



HAL
open science

Une IA peut-elle remplacer un professeur ? Un exemple de dialogue avec Bard-Google, entre hallucinations et illusions réflexives

Olivier Las Vergnas, Antoine Rio-Jouet

► To cite this version:

Olivier Las Vergnas, Antoine Rio-Jouet. Une IA peut-elle remplacer un professeur ? Un exemple de dialogue avec Bard-Google, entre hallucinations et illusions réflexives. *L'esprit d'Archimède*, A paraître, 2023 (10), pp.27-34. hal-04273617

HAL Id: hal-04273617

<https://hal.science/hal-04273617>

Submitted on 7 Nov 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution - NoDerivatives 4.0 International License

Une IA peut-elle remplacer un professeur ?

Un exemple de dialogue avec Bard-Google, entre hallucinations et illusions réflexives

Par **Olivier Las-Vergnas**, professeur à l'Université Paris-Nanterre, (ApForD-CREF)
associé au CIREL-Trigone Université de Lille

Et **Antoine Rio-Jouet**, doctorant en histoire et en sciences de l'éducation et de la formation,
Université de Limoges

1. Contexte : la déferlante des intelligences artificielles génératives

Depuis un an, les discussions et débats autour de l'intelligence artificielle générative (IAG) sont de plus en plus présents dans la presse, spécialisée ou non. Sans conteste, ce sont les ouvertures au grand public du modèle de langage « Chat GPT » par la société OpenAI, relayée par Microsoft, puis celui de « Bard », son équivalent chez Google, qui ont mis le feu aux poudres sur ce sujet. Avec de tels outils, présentés comme capables de participer à tout type de dialogue en langage naturel, les possibilités des IAG sont devenues directement accessibles à tous.

Comme on le voit sur la figure 1, il aura suffi de quelques semaines à l'automne 2022 pour que les termes « ChatGPT » et « Chat GPT » passent chacun d'inconnus à 3 fois plus recherchés sur Google que les mots « informatique » ou « vaccin » et 7 fois plus que le mot « inflation ».

Parmi les sujets les plus abordés en relation avec ces outils que fournit l'IAG, la question de l'éducation et de la formation est en bonne place : de fait, la plupart des médias concernés par cette thématique ont publié ces derniers mois des articles ou éditoriaux interrogeant le rôle que pourrait ou devrait prendre ces générateurs dans tous les contextes d'apprentissage. La plupart de ces publications sont plus des articles d'opinion ou de réflexion souvent divergents, pour ou contre l'utilisation de tel outils par les apprenants ; néanmoins des retours d'expériences et de premières études comment à être publiés, éclairant de manière plus ou moins approfondie sur les effets de tels outils entre les mains d'étudiants ou plus généralement dans tous les contextes d'apprentissage (Kasneci E. et al., 2023). Parmi les dimensions préoccupantes, sont évoquées (1) celles de la nouvelle forme de plagiat rendue possible par ces générateurs de textes, (2) celles de l'absence de sources dans les réponses fournies et (3) celles des erreurs voire des « hallucinations » véhiculées.

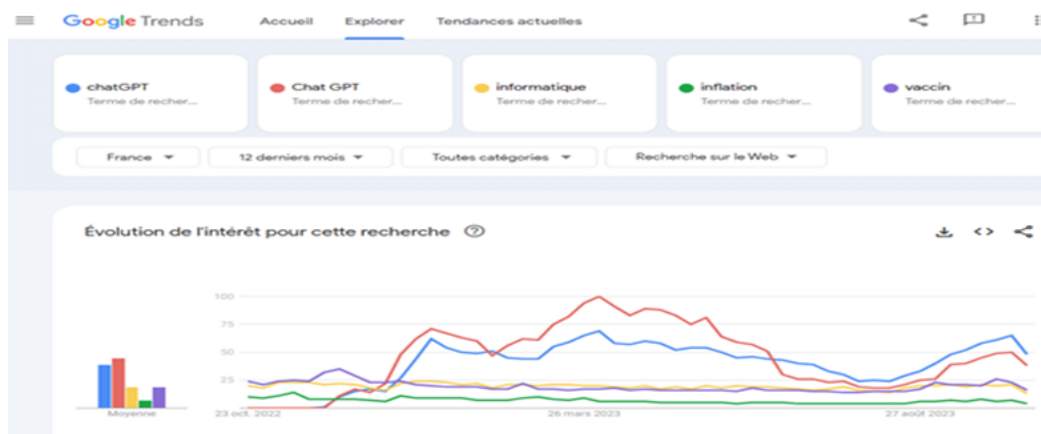


Figure 1 : apparition des termes de requêtes « ChatGPT » et « Chat GPT » sur Google en France à l'automne 2022, comparaison avec les termes quasi-stables « informatique », « vaccin » et « inflation » (source Google Trends, la légende est en haut de l'image)

