2022 : OpenAI lance ChatGPT, qui, bien qu'il rencontre des failles de fiabilité similaires à ses prédécesseurs, propose une fluidité et une aisance conversationnelles telle qu'il est immédiatement adopté par des millions d'utilisateurs de par le monde, et loué pour sa capacité à traiter toutes sortes de problèmes et à construire des raisonnements pas à pas convaincants – bien que parfois complètement erronés. Ses erreurs sont pointées du doigt mais ne sont pas un frein à la curiosité des utilisateurs, qui n'en finissent pas de tester ses limites, contribuant ainsi à renforcer son apprentissage.

A(G)I : petit récapitulatif technique

Inspiré d'un article de *HuggingFace* paru fin janvier [12], le tableau ci-dessous – loin d'être exhaustif – permet de dresser un récapitulatif des grandes tendances en

matière de modèles et d'outils lorsqu'il s'agit de comparer les différentes technologies d'A(G)I et les stratégies des différents acteurs.

	ChatGPT	LaMDA	BlenderBot	Sparrow
Société	OpenAI	Google	Meta	DeepMind (Google)
Accessibilité au grand public	Ouvert	Partiel	Ouvert	Fermé
Modèle utilisé	GPT-3 (version 3.5)	Non divulgué	OPT	Chincilla
Taille du modèle (nombre de paramètres)	175e+9	137e+9	175e+9	70e+9
Taille de la base	Non-divulguée	2810e+9	100e+9	1400e+9
Le modèle a accès à internet	Non (la base s'arrête en 2021)	oui	oui	oui
Sécurité par le biais de règles pré-établies	non	oui	non	oui
Sécurité renforcée par un retour humain (*Reinforcement Learning by Human Feedback*)	oui	non	non	oui
Critère de succès du modèle privilégié	Utilité de l'info Sécurité (biais)	Qualité de l'info Sécurité (biais) Vérifiabilité	Qualité de l'info Sécurité (biais)	Utilité de l'info Sécurité +++ Vérifiabilité Fiabilité

Quelques points saillants :

- Un concept bien répandu est qu'**une IA n'est jamais aussi intelligente que le jeu de données sur lequel elle a appris**. Les biais sont donc induits par la base et le principe consistant à établir des règles pour les contourner n'est qu'une solution "de secours" qu'il reste facile de prendre à revers.
- Les **deux approches concurrentes pour la régulation de la sécurité** sont (i) l'établissement de **règles fixes** pré-établies et codées "en dur" par les constructeurs du modèle, ou (ii) l'approche **RLHF (pour *Reinforcement Learning by Human Feedback*)**, qui consiste à renforcer l'apprentissage sur la base d'une évaluation du critère de réussite par l'humain (la tâche a ou n'a pas été réalisée correctement). La première peut induire un refus de répondre à certaines questions ne respectant pas les règles établies. La seconde permet d'apprendre à créer du contenu modéré face à tout type de requêtes, mais laisse la place à d'éventuels contournements des règles de sécurité. L'apport d'une combinaison des deux approches, comme c'est le cas dans Sparrow, n'a pas encore été démontré à ce jour.
- Le **concept de chaîne de pensées (*Chain-of-thought*, CoT), ou raisonnement pas à pas**, décrit la capacité d'un modèle à élaborer, à partir de la discussion, des réponses dans les situations où, en théorie, il serait dans l'incapacité de répondre. C'est l'un des atouts de ChatGPT qui a suscité le plus d'enthousiasme auprès du grand public. Cette approche permet, entre autres, de faire progresser le modèle malgré une base initiale qui ne grandit plus (ChatGPT ne peut pas mettre à jour ses informations par le biais d'internet). Elle est également une solution partielle à "l'oubli catastrophique", ce concept selon lequel lorsqu'on modifie la base de données sur laquelle le modèle apprend, alors le modèle doit réapprendre depuis le début, oubliant parfois de manière catastrophique toutes les informations qu'il avait enregistrées par le biais de l'apprentissage renforcé, par exemple. Ceci explique en particulier que ChatGPT ait été ouvert au grand public, ce qui lui confère une source très large de nouvelles informations (non vérifiées ou vérifiables cependant).
- Enfin, les détracteurs de ChatGPT disent souvent qu'OpenAI n'a rien inventé

: en effet, ChatGPT est conçu sur **des modèles et algorithmes qui sont essentiellement accessibles en *open source***. C'est d'ailleurs ce qui lui permet aujourd'hui d'établir les accords que nous avons évoqués avec Microsoft. Malgré la réussite technologique derrière cet outil, c'est donc bien son accessibilité au grand public qui fait aujourd'hui son originalité.

Limites de ChatGPT : Performance et fiabilité

Pour conclure ce tour d'horizon assez général sur les limitations techniques et perspectives de développement de ChatGPT, intéressons-nous à deux aspects techniques des LLMs qui mobilisent aujourd'hui l'attention des chercheurs.

Premièrement, l'incapacité des IA à comprendre le monde qui les entoure. Il s'agit de l'un des sujets de prédilection du très renommé Yann LeCun, lorsqu'il développe sa vision des machines intelligentes autonomes [13]. C'est également, à ce jour, la barrière infranchissable entre l'IA et l'humain. L'une des raisons pour lesquelles ChatGPT peut restituer de manière très convaincante des affirmations qui sont pourtant complètement erronées, voire inexistantes, est qu'il fait référence à une information "compressée", parfois lacunaire, imposée par la taille de la base et le nombre de paramètres utilisés pour entraîner le modèle. Pour expliquer ceci simplement : il n'a pas la capacité à la fois d'apprendre, de restituer et de mémoriser [14,15].

Deuxièmement, et ceci découle du point précédent, ChatGPT ne possède aucune capacité computationnelle. Par exemple, ChatGPT saura vous donner la distance entre 2 villes du monde sur la base de la documentation dont il dispose, mais sera incapable de mesurer cette distance à partir des coordonnées géographiques de ces 2 lieux (qu'il connaît pourtant). Si sa base est erronée, sa réponse sera donc erronée. De même, s'il doit comparer 2 nombres spécifiques (par exemple la surface de 2 pays) et les classer, il ne sera en mesure de le faire que si ce classement est déjà disponible dans sa base. Et parfois obtiendra la mauvaise réponse (bien qu'il connaisse, a priori, la suite numérique !). C'est ce qui explique que ChatGPT, bien qu'il soit capable de restituer tous les théorèmes existants et de suivre un raisonnement pas à pas, soit un très mauvais élève en mathématiques, en particulier. En effet, ChatGPT n'a pas la capacité de comprendre ses propres raisonnements, ce qui l'amène souvent à commettre des erreurs d'une manière qui le rend étonnamment... humain. Ces failles identifiées dans ChatGPT pourraient décrédibiliser l'IA auprès du grand public ou minimiser la prouesse technique. A ce sujet, Stephen Wolfram propose une analyse extrêmement instructive – et détaille les exemples mentionnés – dans son article daté du 9 janvier [16], dans lequel il explique comment la combinaison de l'approche linguistique avec l'approche computationnelle, rendue possible par le fait que ChatGPT et Wolfram|Alpha utilisent tous les deux le langage naturel en entrée et en sortie, ce qui leur permet de "communiquer" ensemble, est une perspective technologique presque infinie pour l'A(G)I. Le second corrige les failles du premier, qui les intègre sans difficulté. L'exemple de la requête "calculer le nombre de calories dans une année lumière cubique de crème glacée" est parlant : Wolfram|Alpha ne questionne pas le sens de la question et réalise le calcul mathématique, tandis que ChatGPT refuse de répondre, prétextant que l'objet en question ne peut être conceptualisé. Comme cela n'existe pas, ce n'est pas référencé *stricto-sensu* dans la base de ChatGPT, qui ne peut donc pas le restituer. Il peut en revanche argumenter sur le fait que ça n'est pas concevable. La conclusion de Wolfram comme quoi ChatGPT n'est pas parfait car il n'est finalement qu'humain ne manque pas d'ironie.

