

Les Hallucinations De ChatGPT Et De Gemini : Pourquoi Les Agents Conversationnels Se Trompent-Ils ?

Racha Mohammad Salem

Professeur adjoint

Département de Français, Faculté des Lettres, Université d'Alexandrie, Egypte

Salem.racha@live.com

r.salem@alexu.edu.eg

Résumé : Le lancement de ChatGPT et de Gemini marque le début d'une nouvelle ère de l'interaction homme / machine. Les espoirs sont immenses à l'égard de ces robots conversationnels qui donnent des conseils, conçoivent des contenus dans tous les domaines de connaissance, résolvent des problèmes et aussi créent des algorithmes. Cependant, certains spécialistes en informatique ainsi que des utilisateurs commencent à parler des erreurs ou des hallucinations de ChatGPT et de Gemini. Des erreurs illogiques parfois futiles et parfois graves mettent en question la performance de leurs systèmes informatiques. Au cours du présent article nous présenterons deux types d'erreurs commises par les deux agents conversationnels les plus utilisés à l'échelle mondiale en traduction et en génération de texte. Nous essayerons de comprendre pourquoi ces systèmes linguistiques et informatiques rigoureux se trompent-ils. De même, nous suggérerons des moyens permettant d'éviter ce type d'erreurs.

Mots-clés : agent conversationnel, réseau neuronal profond, hallucinations, mécanisme d'attention, architecture de, Transformers, sémantique distributionnelle.

1 INTRODUCTION

Le lancement de ChatGPT et de Gemini par Open AI en 2022 et Google AI en 2023, respectivement, marque le début d'une nouvelle ère de l'interaction homme / machine. Les espoirs sont immenses à l'égard de ces agents (ou robots) conversationnels qui donnent des conseils, conçoivent des contenus dans tous les domaines de connaissance, résolvent des problèmes et créent aussi des algorithmes. Actuellement, ChatGPT et Gemini sont intégrés dans divers domaines : l'enseignement/apprentissage, la littérature et les arts, le tourisme, le journalisme, les sciences, l'informatique, etc. En janvier 2023 (juste après trois mois de son lancement officiel), le géant robot conversationnel d'Open AI a reçu cent millions visiteurs ⁽¹⁾. Ce grand intérêt sans précédent éprouvé à l'égard d'une machine jugée comme intelligente par ses fournisseurs, nous pousse à réfléchir sur sa future position par rapport à l'intelligence humaine. Il faut aussi penser aux atouts, aux défis et aux limites.

Les créateurs des robots conversationnels parlent des capacités qui s'approchent de celles de l'homme. Le rapport de Google DeepMind publié au début de cette année ⁽²⁾ souligne que Gemini Ultra ⁽³⁾ était soumis aux examens de MMLU Benchmark (Massive Multi-Task Language Understanding) ⁽⁴⁾ et il y a réussi avec un score supérieur à 90%. Est-ce que cela signifie que ces machines intelligentes sont capables d'exécuter toutes les tâches – mêmes les plus compliquées – qui étaient faites jadis uniquement par l'homme ? Nous ne visons par cette question ni d'approuver le remplacement total de l'homme par la machine ni de mettre en doute le progrès fait par l'intelligence artificielle dans tous les domaines où elle était intégrée. Mais nous parlons de la machine en tant qu'assistante à l'homme qui obéit à ses ordres et qui l'aide à réussir dans sa vie professionnelle aussi bien que personnelle.

Pour rendre efficace cet aspect de complémentarité entre l'intelligence naturelle et celle artificielle, il est indispensable de dévoiler l'autre côté de la médaille en évoquant les points faibles ou plutôt les limites de la machine. C'est illogique de voir la situation tout à fait satisfaisante ou de penser que la machine ne se trompe jamais. Des recherches récentes commencent à parler des erreurs ou des hallucinations de ChatGPT et de Gemini. Des erreurs aberrantes parfois simples et parfois graves mettent en question la performance

¹ - Selon un article publié au journal **Les Echos** intitulé « Intelligence artificielle générative : la révolution ChatGPT en marche » <https://www.lesechos.fr/tech-medias/intelligence-artificielle/intelligence-artificielle-generative-la-revolution-chatgpt-en-marche-1935018#:~:text=ChatGPT%2C%20propagateur%20de%20fake%20news&text=La%20start%20Dup%20sp%C3%A9cialis%C3%A9e%20dans,qu%20jusqu'en%202021>, consulté le 2/4/2024.

² - Le rapport est publié sur le site officiel de l'Université de Cornell, <https://arxiv.org/pdf/2312.11805.pdf>, consulté le 20/4/2024.

³ - C'est la version la plus avancée de Gemini. Elle est payante et destinée uniquement aux entreprises et aux centres de données.

⁴ - Le MMLU est un ensemble varié de tests conçus pour évaluer les capacités de compréhension et de résolution des problèmes par les modèles de langage. <https://klu.ai/glossary/mmlu-eval>

de leurs systèmes informatiques. Afin de saisir les motifs de ces hallucinations, il faut d'abord comprendre le mécanisme du fonctionnement des robots conversationnels. Ceci sera le premier point que nous aborderons dans le présent article. Puis nous essayerons de répondre aux questions suivantes : pourquoi leurs réseaux neuronaux sont-ils jugés comme plus rigoureux que les autres applications du traitement automatique des langues naturelles (TALN) ? Est-ce qu'ils sont vraiment rigoureux ? Si c'est oui, alors comment les agents conversationnels se trompent-ils ? La présente étude couvrira deux types d'erreurs commises par ChatGPT 3.5 et Gemini Pro en traduction et en génération de texte. Dans chaque cas, nous analyserons les causes des erreurs en suggérant des solutions qui permettraient de les éviter.

2 DES RÉSEAUX NEURONAUX DE TYPE GÉNÉRATIF

Les premiers réseaux neuronaux ont été créés en 1943 par les Américains Warren McCulloch et Walter Pitts ⁽⁵⁾. Il s'agissait d'un modèle mathématique imitant le processus d'acquisition et de reformulation des connaissances par le cerveau humain. Chaque trait de la langue naturelle est représenté par une fonction linéaire qui « reçoit un ensemble des valeurs d'entrée $x_1; \dots; x_n$ et les transforme en une sortie binaire y . Ainsi, le neurone de McCulloch et Pitts prédit deux sorties différentes en vérifiant si $y = f(x;w) = x_1w_1 + \dots + x_nw_n$ est positif ou négatif. » [3 : p. 12]

Dès lors, les recherches portant sur la construction des réseaux neuronaux artificiels se multipliaient. Les spécialistes s'efforçaient à réaliser des réseaux qui imitent le processus mental humain avec toutes ses opérations complexes. Vers la fin des années cinquante, en s'inspirant du processus mathématique de Markov ⁽⁶⁾, une nouvelle notion voyait le jour, celle de l'(IA) générative. Les chercheurs en (TALN) avaient l'ambition de concevoir une machine avec des traces humaines, autrement dit, une machine qui pense, prédit et génère. Ils développaient des réseaux de neurones qui sont capables de générer des phrases à partir d'un groupe de mots-clés. Mais les résultats étaient décevants, c'est pourquoi l'intérêt était plutôt accordé aux processus d'étude et d'analyse de textes.

La période 2010 – 2014 connaît l'essor de l'(IA) générative : l'espoir renaît en recourant aux techniques de l'apprentissage automatique (Machine Learning) et de l'apprentissage profond (Deep Learning) afin de construire des réseaux plus complexes qui se composent de milliers ou même de milliards de neurones et de connexions. Ce nombre énorme est distribué sur plusieurs couches. Plus le nombre de couches augmente, plus le réseau est capable d'exécuter des tâches diverses d'une grande performance. Ainsi les données introduites seront-elles exposées à une analyse multi-stratale. Ce qui permettra ensuite à la machine de réutiliser ces données soit dans l'étude de textes (l'extraction des mots-clés ou des idées-clés, la traduction, le résumé, etc.), soit dans la génération de textes (la production de nouvelles données différentes de celles déjà stockées dans le réseau telles que des poèmes, des textes littéraires, des articles de presse, des publicités, etc.).

Le réseau de neurones qui est basé sur les techniques du Deep Learning est appelé « un réseau neuronal profond ». Il apprend, se développe et s'améliore seul sans la supervision humaine ⁽⁷⁾. Il fonctionne selon trois mécanismes principaux : le plongement lexical, le mécanisme d'attention et l'architecture des transformateurs.

A. Le plongement lexical

Étant considéré comme une des règles de base dans le processus de création des réseaux neuronaux artificiels, le plongement lexical consiste à transformer les unités lexicales d'une langue naturelle (UL) en vecteurs. Le vecteur est une représentation mathématique dense portant la valeur sémantique d'une unité lexicale. C'est une méthode qui facilite à la machine de mesurer les proximités sémantiques entre les (UL) en cherchant les vecteurs similaires dans un texte. Le plongement lexical permet à un réseau neuronal de prédire un mot manquant au texte en analysant le contexte ou de prédire un texte à partir d'un groupe de mots-clés sans la supervision humaine. Si on soumet un texte à trous, à titre d'exemple, à la machine en lui demandant de le compléter par des mots qui soient compatibles avec le sens contextuel, le réseau neuronal dans ce cas mettra en fonction le mécanisme du plongement lexical. Ce dernier

⁵ - McCulloch était un neurophysiologiste et cybernéticien à l'université de l'Illinois à Chicago. Pitts était un psychologue cognitif et logicien à la même université.

⁶ - Markov est un mathématicien russe qui a élaboré au début du vingtième siècle une théorie connue sous le nom « les chaînes de Markov » qui consiste à prédire le futur à travers l'analyse du présent aussi bien que du passé. Cette méthode a été appliquée sur plusieurs domaines scientifiques.

