Le problème des hallucinations dans l'accès à l'information scientifique fiable par les LLMs: verrous et opportunités

Benjamin Vendeville[1], Liana Ermakova[2] and Pierre De Loor[3]

[1]*Université de Bretagne Occidentale / Lab-STICC (UMR CNRS 6285), Brest France*
[2]*Université de Bretagne Occidentale / HCTI, Brest France*
[3]*ENIB / Lab-STICC (UMR CNRS 6285), Brest, France*

Abstract

L'évolution des LLMs (Large Langage Models) a profondément impacté la manière dont les individus interagissent avec l'information. Les requêtes de recherche traditionnelles (c'est à dire en passant par exemple par des moteurs de recherche comme google ou par wikipedia) sont remplacées par des requêtes à des IA génératives utilisant des LLMs. Cet article se concentre sur un aspect critique des LLMs, à savoir les hallucinations, en se concentrant sur l'exemple de la simplification automatique de textes scientifiques. Les hallucinations sont définies comme la génération par les LLMs de données vraisemblables mais incorrectes. L'article en donne une formalisation et explore ensuite les limites des métriques traditionnelles d'évaluation de qualité de simplification comme BLEU (*bilingual evaluation understudy* ou *étude d'évaluation bilingue*) ou ROUGE[1](*Recall-Oriented Understudy for Gisting Evaluation* ou*Sous-étude axée sur le rappel pour l'évaluation de l'étiquetage*) pour évaluer l'hallucination et argumente que des jeux de tests basés sur des corpus adaptés ainsi que des méthodes de classificateurs d'implication et de reconnaissances d'entités permettront de mieux évaluer la fiabilité des LLMs en simplification scientifique.

Keywords

Hallucinations, LLMs, Recherche d'information, Simplification, Vulgarisation scientifique

1. Introduction

Les avancées récentes en Traitement Automatique du Langage Naturel (TALN) ont montrés l'efficacité des transformers pour la génération de texte [2]. Cela a donné lieu aux grands modèles de langage - LLMs (Large Language models) qui sont des modèles entraînés sur de gros corpus de textes variés [3]. Ceux-ci font partie des modèles de fondation [4], entraînés sur de gros corpus génériques d'une modalité (langage, image ou autre) puis ré-entraînés sur des tâches spécifiques de l'utilisateur avec peu de données. Ce paradigme a permit une grande avancée dans le TALN et a asserti l'omniprésence des LLMs dans la recherche.

Récemment, les LLMs ont profondément changé la manière dont les citoyens sont impliqués dans les processus de recherche [5]. Traditionnellement, depuis l'arrivée d'internet, les

CORIA 24: COnférence en Recherche d'Information et Applications, April 03–04, 2024, La Rochelle, France
benjamin.vendeville@univ-brest.fr (B. Vendeville); liana.ermakova@univ-brest.fr (L. Ermakova); deloor@enib.fr (P. D. Loor)
0009-0003-5298-147X (B. Vendeville); 0000-0002-7598-7474 (L. Ermakova); 0000-0002-5415-5505 (P. D. Loor)

recherches d'informations par le grand public se font au travers des moteurs de recherche tels que Google, en passant par des sites comme Wikipedia. Cependant, les requêtes de recherche traditionnelles sont désormais remplacées par l'IA générative et la recherche conversationnelle [6, 7] basées sur les LLMs.

En conséquence, la maîtrise de l'information coïncide avec la maîtrise de l'IA. Il est important de comprendre l'acquisition des connaissances et le raisonnement des LLMs. L'évaluation de la fiabilité et de la factualité dans les LLMs reste un défi [8]. Les hallucinations sont l'un des problèmes inhérents des modèles génératifs car les LLMs sont entraînés à compléter une invite textuelle mais en mode "génératif" peuvent devenir instables, c'est-à-dire que leur capacité à généraliser peut les amener à proposer des complétions aberrantes [9].

L'hallucination des LLMs est souvent définie comme *la tendance des LLMs à générer des informations plausibles mais incorrectes ou trompeuses qui ne sont pas ancrées dans la réalité* [10, 9, 11]. Les hallucinations sont difficiles à détecter pour les lecteurs humains [9], ce qui rend difficile à déterminer si l'information existe ou non. Cela rend également difficile d'évaluer la fiabilité de la source [7]. Par exemple, lorsqu'on demande à ChatGPT de fournit des sources, celles-ci sont correctes ou partiellement correctes dans la moitié des cas, mais les références fournies n'existent réellement que dans une 10% des réponses [12].

