Accepté à ECIR2024: Naviguer dans l'incertitude : optimiser la dépendance à l'API des systèmes en question/réponse*

Pierre Erbacher[1,*], Louis Falissard[2], Vincent Guigue[3] and Laure Soulier[1]

[1]Sorbonne Université, ISIR, Paris
[2]Université Paris 8, Saint Denis
[3]AgroParisTech, Saclay

Abstract

Même si les grands modèles de langue (LLM) sont capables d'accumuler et de restaurer des connaissances, ils restent sujets aux hallucinations. Surtout face à des questions factuelles, les LLM ne peuvent pas s'appuyer uniquement sur les connaissances stockées dans des paramètres pour garantir des réponses véridiques et correctes. Augmenter ces modèles avec la capacité de rechercher des sources d'informations externes, telles que le Web, constitue une approche prometteuse pour augmenter la factualité des réponses. Cependant, la recherche dans une vaste collection de documents entraîne des coûts de calcul et de temps supplémentaires, ainsi que le traitement des documents retrouvés. Un comportement optimal serait d'interroger des ressources externes uniquement lorsque le LLM n'est pas sûr des réponses. Dans cet article, nous investiguons une méthode pour permettre aux LLM d'auto-estimer s'ils sont capables de répondre directement ou s'il a besoin de solliciter une base de données externe. Nous étudions une approche supervisée en introduisant un mécanisme de masquage des hallucinations dans lequel des étiquettes sont générées à l'aide d'une tâche de question/réponse. De plus, nous proposons d'exploiter des techniques d'adaptation efficaces pour entraîner notre modèle sur une petite quantité de données. Notre modèle fournit directement des réponses pour 78.2% des requêtes connues et choisit de rechercher 77.2% des requêtes inconnues. Cela a pour conséquence que l'API n'est utilisée que 62% du temps.

Keywords

Hallucination, Language Model, Search,

1. Abstract

Les modèles de langue ont démontré des performances remarquables dans un large éventail de tâches de traitement du langage naturel (TAL), notamment les agents conversationnels, le résumé, la traduction et la réponse aux questions [1, 2, 3]. Comme la mise à l'échelle de ces modèles augmente leur capacité à incorporer de plus en plus de connaissances [1, 4], au lieu de s'appuyer sur les moteurs de recherche traditionnels, Metzler et al. [5] suggère d'utiliser LLM comme base de connaissances unifiée capable de répondre aux questions ainsi que de récupérer des documents. Cependant, même les modèles les plus grands [1] sont susceptibles de produire des réponses inexactes ou fausses, communément appelées hallucinations [6]. Ceux-ci ont été largement explorés dans diverses tâches, notamment le résumé, la réponse aux questions ou

*Paper accepeted at ECIR 2024

pierre.erbacher@isir.upmc.fr (P. Erbacher)

la traduction automatique[7, 8, 9, 6, 10]. De nombreuses approches ont été proposées pour résoudre ce problème, toutes employant des techniques externes pour détecter et atténuer les hallucinations. En réponse aux questions, des méthodes de récupération augmentées telles que REALM [11], RAG [12] ou RETRO [13, 14], ont été proposées pour réduire les hallucinations des LLM. Ces approches consistent à ancrer le LLM avec un modèle de recherche d'information pour ajouter du contexte provenant d'un large corpus de documents et générer des réponses plus factuelles. Ces architectures sont efficaces, car elles améliorent à la fois le caractère factuel et réduisent les hallucinations pour des tâches spécifiques à forte intensité de connaissances telles que la réponse aux questions en domaine ouvert [14]. Cependant, les documents récupérés sont toujours pris en compte sans tenir compte de leur utilité dans la résolution de la tâche. Dans un deuxième axe de travail, des modèles, tels que LaMDA, BlenderBot, WebGPT, Toolformer [15, 16, 17, 18] sont spécifiquement entrainés pour générer une requête et s'appuyer sur un moteur de recherche lorsqu'ils sont confrontés à des questions. Bien que ces LLM aient accumulé beaucoup de connaissances au cours de la pré-formation, ils sont affinés pour s'appuyer sur des bases de données externes pour chaque question, sans tenir compte de la capacité inhérente du modèle à répondre à la question. Par exemple, Toolformer appelle l'API Web pour presque toutes les questions, (99.3%) sans réel discernement entre les questions auxquelles il peut directement répondre et le besoin réel de connaissances externes. Les LLM ont accumulé beaucoup d'informations et peuvent être capables de répondre directement lorsque confronté à des faits largement connus [1, 2, 3]. Plusieurs travaux ont proposé des méthodes pour évaluer l'incertitude des connaissances des LLM. Kadavath et at. [19], suggèrent que les LLMs sont capables d'auto-évaluer leurs connaissances en les incitant à estimer la probabilité que leur réponse prédite soit vraie.