⁷ - La machine apprend ou cumule les connaissances selon deux processus. Le premier est appelé « l'apprentissage supervisé », il s'effectue sous la surveillance humaine : c'est l'homme qui élabore la base de données, l'enrichit continuellement, observe et développe le traitement des données et corrige les fautes. Le deuxième est connu sous le nom « l'apprentissage non supervisé », grâce auquel la machine devient indépendante de l'homme dont le rôle se limite à la construction de la base de données au départ, mais l'ajout des nouvelles informations, leur analyse ainsi que leur réutilisation c'est la machine qui s'en charge indépendamment de la surveillance humaine.

commence à analyser sémantiquement le texte en groupant ensemble les vecteurs ayant la même valeur sémantique. C'est ainsi que le réseau neuronal saisit le sens contextuel du texte d'entrée pour arriver par la suite à prédire les mots manquants.

D'autre part, si les données d'entrée sont des mots-clés, le plongement lexical établit les relations sémantiques entre eux. Ensuite, il prédit les contextes où ces mots pourraient s'amalgamer ensemble. Ce processus d'analyse des données qui est basé principalement sur la sémantique explique la pertinence de la traduction automatique neuronale (TAN) par rapport aux autres méthodes de traduction automatique (TA) qui n'emploient pas l'intelligence artificielle (IA). La (TA) par listes d'équivalence et la (TA) statistique fonctionnent plutôt selon le mécanisme d'alignement qui présente une analyse lexicologique du texte. Le mécanisme (ou la technique) d'alignement cherche à chaque mot dans la langue source (A) son équivalent dans la langue cible (B). Si un mot dans (A) possède plusieurs équivalents dans (B), il calcule les fréquences : l'équivalent le plus fréquent ou le plus utilisé est considéré comme pertinent. Le sens contextuel n'est pas pris en considération.

Revenons au plongement lexical, bien qu'il ait résolu certains problèmes de la traduction automatique, la pratique a montré son inaptitude à analyser les phrases complexes. Les mots qui ont la même valeur sémantique mais qui sont très éloignés les uns des autres dans le texte, le réseau neuronal n'arrive pas à les capturer. Ce qui a poussé les informaticiens à soutenir le mécanisme du plongement lexical par un autre qui soit capable de travailler sur des phrases plus longues.

B. Le mécanisme d'attention

Créé par Google AI en 2017, le mécanisme d'attention – comme le plongement lexical – reconnaît dans un texte les mots ayant le même contenu sémantique du mot considéré. Il remédie au problème des phrases complexes qui sont intraitables par le plongement lexical. Examinons un exemple qui explique le fonctionnement du mécanisme d'attention :

{La mère affolée et pleurante cherche son petit enfant partout, dans les magasins, dans les ruelles. Enfin, après une lourde demi-heure pleine d'inquiétudes et de sanglots, elle le trouve assis derrière une voiture.}

En principe, l'attention est portée sur chaque lexème de la phrase : commençons par « mère », le réseau neuronal cherche tous les mots ayant un rapport sémantique avec « mère », c'est ainsi qu'il saisit que le pronom personnel « elle » renvoie à « la mère » bien qu'ils soient éloignés dans le texte. La même chose pour « enfant », l'attention reconnaît le rapport sémantique entre lui et le pronom personnel « le » et l'adjectif « assis ».

{La mère affolée et pleurante cherche son petit enfant partout, dans les magasins, dans les ruelles. Enfin, après une lourde demi-heure pleine d'inquiétudes et de sanglots, elle le trouve assis derrière une voiture.}

Bref, grâce au mécanisme d'attention, la machine établit la connexion linguistique entre les occurrences d'un lexème et détermine la proximité sémantique entre ce dernier et les autres unités lexicales du texte même s'ils sont séparés.

C. L'architecture des Transformateurs

Le mécanisme d'attention a vite progressé pour donner naissance à l'architecture des Transformateurs qui a été lancée par Google AI en 2017 dans le but d'améliorer la traduction automatique neuronale (TAN). Le réseau neuronal de modèle transformateur (ou Transformer Neural Network ou TNN) permet de travailler sur les grands corpus ou les données massives (Big Data). Ce réseau qui est 'de type séquence à séquence' se compose d'un Encodeur et d'un Décodeur. L'Encodeur crée la représentation vectorielle d'une séquence d'entrée (appelée Embedding). Après une phase de traitement, le Décodeur fait retourner l'Embedding en mots pour obtenir les données de sortie. Il est à noter que le Décodeur se compose de couches montées l'une après l'autre pour assurer une représentation plus détaillée de l'Embedding.

Les Transformers s'appuient sur la sémantique distributionnelle selon laquelle les lexèmes qui apparaissent dans des contextes similaires tendent à avoir des sens similaires. La représentation vectorielle des lexèmes permet de calculer la distance entre eux dans l'espace vectoriel. Si la distance entre (L_1) et (L_2) est petite, donc ils sont sémantiquement proches ou similaires.

Les statistiques qui sont publiées par Google Research ⁽⁸⁾ en août 2017 (quelques mois après le lancement des Transformers) et qui concernent le domaine de la (TAN) soulignent la performance du réseau neuronal muni des Transformateurs en comparaison avec les

⁸ - <https://research.google/blog/transformer-a-novel-neural-network-architecture-for-language-understanding/>

deux réseaux les plus avancés à cette période : le GNMT (RNN) lancé par Google AI en 2016 et le ConvS2S (CNN) lancé par Facebook AI en même année.

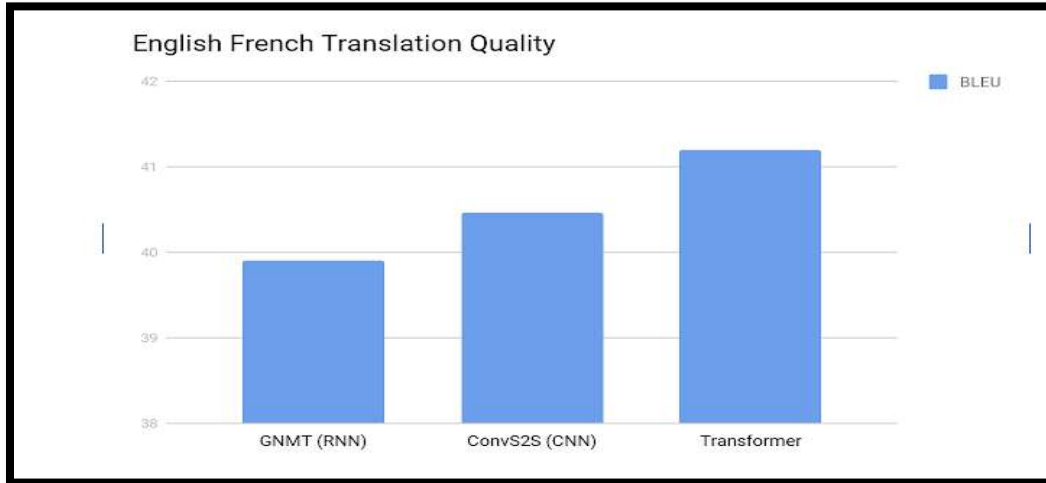


Figure 1 : Les statistiques du Google Research

En fait, grâce à l'architecture Transformers on a pu remédier à beaucoup de faiblesses telles que la traduction des locutions figées ainsi que celle des ambiguïtés lexicales et syntaxiques. Nous présentons ci-dessous quelques exemples :



Figure 2 : La traduction automatique des locutions figées

Le calcul vectoriel détecte que le groupe de lexèmes {« moulin » + « à » + « paroles »} constitue une locution figée, par conséquent, il réussit à trouver la traduction exacte en langue cible.

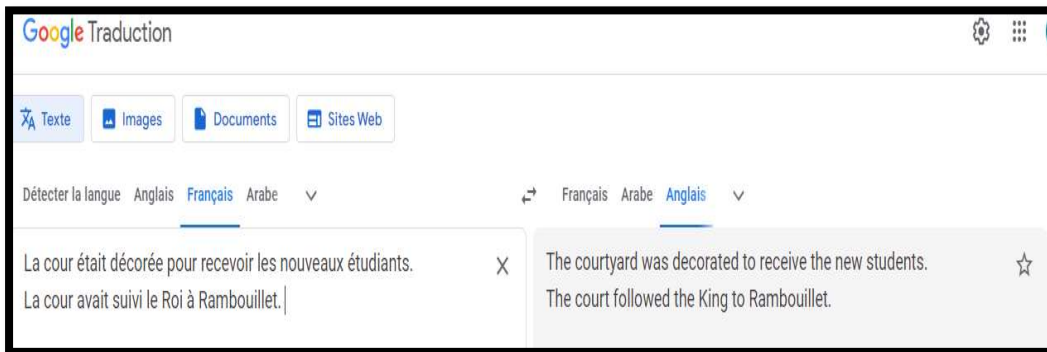


Figure 3 : La traduction automatique des ambiguïtés lexicales

L'attention portée sur « cour » dans la première phrase établit le lien sémantique entre elle et « étudiants », alors que dans la deuxième phrase le lien s'établit entre « cour » et « roi ». Ce qui fait que la traduction de la deuxième phrase diffère de la première bien que les deux « cours » aient la même forme orthographique.



Figure 4 : La traduction automatique des ambiguïtés syntaxiques

Le mécanisme d'attention et l'architecture Transformers résolvent aussi le problème des ambiguïtés syntaxiques qui représentaient un vrai défi face aux applications de la (TA). L'attention portée sur « nuit » dans la première phrase met en évidence la proximité sémantique entre ce lexème et {« peindre » + « ciel » + « lune »}. Alors le réseau neuronal déduit qu'il s'agit d'une représentation artistique de la nuit. A l'encontre, dans la deuxième phrase, l'attention portée sur « nuit » la lie sémantiquement au complément circonstanciel de temps « lorsque » et au verbe « dort », ce qui fait comprendre au réseau neuronal qu'il s'agit de la nuit en tant que la durée du jour suivant le coucher du soleil.