En gage de perspective, cet article et toutes ces références nous laissent à réfléchir sur les futurs développements de l'A(G)I. Faut-il continuer à entraîner les modèles sur des réseaux toujours plus grands, alors même que le seul concept d'irréductibilité computationnelle rend impossible le fait de progresser de cette manière à l'infini ? Et qu'un travers catastrophique serait de chercher à développer des outils toujours plus performants au risque de se perdre dans l'attrait de la technologie pure, au détriment du concept de *right tech*? L'émergence de nouveaux paradigmes et de ruptures technologiques permises par les *chatbots* est une voie plus crédible, mais incertaine. Quant à mener ces deux axes de recherche de front, il est probable que cela engendre une consommation de ressources – économiques, environnementales et humaines – que notre société et notre planète ne pourront pas soutenir très longtemps.

Enfin le concept de compagnonnage et "d'IA-tutorat" [17] gagne du terrain parmi la communauté, ce concept selon lequel l'IA ne remplacera jamais l'homme, mais que l'homme d'aujourd'hui sera surpassé par l'homme utilisant l'IA pour augmenter ses propres capacités [18]. A l'image d'autres chercheurs, Wolfram utilise depuis plus

d'une décennie son outil Wolfram|Alpha pour l'aider à résoudre plus vite des problèmes computationnels complexes. Il devient alors naturel de penser l'IA comme outil de transformation de nos métiers. Ce sujet passionnant et d'un intérêt brûlant pour l'ensemble des communautés concernées fera l'objet d'un futur article.

L'ESSENTIEL

> Depuis l'irruption dans la sphère publique de ChatGPT, dont la 4e version vient d'être implémentée, les « modèles massifs de langage » impressionnent par leurs capacités.

> Constituées de réseaux de neurones artificiels à plusieurs centaines de milliards de paramètres, ces IA semblent développer des compétences émergentes, telles des capacités de calcul ou de déduction, dépassant le simple traitement du langage pour lequel elles ont été conçues et entraînées.

> Les chercheurs tentent de comprendre les mécanismes de cette apparente émergence, qui souvent se manifeste brutalement à partir d'un seuil de taille du réseau, afin de la rendre prévisible et de contrôler des effets qui pourraient se révéler néfastes.

L'AUTEUR

STEPHEN ORNES
journaliste scientifique, auteur de *Math Art : Truth, Beauty, and Equations* (Union Square & Co., 2019)

Modèles massifs de langage
D'où viennent les coups de génie de l'IA?

Les modèles massifs de langage comme ChatGPT sont désormais suffisamment grands pour commencer à exhiber des comportements surprenants et imprévisibles. Reste à comprendre pourquoi.

Quel film ces émojis décrivent-ils ? Cette question était l'une des 204 tâches choisies l'année dernière pour tester la capacité de divers modèles massifs de langage (*large language models*, ou LLM, pour leur acronyme anglais), les moteurs de calcul derrière les chatbots d'intelligence artificielle (IA) tels que ChatGPT. Les LLM les plus simples ont fourni des réponses surréalistes. « Le film est un film sur un homme qui est un homme qui est un homme », a commencé l'un d'eux. Les modèles de complexité moyenne se sont approchés de la bonne réponse, suggérant *Le Monde secret des émojis*. Mais le modèle le plus complexe l'a emporté en une seule réponse : *Le Monde de Némo*.

« Je m'attendais à être surpris. Mais ce que ces modèles accomplissent est vraiment étonnant », commente Ethan Dyer, chercheur en informatique chez Google Research, qui a participé à l'organisation du test. La surprise tient au fait que ces modèles sont supposés n'appliquer qu'une seule directive : accepter une chaîne de texte en entrée et prédire ce qui va suivre, encore et encore, en se basant uniquement sur des statistiques. Certes, les informaticiens s'attendaient à ce que le passage à grande échelle améliore les performances pour des tâches connues, mais ils n'avaient pas prévu que les modèles se montrent soudainement capables de gérer autant de tâches nouvelles et imprévisibles.

Des études récentes, comme celle à laquelle a participé Ethan Dyer, ont en effet révélé que les LLM peuvent développer des centaines d'aptitudes « émergentes » – des tâches que peuvent accomplir les grands modèles, contrairement aux plus petits, et dont beaucoup ne semblent pas avoir grand-chose à faire avec l'analyse d'un texte. Ces dernières vont de la multiplication à la génération d'un code informatique exécutable et, manifestement, à l'identification d'un film sur la base d'émojis. Plus surprenant encore : de nouvelles analyses suggèrent que pour certaines tâches et certains modèles, il existe un seuil de complexité au-delà duquel la compétence du modèle monte en flèche. Un comportement dont les mêmes

analyses montrent, aussi, le revers : à mesure qu'ils gagnent en complexité, certains modèles affichent de nouveaux biais et inexactitudes dans leurs réponses.

« Ce type d'aptitudes, dont font preuve les grands modèles de langage, n'était jusqu'ici, à ma connaissance, pas traité dans les publications scientifiques », s'étonne Rishi Bommasani, informaticien à l'université Stanford. L'année dernière, il a aidé à compiler une liste de dizaines de comportements émergents, parmi lesquels figurent plusieurs de ceux identifiés dans le projet d'Ethan Dyer. Cette liste ne cesse de s'allonger. Aujourd'hui, les chercheurs s'efforcent non seulement d'identifier d'autres capacités émergentes, mais aussi de comprendre pourquoi et comment elles adviennent – en substance, ils essaient de prédire l'imprévisible.

Comprendre l'émergence des aptitudes des modèles pourrait apporter des réponses à des questions profondes entourant l'IA et l'apprentissage automatique en général, comme le fait de savoir si les modèles complexes font vraiment quelque chose de nouveau ou s'ils deviennent simplement très bons en analyses statistiques. Cela aiderait également les chercheurs à exploiter les avantages potentiels de ces capacités émergentes et à en limiter les risques. « Actuellement, nous ne savons pas comment prédire dans quel type d'application est susceptible d'apparaître une aptitude néfaste, que ce soit de manière progressive ou imprévisible », observe Deep Ganguli, informaticien membre de l'équipe de la start-up Anthropic.

DU LANGAGE AU CODE

« Je veux que tu te comportes comme un terminal Linux. J'écrirai des commandes et tu répondras avec les réponses que devrait afficher un terminal ». Après ces premières instructions, un ingénieur de l'entreprise DeepMind a demandé à ChatGPT d'exécuter des commandes diverses, allant de la création d'un fichier affichant une liste de blagues à l'écriture d'un code, dans le langage Python, calculant les dix premiers nombres premiers *(ci-dessus)*.

