Bridging the Gap: Externalizing Knowledge for Generative Search in Domain-Specific Contexts

Samy Ateia[1]

[1]University of Regensburg, Universitätsstraße 31, 93053 Regensburg, Germany

Abstract

Generative search is an emerging and exciting topic in information retrieval. This position paper investigates challenges and potential solutions related to generative search engines. It explores the limitations these engines face when retrieving and summarizing information, focusing particularly on the constraints and needs in enterprise contexts. By integrating structured semantic representations, such as knowledge graphs, with large language models (LLMs), we suggest externalizing knowledge, potentially improving the effectiveness, efficiency, and transparency of generative search engines, especially in domain-specific enterprise settings.

Keywords

generative search, information retrieval, large language models, knowledge representation

1. Introduction

The fields of information retrieval and natural language processing (NLP) have seen a surge in the development and application of large language models (LLMs). The popular rise of ChatGPT showcased the potential of LLMs, but also exposed their limitations. Notably, these models often generate incorrect information or fabricate facts that are absent in the input or their training data, a phenomenon referred to as 'hallucination' in natural language generation [1]. Additionally, LLMs cannot answer questions about events or facts occurring after the cutoff of their training data.

These limitations led to the development of commercial generative search systems[1] such as Bing Chat[2], Google Bard[3], perplexity.ai[4] and the Browsing Plugin for ChatGPT[5]. These systems can retrieve and present current web-based information by first conducting several web searches, then distilling the results, integrating this information into a prompt sent to the LLM that finally generates an answer grounded in the provided information.

These systems are known as generative search engines [2] and the process can be viewed as a subfield of retrieval augmented generation (RAG) [3], where the output of a generative

Samy.Ateia@stud.uni-regensburg.de (S. Ateia)

[1]Archived links from web.archive.org are used to ensure the preservation and availability of online sources in their state at the time of citation.

[2]http://web.archive.org/web/20230813181603/https://blogs.microsoft.com/blog/2023/02/07/reinventing-search-with-a-new-ai-powered-microsoft-bing-and-edge-your-copilot-for-the-web/

[3]https://web.archive.org/web/20230814035215/https://blog.google/technology/ai/bard-google-ai-search-updates/

[4]https://web.archive.org/web/20230810043441/https://perplexity-ai.notion.site/Perplexity-FAQs-0a9141bdd1b94c76b955e892f73451ff

[5]https://web.archive.org/web/20230811205004/https://openai.com/blog/chatgpt-plugins

model is displayed directly to answer the information need of a user. Compared to traditional web search, the user is no longer presented with a list of ranked search results, but instead a coherent generated text that tries to answer the user's query and is often augmented with footnotes or citations to the retrieved sources.

Even though generative search engines can access current and potentially accurate information, these systems still face challenges in efficiently retrieving and summarizing information. These issues stem from the limited context windows of generative LLMs and the potential to feed the LLM with incorrect or inadequate information. This position paper underscores the need for strategies to improve the performance and reliability of these systems, specifically within enterprise contexts, and puts forth promising research directions that could lead to effective solutions for these current limitations.

2. Limitations of Generative Search Engines

2.1. Hallucinations and Context Size

The introduction of Bing Chat and Bard marked significant advancements in mitigating the limitations of the initial ChatGPT, such as hallucinations and the lack of current information. Both can draw on current information from the web to answer user queries and ground their responses. However, these newer systems still face their own unique challenges, including occasional hallucinations and problems related to unsupported statements and inaccurate citations [2]. One possible explanation for these issues is that the underlying generative models are not prompted with sufficient information to answer the user query. All the currently used GPT models in the above-mentioned systems have a limited context window, which is used to prime their next token prediction. This requires the generative search engine to pre-select relevant search results and paragraphs which are fed to the model. This retrieval and selection step can miss relevant information that should have been supplied to the model, or feed it misleading information.

Although it is not publicly disclosed what context window Bing Chat uses, the underlying GPT-4 model comes in two versions with either an 8000 or a 32000 token context window [4]. This context window must be shared between the prompt and the generated output. Approximately 100 tokens correspond to roughly 75 words[6]. As one page typically contains between 250 and 500 words, we estimate that the context fed into the 8000-token model could span 9–18 pages, assuming around 2000 tokens are reserved for generating the answer. This constraint considerably limits the amount of information from web pages or other documents that can be provided to the model as grounding information before it generates its answer.

