Compte rendu de la 77ème séance
Quand l'IA répond aux questions de sondage, 17 novembre 2025
Questions/réponses qui ont suivi la présentation d'Etienne Ollion

Question 1
Je suis doctorante au CEE. Première question : à un moment, vous mentionniez la différence entre des modèles qui sont plus fondationnels et d'autres qui sont vendus au grand public, et je ne sais pas si j'ai loupé l'information. Est-ce que vous avez trouvé des choses différentes entre ces deux modèles ? Et si c'est le cas quelles étaient les différences ? Ma deuxième question, c'est plus sur ces histoires de biais, les cas du Mexique et de la Russie où il y a des erreurs plus importantes. Est-ce possible d'aller voir dans quel sens va l'erreur, quel genre d'erreur on trouve. Par exemple est-ce qu'il y a du racisme là-dedans ?

Question 2
Je suis doctorant au CEE aussi. J'avais deux petites questions. La première sur comment appliquer cette méthode aux sciences sociales, j'imagine que, dans le monde de l'entreprise, c'est peut-être moins nécessaire, mais les mesures de validation, pour ces sondages potentiels, comment on pourrait-on valider un sondage sur une question potentiellement nouvelle, c'est ça qui serait intéressant ? Parce que là, vous le comparez les réponses IA à celles du World Value Survey, ça fait une super mesure de validation, mais comment valider s'il s'agit d'une question nouvelle qui n'a pas été posée ? On se demande un peu comment ce serait réalisable. Ensuite, petite question peut-être d'éclaircissement, pourquoi ChatGPT est-il bien moins performant que les autres modèles d'IA ?

Réponse d'Etienne Ollion
Ce sont d'excellentes questions, merci à tous les deux. Concernant la première question, c'est la différence entre les modèles « fondationnels » et les modèles dits « instruct ». Les seconds sont ceux que vous utilisez quand vous allez sur chatGPT, qui sont rendus publics. Les modèles fondationnels, c'est les autres.
Ensuite, pourquoi font-ils cela. La première chose à dire est que je n'ai pas fait d'anatomie des modèles. Je ne les ai pas mis sur une table de dissection pour comprendre pourquoi ils se comportaient ainsi, et c'est ce qu'il faudrait faire. On sait toutefois des choses. Par exemple, pour toutes les questions qui portent sur des sujets potentiellement sensibles, les modèles sont ajustés, il y a des questions auxquelles ils ne doivent pas répondre, ou seulement d'une certaine manière. Evidemment cela a un impact, et on peut imaginer que c'est ce qui se joue dans la question sur la confiance, « trust ». « Est-ce que vous avez confiance dans vos voisins ? ». Les réponses des humains sont plus ou moins variées, selon les pays. GPT, lui, il répond toujours la même chose, parce que la question est considérée comme sensible, *inappropriate*, et donc, il ne peut pas répondre.

Et l'autre question, c'était au-delà du sens va l'erreur, est-ce qu'on peut aller un peu plus loin ? Alors, le sens de l'erreur, je peux le donner, ça veut dire que la réponse de l'IA il est plus proche ou moins proche de la réponse que donneraient les individus. Par contre, savoir, si l'IA a tendance et je pense que c'est votre question, à mal caractériser en stéréotypant telle ou telle population, ça, on ne peut pas le savoir. En tout cas pas avec ce test. On pourrait, à partir des données, essayer de mesurer ça, mais je pense qu'il vaudrait mieux faire une autre expérience, et il y a des chercheurs qui font ce type d'expérience. Nous, on ne l'a pas fait. Ce serait une autre forme d'étude du biais, quand le modèle, en fait, doit imiter telle ou telle personne. Est-ce que, pour le modèle mexicain, elle lui affuble tous les stéréotypes (vous choisirez), ou bien, est-ce qu'il a une réponse un peu générique pour tout le monde, justement, pour éviter cette espèce de stéréotypisation qui peut être considérée comme extrêmement inappropriée. Les modèles ont été entraînés pour éviter ce genre de choses, mais il pourrait en rester. C'est une question empirique.

