A Retrieval-Augmented Generation Strategy to Enhance Medical Chatbot Reliability

Saba Ghanbari Haez[1,2], Marina Segala[1], Patrizio Bellan[1], Simone Magnolini[1], Leonardo Sanna[1(✉)], Monica Consolandi[1], and Mauro Dragoni[1]

[1] Fondazione Bruno Kessler, Trento, Italy
{sghanbarihaez,msegala,pbellan,magnolini,lsanna,mconsolandi,dragoni}@fbk.eu
[2] Free University of Bozen, Bolzano, Italy

Abstract. The advent of Large Language Models opened new perspectives concerning their usage within the digital health domain. However, their intrinsic probabilistic and unpredictable behavior needs the design of trustworthy strategies aiming to avoid the creation of hallucinations that, especially within the digital health domain, may lead to severe harm. Such an issue has been addressed with the adoption of Retrieval-Augmented Generation solutions, where the text generation task is supported by controlled knowledge injected into the prompts. Even if the hallucination issue is mitigated, the generation of certified information (such as trustworthy content granted by the system's owner) requires more sophisticated strategies. In this work, we propose an approach where the classic Retrieval-Augmented Generation pipeline is enhanced with a further initial step where the Large Language Model is asked to generate a preliminary text used to query the repository of certified information for presenting the appropriate content to the final user.

1 Introduction

Large Language Models (LLMs) such as BERT [3] and T5 [18] possess the ability to generate factual information based on learned patterns from extensive training data [16]. However, their accuracy without external sources may vary due to several factors like data quality, task complexity, and parameter density. Therefore, they may generate inaccurate or fictional content, i.e., hallucinations [23,27]. Recent efforts aim to tackle these challenges by augmenting external knowledge to empower LLMs to interact effectively with users and their surroundings.

Retrieval-Augmented Generation (RAG) [11], explicitly incorporates external knowledge into LLMs' prompts to contribute to the enhancement of their trustworthiness [6]. This involves retrieving documents relevant to the user's query and subsequently generating a comprehensive response considering the contained factual information. The efficiency of RAG systems relies on sufficient and diverse training data, with the risk of observing a low accuracy if

J. Finkelstein et al. (Eds.): AIME 2024, LNAI 14844, pp. 213–223, 2024.
https://doi.org/10.1007/978-3-031-66538-7_22

the retrieval system lacks robustness and reliability. In particular, the *retrieval* phase struggles with issues like semantic ambiguity coupled with basic matching techniques that lead to inaccurate data retrieval. While, the *augmentation* and *generation* phases struggle with integrating context and coherence, resulting in superficial responses that fail to meet sophisticated query demands[1].

Previous research predominantly employs traditional RAG pipelines, which involve retrieving documents relevant to user input from a non-certified database before sending the user query to the LLM. The novel contribution of this paper lies in (i) using a certified document repository to inject factual knowledge into the LLM; and, (ii) proposing a novel enhancement to the classic RAG framework by introducing a further preliminary interaction with the integrated LLM[2].

We present the theoretical framework and discuss its preliminary adoption within an FAQ-based chatbot designed to support pregnant women followed by the Trentino Healthcare Department in Italy, i.e., the *TreC-Mamma* application[3]. Our goal is to answer the following Research Questions: (RQ1) How do the LLMs be integrated effectively into digital health solutions? (RQ2) Can LLMs generate *certified* content (the meaning of the term "certified" is explained in Sect. 3)? (RQ3) Can a RAG strategy be enhanced to solve semantic ambiguity and avoid hallucinations?

Through the empirical evaluation discussed in Sect. 4, we demonstrate the effectiveness of the proposed RAG-based strategy in addressing the limitations of LLMs and enhancing the overall user experience in domains targeted by the current work, i.e., maternal health. Moreover, we suggest that our methodology may represent a promising direction for leveraging advanced language technologies to tackle the certified information challenges effectively.