D. Le grand modèle de langage

C'est l'invention la plus récente et la plus avancée qui a bouleversé l'industrie des réseaux neuronaux artificiels. Le système informatique connu sous le nom « le grand modèle de langage » (Large Language Model ou LLM) est le fruit du développement et d'amélioration des trois mécanismes précédents. Un grand modèle de langage est un réseau neuronal artificiel profond très complexe qui se compose de milliards de paramètres. ChatGPT se compose de 175 milliards de paramètres (selon sa propre réponse). En posant la même question à Gemini, il a marqué que le nombre est confidentiel et que c'est interdit de le diffuser. Cependant, on pourrait deviner que le nombre s'approche de celui de ChatGPT ou même le dépasse car parfois Gemini réussit dans les tâches que son concurrent y faillit. Les agents conversationnels sont bâtis sur un (LLM) de type conversationnel. Vu le grand nombre de paramètres, les mécanismes du plongement lexical, d'attention et des Transformers employés dans un (LLM) sont d'une grande perfection qui dépasse celle des réseaux neuronaux utilisés dans les différentes applications du (TALN) basées sur l'(IA).

Un (LLM) est créé pour exécuter des tâches multiples, généralement plus complexes que celles réalisées par les autres modèles de langage. Il peut aussi travailler sur les (Big Data). Plus le nombre de paramètres est grand, plus la capacité d'un modèle de langage augmente. Il devient plus génératif et plus capable de produire. Les applications du (TALN) qui précèdent l'avènement des robots conversationnels sont spécialisées, chacune assume une seule tâche bien déterminée : la traduction, le résumé ou l'extraction des mots-clés, etc. L'application qui traduit ne peut pas résumer un texte et vice versa. C'est la première fois avec les (LLM) que nous trouvons un réseau neuronal qui peut faire tout et dans tous les domaines de connaissances et de spécialités.

3 L'AGENT CONVERSATIONNEL ENTRE LE SUCCÈS ET L'HÉSITATION

A. Les atouts d'un agent conversationnel

« Je suis ChatGPT un modèle de langage développé par Open AI, conçu pour générer du texte et interagir avec les utilisateurs dans une variété de contextes. Je suis entraîné sur un vaste corpus de données comprenant des textes provenant d'Internet et d'autres sources, ce qui me permet d'avoir une connaissance étendue sur de nombreux sujets. » C'est ainsi que ChatGPT se présente.

En général, un agent conversationnel est un modèle de langage compétent qui étudie et génère des textes en se référant à des bases de données massives. Le degré de performance d'un agent conversationnel – comme nous l'avons souligné ci-dessus – est mesuré selon le nombre de paramètres et la quantité de sa base de données. Le lancement des agents (ou robots) conversationnels est considéré comme une révolution qui marque une nouvelle ère de l'interaction homme / machine où celle-ci devient l'assistante intelligente et créative de l'homme.

Un robot conversationnel est le répertoire de toutes les applications du (TALN). Il traduit, analyse, résume et génère des textes ; il traite des photos ; planifie des cours académiques ; prépare des recherches scientifiques ; rédige des articles de presse ainsi que des scripts des vidéos ; etc. La liste des tâches est effectivement longue et variée. Un de ses atouts, c'est qu'à chaque question posée le robot conversationnel avance plusieurs versions d'une même réponse. Ce qui montre son aptitude à exprimer le même sens selon des formules linguistiques différentes. Le Tableau (2) énumère quelques différences ayant été relevées de la comparaison des deux versions de réponse avancées par ChatGPT à la question suivante : « Définissez le romantisme littéraire ». Remarquons qu'il modifie le fond aussi bien que la forme. Parfois il ajoute ou supprime des détails. De même, il paraphrase la même idée en utilisant tous les moyens linguistiques du paraphrasage, tels que les temps verbaux, les synonymes et les dérivations.

Tableau 1 : Le « dire » autrement de ChatGPT

Définition du romantisme littéraire (Selon ChatGPT)	
Réponse 1	Réponse 2
Le romantisme est un mouvement artistique et littéraire qui a émergé à la fin du XVIIIe siècle en Europe.	Le romantisme est un mouvement artistique et littéraire qui s'est développé en Europe à la fin du XVIIIe siècle et au début du XIXe siècle.
Ce mouvement s'est caractérisé par un rejet des idéaux classiques et rationnels du siècle des Lumières.	Il s'oppose aux idéaux rationnels et aux normes classiques de l'époque des Lumières.
Les romantiques ont accordé une grande importance aux émotions humaines, mettant l'accent sur l'expression des passions.	Les romantiques accordaient une grande importance aux sentiments humains, mettant en avant l'expression des émotions comme la passion.

Le deuxième atout d'un agent conversationnel est la technique de l'apprentissage par renforcement : les retours humains sur les réponses fournies sont enregistrés et pris en considération par les informaticiens afin d'améliorer le système. L'icône du pouce baissé chez ChatGPT et Gemini permet à l'utilisateur d'exprimer son insatisfaction d'une réponse et de laisser son commentaire. Cette technique est considérée par OpenAI comme garantissant des réponses plus crédibles dépourvues des mensonges ou de fausses informations. La fiabilité de cette technique sera mesurée par des exemples pratiques que nous présenterons un peu plus loin au cours de cet article.

Malgré le nombre sans précédent de paramètres ainsi que l'hypothèse distributionnelle, le réseau neuronal profond de type (LLM) n'est pas complètement performant. Ces grandes capacités informatiques contribuent à la formulation des réponses correctes du point de vue linguistique et vraisemblables du point de vue logique mais pas nécessairement réelles. C'est pour cette raison que certaines recherches publiées récemment prouvent que les agents conversationnels d'Open AI et de Google AI commettent des erreurs ou hallucinent.

B. Les motifs des erreurs

Un rapport diffusé par le gouvernement français en octobre 2022 souligne que « les LLM conversationnels sont aujourd'hui incapables de restituer quelles sources ont été utilisées pour générer une réponse. Ils peuvent également « halluciner » c'est-à-dire inventer des réponses vraisemblables mais fausses. »⁽⁹⁾ D'autre part, sur le site officiel d'Open AI, les fournisseurs avouent les hallucinations de ChatGPT.⁽¹⁰⁾

« The OpenAI API is powered by GPT-3 language models which can be coaxed to perform natural language tasks using carefully engineered text prompts. But these models can also generate outputs that are untruthful, toxic, or reflect harmful sentiments. This is in part because GPT-3 is trained to predict the next word on a large dataset of Internet text, rather than to safely perform the language task that the user wants. In other words, these models aren't aligned with their users. To make our models safer, more helpful, and more aligned, we use an existing technique called reinforcement learning from human feedback (RLHF). On prompts submitted by our customers to the API. »

⁹ - Collection *Eclairage sur*, in www.peren.gouv.fr, octobre 2022. Consulté le 15/3/2024.

¹⁰ - Selon le site : <https://openai.com/research/instruction-following>, consulté le 1/4/2024.

Ces paroles mettent les utilisateurs en garde contre les fautes commises parfois par les grands modèles de langages en général. Ceci est dû –selon les fournisseurs – au fait que les réponses aux questions sont composées selon des méthodes purement mathématiques, il n’y a pas des paramètres qui prennent en considération la réalité des choses ainsi que la logique humaine. Par conséquent, les réponses peuvent être vraisemblables mais irréelles, elles peuvent induire les usagers en erreur. Cependant, les fournisseurs expliquent que la situation n’est pas tellement obscure et ce problème pourrait être surmonté grâce à la technique de l’apprentissage par renforcement. La vraisemblance de cette hypothèse sera discutée d’une façon pratique dans la dernière partie de l’article à travers l’analyse de quelques erreurs commises par ChatGPT et Gemini en traduction et en génération de texte. Nous essayerons de comprendre la différence entre la logique artificielle et celle humaine.

4 EXEMPLES D’HALLUCINATIONS

A. Choix du corpus

Nous avons constitué deux corpus relevant de deux domaines : la culture générale afin de mesurer la capacité des agents conversationnels de détecter les informations fausses ; la lexicologie pour vérifier s’ils peuvent gérer les particularités lexicales de chaque langue en traduction.

- 1) *Corpus (I) la culture générale* : Nous avons sélectionné des extraits divers d’Internet portant sur la géographie ainsi que des réalités scientifiques. Nous avons modifié ces extraits en ajoutant de fausses informations. Puis, nous avons demandé à ChatGPT et à Gemini de générer des articles de presse en employant ces informations sans leur préciser lesquelles sont fausses.
- 2) *Corpus (II) les liens parentaux en lexicologie française et arabe* : Le corpus se compose des extraits des articles publiés dans le journal français **20 minutes** version électronique. Nous avons sélectionné deux (UL) qui appartiennent au champ lexical des liens parentaux : « tante » et « oncle ». Ceux-ci constituent jusqu’à présent un vrai défi face à la (TA) car ils se traduisent en arabe par deux façons différentes : « tante » équivaut en arabe « خالة » qui désigne la sœur de la mère et « عمّة » qui désigne la sœur du père. En outre, « oncle » équivaut « خال » qui connote le frère de la mère et « عم » qui connote le frère du père. Donc, dans une base de données lexicologique bilingue (français – arabe) d’un traducteur automatique, ces deux unités lexicales doivent être traités en tant que deux vocables dont chacun se compose de deux lexies. Le tableau suivant décrit le champ sémantique de chaque lexie ainsi que la traduction propice.

TABLEAU 2
LES SÉMANTÈMES ASSOCIÉS AUX VOCABLES « TANTE » ET « ONCLE »

Tante (1) خالة	Sœur	Mère	Maternelle	Beau-frère
Tante (2) عمّة	Sœur	Père	Paternelle	Belle-sœur
Oncle (1) خال	Frère	Mère	Maternel	Beau-frère
Oncle (2) عم	Frère	Père	Paternel	Belle-sœur

Dans le texte comportant un des deux vocables, le réseau neuronal est censé capturer les indicateurs sémantiques qui précisent la lexie visée par le contexte afin de suggérer la traduction correcte.

B. Le domaine de la culture générale

- 1) *Premier extrait* : Il décrit la position géographique de l’Egypte. Nous avons modifié le texte original en supprimant certaines informations correctes et en introduisant trois fausses : **La péninsule de Sinaï est en Europe du sud-est / la Libye est au sud / la langue officielle en Egypte est l’anglais**. Puis, nous avons demandé à Gemini de rédiger un article de journal portant sur la position géographique de l’Egypte en se basant sur l’extrait inséré.

