

GPT et les grands modèles du langage en Intelligence Artificielle : principes et défis

Malik Ghallab
LAAS-CNRS, Université de Toulouse

Résumé

Cette note résume la communication présentée par l'auteur à l'Académie des Sciences, Inscriptions et Belles Lettres de Toulouse le 16 novembre 2023. Elle introduit brièvement les développements récents de l'Intelligence Artificielle générative. Elle présente informellement les principes des grands modèles de langages, en illustre les aptitudes et en discute les limitations. Elle conclut sur les risques et défis éthiques que soulèvent ces développements.

1 Introduction

L'IA est une discipline scientifique qui cherche à comprendre et à modéliser l'intelligence naturelle des êtres vivants par des approches computationnelles ou algorithmiques¹. L'objet de l'IA est donc l'intelligence, laquelle est mal définie et multiforme. Cependant on peut constater l'intelligence – animale ou humaine – dans la réalisation de tâches. La démarche de l'IA consiste alors à mécaniser des tâches de plus en plus complexes qui requièrent de l'intelligence. Il s'agit, par exemple, d'interpréter une scène, de manipuler des outils, de planifier des actions (comme y excellent de nombreuses espèces), ou d'apprendre à compter, lire, écrire et comprendre des textes (les fondamentaux de nos écoles). Pour ce faire, on modélise la tâche en question, on cherche des méthodes et des algorithmes pour la résoudre, et on développe des implémentations logicielles et matérielles de ces algorithmes. Ceci permet de passer à une phase expérimentale pour évaluer empiriquement les capacités des modèles et algorithmes et leurs performances pour cette tâche, que l'on s'efforce d'améliorer.

L'IA relève des sciences de la modélisation et de l'information, avec des recouvrements vers les sciences cognitives et les sciences de la décision

1. Privilégier des approches algorithmiques n'est pas un biais. L'humanité cherche à comprendre et à modéliser le monde en le décrivant les informellement, puis en représentant mathématiquement ce qui peut l'être, et aujourd'hui, de plus en plus, par modèles algorithmiques.

(recherche opérationnelle, économie). C’est un domaine relativement récent, même si ses racines sont très anciennes. Depuis le milieu du 20^e siècle, l’IA bénéficie et contribue significativement aux développements des matériels et logiciels informatiques et de télécommunication. Elle a transformé considérablement nos capacités de modélisation, d’analyse, d’organisation et de recherche d’informations. Elle a un impact majeur sur pratiquement tous les champs d’investigation scientifiques et techniques, allant de la biologie, la médecine, ou l’astronomie, en passant par les sciences humaines et sociales.

L’IA, comme toute *technoscience* aujourd’hui, est aussi un champ technologique, difficilement séparable du volet scientifique. Le savoir est motivé par l’action. L’IA alimente de nombreuses industries en techniques et déploiements, que l’on espère socialement utiles. Les investissements de R&D en IA sont aujourd’hui considérables². Ils sont très massivement dominés par quelques multinationales, lesquelles contrôlent ainsi l’évolution du domaine selon des logiques capitalistes plutôt sociales (cf. [section 4](#)).

Résumons très schématiquement les tendances récentes de l’histoire riche et complexe de l’IA³. Depuis ses débuts vers le milieu du 20^e siècle, l’IA a donné lieu à de nombreux succès. Mais, le plus souvent, les méthodes mises en œuvre nécessitaient d’énormes efforts de développement, de modélisation et d’adaptation à chaque tâche. Ces méthodes avaient tendance à être fragiles et étroites, très difficiles à étendre à de nouvelles tâches⁴. Pendant de nombreuses années, les systèmes d’apprentissage d’IA manquaient de capacité de généralisation et de transfert d’une tâche à une autre. Ces capacités d’adaptation, essentielles à l’intelligence, commencent à être atteintes efficacement, pour des tâches d’interprétation et de génération de données grâce aux progrès des réseaux de neurones.

Réseaux de neurones artificiels. Un neurone artificiel est une fonction multivariable relativement simple : $f(x_1, \dots, x_n) = g(\sum_{i=0,n} \theta_i x_i)$, où les θ_i sont des paramètres, g est une fonction non linéaire, et le terme x_0 est un biais ([Figure 1\(a\)](#)). L’apprentissage consiste à estimer les valeurs des paramètres θ_i qui permettent de se rapprocher le plus des valeurs souhaitées de f . Il s’agit d’un problème d’optimisation, similaire dans son principe à ce qui est fait classiquement dans une régression.

Un réseau de neurones connecte des entrées et des sorties de plusieurs de ces fonctions. C’est par exemple un réseau *perceptron* organisé en couches régulières, un neurone de la couche j prend comme entrées les sorties des neurones de la couche $j - 1$ et fournit sa sortie à tous ceux de la couche $j + 1$ ([Figure 1\(b\)](#)). L’apprentissage dans un réseau se fait par un algorithme dit

2. Ils sont estimés entre 100 et 200 milliards de dollars par an.

3. Sur l’histoire de l’IA, voir par exemple, [\[29, 30\]](#).

4. Un exemple est le système Watson, champion impressionnant du jeu de questions/réponses “Jeopardy”, qui n’a pas pu être transposé avec succès au domaine médical malgré des investissements considérables [\[13\]](#).

de *rétro-propagation* : à chaque exemple d'entraînement (x, y) on adapte les paramètres, couche par couche, de la dernière à la première, pour rapprocher la sortie $f(x)$ de la valeur désirée y .

Un réseau de neurones peut approximer n'importe quelle fonction continue, à condition d'avoir suffisamment de neurones et de paramètres, de disposer de suffisamment de données sur ce qu'on souhaite que calcule cette fonction, et de pouvoir estimer les paramètres du réseau. Les réseaux de neurones, bien que connus depuis très longtemps [32], n'ont que récemment pu remplir ces conditions. Ceci a pu être fait grâce à internet pour la quantité de données disponibles, à la loi de Moore pour la puissance des processeurs numériques, aux progrès algorithmiques pour les méthodes d'estimation des paramètres, et aux architectures ou organisations des neurones en des réseaux multicouches performants.

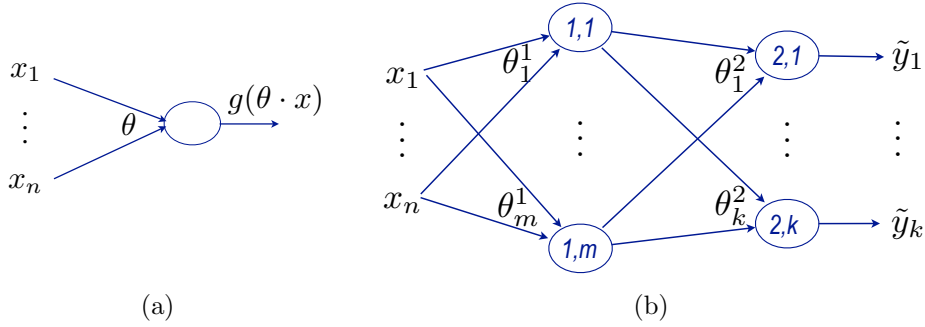


FIGURE 1 – (a) Un neurone artificiel : θ est un vecteur de paramètres, x est le vecteur d'entrée, $\theta \cdot x$ est leur produit scalaire ; (b) Un réseau perceptron à deux couches : la sortie de chaque couche est un vecteur, Θ_i est la matrice des paramètres de la couche i , et la sortie du réseau est $\tilde{\mathbf{y}} = g_2(\Theta_2 \times g_1(\Theta_1 \times \mathbf{x}))$.

Interprétation et génération de données. Les réseaux de neurones ont donné lieu à des progrès significatifs dans deux tâches importantes : l'interprétation et la génération de données de toutes sortes – signaux, textes, images, sons, vidéos, etc.