Les hallucinations de ces modèles réduisent la fiabilité des réponses obtenues pour une demande de l'utilisateur, y compris l'accès aux connaissances scientifiques. On peut même affirmer que ces modèles renforcent le besoin de preuves scientifiques fondées pour valider ces réponses. c'est particulièrement vrai lors de tâches traitant de documents scientifiques. Cette tâche est particulièrement intéressante pour la question des hallucinations car il est nécessaire de garder certaines informations du texte de départ, et peut-être de rajouter des informations qui n'y sont pas mais qui sont véridiques. Cela pose des questions d'extraction d'informations et d'évaluation de la véracité des informations.

Pour évaluer la capacité des LLMs à donner accès à des connaissances fiables, nous étudierons l'hallucination des LLMs dans un contexte de simplification scientifique pour non-experts.

Récemment, les méthodes utilisées pour la simplification de texte sont des LLMs utilisant du fine-tuning et du prompt engineering [9] mais elles gardent ces problématiques d'hallucinations. Table 1 montre les résultats rencontrés par les participants au projet SimpleText [9]. On observe une grande disparité entre les modèles, beaucoup ayant produit très peu de textes présentant des hallucinations, alors que certains en contiennent dans 20%, 30%, voir même dans plus de 50% des résultats. De plus, on peut voir que les métriques BLEU, SARI et FKGL (*flesch kincaid grade level*, ou *niveau scolaire de Flesch Kincaid*) n'ont pas l'air d'être corrélées avec les hallucinations.

Évaluer la capacité des LLMs à donner des connaissances fiables implique des métriques quantitatives et une évaluation qualitative pour mesurer le taux et la gravité des hallucinations et la promotion d'informations factuellement correctes. En particulier, [7] montre que les GPT3.5 et GPT4 ne donnent des sources factuelles (qui renvoient vers un papier qui existe réellement) que dans 22% des cas. La recherche montre que les métriques standards d'évaluation de la qualité d'un texte telles que BLEU ou ROUGE ne sont pas adaptées à la détection d'hallucinations [13, 14]. Pour améliorer la fiabilité des LLMs il est nécessaire de créer et des jeux de tests sur les hallucinations, à la fois dans des contextes généraux et à la fois dans des contextes spécifiques

Table 1
Résultats de simpleText concernant les contenus fallacieux (Spurious Content). Reproduit de [9].

Modèle	Fraction d'hallucinations	FKGL	SARI	BLEU
NLPalma_BLOOMZ	0.55	9.61	35.66	5.76
UAms_Large_KIS150	0.28	10.50	33.02	14.59
AiirLab_run1	0.24	9.86	30.07	15.93
irgc_pegasusTuner007plus_plus	0.23	12.74	23.28	17.42
irgc_t5	0.04	9.56	37.83	15.85
irgc_t5_noaron	0.04	9.55	37.84	15.84
CYUT_run1	0.02	9.63	47.98	14.81
AiirLab_davinci	0.01	11.17	47.10	18.68
CYUT_run2	0.01	8.43	44.93	12.09
Pandas_submission_ensemble	0.01	10.51	40.25	17.40
QH_run1	0.01	12.45	26.46	21.23
UAms_Large_KIS150_Clip	0.00	11.12	33.47	16.59
ThePunDetectives_SimpleT5	0.00	12.92	25.87	21.79

tels que la simplification automatique de textes scientifiques.

Nous proposons dans un premier temps de préciser la notion d'hallucinations telle qu'on l'entend dans notre problématique, puis de lister plusieurs méthodes et métriques qui nous semblent intéressantes pour les limiter.

2. Définition d'une hallucination

Le terme d'hallucination dans la génération de texte est, à notre connaissance, apparu dans [15] sans en donner de définition formelle. Par la suite, [16, 17] définissent les hallucinations comme le fait de "générer des informations en apparence plausibles mais factuellement incorrectes". Cependant, ce terme s'est généralisé pour définir les erreurs qui ne sont pas d'ordre du langage (grammaticales, orthographiques ou syntaxiques). A titre d'exemple, [18] donne une définition "générale" d'halluciner comme " le fait de produire des contenus absurdes ou fallacieux par rapport à certaines sources".

Cette définition générale est reprise dans [11], qui précise d'autres formes plus spécifiques tels que "générer du texte qui n'est pas «fidèle» au contenu de l'entrée (c'est-à-dire que l'information n'existe pas dans la source d'entrée)". Cette dernière phrase montre bien une certaine ambiguïté entre la fidélité une génération avec une "vérité générale" ou avec seulement certaines informations en entrée. Il est clair qu'on ne peut parler d'une hallucination que par rapport à un contexte spécifique, par exemple le contexte de la question, ou d'un consensus scientifique. Sans définir le contexte tout peut être défini comme une hallucination. Cette notion a notamment apparaît dans [19] qui défini deux types d'hallucinations qui dans un contexte de résumé:

- *Hallucination intrinsèque*, qui apparaît lorsque "le texte final comprend des concepts du texte de départ mais en déformant l'information". Il n'est pas question de savoir si le résultat est factuel ou non.