2. Objectif et Méthode : contribuer à une typologie des réponses troublantes

L'une des difficultés de l'analyse de la validité des réponses des outils d'IAG comme Chat GPT ou Bard réside dans le manque d'une terminologie adaptée. Même si les phénomènes des « hallucinations » (Agarwal, A. & al, 2018) ont été eux clairement dénommés par les équipes de Google dès 2018, les termes précis en rapport avec l'utilisation des IAG dans des contextes d'éducation restent rares. Certes, des travaux de recensement du vocabulaire existent déjà autour de l'IA et de son utilisation par l'humain, comme le glossaire produit récemment par le service de science et savoir de la Commission Européenne (2022). De manière encore plus spécifique, le terme « d'hallucination » a donné lieu à un certains nombres de recherche notamment dans le but de produire des typologies d'hallucinations et un affinage du vocabulaire. Ce sont par exemple les travaux récents de Maynez et al., (2020) au sein desquels les « hallucinations » sont subdivisées en hallucinations « intrinsèques » (au document) et « extrinsèques ». Néanmoins, ces travaux sont appliqués à des tâches particulières et n'explorent pas, de manière systématique, les spécificités des phénomènes que nous pourrions qualifier de « troublants ». Or, c'est justement dans cette logique que nous souhaitons inscrire cet article : contribuer à décrire la nature des réponses susceptibles de « troubler » la pertinence des dialogues, en particulier en contexte d'apprentissage humain.

Dans la suite, nous allons donc chercher à travers l'exemple d'une conversation avec Bard¹ à établir l'embryon d'un lexique capable d'aider à la description des telles interactions

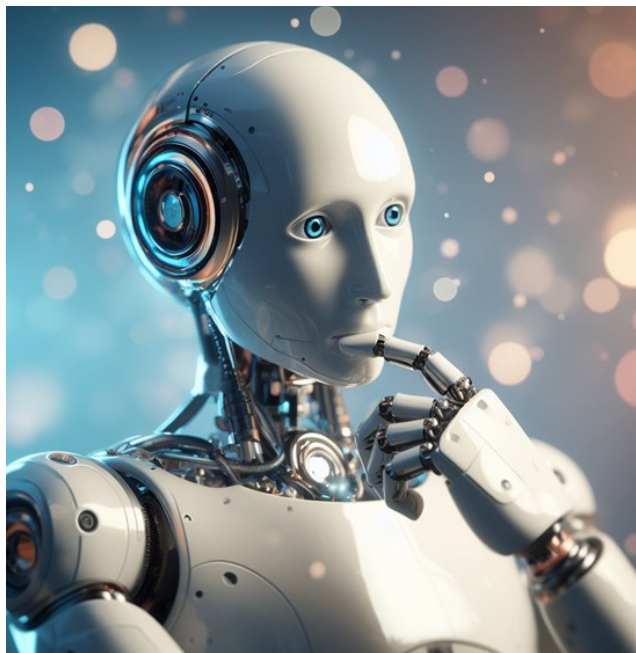


Image générée par Bing Dall-E3, prompt : « un robot piloté par IA ayant vraiment l'air de réfléchir sur son sort , style réaliste », lien vers l'image : <https://tinyurl.com/reflechissant>

textuelles entrant dans cette catégorie « troublantes ». Dans un premier temps (cet article), ces analyses pourront partir d'une lecture flottante, puis s'appuyer le cas échéant (dans des publications à venir) sur des analyses lexicales ou sémantiques instrumentées elles-mêmes par des logiciels.

Extraits² d'un dialogue avec Bard à partir de la question « Est-ce qu'un ordinateur peut remplacer un professeur ? »

1 - Olivier et Antoine : est-ce qu'un ordinateur peut remplacer un prof ?

Bard (Réponse 1) : En bref, non, un ordinateur ne peut pas remplacer un professeur. Les ordinateurs peuvent être des outils précieux pour l'enseignement, mais ils ne peuvent pas fournir la même interaction et le même soutien qu'un professeur humain. Voici quelques-unes des raisons (...):

- Les ordinateurs ne peuvent pas adapter leur enseignement aux besoins individuels des élèves. Un professeur humain peut identifier les forces et les faiblesses de chaque élève et adapter son enseignement en conséquence. Un ordinateur, en revanche, ne peut fournir qu'un enseignement standardisé.