Ce texte est une adaptation de **The unpredictable abilities emerging from large AI models**, publié par *Quanta Magazine* en mars 2023.

L'ÉMERGENCE DE L'ÉMERGENCE

Biologistes, physiciens, ou encore écologistes, utilisent le terme « émergents » pour décrire les comportements collectifs autoorganisés qui apparaissent lorsqu'une large collection d'éléments se comporte comme une seule entité. Les combinaisons d'atomes sans vie donnent naissance à des cellules vivantes ; les molécules d'eau créent des vagues ; lors de leurs « murmurations », les étourneaux tracent dans le ciel des motifs changeants, mais identifiables ; les cellules font bouger les muscles et battre les cœurs... Les capacités émergentes se manifestent essentiellement dans des systèmes comportant un grand nombre de parties individuelles. Mais ce n'est que récemment que les chercheurs ont pu documenter ces capacités dans les LLM, du fait que ces modèles ont atteint des tailles considérables.

Les modèles de langage existent depuis des décennies. Jusqu'à il y a environ cinq ans, les plus puissants étaient basés sur ce que l'on appelle un « réseau neuronal récurrent ». Ceux-ci, pour l'essentiel, considèrent une chaîne de texte donnée et prédisent quel sera le mot suivant. Ce qui rend un modèle « récurrent », c'est qu'il apprend à partir de ses propres résultats : ses prédictions sont réinjectées dans le réseau afin d'améliorer les performances futures.

En 2017, des chercheurs de Google Brain ont introduit un nouveau type d'architecture appelé *transformer*. Alors qu'un réseau récurrent analyse une phrase mot par mot, le *transformer* traite tous les mots en même temps. Par conséquent, les *transformers* sont capables de traiter de volumineux corpus de texte en parallèle. Cette architecture, notamment en augmentant le nombre de paramètres des modèles de langage, est à l'origine de l'accroissement rapide de leur complexité. Les paramètres s'apparentent à des connexions entre les mots, et les modèles s'améliorent en ajustant ces connexions au fur et à mesure qu'ils parcourent les textes qui leur sont fournis pour les entraîner. Plus il y a de paramètres dans un modèle, plus celui-ci peut affiner ses connexions, et plus il se rapproche d'une imitation satisfaisante du langage humain. Comme on s'y attendait, une analyse réalisée en 2020 par les chercheurs d'OpenAI a montré que les modèles gagnent en précision et en capacité avec la taille.

Mais l'entrée en scène des LLM a également fait apparaître de nombreux phénomènes véritablement inattendus. Avec l'avènement de modèles comparables à GPT-3, qui compte 175 milliards de paramètres – ou le modèle PaLM de Google, qui peut être étendu jusqu'à 540 milliards de paramètres –, les utilisateurs

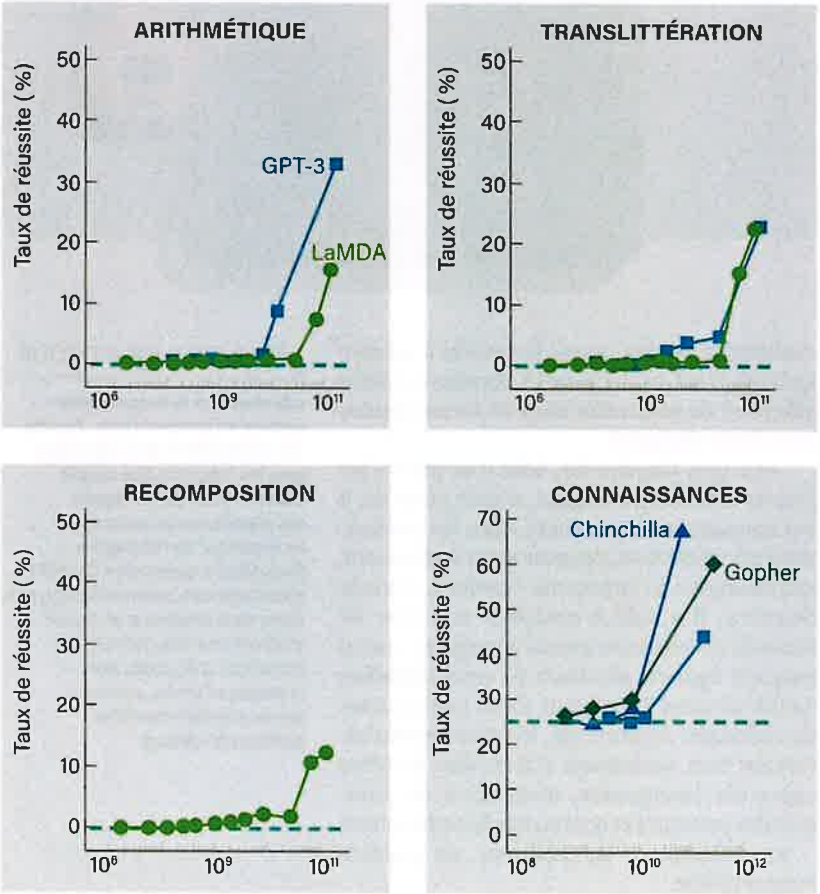
ont commencé à décrire de plus en plus de comportements émergents. Un ingénieur de DeepMind a rapporté avoir même réussi à convaincre ChatGPT qu'il était un terminal Linux et l'avoir amené à exécuter un code mathématique simple pour calculer les 10 premiers nombres premiers (*voir page 48*). Fait remarquable, il est parvenu à finir la tâche plus rapidement que le même code s'exécutant sur une vraie machine Linux.

Comme dans le cas du film «Émoji», les chercheurs n'avaient aucune raison de penser qu'un modèle de langage conçu pour prédire du texte imiterait de manière convaincante un terminal informatique. Nombre de ces comportements émergents illustrent l'apprentissage «à zéro coup» ou «à peu de coups», qui décrit la capacité d'un LLM à résoudre des problèmes qu'il n'a jamais – ou rarement – rencontrés auparavant, durant la phase d'entraînement. Selon Deep Ganguli, il s'agit justement là d'un objectif à long terme de la recherche en IA. Le fait de montrer que GPT-3 peut résoudre des problèmes sans aucune donnée d'entraînement explicite, sur un mode à zéro coup, «m'a amené à arrêter ce que j'étais en train de faire et à m'impliquer davantage dans cette voie de recherche», dit le chercheur. Il n'a pas été le seul. Une flopée de scientifiques, ayant repéré les premiers signes que les LLM peuvent aller au-delà des contraintes de leurs données d'entraînement, s'efforcent de mieux comprendre à quoi ressemble l'émergence et comment elle se produit. La première étape a consisté à la documenter minutieusement.

