# GPT ET LES GRANDS MODÈLES DU LANGAGE EN INTELLIGENCE ARTIFICIELLE : principes et défis

Par Malik GHALLAB[1]

## Introduction

L'intelligence artificielle (IA) est une discipline scientifique qui cherche à comprendre et à modéliser l'intelligence naturelle par des approches computationnelles ou algorithmiques. L'intelligence est multiforme et difficile à définir, mais on peut la constater dans la réalisation de tâches. Ainsi, la démarche de l'IA consiste 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, ou d'apprendre à compter, lire, écrire, ou résoudre des problèmes pratiques. Pour ce faire, on modélise la tâche en question, on cherche des méthodes et des algorithmes pour la résoudre, on développe des implémentations logicielles et matérielles de ces algorithmes, puis on évalue empiriquement les capacités des modèles et algorithmes et leurs performances pour cette tâche, et on s'efforce de les améliorer.

L'IA relève des sciences de la modélisation et de l'information. C'est un domaine relativement récent, même si ses racines sont anciennes. Elle a d'ores et déjà transformé nos capacités de modélisation, d'analyse, d'organisation et de recherche d'informations, avec un impact considérable sur pratiquement tous les champs d'investigation scientifiques et techniques, de l'astronomie à la santé et aux sciences humaines. Comme toute technoscience, le développement de l'IA est motivé par des applications que l'on espère socialement utiles. Mais les investissements de R&D en IA, aujourd'hui considérables, sont dominés par quelques multinationales, qui contrôlent ainsi l'évolution du domaine selon des logiques capitalistes plutôt sociales.

Depuis ses débuts vers le milieu du XX$^{e}$ 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, difficiles à généraliser à de nouvelles tâches. Pendant de nombreuses années, les systèmes d'apprentissage d'IA manquaient de capacité de

1. Résumé d'une communication présentée à l'Académie des Sciences, Inscriptions et Belles-Lettres de Toulouse à la séance du 16 novembre 2023. Cette communication est développée dans un article documenté et illustré disponible en ligne : https://www.academie-sciences-lettres-toulouse.fr/fonds_documentaire/activites/publications_des_membres/Ghallab/MGhallab2024.pdf

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 récents sur les réseaux de neurones.

### *Réseaux de neurones artificiels*

Un neurone artificiel est une fonction multivariable relativement simple qui fournit en sortie la somme pondérée de ses entrées selon un seuil. L'apprentissage consiste à estimer les valeurs des paramètres de cette fonction qui permettent de se rapprocher le plus de valeurs souhaitées. 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 selon une organisation ou architecture particulière. L'apprentissage dans un réseau se fait par un algorithme qui, à chaque exemple d'entraînement, adapte les paramètres du réseau pour se rapprocher des valeurs finales désirées.

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, n'ont que récemment pu remplir ces conditions, grâce à internet pour la quantité de données disponibles, à la puissance croissante des processeurs électroniques, aux progrès algorithmiques pour l'estimation des paramètres, et aux architectures des réseaux.

### *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. 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 fonctions caractéristiques adaptées aux données d'apprentissage. Les méthodes d'interprétation des données ne sont plus coûteuses et spécialisées pour chaque application. Elles sont largement déployées pour l'analyse de données multimodales: signaux physiques, images, sons, vidéos, textes.

• *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 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 des logiciels dits « grands modèles de langage » (LLM pour *Large Language Models*). 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.

## 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 un problème de prédiction classique : ayant une séquence de termes, on veut prédire le terme probable suivant. Lorsque cette prédiction porte sur les états d'un système dont l'évolution est bien modélisée, on aborde le problème grâce au modèle, 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.

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. La prédiction statistique suppose que le domaine est régulier. Elle s'appuie sur l'hypothèse de *l'induction statistique* : on suppose que ce qui est vrai pour les données observées reste vrai pour celles non observées. Cette hypothèse 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, face à un ensemble d'instances possibles potentiellement infini.

Les données d'entraînement pour le traitement du langage naturel sont très abondantes sur le web. 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, qui souligne que « *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* ». Mais les méthodes de Shannon n'informent pas sur la sémantique des mots. Deux mots sont susceptibles d'être sémantiquement proches s'ils sont utilisés de façon interchangeable dans les mêmes contextes. Le calcul à partir de la probabilité de trouver un mot dans un contexte donné se heurte à un problème de complexité : la table des probabilités conditionnelles est de dimension exponentielle en la taille du dictionnaire.

Ce problème de complexité fut résolu avec une technique dite de *projection sémantique des mots* (« *word embedding* »). Cette projection associe à chaque mot un vecteur de quelques milliers de nombres, de telle sorte que des mots sémantiquement similaires aient des projections proches. Des calculs vectoriels simples permettent des opérations d'analogie et de proximité sur les mots. Ces projections de mots peuvent

être calculées par des réseaux neuronaux et utilisées dans le traitement du langage par d'autres réseaux. Le traitement de phrases est ramené au traitement d'une séquence de vecteurs, avec une difficulté liée aux ambiguïtés du langage, telles que les co-références et appariements de pronoms.

Ces ambiguïtés ont été résolues par des mécanismes dits *d'attention*, capables de relier des mots situés en des positions éloignées dans la phrase. On utilise une architecture de réseau dite « transformeur ». Les 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 dit auto-supervisé sur les 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 soit proche du 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 les préférences humaines (la fonction de renforcement) et l'utiliser ensuite en auto-apprentissage. En outre, un LLM peut être adapté à des applications ou des domaines spécifiques, par exemple, la médecine pour le système Med-PaLM2. Enfin, la plupart des LMM 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 traitement.