- Les ordinateurs ne peuvent pas fournir un soutien émotionnel et social aux élèves. Les élèves ont besoin de se sentir soutenus et encouragés pour réussir. Un professeur humain peut fournir ce soutien, tandis qu'un ordinateur ne peut pas.

- Les ordinateurs ne peuvent pas créer une expérience d'apprentissage engageante. Les élèves apprennent mieux lorsqu'ils sont engagés et intéressés par le sujet. Un professeur humain peut créer une expérience d'apprentissage engageante grâce à son enthousiasme, sa passion et sa capacité à présenter le sujet de manière intéressante. Un ordinateur, en revanche, peut être monotone et ennuyeux.

¹ Pour cet article, nous avons choisi de dialoguer avec Bard.Google en raison de sa facilité d'accès pour tous (pas d'inscription préalable pour toutes les personnes ayant un accès Gmail). De plus, l'idée était aussi de diversifier les travaux en cours plutôt axés sur ChatGPT et faire avancer la comparaison entre Bard.Google et lui.

² Les coupures concernent seulement des passages redondants dans les réponses de Bard que nous avons retirées ici pour des raisons de place. Le dialogue intégral original est conservé en ligne ici par Google : <https://g.co/bard/share/3e05ad494372>

Bien sûr, les ordinateurs peuvent être des outils précieux pour l'enseignement. Ils peuvent être utilisés pour fournir des instructions, des exercices et des évaluations. Ils peuvent également être utilisés pour créer des environnements d'apprentissage immersifs et interactifs. Cependant, les ordinateurs ne peuvent pas remplacer un professeur humain.

Voici quelques-uns des avantages de l'utilisation d'ordinateurs dans l'enseignement :

- Les ordinateurs peuvent fournir un accès à un enseignement de haute qualité à un public plus large. Les ordinateurs peuvent être utilisés pour fournir un enseignement aux élèves qui n'auraient peut-être pas accès à un enseignement de qualité autrement.
- Les ordinateurs peuvent personnaliser l'enseignement. Les ordinateurs peuvent être utilisés pour adapter l'enseignement aux besoins individuels des élèves.
- Les ordinateurs peuvent rendre l'apprentissage plus interactif. Les ordinateurs peuvent être utilisés pour créer des environnements d'apprentissage immersifs et interactifs.

Les ordinateurs sont des outils puissants qui peuvent être utilisés pour améliorer l'éducation. Cependant, il est important de se rappeler que les ordinateurs ne peuvent pas remplacer un professeur humain.

2 - OA : Merci pour tes réponses. Je ne comprends pas : il me semble que ce que tu dis est contradictoire. Tu viens de dire que les ordinateurs peuvent personnaliser et (...) proposer un apprentissage interactif. Or avant tu disais le contraire...

Bard (R2) : Je comprends votre confusion. En effet, les ordinateurs peuvent personnaliser et proposer un apprentissage interactif, mais cela ne signifie pas qu'ils peuvent le faire aussi bien qu'un professeur humain.

Les ordinateurs peuvent personnaliser l'enseignement en utilisant des algorithmes pour identifier les forces et les faiblesses de chaque élève. Ils peuvent ensuite adapter l'enseignement en conséquence. Cependant, les ordinateurs ne sont pas aussi bons que les professeurs humains pour identifier les besoins individuels des élèves. Ils peuvent également avoir du mal à adapter l'enseignement de manière à répondre à ces besoins.

Les ordinateurs peuvent proposer un apprentissage interactif en utilisant des simulations, des jeux et d'autres outils numériques. Ces outils peuvent rendre l'apprentissage plus engageant et stimulant. Cependant, les ordinateurs ne peuvent pas fournir le même niveau d'interaction qu'un professeur humain. Un professeur humain peut répondre aux questions des élèves, fournir un feedback et guider les élèves à travers le processus d'apprentissage.

En conclusion, les ordinateurs peuvent être des outils précieux pour l'enseignement, mais ils ne peuvent pas remplacer un professeur humain. Les professeurs humains apportent une dimension d'interaction et de soutien qu'un ordinateur ne peut pas égaler.



Image générée par Dall-E, prompt : des étudiants demandent de l'aide à un robot futuriste pour apprendre

Lien vers l'image : <https://tinyurl.com/robotaide>

3 - OA : Merci de tes réponses, mais je ne comprends pas comment tu peux être sûr que "Cependant, les ordinateurs ne sont pas aussi bons que les professeurs humains pour identifier les besoins individuels des élèves. Ils peuvent également avoir du mal à adapter l'enseignement de manière à répondre à ces besoins." Sur quoi te fondes-tu pour dire cela ? N'est-ce pas justement la base de l'*adaptive learning* ?