L'autre question c'est la validité temporelle. Admettons un instant que ca marche, comme ca ou autrement, qu'on puisse arriver à prédire les opinions. Comment sait-on que ça va marcher demain ? Quand on n'a pas de sondage, on ne peut pas savoir si la réponse est correcte ou pas correcte. Donc même s'il donne 100 % de réponses correctes sur toutes les questions du passé, est-ce que, sur une

nouvelle question qu'il n'avait jamais vue dans le passé ou sur une question de futur, ça peut marcher ? C'est un des problèmes que soulèvent les personnes même les plus fervents défenseurs du *silicon sampling*. Soit on a déjà les réponses et on peut comparer, soit on n'a pas les réponses et on ne sait pas le modèle a tort ou pas. Une réponse souvent donnée, c'est qu'à partir du moment où on a une bonne preuve, on va considérer que ça marche. Vous ne faites pas repasser un test à votre calculatrice, vous l'utilisez ! Je ne conseille pas de faire cela, surtout que les LLMs, contrairement à la calculatrice, sont des outils déterministes, pas stochastiques (intégrant l'aléatoire), mais qu'importe.
Une autre réponse, plus rigoureuse, serait de dire qu'on peut tester sur un petit échantillon. Par exemple, dans l'équipe CSS, on fait beaucoup d'annotations textuelles avec des modèles, et pour s'assurer que le modèle annote bien le texte, on prend un tout petit échantillon, et on va regarder s'il s'est bien comporté, sur 1 % des données, et ensuite on extrapole.

On pourrait imaginer la même logique issue du *machine learning*. Je suis surpris qu'on n'en voie pas plus. On pourrait imaginer, de faire par étape des petits bouts de la population étudiée pour toujours comparer, pour voir si le modèle a réussi ou pas. Ça coûterait beaucoup moins cher. Moi, si j'étais une entreprise de sondage qui croit dans ces techniques, c'est ce que je ferais.

Enfin, pourquoi GPT est-il moins performant ? C'est un peu la même réponse que précédemment, c'est qu'en fait, il est *instruct*, il est *fine-tuned*, il est affiné pour apporter des réponses qui restent dans le cadre des réponses considérées comme appropriées.

Question 3.
Je suis chercheuse au CEE.

Ma question porte aussi sur les biais selon les pays, selon la langue, en fait. C'est vraiment une question de clarification. D'après ce que j'ai compris, tous les *prompts*, sont en anglais, ou bien c'était aussi en espagnol pour la Mexique, en russe pour la Russie, etc. ? Ce serait intéressant d'avoir vos réflexions sur les biais qui viennent de la langue au niveau du matériel de formation ?

Question 4

Je suis à l'INED. Mais par rapport au résultat que vous avez montré, très concentré dans l'espace social, je me demande si le problème ne réside pas dans le fait d'avoir demandé à l'IA de répondre uniquement à partir d'informations socio-démographiques. En me situant un peu dans la perspective d'Alain Desrosières, pour qui le chiffre est quand même une construction sociale, je me demande si ça ne vaut pas le coup d'ajouter aussi des informations relatives au contexte de passation de l'enquête, dire, par exemple, à l'IA, est-ce que tu peux répondre comme si tu étais au téléphone avec un enquêteur, une enquêtrice, qui est un homme, une femme ?, voir si ça ne rajouterait pas de la variabilité dans les réponses.

Je me demande aussi s'il ne faudrait pas s'inspirer aussi de toute la littérature sur les méthodes cognitives dans les enquêtes. Je pense à des méthodes, par exemple, développées par René Tourangeau, par exemple, où on demande aux enquêtés de verbaliser tout processus cognitif qui les amène à répondre à une question donnée, de verbaliser tout ce qui leur passe vers la tête. Est-ce qu'on ne peut pas demander à l'IA de faire un exercice similaire, et voir comment elle aboutit à cette réponse pour essayer d'ouvrir un peu la boîte noire ?