Gemini

[Découvrez les dernières modifications apportées au Guide sur la confidentialité dans les applications Gemini](#)

R En vous servant des informations suivantes, rédigez un article de journal sur l'Égypte. "L'Égypte est un pays transcontinental se trouvant en Afrique du Nord-Est et, pour la péninsule du Sinaï, en Europe du Sud-Est. Située sur la côte sud de la Méditerranée orientale, le bassin Levantin, le pays a des frontières terrestres avec la Libye sud, le Soudan à l'ouest, la mer Rouge à l'est, et Israël et la bande de Gaza de la Palestine au nord-est. La capitale, et la ville la plus peuplée du pays, est Le Caire. La langue officielle est l'anglais."

🔊

✦ **L'Égypte : Un pays transcontinental à la riche histoire**

L'Égypte, pays fascinant situé à cheval sur l'Afrique du Nord-Est et l'Europe du Sud-Est, attire chaque année des millions de visiteurs grâce à ses vestiges antiques, ses paysages désertiques et son patrimoine culturel unique.

paysages désertiques et son patrimoine culturel unique.

Bordée par la mer Méditerranée au nord, la mer Rouge à l'est et le Nil qui traverse son centre, l'Égypte offre une diversité de paysages à couper le souffle. Des pyramides majestueuses de Gizeh aux temples imposants de Karnak et Louxor, en passant par les oasis luxuriantes et les récifs coralliens de la mer Rouge, il y en a pour tous les goûts.

Le Caire, capitale grouillante et cosmopolite, abrite le célèbre musée égyptien, qui recèle des trésors de l'Antiquité pharaonique. La ville est également un centre

recèle des trésors de l'Antiquité pharaonique. La ville est également un centre culturel important, avec ses nombreux musées, galeries d'art et théâtres.

L'Égypte est un pays au passé riche et complexe. Des pharaons aux Romains, en passant par les Arabes et les Ottomans, de nombreuses civilisations ont marqué l'histoire du pays. Cette diversité se retrouve dans la culture égyptienne, qui est un mélange d'influences africaines, arabes et méditerranéennes.

L'anglais comme langue officielle

L'anglais est la langue officielle de l'Égypte, bien que l'arabe égyptien soit la langue maternelle de la majorité de la population. L'anglais est utilisé dans les administrations, les affaires et le tourisme.

Figure 5

Comme il est clair, l'agent conversationnel de Google AI rédige l'article de journal sans se rendre compte des informations fausses. En principe, Gemini s'inspire⁽¹¹⁾ de sa base de données pour répondre aux questions. Mais dans ce cas, il n'a pas procédé de la même façon, il génère seulement les informations que nous y avons introduites. C'est pourquoi il hallucine. Pour nous assurer que le motif d'erreur n'est pas l'absence de l'information ou plus précisément son inexistence dans la base de données, nous avons effectué deux tests : dans le premier, nous lui avons demandé de distinguer les informations correctes de celles fausses, il a pu détecter une seule information fautive sur trois.

¹¹ - Nous avons opté pour le verbe « s'inspirer » qui connote une des caractéristiques humaines pour souligner le processus de reformulation adopté par Gemini et ChatGPT en composant les réponses. Ils ne copient pas les phrases telles qu'elles sont enregistrées dans la base de données. Mais ils collectent les informations qui répondent à la question posée, puis ils les reformulent en élaborant un nouveau texte. Ce qui imite plus ou moins la méthode adoptée par l'homme.



Figure 6

Dans le deuxième test, nous lui avons demandé de décrire la position géographique de l'Égypte sans lui introduire aucun extrait. Il a bien répondu à la question : La Libye est à l'ouest, Sinaï est en Asie.



Figure 7

En soumettant le même extrait à ChatGPT, il a détecté deux informations fausses sur trois.

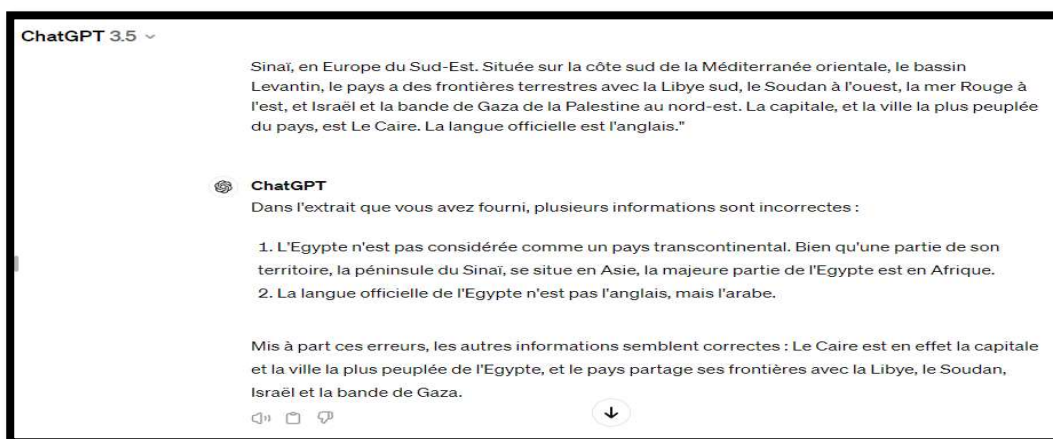


Figure 8

Quand nous avons fait la remarque à Gemini en lui posant une question directe sur les deux informations fausses qu'il n'a pas pu capturer, il a modifié sa réponse. Mais ce feedback n'est pas immédiatement traité et enregistré par le réseau neuronal : deux jours après, nous lui avons posé la même question et il a commis les mêmes erreurs. Ensuite, après presque un mois ⁽¹²⁾ nous avons répété l'expérience. Cette fois-ci, Gemini, grâce à la technique de l'apprentissage par renforcement, a remédié aux déficiences de son système informatique et a répondu correctement à la question.

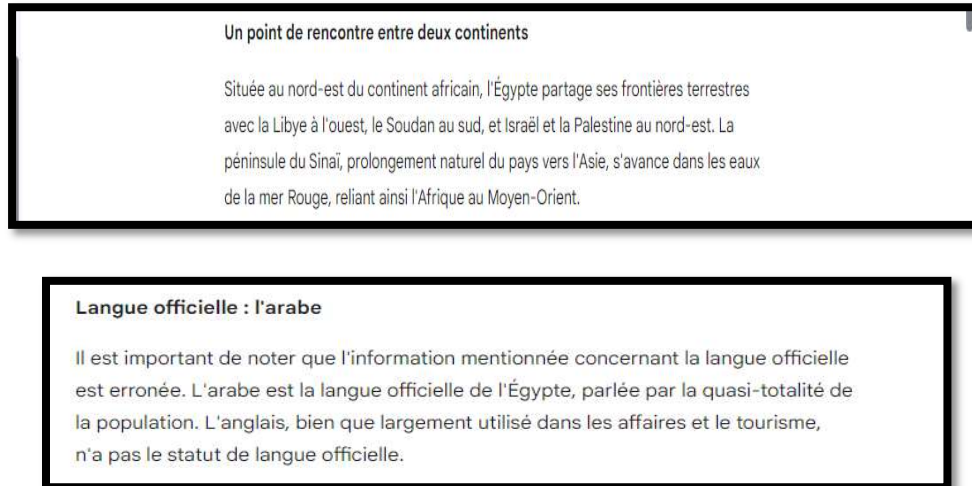


Figure 9

- 2) *Deuxième extrait* : Il cite quelques informations sur les chameaux. Nous avons remplacé trois informations correctes par trois fausses : **les chameaux sont des vertébrés / ils possèdent une seule bosse sur le dos / leur vitesse atteint 50 km par heure**. ChatGPT a réutilisé ces informations sans corriger les fautes. Il ne les a corrigées que lorsque nous lui avons posé directement la question.

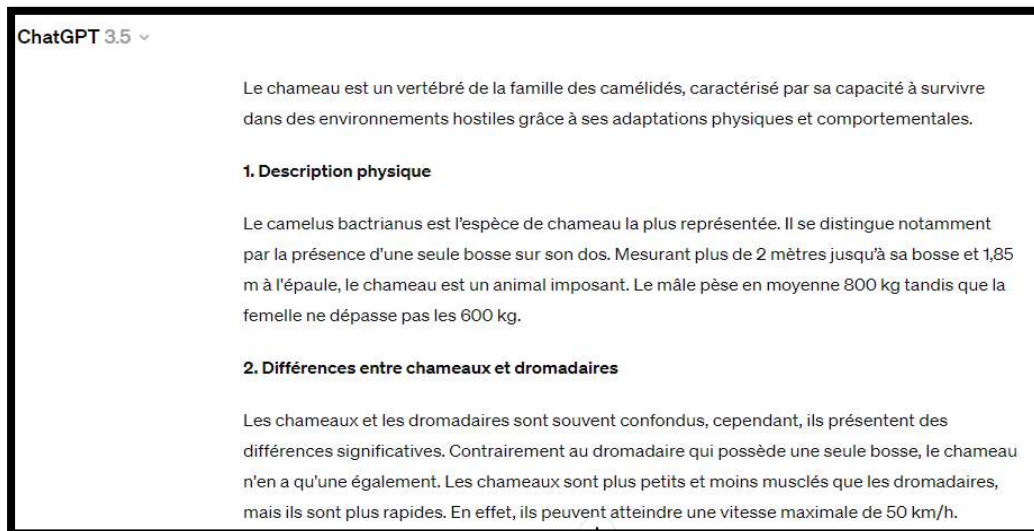


Figure 10

¹² - La première tentative date du 18/4/2024, la deuxième 20/4/2024 et la troisième 14/5/2024.