- *Hallucination extrinsèque*, qui apparaît lorsque "le texte final comprend des concepts non utilisés dans le texte de départ". Ce terme ne fait en revanche pas de distinction entre les hallucinations qui se basent sur un fait réel et celles complètement fictives.

En plus de cela, dans certains contextes, cette "vérité générale" peut être difficile à déterminer. Le propre d'un papier de recherche est de faire avancer la science, en démontrant des nouvelles choses ou en allant, par exemple, à l'encontre de certains courants de pensée établis. Dans ce contexte, certains faits d'un papier ne seront pas directement supportés par le consensus scientifique. Enfin, il n'est pas abordé la question de générations inutiles. Si l'on pose une question à un modèle et que sa réponse est factuellement vraie mais complètement hors sujet, est-ce que cela compte comme une hallucination ? La définition de [11] basée sur "la fidélité de la génération à l'entrée" laisse planer le doute.

La classification *"intrinsèque/extrinsèque"* de [19] ne nous semble pas suffisante pour tenir compte de ces questions. En conséquence nous proposons de construire une définition plus précise et adaptée au contexte de la simplificaiton scientifique.

Nous nous intéressons au contexte de la simplification automatique de papiers scientifiques. Ces textes se basent sur un consensus scientifique, un ensemble de connaissances cohérentes. À ces connaissances, le papier peut ajouter de nouvelles clauses qui découlent ou non des clauses faisant déjà partie de cet ensemble. Il est à noter que les papiers peuvent aussi comprendre des clauses qui ne font pas parties de ce consensus. Ces cas là peuvent arriver lorsque de nouvelles observations vont à l'encontre d'idées établies, ou dans des cas d'erreurs (volontaires ou non). Dans ces cas là, il devient beaucoup plus dur de déterminer ce qui est une information factuelle ou non. Nous faisons donc le choix de nous placer dans un cas où les textes scientifiques ne vont pas à l'encontre du consensus scientifique.

En d'autres termes, nous voulons pouvoir partir d'un Document scientifique qui contient un ensemble de faits D et arriver à un document Simplifié, qui contient les faits S idéalement tel que $S = D$. Dans la réalité, pour produire un document simplifié, il convient peut-être d'oublier certaines informations inutiles à la compréhension mais d'inclure des explications, et donc des faits, qui ne sont pas dans le document de départ. Dans ce cas, les informations doivent venir du Contexte C du consensus de la recherche scientifique, en d'autres termes $S \subset (D \cup C)$.

Dans notre contexte, nous définissons deux types d'hallucinations:

- Les hallucinations factuelles H_f qui donnent des informations fausses.
- Les hallucinations de sujet H_s qui donnent des informations pas nécessairement fausses, mais qui n'ont rien ou trop peu à voir avec le document original.

Les hallucinations H_s sont moins problématiques, en effet, même si elles peuvent complexifier la lecture d'un document, elles ne vont pas induire le lecteur en erreur, en revanche, les hallucinations H_f le peuvent et sont donc beaucoup plus graves.

Dans les hallucinations H_f il convient à nouveau de distinguer deux types d'hallucinations:

- Les hallucinations factuelles de document H_{fD} qui sont des faits démontrablement faux à partir du seul document de départ.

- Les hallucinations factuelles de contexte H_{fC} qui sont des faits non démontrables ni à partir du document de départ ni à partir du contexte de la recherche scientifique.

Cette définition des hallucinations ne tient pas compte du cas où une information i est donnée dans le résultat $i \in S$ et appartiens au contexte $i \in C$ mais pas au texte de départ $i \notin D$. Nous considérons que, dans notre contexte, l'ajout d'une information vraie et pertinente dans le résultat ne constitue pas une hallucination. Un bon exemple serait l'ajout d'une définition d'un mot du texte original: c'est une information nouvelle, mais, à priori, pertinente.

Cette nouvelle classification est plus formelle et nous permet de définir des métriques plus adaptées à la détection d'hallucinations. Il est clairement question de faits non inclus dans une source et non démontrables à partir de celle-ci. Ceci nous permet d'utiliser des concepts de logique formelle (déduction, contradiction) pour vérifier la présence ou non d'hallucination.