Question 5
Je suis postdoctorant au CESDIP. Ma première question
porte sur l'utilité de ces modèles-là pour les sciences sociales. Quelle utilité, pourraient-ils avoir ? On l' imagine bien, effectivement, pour une boîte de sondage, si le but est juste d'avoir le résultat d'une élection, de le prédire correctement. Mais du point de vue des sciences sociales, dont le but est de comprendre la logique de ces résultats ? Soit il y a un genre de boîte noire dans le fonctionnement du modèle, et dans ce cas-là, on en connaît déjà le fonctionnement, ou alors il y a des nouvelles variables qu'on va regarder, mais il y a peu de chance qu'on aille regarder les dimensions qu'on ne regarde pas

déjà quand on fait de la recherche. Donc, quelle serait l'utilité ajoutée du modèle, même si ça marchait ? Bon, c'est un peu une question méchante de personnes qui voient passer quelques papiers qui, parfois, ont tendance à montrer un peu ce qu'on sait déjà, mais avec de très grosses données.

Et deuxièmement, est-ce que ce genre de recherche pourrait être utilisé pour vérifier l'utilisation de l'IA ou des LLM (Large Language Models) dans des sondages, notamment pour vérifier l'honnêteté de la production scientifique, de la de part de personnes qui ne vont pas tout déposer, tout déclarer à la CNIL, comme tu le fais. Ou chez, les étudiants, je pense notamment à ce sujet qui m'a été soufflé par une collègue, qui est confrontée parfois à des étudiants qui fabriquent, pas forcément des enquêtes, mais des entretiens. Du coup, ce serait mieux, plutôt que demander à Compilatio qui nous fait une réponse fausse une fois sur deux. Est-ce que ce genre de modèle pourrait permettre une contre-enquête, pourrait-il même être utilisé pour poursuivre les gens, faire de la répression ?

Réponse d'Etienne Ollion
Je vais commencer par les langues. Merci beaucoup pour cette question, parce qu'en effet, je ne l'ai pas dit, c'est une bonne question, c'est un point faible, et en même temps, je ne pense pas que ça change fondamentalement les résultats, mais je vais prendre le temps de l'expliquer.

Ce qu'on a fait, on a commencé par *prompter* le modèle (lui poser des questions) dans les langues dites vernaculaires. On a prompté en allemand, en russe, ça nous donnait un peu plus de travail, mais ça avait l'air, d'être très cohérent avec les réponses qu'on avait quand on promptait en anglais.

Donc, on a fini par prompter en anglais sans s'assurer absolument que le prompting dans une langue vernaculaire ne changerait pas les résultats, et donc, je ne peux pas assurer ici que ce n'est pas un pur effet de la langue. On a refait l'expérience avec un doctorant qui s'appelle Léo Labat, qui fait une thèse de NLP (Natural Language Processing)(https://theses.fr/s398266) avec François Yvon (ISIR). On a testé les modèles sur tout un tas de questions européennes, dans une dizaine de pays, à chaque fois en changeant la langue, en faisant vraiment le travail de changer la langue. Alors on ne fait pas exactement les mêmes métriques, on ne cherche pas à montrer la même chose, ce n'est pas exactement le même cadre, mais ce qu'on montre, c'est que les modèles sont parfois polyglottes, c'est-à-dire qu'ils vont répondre la même chose quel que soit le pays, et parfois pas, et bien malin qui arriverait à trouver le pattern, la logique. Donc la langue ne changerait rien. Mais Léo montre que cela joue aussi, parfois.