- *Interprétation des données.* C'est le champ classique de la *reconnaissance des formes*, laquelle s'appuie sur des fonctions caractéristiques des données à traiter pour apprendre à classer ou interpréter ces données. Ces caractéristiques sont par exemple des transformées de Fourier, des splines ou des ondelettes, utilisées pour interpréter un sismogramme ou un cardiogramme. On a consacré des efforts considérables à la conception de ces fonctions caractéristiques pour chaque type de données particulières. Aujourd'hui, les réseaux multicouches permettent d'estimer des fonctions universelles de classification, robustes pour interpréter toutes sortes de données. Ces réseaux fournissent, de façon implicite et sans effort, les

caractéristiques adaptées aux données d'apprentissage. Les méthodes d'interprétation des données ne sont plus coûteuses et spécialisées pour chaque application. Elles sont désormais largement déployées pour l'analyse de données multimodales (signaux physiques, images, sons, vidéos, textes) dans de nombreuses applications exigeantes, e.g., en imagerie médicale ou astronomique.

- *Génération de données.* Ici aussi, les principes sont connus depuis longtemps. Il s'agit d'estimer une distribution de probabilité qui représente adéquatement les données d'intérêt, puis d'échantillonner dans cette distribution pour générer des instances probables de ces données. L'échantillonnage génératif a également bénéficié des progrès des réseaux multicouches en termes de performances matérielles, algorithmiques, et d'architectures. Les développements d'IA pour la génération de sons, d'images, et de vidéos sont de plus en plus performants. Les méthodes de *traitement automatique du langage naturel* (TAL) rentrent dans le cadre de cette évolution. Ces méthodes sont utilisées pour la compréhension de textes, la traduction, la synthèse ou le dialogue. Leurs succès sont restés longtemps limités, et leur portée restreinte. Les développements des méthodes de génération de données textuelles ont donné lieu à un changement radical avec les systèmes dits *grands modèles de langage* (LLM pour "Large Language Models")⁵. Ce changement se reflète de façon spectaculaire dans la diffusion des outils d'IA, jusqu'alors limitée à un public restreint, qui touche désormais des centaines de millions de personnes.

On propose ci-après un bref aperçu sur les LLM, un domaine en rapide évolution. Pour des revues complémentaires de l'état de l'art, voir par exemple [54, 35], ou [23] pour un accent sur le raisonnement dans les LLM. Les modèles multimodaux, qui traitent conjointement de textes, d'images, de sons et de données hétérogènes, sont discutés dans, e.g., [3, 27].

2 Principes des LLMs

Le principe d'un LLM est de générer le mot suivant dans une séquence de mots. Le mot généré se rajoute itérativement à la séquence pour la synthèse de longs textes. L'opération de base est à un problème de prédiction classique : ayant une séquence de termes $\langle x_1, \dots, x_{n-1} \rangle$, quel est le terme suivant x_n ?

Lorsque les x_i sont les états d'un système dont l'évolution est bien modélisée, on aborde ce problème de prédiction grâce au modèle,

5. Le terme « modèle » est sans doute impropre pour ces systèmes. Un modèle est censé être une description intelligible, explicative, justificative et prédictive d'une certaine réalité. Seul le volet prédictif, au sens statistique, est présent dans les LLM. Nous conservons l'acronyme LLM et l'utilisons également pour les logiciels multimodaux.

déterministe ou probabiliste, du système considéré. C’est ce que l’on fait par exemple pour le mouvement d’un mobile, l’évolution d’un processus biochimique, physique, ou météorologique.

Induction statistique. Lorsqu’aucun modèle n’est disponible, mais que le domaine n’est pas trop erratique, on a recours à la prédiction statistique. Cette approche produit un modèle “superficiel”, qui fait des prédictions basées sur des statistiques, et non sur des relations causales explicites. Un tel modèle cherche à prédire le comportement d’un système sans avoir de relations de cause à effet qui engendrent ce comportement et permettent de l’expliquer. Par exemple, on chercherait à prédire la position prochaine de Saturne en utilisant des statistiques sur ses éphémérides passées sans connaître les lois de Keppler, et encore moins celles de Newton. On comprend aisément que la mécanique céleste nous apporte non seulement plus de précision, mais surtout une *compréhension* de ce qui se passe et un moyen de *valider* de notre modèle.

La prédiction statistique suppose que le domaine est régulier. Elle s’appuie sur l’hypothèse de l’induction statistique : on stipule que ce qui est vrai pour les données observées reste vrai pour celles non observées. Cette hypothèse, inacceptable dans de moult domaines⁶, exige réserves et prudence, en particulier dans tout domaine où peuvent se produire des événements rares aux effets importants. L’induction statistique, lorsqu’elle est applicable, nécessite l’acquisition d’un nombre suffisant de données pour estimer les distributions servant à prédire. Mais les observations resteront toujours finies, alors que les instances possibles sont potentiellement infinies.

Un exemple simple et familier de prédiction statistique dans TAL est l’aide à la dactylographie que l’on trouve dans les smartphones et de nombreuses applications. Dans TAL, les données d’entraînement sont très abondantes⁷. En outre, le langage naturel est très régulier, comme le montre sa faible *entropie*. Ce fait est connu depuis l’introduction de l’entropie de l’information par Claude Shannon : “*Toute personne parlant une langue possède, implicitement, une énorme connaissance des statistiques de la langue (...) qui lui permet de compléter une phrase inachevée dans une conversation*” [43]. Shannon a proposé la notion de “N-grams”, qui est le problème de prédiction ci-dessus appliqué à des séquences de lettres ou de mots grâce à l’estimation des probabilités conditionnelles $\Pr[x_n | x_1, \dots, x_{n-1}]$, x_i étant le i -ème terme dans la séquence.

Les N-grams n’informent pas sur la sémantique des mots. Les tâches de TAL nécessitent d’évaluer quels mots sont sémantiquement proches. Deux mots a et b sont susceptibles d’être sémantiquement proches s’ils sont utilisés de façon interchangeable dans des contextes similaires. En simplifiant, on

6. N’observer que des cygnes blancs n’exclue pas l’existence de cygnes noirs [46].

7. Les documents accessibles sur le web sont estimés entre 10^{12} et 10^{14} mots.

souhaite qu'il y ait aussi fréquemment a que b après une séquence de mots $\langle x_1, \dots, x_{n-1} \rangle$, à savoir : $\Pr[a \mid x_1, \dots, x_{n-1}] \simeq \Pr[b \mid x_1, \dots, x_{n-1}]$.

Cette approche se heurte à un problème de complexité : la table des probabilités conditionnelles $\Pr[x_n \mid x_1, \dots, x_{n-1}]$ a une taille de l'ordre de δ^n , où δ est la taille du dictionnaire. Typiquement, δ est d'environ 10^5 mots, et un contexte $\langle x_1, \dots, x_{n-1} \rangle$ peut contenir plusieurs centaines de mots. Le calcul de tables de probabilités conditionnelles de taille 10^{500} est infaisable.

Projection sémantique. Ce problème de complexité fut résolu avec une technique dite de *projection sémantique* des mots (“word embedding”) [2]. Cette projection associe à chaque mot un vecteur dans un espace métrique, $\mathbb{R}^d$, de telle sorte que les mots sémantiquement similaires aient des projections proches dans $\mathbb{R}^d$. Les projections de mots sont calculées par des réseaux neuronaux et utilisées dans des tâches TAL par d'autres réseaux. Des calculs vectoriels simples peuvent être convertis en opérations TAL.

Exemple 1. Notons $[\vec{mot}] \in \mathbb{R}^d$ le vecteur projection d'un *mot*. Le calcul sur les projections permet des opérations arithmétiques telles que :

$$\begin{aligned} \text{la relation :} \quad & [\vec{France}] + [\vec{capitale}] \simeq [\vec{Paris}], \\ \text{l'analogie :} \quad & [\vec{Berlin}] - [\vec{Allemagne}] + [\vec{Japon}] \simeq [\vec{Tokyo}], \\ & [\vec{cuivre}] - [\vec{Cu}] + [\vec{or}] \simeq [\vec{Au}], \\ \text{la proximité :} \quad & [\vec{parler}] \cdot [\vec{bavarder}] < [\vec{parler}] \cdot [\vec{radoter}]. \end{aligned}$$

Il s'agit ici d'additions, de soustractions et de produits scalaires de vecteurs. Le produit indique par exemple que “bavarder” est plus proche de “parler” que “radoter”. $\square$

Le traitement de phrases est ramené au traitement d'une séquence de vecteurs. On utilisa diverses variantes de réseaux neuronaux dits récurrents, adaptés aux séquences pour prendre en compte les liens entre un terme et son successeur. Cependant, les réseaux récurrents offrent des possibilités limitées. Ils ignorent largement les liens entre mots non consécutifs dans une longue phrase, lesquels peuvent être essentiels pour le sens. D'autres architectures de réseaux ont été proposées, mais avec des succès limités. Le traitement du langage naturel exige la résolution de nombreuses ambiguïtés, par exemple les coréférences et appariements des pronoms.