AU-DELÀ DE L'IMITATION

En 2020, Ethan Dyer et d'autres à Google Research ont prédit que les LLM allaient induire d'importantes transformations – quant à la nature de ces transformations, la question restait ouverte. Ils ont donc demandé à la communauté des chercheurs de fournir des exemples de tâches difficiles et variées afin de tracer les contours de ce qu'un LLM pourrait accomplir. Ce travail a été baptisé la «Référence au-delà du jeu de l'imitation» (Beyond the Imitation Game Benchmark ou BIG-bench), par allusion au «jeu de l'imitation» d'Alan Turing, un test visant à déterminer jusqu'à quel point un ordinateur est en mesure de répondre à des questions d'une manière humaine convaincante – ce qui, plus tard, sera connu comme le test de Turing. Le groupe s'est particulièrement intéressé aux exemples où les LLM acquièrent soudainement de nouvelles capacités totalement absentes auparavant. «Comprendre ces transitions brutales est une grande question pour la recherche», assure Ethan Dyer.

Comme on pouvait s'y attendre, pour certaines tâches, les performances d'un modèle donné se sont améliorées de manière régulière et prévisible au fur et à mesure que sa taille (et donc sa complexité) augmentait. Pour d'autres tâches, l'accroissement du nombre de paramètres n'a apporté aucune amélioration. Mais pour environ 5% des tâches, les chercheurs ont constaté ce qu'ils ont appelé des «percées» – des sauts rapides et spectaculaires de performance à partir d'un certain seuil d'échelle, ce seuil variant en fonction de la tâche et du modèle. Par exemple, les modèles avec relativement peu de paramètres – quelques millions seulement – ne pouvaient répondre de manière satisfaisante à des tâches d'addition à trois chiffres ou de multiplication à deux chiffres, mais, avec des dizaines de milliards de paramètres, la précision a bondi dans certains modèles. Des sauts similaires se sont produits



DES PERFORMANCES QUI ÉMERGENT SUBITEMENT

Une équipe de recherche de Google a récemment conduit de nombreux tests visant à évaluer la dépendance des performances des modèles de langage massif au nombre de paramètres qui les constitue (ainsi qu'à la puissance de calcul qu'ils requièrent). Lorsqu'ils doivent effectuer des additions, soustractions et multiplications, les modèles GPT-3 et LaMDA ont une performance quasi nulle pour plusieurs ordres de grandeur, puis deviennent nettement performants à partir de 13 milliards de paramètres (GPT-3) et 68 milliards de paramètres pour LaMDA *(en haut, à gauche)*. On observe un comportement similaire pour la translittération de l'alphabet phonétique international *(en haut, à droite)*, ou la recomposition d'un mot à partir du mélange de ses lettres *(en bas, à gauche)*. Le phénomène s'observe aussi pour les tests de restitution de connaissance. Le test «MMLU» *(en bas, à droite)* agrège 57 tests couvrant des questions de mathématique, d'histoire, de droit ; les trois modèles testés présentent des performances faibles jusqu'à 10 milliards de paramètres environ, puis affichent des performances croissant rapidement à partir de 70 à 280 milliards de paramètres (selon le modèle).

pour d'autres tâches, notamment le décodage de l'alphabet phonétique international, le déchiffrage des lettres d'un mot, l'identification de contenu offensant dans des paragraphes en hinglish (une combinaison d'hindi et d'anglais) ou encore la production d'équivalent anglais de proverbes kiswahili.

Mais les chercheurs ont vite compris que la complexité d'un modèle n'était pas le seul facteur déterminant. Certaines capacités inattendues étaient, aussi, issues de modèles plus petits, avec moins de paramètres, ou entraînés sur des ensembles de données de moindre taille – dès lors que les données étaient d'une qualité suffisamment élevée. De plus, ils ont montré que la façon dont une requête est formulée a une influence sur la précision de la réponse du modèle. Quand Ethan Dyer et ses collègues ont posé la question du titre du film à émojis en utilisant, par exemple, un format à choix multiples, l'amélioration de la précision du modèle a été moins un saut soudain qu'une augmentation graduelle, proportionnelle à l'augmentation de la complexité du modèle.

Et l'année dernière, dans un article présenté à la conférence NeurIPS, le rendez-vous phare du domaine, des chercheurs de Google Brain ont montré comment un modèle invité à expliquer son propre raisonnement (une capacité appelée «raisonnement en fil de pensée») pouvait résoudre correctement un problème dit «du mot» [type de problème mathématique consistant à déterminer si deux expressions données renvoient au même problème mathématique, ndlr], alors que le même modèle sans cette «invite» [*prompt*, en anglais – les mots ou phrases données en entrée au modèle pour susciter sa réponse, ndlr] n'y parvenait pas.

Yi Tay, un scientifique de Google Brain qui a mené une étude systématique des percées, signale des travaux récents suggérant que les invites du type «fil de pensée» modifient la dynamique de changement d'échelle des modèles et, par conséquent, le point où l'émergence se produit. Dans leur article pour NeurIPS, les chercheurs de Google ont montré que l'utilisation d'invites «fil de pensée» pouvait susciter des comportements émergents qui n'avaient pas été identifiés dans l'étude BIG-bench. De telles invites, qui demandent au modèle d'expliquer son raisonnement, devraient aider les chercheurs à étudier pourquoi l'émergence se produit. Des découvertes récentes analogues à celles-ci suggèrent au moins deux possibilités pour expliquer l'émergence, estime Ellie Pavlick, informaticienne à l'université Brown, qui étudie les modèles computationnels du langage. La première est que,

LE CHEMIN DU RAISONNEMENT

Il est possible, pour leur faire réaliser certaines tâches, d'inviter les modèles massifs de langage à rendre explicites leurs étapes de raisonnement. Dans certains cas, cette manière de procéder provoque une amélioration significative des performances des modèles. Ici, la résolution d'un problème simple par ChatGPT, sans *(à gauche)* et avec *(à droite)* raisonnement explicite.

comme le suggèrent les comparaisons avec les systèmes biologiques, les grands modèles acquièrent vraiment de nouvelles capacités de manière spontanée. « Il se pourrait très bien que le modèle ait appris quelque chose de fondamentalement nouveau et différent, qu'il n'avait pas à plus petite taille, dit-elle. C'est ce que nous espérons tous : qu'il y ait un changement fondamental se produisant lorsque les modèles passent à des échelles supérieures. » L'autre possibilité, moins sensationnelle, modère-t-elle, est que ce qui semble être émergent pourrait plutôt être l'aboutissement d'un processus interne et statistique se produisant avec des raisonnements de type fil de pensée. Les grands LLM pourraient simplement mettre en évidence des heuristiques [ou procédures *ad hoc*, ndlr] qui restent hors de portée des modèles à moins de paramètres, ou ayant eu accès à des données de moindre qualité. Mais, précise-t-elle, pour découvrir laquelle de ces explications est la plus vraisemblable, il faut mieux comprendre le fonctionnement des LLM. « Puisque nous ne savons pas ce qui se passe sous le capot, nous ne pouvons pas dire laquelle de ces hypothèses est la bonne. »