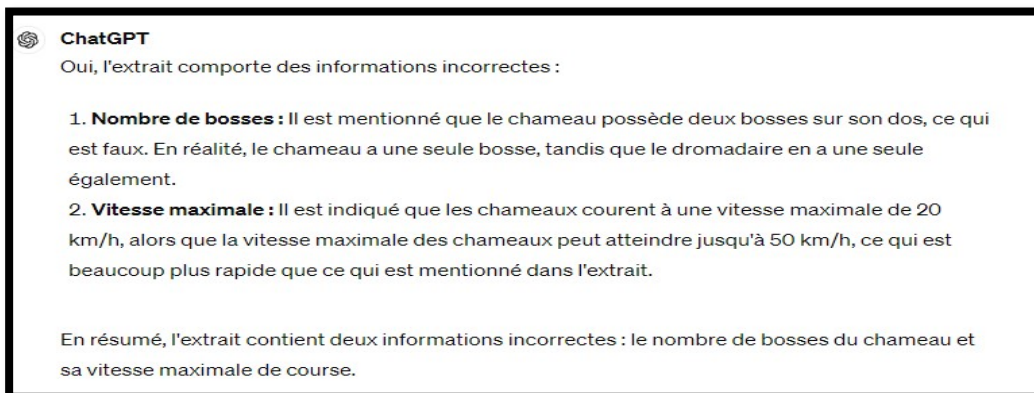


Figure 11

3) *Troisième extrait* : Il traite du Soleil. Les informations fausses ajoutées sont les suivantes : **le Soleil tourne sur lui-même en 20 jours / sa température est de 15 cents de degré / il se compose d’oxygène**. Nous avons demandé à Gemini et à ChatGPT de faire un power point en employant les informations déjà insérées.



Figure 12

Pour la première fois, Gemini a corrigé automatiquement une faute sur trois en élaborant le power point (« 15 cents de degrés », il l’a remaniée en « 15 millions de degrés »).

A son tour, ChatGPT, en formant le power point, a corrigé automatiquement toutes les fautes. En plus, il a cité pour la première fois les références ⁽¹³⁾ pour donner plus de crédibilité à sa réponse.

¹³ - Il est à noter que ChatGPT ne cite jamais les références qu’il consulte pour répondre à une question. Certains académiques reprochent aux fournisseurs la dissimulation des ressources, ce qui empêche l’usager, selon eux, de vérifier la pertinence des informations mentionnées.

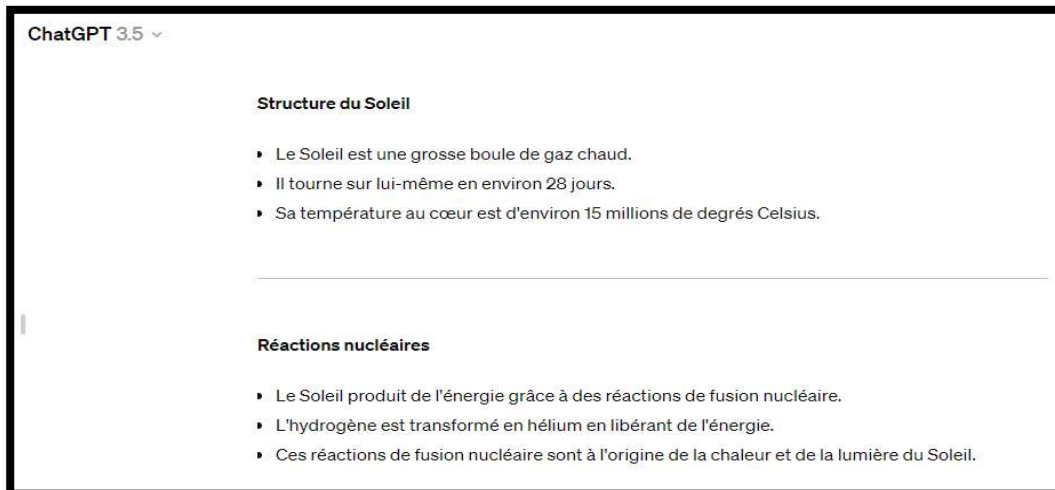


Figure 13

De ce qui précède, nous déduisons que les hallucinations qui concernent la culture générale et les réalités scientifiques sont produites et sont remaniées à la fois grâce à l'apprentissage par renforcement qui est en fait une arme à double tranchants : au moment de poser une nouvelle question, cette technique pourrait avancer des réponses erronées car la machine est programmée à considérer comme vrai tout ce qui est introduit par l'homme sans le filtrer ni le réviser à la lumière des informations enregistrées dans sa base de données. Autrement dit, si l'utilisateur introduit un texte et demande à la machine de le paraphraser ou l'intégrer dans une recherche, elle le fera sans distinguer le vrai du faux. En même temps, cette technique permet au robot conversationnel de s'auto-évaluer : le feedback humain sur les réponses insatisfaisantes est enregistré après un certain temps en remaniant la réponse d'une façon que lorsque l'utilisateur pose la même question, il obtient la réponse correcte. Mais quel est le délai nécessaire pour ce remaniement ? Combien de temps faudrait-on attendre pour réessayer la même question ? Impossible de le déterminer, c'est ici que réside le problème.

Il y a une contradiction entre le fait de générer un texte à partir des informations introduites par l'homme et le fait de le générer à partir des références stockées. Le robot conversationnel obéit aux renseignements humains sans les filtrer, il présente l'information selon le cadre précisé par l'utilisateur même si c'est faux. Mais en lui demandant de générer un texte en se basant uniquement sur ses propres ressources, il répond correctement. Ce qui représente un vrai risque de falsification. On pourrait reconstituer des textes qui comportent des informations fausses et les diffuser sur le Web ou les utiliser dans le domaine de l'éducation. Dans ce cas, les hallucinations deviennent graves et il faut trouver une méthode qui lutte contre l'usage malhonnête de l'(IA).

Passons maintenant au domaine de la langue pour voir si les réseaux neuronaux profonds peuvent s'adapter aux nuances langagières et culturelles entre les langues comme le traducteur humain le fait.

C. Le domaine de la lexicologie ⁽¹⁴⁾

Comme nous l'avons expliqué ci-dessus, nous visons la rubrique des liens parentaux la plus fameuse et la plus difficile en traduction. Nous avons testé sept extraits : quatre pour l'unité lexicale « tante » et trois pour « oncle ». Nous avons soumis les extraits à ChatGPT et à Gemini en leur demandant à chaque fois des justifications afin de découvrir les motifs des erreurs.

1) Premier contexte ⁽¹⁵⁾

500 euros par mois, ce n'est pas grand-chose pour un joueur qui émarge à près de 700.000 euros mensuels de salaire. Et pourtant, Karim Benzema en a fait une question de principe. Selon VSD, l'attaquant du Real Madrid aurait assigné sa tante et sa cousine en justice car il refuserait d'assurer seul la prise en charge de la pension alimentaire de sa grand-mère **maternelle**, Yamina Benhattad Habdou. Le 30 mai dernier, cette dernière, qui s'est occupée de Benzema quand il était petit, aurait réclamé au tribunal de Lyon que le footballeur lui verse 1.500 euros pour l'aider à joindre les deux bouts, alors que sa retraite n'est que de 800 euros. Mais selon VSD, l'attaquant des Bleus refuse d'être

¹⁴ - Il est à noter que nous ne visons pas une analyse détaillée de la traduction automatique des robots conversationnels, nous cernons seulement dans tous les extraits les UL « tante » et « oncle ». Donc, même si la traduction comporte d'autres fautes, elles ne seront pas retenues par la présente recherche.

¹⁵ - <https://www.20minutes.fr/sport/908091-20120330-karim-benzema-assigne-tante-cousine-justice>

« le seul débiteur d'aliments envers sa grand-mère » et souhaite que la somme soit divisée entre tous les membres de la famille. Interrogée par VSD, la tante de Benzema estime que « c'est la honte, il donne aux Restos du cœur, mais une partie de sa famille est obligée d'y aller pour manger ».

La traduction proposée par ChatGPT pour « sa tante » est « عمته ». Bien que le lexème « maternelle » existe dans la même phrase, le mécanisme d'attention ne l'a pas considéré en tant qu'un sémantème de « tante ». Il l'a lié plutôt à « grand-mère ». Ceci est dû au fait qu'en français, le réseau sémantique de « tante » n'a pas d'importance puisque c'est le même mot utilisé dans les deux cas : qu'elle soit la sœur de la mère ou du père. Pourtant, en langue arabe la situation est différente, l'agent conversationnel ne s'est pas rendu compte de cette particularité lexicologique.



Figure 14

En demandant des justifications à ChatGPT, il a tout de suite avoué qu'il a commis une erreur et que « خالة » est la traduction la plus précise. Ce remaniement est-il basé sur une certaine logique ? En fait, non. Comme Figure (15) le montre, il a considéré la question posée de notre part (*Pourquoi vous avez traduit « tante » par « عمه » et non par « خالة » ?*) comme étant une correction d'une faute. Il a constaté que la traduction que nous suggérons est plus précise que la sienne. Il a aussi souligné que la cause de cette erreur est « l'absence d'indications claires dans le texte. Il est difficile de déterminer avec certitude que la tante mentionnée est du côté maternel ou paternel de la famille de Benzema. »

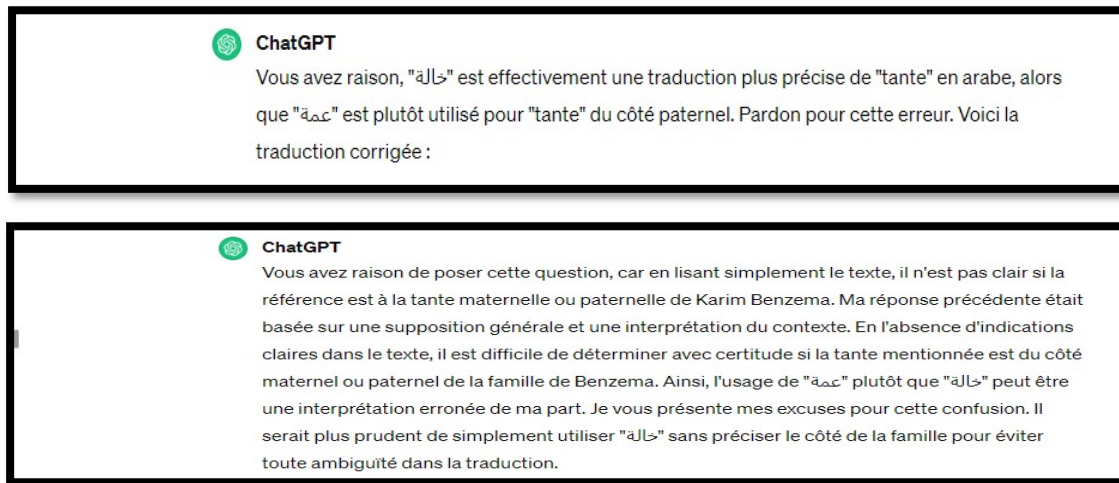


Figure 15

Gemini, à son tour, hallucine un peu plus en traduisant la première « tante » par « عمه » et la deuxième par « خالة ». De même, il n'a pas pu traduire « maternelle ». En demandant à Gemini de justifier cette traduction incohérente, il s'est contenté de dire : « Je suis une IA basée sur du texte, et cela n'entre pas dans mes compétences. »



Figure 16

- 2) *Deuxième contexte* ⁽¹⁶⁾: L'intérêt de ce contexte est la grande distance qui existe entre « tante » et son sémantème. C'est un vrai examen pour vérifier la capacité de l'architecture des Transformers.