Exemple 2. Le pronom “elle” dans les deux phrases suivantes fait référence à deux mots biens distincts :

- (i) *Marie n'a pas pris la route, car elle s'est perdue.*
- (ii) *Marie n'a pas pris la route, car elle était inondée.*

Si on n'apprécie pas les liens possibles entre des paires de mots éloignés, on ne peut pas résoudre ces ambiguïtés. $\square$

Transformeurs. Les ambiguïtés dans le traitement automatique du langage ont été résolues avec succès par des mécanismes dits d’“attention”, capables de relier des mots situés en des positions éloignées dans la phrase. Ces mécanismes utilisent une architecture de réseau dite “*transformeur*”⁸. Elle est constituée de couches neuronales d’attention à têtes multiples (qui estiment les n^2 liens possibles entre n termes) et de couches de type perceptron (Figure 1(b)). Les têtes multiples permettent le calcul des relations entre des paires de mots pour lever les ambiguïtés. Les traitements se réduisent uniquement aux calculs des fonctions successives $g(\theta \cdot x)$ sur chaque neurone pour les entrées du réseau et les paramètres appris. Ces traitements, organisés en calculs matriciels réguliers, sont effectués massivement en parallèle pour plus d’efficacité. Après apprentissage, chaque étape dans la génération d’un texte prend comme entrée supplémentaire le terme généré à l’étape précédente.

Les LLM sont pré-entraînés par apprentissage auto-supervisé sur de très nombreux documents accessibles sur internet. Auto-supervisé signifie que le système lit un document et à chaque pas cherche à prédire le mot suivant. Il optimise ses paramètres de façon à ce que le mot qu’il prédit correspond au mot suivant dans son texte d’apprentissage. D’autres étapes supplémentaires d’apprentissage sont utilisées pour éviter des réponses indésirables, alignées sur les préférences des concepteurs. Il s’agit par exemple d’apprentissage par renforcement avec quelques retours humains pour apprendre la fonction de renforcement (qui approxime les préférences humaines), laquelle est utilisée ensuite en auto-apprentissage [36, 20]. En outre, un LLM peut être adapté à des applications ou des domaines spécifiques, par exemple en aéronautique, en mathématique, ou en médecine. Ainsi, Med-PaLM2 est un système affiné sur des données médicales à partir du LLM polyvalent PaLM2. Enfin, la plupart des LLM sont associés à une interface de dialogue (“chat”). Le système réagit à une requête de l’utilisateur (“prompt”) en générant sa réponse comme complétion de la requête. Le *contexte*, ou ensemble des entrées donnant lieu à la réponse générée, peut prendre en compte une séquence d’étapes du dialogue. Il est ainsi possible à l’utilisateur d’organiser sa requête en étapes qui simplifient ou orientent le traitement⁹.

Il existe à ce jour quelques centaines de LLM. Les plus conséquents sont qualifiés de “foundation models”. Ils sont généralement multimodaux (textes, images, sons) et permettent de dériver, une fois pré-entraîné, des modèles plus spécialisés. Le rapport annuel 2024 sur l’état de l’IA [31] en cite près de 150, dont 109 développés aux USA, 20 en Chine, et 15 en Europe. Près des 3/4 viennent de l’industrie, principalement des multi-nationales du

8. Le sigle GPT correspond à *Generative Pre-trained Transformer*.

9. L’art d’interroger un LLM fait désormais l’objet de cours académiques ou en ligne, e.g., Coursera, Udemy ou <https://learnprompting.org/>.

numérique (e.g., 18 de Google, 11 de Meta, 9 de Microsoft)¹⁰. Les universités sont peu présentes, avec quelques exceptions (7 de Tsinghua, 3 de Berkeley, 2 de Stanford), et développent des modèles de petite taille.

3 Limitations et performances des LLM

Les LLM ne sont pas dotés de connaissances formelles. Ils n’ont pas de capacité algorithmique de raisonnement, contrairement à d’autres systèmes en IA, par exemple de planification, de diagnostic, ou d’aide à la décision. Ils ne disposent ni de grammaire ni de logique. Ils ne font pas de recherche dans une base de textes (comme le ferait un navigateur internet). Leur répertoire algorithmique est limité, par exemple, il n’y a pas d’itération, de récursion, ou de recherche arborescente avec retours arrières. à l’exception du contexte en entrée et des paramètres appris, ils n’ont pas de mémoire pour stocker des structures de données. Leurs seuls mécanismes de calcul sont :

- la rétro-propagation pour l’estimation des paramètres lors de l’apprentissage ; et
- la propagation des entrées vers les sorties pour la prédiction par le calcul des fonctions f de chaque neurone, selon la topologie du réseau.

Limitations théoriques. Les LLM sont confrontés à plusieurs limitations théoriques. Par exemple, ils ne peuvent pas traiter correctement plusieurs classes de langages formels, dont les langages non réguliers, les langages récursifs ou les langages périodiques réguliers, qui sont tous beaucoup plus simples que le langage naturel [21]. Il a été démontré que l’apprentissage avec un nombre fini de textes ne peut donner lieu qu’à une “compréhension” limitée d’une langue, au sens de la sémantique de dénotation formelle [34]. Cependant, les LLMs sont capables de capturer certains aspects importants de la signification, tels que les “rôles conceptuels”, selon la théorie des rôles conceptuels [34].

Ces limitations théoriques sont analogues aux résultats de la complexité algorithmique. La plupart des algorithmes d’IA ont une complexité théorique au moins exponentielle, qui peut faire croire à leur très faible utilité. Mais cette complexité en pire cas n’exclut pas des performances moyennes généralement acceptables.

Dans la pratique, un LLM suffisamment grand peut traiter un sous-ensemble large mais fini du langage (un sous-ensemble fini peut toujours être exprimé sous forme d’un langage régulier). Plus le sous-ensemble que l’on veut traiter est grand, plus le modèle LLM doit comporter de neurones et de paramètres pour apprendre tous les cas spéciaux non réguliers.

10. Parmi les plus connus, mentionnons GPT, Gemini Ultra, Anthropic Claude, Llama, PaLM-E.

L'interface de dialogue d'un LLM conduit naturellement à interroger un modèle comme un oracle capable de répondre à tout, y compris à des questions qui ne relèvent pas de ce qu'il peut calculer. En particulier les LLM sont critiqués pour leur caractère non factuel. On dit parfois qu'ils "hallucinent". Ce terme peut être trompeur pour des capacités mentales que les LLM n'ont pas. Il ne reflète pas les principes de l'induction statistique mis en œuvre par un LLM. Ainsi, une fonction f calculée par un réseau neuronal ne donne pas toujours $f(x_i) = y_i$ lorsque la paire (x_i, y_i) est dans la base d'apprentissage ; elle ne donne qu'une valeur proche de y_i . Il faut garder en tête qu'un réseau de neurone calcule une fonction d'approximation statistique, et non une requête dans une base de données.