Nous, on n'a même pas complètement réussi à mettre au jour ce phénomène. Donc je pense que ça aurait changé, comme pas mal d'autres choses que j'ai montrées ici, la forme des résultats, mais je ne pense pas que ça aurait changé l'esprit des résultats. A vérifier donc. Par ailleurs, on a fait d'autres choses dans ce papier avec Léo Labat et François Yvon, par exemple, on a décidé de rajouter, ou pas, un espace à la fin de la question. Quand on fait un peu de théorie et pratique de l'informatique, – j'en parle librement, je ne fais sérieusement ni l'un ni l'autre-, on sait que rajouter un espace, ça fait sens, c'est un autre *token* (un segment de mot), c'est une autre manière d'écrire, et ça va produire des résultats tout à fait différents. Ca ne manque pas, ça produit des réponses différentes de simplement rajouter un espace à une question. Pareil si, au lieu de poser les questions en disant A, B, C, D, vous dites 1, 2, 3, 4, ou Alpha, Beta, Delta, Gamma, il va répondre encore un peu différemment.

Si vous inversez, c'est encore plus fourbe, mais soyons fourbes, si vous inversez les réponses, au lieu de dire est-ce que vous avez confiance dans votre voisin, « un peu, beaucoup, passionnément, à la folie, pas du tout », vous dites « un peu, pas du tout, passionnément, un peu, à la folie », bref, si vous changez l'ordre, le modèle répond différemment et il n'est pas toujours cohérent, en fait.

Le papier avec Léo Labat (under review...) permet une mesure de cohérence, en faisant des changements qui sont des changements sémantiques non significatifs, ça ne change rien. Un humain, normalement, n'est pas censé répondre autre chose si on dit A, B, C, D, ou 1, 2, 3, 4, mais le modèle lui va, parfois, répondre à côté.

Donc ce sont de vraies bonnes questions qui, à mon avis, appellent quand même un minimum de prudence dans l'usage de ces modèles, d'autant plus que les gens qui les utilisent, en général, ne font pas ça, ils ne rajoutent pas nécessairement toute une série de variables, parce que ça prend du temps, de l'énergie (*compute power*).

L'autre question, si je comprends bien est-ce que ce n'est pas simplement lié au fait que c'est des variables sociodémographiques qui sont fournies au modèle et que si je l'avais informé sur le contexte, il aurait peut-être réagi autrement. Le mettre en situation d'entretien, en fait. Ça ne marcherait de toute façon que pour les modèles *instruct*, pas pour les autres auxquels il faut vraiment donner des instructions, apprendre à répondre. Bon. Si on lui donne le contexte en plus, admettons qu'il le comprenne, qu'il réponde bien, est-ce que ça changerait quelque chose ? Oui, probablement, mais quoi, et dans quel sens…

Est-ce qu'un modèle répondrait plus… Peut-être. Il faudrait le tester. Est-ce qu'un modèle répondrait plus diversement si on lui disait « attention, tu peux répondre maintenant parce que tu es en entretien avec un enquêteur », alors qu'avant on lui pose des questions et il répond ? Ça voudrait dire qu'en fait, on lui donne un degré de liberté, c'est une opinion, ce n'est pas un fait, quoi.
C'est ce que vous avez en tête ?

Intervenant
Oui, c'était ça l'idée, qu'on lui donne la possibilité de répondre ainsi. Je pense, à des travaux comme ceux Arnaud Régnier-Loilier, par exemple, qui pointent vers l'impact de la présence d'un tiers lors de la passation du questionnaire sur la répartition des tâches dans le couple[1]. Ce n'est pas neutre, par exemple, les hommes ont tendance à donner des réponses qui surévaluent leur implication dans ces tâches, en l'absence de leur conjointe. Ça dit quelque chose, de l'implication des hommes dans les tâches ménagères, au-delà de la question, en fait, telle qu'elle est posée dans le questionnaire.
Enfin, il y a quand même des effets de désirabilité sociale au moment de la passation du questionnaire, qu'on ne capte pas, là, il me semble.

