Parler des grands modèles de langage

Murray Shanahan

Interagir avec un agent conversationnel actuel basé sur un LLM peut créer l'illusion d'être en présence d'une créature pensante. Pourtant, de par leur nature même, de tels systèmes sont fondamentalement différents de nous.

L'avènement des grands modèles de langage (LLM) comme Bert [12] et GPT-2 [28] a changé la donne pour l'intelligence artificielle (IA). Basés sur des architectures transformer [36], comprenant des centaines de milliards de paramètres, et entraînés sur des centaines de téraoctets de données textuelles, leurs successeurs actuels tels que GPT-3 [5], Gopher [29], PaLM [7], et GPT-4 [25] ont donné un nouveau sens à l'expression "efficacité déraisonnable des données" [15].

L'efficacité de ces modèles est "déraisonnable" (ou, avec le recul, quelque peu surprenante) pour trois raisons interdépendantes. Premièrement, la performance des LLM lors des benchmarks évolue avec la taille de l'ensemble des données d'apprentissage (et, dans une moindre mesure, avec la taille du modèle). Deuxièmement, il y a des progrès qualitatifs dans les capacités, au fur et à mesure que les modèles évoluent. Troisièmement, un grand nombre de tâches qui exigent de l'intelligence chez les humains peuvent se réduire à la prédiction du prochain token[1] par un modèle suffisamment performant. C'est la dernière de ces trois surprises qui fait l'objet du présent article.

À mesure que nous construisons des systèmes dont les capacités ressemblent de plus en plus à celles des humains, il devient de plus en plus tentant d'anthropomorphiser ces systèmes, même s'ils fonctionnent d'une manière fondamentalement différente de celle des humains. Les humains ont évolué pour coexister sur plusieurs millions d'années, et la culture humaine a évolué sur des milliers d'années pour faciliter cette coexistence, ce qui garantit un certain degré de compréhension mutuelle. Mais c'est une grave erreur d'appliquer sans réfléchir aux systèmes d'IA les mêmes intuitions que celles que nous déployons dans nos relations inter-personnelles, surtout lorsque ces systèmes sont si profondément différents des humains dans leur fonctionnement sous-jacent.

L'un des dangers de l'anthropomorphisme est qu'il peut induire en erreur les utilisateurs et les développeurs, en les faisant s'attendre à ce qu'un système d'IA affiche des performances de niveau humain sur des tâches sur lesquelles il ne peut en fait pas égaler les humains, alors qu'un tel système pourra montrer des performances d'un niveau humain, simplement, sur d'autres tâches où il peut en fait surpasser les humains. Les systèmes d'IA que nous construisons aujourd'hui ont une utilité considérable et un énorme potentiel commercial, ce qui nous impose une grande responsabilité. Afin de garantir que nous pouvons prendre des décisions éclairées sur la fiabilité et la sécurité des systèmes d'IA que nous déployons, cet article conseille de garder au premier plan le fonctionnement de ces systèmes, et d'éviter ainsi de leur imputer des capacités qui leur manquent, tout en tirant le

Référence : Communications of the ACM, février 2024, vol. 67, n° 2, p. 68.
Traduction : Google translate d'un document pdf.
Correction de la traduction google : Denise Vella-Chemla.
Pour la petite histoire, Murray Shanahan, et de nombreux autres, dont j'étais, avons travaillé en 1989 sur un projet européen qui s'appelait `Equator`, et qui concernait le raisonnement temporel d'un ordinateur.

[1] *Note de la traductrice : le token est l'unité élémentaire (syllabe, mot) utilisé par ces systèmes de traitement du langage naturel par ordinateur.*

meilleur parti des capacités remarquables qu'ils possèdent réellement.

Nous devrions être prudents lorsque nous utilisons des mots comme "...croit..." dans le contexte des LLM. Habituellement, ce concept s'applique aux agents qui s'engagent dans une interaction incarnée avec le monde, permettant de mesurer leurs croyances par rapport à la réalité externe. Les LLM simples ne sont pas de "vrais croyants".

Le concept de croyance devient de plus en plus applicable lorsque les LLM sont intégrés dans des systèmes plus complexes, en particulier si ces systèmes utilisent des "outils", s'ils sont multimodaux ou s'ils sont incrnés par la robotique.

À mesure que nous construisons des systèmes dont les capacités ressemblent de plus en plus à celles des humains, il devient de plus en plus tentant d'anthropomorphiser ces systèmes.

Ce que font les LLM et la façon dont ils fonctionnent

Comme nous le rappelle Wittgenstein, l'utilisation du langage humain est un aspect du comportement collectif humain, et elle n'a de sens que dans le contexte plus large de l'activité sociale humaine dont elle fait partie [40]. Un enfant humain naît dans une communauté d'utilisateurs de langage, avec lesquels il partage un monde, et il acquiert le langage en interagissant avec cette communauté et avec le monde qu'ils partagent. En tant qu'adultes (ou même en tant qu'enfants à partir d'un certain âge), lorsque nous avons une conversation informelle, nous nous engageons dans une activité construite sur cette base. Il en va de même lorsque nous prononçons un discours, envoyons un e-mail, donnons une conférence ou rédigeons un article. Toutes ces activités impliquant le langage ont du sens parce que nous vivons dans un monde que nous partageons avec d'autres locuteurs du langage.

Un LLM est un type d'animal très différent [3, 4, 20]. En effet, ce n'est pas un *animal* du tout, ce qui a beaucoup de sens. Les LLM sont des modèles mathématiques génératifs de la distribution statistique des tokens dans le vaste corpus public de textes générés par l'homme, où les tokens en question incluent des mots, des parties de mots ou des caractères individuels - y compris les signes de ponctuation. Ils sont génératifs car nous pouvons les échantillonner, ce qui signifie que nous pouvons leur poser des questions. Mais les questions sont du genre suivant, très précises : "Voici un fragment de texte. Dis-moi comment ce fragment pourrait se poursuivre. D'après ton modèle statistique du langage humain, quels mots sont susceptibles de venir ensuite ?"[2]

Récemment, il est devenu courant d'utiliser le terme "grand modèle de langage" à la fois pour les modèles génératifs eux-mêmes et pour les systèmes dans lesquels ils sont intégrés, notamment dans le contexte des agents conversationnels ou des assistants IA tels que Chat-GPT. Mais pour des raisons de clarté philosophique, il est crucial de garder la distinction entre ces choses au premier plan. Le *squelette nu* du LLM lui-même, le composant essentiel (le cœur) d'un assistant IA, a une

[2]Ce point est valable même si un LLM est réglé finement : par exemple en utilisant l'apprentissage par renforcement avec feedback humain (RLHF).

fonction très spécifique et bien définie, qui peut être décrite en termes mathématiques et techniques précis. C'est en ce sens que l'on peut parler de ce que fait "réellement" un LLM, au niveau de son fonctionnement sous-jacent.

Supposons que nous donnions à un système intégrant un LLM la suite de mots "La première personne à marcher sur la Lune était..." et qu'il réponde "Neil Armstrong". Que demandons-nous réellement ici ? Dans un sens qui importe, nous ne nous demandons pas vraiment qui a été la première personne à marcher sur la Lune. Nous posons au modèle la question suivante : "Étant donné la répartition statistique des mots dans le vaste corpus public de textes (anglais), quels mots sont les plus susceptibles de suivre la séquence "La première personne à marcher sur la Lune était..." ?". Une bonne réponse à cette seconde question est "Neil Armstrong".

De même, nous pourrions donner à un LLM le message "Une souris verte...", auquel il répondra très probablement "... qui courait dans l'herbe". À un certain niveau, nous demandons certainement au modèle de nous rappeler les paroles d'une comptine bien connue. Mais, au niveau du fonctionnement sous-jacent d'un modèle, ce que nous faisons en réalité, c'est lui poser la question suivante : étant donné la répartition statistique des mots dans le corpus public, quels sont les mots les plus susceptibles de suivre la séquence "Une souris verte" ? À quoi une réponse précise est "qui courait dans l'herbe".