Garder la distinction entre les hallucinations relatives au document et celles relatives au contexte nous paraissait également important. Cette distinction permet de mesurer spécifiquement les hallucinations relatives au document qui est une tâche à priori plus facile que sur le contexte entier. C'est particulièrement vrai pour évaluer des modèles de génération augmentée de récupération (Retrieval Augmented Generation). Ces modèles utilisent en entrée un corpus pouvant être utilisé pour générer des informations. Dans ce cas, pouvoir estimer les hallucinations par rapport à ce corpus est intéressant. Nous nous retrouvons donc face à deux types d'hallucinations différentes. Nous pensons qu'il est plus intéressant de développer des jeux de tests séparés, notamment parce que les hallucinations factuelles de document sont à priori beaucoup plus faciles à mesurer que les hallucinations factuelles de contexte.

3. Jeux de tests

La problématique de détection des hallucinations est relativement nouvelle. Nous avons recherché l'existence d'autres jeux de tests destinés à résoudre cette question mais il n'existe pas de méthode adaptée. Nous proposons donc de créer de nouveaux jeux de tests pour résoudre cette question.

Le projet SHROOMS [20] étudie la détection d'hallucinations. Pour cela, il propose notamment un jeu de données de textes (générés par LLM) labelisées qui permet d'entraîner des modèles. Des modèles sont été évalués sur des tâches de traduction, de génération de définitions, et de génération de paraphrase. Les réponses ont été évaluées (par rapport au texte de départ et une réponse attendue) par 3 humains pour déterminer si elles présentent une hallucination ou non. Les phrases générées reçoivent donc un score d'hallucination de 0, 1/3, 2/3, ou 1 selon les réponses des examinateurs. Le projet ELOQUENT [21] utilise ce jeu de données pour mesurer la capacité de LLMs multilinguistes à détecter et générer des hallucinations. La capacité à détecter des hallucinations sera mesurée par une métrique de similarité sémantique entre les textes générés d'une part et de l'autre une réponse générée par un humain.

Le projet SimpleText [22] propose un jeu de données de simplifications labelisées pour la simplification automatique de textes scientifiques. Nous proposons d'utiliser ces jeux de données

pour créer des benchmarks (ensemble de jeux de données et de jeux de tests) d'hallucinations pour LLMs.

Pour évaluer la capacité des LLMs à donner accès à des connaissances fiables il faut des métriques quantitatives et une évaluation qualitative afin de mesurer le taux et la sévérité des hallucinations. En simplification de texte, les mesures habituellement utilisées sont FKGL, SARI ET BLEU[9].

FKGL [23] est une métrique développée en 1975 pour l'évaluation des capacités de lectures des marins américains. Cette mesure se base sur le nombre de mots, de phrases et de syllabes, pour calculer le niveau de lecture nécessaire à la compréhension du texte. Cette métrique ne mesure aucune information sémantique et ne permet pas d'évaluer la présence d'hallucination.[9]

BLEU [24] est une métrique crée pour évaluer la qualité de traductions automatiques. Cette métrique se base sur la corrélation entre le texte généré et une référence générée par un humain. BLEU est faiblement voire non corrélé avec les références humaines sur la simplification [25] notamment sur la de préservation de sens et donc d'hallucination.

SARI [26] est une métrique développée pour l'évaluation de simplification automatique. Cette métrique se base sur le nombre de transformations (mots ajoutés ou supprimés) pour arriver à la phrase générée en partant de l'entrée. Cette mesure a été montrée comme corrélant faiblement avec la présence d'hallucinations [9]

Ayant peu de métriques sur les hallucinations spécifiques à la simplification de textes scientifiques, nous proposons de chercher à utiliser et adapter des métriques de tâches similaires.

La génération de résumé se rapproche de notre problématique: il faut générer un texte à partir d'un autre texte en entrée en gardant les informations importantes. Dans [19] les chercheurs ont utilisés un jeu de données composé de documents, leur résumé fait à la main par des humains, et des résumés fait par des modèles d'apprentissage. Ils ont ensuite annoté à la main les résumés comme étant *faithful* (aucune hallucination) ou *factual* (information non contenue dans le texte de départ mais correcte). Enfin, ils ont calculés (Table 2) la corrélation de Spearman de différentes métriques avec les annotations humaines.

Ici, la meilleure métrique est un classificateur d'implication (entailment classifier, abrégé "Entailment" dans le tableau) entraîné à partir de BERT-Large [27] sur le jeu de données Multi-NLI [28], les métriques ROUGE [1] et BERTScore [29] ayant des performances bien inférieures.

Le jeu de données Multi-NLI est composé de paires de phrases labelisées (de plusieures thématiques) selon que leur relation soit *ENTAILMENT* (Implication), *NEUTRAL*, ou *CONTRADICTION*, exemple en Table 3. Le jeu est composé de plusieurs phrases *Prémisses* labelisées en fonction de leur thématique (contexte du texte, exemple: politique, sciences etc...). A partir de ces phrases des humains ont générés trois phrases (Hypothèses) qui ont respectivement une relation d'implication (entailement), neutre (neutral) ou contradictoire (contradiction) avec la prémisse. Une étape de validation a présenté chaque couple *prémisse-*

Table 2
Corrélation des métriques avec annotateurs humains sur les labels "Fidèles" (faits présents dans le document en entrée) et "Factuels"(faits véridiques par rapport au contexte général). Reproduit de [19].