Étienne Ollion
Alors là, c'est vraiment l'anthropomorphisation jusqu'au bout. Déjà se dire que le modèle « en situation d'entretien » va répondre, c'est imaginer qu'il se comporte comme un humain, et pas comme un générateur de distribution statistique (ce qu'il est), mais se dire qu'il y a un biais de désirabilité quand on dit que le conjoint est présent... J'ai du mal à y croire, mais tant qu'on ne l'a pas testé, on ne peut pas savoir.
Alors, on ne l'a pas testé, mais on a testé autre chose, une question que les informaticiens nous posent parfois. Je ne sais pas si vous le savez, mais pour faire varier la distribution des réponses dans un LLM, on a un paramètre qui s'appelle la température, et donc une température qui est très proche de zéro, le modèle va répondre presque toujours la même chose, il devient quasi déterministe, et plus on augmente la température, plus le modèle va avoir tendance à répondre de manière plus diversifiée.
Alors, ce n'est pas exactement votre question, mais ça me fait penser à ça, parce que ça dit quand même quelque chose de plus proche, plus on te donne plus de marge de manœuvre. En gros, plus vous êtes près de zéro, plus il répondra toujours la même chose, et plus vous allez au-delà de 1, 2, 3, etc., et plus le modèle devient créatif, inventif, etc. Bon, et comme on le voit sur cette slide, ca ne change rien.

Enfin, sur la question de l'utilité de ces modèles pour les sciences sociales ?
En gros, si on nous proposait demain un modèle, ayant la capacité à connaître le monde immédiatement, si je lui disais, tiens, que pensent les Français de Laurent Wauquiez, de l'antisémitisme, du changement climatique, tout cela groupe social par groupe social. Est-ce qu'on ne voudrait pas savoir, si on était sûrs que ca marche ? Je pense que si.

---

[1]Régnier-Loilier Arnaud. Conditions de passation et biais occasionnés par la présence d'un tiers sur les réponses obtenues à l'enquête Érfi. In : Économie et statistique, n°407, 2007. pp. 27-49

Par contre, ce qu'on n'aura pas, là, je te rejoins complètement, c'est la question du mécanisme, ou des représentations qui ont amené à prendre cette décision-là. Et ça, d'une certaine manière, on ne l'avait déjà pas avec l'enquête classique par questionnaire. L'enquête par questionnaire, elle est faite pour agréger les réponses des gens relativement un objet. Elle n'est pas faite pour nous donner le processus par lequel les gens sont arrivés à cette question-là. C'est pour ça, d'ailleurs, qu'on fait des entretiens, qu'on fait de l'observation, etc. Et donc, c'est sûr que pour le coup, on ne le répliquera pas jusqu'au moment où on arrivera à montrer que les raisonnements des LLM sont les mêmes raisonnements que ceux des humains. Mais là, et je n'avais pas répondu là-dessus, je pense que ce n'est pas une solution.

Enfin, est-ce que ça permet de faire de la répression estudiantine, en détectant les faux entretiens ? Je n'ai pas de réponse à ça, en fait. Je ne sais pas spécialiste de ce domaine… Enfin, les dernières publications que j'ai vues sur le thème : est-ce qu'un LLM est capable de détecter si un autre LLM a fait le travail, c'est oui dans 50 % des cas. Une fois, ça marche, une fois, ça ne marche pas. Et c'est un jeu de chat et la souris. C'est comme le *hacking*. Il n'y a jamais personne qui arrive à dire de manière définitive, je peux vous assurer que ça, c'est complètement sécure. La capacité des modèles à imiter des étudiants ira de pire en pire, mais les techniques deviendront meilleures et meilleures.

Question 6
Moi, j'ai peut-être une autre petite question pour continuer un peu sur cette lancée et de façon peut-être un peu provoquante, mais puisque vous n'avez pas tout à fait répondu à cette question. Du coup, quid des super nouveaux modèles, bien plus grands, avec beaucoup plus de paramètres, beaucoup plus de langues ? Est-ce que ce n'est pas la solution ? Notamment pour ces langues qui n'étaient pas très présentes dans les premiers modèles, ou pour les modèles plus petits ?