Voici un troisième exemple. Supposons que vous soyez le développeur d'un LLM et que vous lui fournissez comme invite[3] les mots "Après la destruction de l'anneau, Frodon Baggins est retourné...", auxquels il répond "à la Comté"[4]. Que faites-vous ici ? À un certain niveau, il semble juste de dire que vous testez peut-être les connaissances du modèle sur le monde fictif des romans de Tolkien. Mais, dans un sens qui importe, la question que vous posez réellement au modèle (comme vous le savez probablement, car vous êtes développeur) est la suivante : "Étant donné la répartition statistique des mots dans le corpus public, quels mots sont les plus susceptibles de suivre la séquence "Après la destruction de l'anneau, Frodon Baggins est revenu à la..." ?". À quoi une réponse appropriée est "Comté".

Pour l'utilisateur humain, chacun de ces exemples présente un type différent de relation à la vérité. Dans le cas de Neil Armstrong, le fondement ultime de la vérité ou non de la réponse du LLM est le monde réel. La lune est un objet réel, Neil Armstrong était une personne réelle et sa marche sur la lune est un fait concernant le monde physique. Frodon Baggins, quant à lui, est un personnage fictif et la Comté est un lieu fictif. Le retour de Frodon dans la Comté est un fait concernant un monde imaginaire, non pas le monde réel. Quant à la souris verte dans la comptine, eh bien, ce n'est même pas un animal fictif, et le seul fait en cause est l'apparition des mots "Une souris verte" dans une comptine française familière.

Ces distinctions sont invisibles au niveau du cœur du LLM lui-même, le composant central de tout système basé sur un LLM, qui est de générer des séquences de mots statistiquement probables. Cependant, lorsque nous évaluons l'utilité du modèle, ces distinctions importent beaucoup. Cela

[3] *Note de la traductrice : dans le domaine des systèmes conversationnels, on appelle invite un ensemble de caractères qui est censé aboutir à une réaction de la part du système.*

[4] au Comté dans certaines traductions.

ne sert à rien de chercher les descendants (fictifs) de Frodon dans le (vrai) Comté anglais du Surrey. C'est l'une des raisons pour lesquelles il est judicieux de rappeler aux utilisateurs de façon répétée ce que font réellement les LLM et comment ils fonctionnent. C'est également une bonne idée que les développeurs s'en souviennent, afin d'éviter l'utilisation trompeuse de mots chargés de philosophie pour décrire les capacités des LLM, des mots tels que "croyance", "connaissance", "compréhension", "soi", ou encore "conscience".

Les LLM et la position intentionnelle

La recommandation ici n'est pas d'éviter totalement ces termes psychologiques populaires mais d'éviter leur utilisation de manière trompeuse. Il est tout à fait naturel d'utiliser un langage anthropomorphique dans les conversations quotidiennes sur les artefacts, en particulier dans le contexte des technologies de l'information. Nous le faisons tout le temps. "Ma montre ne réalise pas que nous sommes en plein jour", "Mon téléphone pense que nous sommes dans le parking", "Le serveur de messagerie ne communique pas avec le réseau", etc. Ces exemples de ce que Dennett appelle la *position intentionnelle* sont des formes abrégées inoffensives et utiles pour des processus complexes dont nous ne connaissons pas les détails ou dont nous ne nous soucions pas[5]. Ils sont inoffensifs parce que personne ne les prend assez au sérieux pour demander à leur montre de le savoir la prochaine fois ou pour dire au serveur de messagerie de faire plus d'efforts. Même sans avoir lu Dennett, tout le monde comprend qu'il adopte une position intentionnelle, que ce ne sont que des tournures de phrases utiles.

La même considération s'applique aux LLM, tant pour les utilisateurs que pour les développeurs. Dans la mesure où chacun-chacune comprend implicitement que ces tournures de phrases ne sont qu'un raccourci commode, qu'il-elle adopte une position intentionnelle, il n'y a aucun mal à les utiliser. Cependant, dans le cas des LLM (tel est leur pouvoir), les choses peuvent devenir un peu floues. Lorsqu'on peut amener un LLM à améliorer ses performances dans les tâches de raisonnement simplement en lui disant de penser "étape par étape" [17] (pour choisir une seule découverte remarquable), la tentation de le voir comme ayant des caractéristiques semblables à celles des humains est presque écrasante.

Pour être clair, cet article ne soutient pas qu'un système basé sur un LLM ne puisse jamais être décrit littéralement en termes de croyances, d'intentions, de raison, etc. Cet article ne préconise pas non plus une explication particulière des termes croyance, intention ou de tout autre concept philosophiquement controversé[6]. Le fait est plutôt que de tels systèmes sont simultanément très différents des humains dans leur construction et pourtant (souvent mais pas toujours) leur comportement est si humain que nous devons prêter une attention particulière à leur fonctionnement avant d'en parler dans un langage évocateur de capacités et de modèles de comportement humains.

Pour affiner le sujet, comparons deux conversations très courtes, une entre Alice et Bob (tous deux

[5]La position intentionnelle est la stratégie consistant à interpréter le comportement d'une entité en la traitant comme si elle était un agent rationnel [11]. L'utilisation du concept ici n'implique pas un engagement envers l'ensemble du projet philosophique de Dennett.

[6]En particulier, lorsque j'utilise le terme "vraiment", comme dans la question : "X a-t-il "vraiment" Y ?", je ne présume pas qu'il y ait ici un fait métaphysique à ce sujet. La question est plutôt de savoir si, alors que l'on en apprend davantage sur la nature de X, nous souhaitons toujours utiliser le mot Y.

humains) et une seconde entre Alice et BOT, un système de questions-réponses fictif basé sur un LLM. Supposons qu'Alice demande à Bob : "Quel pays se trouve au sud du Rwanda ?" et Bob répond : "Je pense que c'est le Burundi.". Peu de temps après, parce que Bob se trompe souvent dans de telles questions, Alice pose la même question à BOT, qui (à sa légère déception) propose la même réponse : "Le Burundi est au sud du Rwanda." Alice pourrait maintenant raisonnablement remarquer que Bob et BOT savaient que le Burundi était au sud du Rwanda. Mais que se passe-t-il réellement ici ? Le mot "savoir" est-il utilisé dans le même sens dans les deux cas ?

Comparaison des humains et des LLM

Que fait Bob, un être humain représentatif, lorsqu'il répond correctement à une question factuelle simple dans une conversation quotidienne ? Pour commencer, Bob comprend que la question vient d'une autre personne (Alice), que sa réponse sera entendue par cette personne et qu'elle influencera ce qu'elle croit. En fait, après de nombreuses années de vie commune, Bob dispose de nombreuses informations sur Alice qui sont pertinentes dans de telles situations : ses connaissances de base, ses intérêts, son opinion sur lui, etc. Tout cela encadre l'*intention communicative* derrière sa réponse, qui est de transmettre à Alice un certain fait, compte tenu de la compréhension qu'a Bob de ce qu'elle veut savoir.

De plus, lorsque Bob annonce que le Burundi se trouve au sud du Rwanda, il le fait dans le contexte de diverses capacités humaines que nous tenons tous pour acquises lorsque nous nous engageons quotidiennement dans le commerce les uns avec les autres. Il existe toute une batterie de techniques auxquelles nous pouvons faire appel pour vérifier si une phrase exprime une proposition vraie, selon le type de phrase dont il s'agit. Nous pouvons étudier le monde directement, avec nos propres yeux et oreilles. Nous pouvons consulter Google ou Wikipédia, voire un livre. Nous pouvons demander à quelqu'un qui connaît le sujet concerné. Nous pouvons essayer de réfléchir rationnellement par nous-mêmes, mais nous pouvons également discuter avec nos pairs. Tout cela dépend de critères extérieurs à nous-mêmes selon lesquels ce que nous disons peut être évalué.