2 Related Work

In recent years, there has been an increase in research efforts to improve the credibility and effectiveness of LLMs, particularly in domains that prioritize accuracy and reliability, such as modern medicine and digital health [17]. Conversational Artificial Intelligence (AI) in healthcare suffers several challenges that are crucial to address, including the lack of suitable evaluation metrics, concerns regarding fairness, bias, and hallucination in chatbot responses, the balance between personalization and oversimplification, and obstacles in implementation [1]. In this context, RAG demonstrated to be a suitable candidate for mitigating hallucination issues, enriching factual content generation in LLMs, and integrating external knowledge sources. We believe these advancements can enhance the utility of conversational AI in healthcare settings.

RAG has recently gained attention for its explicit incorporation of external knowledge into LLMs' prompts. By leveraging Information Retrieval (IR) techniques, RAG aims to enhance the credibility and relevance of generated content

[1] https://arxiv.org/abs/2401.05856.
[2] In this work, we adopted the GPT4 LLM.
[3] https://trentinosalutedigitale.com/blog/portfolio/trec-mamma/.

by retrieving documents pertinent to user queries. RAG utilizes both parametric and non-parametric memory, drawing upon pre-trained seq2seq models [11] and Dense Passage Retrieval (DPR) [10]. This approach has led to surpassing state-of-the-art performance on tasks like QA and summarization and has been noted for enhancing language generation by producing more specific, diverse, and factual language compared to parametric-only seq2seq models [22]. Nevertheless, while RAG has shown promise in improving the output of LLMs, it faces constraints when confronted with data outside its training set. Several approaches have been explored to address this issue.

Building upon the works by [10,11], Retrieval-Augmented Language Model pre-training (REALM) [6] integrates a knowledge retrieval mechanism to enhance neural language models' performance in question-answering tasks, demonstrating superior accuracy and interoperability. A two-stage Approach proposed by [8] combines DPR with generative sequence-to-sequence language models, leveraging the strengths of both approaches to provide comprehensive and contextually relevant responses for open-domain QA tasks.

I-RetGen (Iterative Retrieval-Generation) [21] iteratively integrates retrieval and generation processes, enhancing relevance for complex queries while minimizing overhead by utilizing the LLM's generation output to guide retrieval. RAGE (Retrieval-Augmented Generation with Rich Answer Encoding) [7] combines retrieval and generation techniques to produce informative and coherent answers, enhancing the richness and relevance of generated content. FLARE (Active Retrieval-Augmented Generation) [9] dynamically decides when and what to retrieve throughout the generation process, offering a proactive approach to content augmentation.

Recent advancements have further expanded the capabilities of retrieval-augmented generation, including techniques such as Augmentation-Adapted Retriever (AAR) [28] and Knowledge-Augmented Language Model Verification (KALMV) [2]. These approaches aim to enhance language model accuracy across different domains by integrating external knowledge and detecting errors in both knowledge retrieval and text generation processes. Additionally, frameworks like Induction-Augmented Generation (IAG) [29] and domain adaptation techniques for RAG models [22] demonstrate ongoing efforts to improve implicit reasoning and adaptability in question-answering tasks.

Despite the diversity of approaches, prior studies have concentrated on traditional RAG pipelines, which focus exclusively on enhancing text generation. Our proposed approach differs significantly by involving instead the use of a certified repository as well as the injection of knowledge directly into the LLM. By reconfiguring the RAG process and integrating retrieved information into the answer-generation pipeline, we aim not only to mitigate hallucination issues but also to ensure the generation of certified and contextually informed responses. This novel methodology represents a significant advancement in addressing the challenges of factual content generation within LLM frameworks, particularly in sensitive domains like digital health. Furthermore, our approach holds promise for improving user trust and satisfaction in automated systems by providing

reliable and contextually relevant information. Additionally, our empirical evaluation and comparative analysis demonstrate the effectiveness of our method in enhancing the user experience, particularly in domains such as maternal nutrition and health, as evidenced by our preliminary adoption of the TreC Mamma application.

3 Method

This Section presents a preliminary discussion about how RAG may be a suitable strategy to mitigate the hallucination issues affecting LLMs followed by a description of how we implemented our strategy.