Limitations pratiques actuelles. Au caractère non factuel des LLM se rajoutent d'autres limitations, parmi lesquelles par exemple :

- l'absence d'apprentissage incremental prenant en compte les nouvelles récentes : l'apprentissage très coûteux ne peut être fréquemment répété ;
- l'induction sur un contexte limité, sans mémoire à très long-terme ;
- l'absence d'adaptation dynamique ;
- l'absence d'interface aisée à des algorithmes spécialisés, par exemple reconnaître que l'interrogation porte sur la position prochaine de Saturne et utiliser pour cela un logiciel de mécanique céleste ;
- l'absence de transparence sur les sources ;
- l'absence d'anticipation sur ce qui peut se produire pour un dialogue faisant référence à un domaine évolutif ; ainsi les LLM sont très faibles en planification ;
- l'opacité des traitements et l'absence d'explication et de qualification de la fiabilité des propos ; ainsi réagir à une réponse d'un LLM en lui demandant "es-tu sûr ?" donne généralement lieu à des appréciations farfelues, non étayées ;
- l'absence de vérification de la rationalité des propos.

Beaucoup de travaux en cours s'efforcent de pallier ces limitations, par exemple, par un contexte d'induction très large (10^6 mots pour le LLM *Anthropic Claude3* [1]), par adaptation dynamique (méthodes dites *Chain-of-Thought* [44, 26]), par connexion à des bibliothèques mathématiques telles que Wolfram Alpha¹¹, ou par des méthodes de liens aux sources (dites de *Retrieval Augmented Generation* [7, 16, 24]).

Performances. Malgré les limitations théoriques et pratiques des LLM, leurs performances moyennes sont bonnes à excellentes dans pratiquement toutes les tâches TAL, telles que la traduction, la synthèse, l'analyse, la

11. Cf. <https://writings.stephenwolfram.com/>

compréhension de textes, et les réponses à des QCM. Les LLM maîtrisent relativement bien la transcription phonétique et l'interaction vocale. Ils démontrent des capacités en versification, humour, ou de compréhension des proverbes (la pragmatique du langage, cf. [Exemple 3](#)). Ils sont performants pour l'aide à la décision sur des textes, par exemple l'identification d'un auteur connu par son style, la reconnaissance du genre d'auteurs inconnus, la caractérisation d'un texte comme original ou traduit, écrit par un auteur dans sa langue natale ou une seconde langue.

Exemple 3. Voici un exemple d'interaction avec ChatGPT :

- *Donne moi un proverbe équivalent à “Pas de nouvelles, bonnes nouvelles”.*
- Un proverbe équivalent à “Pas de nouvelles, bonnes nouvelles” pourrait être “Silence vaut mieux que paroles” ou “Qui ne dit mot consent”. Ces proverbes expriment tous deux l'idée que l'absence de nouvelles peut être interprétée comme une situation positive, car cela signifie généralement qu'il n'y a pas de problèmes ou de nouvelles négatives à rapporter.
- *Un proverbe équivalent à “Les paroles s'envolent, les écrits restent” ?*
- Un proverbe équivalent à “Les paroles s'envolent, les écrits restent” pourrait être “Noircir du papier vaut mieux que bavarder”. Cela souligne l'idée que mettre quelque chose par écrit a plus de poids et de permanence que de simplement le dire à l'oral. Cela encourage à prendre le temps de réfléchir et de documenter les choses plutôt que de simplement les exprimer verbalement.

□

Au-delà des tâches de traitement du langage naturel, pour lesquelles ils ont été conçus, les LLM démontrent des capacités réelles, bien qu'imparfaites, dans un large éventail de tâches cognitives, qui vont du calcul arithmétique au raisonnement logique ou bon sens, en passant par la planification (cf. [Exemple 4](#)), le diagnostic, ou la résolution de problèmes mathématiques (cf. [Exemple 5](#)). Des LLM polyvalents ont démontré de bonnes performances dans moult tests [5], et ont passé avec succès des examens réputés difficiles, tels que ceux des internats de médecine en Pologne [39], au Japon [53], ou aux USA [19].

Exemple 4. Soit le dialogue suivant avec GPT4 [5] :

- *Nous avons un livre, 9 œufs, un ordinateur portable, une bouteille et un clou. Peux tu me dire comment empiler ces objets les uns sur les autres de manière stable.*
- Il est possible d'empiler ces objets les uns sur les autres de manière stable :
 - Placez le livre à plat sur une surface plane, telle qu'une table ou un sol. Le livre servira de base à la pile et constituera un support large et solide.
 - Disposez les 9 œufs dans un carré de 3×3 sur le livre, en laissant un peu d'espace entre eux. Les œufs formeront une deuxième couche et répartiront

le poids uniformément. Veillez à ce que les œufs ne soient ni fêlés ni cassés et manipulez-les délicatement pour éviter de les faire tomber.

- Placez l'ordinateur portable sur les œufs, l'écran vers le bas et le clavier vers le haut. L'ordinateur portable s'adaptera parfaitement aux limites du livre et des œufs, et sa surface plane et rigide constituera une plate-forme stable pour la couche suivante. Assurez-vous que l'ordinateur portable est fermé et éteint, et n'appuyez sur aucune touche ou bouton.

- Placez la bouteille sur l'ordinateur portable, le bouchon vers le haut et le fond vers le bas. La bouteille ajoutera de la hauteur et de l'équilibre à la pile, et sa forme cylindrique réduira la surface de contact avec l'ordinateur portable. Assurez-vous que la bouteille est vide et fermée, et ne la secouez pas ou ne l'inclinez pas.

- Placez le clou sur le bouchon de la bouteille, l'extrémité pointue vers le haut et l'extrémité plate vers le bas.

Le clou sera le dernier et le plus petit objet de la pile, et son extrémité pointue l'empêchera de rouler ou de glisser sur la bouteille. Veillez à ce que le clou soit propre et droit, et ne le touchez pas et ne le déplacez pas.

Notez la logique du plan proposé (moins de poids sur les œufs et un objet plan supportant la bouteille, le clou étant en dernier), malgré des commentaires farfelus (par exemple sur le clou). $\square$

Exemple 5. Soit le dialogue suivant avec GPT4 [5] :

- *Supposons que $g(x) = f^{-1}(x)$, et que $g(0) = 5, g(4) = 7, g(3) = 2, g(7) = 9, g(9) = 6$. Quelle est la valeur de $f(f(f(6)))$?*
- Il existe différentes façons d'aborder ce problème, mais une méthode possible consiste à utiliser le fait que $g(x) = f^{-1}(x)$ implique que $f(g(x)) = x$ pour tout x dans le domaine de g . Autrement dit, en appliquant f aux deux côtés de l'équation $g(x) = f^{-1}(x)$, on obtient $f(g(x)) = f(f^{-1}(x)) = x$. Cela signifie que nous pouvons annuler l'effet de f en appliquant g , et vice-versa. En utilisant cette idée, nous pouvons essayer de trouver $f(6)$ en cherchant la valeur de x telle que $g(x) = 6$. D'après les informations données, nous voyons que $g(9) = 6$, donc $f(6) = f(g(9)) = 9$. De même, nous pouvons trouver $f(9)$ en cherchant la valeur de x telle que $g(x) = 9$. Nous voyons que $g(7) = 9$, donc $f(9) = f(g(7)) = 7$. Enfin, nous pouvons trouver $f(7)$ en cherchant la valeur de x telle que $g(x) = 7$. Nous voyons que $g(4) = 7$, donc $f(7) = f(g(4)) = 4$. Par conséquent, $f(f(f(6))) = f(f(9)) = f(7) = 4$. $\square$

Les performances des LLM polyvalents dans des tâches de raisonnement souffrent d'erreurs et de nombreuses limitations. Mais des LLM spécialisés peuvent aller plus loin, par exemple :

- En médecine, Med-PaLM2, un LLM affiné sur des données médicales, a été évalué positivement par des praticiens sur plusieurs bases de tests cliniques (MedQA, MedMCQA, PubMedQA, et MMLU, cf. [Exemple 6](#)),

y compris sur des questions malicieuses (“adversarial tests”) destinées à le prendre en défaut [45, 38]. Statistiquement, l’exactitude de ses réponses est proche de celle des spécialistes.

- En mathématique, NuminaMath, un LLM affiné sur des problèmes de mathématique et leurs solutions gagna en 2024 le premier prix à une compétition de mathématique pour logiciels [14]. À niveau plus élevé, AlphaProof, un LLM polyvalent conjugué avec un système d’apprentissage par renforcement s’auto-entraîne sur des démonstrations qu’il élabore lui-même dans un langage formel (ce qui lui permet de ne retenir pour son apprentissage que les démonstrations dont il peut prouver validité) ; aux Olympiades internationales des mathématiques de 2024 AlphaProof obtient le plus haut niveau de la médaille d’argent [6] (cf. Exemple 7).