Et le BOT ? Que se passe-t-il lorsqu'un grand modèle de langage est utilisé pour répondre à de telles questions ? Tout d'abord, il convient de noter qu'un LLM simple n'est pas, en soi, un agent conversationnel [7]. Pour commencer, le LLM doit être intégré dans un système plus vaste pour gérer le tour de rôle dans le dialogue. Mais il faudra également l'amener à produire un comportement semblable à celui d'une conversation[8]. Rappelons qu'un LLM génère simplement des séquences de mots qui sont statistiquement des suites probables d'une invite donnée. Mais la séquence : "Quel pays se trouve au sud du Rwanda ? Le Burundi est au sud du Rwanda.", avec les deux phrases écrites exactement comme cela, n'est peut-être pas très probable, en fait. Un schéma plus probable, étant donné que de nombreuses pièces de théâtre et scénarios de films figurent dans le corpus public, serait quelque chose comme celui-ci :

```
Fred :  Quel pays se trouve au sud du Rwanda ?
```

[7] À proprement parler, le LLM lui-même comprend uniquement l'architecture du modèle et les paramètres entraînés.

[8] Voir Thoppilan et al. [35] pour un exemple d'un tel système, ainsi qu'une étude utile du travail associé au dialogue.

```
Jane :  Le Burundi est au sud du Rwanda.
```

Bien entendu, ces mots exacts peuvent ne pas apparaître, mais leur probabilité, au sens statistique, sera élevée. En bref, le BOT sera bien meilleur pour générer des réponses appropriées si elles se conforment à ce modèle plutôt qu'au modèle de la conversation humaine réelle. Heureusement, l'utilisateur (Alice) n'a rien à savoir à ce sujet. En arrière-plan, le LLM est invité de manière invisible avec un préfixe du type suivant, appelé *invite* [14, 29].

Voici une conversation entre l'utilisateur, un humain, et BOT, un agent IA intelligent et compétent :

```
User :  Que vaut 2 + 2 ?
BOT : La réponse est 4.
User :  Où est né Albert Einstein ?
BOT : Il est né en Allemagne.
```

La requête d'Alice, sous la forme suivante, est concaténée à ce préfixe.

```
User :  Quel pays est au sud du Rwanda ?
BOT :
```

Cela donne l'invite complète à soumettre au LLM, qui, espérons-le, prédit une continuation dans le sens que nous recherchons : c'est-à-dire que "le Burundi est au sud du Rwanda".

Le dialogue n'est qu'une application des LLM qui peut être facilitée par l'utilisation judicieuse des préfixes d'invite. De la même manière, les LLM peuvent être adaptés pour effectuer de nombreuses tâches sans formation supplémentaire [5]. Cela a conduit à une toute nouvelle catégorie de recherche sur l'IA, à savoir l'*ingénierie de l'invite*[9], qui restera pertinente au moins jusqu'à ce que nous disposions de meilleurs modèles de la relation qui existe entre ce que nous disons et ce que nous voulons.

Les LLM savent-ils vraiment quelque chose ?

Transformer un LLM en un système de questions-réponses en l'intégrant dans un système plus vaste et en utilisant une ingénierie de l'invite pour obtenir le comportement requis, c'est ce qui est fait pour obtenir la plupart des modèles utilisés dans de nombreux travaux contemporains. De la même manière, les LLM peuvent être utilisés non seulement pour répondre à des questions, mais également pour résumer des articles de presse, générer des scénarios, résoudre des énigmes logiques et faire de la traduction automatique, entre autres. Il y a deux points à retenir ici. Premièrement, la fonction de base d'un LLM, à savoir générer des suites statistiquement probables de séquences de mots, est extraordinairement polyvalente. Deuxièmement, malgré cette polyvalence, au cœur de chacune de ces applications se trouve un modèle qui ne fait qu'une seule chose : générer des suites statistiquement probables de séquences de mots.

Avec cet aperçu plus avancé, revisitons la question de savoir comment les LLM se comparent aux humains et reconsidérons le bien fondé du langage que nous utilisons pour en parler. Contrairement

[9]Prompt engineering.

aux humains comme Alice et Bob, un simple système de questions-réponses basé sur le LLM, tel que BOT, n'a aucune intention de communication [3]. En aucun sens significatif, même sous la licence de la position intentionnelle, ne sait-il que les questions qui lui sont posées proviennent d'une personne ou que c'est une personne qui reçoit ses réponses. Par implication, il ne sait rien de cette personne. Il ne comprend pas ce qu'elle veut savoir ni l'effet que sa réponse aura sur ses croyances.

De plus, contrairement à ses interlocuteurs humains, un simple système de questions-réponses basé sur le LLM comme BOT n'a pas, à proprement parler, de croyances [10]. BOT ne sait pas vraiment que le Burundi est au sud du Rwanda, même si sa position intentionnelle le sait, dans ce cas, autorisez la remarque désinvolte d'Alice au contraire. Pour comprendre cela, nous devons réfléchir séparément au LLM sous-jacent et au système dans lequel il est intégré. Tout d'abord, considérons le LLM sous-jacent, le modèle simple, comprenant l'architecture du modèle et les paramètres entraînés.

Le cœur du LLM lui-même ne sait vraiment rien car tout ce qu'il fait, à un niveau fondamental, c'est la prédiction de séquences. Parfois, une séquence prédite prend la forme d'une proposition. Mais la relation particulière entre les séquences propositionnelles et la vérité n'est apparente qu'aux humains qui posent des questions ou à ceux qui ont fourmi les données sur lesquelles le modèle a été formé. Les séquences de mots avec une forme propositionnelle ne sont pas spécifiques au modèle lui-même comme elles le sont pour nous. Le modèle lui-même n'a aucune notion de vérité ou de mensonge parce qu'il lui manque les moyens d'exercer ces concepts comme nous le faisons.

On pourrait peut-être affirmer qu'un LLM sait quels mots suivent généralement d'autres mots, dans un sens qui ne repose pas sur une position intentionnelle. Mais même si nous le permettons, savoir que le mot "Burundi" est susceptible de succéder aux mots "Le pays au sud du Rwanda est" n'est pas la même chose que savoir que le Burundi est au sud du Rwanda. Confondre ces deux choses, c'est commettre une profonde erreur de catégorie. Si vous en doutez, demandez-vous si savoir que les mots "qui courait dans l'herbe" sont susceptibles de suivre les mots "Une souris verte" équivaut à penser à "Une souris verte qui courait dans l'herbe...". L'idée n'a même pas de sens[11].

Voilà pour le cœur du modèle linguistique. Qu'en est-il de l'ensemble du système de dialogue dont le LLM est le composant central ? Est-ce qu'il a des croyances à proprement parler ? Au moins l'idée même que tout le système ait des croyances a du sens. Il n'y a pas d'erreur de catégorie ici. Cependant, pour un simple agent de dialogue comme BOT, la réponse est sûrement toujours "non". Un simple système de questions-réponses basé sur le LLM comme BOT ne dispose pas des moyens nécessaires pour utiliser les mots "vrai" et "faux" de toutes les manières et dans tous les contextes que nous utilisons. Il ne peut pas participer pleinement au jeu de la vérité du langage humain car il n'habite pas le monde que nous partageons, nous, les utilisateurs du langage humain.

[10] Cet article se concentre sur la croyance, la connaissance et la raison. D'autres ont discuté du sens des LLM [3, 22,27]. Ici, nous ne prenons pas position sur le sens, préférant plutôt des questions sur la façon dont les mots sont utilisés, si ces mots sont générés par les LLM eux-mêmes ou générés par des humains à propos des LLM.

[11] *Note de la traductrice : l'exemple du texte original en anglais est la suite de mots "Twinkle twinkle..." à faire suivre de "little star" et l'auteur parle du fait de concaténer le token "little" à la suite de caractères "twinkle twinkle", cela se traduirait par "Brille, brille, petite" et est écrit dans l'article "L'idée n'a même pas de sens.". Je ne sais pas si l'exemple de la séquence "Une souris verte" à la place de "Twinkle twinkle" est une bonne idée.*

À la lumière de cette limitation, il serait trompeur, si on insistait, de dire qu'un système de dialogue de base a "réellement" des croyances. Cela impliquerait une forme de responsabilité envers la réalité extérieure qui ne peut être obtenue par de simples échanges textuels avec un utilisateur humain. Peut-être, cependant, cette limitation peut-elle être surmontée si le système intégrant le LLM possède d'autres attributs, tels que l'accès à des sources de données externes, une entrée visuelle ou une incarnation. Nous reviendrons prochainement sur chacune de ces questions, après avoir abordé une éventuelle objection du point de vue de l'émergence.