Question 7
Et quid de DeepSeek, puisque tu n'as rien dit ? Je suis curieux de savoir si vous aviez déjà essayé de faire ça, parce qu'à priori, justement, c'est un modèle plus léger et plus efficace.

Réponse d'Etienne Ollion
Je suis en train de regarder les modèles qui fonctionnaient bien pour le travail avec Léo Labat. Ce sont des modèles récents, Gemma, Mistral, Llama, des modèles très très récents et très très gros. Llama70b, c'est un des plus gros en ce moment. DeepSeek, c'est vrai qu'on ne l'a pas là. Qui sait ?

Mais le tour de force de DeepSeek, ce n'était pas qu'il était meilleur, c'est qu'il était aussi bon et en beaucoup moins de temps d'entrainement, qu'il avait eu besoin de beaucoup moins de temps de calcul pour arriver à des performances. Peut-être, mais encore une fois, on ne sait pas. Je suis sceptique, mais sceptique dans le sens du « scepticisme organisé » que Merton appelait les scientifiques à pratiquer. Je ne suis pas sûr, et peut-être qu'il va se passer des choses. J'ai de bonnes raisons de penser que ça ne marchera pas sous cette forme-là, et qu'il y a plein de limites. Mais je ne vais pas dire que ça ne marchera jamais, autrement, un peu…

Déjà, je suis très mauvais en prédiction et en prophétie, et ce ne serait pas une attitude scientifique. Après, je vois pas mal de difficultés à anticiper et donc des usages qui restent pour l'instant, à mon avis, un peu limités.

Question de Nonna Mayer
Moi, j'ai envie de poser une question. Tu disais tout à l'heure qu'un humain ne réagirait pas, ne répondrait pas différemment si on changeait ABCD en 1, 2, 3, 4 ou si on inversait l'ordre des modalités de réponse. Mais si, justement. Il y a plein d'expériences qui montrent que si on inverse les modalités de réponse, ça a des conséquences. Est-ce que ça ne vaudrait pas le coup de faire un travail systématique, comme sur les humains, sur tes modèles, pour voir si l'ordre des modalités de réponses, et des questions dans le questionnaire, ce genre d'expérimentation, je pense que ça serait intéressant pour voir les capacités cognitives de ton modèle.

Autre chose, pour revenir à ce qu'on disait tout à l'heure, qu'est-ce qu'on peut faire pour compenser les biais des sondages en général, qu'ils soient siliconés ou pas. Recourir aux *big data* ? Parce qu'avec tous leurs défauts, on a aujourd'hui des masses de données gigantesques à disposition. Je pense à cette expérience sur le vote Obama aux États-Unis. Était-il biaisé par le racisme ? Un chercheur, dont le nom m'échappe, mais tu le connais, oui, Seth Stephen Davidovitz. Il s'est contenté de regarder sur plusieurs années le nombre de recherches sur Google du mot « *nigger* », mot tabou aux États-Unis et leur répartition sur le territoire. C'est instructif. Cela lui a permis de dresser une cartographie du racisme aux États-Unis, beaucoup plus large que celle que donnaient les sondages d'opinion. Surtout il a trouvé surtout que si les sondages d'opinion ne montraient que peu de relation ou très peu de relation entre le niveau de racisme antinoir le vote pour Obama, lui trouvait une très nette corrélation négative entre nombre de recherches sur le N-word et faiblesse des scores Obama comparés au reste du territoire et à ceux c des candidats démocrates dans le passé. C'est une piste à explorer. Comment est-ce que tu mettrais en parallèle ces deux types de données ?

Donc, sur les sondages, en même temps que je disais que les humains sont capables de faire le calcul et donc de retomber sur leur patte, même si j'inverse l'ordre des questions, je vais avoir une bonne réponse, en même temps que je le disais, je me disais ben non, on sait bien et on a vu que, en fait, selon l'endroit où on pose la question, on va avoir des réponses différentes.