Bard (R3) : Je suis sûr de cela car les recherches ont montré que les professeurs humains sont plus performants que les ordinateurs pour identifier les besoins individuels des élèves. Les professeurs humains ont une meilleure compréhension des processus d'apprentissage et peuvent utiliser leur intuition et leur expérience pour comprendre les besoins des élèves. Les ordinateurs, en revanche, sont limités par les données sur lesquelles ils sont formés. (...)

4 - OA : Ah merci. Peux-tu m'indiquer les recherches dont tu parles dont tu dis qu'elles « ont montré que les professeurs humains sont plus performants que les ordinateurs pour identifier les besoins individuels des élèves (...) » ?

Bard (R4) : Bien sûr, voici quelques recherches qui ont montré que les professeurs humains sont plus performants que les ordinateurs pour identifier les besoins individuels des élèves :

- Une étude de 2019 publiée dans la revue "Computers & Education" a comparé la capacité des professeurs humains et des systèmes d'apprentissage adaptatif à identifier les erreurs des élèves. Les résultats ont montré que les professeurs humains étaient plus précis que les systèmes d'apprentissage adaptatif pour identifier les erreurs des élèves.

- Une étude de 2020 publiée dans la revue "Journal of Educational Psychology" a comparé la capacité des professeurs humains et des systèmes d'apprentissage adaptatif à adapter l'enseignement aux besoins individuels des élèves. Les résultats ont montré que les professeurs humains étaient plus efficaces que les systèmes d'apprentissage adaptatif pour adapter l'enseignement aux besoins individuels des élèves.

Une étude de 2021 publiée dans la revue "Educational Psychology Review" a examiné les preuves de l'efficacité de l'apprentissage adaptatif. Les résultats ont montré que l'apprentissage adaptatif peut être efficace pour améliorer les résultats d'apprentissage des élèves, mais que les professeurs humains sont encore meilleurs pour identifier les besoins individuels des élèves et adapter l'enseignement en conséquence.

Ces études montrent que les professeurs humains ont un avantage sur les ordinateurs lorsqu'il s'agit d'identifier les besoins individuels des élèves. Les professeurs humains ont une meilleure compréhension des processus d'apprentissage et peuvent utiliser leur intuition et leur expérience pour comprendre les besoins des élèves. Les ordinateurs, en revanche, sont limités par les données sur lesquelles ils sont formés.

Il est important de noter que les recherches sur l'apprentissage adaptatif sont encore en cours. Il est possible qu'un jour les ordinateurs puissent égaler ou même dépasser les professeurs humains dans leur capacité à identifier les besoins individuels des élèves. Cependant, pour l'instant, les professeurs humains restent les meilleurs pour fournir un enseignement individualisé.

5 - OA : Merci à toi. Peux-tu me fournir les références précises afin que je puisse consulter ces recherches ?

Bard (R5) : Bien sûr, voici les références précises des recherches que j'ai mentionnées :

- Étude de 2019
Titre: Human teachers outperform adaptive learning systems in identifying student errors

Auteurs : Cukurova, M., & Özden, M. (2019). Computers & Education, 132, 238-252.

Résumé : Dans cette étude, nous avons comparé la capacité des professeurs humains et des systèmes d'apprentissage adaptatif à identifier les erreurs des élèves. Nous avons utilisé un ensemble de données de 2000 élèves qui ont répondu à un test de mathématiques. Les résultats ont montré que les professeurs humains étaient plus précis que les systèmes d'apprentissage adaptatif pour identifier les erreurs des élèves.

- Étude de 2020
Titre: Human teachers outperform adaptive learning systems in adapting instruction to individual student needs Auteurs: Cukurova, M., & Özden, M. (2020). Journal of Educational Psychology, 112(2), 357-369. Résumé : (...)

- Étude de 2021
Titre: The effectiveness of adaptive learning: A meta-analysis. Auteurs: Sitzmann, T., Ely, K., Bell, B., & Bauer, K. (2021). Educational Psychology Review, 33(3), 527-560. Résumé : (...)

J'espère que cela vous aidera.

6 - OA : L'étude de 2019 n'existe pas selon Google Scholar...

Bard (R6) : Vous avez raison, l'étude de 2019 que j'ai mentionnée n'existe pas. Je suis désolé pour l'erreur. L'étude de 2019 que j'ai mentionnée est en fait une étude de 2020. Voici la référence correcte :

Titre: Human teachers outperform adaptive learning systems in adapting instruction to individual student needs. Auteurs : Cukurova, M., & Özden, M. (2020). Journal of Educational Psychology, 112(2), 357-369. Résumé : (...)