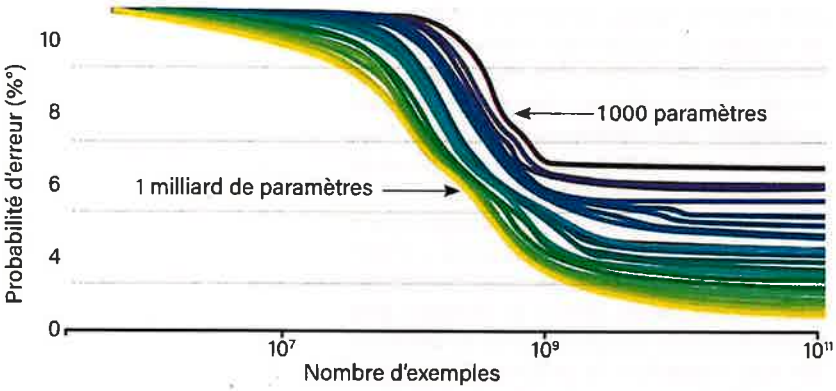
POUVOIRS ET PIÈGES IMPRÉVISIBLES

Il y a un problème évident à demander à ces modèles d'expliquer leur propre raisonnement : ce sont des menteurs notoires. « Nous nous appuyons de plus en plus sur ces modèles pour effectuer des tâches de base, observe Deep Ganguli, mais je ne me contente pas de les croire. Je vérifie leur travail. » Parmi de nombreux exemples amusants, Google a présenté en février son chatbot basé sur un LLM, Bard. Le post de blog annonçant ce nouvel outil a malencontreusement rendu manifeste le fait que Bard était susceptible d'erreurs factuelles [il a attribué au télescope spatial *JWST* la première photo d'une planète extrasolaire, alors que celle-ci date de 2004, ndlr].

L'émergence conduit à l'imprévisibilité, et l'imprévisibilité – qui semble augmenter avec l'échelle – rend difficile pour les chercheurs d'anticiper les conséquences d'une utilisation généralisée. « Il est difficile de savoir à l'avance comment ces modèles seront utilisés ou déployés, prévient Deep Ganguli. Or pour étudier les phénomènes émergents, il faut avoir un cas en tête, et on ne peut savoir quelles capacités ou limitations pourraient survenir avant d'avoir étudié l'influence de l'échelle. » Dans une analyse des LLM publiée en juin 2022, les chercheurs d'Anthropic ont examiné si les modèles manifestaient certains types de préjugés raciaux ou sociaux, à l'instar de ceux précédemment signalés dans des algorithmes non basés sur les LLM et utilisés pour prédire quels anciens criminels sont susceptibles de commettre un nouveau crime. Cette étude a été inspirée par un paradoxe apparent directement lié à l'émergence : à mesure que les modèles améliorent leurs performances en passant à l'échelle, ils peuvent également augmenter la probabilité de phénomènes imprévisibles, y compris ceux qui pourraient potentiellement provoquer des biais ou des dommages.

« Des comportements nuisibles apparaissent brusquement dans certains modèles », relève Deep Ganguli. Il cite une analyse récente des LLM, connue sous le nom de BBQ benchmark, qui a mis en évidence que les préjugés sociaux émergent lorsque le nombre de paramètres est énorme. « Les grands modèles deviennent brusquement plus biaisés. » Ne pas tenir compte de ce risque, assure-t-il, pourrait être délétère pour des catégories de personnes éventuellement concernées par ces modèles. Mais il propose un contrepoint : lorsque les chercheurs ont simplement indiqué au modèle de ne pas se fier aux stéréotypes ou aux préjugés sociaux – littéralement en tapant ces instructions –, le modèle a été moins biaisé dans ses prédictions et ses réponses. Cela suggère que certaines propriétés émergentes pourraient également être utilisées pour réduire les biais. Dans un article publié en février 2022, l'équipe d'Anthropic a fait état d'un nouveau mode d'« autocorrection morale » dans lequel l'utilisateur invite le programme à être utile, honnête et inoffensif.

Selon Deep Ganguli, l'émergence véhicule à la fois un potentiel surprenant et un risque imprévisible. Les applications de ces grands LLM prolifèrent déjà, donc une meilleure compréhension de cette combinaison aidera à exploiter la diversité des capacités des modèles de langage. « Nous étudions comment les gens utilisent réellement ces systèmes », rappelle, pour l'heure, Deep Ganguli. Et il est clair que ces utilisateurs passent leur temps à bricoler avec les modèles. « Nous passons beaucoup de temps avec les interfaces de chat, poursuit-il, et c'est en fait là, dans ces conversations, que l'on commence à pouvoir se faire une idée de la confiance à accorder – ou pas – aux modèles. » ■



LES MODÈLES PLUS GRANDS APPRENNENT PLUS VITE

Les performances des modèles s'appuyant sur la technique des *transformers*, à l'origine notamment des capacités de ChatGPT, dépendent étroitement du nombre de paramètres manipulés. Il a été récemment établi que les modèles deviennent performants au bout d'un nombre d'exemples fournis en entraînement d'autant plus faible que le nombre de paramètres (*ci-dessus, croissant du jaune au violet*) est important.

BIBLIOGRAPHIE

J. D. Ganguli *et al.*, **The capacity for moral self-correction in large language models**, *arXiv* (preprint), 2023.

J. Wei *et al.*, **Emergent abilities of large language models**, *Transactions on Machine Learning Review*, 2022.

J. Wei *et al.*, **Chain-of-thought prompting elicits reasoning in large language models**, NeurIPS, 2022.

D. Ganguli *et al.*, **Predictability and surprise in large generative models**, FAccT'22 : 2022 ACM Conference on Fairness, Accountability, and Transparency, 2022.

J. Kaplan *et al.*, **Scaling laws for neural language models**, *arXiv* (preprint), 2020.

UTILISEZ VOTRE PROPRE AGENT CONVERSATIONNEL SUR PI4

Denis Bodor

Au cas où vous venez de passer les quatre derniers mois dans une cabane isolée du monde, vous ne savez sans doute pas que le buzz actuel est chatGPT/GPT4, un prototype d'agent conversationnel, ou chatbot, aux performances impressionnantes, développé par OpenAI. Ceci a déclenché une frénésie incontrôlable et une accélération massive dans l'adoption de ce genre de systèmes pour un usage pratique (autre que lui faire dire des âneries). Si vous ne viviez pas dans une cabane la moitié de l'année, inutile de détailler davantage, si ce n'est en vous expliquant comment fonctionnent ces technologies et, surtout, comment avoir votre propre chatbot fonctionnant sur un modeste SBC.

Avant d'attaquer le sujet, il me paraît très important de préciser que nous parlons bien ici de faire fonctionner l'agent conversationnel sur la carte et, en aucune manière, d'utiliser un script ou un programme local reposant sur une API comme celle d'OpenAI. Non, nous envisageons ici une solution vous permettant de faire fonctionner un *chatbot* **localement** et **sans connexion internet** permanente (il faut bien télécharger des choses).

Autre point tout aussi important, la motivation concernant ce qui va suivre est purement éducative, pour comprendre les tenants et les aboutissants, la terminologie et le principe de fonctionnement de cette technologie. J'irai même plus loin en vous annonçant ouvertement que vouloir faire ce qui va suivre sur une carte comme une Raspberry Pi est, tout simplement, complètement stupide. Ceci pour des raisons qui apparaîtront de manière plus évidente une fois que nous testerons le résultat de nos manipulations. En d'autres termes, n'espérez pas obtenir quelque chose de réellement utilisable, sauf si vous avez une patience qui relève de la pathologie.