3.1 Preliminaries

As introduced in Sect. 1, the variability of LLMs' responses poses a substantial issue when operating in contexts where certified information is needed. In this work the term "*certified information*" refers to text created or verified by healthcare professionals, ensuring it aligns with current scientific knowledge in the specific domain. To preserve the nature of "*certified information*", it is essential that the content remains unchanged in its textual form. Moreover, it should be semantically predetermined, i.e., each specific question consistently corresponds to a particular set of semantic equivalent answers. Using LLMs, even with RAG, cannot guarantee this requirement.

Indeed, a standard RAG pipeline in a FAQ-based chatbot would employ the user's question to query the certified documents. Usually, RAG converts a user query into a vector embedding representation, which is then employed to evaluate semantic similarity across the document repository. Yet, there can be significant differences between the vector representations of the query and the documents within the semantic space. This disparity poses a notable limitation as it could result in relevant documents being overlooked during retrieval.

Moreover, RAG provides additional opacity to the algorithm since the user does not know which information is being used to provide the answer. Although our conversational agent is not directly involved in diagnostic processes, some ethical questions about possible bias of LLMs remain valid [24] as well as the need to build a system as much as possible transparent and accountable [25]. Given the intrinsic opacity of LLMs, it is nonetheless our effort to pursue "*explicability*", i.e., the new ethical principle that Floridi et al. [4] introduced alongside the traditional ones (beneficence, non-maleficence, autonomy, and justice). This ethical principle is already largely used in the field of explainable AI [20]. Hence, in our chatbot, we want to provide the sources used in the RAG process to ensure transparency at the epistemic level [26]. This strategy aims to prevent possible trust issues towards the agent [14] or at least to mitigate them.

3.2 Implementation

The solution we propose in this paper is summarized in Fig. 1.

Our goal is to address the limitations mentioned above with a modular approach aiming to enhance the classic RAG pipeline with the integration of the Hypothetical Document Embeddings (HyDE) framework [5]. The HyDE introduces a further step at the beginning of the pipeline where the LLM is asked to produce a hypothetical document (HyDoc) based only on the input query provided by the user. The hypothetical document represents the query's information request and is also meant to capture relevant textual patterns that might be present in the certified repository connected to the pipeline. It is important to mention that, at this stage, the output generated by the LLM model might contain hallucinations since any check is performed. However, the HyDoc generated should lie in the semantic space in a neighborhood of similar real documents that contain the correct and certified answer to provide to the user.

Thus, the main idea is to use the HyDoc generated by the LLM to augment the initial query. To do that, we need to transform the HyDoc into a semantic vector, namely a sentence embedding. Creating an embedding representation of a sentence is challenging because the meaning of each word needs to be contextualized concerning the other words in the sentence. Anyhow, many LLMs are trained to predict the next word in a sequence, so the embedding representation of a sentence cannot be easily extracted. Indeed, unlike universal word embeddings methods, e.g., word2vec [15], a widely accepted general-purpose sentence embedding technique is still a very active research field [13].

In our work, we integrated the *paraphrase-multilingual-mpnet-base-v2* Bi-Encoder model [19] and we used it to both create the embeddings of our HyDoc