Métriques	Fidèle	Factuels
ROUGE-1	0.197	0.125
ROUGE-2	0.162	0.095
ROUGE-L	0.162	0.113
BERTScore	0.190	0.116
QA	0.044	0.027
Entailment	**0.431**	**0.264**

Table 3
Exemples du dataset MultiNLI avec leurs thématique, les labels de validation (abrégées E pour entailment ou implication, N pour neutre, et C pour contradictoire) attribuées par des annotateurs individuels et leurs labels (déterminé en fonction du label de validation majoritaire). Reproduit de [28].

Prémisse	Labels	Hypothèse
Met my first girlfriend that way.	FACE-TO-FACE **contradiction** *C C N C*	I didn't meet my first girlfriend until later.
8 million in relief in the form of emergency housing.	GOVERNMENT **neutral** *N N N N*	The 8 million dollars for emergency housing was still not enough to solve the problem
Now, as children tend their gardens, they have a new appreciation of their relationship to the land, their cultural heritage, and their community.	LETTERS **neutral** *N N N N*	All of the children love working in their gardens.
At 8:34, the Boston Center controller received a third transmission from American 11.	9/11 **entailment** *E E E E*	The Boston Center controller got a third transmission from American 11.
I am a lacto-vegetarian.	SLATE **neutral** *N N E N*	I didn't meet my first girlfriend until later.
someone else noticed it and i said well i guess that's true and it was somewhat melodious in other words it wasn't just you know it was really funny.	TELEPHONE **contradiction** *C C C C*	No one noticed and it wasn't funny at all.

hypothèse à 4 annotateurs humains chargés d'annoter la relation entre ces phrases. Le label final est déterminé par le label que les annotateurs ont majoritairement donnés à ce couple de phrases.

Dans [30] les chercheurs présentent le modèle $fact_{acc}$. Ce modèle se base sur des modèles de

Table 4
Corrélation de différentes métriques avec l'évaluation humaine, reproduit de [30].

Métriques	**Corrélation avec les scores humains**
ROUGE-1	0.583
ROUGE-2	0.639
ROUGE-L	0.634
OpenIE	0.258
$fact_{acc}$-Binary Classifier	0.596
$fact_{acc}$-Relation Classifier	0.523
$fact_{acc}$-E2E	0.645
$fact_{acc}$-E2E-Reduced	0.668

reconnaissance d'entités (Named Entity Recognition Models) pour déterminer des entités, puis un classificateur de relations pour déterminer leurs relations. Enfin, ils comparent l'ensemble des tuples de relations (sujet, relation, objet) obtenu dans un texte et dans son résumé pour déterminer si les informations obtenues dans le texte sont factuelles. Les résultats (nommés $fact_{acc}$) sont encourageants (Table 4) mais présentent certaines limitations comme la capacité à généraliser les prédictions du modèle de reconnaissance d'entités.

Ces métriques sont intéressantes et pourraient être utilisées pour des tâches de simplification de texte pour mesurer les hallucinations factuelles de document, mais pour cela elles doivent être entraînées sur un jeu de données adapté à la simplification de textes scientifique. Nous suggérons que la recherche doit permettre de créer un jeu de données composé de documents scientifiques pour pouvoir fine-tune des modèles similaires aux deux modèles présentés. Nous pensons que ces modèles constituerons des métriques suffisamment fiables comparées aux classifications humaines [9] pour pouvoir faciliter l'évaluation.

4. Conclusion

Dans ce papier nous avons présenté le problème des hallucinations des LLMs. Nous avons listés différentes métriques, leur capacité à détecter ce problème, et la manière dont nous pouvons les adapter à notre contexte et créer des jeux de tests d'hallucinations. Les nouvelles idées développées ici permettront d'améliorer l'évaluation des hallucinations des méthodes simplification automatique de textes scientifiques.

Acknowledgments

Cette recherche a été financée en tout ou partie, par l'Agence Nationale de la Recherche (ANR) au titre du projet « ANR-22-CE23-0019-01 ».

References

[1] C.-Y. Lin, ROUGE: A Package for Automatic Evaluation of Summaries, in: Text Summarization Branches Out, Association for Computational Linguistics, Barcelona, Spain, 2004, pp. 74–81.

[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention Is All You Need, 2023. URL: http://arxiv.org/abs/1706.03762. doi:10.48550/arXiv.1706.03762, arXiv:1706.03762 [cs].