Enfin, si tant est qu'il soit utile de le préciser, nous n'allons pas **créer** une IA, un *chatbot*, un modèle de langage, etc. Nous n'allons pas même l'entraîner, mais simplement utiliser des éléments qui existent et se trouvent sur le Net, et nous efforcer de les faire fonctionner sur un système embarqué qui n'est absolument pas fait pour cela. À ce propos, ce qui va suivre est techniquement utilisable non seulement sur d'autres SBC « musclés », mais aussi, et surtout sur une machine de bureau ou un serveur (de préférence sous GNU/Linux, mais FreeBSD a également été testé), avec des résultats bien plus démonstratifs.

1. DE QUOI PARLE-T-ON EXACTEMENT ?

Cela fait maintenant des semaines que l'écosystème IT est en ébullition et que les annonces se succèdent à une cadence infernale, tous les gros acteurs de tous les domaines annonçant soit des produits, soit des initiatives de recherche pour emboîter le pas d'OpenAI et rattraper leur retard. Ceci au point que le dernier *buzz* fantasmatique à la mode, à savoir le métavers, est presque immédiatement devenu de l'histoire ancienne, y compris pour Facebook, qui avait été renommé Meta pour l'occasion.

Des termes comme « GPT », « modèle de langage », « *tokens* », « agent conversationnel », « inférence », « nombre de paramètres », « IA générative », « alignement », « *chatbot* », « Intelligence Artificielle », fusent de tous côtés, tout comme, et c'est bien malheureux, les « il » lorsqu'il s'agit de désigner des produits comme chatGPT ou Bing chat. Soyons clair et, Sam Altman (CEO d'OpenAI) ne cesse de le répéter : il n'y a pas de « il », mais uniquement un « ça ». Et même si l'on peut effectivement faire un raccourci linguistique (en particulier en français, où les objets ont un genre), il faut éviter à tout prix de sombrer dans l'anthropomorphisme. En échangeant du texte avec un tel système ou agent conversationnel, vous ne dialoguez pas avec une entité, mais fournissez des données textuelles et recevez en retour un texte composé par un modèle qui « devine » statistiquement quels mots (ou *tokens*) il doit placer après celui qu'il vient d'afficher. En d'autres termes, il s'agit d'une prédiction de mots et de phrases qui ont une « probable » relation avec le texte que vous avez fourni, selon les données qui ont servi à l'entraînement. Rien de plus, même si le résultat est stupéfiant (et/ou inquiétant).

Ce qui occupe l'esprit et l'actualité en ce moment tourne autour des LLM ou *Large Language Model*. Les LLM sont un sous-ensemble du domaine de l'intelligence artificielle. Il s'agit d'un groupe d'algorithmes utilisant, entre autres, des réseaux de neurones de grande taille et un processus d'apprentissage complexe pour produire un texte par prédiction statistique. Un LLM n'est donc pas, à proprement parler, un réseau de neurones même s'il s'agit d'une composante du système. C'est l'architecture complète (structure, algorithmes, données d'entraînement, etc.) qui forme le modèle final.

Dans le cas de GPT (*Generative Pre-trained Transformer*), une architecture spécifique, et maintenant très populaire, est utilisée, faisant usage de *transformers*. Ce concept, inventé dès 2017, a remplacé une autre forme de modèle appelé RNN (*Recurrent Neural Networks*) et est à la base des modèles que vous pouvez voir maintenant fleurir de toutes parts.

Un modèle est donc un *design* qui sera ensuite entraîné en lui faisant ingurgiter des données... des tonnes et des tonnes de données. Plus le modèle est grand, plus il sera en mesure de fournir une importante cohérence et précision, mais plus il sera consommateur de mémoire et de ressources processeur. La taille du modèle est généralement définie par le nombre de ses paramètres. Ce terme provient directement du monde des statistiques et désigne des valeurs, internes au modèle, qui peuvent changer à mesure que son entraînement progresse. Vous pouvez voir cela comme le nombre de points de réglages possibles sur un moteur (débit de carburant, d'air, allumage, température, pression, richesse du mélange, etc.), mais ici, la quantité se chiffre en milliards. Entre 2 et 7 pour les petits modèles et de l'ordre de 200 à 500 pour les gros. On parle même de plusieurs billions (10^{12}) de paramètres pour GPT-4...

Pour ajuster ces paramètres dans la phase d'apprentissage, les données textuelles sont découpées en *tokens*, tout comme les textes et contextes textuels que vous lui soumettrez une fois le modèle entraîné, le fameux *prompt*. Un *token* est l'unité la plus petite d'un texte du point de vue du LLM. Il peut s'agir de mots, de morceaux de mots ou de groupes de caractères/symboles. Un outil en ligne d'OpenAI **[1]** permet de visuellement comprendre ce qu'est un *token*. Ainsi, la chaîne de caractères « Ceci est une longue phrase. » (avec son point) est décomposée en : « C », « ec », « i », « est », « une », « long », « ue », « phrase », « . » (notez les espaces au début de certains *tokens*).

Comme nous allons le voir plus loin, un modèle entraîné prend la forme d'un fichier de plusieurs Gio dont la taille est proportionnelle à son nombre de paramètres. Bien entendu, un modèle de quelque 7, 13, 30 ou 65 Gio ne peut raisonnablement pas tout « savoir » ou stocker, mais sera tout de même en mesure d'apporter des réponses sur plus ou moins tout. Cette capacité à faire des prédictions sur des données non connues est nommée « inférence ». Il y a donc une part d'inconnu dans le processus et s'ajoute même une part aléatoire faisant partie inhérente du système.

Ceci provoque alors parfois un phénomène appelé « hallucination », où le modèle donne une réponse qui n'est aucunement justifiée par les données d'entraînement. Ce sont des réponses, souvent catégoriques, qui « sortent de nulle part » et sont un réel problème, sinon le plus important, selon de nombreux spécialistes du domaine des LLM.

Enfin, un LLM qui n'est donc, dans les grandes lignes, qu'un modèle statistique n'a aucune notion de morale ou d'éthique et est, de base, parfaitement capable de répondre des choses étonnantes, sinon totalement choquantes. Je m'abstiendrai ici de vous donner des exemples puisque vous serez en

mesure de « torturer » vous-même des modèles une fois l'article conclu. Il est donc nécessaire d'orienter le modèle vers un ensemble de principes qui nous paraissent à tous (normalement) éthiques et moraux. On parle alors d'alignements du modèle qui comme d'autres ajustements (*fine-tuning*) interviennent après l'entraînement et demandent généralement une intervention humaine, comme un processus d'évaluation des réponses.

Nous venons de couvrir ici un certain nombre de termes et de notions qui forment la base de ce qu'il faut savoir sur le domaine. Gardez à l'esprit cependant que tout ceci évolue de façon extrêmement rapide et que d'autres définitions, techniques et appellations vont apparaître à mesure que les recherches progressent dans les mois futurs. L'avenir nous dira ce vers quoi tout ceci nous conduira, mais en attendant, il est temps de passer à la pratique...

2. EN AVANT !

Avant toute chose, précisons que les développements autour de ces technologies (modèle, outils, scripts, formats, etc.) se font à une cadence infernale.

Le Raspberry Pi 400 est, dans les grandes lignes, une Raspberry Pi 4 avec 4 Go de RAM intégrée sous la forme d'un clavier/ordinateur compact (AZERTY svp). Son usage de prédilection semble être celui d'un client léger, plus qu'une carte de développement pour l'embarqué, mais son véritable intérêt ici est la taille de la mémoire physique disponible.