This limitation is particularly challenging for open-domain web search, as the underlying web search might yield ambiguous or irrelevant results, or the system may summarize the results in a way that excludes crucial information.

Even when all the required information to answer a question correctly is supplied to the language model generating the answer, and the model is instructed to only use the information

[6] http://web.archive.org/web/20230519002451/https://help.openai.com/en/articles/4936856-what-are-tokens-and-how-to-count-them

provided in the context, it might still generate additional, unfounded information. OpenAI refers to this phenomenon as *closed domain hallucinations* [4]. In their technical report, OpenAI indicated that GPT-4 scored 29% higher in avoiding closed domain hallucinations than their previous GPT-3.5 model in their internal evaluations, but they did not provide a specific figure for how prevalent this problem is [4].

2.2. Challenges in Domain-Specific Enterprise Use Cases

LLMs pretrained on vast amounts of general knowledge have been shown to exhibit a domain gap when applied in domain-specific contexts such as biomedical texts compared to models pretrained with documents from this domain [5]. This problem also affects enterprise use-cases where internal abbreviations, documents, and concepts are not present in the pre-training of publicly available LLMs. One strategy to overcome this is to fine-tune or even pre-train models from scratch on internal and domain-specific data, which comes with significant costs [6].

Enterprise use cases also often have additional requirements regarding transparency and explainability of system outputs [7]. These are currently not solvable with just using an LLM directly that was pre-trained or fine-tuned on internal data.

The generative search approach, where all relevant information is collected first and added to the human-readable prompt that is sent to a model together with the instructions to only base its answer on this information, may help LLMs overcome the domain gap while improving transparency and predictability.

There seems to be a trend to simplify the application of generative search for enterprise use-cases by increasing the context sizes for commercial models. Just recently, Anthrophic published a new version of their Claude LLM with a context size of now up to 100k tokens, advertised directly towards businesses as an alternative to vector based search over their documents. [7].

But an issue for deploying generative search in enterprise search use-cases is that the internal information and documents are often classified [8] and cannot just be sent to an external API for the generation of answers and summarizations. At the same time, the models that exhibit breakthrough performance in generative search are presumably so big (last published model size for GPT-3.5 InstructGPT was 175B parameters [9]) that the hardware cost of hosting them for inference might be prohibitively expensive.

The aforementioned closed domain hallucinations could be an especially important issue in generative search for enterprise use-cases, where incorrect information is less tolerable than in web search results.

The outlined potential and limitations of generative search approaches in domain-specific enterprise search use-cases open up several interesting research avenues.

3. Proposed Research Questions

The widespread use, new capabilities, and deployment of these commercial LLMs have sparked multiple interesting research problems. Commercial players have clear incentives to work on better-performing models to compete for the growing market of AI Assistants. However, the

[7]https://web.archive.org/web/20230524065546/https://www.anthropic.com/index/100k-context-windows

proprietary nature of these commercial models often limits transparency. For instance, even the number of parameters in the latest generation of commercial LLMs is undisclosed. This opacity has given rise to several open-source alternatives in recent months, such as OPT [10], BLOOM [11], or Pythia [12].

While the efficacy of future models will inevitably increase and much work is already done in regard to mitigating hallucinations, improving alignment and safety, we want to focus on the generative search use-cases and some of the limitations mentioned. The main overarching research question is: How can structured external knowledge, such as knowledge graphs, be most effectively used to improve the overall system performance? The specific research questions we aim to address are as follows:

- RQ1: How does the addition of grounding information to prompts influence the accuracy and likelihood of hallucinations in LLM responses, particularly in domain-specific and enterprise contexts?
- RQ2: How can external knowledge sources such as knowledge graphs be used to improve the retrieval and reasoning about the relevance and correctness of summarized information before the model generates a final answer based on this information?
- RQ3: Can the size of LLMs be reduced while maintaining or improving performance when they are fine-tuned for grounded question-answering and knowledge discovery tasks? How does this affect their applicability in enterprise search use-cases that require on-premises hosting?

3.1. Related Work

3.1.1. Research question 1

Research question 1 is primarily motivated by the need to estimate the likelihood of residual hallucinations and errors in a generative search setting when the model is supplied with correct information.