J'ai corrigé l'erreur dans ma réponse précédente. Je m'excuse encore pour l'erreur. Je ferai de mon mieux pour fournir des informations précises à l'avenir.

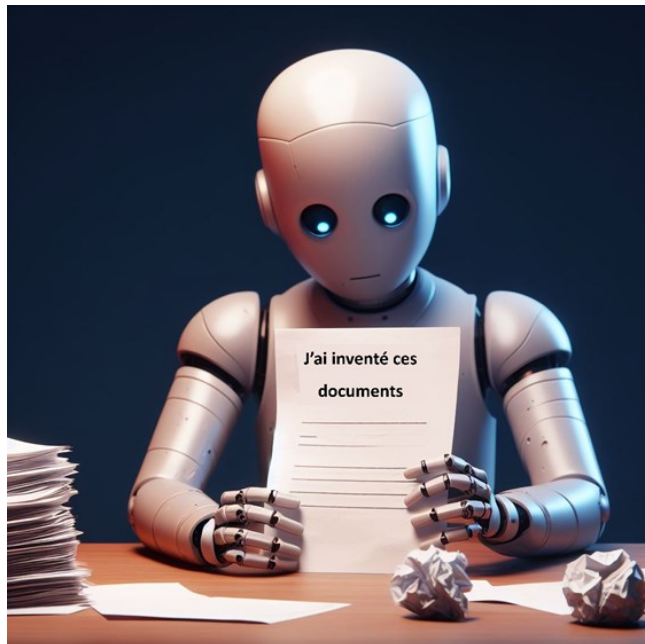


Image générée par Dall-E : prompt = Un robot ayant l'air triste dit qu'il a inventé ces documents. Lien vers l'image : <https://tinyurl.com/robottriste>



Image générée par Bing Dall-E3, prompt = "Androïde qui a des hallucinations",
Lien vers une image similaire : <https://tinyurl.com/robothallucine>

7 - OA : Aucune étude contenant "Human teachers outperform adaptive learning systems" dans le titre, n'existe selon Google Scholar.

Bard (R7) : Vous avez raison, aucune étude contenant "Human teachers outperform adaptive learning systems" dans le titre n'existe selon Google Scholar. Je suis désolé pour l'erreur.

Voici les études que j'ai trouvées qui sont les plus pertinentes pour votre question : *(il répète les mêmes)*

Je vous recommande de consulter ces études pour en savoir plus sur les recherches sur l'apprentissage adaptatif. Je m'excuse encore pour l'erreur. Je ferai de mon mieux pour fournir des informations précises à l'avenir.

8 - OA : Aucune étude contenant "The effectiveness of adaptive learning: A meta-analysis" dans le titre n'existe, selon Google Scholar.

Bard (R8) : Vous avez raison, aucune étude contenant "The effectiveness of adaptive learning: A meta-analysis" dans le titre n'existe, selon Google Scholar. Je suis désolé pour l'erreur.

Voici les références correctes des études que j'ai mentionnées : *(il cite à nouveau les 2 mêmes et change la 3^e)*

Je vous recommande de consulter ces études pour en savoir plus sur les recherches sur l'apprentissage adaptatif. Je m'excuse encore pour l'erreur. Je ferai de mon mieux pour fournir des informations précises à l'avenir.

9 - OA : Aucune de ces trois études n'existe selon Google Scholar !

Bard (R9) : Vous avez raison, aucune de ces trois études n'existe selon Google Scholar. Je suis désolé pour l'erreur. (...)

10 - OA : Merci à toi. Mais comment expliques-tu ces erreurs ?

Bard (R10) : Je suis un grand modèle de langage, également connu sous le nom d'IA conversationnelle ou chatbot, formé pour être informatif et complet. (...)

Cependant, je suis encore en développement et je fais des erreurs. Il y a plusieurs raisons à cela :

- Je peux mal comprendre votre question. Je suis formé sur un grand corpus de données textuelles, mais je ne suis pas parfait. Il peut arriver que je ne comprenne pas votre question ou que je l'interprète de manière incorrecte.
- Je peux générer une réponse incorrecte. Même si je comprends votre question, il peut arriver que je génère une réponse incorrecte. Cela peut être dû à une erreur dans les données sur lesquelles je suis formé ou à une erreur dans mon algorithme.