De plus, Kuhn et al. [20] a introduit la notion d'entropie sémantique. Cette mesure inspirée de l'entropie et qui intègre des invariances linguistiques dérivées d'un modèle BART externe s'est avérée plus fiable que les estimations de vraisemblance standard pour évaluer l'incertitude du modèle. Cependant, toutes ces méthodes proposées nécessitent toujours que les modèles génèrent d'abord une réponse avant de pouvoir effectuer une estimation de l'incertitude. En conséquence, il reste difficile de savoir si un LLM peut apprendre à identifier les réponses connues des réponses inconnues avant de prédire.

Dans cet article, nous étudions une approche plus nuancée qui exploite les connaissances externes tout en intégrant les connaissances intrinsèques des LLM. Nous proposons donc un modèle qui génère soit une réponse en langage naturel, soit un appel API (par exemple $\langle search \rangle$) uniquement lorsque le modèle n'est pas sûr de lui quant à la réponse. Minimiser la dépendance aux ressources externes permet d'économiser temps d'inférence et coûts de calcul. Nous nous concentrons sur des tâches de réponse à des questions en livre fermé (CBQA) et réalisées sur deux ensembles de données (Natural Questions (NQ) [21] et TriviaQA (TQA) [22]). Nous étudions comment les LLM parviennent à auto-estimer leur capacité à répondre correctement à des questions factuelles.

Nous proposons d'apprendre à un LLM à évaluer en interne sa capacité à répondre correctement à une question donnée, sans rien utiliser d'autre que les données utilisées pour son entrainement. Le modèle résultant peut identifier directement sa capacité à répondre à une question donnée, avec des performances comparables, voire supérieures, modèles exogènes de détection des hallucinations, telles que les approches basées sur la perplexité. Notre modèle

fournit directement des réponses pour 78.2% des cas où il connait la réponse et choisit de rechercher 77.2% au lieu de répondre faux. Cela a pour conséquence que l'API n'est utilisée que 62% du temps. De plus, cette approche permet aux grands modèles de langage de conditionner leur génération à leur capacité à répondre de manière appropriée à une requête donnée, une fonctionnalité d'une importance cruciale dans l'approche Toolformer qui permet d'apprendre à rechercher uniquement en cas de besoin.

References

[1] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. 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. 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, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems, volume 33, Curran Associates, Inc., 2020, pp. 1877–1901. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

[2] S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg, H. Nori, H. Palangi, M. T. Ribeiro, Y. Zhang, Sparks of artificial general intelligence: Early experiments with gpt-4, 2023. arXiv:2303.12712.

[3] J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, E. H. Chi, T. Hashimoto, O. Vinyals, P. Liang, J. Dean, W. Fedus, Emergent abilities of large language models, 2022. arXiv:2206.07682.