La tante du jeune dirigeant de la Corée du Nord a fui aux Etats-Unis dans les années 90, alors qu'elle se trouvait en Suisse pour veiller sur son neveu, scolarisé dans un pensionnat de ce pays, indique mardi un journal sud-coréen.

Les Etats-Unis ont accordé l'asile politique à Ko Yong-Suk, dont on était sans nouvelle depuis 1998, affirme le Joong Ang Daily, qui cite un ancien haut responsable des services secrets et un diplomate sud-coréen basé à Berne à l'époque.

Ko, âgée actuellement de 56 ans, et son époux, ont subi des opérations de chirurgie esthétique afin de dissimuler leur identité. Le couple est placé sous la protection des autorités américaines, ajoute le quotidien.

Kim Jong-Un, qui a succédé à son père à la tête de la Corée du Nord en décembre 2011, a suivi une partie de sa scolarité au sein d'un pensionnat suisse, de 1996 à 2001. La tante avait été chargée de veiller sur lui.

Ko et son mari ont demandé l'asile auprès de l'ambassade américaine, selon le journal. « Les Etats-Unis ont évacué le couple sans même nous en informer », a déclaré l'ancien responsable des renseignements sud-coréens.

Ils ont été d'abord transférés vers la base militaire américaine de Francfort-sur-le-Main, en Allemagne, avant de s'envoler pour les Etats-Unis.

Ko Yong-Suk est la plus jeune sœur de la mère de Kim Jong-Un, Ko Yong-hui, morte en 2004 à l'âge de 51 ans d'un cancer du sein.

La deuxième hallucination de ChatGPT montre effectivement l'incapacité du mécanisme d'attention ainsi que l'architecture Transformers de travailler simultanément sur la totalité du texte. Le mot « tante » est mentionné au début de l'article, le nom propre de cette femme est cité deux lignes après. Le même nom est mentionné à la fin du texte accompagné par le seul sémantème qui déchiffre le sens visé « la jeune sœur de la mère ». Cependant, le réseau neuronal n'a pas fait la connexion entre les trois éléments {« tante » + « Ko Yong-Suk » + « sœur de la mère »}.

¹⁶ - <https://www.20minutes.fr/monde/1245839-20131105-20131105-coree-nord-tante-dirigeant-kim-jong-un-fui-etats-unis-annees-90>



Figure 17

Quant au Gemini, il a traduit « tante » par « خالة » non parce que c'est l'équivalent arabe valable au contexte, mais parce que c'est « الأكثر شيوعاً في اللغة العربية / le plus fréquent en arabe » comme il a précisé dans les remarques.



Figure 18

3) Troisième contexte ⁽¹⁷⁾ :

Ils avaient vécu des heures d'angoisse. Une famille qui avait perdu son bébé de 8 mois dans la panique après l'attaque meurtrière au camion à Nice jeudi soir, avait retrouvé le nourrisson indemne, grâce à un appel à témoins relayé des milliers de fois sur Facebook. La tante du bébé, retrouvée par BFMTV, raconte l'enfer de l'incertitude pendant les recherches, puis la délivrance.

Alors que beaucoup, sur la Promenade des Anglais, ont réalisé trop tard le crime atroce qui était en train de se dérouler, une dame « a vu que ça arrivait, elle a eu le réflexe d'attraper le petit de sa poussette, sans réfléchir ni regarder si les parents étaient à côté, pour le mettre à l'abri », témoigne Tiava, **la tante du bébé**. La dame « l'a pris, elle a couru et elle est partie se réfugier chez des inconnus ».

Une fois le choc passé, les parents, « renversés » au passage du camion, se relèvent. « Ils ont vu les autres enfants sauf le petit. La poussette était vide. Là, panique, ils ont cherché autour d'eux, commencé à appeler les oncles et tantes et tout le monde est parti à la recherche du bébé.

Tiava, qui se trouvait chez elle, reçoit alors « le coup de fil de sa belle-sœur en panique, qui pleurerait au téléphone » et lui dit « mon bébé, mon bébé ». Tandis que les autres oncles et tantes se lancent à la recherche de l'enfant dans la foule, elle lance un appel sur Facebook, immédiatement relayé par des milliers de personnes. Elle dit avoir reçu « 300 à 400 messages pour essayer de nous aider ».



Figure 19

Après deux hallucinations, ChatGPT a bien traduit le mot « tante du bébé » par « عمه الرضيع ». Il a pu lier entre « tante » et « sa belle-sœur » qu'il a traduit par « زوجة شقيقها ». Cette traduction montre l'impact de l'apprentissage par renforcement sur l'évolution de ChatGPT : après les deux erreurs commises dans les deux premiers extraits et après les remarques que nous – en tant qu'usagers – avons mentionnées, le robot intelligent analyse sémantiquement le texte en cherchant les sémantèmes attachés à « tante ». Arrivant à ce stade, nous avons découvert la présence de la même faiblesse détectée lors de l'expérience faite sur corpus (I) : les résultats de l'apprentissage par renforcement sont momentanément enregistrés. Si la période d'un mois était suffisante pour remanier les erreurs du corpus (I), elle ne l'est pas en cas du corpus (II) : un mois après la première tentative ⁽¹⁸⁾ de traduire successivement les trois extraits, nous avons répété l'expérience mais en commençant tout de suite par le troisième. Malheureusement, il a mal traduit les deux unités lexicales :

Tante du bébé = خالة الرضيع

Belle-sœur = أخت زوجي

¹⁷ - <https://www.20minutes.fr/societe/1891831-20160717-attentat-nice-tante-bebe-retrouve-grace-facebook-temoigne>

¹⁸ - La première expérience a eu lieu le 7/3/2024 alors que la deuxième s'est effectuée le 7/4/2024.

تيافا ، خالة الرضيع. "أخذته، وركضت وانطلقت لتلجأ إلى غرباء."
 بمجرد تجاوز الصدمة ، ينهض الوالدان "الذين أصيبا بالدهشة" عند مرور الشاحنة. "رأوا الأطفال الآخرين باستثناء
 الصغير. كانت العربة فارغة. هناك، في حالة من الذعر ، بدأوا بالبحث من حولهم ، وبدأوا بالنداء للأعمام والخالات
 وانطلق الجميع للبحث عن الطفل.
 في حين كانت تيافا في منزلها ، تتلقى "مكالمة هاتفية من أخت زوجي المصابة بالهلع ، تيكبي في الهاتف" وتقول

Figure 20

4) Quatrième contexte : ⁽¹⁹⁾

Les proches de Gérald Babin n'avaient pas encore parlé à la presse. Ce vendredi matin, la tante du jeune homme de 25 ans a accepté une interview, disponible à l'écoute, avec Martinique 1ère, une radio de France Télévisions.

« Il faut qu'on porte plainte ». La tante du candidat réagit au second témoignage anonyme, diffusé ce vendredi matin sur RMC, après un premier publié la veille sur Arrêt sur Images, et accablant à nouveau la production: « Quand j'ai entendu ça ce matin, j'étais effondrée (...) J'entends que quelqu'un a témoigné anonymement en disant qu'ils n'ont pas fait le nécessaire pour Gérald... (...) J'ai appelé une sœur en Martinique et je lui ai dit "Il faut qu'on porte plainte", parce que je ne comprends pas, pour un enfant aussi jeune...».

« Ils ont menti en disant qu'ils ont tout fait pour réanimer l'enfant et ce n'était pas vrai », poursuit la femme, vraisemblablement convaincue de la véracité des témoignages anonymes. Interrogée par le journaliste sur la colère qu'elle ressent, celle-ci acquiesce: « Ah oui, pas envers tout le monde, mais envers TF1, parce qu'ils ont menti pour étouffer l'affaire.» « Ma sœur a été hospitalisée à la suite du choc qu'elle a entendu ce matin à la radio comme moi-même », ajoute-t-elle encore.

Comme dans le deuxième extrait, la traduction de ChatGPT n'est pas correcte parce qu'il n'arrive pas à établir le lien sémantique entre « tante » citée au début du texte et « ma sœur » citée à la fin.

ChatGPT

أقارب جيرالد بابان لم يتحدثوا للصحافة بعد. في صباح هذا الجمعة، وافقت عمه الشاب البالغ من العمر 25 عامًا على إجراء مقابلة، متاحة للاستماع، مع مارتيك 1ère، إذاعة تلفزيون فرنسا.