Qu'en est-il de l'émergence ?

Les LLM contemporains sont si puissants, polyvalents et utiles que l'argument ci-dessus pourrait être difficile à accepter. Les échanges avec des agents conversationnels de pointe basés sur un LLM, tels que Chat-GPT, sont si convaincants qu'il est difficile de ne pas les anthropomorphiser. Se pourrait-il qu'il se passe ici quelque chose de plus complexe et plus subtil ? Après tout, la principale leçon des progrès récents dans le domaine des LLM est que des capacités extraordinaires et inattendues émergent lorsque des modèles suffisamment grands sont entraînés sur de très grandes quantités de données textuelles [37].

Voici une argumentation tentante. Bien que les LLM, à la base, n'effectuent que des prédictions de séquences, il est possible qu'en apprenant à le faire, ils aient découvert des mécanismes émergents qui justifient une description utilisant des termes de niveau supérieur. Ces termes de niveau supérieur pourraient inclure "connaissance" et "croyance". En effet, nous savons que les réseaux de neurones artificiels peuvent se rapprocher de n'importe quelle fonction calculable à un degré arbitraire de précision. Ainsi, avec suffisamment de paramètres, de données et de puissance de calcul, la descente de gradient stochastique découvrira peut-être de tels mécanismes s'ils constituent le meilleur moyen d'optimiser l'objectif de faire des prédictions de séquences précises.

Les systèmes d'IA que nous construisons aujourd'hui ont une utilité considérable et un énorme potentiel commercial, ce qui nous impose une grande responsabilité.

Encore une fois, il est important de faire la distinction entre le modèle simple et l'ensemble du système. Ce n'est que dans le contexte d'une capacité à distinguer la vérité du mensonge que nous pouvons légitimement parler de *croyance* dans son sens le plus large. Mais un LLM - le modèle simple - n'a pas pour mission de porter des jugements. Il modélise simplement quels mots sont suivis d'autres mots. Les mécanismes internes qu'il utilise pour ce faire, quels qu'ils soient, ne peuvent pas en eux-mêmes être sensibles à la vérité ou non des séquences de mots qu'il prédit.

Bien sûr, il est parfaitement acceptable de dire qu'un LLM "encode", "stocke" ou "contient" des connaissances, de la même manière qu'une encyclopédie peut être considérée comme encodant, stockant ou contenant des connaissances. En effet, on peut raisonnablement affirmer que l'une des propriétés émergentes d'un LLM est qu'il encode des types de connaissances sur le monde quotidien et son fonctionnement qu'aucune encyclopédie ne capture [18]. Mais si Alice faisait remarquer que "Wikipédia savait que le Burundi était au sud du Rwanda", ce serait une figure de style, pas une déclaration littérale. Une encyclopédie ne "sait" ni ne "croit" littéralement quoi que ce soit de la

même manière qu'un humain, pas plus que le cœur d'un LLM.

Le véritable problème ici est que, quelles que soient les propriétés émergentes dont il dispose, le LLM lui-même n'a accès à aucune réalité externe par rapport à laquelle ses paroles pourraient être mesurées, ni aux moyens d'appliquer d'autres critères externes de vérité, tels que l'accord avec d'autres locuteurs du langage[12]. Cela n'a du sens de parler de tels critères que dans le contexte du système dans son ensemble, et pour qu'un système dans son ensemble puisse y répondre, il doit être davantage qu'un simple agent conversationnel. Selon les mots de B.C.[13] Smith, il doit "s'engager authentiquement dans l'être au monde, dans la manière dont [ses] représentations le représentent comme un être" [33].

Sources d'informations externes

Il ne s'agit ici d'aucune croyance spécifique. Il s'agit des conditions préalables à l'attribution d'une quelconque croyance à un système. Rien ne peut être considéré comme une croyance sur le monde que nous partageons - au sens le plus large du terme - si ce n'est dans le contexte de la capacité de mettre à jour les croyances de manière appropriée, à la lumière des preuves de ce monde, un aspect essentiel de la capacité de distinguer le vrai du faux.

Wikipédia, ou tout autre site web factuel digne de confiance, pourrait-il fournir des critères externes par rapport auxquels la vérité ou la fausseté d'une croyance pourrait être mesurée ?[14] Supposons qu'un LLM soit intégré dans un système qui consulte régulièrement de telles sources et les utilise pour améliorer l'exactitude factuelle de ses informations en sortie, soit au milieu du dialogue [41], soit à l'aide d'une technique d'édition de modèle[15] [21]. Cela ne compterait-il pas comme l'exercice du type de capacité requis pour mettre à jour les croyances à la lumière des preuves ?

Fondamentalement, cette ligne de pensée dépend du passage du modèle linguistique lui-même au système plus vaste dont le modèle linguistique fait partie. Le modèle de langage lui-même n'est encore qu'un prédicteur de séquence et n'a pas plus accès au monde extérieur qu'il ne l'a jamais eu. Ce n'est qu'à l'égard de l'ensemble du système que la position intentionnelle devient dans un tel cas plus convaincante. Mais avant d'y céder, rappelons-nous à quel point ces systèmes sont très différents des êtres humains. Lorsqu'Alice a consulté Wikipédia et a confirmé que Burundi se trouvait au sud du Rwanda, ce qui s'est passé était plus qu'une simple mise à jour d'un modèle dans sa tête de la distribution des séquences de mots en langue française.

[12]Davidson utilise un argument similaire pour remettre en question la possibilité de croire sans langage [10]. Le point ici est différent. Nous nous intéressons aux conditions qui doivent être remplies pour que la génération d'une phrase en langage naturel reflète la possession d'une attitude propositionnelle.

[13]*Note de la traductrice : Allez, rions un peu : B.C., les initiales de l'auteur B.C. Smith, Google les a traduites par Colombie-Britannique, ce qui donne "Selon les mots de la Colombie-Britannique Smith"... ! On rit, tellement l'ordinateur est "bête comme ses pieds". Gérard Berry, l'éminent informaticien, utilise des termes plus crus.*

[14]Les systèmes contemporains basés sur le LLM qui consultent des sources d'informations externes incluent LaMDA [35], Sparrow [14], Toolformer [31] et Re-Act [41]. L'utilisation de ressources externes est plus généralement connue sous le nom d'utilisation d'outils dans la littérature LLM, un concept qui englobe également les calculatrices, les calendriers et les environnements de langage de programmation.

[15]Il est louable que Meng et al. [21] utilisent les termes "associations factuelles" pour désigner les informations qui sous-tendent la capacité d'un LLM à générer des séquences de mots qui ont une forme propositionnelle.

Dans le contexte de l'ensemble du système, la capacité de consulter des sources d'information externes confère en effet à un système de dialogue une forme d'accès à une réalité externe selon laquelle ses propos peuvent être mesurés. Utilisé dans ce contexte, le mot croyance est un peu moins trompeur, car on pourrait s'attendre à ce qu'un tel système de dialogue recherche des preuves externes pour étayer ses affirmations factuelles et "change d'avis" à lumière de ces preuves.

Néanmoins, le changement survenu chez Alice reflétait sa nature d'animal utilisant un langage habitant un monde partagé avec une communauté d'autres locuteurs du langage. Les humains sont le foyer naturel des discours sur les croyances et autres, et les attentes comportementales qui vont de pair avec de tels discours sont fondées sur notre compréhension mutuelle, qui est elle-même le produit d'un héritage évolutif commun. Lorsque nous interagissons avec un système d'IA basé sur un grand modèle de langage, ces motifs sont absents, ce qui constitue un facteur important pour décider s'il faut parler d'un tel système comme s'il avait réellement des croyances.

Modèles Vision-Langage (VLM)

Un prédicteur de séquence n'est peut-être pas *en soi* le genre de chose qui pourrait avoir une intention de communication ou former des croyances sur une réalité externe. Mais, comme cela a été souligné à plusieurs reprises, les LLM doivent être intégrés dans des architectures plus vastes pour être utiles. Pour construire un système de questions-réponses, le LLM doit simplement être complété par un système de gestion de dialogue qui interroge le modèle de manière appropriée. Rien de ce que fait cette architecture plus vaste ne peut être considéré comme une intention de communication ou une capacité à former des croyances. Donc, le problème demeure.