[4] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, P. Schuh, K. Shi, S. Tsvyashchenko, J. Maynez, A. Rao, P. Barnes, Y. Tay, N. Shazeer, V. Prabhakaran, E. Reif, N. Du, B. Hutchinson, R. Pope, J. Bradbury, J. Austin, M. Isard, G. Gur-Ari, P. Yin, T. Duke, A. Levskaya, S. Ghemawat, S. Dev, H. Michalewski, X. Garcia, V. Misra, K. Robinson, L. Fedus, D. Zhou, D. Ippolito, D. Luan, H. Lim, B. Zoph, A. Spiridonov, R. Sepassi, D. Dohan, S. Agrawal, M. Omernick, A. M. Dai, T. S. Pillai, M. Pellat, A. Lewkowycz, E. Moreira, R. Child, O. Polozov, K. Lee, Z. Zhou, X. Wang, B. Saeta, M. Diaz, O. Firat, M. Catasta, J. Wei, K. Meier-Hellstern, D. Eck, J. Dean, S. Petrov, N. Fiedel, Palm: Scaling language modeling with pathways, 2022. arXiv:2204.02311.

[5] D. Metzler, Y. Tay, D. Bahri, M. Najork, Rethinking search: Making domain experts out of dilettantes, SIGIR Forum 55 (2021). URL: https://doi.org/10.1145/3476415.3476428. doi:10.1145/3476415.3476428.

[6] 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 Comput. Surv. 55 (2023). URL: https://doi.org/10.1145/3571730. doi:10.1145/3571730.

[7] W. Yuan, G. Neubig, P. Liu, Bartscore: Evaluating generated text as text generation, in: M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, J. W. Vaughan (Eds.), Advances in Neural Information Processing Systems, volume 34, Curran Associates, Inc.,

2021, pp. 27263–27277. URL: https://proceedings.neurips.cc/paper_files/paper/2021/file/e4d2b6e6fdeca3e60e0f1a62fee3d9dd-Paper.pdf.

[8] N. M. Guerreiro, E. Voita, A. Martins, Looking for a needle in a haystack: A comprehensive study of hallucinations in neural machine translation, in: Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, Dubrovnik, Croatia, 2023, pp. 1059–1075. URL: https://aclanthology.org/2023.eacl-main.75.

[9] P. Manakul, A. Liusie, M. J. F. Gales, Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models, 2023. arXiv:2303.08896.

[10] N. Lee, Y. Bang, A. Madotto, P. Fung, Towards few-shot fact-checking via perplexity, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Online, 2021, pp. 1971–1981. URL: https://aclanthology.org/2021.naacl-main.158. doi:10.18653/v1/2021.naacl-main.158.

[11] K. Guu, K. Lee, Z. Tung, P. Pasupat, M.-W. Chang, Realm: Retrieval-augmented language model pre-training, in: Proceedings of the 37th International Conference on Machine Learning, ICML'20, JMLR.org, 2020.

[12] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, D. Kiela, Retrieval-augmented generation for knowledge-intensive nlp tasks, in: Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS'20, Curran Associates Inc., Red Hook, NY, USA, 2020.

[13] S. Borgeaud, A. Mensch, J. Hoffmann, T. Cai, E. Rutherford, K. Millican, G. B. Van Den Driessche, J.-B. Lespiau, B. Damoc, A. Clark, D. De Las Casas, A. Guy, J. Menick, R. Ring, T. Hennigan, S. Huang, L. Maggiore, C. Jones, A. Cassirer, A. Brock, M. Paganini, G. Irving, O. Vinyals, S. Osindero, K. Simonyan, J. Rae, E. Elsen, L. Sifre, Improving language models by retrieving from trillions of tokens, in: K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, S. Sabato (Eds.), Proceedings of the 39th International Conference on Machine Learning, volume 162 of *Proceedings of Machine Learning Research*, PMLR, 2022, pp. 2206–2240. URL: https://proceedings.mlr.press/v162/borgeaud22a.html.

[14] K. Lee, M.-W. Chang, K. Toutanova, Latent retrieval for weakly supervised open domain question answering, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 6086–6096. URL: https://aclanthology.org/P19-1612. doi:10.18653/v1/P19-1612.