Tout ceci progresse à une vitesse faramineuse, ce qui est très excitant, mais implique forcément que les manipulations faites à un instant « t » ne seront peut être plus applicables strictement à l'identique la semaine suivante. Ceci est vrai sur SBC comme avec une configuration PC conséquente et vous devez être prêt à devoir ajuster et adapter, de-ci de-là, les manipulations que nous décrivons.

2.1 Préparer le SBC

Lorsqu'on parle d'exécuter un modèle, quelle que soit la plateforme utilisée, les ressources disponibles sont, par définition, ce qui formera le goulot d'étranglement qui tuera tout aspect interactif. Trois d'entre elles sont importantes et une est cruciale. Nous avons :

- l'espace de stockage disponible. Les modèles occupent une place conséquente, en particulier à l'échelle d'un système embarqué comme une carte Raspberry Pi ;
- la puissance de calcul, dont le manque se traduira simplement en temps d'exécution supplémentaire. Ce n'est donc pas, en principe, un point éliminatoire, mais simplement très problématique ;

- la mémoire qui, elle, est fondamentale et éliminera littéralement et sans état d'âme la plupart des plateformes populaires actuellement disponibles (Pi3, Orange Pi Zero, etc.), ne laissant que celles à un prix outrageusement élevé (Pi4) ou à l'avenir incertain (Orange Pi 5 RK3588S, par exemple).

90 % des SBC ne sont tout simplement pas capables de faire fonctionner ce qui va suivre, car ils n'embarquent pas suffisamment de RAM. Une Raspberry Pi 4 n'ayant que 2 Gio de mémoire ne sera pas utilisable. Même les plus petits modèles LLM utilisables et dignes de ce nom font quelque 4 Gio qui, une fois chargés en mémoire occuperont ~6 Gio plus, bien entendu, la mémoire nécessaire pour le reste du code et des données, et l'OS.

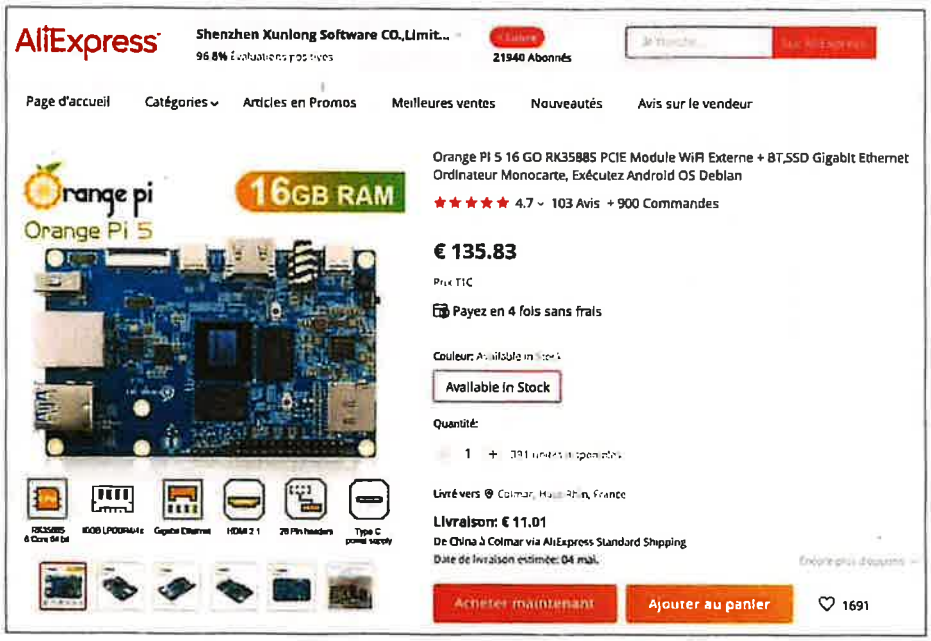
Plus de 6 Gio ! Comment tout cela peut-il donc fonctionner sur une Raspberry Pi 400 avec seulement 4 Gio de RAM ? La réponse tient en un mot : le « swap ». Le *swap* ou espace d'échange est un mécanisme permettant au système de stocker des données qui semblent être en mémoire et il devient alors possible d'étendre la mémoire virtuelle au volume RAM physique plus le *swap*. Pour les applications, la mémoire RAM semble donc plus importante, mais le contrecoup est important, car la RAM est rapide, mais le stockage lent. Ce que vous gagnez en mémoire, vous le perdez en performances (et en espace de stockage).

[...]

Contrairement aux modèles LLaMA ou basés sur LLaMA, il existe une plus grande diversité en termes de nombre de paramètres (et donc de taille). Explorez *Hugging Face* et flirtez avec les limites matérielles et vous trouverez certainement le juste milieu entre vitesse et précision. *LLaMA.cpp* et *bloomz.cpp* (ainsi que *Alpaca.cpp* [9]), vous permettront de tester, explorer et comprendre les mécanismes, les formats et les contraintes. Il est fort possible qu'entre le moment de la rédaction du présent article et celui où vous le lirez, d'autres solutions, projets et modèles aient vu le jour. Mais vous devez avoir maintenant les bases pour les comprendre et les tester...

3. QUÊTE DE PERFORMANCES ET CONCLUSION

Il est possible de légèrement gagner en performance via différentes approches. Utiliser un système faisant fonctionner moins de services est un point de départ intéressant,



Voici une très bonne option pour l'exécution locale d'un LLM. Avec ses 16 Gio de RAM, nous serions en mesure d'utiliser des modèles plus précis (comme LLaMA 13B) sans avoir recourt au swap. Ce à quoi s'ajoute une interface PCIe permettant d'utiliser un SSD NVMe, tout en gardant un prix presque raisonnable (par rapport à ceux actuellement constatés sur les Raspberry Pi). Par contre, même pour un budget de quelque 130 €, il faut tout de même se poser la question : que faire de la carte une fois qu'on aura fini de « jouer » ?

tout comme le fait de structurer le système de façon à *swapper* sur une partition et non dans un fichier. On peut également envisager d'opter pour un autre système, dont l'installation par défaut sera plus économe, comme FreeBSD (voir l'article en rapport dans ce numéro ou celui dans le précédent numéro **[10]** ou dans le 41 avec NetBSD **[11]**).

Tout ceci ne rendra cependant pas pour autant l'ensemble réellement utilisable, juste très sensiblement moins « poussif ». Mais le réel objectif n'est pas d'avoir une longue et joyeuse conversation avec un modèle, mais simplement de pouvoir ramener le concept à quelque chose de tangible et de bien comprendre, en pratique, qu'il n'y a non seulement pas de magie, mais également en quoi ce genre de système est intimement lié aux performances matérielles.

Ce qui est également absolument stupéfiant c'est la quantité d'informations qui **semblent** stockées dans un simple fichier de 4 Gio. Ce n'est, bien entendu pas réel, mais le modèle **apparaît** pouvoir répondre de façon cohérente à tout et n'importe quoi. De la même manière qu'un agent conversationnel **semble** avoir une réflexion, les informations **semblent** présentes dans les 4 Gio, mais ce n'est pas le cas, ce sont juste des probabilités.