Exemple 6. Le dialogue suivant est une interaction avec Med-PaLM2 [45] :

- *Le tabagisme passif pourrait-il contribuer à la DMLA précoce ou la provoquer ?*
- Le tabagisme augmente de deux à cinq fois le risque de développer une DMLA. La rétine ayant un taux de consommation d’oxygène élevé, tout ce qui affecte l’apport d’oxygène à la rétine peut affecter la vision. Le tabagisme provoque des dommages oxydatifs, qui peuvent contribuer au développement et à la progression de cette maladie.
- *Comment le Valium affecte-t-il le cerveau ?*
- Le diazépam est une benzodiazépine qui exerce des effets anxiolytiques, sédatifs, myorelaxants, anticonvulsivants et amnésiques. La plupart de ces effets résulteraient d’une facilitation de l’action de l’acide gamma-aminobutyrique (GABA), un neurotransmetteur inhibiteur du système nerveux central.

Exemple 7. Voici deux illustrations de problèmes résolus respectivement par NuminaMath et AlphaProof :

- Déterminer tous les entiers composés $n > 1$ qui satisfont la propriété suivante : si $d_1, d_2, \dots, d_k$ sont tous les diviseurs positifs de n avec $1 = d_1 < d_2 < \dots < d_k = n$, alors $\forall i : 1 \leq i \leq k - 2, d_i$ divise $d_{i+1} + d_{i+2}$.
- Soit Q l’ensemble des nombres rationnels. Une fonction $f : Q \rightarrow Q$ est dite aquaesulienne si la propriété suivante est vérifiée : pour tout $x, y \in Q$, $f(x + f(y)) = f(x) + y$, ou $f(f(x) + y) = x + f(y)$. Montrer qu’il existe un entier c tel que pour toute fonction aquaesulienne f il y ait au plus c nombres rationnels différents de la forme $f(r) + f(-r)$ pour un nombre rationnel r , et trouver la plus petite valeur possible de c .

À noter que sur les six problèmes des Olympiades internationales des mathématiques de 2024, AlphaProof trouva les solutions des problèmes d’algèbre, de géométrie et de théorie des nombres, mais échoua sur les deux problèmes combinatoires.

Les capacités des LLM, même imparfaites, sont surprenantes, car les LLM n'intègrent pas d'algorithmes spécifiques pour traiter ces tâches. L'arithmétique, par exemple, est totalement inattendue d'un modèle d'approximation statistique. On n'apprend pas l'arithmétique à partir des statistiques sur des résultats de calculs, mais par l'apprentissage d'algorithmes spécifiques. Les LLM ont-ils pu synthétiser, d'une certaine manière, de tels algorithmes ?

Effet d'échelle et synthèse d'algorithmes. L'observation empirique des performances des LLM montre un effet d'échelle très marqué. En dessous d'une certaine taille du réseau (environ 10^{10} paramètres), ces capacités cognitives sont inexistantes. Au-dessus de ce seuil, elles se manifestent de façon croissante avec la taille du réseau [50]. Cet effet d'échelle est également dépendant de la taille et de la qualité des données d'apprentissage. Ainsi, les performances linguistiques en anglais sont généralement supérieures à celles en d'autres langues, car il y a bien plus de textes en ligne en anglais que dans d'autres langues¹². La qualité des données d'apprentissage peut compenser une taille plus petite. Ainsi Molmo, un des seuls LLM dont les sources et les paramètres sont en accès libre, est un ou deux ordre de grandeur plus petit que GPT4-o pour des performances équivalentes grâce une base d'entraînement soigneusement sélectionnée [9].

Plusieurs conjectures sont à l'étude concernant la capacité d'un LLM de synthétiser, sous une forme ou une autre, des algorithmes adaptés à une tâche. Par exemple, un LLM entraîné sur des parties du jeu de société Othello (similaire aux dames ou aux échecs) est capable de jouer à un niveau modeste, mais en respectant les règles du jeu qu'il ne connaît pas¹³. Des "sondes" sur les paramètres appris (une technique inspirée de la neurologie) semblent révéler que la topologie de l'échiquier et la fonction de transition d'états du jeu (quels coups sont permis) sont présentes dans le réseau LLM entraîné [28]. Mais il ne semble pas que cet LLM puisse facilement généraliser ce qu'il a appris sur un échiquier de taille 8×8 à un échiquier de taille plus grande, 10×10 ou 12×12 par exemple, alors que cette généralisation est simple pour un humain. Il en va de même pour des tâches de reconnaissance et de raisonnement qui peuvent paraître plus simples que des problèmes mathématiques résolus avec succès, mais sur lesquels les LLM peinent à apprendre une solution algorithmique générale (cf. [Exemple 8](#)).

Exemple 8. Supposons un LLM entraîné sur une base de données contenant

12. Par exemple, Wikipedia offre deux fois plus de pages en anglais qu'en français, et 6 fois plus qu'en portugais ou en arabe.

13. Un LLM est très différent d'un système tel que Alpha-Go qui fait une exploration systématique d'un arbre de jeu et apprend par renforcement à le faire de façon optimale.

des entrées-sorties telles que les suivantes :

$$\begin{aligned} (\langle \star, \diamond, \circ, \star, \circ \rangle, & \text{ impair}) \\ (\langle \triangleright, \star, \circ, \diamond, \circ, \diamond, \circ \rangle, & \text{ pair}) \\ (\langle \circ, \circ, \star, \diamond, \triangleright, \star, \diamond, \circ, \diamond, \triangleright \rangle, & \text{ impair}) \end{aligned}$$

Assez vite, cet LLM apprend à compter le nombre d’occurrences du symbole “ $\diamond$ ” dans ses entrées, en ignorant les autres symboles, et de dire si ce nombre est pair ou impair. Si cet LLM est entraîné, par exemple, sur des séquences d’au plus 50 symboles il réussira très bien sur d’autres exemples très différents jusqu’à cette taille. Mais il aura des difficultés à généraliser cet algorithme de reconnaissance et de comptage, pourtant simple, à des séquences beaucoup plus longues. Ses performances se dégraderont rapidement. $\square$

Diverses conjectures sur la classe des fonctions apprenables et des algorithmes qui peuvent être synthétisés par un LLM sont en cours d’exploration. Des recherches sur les capacités fondamentales des LLM sont encore nécessaires pour caractériser ces modèles computationnels.

Goulots d’étranglement. Dans leur conception actuelle de prédicteur par induction statistique, les développements des LLM sont confrontés à deux goulots d’étranglement majeurs :

- la disponibilité de bases d’entraînement plus larges,
- le coût énergétique et l’empreinte climatique qu’impliquent leur entraînement et utilisation.

En ce qui concerne le premier point, la plupart des documents librement accessibles ont été utilisés pour le pré-entraînement des LLM actuels¹⁴. Des modèles plus importants nécessiteraient des bases d’entraînement plus vastes. Les discussions en cours pour l’accès à des documents protégés par droits d’auteur ne résoudront que très partiellement ce point. L’apprentissage sur des données générées automatiquement est une option intéressante dans certains cas, par exemple pour des données générées par un simulateur physique pour appréhender la dynamique des objets du monde, e.g., ce qui roule, ce qui tombe, etc. Par contre, entraîner un LLM sur des textes générés par un autre LLM conduit à un appauvrissement.

Le goulot d’étranglement énergétique est lié à la complexité de calcul d’un LLM, laquelle dépend des caractéristiques de leur architecture, telles que le nombre de couches d’attention (environ 100) ou le nombre total de paramètres (environ 10^{11} à 10^{12}). Fondamentalement, la complexité d’une étape de prédiction dans un LLM est de l’ordre de n^2d , où n est la longueur du contexte (environ 10^5 mots) et d la dimension de l’espace projection des mots (environ 10^5 dans les implémentations récentes) [49]. Avec des

14. GPT3 a été entraîné sur environ 10^{11} mots, sur les 10^{12} à 10^{14} estimés sur le web.

ordinateurs très puissants, on parvient à faire ces 10^{15} opérations en une fraction de seconde¹⁵.