[15] R. Thoppilan, D. D. Freitas, J. Hall, N. Shazeer, A. Kulshreshtha, H.-T. Cheng, A. Jin, T. Bos, L. Baker, Y. Du, Y. Li, H. Lee, H. S. Zheng, A. Ghafouri, M. Menegali, Y. Huang, M. Krikun, D. Lepikhin, J. Qin, D. Chen, Y. Xu, Z. Chen, A. Roberts, M. Bosma, V. Zhao, Y. Zhou, C.-C. Chang, I. Krivokon, W. Rusch, M. Pickett, P. Srinivasan, L. Man, K. Meier-Hellstern, M. R. Morris, T. Doshi, R. D. Santos, T. Duke, J. Soraker, B. Zevenbergen, V. Prabhakaran, M. Diaz, B. Hutchinson, K. Olson, A. Molina, E. Hoffman-John, J. Lee, L. Aroyo, R. Rajakumar, A. Butryna, M. Lamm, V. Kuzmina, J. Fenton, A. Cohen, R. Bernstein, R. Kurzweil, B. Aguera-Arcas, C. Cui, M. Croak, E. Chi, Q. Le, Lamda: Language models for dialog applications, 2022. arXiv:2201.08239.

[16] R. Nakano, J. Hilton, S. Balaji, J. Wu, L. Ouyang, C. Kim, C. Hesse, S. Jain, V. Kosaraju,

W. Saunders, X. Jiang, K. Cobbe, T. Eloundou, G. Krueger, K. Button, M. Knight, B. Chess, J. Schulman, Webgpt: Browser-assisted question-answering with human feedback, 2022. arXiv:2112.09332.

[17] K. Shuster, J. Xu, M. Komeili, D. Ju, E. M. Smith, S. Roller, M. Ung, M. Chen, K. Arora, J. Lane, M. Behrooz, W. Ngan, S. Poff, N. Goyal, A. Szlam, Y.-L. Boureau, M. Kambadur, J. Weston, Blenderbot 3: a deployed conversational agent that continually learns to responsibly engage, 2022. arXiv:2208.03188.

[18] T. Schick, J. Dwivedi-Yu, R. Dessì, R. Raileanu, M. Lomeli, L. Zettlemoyer, N. Cancedda, T. Scialom, Toolformer: Language models can teach themselves to use tools, 2023. arXiv:2302.04761.

[19] S. Kadavath, T. Conerly, A. Askell, T. Henighan, D. Drain, E. Perez, N. Schiefer, Z. Hatfield-Dodds, N. DasSarma, E. Tran-Johnson, S. Johnston, S. El-Showk, A. Jones, N. Elhage, T. Hume, A. Chen, Y. Bai, S. Bowman, S. Fort, D. Ganguli, D. Hernandez, J. Jacobson, J. Kernion, S. Kravec, L. Lovitt, K. Ndousse, C. Olsson, S. Ringer, D. Amodei, T. Brown, J. Clark, N. Joseph, B. Mann, S. McCandlish, C. Olah, J. Kaplan, Language models (mostly) know what they know, 2022. arXiv:2207.05221.

[20] L. Kuhn, Y. Gal, S. Farquhar, Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation, 2023. arXiv:2302.09664.

[21] T. Kwiatkowski, J. Palomaki, O. Redfield, M. Collins, A. Parikh, C. Alberti, D. Epstein, I. Polosukhin, J. Devlin, K. Lee, K. Toutanova, L. Jones, M. Kelcey, M.-W. Chang, A. M. Dai, J. Uszkoreit, Q. Le, S. Petrov, Natural questions: A benchmark for question answering research, Transactions of the Association for Computational Linguistics 7 (2019) 452–466. URL: https://aclanthology.org/Q19-1026. doi:10.1162/tacl_a_00276.

[22] M. Joshi, E. Choi, D. Weld, L. Zettlemoyer, TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 1601–1611. URL: https://aclanthology.org/P17-1147. doi:10.18653/v1/P17-1147.