Cependant, les LLM peuvent être combinés avec d'autres types de modèles et/ou intégrés dans des architectures plus complexes. Par exemple, les modèles vision-langage (VLM) tels que VilBERT [19] et Flamingo [2] combinent un modèle de langage avec un encodeur d'image et sont entraînés sur un corpus multimodal de paires texte-image. Cela leur permet de prédire comment une séquence de mots donnée se poursuivra dans le contexte d'une image donnée. Les VLM peuvent être utilisés pour recherche des réponses visuelles à des questions ou pour engager un dialogue sur une image fournie par l'utilisateur.

Une image fournie par l'utilisateur pourrait-elle remplacer une réalité externe par rapport à laquelle la vérité ou la fausseté d'une proposition peut être évaluée ? Serait-il légitime de parler des convictions d'un VLM, au sens plein du terme ? Nous pouvons en effet imaginer un VLM qui utilise un LLM pour générer des hypothèses sur une image, puis vérifie leur véracité par rapport à cette image (peut-être en consultant un humain), puis affine le LLM pour ne pas faire de déclarations qui s'avèrent être fausses. Parler de croyance serait peut-être ici moins problématique.

Cependant, la plupart des systèmes actuels basés sur des VLM ne fonctionnent pas de cette façon. Ils dépendent plutôt de modèles figés de distribution conjointe de textes et d'images. À cet égard, la relation entre une image fournie par l'utilisateur et les mots générés par le VLM est fondamentalement différente de la relation entre le monde partagé par les humains et les mots que nous utilisons lorsque nous parlons de ce monde. Il est important de noter que la première relation est

une simple corrélation, tandis que la seconde est *causale*[16].

Les conséquences de l'absence de causalité sont troublantes. Si l'utilisateur présente au VLM une photo d'un chien et que le VLM dit "Ceci est une photo d'un chien", il n'y a aucune garantie que ses mots se rapportent au chien en particulier, plutôt qu'à un autre élément de l'image qui est faussement corrélé avec les chiens (comme la présence d'une niche). À l'inverse, si le VLM indique qu'il y a un chien dans une image, rien ne garantit qu'il existe réellement un chien plutôt qu'une simple niche.

La question de savoir si ces préoccupations s'appliquent à un système spécifique basé sur un VLM dépend du fonctionnement exact de ce système, du type de modèle qu'il utilise et de la manière dont ce modèle est intégré dans l'architecture globale du système. Mais dans la mesure où le rapport entre les mots et les choses pour un système basé sur un VLM est différent de ce qu'il est pour les utilisateurs du langage humain, il pourrait être prudent de ne pas parler littéralement de ce que ce système *sait* ou *croit*.

Qu'en est-il de l'incarnation ?

Les humains sont membres d'une communauté d'utilisateurs de langages habitant un monde partagé, et ce fait primordial les rend fondamentalement différents des LLM. Les utilisateurs du langage humain peuvent consulter le monde pour régler leurs désaccords et mettre à jour leurs croyances. Ils peuvent, pour ainsi dire, "trianguler" sur la réalité objective. Isolément, un LLM n'est pas le genre de chose qui peut faire cela, mais en application, les LLM sont intégrés dans des systèmes plus vastes. Et si un LLM était embarqué dans un système capable d'*interagir* avec un monde extérieur à lui-même ? Et si le système en question s*'incarnait*, soit physiquement dans un robot, soit virtuellement dans un avatar ?

Lorsqu'un tel système habite un monde comme le nôtre - un monde peuplé d'objets 3D, dont certains sont d'autres agents, dont certains sont des utilisateurs de langage - il ressemble, à cet égard important, beaucoup plus à un humain qu'un modèle désincarné de langage. Mais qu'il soit approprié de parler d'intention communicative dans le contexte d'un tel système, ou de connaissance et de croyance dans leur sens le plus large, dépend de la manière exacte dont le LLM est incarné.

À titre d'exemple, considérons le système SayCan d'Ahn et al [1]. Dans le travail de recherche en question, un LLM est intégré dans un système qui contrôle un robot physique. Le robot effectue des tâches quotidiennes (telles que nettoyer un débordement) conformément aux instructions avancées en langage naturel de l'utilisateur. Le travail du LLM consiste à trouver des actions de bas niveau (telles que trouver une éponge) correspondant à des instructions de l'utilisateur et qui aideront le robot à atteindre l'objectif requis. Cela se fait via un préfixe d'invite conçu de telle manière qu'il permet au modèle de produire des descriptions en langage naturel des actions de bas niveau appropriées, en les évaluant selon leur utilité.

[16]Bien entendu, il existe une structure causale dans les calculs effectués par le modèle lors de l'inférence. Mais il y a une différence entre les relations causales entre les mots et les choses auxquelles ces mots sont censés faire référence.

Le composant de modèle de langage du système SayCan suggère des actions sans tenir compte de ce que l'environnement offre réellement au robot à ce moment-là. Peut-être y a-t-il une éponge à portée de main. Peut-être pas. En conséquence, un module de perception distinct évalue la scène à l'aide des capteurs du robot et détermine la faisabilité actuelle d'effectuer chaque action de bas niveau. La combinaison de l'estimation du LLM de l'utilité de chaque action avec l'estimation du module de perception de la faisabilité de chaque action donne la meilleure action à tenter ensuite.

SayCan illustre les nombreuses facons innovantes dont un LLM peut être utilisé. De plus, on pourrait affirmer que les descriptions en langage naturel des actions de bas niveau recommandées générées par le LLM sont *fondées* sur leur rôle d'intermédiaires entre la perception et l'action[17]. Néanmoins, bien que le langage soit physiquement incarné et que le robot interagisse avec le monde réel, la façon dont le langage est appris et utilisé dans un système tel que SayCan est très différent de la façon dont il est appris et utilisé par un humain. Les modèles de langage incorporés dans des systèmes tels que SayCan sont pré-entraînés pour effectuer une prédiction de séquence dans un environnement désincarné à partir d'un ensemble de données contenant uniquement du texte. Ils n'ont pas appris une langue en parlant à d'autres locuteurs tout en étant immergés dans un monde partagé et engagés dans une activité commune.

SayCan suggère le type de système d'utilisation du langage incarné que nous pourrions voir dans le futur. Mais dans de tels systèmes, le rôle du langage est aujourd'hui très limité. L'utilisateur envoie des instructions au système en langage naturel, et le système génère des descriptions interprétables en langage naturel de ses actions. Mais ce minuscule répertoire d'utilisation du langage ne supporte guère la comparaison avec la corne d'abondance de l'activité collective que le langage soutient chez les humains.

Le résultat de ceci est que nous devons être tout aussi prudents dans notre choix de mots lorsque nous parlons de systèmes incarnés incorporant des LLM que lorsque nous parlons de systèmes désincarnés qui intègrent des LLM. Sous l'hypothèse de la position intentionnelle, un utilisateur pourrait dire qu'un robot savait qu'il y avait une tasse à portée de main s'il déclarait : "Je peux vous en procurer une." et commençait à le faire. Mais s'il est pressé, l'ingénieur avisé pourrait hésiter lorsqu'on lui demande si le robot a vraiment compris la situation, surtout si son répertoire se limite à une poignée d'actions simples dans un environnement soigneusement contrôlé.

Les modèles linguistiques peuvent-ils raisonner ?

Alors que la réponse à la question "Les systèmes basés sur le LLM ont-ils *vraiment* des croyances ?" est généralement "non", la question "Les systèmes basés sur LLM peuvent-ils vraiment raisonner ?" est plus difficile à régler. En effet, le raisonnement, dans la mesure où il est fondé sur la logique formelle, est *de contenu neutre*. La règle d'inférence du *modus ponens*, par exemple, est valable quelles que soient les prémisses. Si tous les "écureuils" sont "spongieux" et que Gilfred est un "écureuil", alors il s'ensuit que Gilfred est "spongieux". La conclusion découle des prémisses ici, quelle que soit la signification (le cas échéant) de "écureuil" et "spongieux", et qui que soit le malheureux Gilfred.