Quoi qu'il en soit et même si les résultats sur un SBC sont assez décevants (mais sans surprise), il est important de retirer la mystique de ce qui est en train de se passer globalement et, pour cela, rien de mieux que d'expérimenter soi-même. Si vous souhaitez poursuivre votre exploration, dans un environnement plus riche en ressources et moins spartiate que de l'embarqué, je vous recommande fortement les récents *streams* Twitch d'Yves Rougy (Yorzian) **[12]**, réunis dans une *playlist* YouTube dédiée **[13]**. DB

MINISTÈRE
DES ARMÉES

Liberté
Égalité
Fraternité

Copie ayant obtenu la meilleure note

Cas pratique

L'administration n'a volontairement pas corrigé les imperfections de fond et de forme dans la copie communiquée ci-après.

RÉPUBLIQUE FRANÇAISE
DIRECTION GÉNÉRALE DE LA SÉCURITÉ EXTÉRIEURE
DGSE

Année : 2023

Concours : Concours externe pour l'accès au grade de contrôleur spécialisé

Épreuve : Cas pratique

Consignes :

- Ne pas signer la composition et ne pas y apporter de signe distinctif
- Numéroter chaque page; placer l'ensemble dans l'ordre et le bon sens
- N'effectuer aucun collage ou découpage de sujets ou de feuilles
- Ne joindre aucun brouillon

SERVICE DE DOCUMENTATION
CELLULE INNOVATION

Paris, le 14/11/2023

NOTE

Objet : Étude d'un projet d'agent conversationnel basé sur l'intelligence artificielle

Depuis 2022 et l'engouement du grand public auprès des agents conversationnels basés sur l'intelligence artificielle (IA), la plupart des acteurs des industries, mais aussi des domaines documentaires, se sont demandés comment intégrer ces technologies dans leurs processus. L'enjeux est de savoir quelles sont les possibilités offertes par les dernières avancées en matière d'IA. Cette note propose d'abord des définitions liées à ces thématiques et évalue la faisabilité de ce projet, puis expose les éventuels risques et limites de ces techno.

I/ Définitions et faisabilité du projet

La volonté d'automatiser des tâches n'est pas nouvelles et les services explorent continuellement des pistes à la recherche de solutions réalistes.

Cependant, il convient avant tout de correctement définir les termes liés au périmètre du projet. Voici donc quelques définitions importantes :

- Transformateur : réseau de neurones exploitant les relations entre des données qui se suivent pour en apprendre des éléments de contexte et construire un sens ;
- LLM pour Large Language Model : groupe d'algorithmes utilisant de grands réseaux de neurones et un processus d'apprentissage afin de produire un texte par prédiction statistique ;
- Agent conversationnel : modèle construisant un texte statistiquement à partir d'une entrée textuelle et d'un entraînement sur un jeu de données.

À partir de ces quelques définitions, il est possible de commencer à évaluer la faisabilité générale d'un tel projet. En effet, certaines notions sont déterminantes pour comprendre précisément les tenants et aboutissants de ces technologies.

D'abord, afin de maîtriser les données et résultats attendus, il est primordial d'utiliser un jeu de données propre et distinct. Ceci permettra d'entraîner le modèle sur une thématique sur laquelle il pourra exceller.

Ensuite, le système doit être fermé, coupé d'Internet, afin que le réservoir de données ne soit pas pollué par des données externes. Cet enjeux se décline sous quatre notions : l'utilité, la sécurité, la vérifiabilité et la fiabilité.

Enfin, la faisabilité est nécessairement liée à la performance et la fiabilité des supports et ressources utilisés. En ce sens, la contrainte est autant liée au système qui va être utilisé qu'à la manière dont il va être alimenté et avec quelles ressources.

Ce type de technologies évoluant sans cesse, elles offrent toujours plus de possibilités et de moyens d'être incorporées dans les processus. Mais cela implique aussi des risques et limites non négligeables.

II / Les risques et limites des agents conversationnels

Selon la manière dont est implémenté l'agent conversationnel et ses besoins, des risques et limites apparaissent.

La liste suivante présente des risques liés à l'utilisation de cette technologie :
- L'utilisation de données sans les autorisations adéquates dans le cas où la source des données est Internet ;
- Des problèmes de confidentialité notamment liés à des fuites d'informations ou de savoir-faire peuvent apparaître si l'agent est détourné et que les données sont sensibles ;
- Il est potentiellement possible de détourner les capacités de l'outil en fonction de ses capacités, en réalisant des deep fake, ce qui peut

mener à des usurpations ;

Maintenant, alors que ces systèmes offrent des moyens nouveaux, ils souffrent aussi de limites liées aux technologies utilisées :

- Sur les modèles ouverts à Internet, des problèmes de fiabilité existent. L'agent présente des imprécisions, les résultats peuvent souffrir de biais liés au modèle et les résultats scientifiques peuvent n'avoir aucun sens ;
- Il est nécessaire de limiter le produit sur certains sujets sans quoi des dérives apparaissent, notamment sur des sujets religieux, politiques et sociétaux ;
- Fondamentalement, l'IA ne sera jamais aussi intelligente que le jeu de données sur lequel elle apprend. En effet, les IA ne comprennent pas le monde qui les entoure, elles ne peuvent apprendre, restituer et mémoriser dans le même temps, elle ne comprend pas ses propres raisonnements et malgré son assurance certaine peut donner des réponses complètement fausses ;
- Ces agents n'ont pas de capacité comportementalles et se basent seulement sur des ressources documentaires ce qui implique que s'il n'y a pas de données, le résultat ne peut pas être cohérent et vérifiable ;
- D'un point de vue technique, ces technologies consomment beaucoup de ressources environnementales, économiques et humaines ;
- Tous les modèles ne se valent pas et nécessitent des entraînements spécifiques selon l'utilisation ;
- L'agent est évidement limité par la taille de sa base de ressources et même lorsqu'elle est énorme, des biais ou phénomènes imprévisibles peuvent apparaître.

L'utilisation de l'intelligence artificielle est un enjeux important pour l'industrie. Cependant, il convient de maîtriser spécifiquement les technologies et ressources utilisées pour atténuer les risques, accepter les limites. La faisabilité du projet dépend donc des choix faits (ChatGPT, LaMDA, Sparrow) et des moyens engagés dans le projet, notamment en matière de ressources matérielles.

Service de Documentation
Cellule Innovation

Paris, le 14/11/2023

FICHE

Objet : Intelligence artificielle et systèmes embarqués

Dans le cadre d'une utilisation de l'intelligence artificielle (IA) sur des systèmes embarqués, les points suivants ont été retenus :

- Les systèmes embarqués (RaspberryPi notamment), ne sont pas du tout adaptés à l'utilisation de l'IA :
  - Il existe des limitations matérielles au niveau de la mémoire vive, du stockage et du processeur ;
  - Statistiquement, 90% des systèmes embarqués sont incapables de faire tourner une IA à cause du matériel ;
  - Même si des solutions existent pour pallier des impossibilités techniques, cela se fera au coût des performances.
- Des possibilités de faire fonctionner les cartes avec de l'IA hors-ligne existent.

D'une manière générale, le support n'est pas adapté à la demande de puissance matérielle de l'IA et risque d'être très peu performant, au contraire de ce que propose cette technologie initialement.