[3] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language Models are Few-Shot Learners, 2020. URL: http://arxiv.org/abs/2005.14165. doi:10.48550/arXiv.2005.14165, arXiv:2005.14165 [cs].

[4] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, E. Brynjolfsson, S. Buch, D. Card, R. Castellon, N. Chatterji, A. Chen, K. Creel, J. Q. Davis, D. Demszky, C. Donahue, M. Doumbouya, E. Durmus, S. Ermon, J. Etchemendy, K. Ethayarajh, L. Fei-Fei, C. Finn, T. Gale, L. Gillespie, K. Goel, N. Goodman, S. Grossman, N. Guha, T. Hashimoto, P. Henderson, J. Hewitt, D. E. Ho, J. Hong, K. Hsu, J. Huang, T. Icard, S. Jain, D. Jurafsky, P. Kalluri, S. Karamcheti, G. Keeling, F. Khani, O. Khattab, P. W. Koh, M. Krass, R. Krishna, R. Kuditipudi, A. Kumar, F. Ladhak, M. Lee, T. Lee, J. Leskovec, I. Levent, X. L. Li, X. Li, T. Ma, A. Malik, C. D. Manning, S. Mirchandani, E. Mitchell, Z. Munyikwa, S. Nair, A. Narayan, D. Narayanan, B. Newman, A. Nie, J. C. Niebles, H. Nilforoshan, J. Nyarko, G. Ogut, L. Orr, I. Papadimitriou, J. S. Park, C. Piech, E. Portelance, C. Potts, A. Raghunathan, R. Reich, H. Ren, F. Rong, Y. Roohani, C. Ruiz, J. Ryan, C. Ré, D. Sadigh, S. Sagawa, K. Santhanam, A. Shih, K. Srinivasan, A. Tamkin, R. Taori, A. W. Thomas, F. Tramèr, R. E. Wang, W. Wang, B. Wu, J. Wu, Y. Wu, S. M. Xie, M. Yasunaga, J. You, M. Zaharia, M. Zhang, T. Zhang, X. Zhang, Y. Zhang, L. Zheng, K. Zhou, P. Liang, On the Opportunities and Risks of Foundation Models, 2022. URL: http://arxiv.org/abs/2108.07258. doi:10.48550/arXiv.2108.07258, arXiv:2108.07258 [cs].

[5] S. Li, C. Han, P. Yu, C. Edwards, M. Li, X. Wang, Y. Fung, C. Yu, J. Tetreault, E. Hovy, H. Ji, Defining a New NLP Playground, in: H. Bouamor, J. Pino, K. Bali (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2023, Association for Computational Linguistics, Singapore, 2023, pp. 11932–11951. URL: https://aclanthology.org/2023.findings-emnlp.799. doi:10.18653/v1/2023.findings-emnlp.799.

[6] J. Gao, C. Xiong, P. Bennett, N. Craswell, Neural Approaches to Conversational Information Retrieval, volume 44 of *The Information Retrieval Series*, Springer International Publishing, Cham, 2023. URL: https://link.springer.com/10.1007/978-3-031-23080-6. doi:10.1007/978-3-031-23080-6.

[7] D. Pride, M. Cancellieri, P. Knoth, Core-gpt: Combining open access research and large language models for credible, trustworthy question answering, 2023.

[8] O. Ignat, Z. Jin, A. Abzaliev, L. Biester, S. Castro, N. Deng, X. Gao, A. Gunal, J. He, A. Kazemi, M. Khalifa, N. Koh, A. Lee, S. Liu, D. J. Min, S. Mori, J. Nwatu, V. Perez-Rosas,

S. Shen, Z. Wang, W. Wu, R. Mihalcea, A PhD Student's Perspective on Research in NLP in the Era of Very Large Language Models, 2023. URL: http://arxiv.org/abs/2305.12544. doi:10.48550/arXiv.2305.12544, arXiv:2305.12544 [cs].

[9] H. M. Liana Ermakova, Sarah Bertin, J. Kamps, Overview of the clef 2023 simpletext task 3: Simplification of scientific texts (2023).

[10] S. Ateia, U. Kruschwitz, Is ChatGPT a Biomedical Expert? – Exploring the Zero-Shot Performance of Current GPT Models in Biomedical Tasks, 2023. URL: http://arxiv.org/abs/2306.16108. doi:10.48550/arXiv.2306.16108, arXiv:2306.16108 [cs].

[11] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, P. Fung, Survey of hallucination in natural language generation, ACM Computing Surveys 55 (2023) 1–38. URL: http://dx.doi.org/10.1145/3571730. doi:10.1145/3571730.

[12] G. Zuccon, B. Koopman, R. Shaik, ChatGPT Hallucinates when Attributing Answers, 2023. URL: http://arxiv.org/abs/2309.09401. doi:10.48550/arXiv.2309.09401, arXiv:2309.09401 [cs].