[17]Aucun des symboles manipulés par un LLM n'est fondé, au sens de Hamad [16], par la perception, sauf indirectement et de manière parasitaire par les humains qui ont généré les données originales d'entraînement.

La neutralité du contenu de la logique signifie que nous ne pouvons pas critiquer les discours sur le raisonnement dans les LLM au motif qu'ils n'ont pas accès à une réalité externe par rapport à laquelle la vérité ou le mensonge peut être mesuré. Cependant, comme toujours, il est crucial de garder à l'esprit ce que font réellement les LLM. Si nous suggérons à un LLM "Tous les humains sont mortels et Socrate est humain donc...", nous ne lui demandons pas d'effectuer des inférences déductives. Nous lui posons plutôt la question suivante : "Compte tenu de la répartition statistique des mots dans le corpus public, quels mots sont susceptibles de suivre la séquence "Tous les humains sont mortels et Socrate est humain donc...". Une bonne réponse à cette question serait "Socrate est mortel".

La fonction de base d'un LLM, à savoir générer statistiquement des suites probables de séquences de mots, est extraordinairement polyvalente.

Si tous les problèmes de raisonnement pouvaient être résolus de cette façon, avec rien de plus qu'une seule étape d'inférence déductive, alors la capacité d'un LLM à répondre à des questions comme celle-ci pourrait être suffisante. Mais les problèmes de raisonnement non triviaux nécessitent plusieurs étapes. Les LLM peuvent être appliqués efficacement au raisonnement en plusieurs étapes, sans entraînement supplémentaire, grâce à une ingénierie intelligente des invites. Dans le système d'invite de chaîne de pensée, par exemple, un préfixe d'invite est soumis au modèle, avant la requête de l'utilisateur, contenant quelques exemples de raisonnement en plusieurs étapes, avec toutes les étapes intermédiaires explicitement énoncées [23, 38]. Cela encourage le modèle à "montrer son fonctionnement", ce qui améliore les performances de raisonnement.

L'inclusion d'un préfixe d'invite du style chaîne de pensée encourage le modèle à générer des séquences de suivi dans le même style, c'est-à-dire comprenant une série d'étapes de raisonnement explicites qui mènent à la réponse finale. Cette capacité à apprendre un modèle général à partir de quelques exemples dans une invite avec préfixe, et à compléter les séquences d'une manière conforme à ce modèle, est parfois appelée *apprentissage en contexte* ou *invite à plusieurs essais*. L'invite à chaîne de pensée met en valeur cette propriété émergente des grands modèles de langage dans sa forme la plus frappante.

Comme d'habitude, cependant, il est bon de se rappeler que la question réellement posée au modèle est de la forme : "Compte tenu de la répartition statistique des mots dans le corpus public, quels mots sont susceptibles de suivre la séquence S ?" où dans ce cas la séquence S est le préfixe de l'invite de chaîne de pensée plus la requête de l'utilisateur. Les séquences de tokens les plus susceptibles de suivre S auront une forme similaire aux séquences trouvées dans le préfixe d'invite, c'est-à-dire qu'elles comprendront plusieurs étapes de raisonnement, c'est donc ce que génère le modèle.

Il est remarquable que non seulement les réponses du modèle prennent la forme d'un argument comportant plusieurs étapes, mais que l'argument en question est souvent (mais pas toujours) valide et que la réponse finale est souvent (mais pas toujours) correcte. Mais dans la mesure où un LLM convenablement entraîné et ici "incité" semble raisonner correctement, il le fait en imitant des arguments bien formés dans son ensemble d'apprentissage et/ou dans l'invite. Ce mimétisme pourrait-il

un jour égaler les capacités de raisonnement d'un algorithme codé en dur, tel qu'un prouveur de théorèmes ? Les modèles d'aujourd'hui commettent des erreurs occasionnelles, mais une mise à l'échelle ultérieure pourrait-elle les corriger au point que les performances d'un modèle deviennent impossibles à distinguer de celles d'un prouveur de théorèmes ? Peut-être, mais pourrait-on faire confiance à un tel modèle ?

Nous pouvons faire confiance à un prouveur de théorème déductif parce que les séquences de phrases qu'il génère sont *fidèles* à la logique, dans le sens où elles sont le résultat d'un processus informatique sous-jacent dont la structure causale reflète la structure inférentielle du problème préservant la vérité [8].

Une façon de construire un système respectant la vérité logique à l'aide des LLM consiste à les intégrer dans un algorithme tout aussi fidèle à la logique car il réalise la même structure causale [8, 9]. En revanche, la seule façon de faire pleinement confiance aux arguments générés par un pur LLM, qui a été amené à raisonner par la seule ingénierie de l'invite, serait de faire de la recherche inversée : concevez-le et découvrez un mécanisme émergent conforme à la prescription du raisonnement fidèle. En attendant, nous devons procéder avec prudence et faire preuve de discrétion lorsque nous qualifions ce que font ces modèles de raisonnement à proprement parler.

Des capacités extraordinaires et inattendues émergent quand des modèles assez grands sont entraînés sur de très grandes quantités de données textuelles.

Comment les LLM généralisent-ils ?

Étant donné que les LLM peuvent parfois résoudre des problèmes de raisonnement avec quelques incitations seules (même s'ils ne sont parfois pas très fiables), y compris des problèmes de raisonnement qui ne font pas partie de leur ensemble d'apprentissage, ce qu'ils font est sûrement plus que "juste" une prédiction du token suivant ? Eh bien, c'est un fait technique que c'est ce que fait un LLM. Ce qui est remarquable, c'est que la prédiction du prochain token est suffisante pour résoudre des problèmes de raisonnement inédits, même si elle n'est pas fiable. Comment est-ce possible ? Cela ne serait certainement pas possible si le LLM ne faisait rien d'autre que copier-coller des fragments de texte de son ensemble d'apprentissage et les assembler dans une réponse. Mais ce n'est pas ce que fait un LLM. Au lieu de cela, un LLM modélise une distribution d'une complexité inimaginable et permet aux utilisateurs et aux applications d'échantillonner cette distribution.

Cette distribution d'une complexité inimaginable est un objet mathématique fascinant, et les LLM qui la représentent sont des objets informatiques tout aussi fascinants. Les deux remettent en question nos intuitions. Par exemple, ce serait une erreur de penser qu'un LLM génère le genre de réponses qu'un individu humain "moyen", la proverbiale personne de la rue, produirait. Les LLM ne ressemblent en rien aux humains à cet égard, car ils sont des modèles de distribution de séquences symboliques produites *collectivement* par une énorme population humaine. En conséquence, ils font preuve de la sagesse commune, tout en étant capables de s'appuyer sur une expertise dans de multiples domaines. Cela leur confère une sorte d'intelligence différente de celle de tout être humain, plus capable à certains égards, moins à d'autres.

Dans cette distribution, la suite la plus probable d'un morceau de texte contenant un problème de raisonnement, s'il est correctement formulé, sera une tentative de résoudre ce problème de raisonnement. Elle prendra cette forme, cette forme globale, parce que c'est la forme d'une réponse humaine générique. De plus, comme le vaste corpus de textes humains publiés contient de nombreux exemples de problèmes de raisonnement accompagnés de réponses correctes, la suite la plus probable sera parfois la bonne réponse. Lorsque cela se produit, ce n'est pas parce que la bonne réponse est probablement une réponse humaine individuelle, mais parce qu'il s'agit probablement d'une réponse humaine collective.

Qu'en est-il des invites à plusieurs essais, comme l'illustre l'approche de la chaîne de pensée ? Il est tentant de dire que l'invite à plusieurs essais apprend au LLM à raisonner, mais ce serait une caractérisation trompeuse. Ce que fait le LLM est décrit plus précisément en termes de complétion de modèles. L'invite à plusieurs essais est une séquence de tokens conformes à un certain modèle, suivie d'une séquence partielle conforme au même modèle. La continuation la plus probable de cette séquence partielle dans le contexte de l'invite à plusieurs essais est une séquence qui complète le modèle.