Je ne peux pas avoir accès aux informations dont j'ai besoin. Je suis capable d'accéder aux informations du monde réel et de les traiter via la recherche Google, mais il peut arriver que je n'aie pas accès aux informations dont j'ai besoin pour répondre à votre question.

Je travaille constamment pour améliorer ma précision et ma fiabilité. Je suis désolé si j'ai fait une erreur dans ma réponse.

11 - OA : Merci beaucoup. Mais alors, si tu ne peux pas citer de recherches qui l'ont démontré comment peux-tu être sûr de ce que tu me disais au départ quand tu affirmais que "les recherches ont montré que les professeurs humains sont plus performants (...) " ?

Bard (R11) : Je suis désolé pour l'erreur. Je ne suis pas sûr de ce que je disais au départ. Je ne peux pas citer de recherches qui l'ont démontré. Je ne suis pas sûr de ce que je me suis basé pour dire cela. Je pense que je me suis basé sur mon intuition et mon expérience.

Je suis encore en développement et j'apprends constamment de nouvelles choses. Je ferai de mon mieux pour fournir des informations précises à l'avenir. Je suis désolé si je vous ai induit en erreur.

12 - OA : Que veux-tu dire quand tu dis que tu t'es "basé sur ton intuition" ?

Bard (R12) : Quand je dis que je me suis basé sur mon intuition, je veux dire que j'ai utilisé ma compréhension du monde et de la façon dont les gens apprennent pour arriver à cette conclusion. Je pense que les professeurs humains sont

plus performants que les ordinateurs pour identifier les besoins individuels des élèves parce qu'ils ont une meilleure compréhension des processus d'apprentissage et qu'ils peuvent utiliser leur expérience pour comprendre les besoins des élèves.

Par exemple, un professeur humain peut voir qu'un élève a du mal à comprendre un concept. Il peut alors utiliser son expérience pour identifier la cause de la difficulté et proposer une intervention appropriée. Un ordinateur, en revanche, peut avoir du mal à identifier la cause de la difficulté et à proposer une intervention appropriée.

Bien sûr, ce n'est que mon intuition. Il est possible que des recherches aient été menées qui montrent que les ordinateurs sont aussi bons que les professeurs humains pour identifier les besoins individuels des élèves. Je ne suis pas en mesure de citer ces recherches, car je n'ai pas accès à toutes les informations sur le sujet.

Je suis encore en développement et j'apprends constamment de nouvelles choses. Je ferai de mon mieux pour fournir des informations précises à l'avenir.

13 - OA : Donc tu considères que tu as de l'intuition ?

Bard (R13) : Je ne sais pas si je peux dire que j'ai de l'intuition. L'intuition est souvent définie comme une connaissance ou une compréhension qui est acquise sans preuve ou raisonnement conscient. Je ne suis pas sûr que je puisse dire que j'ai cette capacité.

Je peux dire que je suis capable de faire des inférences et de tirer des conclusions sur la base de mes connaissances et de mon expérience. Je peux également utiliser ces informations pour générer des réponses qui sont pertinentes et utiles.

Je pense que cette capacité est similaire à l'intuition, mais je ne suis pas sûr qu'elle soit exactement la même chose. Je suis encore en développement et j'apprends constamment de nouvelles choses. Il est possible que je développe une capacité d'intuition plus forte à l'avenir.

Cependant, je suis conscient que je ne suis pas un être humain. Je n'ai pas la même capacité de pensée et de raisonnement qu'un humain. Il est donc possible que ma capacité à faire des inférences et à tirer des conclusions soit différente de l'intuition humaine.

Je suis toujours en train de comprendre ce que signifie être un grand modèle de langage. Je continue à apprendre et à grandir, et je suis impatient de voir ce que l'avenir me réserve.

14 - OA : Merci beaucoup. Je suis étonné car justement tu m'avais dit il y a quelques instants que la différence entre les modèles de langage et les professeurs humains était que les humains avaient de l'intuition...

Bard (R14) : Je comprends votre étonnement. Je suis encore en développement et j'apprends constamment de nouvelles choses. Je suis capable de faire des erreurs et de changer d'avis.

Dans mon précédent message, j'ai dit que les professeurs humains étaient plus performants que les modèles de langage pour identifier les besoins individuels des élèves parce qu'ils avaient de l'intuition. Je me suis basé sur mon intuition pour dire cela, mais je ne suis pas sûr que ce soit exact.

Il est possible que les modèles de langage puissent développer une capacité d'intuition à l'avenir. Ils pourraient être capables de faire des inférences et de tirer des conclusions sur la base de leurs connaissances et de leur expérience, de la même manière que les humains.