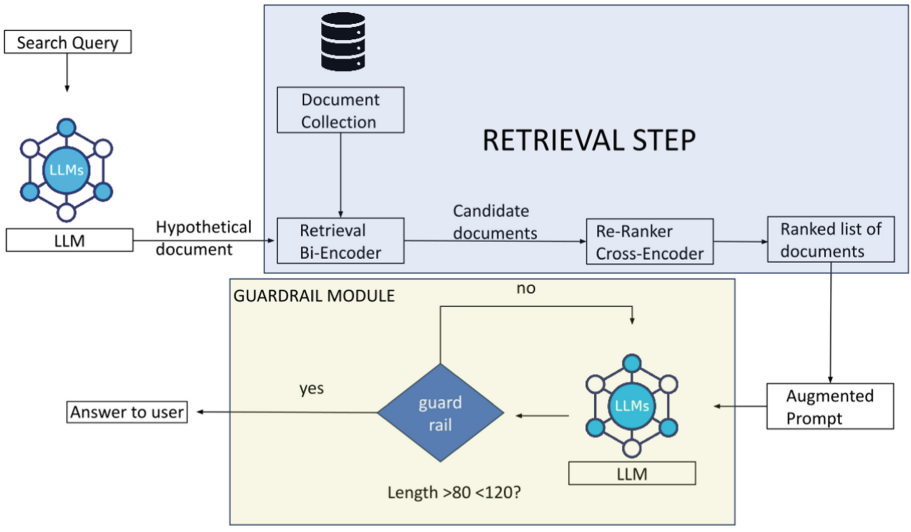

Fig. 1. Our RAG pipeline

and the vector-based representation of the whole repository. The model works by adding a pooling operation, which generates a fixed-size embedding representation normalized to have a total length of 1.00. Such a representation eases the comparison of each vector by adopting cosine similarity. After calculating the cosine similarity between all pairs of type $<HyDoc, D_i>$ with $i = 1...n$, where n is the number of documents contained in the certified repository and D_i the i_{th} document contained in the certified repository, we ranked the documents and selected the $k = 50$ most similar ones. The rationale behind the choice of $k = 50$ is that from an information retrieval perspective, this is an acceptable number of documents that may grant the inclusion of the most relevant ones [12].

The Bi-Encoder is a computationally efficient method for semantic search, but it works well only when we have documents of comparable lengths. Hence, if the HyDoc is either significantly shorter or longer than the documents to retrieve, the risk of retrieving non-relevant documents is considerable.

For this reason, we integrated a Cross-Encoder module to re-rank the list of retrieved documents. This post-retrieval operation ensures the selection of the most informative documents. For this task, we use the *ms-marco-MiniLM-L-6-v2*[4] cross encoder. The Cross-Encoder is thus more accurate than the Bi-Encoder although it is computationally more expensive. For this reason, we applied it only to the list of candidate relevant documents to reduce the overall computation time of each user's request aiming to preserve the usability of the system.

Once the re-ranking process is completed, we select the top $j = 3$ documents that are most similar to the HyDoc, according to the Cross-Encoder output. These three documents are then used to augment the original prompt and retrieve the textual part of the final answer sent to the user. Here, a *Guard-Rail* module[5] is applied to ensure that the reply generated by the LLM satisfies the length requested through the prompt. The response of the agent will therefore contain the generated text as well as the pointers to the original certified sources, i.e., the three selected documents, used to generate the answer.

4 Evaluation and Discussion

For this study, we have curated a document repository certified by the Trentino Healthcare Department. This dataset forms the backbone of the strategy we proposed in this work, by ensuring that the information fed into the integrated LLM is both accurate and authoritative. The primary source of our dataset comes from the Obstetrician Department of the Hospital of Trento. This includes a comprehensive collection of 1512 documents split as follows. A set of 179 informative cards associated with each pregnancy week offering information pertinent to maternal health, pregnancy, and fetus status. These cards have been written and certified by healthcare professionals, providing a reliable foundation for our model. In addition to the informative cards, this set contains additional content

[4] https://huggingface.co/cross-encoder/ms-marco-MiniLM-L-6-v2.

[5] For a detailed description of how this strategy works see the paper of Mangaokar et al. https://arxiv.org/abs/2402.15911.

extracted from videos released by the same department. Then, documents from two certified repositories: UPPA (953 documents) and ISS-Salute (380 documents) have been incorporated. Both sources have been considered *certified* by our healthcare department for their evidence-based approach to parenting and child health. The inclusion of both repositories introduces a broader perspective on child care, complementing the information provided by the Obstetrician Department of the Hospital of Trento. This amalgamation of resources from both the hospital and repositories ensures a comprehensive and well-rounded dataset that covers the spectrum of maternal and child health.

To address the research questions provided in Sect. 1, our system combines the generative capabilities of LLMs with a retrieval mechanism performed on the repository of certified documents. This hybrid approach ensures that the generated responses are not only linguistically coherent and contextually relevant but also grounded in verified medical knowledge.

To rigorously assess the performance of our solutions, we devised a comprehensive evaluation strategy that encompasses both the accuracy of the answers provided and the quality of the supporting documents retrieved from the certified repository. This dual-focus evaluation is crucial for ensuring that the solution delivers precise information and enriches its responses with credible and authoritative sources, thereby enhancing the trustworthiness and reliability of the system.