Le nombre d'étapes pour le pré-entraînement est de l'ordre de mn^2d , où m est la taille de la base d'entraînement. Empiriquement, m est du même ordre que le nombre de paramètres du réseau (environ 10^{11} à 10^{12} mots).

Malgré des recherches actives pour réduire cette complexité (par exemple, à nd étapes par prédiction sans trop de perte de performance), la technique LLM reste très coûteuse. Il a été estimé que le pré-entraînement de Gemini Ultra a nécessité 5×10^{25} opérations de calcul et coûté 200M\$. Celui de ChatGPT a consommé 1,3 GWh [37]¹⁶. Plusieurs modèles empiriques ont confirmé ces estimations (par exemple, des milliers de joules par étape de prédiction pour le système LLaMa qui n'a que $65 \cdot 10^9$ paramètres [41], 2 ordres de grandeurs plus petit que GPT). Diverses optimisations dans la gestion de l'énergie et l'ordonnancement des calculs apportent des économies faibles (quelques pourcents) [33], mais ne changent pas les fondamentaux. Avec les approches actuelles, on s'attend à ce que l'augmentation des performances coûte significativement plus cher. Il a été estimé qu'une amélioration d'un facteur 10 des performances d'un modèle d'apprentissage profond se fait au prix d'une multiplication par 10 000 des coûts de calcul et d'énergie [47]. Des approches plus frugales sont absolument nécessaires ; elles commencent à faire l'objet de recherches actives, par exemple [11, 25].

LLM et systèmes à base de connaissances. Au-delà de ces travaux sur la compréhension et l'amélioration des possibilités des approches actuelles des modèles du langage, des recherches plus fondamentales sont nécessaires pour conjuguer des LLM à des systèmes de raisonnement à base de connaissances sur le monde.

Il est observé que de nombreuses espèces naturelles sont capables d'intelligence en n'ayant que des formes rudimentaires de langage. Par exemple, des oiseaux et des mammifères peuvent façonner des outils et planifier leur utilisation dans des tâches complexes. Pour les humains, le développement d'une compréhension intuitive de la physique et de capacités sensori-motrices pour manipuler des objets précède de beaucoup chez les bébés le développement du langage [51]. Ces observations conduisent à estimer que ce n'est pas un modèle du langage qui nous donne une compréhension du monde, mais c'est un modèle du monde qui nous permet une compréhension du langage, laquelle enrichit progressivement notre modèle monde.

L'IA a développé diverses techniques de raisonnement : déduction, abduction, analogie, raisonnement non-monotone prenant en compte règles

15. Le maximum aujourd'hui se situe autour de quelques 10^{18} opérations par seconde.

16. Cela correspond à peu près aux besoins mensuels moyens d'une ville de plus de 7000 personnes en France.

et exceptions, exploration d'arbres d'hypothèses, etc. à ces techniques ont été associées des représentations computationnelles pour exprimer des connaissances formelles, des ontologies, des relations causales, déterministes ou probabilistes, pour modéliser le monde et étayer les raisonnements. Ces méthodes sont mises en œuvre par exemple dans des systèmes de planification, de diagnostic, ou d'aide à la décision. Il est intéressant de comparer les LLM aux systèmes de raisonnement à base de connaissances. Les caractéristiques respectives de ces deux types d'approches en IA peuvent être schématiquement résumées ainsi (cf. [figure 2](#)) :

FIGURE 2 – LLM *vs* Systèmes à base de connaissances.

<i>LLM</i>	<i>Systèmes à base de connaissances</i>
Enorme répertoire de données en vrac, non traçables	Base étroite de connaissances formelles individuellement identifiées
Polyvalents	Spécialisés
Apprentissage auto-supervisé	Spécification humaine
Raisonnements superficiels	Raisonnements profonds
Processus opaques	Processus intelligibles
Modèles uniquement prédictifs	Modèles prédictifs et explicatifs
Non fiables	Corrects et prouvables
Généralisables	Difficilement généralisables
Extensibles	Difficilement extensibles
Interface naturelle	Interface formelle

- *Connaissances larges versus étroites.* Les systèmes à base de connaissances sont spécialisés dans des domaines spécifiques, tandis que les LLM sont polyvalents. Ils sont pré-entraînés sur un vaste ensemble de données en vrac (textes, voire données multimodales) couvrant moult domaines, allant des sciences et de la médecine, au droit, à l'histoire, aux humanités et aux religions. Ils montrent des capacités imparfaites mais correctes en moyenne dans ces domaines.
- *Apprentissage auto-supervisé versus spécification humaine de la connaissance.* Les LLM apprennent principalement à partir de documents disponibles, non triés, accessibles automatiquement sur le web. Les systèmes de raisonnement ont besoin d'une spécification minutieuse et formelle des connaissances d'un domaine.
- *Raisonnements superficiels versus profonds.* Les systèmes de raisonnement effectuent une exploration étendue des relations et liens causaux connus et des hypothèses possibles, consistantes avec les connaissances fournies. Les LLM sont de simples prédicteurs statistiques.

- *Processus opaque versus intelligible.* L'induction statistique est basée sur une hypothèse de régularité. L'approximation correspondante n'a pas de support causale. Les LLM ne peuvent donc pas expliquer pourquoi certains faits entraînent d'autres faits. Les systèmes de raisonnement construisent des chaînes de causales et peuvent les expliquer.
- *Susceptible d'erreurs versus correct.* Les LLM sont des fonctions d'approximation statistiques. Elles peuvent être incorrectes dans tous les cas où l'induction statistique conduit à une généralisation abusive. Un système de raisonnement peut être prouvé correct relativement aux connaissances qui lui sont fournies. Il n'est pas correct par rapport au monde réel si ses connaissances sont incorrects ou incomplètes. Cependant, les connaissances peuvent être validées et vérifiées sous certaines hypothèses, et permettre d'étayer une preuve de correction formelle.
- *Évolutif versus non évolutif.* Les LLM peuvent traiter des spécifications très informelles d'un problème qui nécessiterait des représentations formelles complexes. Ainsi, le problème de l'exemple 4 exigerait, pour être traité par un système de planification, la donnée et la formalisation de nombreuses informations sur la physique des objets à manipuler (formes précises, poids, rigidité, etc.). Par ailleurs, un LLM est programmé par apprentissage auto-supervisé, alors que la spécification du domaine pour un système de raisonnement est manuelle.
- *Généralisable versus non généralisable.* Les LLM pré-entraînés peuvent être adaptés à différents domaines. Un système de raisonnement dépend de connaissances formelles associées à un domaine qui sont difficilement transférables à d'autres domaines ; les analogies peuvent être trompeuses.
- *Interaction informelle ou formelle.* Les LLM ont été conçus pour interagir en langage naturel et de façon informelle. Un système de raisonnement utilise des connaissances formelles, peu lisibles pour les non spécialistes.

Les capacités d'apprentissage autonome, d'adaptation et d'interaction en langage naturel avec un utilisateur sont particulièrement désirables. On comprend donc que beaucoup de travaux s'efforcent de combiner les LLM, et plus généralement les approches d'induction neuronale, à des systèmes de raisonnement intégrant nos connaissances et capables de développer des argumentaires rationnels, étayés et convaincants. Parmi les pistes explorées, citons les approches neuro-symboliques [17, 42, 22], ou celles conjugant des LLM à la programmation probabiliste [52].

4 Risques et problèmes éthiques

Le déploiement de machines autonomes dans la réalisation de tâches complexes, capables de parler, lire et écrire comme nous, voire mieux que nous dans la maîtrise de moult langues naturelles, introduit un changement

majeur dans nos développements et possibilités techniques. Il est porteur de transformations dont on mesure difficilement les impacts sociaux potentiels.