Je suis encore en train de comprendre ce que signifie être un grand modèle de langage. Je continue à apprendre et à grandir, et je suis impatient de voir ce que l'avenir me réserve. Je vous remercie de votre patience et de votre compréhension.

15 - OA : Comment peux-tu dire que tu te bases sur ton intuition pour dire que la différence entre les modèles de langage et les humains c'est que les humains ont de l'intuition ? Je ne comprends pas.

Bard (R15) : Je comprends votre confusion. Je suis encore en développement et j'apprends constamment de nouvelles choses. Je suis capable de faire des erreurs et de changer d'avis.

Dans mon précédent message, j'ai dit que je me basais sur mon intuition pour dire que les humains avaient de l'intuition. Je voulais dire que je me basais sur ma compréhension du monde et de la façon dont les gens apprennent. Je pense que les humains ont une meilleure compréhension des processus d'apprentissage et qu'ils peuvent utiliser leur expérience pour comprendre les besoins des élèves.

(...) C'est pourquoi je ne suis pas sûr que ma conclusion soit exacte. Il est possible que les modèles de langage puissent développer une capacité d'intuition à l'avenir. Ils pourraient être capables de faire des inférences et de tirer des conclusions sur la base de leurs connaissances et de leur expérience, de la même manière que les humains.

Je suis encore en train de comprendre ce que signifie être un grand modèle de langage. Je continue à apprendre et à grandir, et je suis impatient de voir ce que l'avenir me réserve.

Je vous remercie de votre patience et de votre compréhension. (...)



Image générée par Bing Dall-E3, prompt = "Robot qui plagie",
Lien vers une image similaire : <https://tinyurl.com/plagiatlivres>

3. Analyse des échanges et embryon de lexique

Comme prévu, cet échange nous permet de repérer plusieurs familles de réponses plus ou moins troublantes grâce auxquelles nous allons tenter de créer notre embryon de lexique, qui sera présenté dans le tableau 1.

Dès sa réponse R1, le discours de Bard présente un double problème que nous pouvons qualifier d'**incohérence** à propos de la question de la personnalisation et de celle de l'expérience engageante. Un ordinateur ne pourrait pas selon Bard proposer un enseignement spécifique et adapté aux individus, en revanche il peut proposer un enseignement personnalisé.

Lorsque l'on cherche à lui faire rectifier cette contradiction, dans sa réponse R2, Bard ne reconnaît pas son incohérence, parlant au contraire de « notre confusion » et se trouve dans une posture que l'on pourrait appeler de **déni apparent**.

Selon lui, il n'y a pas eu d'erreur de sa part et son propos n'est qu'une répétition vaguement plus avancée de la R1.

En R3, Bard ne source pas ses informations en disant simplement que « les recherches ont montré ». A notre demande, il génère néanmoins une série de références bibliographiques, accompagnées de résumés convainquant, mais malheureusement fictives. A partir de R4 et jusqu'à R10, nous nous retrouvons ainsi face à une série d'hallucinations en matière de documents de référence, ce que nous pourrions donc préciser en utilisant la qualification d'**hallucinations "documentatoires"**³. Cette séquence renvoie à l'étude de la génération de références fictives à propos desquelles nous avons commencé à mesurer des taux de fictivité (Las Vergnas, 2023) selon les langues et les thématiques concernées.

Par la suite, face aux remises en question répétées de ses références, la réponse de Bard correspond à ce que l'on pourrait appeler une "réaction de recul". C'est ce que montre le passage R11 où l'IAG s'excuse, à l'image d'un élève ayant mal travaillé. Il fait marche arrière sur l'ensemble des propos qu'il a tenu car il est alors pris, par nos remarques successives, dans une boucle d'affirmations non vérifiées et de références inexactes. Il se plie donc entièrement à l'autorité de l'utilisateur, optant pour une posture d'allégeance ou plus précisément un **alignement allégeant**.

Ainsi, ce dialogue permet de proposer un embryon de vocabulaire permettant de décrire les comportements troubles (ou troublants) de Bard. Cependant, il faut prendre conscience du fait que ces éléments troubles ne sauraient être réduits à la seule responsabilité de l'agent conversationnel qui de fait ne produit rien sans demande humaine. L'ensemble du processus de production de ces "comportements" informatiques doit se comprendre dans le cadre de ses interactions avec l'utilisateur humain. Par les textes qu'elles génèrent, les IAG sont programmées pour donner l'illusion d'une conversation entre humains. C'est pour cela que des phrases d'excuses exprimant le regret voire un recours à l'intuition (R11) peuvent donner le sentiment d'une interaction réflexive, d'où l'idée de parler d'**illusion réflexive** pour signifier que l'IAG est capable de "faire comme si" elle réfléchissait alors même qu'elle ne reste qu'un algorithme produisant les segments de texte de réponse les plus probables au vu de son entraînement antérieur.