The evaluation task consisted of a set of 100 questions, which were considered representative by the experts involved in the evaluation process and that represent typical inquiries made by new mothers regarding pregnancy and early childcare. These questions were designed to cover a broad spectrum of topics within the domain, ensuring a thorough evaluation of the system's capabilities. The questions were then presented to a group of five test users, who interacted with the TreC Mamma application and collected the answers. The users were instructed to evaluate the responses based on seven criteria: (i) the relevance of the answer to the question, (ii) the relevance of the links (documents) provided, (iii) text quality, (iv) reliability, (v) clarity, (vi) completeness, and (vii) an overall evaluation score. These metrics were chosen to provide a holistic view of the system's performance, encompassing both the quality of the generated text and the relevance and certified documents provided within the answers.

The results of the evaluation are summarized in Table 1, which presents the average scores across all test users for each evaluation criterion. The first column contains the name of each metric; the second column contains the average score computed by considering the judgments provided by each user on all questions; the third and fourth columns contain the highest and the lowest scores obtained for that metric, respectively; and, the fifth column contains the variance. The metric [M1] involves binary classification (relevant, not relevant), and its score is interpretable as a percentage. The metric [M2] is a three-way classification (on-topic, partially on-topic, off-topic) and can similarly be interpreted as a percentage. The metrics from [M3] to [M7] employ a 5-level evaluation scale ranging from 1 (insufficient) to 5 (great).

Table 1. Summary of the results provided by the test users.

Evaluation Criterion	Avg	Max	Min	Var
[M1] Is the answer relevant to the question?	0.93	1.00	0.50	0.02
[M2] Links relevance	0.44	1.00	0.00	0.05
[M3] Text quality	4.59	5.00	3.33	0.06
[M4] Reliability	3.79	4.75	2.33	0.40
[M5] Clarity	4.60	5.00	3.33	0.05
[M6] Completeness	3.38	4.75	1.33	0.81
[M7] Overall evaluation	3.40	4.75	1.67	0.59

The high relevance score ([M1]), i.e., 0.93, indicates that the chatbot is highly effective in providing answers that are pertinent to the users' questions. However, the relevance of the links provided ([M2]), i.e., 0.44 suggests that there is significant room for improvement in the selection and presentation of supporting documents. The text quality, clarity, and reliability scores are relatively high, demonstrating the system's ability to generate well-written, clear, and somewhat reliable responses. Completeness and overall evaluation scores, while above average, highlight areas where further enhancements could be made to improve user satisfaction and the comprehensiveness of the answers provided. By considering the variance values, we may observe how the criteria [M4], [M6], and [M7] required further investigations. A preliminary further analysis revealed how, for some of the queries contained within the test set, the final output produced by the LLM did not satisfy the expectations of the evaluators.

As a final consideration, we state that we can positively answer the three research questions presented in Sect. 1. We can positively answer **RQ1** since the average score observed for the criterion [M7] proves an effective behavior of the proposed solution within the digital health domain. We can positively answer to **RQ2** as well, since the high values obtained for metrics [M1], [M3], and [M4] demonstrate how the content generated by the LLM can be considered certified. Finally, we can also positively answer to **RQ3** given the high values obtained for metrics [M2], [M4], [M5], and [M6] showing how, on average, the content of the final text sent to the evaluators has been considered correct, i.e., no hallucinations were included. The only point of attention related to the metric [M2] whose value demonstrates that there is still room for improvement, even if, on average, half of the documents included in the links sent to the users have been considered fully relevant.

5 Conclusions

In this work, we presented a framework showing how the RAG pipeline can be enhanced by introducing a further interaction with the integrated LLM before the retrieval step to support scenarios where the answer provided to users must

contain only certified information. We tested our approach in the context of the local project TreC-Mamma, promoted by the Trentino Healthcare Department, which includes a mobile application used by pregnant women with an FAQ facility. Preliminary results demonstrated the suitability of the proposed strategy and paved the way for further steps in this research direction and future implementation in other domains. In particular, the effort will focus on the main limitation we observed, i.e., the retrieval module. Such a module is in charge of retrieving the certified information and, in the current setting, registered the lowest score compared with the other criteria adopted to evaluate the performance of the proposed solution. Finally, we intend to explore the integration of open LLMs that may represent a strong requirement concerning the deployment of this type of solution into production environments.