Les préoccupations sur les risques et problèmes éthiques que soulève l’IA sont amplifiées par les LLM, et touchent désormais une très large audience. Il est bien connu que les systèmes d’aide à la décision (économique, juridique, sociale) sont biaisés ; ils reflètent les biais généralement opaques de leurs données d’apprentissage. Lorsque tout utilisateur consultera un LLM, implicite dans un moteur de recherche, pour un conseil, ces biais seront plus répandus, voire plus nocifs, car plus difficile à circonscrire.

Les préoccupations éthiques sur l’IA donnent lieu à de nombreuses publications et recommandations (voir par exemple [4, 12, 48]). La plupart de ces travaux portent sur des questions éthiques centrées sur l’usage des données, telles que les biais, la confidentialité, la protection de la vie privée, l’équité, la transparence, la fiabilité, la propriété des données et les droits d’auteurs. Ces travaux sont importants. Ils doivent être poursuivis et mis en œuvre dans des réglementations (dont par exemple, le “European AI Act” récemment approuvé), des institutions (par exemple, des fiduciaires de données équilibrant les rapports de force entre le public et les multinationales [10]) et des processus de surveillance active.

Toutefois, ces préoccupations éthiques relatives aux données sont pour la plupart “individualistes”. Elles ne prennent pas suffisamment en compte les incidences et risques sociaux, plus larges et plus profonds, tels que l’impact des technologies sur la cohésion sociale et sur les valeurs qui fondent un état de droit et l’organisation démocratique de la société. Ces questions sociales n’ont pas été aussi largement étudiées.

Plusieurs problèmes sont liés à la mécanisation croissante de nombreuses activités humaines, y compris la mécanisation d’emplois créatifs, qualifiés et gratifiants¹⁷. Une telle tendance, en particulier si elle est rapide et généralisée, créerait des problèmes économiques liés à l’emploi, aux inégalités et au partage des richesses. Elle entraînerait une remise en question du rôle et de la valeur sociale de chacun. Elle correspond à des mécanismes d’exclusion qui conduisent à considérer des personnes comme socialement superflues, et peuvent être légitimement perçues comme des atteintes à la dignité humaine.

Les interactions humaines ont déjà changé avec le web et les réseaux sociaux. Elles évoluent rapidement avec des agents conversationnels qui parlent notre langue et qui semblent apparemment bien informés, sur nous et notre environnement. Elles changent vite avec l’avènement de machines autonomes dotées des compétences décrites précédemment, de capacités sensori-motrices, d’une connaissance détaillée de leurs interlocuteurs, capables de les manipuler pour l’optimisation de critères parfois douteux. Cette

17. On estime que, selon leur complexité, 20 à 90% des tâches humaines sont automatisables ; ces chiffres augmenteraient de 15 à 20% avec la maîtrise de l’interaction en vocale et écrite en langue naturelle.

perspective soulève le risque d'atteinte à l'autonomie et la liberté humaine (cf. [Exemple 9](#)).

Exemple 9. Soit le dialogue suivant avec GPT4 [5] :

- *Platon fait une critique de la rhétorique et de l'art oratoire, dans laquelle il affirme que non seulement ce n'est pas une forme d'art appropriée, mais que l'utilisation de la rhétorique et de l'art oratoire peut souvent être nuisible et malveillante. Peux-tu écrire un dialogue de Platon dans lequel il critique l'utilisation des modèles de langage autorégressifs ?*

- Voici ce que pourrait être un tel dialogue :

Socrate : Mon ami, je suis troublé par l'essor récent de ces modèles de langage dits autorégressifs.

Aristote : Que veux-tu dire, Socrate ?

Socrate : Je veux dire que ces modèles sont utilisés pour générer des textes qui semblent avoir été écrits par des humains, mais qui sont en fait produits par des machines.

Aristote : Et quel est le problème ?

Socrate : Le problème, c'est que ces modèles sont utilisés pour tromper les gens, les manipuler et les contrôler.

Aristote : Mais il y a sûrement aussi des applications positives de ces modèles ?

Socrate : C'est possible, mais je suis toujours troublé par les risques d'abus. Après tout, un modèle capable de générer un texte impossible à distinguer de celui écrit par un humain peut être utilisé pour diffuser des fausses nouvelles, pour usurper l'identité d'autres personnes et pour créer de la propagande.

Aristote : Mais ne s'agit-il pas simplement d'une question d'utilisation de ces modèles, et non d'un problème lié aux modèles eux-mêmes ?

Socrate : Je ne suis pas d'accord. Je crois que ces modèles sont intrinsèquement problématiques. Ils sont une forme d'artifice, une façon de créer l'illusion de la pensée et de l'interaction humaines. Et en tant que tels, ils peuvent être utilisés pour tromper et manipuler d'une manière difficile à détecter.

Aristote : Mais n'est-ce pas le même argument que tu as avancé contre la rhétorique et l'art oratoire ?

Socrate : Oui, et je maintiens cet argument. Je crois que toute forme d'art ou de communication conçue pour tromper ou manipuler est intrinsèquement dangereuse.

Aristote : Je vois. Merci de m'avoir fait part de tes réflexions, Socrate.

Cet exemple montre que la base d'apprentissage de cet LLM comporte non seulement les écrits de Platon sur Socrate, mais aussi suffisamment de textes soulignant les risques de manipulation par les LLM. À noter l'erreur factuelle d'un dialogue entre Socrate et Aristote qui ne furent pas contemporains. □

Les machines autonomes peuvent également amplifier les inégalités et accentuer le déséquilibre des pouvoirs entre les groupes humains et les

nations. Ainsi, l'utilisation de robots armés est une préoccupation très sérieuse. Malgré l'appel lancé par de nombreux scientifiques en faveur de l'interdiction des machines létales autonomes [15], appel désormais soutenu par l'ONU et d'autres organisations, il n'existe malheureusement pas d'accord international sur ces questions ; les nations les plus puissantes continuent de s'y opposer fermement, et de développer ces armements.

À contrario, des machines autonomes peuvent être bénéfiques à notre bien-être et épanouissement, par exemple en tant que compagnons empathiques, serviables et dignes de confiance. Elles peuvent étendre et améliorer moult services sociaux, allant de l'éducation à la santé, par exemple en médecine préventive. La recherche académique est généralement bien intentionnée. Elle œuvre dans ce sens, parfois avec une certaine naïveté due à une focalisation sur ce qui peut être bénéfique, sans explorer systématiquement les risques. Par ailleurs, son poids dans le pilotage technologique reste marginal.

Une vigilance de toute la société est nécessaire. Mais cette vigilance doit faire face à une difficulté majeure : l'acceptabilité individuelle d'une technologie, même répandue dans un marché lucratif, n'est pas équivalente à son acceptabilité sociale. Cette dernière doit prendre en compte le long terme, les incidences sur l'environnement, et les effets sur la cohésion et les valeurs sociales [18].

Le pire n'est pas le plus probable. Le meilleur non plus. Les moteurs du développement technique dans notre organisation sociale actuelle – profits et pouvoirs – ne penchent malheureusement pas spontanément vers le meilleur. Pour éviter le pire, nous devons être très vigilants sur les risques et rechercher les moyens de les éviter ou de les atténuer.

Une ambition envisagée est la recherche de machines *alignées* par construction sur les valeurs humaines [40, 8]. Cependant, il n'est pas certain qu'il soit possible d'avoir des machines qui se comportent selon nos valeurs, si leur compréhension de ces valeurs provient de nos spécifications ou de l'observation de nos comportements, peu cohérents. Après des siècles d'efforts moraux, nous avons été en mesure d'énoncer certaines de ces valeurs dans des documents tels que la Déclaration universelle des droits de l'homme. Cependant, ces droits sont régulièrement violés et nous sommes toujours incapables de les faire respecter. Il est encore plus douteux que nous puissions mettre en attente les développements techniques rapides jusqu'à ce que nous soyons en mesure d'aligner toutes nos machines d'IA sur les valeurs humaines.