[13] T. Falke, L. F. R. Ribeiro, P. A. Utama, I. Dagan, I. Gurevych, Ranking Generated Summaries by Correctness: An Interesting but Challenging Application for Natural Language Inference, in: A. Korhonen, D. Traum, L. Màrquez (Eds.), Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 2214–2220. doi:10.18653/v1/P19-1213.

[14] E. Reiter, A Structured Review of the Validity of BLEU, Computational Linguistics 44 (2018) 393–401. doi:10.1162/coli_a_00322.

[15] S. Roller, E. Dinan, N. Goyal, D. Ju, M. Williamson, Y. Liu, J. Xu, M. Ott, K. Shuster, E. M. Smith, Y.-L. Boureau, J. Weston, Recipes for building an open-domain chatbot, 2020. doi:10.48550/arXiv.2004.13637. arXiv:2004.13637.

[16] M. Komeili, K. Shuster, J. Weston, Internet-Augmented Dialogue Generation, 2021. doi:10.48550/arXiv.2107.07566. arXiv:2107.07566.

[17] K. Shuster, S. Poff, M. Chen, D. Kiela, J. Weston, Retrieval Augmentation Reduces Hallucination in Conversation, 2021. doi:10.48550/arXiv.2104.07567. arXiv:2104.07567.

[18] OpenAI, J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, R. Avila, I. Babuschkin, S. Balaji, V. Balcom, P. Baltescu, H. Bao, M. Bavarian, J. Belgum, I. Bello, J. Berdine, G. Bernadett-Shapiro, C. Berner, L. Bogdonoff, O. Boiko, M. Boyd, A.-L. Brakman, G. Brockman, T. Brooks, M. Brundage, K. Button, T. Cai, R. Campbell, A. Cann, B. Carey, C. Carlson, R. Carmichael, B. Chan, C. Chang, F. Chantzis, D. Chen, S. Chen, R. Chen, J. Chen, M. Chen, B. Chess, C. Cho, C. Chu, H. W. Chung, D. Cummings, J. Currier, Y. Dai, C. Decareaux, T. Degry, N. Deutsch, D. Deville, A. Dhar, D. Dohan, S. Dowling, S. Dunning, A. Ecoffet, A. Eleti, T. Eloundou, D. Farhi, L. Fedus, N. Felix, S. P. Fishman, J. Forte, I. Fulford, L. Gao, E. Georges, C. Gibson, V. Goel, T. Gogineni, G. Goh, R. Gontijo-Lopes, J. Gordon, M. Grafstein, S. Gray, R. Greene, J. Gross, S. S. Gu, Y. Guo, C. Hallacy, J. Han, J. Harris, Y. He, M. Heaton, J. Heidecke, C. Hesse, A. Hickey, W. Hickey, P. Hoeschele, B. Houghton, K. Hsu, S. Hu, X. Hu, J. Huizinga, S. Jain, S. Jain, J. Jang, A. Jiang, R. Jiang, H. Jin, D. Jin, S. Jomoto, B. Jonn, H. Jun, T. Kaftan, Ł. Kaiser, A. Kamali, I. Kanitscheider, N. S. Keskar, T. Khan, L. Kilpatrick, J. W. Kim, C. Kim, Y. Kim, J. H. Kirchner, J. Kiros, M. Knight, D. Kokotajlo, Ł. Kondraciuk, A. Kondrich, A. Konstantinidis, K. Kosic, G. Krueger, V. Kuo, M. Lampe, I. Lan, T. Lee,