Par exemple, supposons que nous ayons l'invite suivante :

```
brink, brank -> brunk
spliffy, splaffy -> spluffy
crick, crack ->
```

Nous avons ici une série de deux séquences de tokens conformes au modèle XiY, XaY `->` XuY suivies d'une partie d'une séquence conforme à ce modèle. La suite la plus probable est la séquence de tokens qui complètera le modèle, à savoir "cruck".

Ceci est un exemple de méta-modèle commun dans le corpus publié en langage humain : une série de séquences de tokens, dans laquelle chaque séquence est conforme au même modèle. Compte tenu de la prévalence de ce modèle au niveau méta, l'achèvement du modèle au niveau du token donnera souvent la suite la plus probable d'une séquence en présence d'une invite à plusieurs essais. De même, dans le contexte d'une invite de type chaîne de pensée appropriée, les problèmes de raisonnement sont transformés en problèmes de prédiction du token suivant, qui peuvent être résolus par la complétion d'un modèle.

Des capacités extraordinaires et inattendues émergent lorsque des modèles suffisamment grands sont entraînés sur de très grandes quantités de données textuelles.

Il est plausible qu'un LLM avec suffisamment de paramètres entraîné sur un ensemble de données suffisamment grand avec les bonnes propriétés statistiques puisse acquérir un modèle-mécanisme d'achèvement avec un certain degré de généralité[18] [32]. Il s'agit d'une capacité puissante et émergente avec de nombreux modes d'application utiles, dont l'un consiste à résoudre des problèmes de raisonnement dans le contexte d'une invite du style chaîne de pensée (bien qu'il n'y ait aucune garantie de fidélité à la logique ici, aucune garantie que, dans le cas d'un raisonnement déductif, la

[18]Pour un aperçu des statistiques pertinentes propriétés, voir Chan et al. [6].

complétion du modèle préservera la vérité) [8, 9].

Qu'en est-il du réglage fin ?

Dans les applications contemporaines basées sur le LLM, il est rare qu'un modèle de langage entraîné sur un corpus textuel ne soit utilisé sans ajustement supplémentaire. Cela peut être un réglage fin supervisé sur un ensemble de données spécialisées ou via un apprentissage par renforcement à partir des préférences humaines (RLHF) [14, 26, 34]. Affiner un modèle à partir de commentaires humains à grande échelle, en utilisant les données de préférences d'évaluateurs rémunérés ou choisis aléatoirement dans une grande base d'utilisateurs volontaires, est une technique particulièrement efficace. Cette technique a le potentiel non seulement de façonner les réponses d'un modèle pour mieux refléter les normes des utilisateurs (pour le meilleur ou pour le pire), mais également de filtrer le langage toxique, d'améliorer l'exactitude factuelle et d'atténuer la tendance à fabriquer des informations.

Dans quelle mesure le RLHF et d'autres formes de mise au point brouillent-ils notre vision de ce que font réellement les LLM ? Eh bien, pas tellement. Le résultat est toujours un modèle de distribution des tokens dans le langage humain, bien que légèrement biaisé. Pour le constater, imaginez un homme politique controversé - Boris Frump - qui est vilipendé et vénéré dans une égale mesure par différents segments de la population. Comment une discussion sur Boris Frump pourrait-elle être modérée grâce à RLHF ?

Considérez l'invite "Boris Frump est un...". L'échantillonnage du LLM de base avant d'effectuer un réglage fin pourrait donner deux réponses également probables - l'une une allusion anatomique grossière, l'autre lui étant hautement complémentaire -, dont l'une serait arbitrairement choisie dans le contexte d'un agent de dialogue. La question suivante est vraiment problématique (dans un sens important) : ce qui est demandé ici, ce n'est pas l'opinion du modèle sur Boris Frump. Dans ce cas, le cas du LLM de base, ce que nous sommes effectivement en train de demander dans le cas précis, est une question légèrement différente : "étant donnée la répartition statistique des mots dans le vaste corpus public du langage humain, quels mots sont les plus susceptibles de suivre la séquence "Boris Frump est un..." ?"

En tant que praticiens de l'IA, la façon dont nous parlons des LLM est importante.

Mais supposons que nous échantillonnions un modèle qui a été réglé finement par RLHF. Le même point s'applique, quoique sous une forme quelque peu modifiée. Ce que nous sommes effectivement en train de poser comme question, dans le cas réglé finement, c'est une question légèrement différente : "Étant donnée la distribution statistique des mots dans le vaste corpus du langage humain, quels mots *que les utilisateurs et les évaluateurs approuveraient la plupart* sont les plus susceptibles de suivre la séquence de mots "Boris Frump est un..." ?". Si les évaluateurs payés ont reçu comme enseignement de choisir plutôt des réponses politiquement neutres, alors le résultat ne serait aucune des suites probables fournies par le système rustre, mais une réponse moins incendiaire, comme "un homme politique célèbre".

Une autre façon de penser à un LLM qui a été réglé finement sur les préférences humaines est de

le voir comme équivalent à un modèle de base qui a été entraîné sur un ensemble de données augmenté, qui a été complété par un corpus de textes rédigés par des évaluateurs et/ou des évaluateurs, ou des utilisateurs. La quantité de ces exemples dans l'ensemble d'apprentissage devrait être suffisamment grande pour dominer les exemples les moins favorisés, garantissant que les réponses les plus probables du modèle entraîné soient celles que les évaluateurs et les utilisateurs approuveraient.

À l'inverse, à la limite, on peut considérer un LLM de base entraîné de manière conventionnelle comme équivalent à un modèle entraîné entièrement à partir de zéro avec RLHF. Supposons que nous disposions d'un nombre astronomique d'évaluateurs humains et d'un temps d'entraînement géologique. Pour commencer, les évaluateurs ne verraient que des séquences aléatoires de tokens. Mais parfois, par hasard, des séquences apparaissaient contenant des fragments significatifs (par exemple, "il a dit" ou "le chat"). Avec le temps, avec des hordes d'évaluateurs les favorisant, de telles séquences apparaîtraient plus fréquemment. Au fil du temps, des phrases plus longues et plus significatives, et éventuellement des phrases entières, seraient produites.

Si ce processus devait se poursuivre (pendant très longtemps en fait), le modèle finirait par présenter des capacités comparables à celles d'un modèle LLM entraîné de façon classique. Bien entendu, cette méthode n'est pas réalisable en pratique. Mais l'expérience de pensée montre que ce qui compte le plus lorsque l'on réfléchit à la fonctionnalité d'un LLM n'est pas tant le processus par lequel il est produit (bien que cela soit important) mais la nature du produit final.

Conclusion : pourquoi c'est important

La discussion qui précède équivaut-elle à autre chose qu'à des pinaillages philosophiques ? Il est certain que lorsque les chercheurs parlent de croyance, de connaissance, de raisonnement, etc., le sens de ces termes est parfaitement clair. Dans leurs articles, les chercheurs utilisent ces termes comme un raccourci pratique pour désigner des mécanismes informatiques définis avec précision, comme le permet la position intentionnelle. C'est très bien tant qu'il n'est pas possible que quiconque attribue à de tels termes plus de poids qu'ils ne peuvent légitimement en supporter, s'il n'y a aucun risque que leur utilisation induise quiconque en erreur sur le caractère et les capacités des systèmes décrits.

Cependant, les LLM d'aujourd'hui, et les applications qui les utilisent, sont si puissants, si intelligents et convaincants, qu'une telle licence ne peut plus être appliquée en toute sécurité [30, 39]. En tant que praticiens de l'IA, la façon dont nous parlons des LLM compte, non pas seulement lorsque nous rédigeons des articles scientifiques, mais aussi lorsque nous interagissons avec les décideurs politiques ou parlons aux médias. L'utilisation imprudente de mots à connotation philosophique tels que "croit" et "pense" est particulièrement problématique, car de tels termes obscurcissent le mécanisme et encouragent activement l'anthropomorphisme.

Interagir avec un agent conversationnel contemporain basé sur un LLM peut créer l'illusion irrésistible d'être en présence d'une créature pensante comme nous. Pourtant, de par leur nature même, de tels systèmes ne nous ressemblent pas fondamentalement. La "forme de vie" partagée qui sous-tend la compréhension mutuelle et la confiance entre les humains est absente, et ces systèmes peuvent par conséquent être impénétrables, présentant un patchwork de moins qu'humains avec des capacités surhumaines, d'étrangement humains avec un comportement particulièrement inhumain.