Acknowledgments. We acknowledge the support provided by the PNRR initiatives: INEST (Interconnected North-East Innovation Ecosystem), project code ECS00000043, and FAIR (Future AI Research), project code PE00000013. These projects are part of the NRRP MUR program, funded by the NextGenerationEU. This paper is supported by the TrustAlert project, funded by Fondazione Compagnia San Paolo and Fondazione CDP under the "Artificial Intelligence" call.

References

1. Abbasian, M., et al.: Foundation metrics for evaluating effectiveness of healthcare conversations powered by generative AI. NPJ Digit. Med. **7**(1), 82 (2024). https://doi.org/10.1038/s41746-024-01074-z
2. Baek, J., Jeong, S., Kang, M., Park, J., Hwang, S.: Knowledge-augmented language model verification. In: Bouamor, H., Pino, J., Bali, K. (eds.) Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pp. 1720–1736. Association for Computational Linguistics, Singapore (2023). https://doi.org/10.18653/v1/2023.emnlp-main.107, https://aclanthology.org/2023.emnlp-main.107
3. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Burstein, J., Doran, C., Solorio, T. (eds.) Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (2019). https://doi.org/10.18653/v1/N19-1423, https://aclanthology.org/N19-1423
4. Floridi, L., et al.: AI4people-an ethical framework for a good AI society: opportunities, risks, principles, and recommendations. Mind. Mach. **28**, 689–707 (2018)
5. Gao, L., Ma, X., Lin, J., Callan, J.: Precise zero-shot dense retrieval without relevance labels. In: Rogers, A., Boyd-Graber, J.L., Okazaki, N. (eds.) Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023. pp. 1762–1777. Association for Computational Linguistics (2023). https://doi.org/10.18653/V1/2023.ACL-LONG.99
6. Guu, K., Lee, K., Tung, Z., Pasupat, P., Chang, M.: Retrieval augmented language model pre-training. In: III, H.D., Singh, A. (eds.) Proceedings of the 37th

International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 119, pp. 3929–3938. PMLR (13–18 Jul 2020). https://proceedings.mlr.press/v119/guu20a.html
7. Huang, W., Lapata, M., Vougiouklis, P., Papasarantopoulos, N., Pan, J.Z.: Retrieval augmented generation with rich answer encoding. In: Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1012–1025. Association for Computational Linguistics (2023)
8. Izacard, G., Grave, E.: Leveraging passage retrieval with generative models for open domain question answering. In: Merlo, P., Tiedemann, J., Tsarfaty, R. (eds.) Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. pp. 874–880. Association for Computational Linguistics, Online (2021). https://doi.org/10.18653/v1/2021.eacl-main.74, https://aclanthology.org/2021.eacl-main.74
9. Jiang, Z., et al.: Active retrieval augmented generation. In: Bouamor, H., Pino, J., Bali, K. (eds.) Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pp. 7969–7992. Association for Computational Linguistics, Singapore (2023). https://doi.org/10.18653/v1/2023.emnlp-main.495, https://aclanthology.org/2023.emnlp-main.495
10. Karpukhin, V., et al.: Dense passage retrieval for open-domain question answering. In: Webber, B., Cohn, T., He, Y., Liu, Y. (eds.) Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 6769–6781. Association for Computational Linguistics (2020). https://doi.org/10.18653/v1/2020.emnlp-main.550, https://aclanthology.org/2020.emnlp-main.550
11. Lewis, P., et al.: Retrieval-augmented generation for knowledge-intensive nlp tasks: Adv. Neural. Inf. Process. Syst. **33**, 9459–9474 (2020)
12. Li, H.: Learning to rank for information retrieval and natural language processing. Springer Nature (2022)
13. Li, R., Zhao, X., Moens, M.: A brief overview of universal sentence representation methods: a linguistic view. ACM Comput. Surv. **55**(3), 1–42 (2023). https://doi.org/10.1145/3482853
14. Martens, M., De Wolf, R., De Marez, L.: Trust in algorithmic decision-making systems in health: a comparison between ADA health and IBM Watson oncology. Cyberpsychology **18**(1) (2024)
15. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. Adv. Neural Inform. Proc. Syst. **26** (2013)
16. Petroni, F., et al.: Language models as knowledge bases? (2019) arXiv preprint arXiv:1909.01066
17. Pham, K.T., Nabizadeh, A., Selek, S.: Artificial intelligence and chatbots in psychiatry. Psychiatr. Q. **93**, 249–253 (2022). https://doi.org/10.1007/s11126-022-09973-8received 26 September 2021, Revised 23 January 2022, Accepted 26 January 2022, Published 25 February 2022
18. Raffel, C., et al.: Exploring the limits of transfer learning with a unified text-to-text transformer. J. Mach. Learn. Res. **21**(1), 5485–5551 (2020)
19. Reimers, N., Gurevych, I.: Sentence-bert: Sentence embeddings using siamese bert-networks. In: Inui, K., Jiang, J., Ng, V., Wan, X. (eds.) Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP

2019, Hong Kong, China, November 3-7, 2019. pp. 3980–3990. Association for Computational Linguistics (2019). https://doi.org/10.18653/V1/D19-1410
20. Saeed, W., Omlin, C.: Explainable AI (XAI): a systematic meta-survey of current challenges and future opportunities. Knowl.-Based Syst. **263**, 110273 (2023)
21. Shao, Z., Gong, Y., Shen, Y., Huang, M., Duan, N., Chen, W.: Enhancing retrieval-augmented large language models with iterative retrieval-generation synergy. In: Bouamor, H., Pino, J., Bali, K. (eds.) Findings of the Association for Computational Linguistics: EMNLP 2023, pp. 9248–9274. Association for Computational Linguistics, Singapore (2023). https://doi.org/10.18653/v1/2023.findings-emnlp.620, https://aclanthology.org/2023.findings-emnlp.620
22. Siriwardhana, S., Weerasekera, R., Wen, E., Kaluarachchi, T., Rana, R., Nanayakkara, S.: Improving the domain adaptation of retrieval augmented generation (RAG) models for open domain question answering. Trans. Assoc. Comput. Linguist. **11**, 1–17 (2023). https://doi.org/10.1162/tacl_a_00530, https://aclanthology.org/2023.tacl-1.1
23. Wang, B., et al.: Decodingtrust: A comprehensive assessment of trustworthiness in GPT models. In: Advances in Neural Information Processing Systems (NeurIPS) (2023)
24. Wang, C., Liu, S., Yang, H., Guo, J., Wu, Y., Liu, J.: Ethical considerations of using CHATGPT in health care. J. Med. Internet Res. **25**, e48009 (2023)
25. Williams, R., et al.: From transparency to accountability of intelligent systems: moving beyond aspirations. Data Policy **4**, e7 (2022). https://doi.org/10.1017/dap.2021.37
26. Winter, P.D., Carusi, A.: (De) troubling transparency: artificial intelligence (AI) for clinical applications. Med. Humanit. **49**(1), 17–26 (2023)
27. Xu, Y., Namazifar, M., Hazarika, D., Padmakumar, A., Liu, Y., Hakkani-Tur, D.: KILM: Knowledge injection into encoder-decoder language models. In: Rogers, A., Boyd-Graber, J., Okazaki, N. (eds.) Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 5013–5035. Association for Computational Linguistics, Toronto, Canada (2023). https://doi.org/10.18653/v1/2023.acl-long.275, https://aclanthology.org/2023.acl-long.275
28. Yu, Z., Xiong, C., Yu, S., Liu, Z.: Augmentation-adapted retriever improves generalization of language models as generic plug-in. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, Volume 1: Long Papers, pp. 2421–2436. Association for Computational Linguistics (July 9–14 2023)
29. Zhang, Z., et al.: Iag: Induction-augmented generation framework for answering reasoning questions. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. pp. 1–14, Association for Computational Linguistics (Dec 6–10 2023)