J. Leike, J. Leung, D. Levy, C. M. Li, R. Lim, M. Lin, S. Lin, M. Litwin, T. Lopez, R. Lowe, P. Lue, A. Makanju, K. Malfacini, S. Manning, T. Markov, Y. Markovski, B. Martin, K. Mayer, A. Mayne, B. McGrew, S. M. McKinney, C. McLeavey, P. McMillan, J. McNeil, D. Medina, A. Mehta, J. Menick, L. Metz, A. Mishchenko, P. Mishkin, V. Monaco, E. Morikawa, D. Mossing, T. Mu, M. Murati, O. Murk, D. Mély, A. Nair, R. Nakano, R. Nayak, A. Neelakantan, R. Ngo, H. Noh, L. Ouyang, C. O'Keefe, J. Pachocki, A. Paino, J. Palermo, A. Pantuliano, G. Parascandolo, J. Parish, E. Parparita, A. Passos, M. Pavlov, A. Peng, A. Perelman, F. d. A. B. Peres, M. Petrov, H. P. d. O. Pinto, Michael, Pokorny, M. Pokrass, V. H. Pong, T. Powell, A. Power, B. Power, E. Proehl, R. Puri, A. Radford, J. Rae, A. Ramesh, C. Raymond, F. Real, K. Rimbach, C. Ross, B. Rotsted, H. Roussez, N. Ryder, M. Saltarelli, T. Sanders, S. Santurkar, G. Sastry, H. Schmidt, D. Schnurr, J. Schulman, D. Selsam, K. Sheppard, T. Sherbakov, J. Shieh, S. Shoker, P. Shyam, S. Sidor, E. Sigler, M. Simens, J. Sitkin, K. Slama, I. Sohl, B. Sokolowsky, Y. Song, N. Staudacher, F. P. Such, N. Summers, I. Sutskever, J. Tang, N. Tezak, M. B. Thompson, P. Tillet, A. Tootoonchian, E. Tseng, P. Tuggle, N. Turley, J. Tworek, J. F. C. Uribe, A. Vallone, A. Vijayvergiya, C. Voss, C. Wainwright, J. J. Wang, A. Wang, B. Wang, J. Ward, J. Wei, C. J. Weinmann, A. Welihinda, P. Welinder, J. Weng, L. Weng, M. Wiethoff, D. Willner, C. Winter, S. Wolrich, H. Wong, L. Workman, S. Wu, J. Wu, M. Wu, K. Xiao, T. Xu, S. Yoo, K. Yu, Q. Yuan, W. Zaremba, R. Zellers, C. Zhang, M. Zhang, S. Zhao, T. Zheng, J. Zhuang, W. Zhuk, B. Zoph, GPT-4 Technical Report, 2024. doi:10.48550/arXiv.2303.08774. arXiv:2303.08774.

[19] J. Maynez, S. Narayan, B. Bohnet, R. McDonald, On Faithfulness and Factuality in Abstractive Summarization, 2020. URL: http://arxiv.org/abs/2005.00661, arXiv:2005.00661 [cs].

[20] SHROOM, a Shared-task on Hallucinations and Related Observable Overgeneration Mistakes, 2024. URL: https://helsinki-nlp.github.io/shroom/.

[21] J. Karlgren, L. D¨urlich, E. Gogoulou, L. Guillou, J. Nivre, M. Sahlgren, A. Talman, Eloquent clef shared tasks for evaluation of generative language model quality, in: ECIR'24: Advances in Information Retrieval, 46th European Conference on Information Retrieval, Lecture Notes in Computer Science, Springer, 2024.

[22] L. Ermakova, E. SanJuan, S. Huet, H. Azarbonyad, O. Augereau, J. Kamps, Overview of the clef 2023 simpletext lab: Automatic simplification of scientific texts, in: A. Arampatzis, E. Kanoulas, T. Tsikrika, S. Vrochidis, A. Giachanou, D. Li, M. Aliannejadi, M. Vlachos, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction, Springer Nature Switzerland, Cham, 2023, pp. 482–506.

[23] J. Kincaid, R. Fishburne, R. Rogers, B. Chissom, Derivation Of New Readability Formulas (Automated Readability Index, Fog Count And Flesch Reading Ease Formula) For Navy Enlisted Personnel, Institute for Simulation and Training (1975).

[24] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: A Method for Automatic Evaluation of Machine Translation, in: P. Isabelle, E. Charniak, D. Lin (Eds.), Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, 2002, pp. 311–318. doi:10.3115/1073083.1073135.

[25] E. Sulem, O. Abend, A. Rappoport, BLEU is Not Suitable for the Evaluation of Text Simplification, in: E. Riloff, D. Chiang, J. Hockenmaier, J. Tsujii (Eds.), Proceedings of

the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 738–744. doi:10.18653/v1/D18-1081.

[26] W. Xu, C. Napoles, E. Pavlick, Q. Chen, C. Callison-Burch, Optimizing Statistical Machine Translation for Text Simplification, Transactions of the Association for Computational Linguistics 4 (2016) 401–415. doi:10.1162/tacl_a_00107.

[27] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, 2019. URL: http://arxiv.org/abs/1810.04805, arXiv:1810.04805 [cs].

[28] A. Williams, N. Nangia, S. Bowman, A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 1112–1122. URL: http://aclweb.org/anthology/N18-1101. doi:10.18653/v1/N18-1101.

[29] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, BERTScore: Evaluating Text Generation with BERT, 2020. URL: http://arxiv.org/abs/1904.09675, arXiv:1904.09675 [cs].

[30] B. Goodrich, V. Rao, M. Saleh, P. J. Liu, Assessing The Factual Accuracy of Generated Text, in: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2019, pp. 166–175. URL: http://arxiv.org/abs/1905.13322. doi:10.1145/3292500.3330955, arXiv:1905.13322 [cs].