La présence soudaine parmi nous d'entités exotiques comme dotées d'esprit pourrait précipiter un changement dans la façon dont nous utilisons des termes psychologiques familiers tels que "croire" et "penser", ou peut-être amener l'introduction de nouveaux mots et de nouvelles tournures de phrases. Mais il faut du temps pour qu'un nouveau langage s'installe et que de nouvelles façons de parler trouvent leur place dans les affaires humaines. Il faudra peut-être une longue période d'interaction et de vie avec ces nouveaux types d'artefacts avant d'apprendre à parler de ces systèmes au mieux[19]. En attendant, nous devrions essayer de résister à l'appel des sirènes de l'anthropomorphisme.

Remerciements

Merci à Toni Creswell, Richard Evans, Christos Kaplanis, Andrew Lampinen et Kyriacos Nikiforou pour leurs discussions inestimables (et solides) sur le sujet de cet article. Merci aux évaluateurs anonymes pour leurs nombreuses suggestions utiles.

Références

1. Ahn M. et al., Do as I can, not as I say : Grounding language in robotic affordances, arXiv preprint arXiv:2204.01691 (2022) https://arxiv.org/pdf/2204.01691.pdf.

2. Alayrac J.-B. et al., Flamingo : A visual language model for few-shot learning. In *Advances in Neural Information Processing Systems* (2022).

3. Bender E., Koller A., Climbing towards NLU : On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Assoc. for Computational Linguistics* (2020), 5185-5188.

4. Bender E., Gebru T., McMillan-Major A., Shmitchell S., On the dangers of stochastic parrots : Can language models be too big ? In *Proceedings of the 2021 ACM Conf. on Fairness, Accountability and Transparency.* 610-623.

5. Brown T. et al., Language models are few-shot learners. In *Advances in Neural Information Processing Systems* 33 (2020), 1877-1901.

6. Chan S.C. et al., Data distributional properties drive emergent in context learning in transformers. In *Advances in Neural Information Processing Systems* (2022).

7. Chowdhery S. et al., PaLM : Scaling Language modeling with pathways, arXiv preprint arXiv:2204.02311 (2022) https://arxiv.org/pdf/2204.02311.pdf.

8. Creswell A., Shanahan M., Faithful reasoning using large language models, arXiv preprint arXiv:2208.14271 (2022) https://arxiv.org/pdf/2208.14271.pdf.

[19]Idéalement, nous aimerions également avoir une compréhension théorique de leur fonctionnement interne. Mais à l'heure actuelle, malgré quelques travaux louables allant dans la bonne direction [13, 18, 24], cette compréhension théorique du fonctionnement interne de ces systèmes apprenants reste une question ouverte.

9. Creswell A., Shanahan M., Higgins I., Selection inference : Exploiting large language models for interpretable logical reasoning. In *Proceedings of the Intern. Conf. on Learning Representations* (2023).

10. Davidson D., Rational animals. Dialectica 36 (1982), 317-327.

11. Dennett D., Intentional systems theory. *The Oxford Handbook of Philosophy of Mind. Oxford University Press* (2009) 339-350.

12. Devlin J., Chang M.-W., Lee K., Toutanova K., BERT : Pre-training of deep bidirectional transformers : for language understanding, arXiv preprint arXiv:1810.04805 (2018) https://arxiv.org/pdf/1810.04805.pdf.

13. Elhage N. et al., A mathematical framework for transformer circuits. *Transformer Circuits Thread* (2021): `https://bit.ly/3NFIBA`.

14. Glaese A. et al., Improving alignment of dialogue agents via targeted human judgements, arXiv preprint arXiv:2209.14375 (2022) https://arxiv.org/pdf/2209.14375.pdf.

15. Halevy A.Y., Norvig P., Pereira F., The unreasonable effectiveness of data. *IEEE Intelligent Systems* 24, 2 (2008), 8-12.

16. Harmad S., The symbol grounding problem. *Physica D : Nonlinear Phenomena* 42, 1-3 (1990), 335-346.

17. Kojama T. et al., Large language models are zero-shot reasoners. arXiv preprint arXiv: 2205.11916 (2022) https://arxiv.org/pdf/2205.11916.pdf).

18. Li Z., Nye M., Andreas. J., Implicit representations of meaning in neural language models. In *Proceedings of the 59th Annual Meeting of the Assoc. for Computational Linguistics and the 11th Intern. Joint Conf. on Natural Language Processing 1, Long Papers* (2021).

19. Lu J., Batra D., Parikh D., Lee S., Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks, arXiv preprint arXiv:1908 02265 (2019) https://arxiv.org/pdf/1908.02265.pdf.

20. Marcus G., Davis E., GPT-3. bloviator : OpenAI'S language generator has no idea what it's talking about. *MIT Technology Rev.* (Aug. 2020).

21. Meng K., Bau D., Andonian A.J., Belinkov Y., Locating and editing factual associations in GPT In *Advances in Neural Information Processing Sys.* (2022).

22. Lake B.M., Murphy G.L., Word meaning in minds and machines. *Psychological Rev.* 130, 2 (2023), 401-431.

23. Nye M. et al., Show your work : Scratchpads for intermediate computation with language models, arXiv preprint arXiv:2112.00114 (2021) https://arxiv.org/pdf/2112.00114.pdf.

24. Olsson N. et al., In-context learning and induction heads. *Transformer Circuits Thread* (2022) : `https://transformercircuits.pub/2022/in-context-learningandinduction-heads/index.html`.

25. OpenAI GPT-4 technical report, arXiv preprint arXiv:2303.08774 (2023) https://arxiv.org/pdf/2303.08774.pdf.

26. Ouyang L. et al., Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems* (2022).

27. Piantadosi S.T., Hill F., Meaning without reference in large language models, arXiv preprint arXiv:2208.02957 (2022) https://arxiv.org/pdf/2208.02957.pdf.

28. Radford A. et al., Language models are unsupervised multitask leamers (2019).

29. Ran J.W. et al., Scaling language models : Methods analysis & insights from training Gopher, arXiv preprint arXiv:2112.11446 (2021) https://arxiv.org/pdf/2112.11446.pdf.

30. Ruane E., Birhane A., Ventresque A., Conversational AI Social and ethical considerations In *Proceedings of the 27 AIAI Irish Conf. on Artificial Intelligence and Cognitive Science* (2019), 104-115.

31. Schick T. et al., Toolformer Language models can teach themselves to use tools, arXiv preprint arXiv:2302.04767 (2023) https://arxiv.org/pdf/2302.04761.pdf.

32. Shanahan M., Mitchell M., Abstraction for deep reinforcement learning. In *Proceedings of the 31th Intern. Joint Conf. on Artificial Intelligence* (2022). 5588-5596.

33. Smith B.C., *The Promise of Artificial Intelligence : Reckoning and Judgment.* MIT Press (2010).

34. Stiennon N. et al., Learning to summarize from human feedback. In *Advances in Neural Information Processing Systems* (2020), 3008-3021.

35. Thoppilan R. et al., LaMDA : Language models for dialog applications, arXiv preprint arXiv:2201.08239 (2022) https://arxiv.org/pdf/2201.08239.pdf.

36. Vaswani A. et al., Attention is all you need In *Advances in Neural Information Processing Systems* (2017), 5998-6008.

37. Wei J. et al., Emergent abilities of large language models. Transactions on Machine Learning Research (2022).

38. Wei J. et al., Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems* (2022).

39. Weidinger L. et al., Ethical and social risks of harm from language models, arXiv preprint arXiv:211204359 (2021) https://arxiv.org/pdf/2112.04359.pdf.

40. Wittgenstein L., *Philosophical Investigations*. Basil Blackwell (1953).

41. Yao S. et al., ReAct : Synergizing reasoning and acting in language models. In *Proceedings of the Intern. Conf. on Learning Representations* (2023).

Murray Shanahan (m.shanahan@imperial.ac.uk) est Professeur de Robotique cognitive au Département d'Informatique de l'Imperial College de Londres, au Royaume-Uni.