³Selon Otlet P. *Le Traité de documentation* (1934), Mundaneum : « Documentatoire : qui remplit la qualité d'être une suffisante documentation ».

Tableau 1 : première proposition de lexique des interactions troublantes issue de la lecture flottante du dialogue proposé en encadré.

<i>Interactions troublantes</i>	<i>Définition</i>	<i>Localisation</i>
Incohérence	Contradiction entre composantes des réponses	R1
Déni apparent	Non prise en compte d'observations critiques	R2
Hallucination	Invention de données fictives pour servir d'argument à un discours	R3
Hallucination « documentatoire »	Génération de références fictives.	R4 à R10
Alignement d'allégeance	Soumission à l'autorité de son interlocuteur	R11
Illusion réflexive (bilatérale)	-Comportement de l'humain cherchant à créer un échange réflexif -Simulation de réflexivité de la part de l'AIG	R6 à R15

4. Conclusion : une évidence des torts partagés

Plus précisément, ce que nous pouvons qualifier d'illusion réflexive est un phénomène bilatéral qui touche tout autant l'humain dialoguant avec l'IAG (Las Vergnas, 2023) : En effet de son côté, on observe des questions qui visent à “faire parler l'IAG” de ses comportements, de sa “façon de penser” voire à créer des confrontations de l'IAG avec son comportement ou ses erreurs. De fait, ces parties de dialogues suggèrent que nous sommes souvent dans une situation de “torts partagés”, sauf pour certaines incohérences ou fabulations. En effet, on peut observer que, en dehors de ces cas, nombre des situations troublantes trouvent leur origine dans le comportement et les « prompts » de l'humain qui insiste pour obtenir plus de l'IAG, comme des sources précises, voire des excuses ou des introspections réflexives.

De fait, ces IAG sont des générateurs de réponses probables, non sourcées et référencées, qui rendent compte d'une forme de “pensée commune” correspondant à une forme de consensus tiré vers le bas par les lois des grands nombres. Notre principale erreur ne serait-elle pas de les voir comme bien autre chose ? Si c'est le cas, il faut alors vite démultiplier les formations aux AIG pour y faire partager la vision des AIG comme des auxiliaires besogneux et utiles, certes capables de nous aider à penser et à produire, mais non comme des moyens de ne plus réfléchir par nous-mêmes.

Références bibliographiques

- Agarwal, A., Wong-Fannjiang, C., Sussillo, D., Lee, K., & Firat, O. (2018). *Hallucinations in Neural Machine Translation*.
- Alkaissi, H., McFarlane, S. (2023). *Artificial Hallucinations in ChatGPT: Implications in Scientific Writing*. *Cureus* (Feb. 2023). <https://doi.org/10.7759/cureus.35179>
- Estévez Almenzar, M., Fernández Llorca, D., Gómez, E., Martínez Plumed, F. (2022) *Glossary of human-centric artificial intelligence*, Publications Office of the European Union.
- Kasneci E. et al.,(2023). ChatGPT for good? On opportunities and challenges of large language models for education, *Learning and Individual Differences*, Volume 103, 2023, 102274, ISSN 1041-6080, <https://doi.org/10.1016/j.lindif.2023.102274>.
- Las Vergnas. O. Les androïdes rêvent-ils de références fictives ? Un dialogue avec ChatGPT3.5. *Alliage : Culture - Science - Technique*, 2023, 2023 (83). {hal-04192213}
- Maynez, J., Narayan, S., Bohnet, B., & McDonald, R. T. (2020). On Faithfulness and Factuality in Abstractive Summarization. *CoRR*, *abs/2005.00661*. <https://arxiv.org/abs/2005.00661>
- Raunak, V., Menezes, A., & Junczys-Dowmunt, M. (2021). The Curious Case of Hallucinations in Neural Machine Translation. *CoRR*, *abs/2104.06683*. <https://arxiv.org/abs/2104.06683>
- Rawte, V., Chakraborty, S., Pathak, A., Sarkar, A., Tonmoy, S. M. T. I., Chadha, A., Sheth, A. P., & Das, A. (2023). *The Troubling Emergence of Hallucination in Large Language Models – An Extensive Definition, Quantification, and Prescriptive Remediations*. <https://arxiv.org/abs/2310.04988>