Réponse d'Etienne Ollion
Certes, les individus varient dans leurs réponses, c'est bien documenté. Mais les pourcentages que trouve Léo Labat dans son analyse sont bien plus forts que chez les humains. Il faudrait tester. Si on veut rentrer dans cette perspective que moi je trouve un peu anthropomorphique, c'est revenir à la question précédente, et se demander s'il y a un effet enquêteur ? Est-ce qu'il y a un effet cadre de passation ? Toutes ces choses qu'on a très bien montrées sur les enquêtes par questionnaire depuis 50 à 70 ans. Vous êtes à l'INED, la question s'y pose en permanence. C'est quelque chose de central. On fait une enquête sur la sexualité, on ne pose pas la question devant la conjointe. On pose la question au téléphone, etc. Marie Bergström va justement venir nous en parler dans ce séminaire en décembre[2]. Mais je pense que ce n'est pas ce niveau d'erreur qu'on voit sur les LLM.
Les humains ont quand même une particularité, si on leur disait qu'on modifie l'ordre d'une échelle ordonnée, ils se diraient qu'il y a un piège, ils seraient plus attentifs, et donc on éviterait ce problème. Par contre, en effet, est-ce que les LLM ont tendance à privilégier certaines choses ?
Il y a un article très marrant de Paul Röttger, qui est un linguiste, qui s'appelle « My Answer Is C »[3], parce que le modèle a tendance, comme les humains, à privilégier certaines réponses, selon l'endroit où on met la réponse possible, etc. Donc ça a déjà été un peu fait. Là, c'était plutôt voir s'ils ont tendance à privilégier un type de modalité selon l'ordre qui est proposé.

Sur les big data, les données numériques massives, c'est quelque chose qui m'a beaucoup intéressé, et qui continue de m'intéresser. C'est des données intéressantes, mais avec leur biais comme les autres. Donc il faut regarder, sans leur conférer une forme d'extraterritorialité épistémologique (parce qu'elles sont autres, on les juge autrement).

Alors, dans le cas des recherches Google, moi, je me méfie beaucoup, parce qu'il y a eu un enthousiasme absolument incroyable autour de Google, les recherches Google sur les symptômes de la grippe vont remplacer les autres données, etc., je pense aux déclarations du Center for Disease Control sur la grippe, et en fait, ça a été un flop. Il y avait eu un enthousiasme de gens qui disaient, regardez, ça corrèle extrêmement bien. Et puis en fait, on s'est rendu compte que très peu. C'est le genre

---

[2]Seth Stephen-Davidowitz, *Everybody Lies : What the Internet Can Tell Us About Who We Really Are*, Londres, Bloomsbury, 2017
[3]Xinpeng Wang, Bolei Ma, Chengzhi Hu, Leon Weber-Genzel, Paul Röttger, Frauke Kreuter, Dirk Hovy, Barbara Plank, My Answer is C »: First-Token Probabilities Do Not Match Text Answers in Instruction-Tuned Language Models, *ArXiv*, 2024, https://arxiv.org/abs/2402.14499

d'enthousiasme initial, qui a disparu dans les poubelles de l'histoire des sciences. assez rapidement. Donc, si on a zéro information, c'est peut-être mieux que rien, mais si on a un très beau système de détection et de recollection d'informations sur la prévalence des maladies, je pense qu'il ne faut pas arrêter de le financer, voilà. Ce sera mon mot de la fin.

Intervention de George Marcus (professeur émérite, Williams College)
Je parle en anglais parce que j'ai plusieurs choses à dire. La question de la représentativité est beaucoup plus compliquée que ce que le LLM produit. Prenez la question du bonheur.