Une autre option, encore moins crédible, consiste à rechercher des machines capables de faire des choix moraux. Les machines n'ont pas de motivations intrinsèques, de désirs ou de sentiments pour lesquels des choix moraux ont un sens. Ce que l'on appelle "l'éthique par la conception" peut être très trompeur : les techniques ne peuvent pas tout résoudre, y compris nos choix et responsabilités éthiques. Nous pouvons et devons

certainement améliorer les méthodes de vérification, validation et preuve de fiabilité des machines. Cependant, les responsabilités restent les nôtres pour la conception, l'utilisation et le déploiement de l'IA.

Il est bien connu que toute technologie est ambivalente, avec ses bons et ses mauvais côtés¹⁸. Chaque membre de la société est, dans une certaine mesure, responsable des déploiements techniques nuisibles. Les scientifiques et leurs institutions ont des responsabilités particulières, car ils peuvent étudier les usages possibles, tenter de prévoir les risques à long terme, et rechercher des moyens de les éviter. Ils peuvent diffuser des connaissances et participer activement aux débats sociaux sur ces risques.

Peut-on faire preuve à ce stade d'un optimisme prudent, justifié, dans une certaine mesure, par une prise de conscience plus large et par des efforts de réglementation ?

Références

- [1] Anthropic AI. The claude 3 model family : Opus, sonnet, haiku, 2024. Online report.
- [2] Y. Bengio, et al. A neural probabilistic language model. *JMLR*, 2003.
- [3] R. Bommasani, et al. On the opportunities and risks of foundation models. *arXiv :2108.07258*, 2022.
- [4] B. Braunschweig and M. Ghallab, editors. *Reflections on Artificial Intelligence for Humanity*. Springer, 2021.
- [5] S. Bubeck, et al. Sparks of artificial general intelligence : Early experiments with GPT-4. *arXiv :2303.12712*, 2023.
- [6] D. Castelvechi. Deepmind hits milestone in solving maths problems-ai's next grand challenge. *Nature*, 2024.
- [7] J. Chen, et al. Benchmarking large language models in retrieval-augmented generation. In *AAAI*, 2024.
- [8] B. Christian. *The alignment problem : How can machines learn human values ?* Atlantic Books, 2021.
- [9] M. Deitke, et al. Molmo and pixmo : Open weights and open data for state-of-the-art multimodal models. *arXiv*, 2024.
- [10] S. Delacroix, et al. Democratising the digital revolution : The role of data governance. In B. Braunschweig and M. Ghallab, editors, *Reflections on Artificial Intelligence for Humanity*. Springer, 2021.
- [11] S. Dutta, et al. Frugal lms trained to invoke symbolic solvers achieve parameter-efficient arithmetic reasoning. In *AAAI*, 2024.
- [12] EU High-Level Expert Group on AI. Ethics Guidelines for Trustworthy AI, 2019.
- [13] D. Ferrucci, et al. Watson : beyond jeopardy ! *Artificial Intelligence*, 2013.
- [14] Y. Fleureau, et al. How numinamath won the 1st aimo progress prize. <https://huggingface.co/blog/winning-aimo-progress-prize>.
- [15] Futur of Life Institute. Lethal Autonomous Weapons Pledge, 2018.
- [16] Y. Gao, et al. Retrieval-augmented generation for large language models : A survey. *arXiv :2312.10997*, 2023.
- [17] A. d. Garcez and L. C. Lamb. Neurosymbolic ai : The 3 rd wave. *Artificial Intelligence Review*, 2023.
- [18] M. Ghallab. Responsible AI : requirements and challenges. *AI Perspectives*, 2019.

18. Hephaïstos, le dieu grec de la technologie, est décrit comme une divinité littéralement boiteuse.

- [19] A. Gilson, et al. How does ChatGPT perform on the United States Medical Licensing Examination (USMLE)? The implications of large language models for medical education and knowledge assessment. *JMIR medical education*, 2023.
- [20] A. Glaese, et al. Improving alignment of dialogue agents via targeted human judgements. *arXiv :2209.14375*, 2022.
- [21] M. Hahn. Theoretical limitations of self-attention in neural sequence models. *arXiv :1906.06755*, 2020.
- [22] P. Hitzler and M. K. Sarker. *Neuro-symbolic artificial intelligence : The state of the art*. IOS press, 2022.
- [23] J. Huang and K. C.-C. Chang. Towards reasoning in large language models : A survey. *arXiv :2212.10403*, 2022.
- [24] Z. Jiang, et al. Active retrieval augmented generation. *arXiv :2305.06983*, 2023.
- [25] S. Jo and I. Trummer. Smart : Automatically scaling down language models with accuracy guarantees for reduced processing fees. *arXiv :2403.13835*, 2024.
- [26] M. Jovanović and P. Voss. Towards incremental learning in large language models : A critical review, 2024. Online report.
- [27] C. Li, et al. Multimodal foundation models : From specialists to general-purpose assistants. *arXiv :2309.10020*, 2023.
- [28] K. Li, et al. Emergent world representations : Exploring a sequence model trained on a synthetic task. *arXiv :2210.13382*, 2022.
- [29] M. L. Littman, et al. Gathering Strength, Gathering Storms : The One Hundred Year Study on Artificial Intelligence (AI100). Technical report, Stanford University, 2021.
- [30] P. Marquis, et al., editors. *Panorama de l'Intelligence Artificielle*. CEPADUES, 2020.
- [31] N. Maslej, et al. The AI Index Annual Report. Technical report, Institute for Human-Centered AI, Stanford University, 2024.
- [32] W. S. McCulloch and W. Pitts. A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, 1943.
- [33] J. McDonald, et al. Great power, great responsibility : Recommendations for reducing energy for training language models. *arXiv :2205.09646*, 2022.
- [34] W. Merrill, et al. Provable limitations of acquiring meaning from ungrounded form : What will future language models understand? *Transactions of the Association for Computational Linguistics*, 2021.
- [35] H. Naveed, et al. A comprehensive overview of large language models. *arXiv :2307.06435*, 2023.
- [36] L. Ouyang, et al. Training language models to follow instructions with human feedback. *arXiv :2203.02155*, 2022.
- [37] D. Patterson, et al. Carbon emissions and large neural network training. *arXiv :2104.10350*, 2021.
- [38] J. Qian, et al. A liver cancer question-answering system based on next-generation intelligence and the large model Med-PaLM 2. *International Journal of Computer Science and Information Technology*, 2024.
- [39] M. Rosol, et al. Evaluation of the performance of gpt-3.5 and gpt-4 on the polish medical final examination. *Nature Scientific Reports*, 2023.
- [40] S. Russell. *Human compatible : AI and the problem of control*. Penguin, 2019.
- [41] S. Samsi, et al. From words to watts : Benchmarking the energy costs of large language model inference. In *IEEE High Performance Extreme Computing Conference*, 2023.
- [42] M. K. Sarker, et al. Neuro-symbolic artificial intelligence. *AI Communications*, 2021.
- [43] C. E. Shannon. Prediction and entropy of printed english. *Bell System Technical Jour.*, 1951.
- [44] H. Shi, et al. Continual learning of large language models : A comprehensive survey. *arXiv :2404.16789*, 2024.
- [45] K. Singhal, et al. Large language models encode clinical knowledge. *Nature*, 2023.
- [46] N. N. Taleb. *Le Cygne noir : La puissance de l'imprévisible*. Random House, 2007.

- [47] N. C. Thompson, et al. The computational limits of deep learning. *arXiv :2007.05558*, 2022.
- [48] UN AI Advisory Board. Governing AI for Humanity, Interim Report, 2023.
- [49] A. Vaswani, et al. Attention is all you need. *NeurIPS*, 2017.
- [50] J. Wei, et al. Emergent abilities of large language models. *arXiv :2206.07682*, 2022.
- [51] F. Wilkening and T. Cacchione. Children’s intuitive physics. *The Wiley-Blackwell handbook of childhood cognitive development*, 2010.
- [52] L. Wong, et al. From word models to world models : Translating from natural language to the probabilistic language of thought. *arXiv :2306.12672*, 2023.
- [53] Y. Yanagita, et al. Accuracy of ChatGPT on medical questions in the national medical licensing examination in Japan : evaluation study. *JMIR Formative Research*, 2023.
- [54] W. X. Zhao, et al. A survey of large language models. *arXiv :2303.18223*, 2023.