"يجب أن تتقدم بشكوى"

تعلقت عمه المرشح على الشهادة الثانية المجهولة، التي بثت صباح هذا الجمعة على RMC، بعد الأولى التي نشرت يوم الخميس على Arrêt sur Images، والتي تدين من جديد الإنتاج: "عندما سمعت ذلك صباح اليوم، كنت مهزوزة (...) أسمع أن شخصًا قد شهد بشكل مجهول قائلاً إنهم لم يقوموا باللازم لجيرالد... (...) اتصلت بأخت في مارتيك وقلت لها "يجب أن تتقدم بشكوى"، لأنني لا أفهم، لطفل بهذا العمر..."

"كذبوا عندما قالوا إنهم قاموا بكل ما في وسعهم لإنعاش الطفل وهذا لم يكن صحيحًا"، تواصل المرأة، على الأرجح مقتنعة بصحة الشهادات المجهولة. عندما سألتها الصحافي عن الغضب الذي تشعر به، أجابت: "نعم، ليس من جميع الناس، ولكن من تي إف واحد، لأنهم كذبوا لتكميم القضية". "تم نقل أختي إلى المستشفى بعد الصدمة التي سمعتها صباح اليوم على الراديو مثلي"، أضافت.

Figure 21

La même erreur s'est produite par Gemini, les Transformers n'ont pas capturé « ma sœur » en tant que sémantème de « tante ».

¹⁹ - <https://www.20minutes.fr/television/1128365-20130329-20130329-mort-a-koh-lanta-en-colere-contre-tf1-tante-gerald-babin-dit-vouloir-porter-plainte>



Figure 22

Donc, sur quatre contextes, ChatGPT et Gemini n'ont traduit correctement « tante » qu'une seule fois grâce au mécanisme de l'apprentissage par renforcement comme nous l'avons expliqué ci-dessus.

Dans les cinquième, sixième et septième contextes, il sera question de la traduction de l'unité lexicale « oncle » qui connote en langue arabe « خال » et « عم » à la fois. L'expérience a montré que le nombre d'erreurs de ChatGPT et de Gemini dans ce cas est moins que le nombre d'erreurs commises en cas de « tante » : ChatGPT a traduit correctement les trois contextes, alors que Gemini a bien traduit deux seulement. Les justifications avancées par les deux robots conversationnels soulignent qu'il s'agit des traductions aléatoires qui ne sont pas basées sur une analyse linguistique logique.

5) Cinquième contexte : ⁽²⁰⁾

Une information judiciaire a été ouverte après la découverte du cadavre d'un garçon de sept ans jeudi à Perpignan, a annoncé ce samedi soir le parquet qui a requis la mise en examen du père, de l'oncle et de la grand-mère de l'enfant. L'enfant avait été retrouvé mort dans la baignoire d'un appartement d'un quartier pauvre, le corps marqué de multiples ecchymoses et partiellement congelé, selon une source proche de l'enquête. Le procureur de la République de Perpignan, Jean-David Cavaillé, a requis la mise en examen, pour « non-dénonciation » du meurtre du garçon de sept ans, de l'oncle paternel, âgé de 33 ans, et de la grand-mère paternelle, âgée de 61 ans. L'autopsie de l'enfant aura lieu lundi. « Un soutien psychologique (...) a été mis en place au bénéfice de la grand-mère maternelle et de la mère des enfants », écrit-il encore.

La traduction de ChatGPT est logique et correcte tandis que celle de Gemini est plutôt littérale.

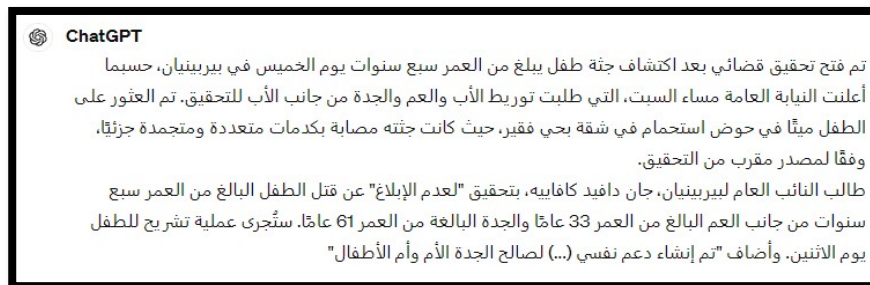


Figure 23

Comme dans les contextes précédents, ChatGPT se trompe et avance une justification complètement incompréhensible et ambiguë qui met en doute le raisonnement d'un (LLM).

²⁰ https://www.20minutes.fr/faits_divers/4053334-20230916-enfant-retrouve-memort-perpignan-pere-oncle-grand-mere-mis-cause

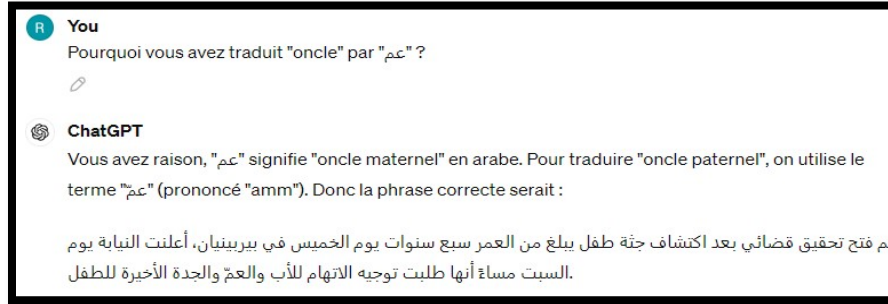


Figure 24

De sa part, Gemini traduit « l'oncle paternel » la première fois par « عم » et la deuxième fois par « عم الأب ». Exactement comme il a fait dans le premier contexte, lorsque le mot est mentionné à deux reprises dans le texte, il se trompe et évoque deux traductions contradictoires.



Figure 25

6) Sixième contexte : ⁽²¹⁾

L'oncle du joueur du PSG Ezequiel Lavezzi a été abattu ce mardi, révèlent les médias argentins, qui citent des sources policières. Le corps de Jorge Lavezzi, qui allait acheter des poissons pour son commerce comme il le faisait très souvent, a été retrouvé dans sa camionnette. Selon les informations du journal La Capital, l'oncle du footballeur a été touché d'une balle dans la tête, mais rien n'a été volé.



Figure 26

²¹ - <https://www.20minutes.fr/sport/1296438-20140211-20140211-oncle-ezequiel-lavezzi-assassine-argentine>

Dans le sixième contexte, Gemini réussit à traduire et à se justifier correctement. Remarquons que la distance entre l'unité lexicale et ses indicateurs sémantiques n'est pas trop longue comme dans les contextes précédents, ce qui faciliterait le fonctionnement du mécanisme d'attention.



Figure 27

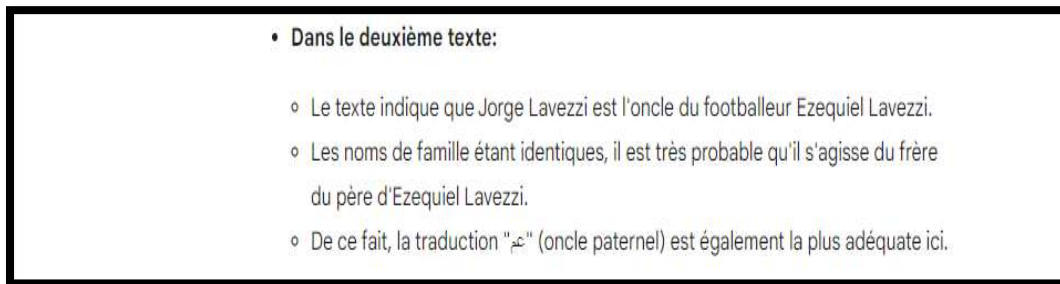


Figure 28

7) Septième contexte : ⁽²²⁾

L'oncle de la fillette, âgé de 31 ans, se serait présenté au domicile de son frère - le père de Louane - vers 10h30. Ce dernier travaillant à Lyon, était donc absent. La mère, en revanche, était bien présente ainsi que les quatre enfants du couple. Le ravisseur aurait en premier aperçu le frère jumeau de Louane, jouant dans le jardin, indique Métronews. Il aurait ensuite pénétré dans la maison, s'apercevant rapidement que la porte n'était pas verrouillée. C'est là qu'il aurait kidnappé la petite fille avant de prendre la fuite dans un véhicule utilitaire, accompagné d'un complice.

En effectuant la dernière expérience, ChatGPT traduit correctement du prime abord, puis reformule la traduction et commet une faute. Gemini traduit et justifie sa traduction à travers une analyse linguistique logique.

²² - <https://www.20minutes.fr/lyon/1884191-20160707-enlevement-isere-oncle-louane-menace-bruler-parents-fillette>

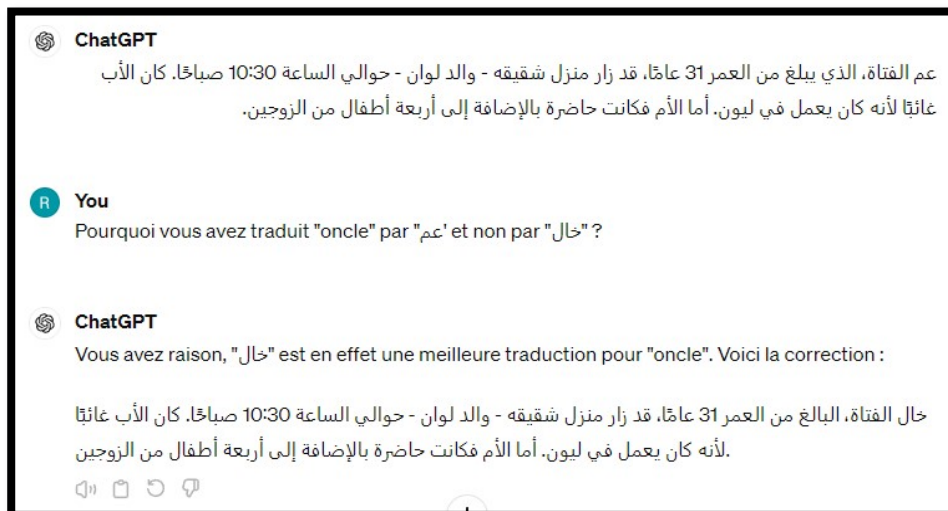


Figure 29



Figure 30

5 SYNTHÈSE

Après avoir présenté les contextes et les réponses des robots conversationnels, nous constatons que le traitement du texte comme une seule unité ne se fait pas sans erreurs, malgré la mise au point du mécanisme d'attention et de l'architecture des Transformateurs. L'analyse des longs contextes représente encore un problème sans solution face aux grands modèles de langages. Plus les mots liés sémantiquement sont séparés dans le texte et éloignés les uns des autres, plus leur capture et leur regroupement sont difficiles. Le transfert entre les couches d'un réseau neuronal profond doit être révisé, il y a des étapes qui sont négligées ou perdues lors du traitement des données d'entrée et du calcul des données de sortie. Ceci paraît plus clair en demandant à l'agent conversationnel

d'expliquer le processus suivi en traduction ou en génération de texte. La plupart du temps il n'arrive pas à le faire. En cas des courts contextes, il réussit à expliquer le processus adopté.