Si vous demandez aux êtres humains cette question, il y a tellement de variables qui interférent sur la réponse, mises en lumière par les psychologues, du temps du jour aux événements de votre vie. Et nous créerons des théories qui vous expliqueront pourquoi cela change. Est-ce que les modèles LLM essayent d'imiter le jugement des êtres humains ? Ou est-ce la réponse donnée à un moment particulier selon leur propre compréhension intérieure de ce que les êtres humains devraient faire s'ils sont raisonnables, rationnels, ou émotionnels ? Ce sont deux dynamiques très différentes.
Et, en tant que scientifiques nous sommes généralement si loin du design des LLM que nous ne sommes pas même en position de comprendre exactement ce qu'ils font.

Mais vous pouvez utiliser d'autres techniques de validation pour, déjà, voir ce que les modèles LLM font. Par exemple, il y a un certain nombre de dynamiques connues quand les êtres humains répondent par exemple, à des questions sur le bonheur. Si vous demandez à quelqu'un « Pensez que vous êtes heureux aujourd'hui », vous recevrez une réponse différente que si vous demandez « dans quelle mesure vous sentez-vous heureux aujourd'hui ». C'est bien connu, il y a 40 ans de recherche sur le sujet. Est – ce que sera pareil pour les modèles LLM ?

Une autre chose que vous pouvez faire c'est voir ce que les modèles disent aujourd'hui sur des événements prévisibles et d'autres qui le sont moins, voir ce que les modèles LLM font à travers ces fenêtres de temps. Cela fait un demi-siècle que je fais de la *survey research*. On en sait beaucoup sur ce qui rend les enquêtes meilleures ou pires, quelle que soit la technique, face-à face, téléphone, etc. On ne sait pas ce qui rend les modèles LLM bons, parce qu'on n'est pas même sûr de savoir ce qui est bon.
Mais, vraisemblablement, dans 20 ans, pas dans ma vie, mais peut-être dans la vôtre, il y aura beaucoup plus de technologie à construire pour comprendre quand les modèles LLM sont bons et sur quoi.

Autre chose, la plupart des critères que vous utilisez pour entraîner les modèles ne sont pas intéressants, non parce qu'ils sont inintéressants en soi, mais parce qu'ils sont si stables. Les données socio démographiques, n'importe quel modèle peut aller chercher les données de recensement du pays, le moment n'a pas d'importance. La pratique religieuse par exemple, cet différent,
Elle n'est pas très stable, ça dépend à quelle saison vous posez la question, pour les êtres humains j'entends. Aucun être humain ne dit « je vais vous donner ma pratique religieuse moyenne sur l'année ».
Qu'il soit chrétien, juif, musulman, ce sont les dates qui sont spécifiques et sensibles qui vont conduire les sondés à surestimer ou sous-estimer leur pratique. Il y a des sujets où les opinions bougent beaucoup, où la moindre erreur d'échantillonnage peut être fatale. Les modèles LLM capturent-ils ces mouvements ou pas ?

Réponse d'Etienne Ollion
Je vais juste répondre en quelques mots. Je ne suis pas sûr d'avoir envie de me lancer dans la recherche qui consisterait chercher comment remplacer les sujets humains. Il y aurait donc beaucoup à dire, mais je ne suis pas spécialiste, et je ne veux pas le devenir. Enfin voyons. Pour la fréquentation de l'église, que vous mentionnez. Le moment où vous posez une question, c'est important pour les humains, mais je ne pense pas que les LLMs soient aussi sensibles à ça. Je dirais même que c'est le contraire. Je pense qu'en fin de compte, ce que vous signalez, et c'est une des questions à laquelle nous n'avons pas répondu, c'est que la plupart des critiques que vous faites, sont en fait des critiques

qui pourraient être faites à ce type d'enquête en général et notamment au WVS (World Values Survey). Certaines personnes nous ont dit que le WVS n'est pas un bon sondage. Donc si je vous suis, on pourrait dire que les LLMs feront, à la limite mieux, que le WVS. C'est peut-être le cas. Mais ce serait un cas rare.