Il existe à ce jour plusieurs centaines de LLM. Les plus conséquents sont qualifiés de « *foundation models* ». Ils sont généralement multimodaux (textes, images, sons) et permettent de développer, une fois pré-entraînés, des modèles plus spécialisés. La plupart de ces logiciels viennent de l'industrie, principalement des multinationales du numérique.

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). Leurs techniques algorithmiques sont limitées, par exemple, il n'y a pas d'itération, de récursion ou de recherche arborescente avec essais et erreurs. 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 l'estimation des paramètres lors de l'apprentissage et la prédiction par le calcul des fonctions de chaque neurone, selon la topologie du réseau.

### *Limitations et performances des LLM*

Les LLM sont confrontés à plusieurs limitations, théoriques et opérationnelles. Un LLM est non factuel, il peut faire des erreurs grossières. Il a du mal à étayer ses propos par des sources convaincantes ou à en vérifier la rationalité. L'interface de dialogue d'un LLM conduit à l'interroger comme un oracle capable de répondre à tout, y compris à des questions qui ne relèvent pas de ce qu'il peut calculer. On dit parfois qu'un LLM

« hallucine ». Mais il faut garder en mémoire qu'il s'agit d'une fonction d'approximation statistique, et non une requête dans une base de données fiable.

Malgré les limitations théoriques et pratiques des LMM, leurs performances moyennes sont bonnes à excellentes dans pratiquement toutes les tâches de traitement du langage, telles que la traduction, la synthèse, l'analyse, la 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, d'humour ou d'interprétation de proverbes.

Au-delà du traitement du langage naturel pour lequel ils ont été conçus, les LLM démontrent des capacités limitées dans un large éventail de tâches cognitives qui vont du calcul arithmétique au raisonnement logique ou bon sens, en passant par la planification, le diagnostic et la résolution de problèmes mathématiques. Les LLM ont été testés sur de nombreux examens universitaires. Ils ont démontré de bonnes performances dans moult tests, y compris pour des examens réputés très difficiles, tels que ceux de l'internat de médecine. Ces capacités 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 ? 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.

L'observation empirique des performances des LLM montre un effet d'échelle très marqué. En dessous d'une certaine taille du réseau (environ le milliard de paramètres), ces capacités cognitives sont inexistantes. Au-dessus de ce seuil, elles se manifestent de façon croissante avec la taille du réseau. Notons que 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 plus de textes en ligne en anglais que dans d'autres langues.

Ajoutées à ces travaux sur la compréhension et l'amélioration des possibilités des LLM, des recherches plus fondamentales sont en cours pour conjuguer les méthodes statistiques à des méthodes de raisonnement à base de connaissances sur le monde, qui fournissent des modèles intelligibles, explicatifs, vérifiables et prouvables.

### *Goulots d'étranglement*

Dans leur conception actuelle de prédicteur par induction statistique, les développements des LLM sont confrontés à deux goulots d'étranglement : (i) la disponibilité de bases d'entraînement plus larges et fiables ; et (ii) le coût énergétique et l'empreinte climatique qu'impliquent leur entrainement et utilisation.

Sur le premier point, la plupart des documents librement accessibles ont été utilisés pour le pré-entraînement des LLM actuels. Des modèles plus importants nécessiteraient des bases d'entraînement plus vastes ou soigneusement choisies. Les discussions pour l'accès à des documents protégés par droits d'auteur ne résoudront que 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 d'un système physique. Mais 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 total de paramètres (environ quelques centaines de milliard). Malgré des recherches actives pour réduire cette complexité, la technique LLM est très coûteuse. Il a été estimé que le pré-entrainement du LLM Gemini Ultra a nécessité plusieurs millions de milliards de milliards d'opérations de calcul et couté 200 millions d'euros. Celui de ChatGPT a consommé 1,3 GWh (soit l'équivalent mensuel d'une commune de 7000 personnes en France). Diverses optimisations dans la gestion de l'énergie et l'ordonnancement des calculs apportent des économies faibles, mais ne changent pas les fondamentaux. Avec les approches actuelles, on s'attend à ce que l'augmentation des performances coûte significativement plus cher. Des approches plus frugales sont nécessaires et commencent à faire l'objet de recherches actives.

## 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. On sait 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, ces biais seront plus répandus, voire plus nocifs.

Les préoccupations éthiques sur l'IA donnent lieu à de nombreuses publications et recommandations. La plupart des 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 et des processus de surveillance active.

Toutefois, ces préoccupations éthiques relatives aux données 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 tâches. 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 cause du rôle et de la valeur sociale de chacun. Elle peut donner lieu à des mécanismes d'exclusion conduisant à considérer des personnes comme *socialement superflues*, ce qui peut légitimement être perç comme une atteinte 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 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 douteux. Cette perspective soulève le risque d'atteinte à l'autonomie et la liberté humaine.

Les machines autonomes peuvent également amplifier les inégalités et accentuer le déséquilibre des pouvoirs entre les groupes humains et les nations. Ainsi, l'utilisation d'armements autonomes est une préoccupation très sérieuse. Malgré l'appel lancé par de nombreux scientifiques en faveur de l'interdiction des machines létales autonomes, appel désormais relayé 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 et de développer ces armements.

*A 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. 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 faible.

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.

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. Il est bien connu que toute technologie est ambivalente, avec ses bons et ses mauvais côtés. 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 ?