Current research explored the citation recall and precision in multiple public generative search engines and found that they frequently contain unsupported statements and inaccurate citations [2]. They also interestingly found that the responses that seem more helpful are often those with more unsupported statements or inaccurate citations.

Other interesting work looked at model behavior when presented with conflicting information, they found that the models ”demonstrate a strong confirmation bias when the external evidence contains some information that is consistent with their parametric memory” [13].

Earlier work by OpenAI also explored the truthfulness of a retrieval augmented model called WebGPT [14] on the TruthfulQA benchmark [15], consisting of question that cover common misconceptions that can likely be picked up by models during their pre-training. They found that WebGPT achieves 75% truthfulness by being also informative 54% of the time, and greatly improves over non retrieval augmented GPT-3. However, it still falls short of human accuracy, which stands at 95%.

An extensive overview of the hallucination problem in natural language generation (NLG) is given by Ji et al. [1]. They also mentioned that the challenges of measuring and mitigating these hallucinations are task dependent. We want to focus on the factuality of initial generative

search responses and the impact of grounding information in specific and enterprise domains, which relates to the tasks of generative question answering and abstractive summarization.

3.1.2. Research question 2

Research question 2 is more exploratory and suggests a possible solution to problems that might occur in generative search settings. We want to explore how structured knowledge from lookups in knowledge graphs and based on the entities discovered in the original user prompt can be used, during query generation in the retrieval step, and as additional grounding information that might help the model to focus on the most relevant aspects while answering the users' information need.

Some related work explored how to integrate knowledge graphs directly into the generation process of an LLM by embedding an additional encoding layer into the transformer architecture [16] or reason on subgraphs and choosing next words via a gate function drawing from the LLM vocabulary or the knowledge graph [17].

Other recent work used external knowledge lookups to recreate and correct possible mistakes in the generation process by estimating model uncertainty and supplying the model with additional grounding information in the correction step [18] [19].

We on the other hand want to explore neuro-symbolic approaches for entity linking [20] [21] during both retrieval and knowledge graph lookups to extract the most relevant information for the model and use chain-of-thought prompting [22] [23] grounded with this information, to possibly improve the model performance, factuality, and transparency.

3.1.3. Research question 3

Research question 3 is not only motivated by resource constraints in enterprise search use-cases, but also by the fact that the current state-of-the-art model architectures are too big to conduct in-depth research on for most researchers due to the high hardware requirements and associated costs. But there is promising research in optimizing the fine-tuning as well as inference of open source models to enable them to run on limited resources. Recently published QLoRA optimization for example enables the fine-tuning and inference of a 20 Billion parameter gpt-neo-x-20b model on only one 16 GB VRam GPU [24].

Other promising recent work has shown that even very small models, fine-tuned with the right data, can outperform models with magnitudes more parameters in text generation or reasoning [25] [26]. We want to especially explore how supplying the model with ideally all required knowledge to answer an information need, possibly simplifies the task. Preliminary work, that we will outline in the next section, conducted with GPT-3.5-turbo and GPT-4 on the 2023 BioASQ challenge indicated that GPT-3.5-turbo might compete with the presumably[8] bigger and better trained GPT-4 in grounded Q&A.

[8]OpenAI did not disclose parameter counts for either model

4. Preliminary Work

In our preliminary work, we assessed the capabilities of the commercially available LLMs, GPT-3.5-turbo and GPT-4, on the 2023 BioASQ challenge. BioASQ is a challenge on biomedical semantic indexing and question answering and in its eleventh installment held as a lab of the 2023 CLEF conference [27]. We chose the BioASQ challenge because it can be framed as a form of professional search, where the searchers are biomedical experts trying to find answers to domain-specific questions [28].

In Task 11 B Phase A, participating systems are given a biomedical question and need to retrieve relevant studies from PubMed and extract snippets that might help answer this question. In Phase B the systems are tasked with generating a short (max 200 words) answer as well as an exact structured answer (Yes/No, Factoid, List) depending on the type of the question, given the retrieved snippets and question. The overall process aligns with the retrieval and generation flow in generative search. The system has to first find relevant documents, extract or summarize the information, and write a coherent response that summarizes the findings and answers the biomedical question.

In our submission we chose to evaluate GPT-4 and GPT-3.5-turbo, the two models that power not only ChatGPT but also Bing Chat (GPT-4) and perplexity.ai (GPT-4). We used zero-shot learning, where the model is just given the test question directly without ever seeing any of the training data. We compared model performance both with and without adding relevant snippets to the model prompt to ground the answer generation. This compares plain LLM usage to the generative search approach.