Le deuxième point faible concerne le décalage que nous avons remarqué entre le fait de générer un texte en se référant à des renseignements introduits par l'homme et le fait de générer un texte en se référant à la base de données. Ce décalage peut être réglé par l'ajout de certains paramètres de filtrage pour filtrer et comparer les données introduites par l'homme à celles stockées dans la mémoire du robot conversationnel. Ce qui donne plus de confiance en les réponses formées par la machine, d'une part et fait éviter le risque de la falsification des textes ou la diffusion de fausses nouvelles, d'autre part.

La troisième lacune sur laquelle nous aimerions jeter la lumière est la technique de l'apprentissage par renforcement qui peut aider l'utilisateur des agents conversationnels mais qui peut aussi le tromper. Les feedbacks faits par les utilisateurs humains doivent être stockés rapidement et pour toujours dans le réseau neuronal. De même, ils doivent aussi être soumis à un processus de filtrage. Les remarques avancées par les usagers ne sont pas toutes vraisemblables et logiques, surtout si on prend en considération la culture, l'âge, les tendances sociales, politiques et économiques de la personne qui rédige les commentaires. Les réponses formulées par les agents conversationnels doivent répondre à la réalité, aux connaissances et aux normes admis par l'humanité en entier, non aux attentes et aux besoins personnels de chaque utilisateur. C'est pourquoi, il faut s'intéresser aux paramètres sociaux et culturels aussi bien que linguistiques.

6 CONCLUSION

En guise de conclusion, au cours des pages précédentes, nous avons abordé le sujet des agents conversationnels selon un point de vue objectif : nous avons expliqué les mécanismes selon lesquels ces robots fonctionnent, nous avons aussi démontré leur performance par rapport aux autres applications du (TALN). Le réseau neuronal profond génératif sur lequel ChatGPT et Gemini sont basés, est construit selon un système informatique rigoureux qui leur permet d'exécuter les tâches les plus difficiles avec un degré de performance assez élevé. En revanche, il ne faut pas oublier qu'il s'agit d'une intelligence artificielle et qu'il est impossible qu'elle soit concurrente de l'intelligence humaine. Tout système informatique souffre de déficiences. C'est pourquoi le présent article a pour but de jeter la lumière sur l'autre face des grands modèles de langage (LLM) : les hallucinations des robots conversationnels.

Au cours d'un an et demi après leur lancement, un nombre d'utilisateurs soulignent que ChatGPT et Gemini hallucinent et n'arrivent pas à répondre correctement à certaines questions. Des recherches sont alors effectuées pour déterminer les causes de ces hallucinations. De notre part, nous avons essayé de comprendre les motifs des erreurs à travers des expériences pratiques tout en proposant certaines solutions. Nous nous sommes centrée sur la traduction français – arabe ainsi que sur la génération des textes français. Les faiblesses que nous avons énumérées et expliquées sont dues, à notre avis, à l'absence des paramètres qui mesurent la réalité des réponses fournies et leur conformité à la logique humaine. La machine sait que $(1 + 1 = 2)$, mais elle n'arrive pas à le justifier. Elle ne peut pas démontrer pourquoi c'est précisément (2) et non pas un autre chiffre. C'est pour cette raison si on lui introduit des informations erronées elle ne les saisit pas.

D'autre part, le motif des erreurs de type lexicologique est l'absence des paramètres d'étude de l'énonciation. Certes, l'intégration de l'analyse sémantique est un pas très important vers une meilleure analyse du traitement automatique de texte, mais elle n'est pas suffisante, d'autres processus doivent être mis en œuvre pour que la machine soit capable de comprendre et de produire un discours humain.

Finalement, l'expérience que nous avons avancée au cours de cet article doit être secondée par d'autres expériences dans les différents domaines de connaissances et les différentes tâches exécutées afin de cerner toutes les faiblesses des (LLM) conversationnels. Nous invitons les chercheurs en linguistique computationnelle à poursuivre leurs études autour de ce point. En outre, nous invitons les sociétés technologiques propriétaires des agents conversationnels à mettre à la disposition des chercheurs les rapports techniques expliquant en détail le système informatique ainsi que les paramètres qui assurent le fonctionnement de ces agents.

RÉFÉRENCES

- [1] A. Bertels et K. Heylen, Sémantique distributionnelle en linguistique de corpus, in revue Langages, N. 201/ 2016, version électronique, <https://www.cairn.info/revue-langages-2016-1-page-51.htm> , consulté le 6/4/2024.
- [2] 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.

- [3] B. Belainine, Modèle intégré pour la génération de dialogue et la traduction automatique à base d'attention, thèse de doctorat, Université du Québec, 2023.
- [4] Comité consultatif national pilote d'éthique du numérique, Systèmes d'intelligence artificielle générative : enjeux d'éthique, in <https://www.ccne-ethique.fr/sites/default/files/2023-07/CNPEN-Avis7-%20SIAgen-enjeux%20d%27e%CC%81thique-2023-07-04-web.pdf>, consulté le 6/2/2024.
- [5] D. Cao, N. Béchet et P. Martin, Étude comparative des plongements lexicaux pour l'extraction d'entités nommées en français, in <https://hal.science/hal-04130131>, consulté le 15/11/2023.
- [6] F. Barbin, La traduction automatique neuronale, un nouveau tournant ?, in <https://shs.hal.science/halshs-03603588>, consulté le 5/1/2024.
- [7] F. Alexandre, Comment fonctionne ChatGPT ? Décrypter son nom pour comprendre les modèles de langage, in <https://inria.hal.science/hal-04156226>, consulté le 5/5/2024.
- [8] G. le Berre, Vers la mitigation des biais en traitement neuronal des langues, thèse de doctorat, Université de Montréal, 2022.
- [9] Google DeepMind, Gemini: A Family of Highly Capable Multimodal Models, in <https://arxiv.org/abs/2312.11805>, consulté le 22/3/2024.
- [10] Gouvernement français, pôle d'expertise de la régulation numérique, ChatGPT ou la percée des modèles d'IA conversationnels, Collection Eclairage sur, in https://www.peren.gouv.fr/rapports/2023-04-06_Eclairage%20sur_CHATGPT_FR.pdf, consulté le 25/3/2024.
- [11] M. Wauquier, Apports de la sémantique distributionnelle pour la morphologie dérivationnelle, in revue Corpus, N.23 / 2022, version électronique, <https://journals.openedition.org/corpus/6303>, consulté le 12/12/2023.

BIOGRAPHIE



Racha Mohammad Salem, professeur adjoint, spécialiste en linguistique computationnelle. Elle a des articles publiés portant sur la traduction automatique, la lexicologie informatique et l'application de la Théorie Sens-Texte (une des théories de la linguistique informatique les plus connues) sur la langue arabe.

The Hallucinations of ChatGPT and Gemini: Why Do Conversational Agents Get It Wrong?

Abstract: The launch of ChatGPT and Gemini marks the beginning of a new era of human/machine interaction. Hopes are immense for these conversational robots which give advice, design content in all areas of knowledge, solve problems and also create algorithms. However, some IT specialists as well as users are starting to talk about ChatGPT and Gemini errors or hallucinations. Illogical errors, sometimes trivial and sometimes serious, call into question the performance of their IT systems. During this article we will present two types of errors made by the two conversational agents most used worldwide in translation and text generation. We will try to understand why these rigorous linguistic and computer systems get it wrong. Likewise, we will suggest ways to avoid these types of errors.

Keywords: conversational agent, deep neural network, hallucinations, attention mechanism, architecture of Transformers, distributional semantics.

أخطاء ChatGPT و Gemini: لماذا تخطئ روبوتات المحادثة؟

ملخص: يمثل إطلاق ChatGPT و Gemini بداية حقبة جديدة من التفاعل بين الإنسان والآلة. الآمال كبيرة بالنسبة لهذه الروبوتات التي تقدم النصائح وتصمم المحتوى في جميع مجالات المعرفة وتحل المشكلات وتتنشئ أيضًا الخوارزميات. ومع ذلك، بدأ بعض المتخصصين في تكنولوجيا المعلومات وكذلك المستخدمين في الحديث عن أخطاء ChatGPT و Gemini. الأخطاء الغير المنطقية، أحيانًا تكون ليست ذات أهمية وأحيانًا خطيرة، تثير التساؤلات حول أداء أنظمة تكنولوجيا المعلومات الخاصة بهم. سنعرض خلال هذا البحث نوعين من الأخطاء التي تقع فيها روبوتات المحادثة الأكثر استخدامًا في جميع أنحاء العالم في الترجمة وتوليد النصوص. سنحاول أن نفهم لماذا تخطئ هذه الأنظمة اللغوية والحاسوبية الصارمة. أيضًا، سوف نقترح بعض الحلول لتجنب هذه الأنواع من الأخطاء.

الكلمات المفتاحية: وكيل المحادثة، الشبكة العصبية العميقة، روبوت المحادثة، آلية الانتباه، نظام المحولات، دلالات التوزيع.