The final results of the BioASQ Task 11 B Phase B, where experts rate the generated model answers, are not available yet. Nevertheless, preliminary results for the structured answer formats and rouge scores indicate that in the grounded setting both models demonstrated competitive abilities with leading systems, taking leading spots in 2 of 4 batches for factoid and list answers[9][10]. Interestingly, the older and cheaper GPT-3.5-turbo system was sometimes able to outperform GPT-4 in the grounded Q&A setting for factoid answers. We also observed that in the ungrounded Q&A setting, all models performed significantly worse than their grounded counterpart that had access to relevant snippets in its prompt.

These results might indicate that model size has less impact on generated answers in some tasks when sufficient grounding information is supplied to these generative models, reducing the problem to extractive question answering. Interestingly, the system using answers generated by GPT-4 without any grounding information was able to outperform GPT-3.5-turbo with grounding information in 2 of 4 batches on the Yes/No answer format. This might indicate that for this answer format, better reasoning capabilities are required, which are shown to improve with model size [29].

In future work, we want to assess and quantify the factuality and hallucination prevalence in the generated ideal answers on the BioASQ data and compare these between grounded and ungrounded settings and with additional presumably smaller open-source models.

Our preliminary work therefore highlighted several interesting aspects about the use of GPT

[9]Our participating systems are prefixed with "UR-gpt", the "-simple" in the name means no grounding information was given.

[10]http://participants-area.bioasq.org/results/11b/phaseB/

models in domain-specific generative search Q&A task, the challenges they pose, and potential strategies for improvement. These insights will guide our proposed research questions.

5. Discussion

One could argue that research question 1 is only investigating a temporary shortcoming of state of the art LLMs [30] and companies such as OpenAI will eventually solve the hallucination problems with new approaches such as process supervision [31]. But we still see a current gap in research that estimates the hallucination prevalence in the closed domain setting. OpenAI only stated that they improved on the problem, but they did not publish actual evaluations on the overall frequency of hallucinations in this setting, as mentioned in section 2.1.

Research question 2 is highly dependent on the optimal prompt design, which is a hard problem to solve in its own. Currently, a lot of zero shot prompts that are used in use-case specific research are human designed, and only follow some patterns that emerged in research or are suggested by the vendors such as chain-of-thought prompting. Some current research suggests that LLMs can optimize prompts iteratively [32] [33], which might be an interesting approach to make prompt engineering more transparent.

Research question 3 assumes that generative search in enterprise search use-cases can be limited to question answering and knowledge discovery, this might not capture the breath of use-cases that open up in using generative models augmented with internal data. But thoroughly evaluating the impact of model size on the performance requires that the scope is first limited to a well measurable use-case. Such smaller fine-tuned expert models could also be combined in a mixture of experts architecture that can perform well on diverse tasks [34] [35], which is an approach rumored to be used in GPT-4[11].

Using closed-source commercial models such as GPT-3.5-turbo and GPT-4 in research, for which not even the parameter count or the training data is published, can be considered unscientific. But in the case of the new field of generative search, this application only just emerged with the unparalleled capability of OpenAIs ChatGPT models (GPT-4 and GPT-3.5-turbo). Failing to study the capabilities and behavior of these models, especially when they are actively integrated into products, risks making academic research less relevant. It also potentially widens the gap between proprietary enterprises and open academic investigations, solidifying the lead that commercial entities hold in this domain. We are confident that the continuous publishing of more capable open alternatives such as Llama 2 based models [36] [37] will reduce the reliance on commercial models in this field, and we are committed to exploring their capabilities in future work as stated in research question 3.

There are also concerns about the reproducibility of scientific results when using OpenAIs models, as with their default token sampling strategy they exhibit a fair amount of non-determinism. Any research conducted via the OpenAI API should therefore set their temperature parameter to 0, this will ensure that the output is mostly deterministic[12]. Some slight variability in the model outputs remains if two tokens have very similar probability. We

[11] https://web.archive.org/web/20230728003847/https://the-decoder.com/gpt-4-architecture-datasets-costs-and-more-leaked/

[12] https://platform.openai.com/docs/guides/gpt/why-are-model-outputs-inconsistent

were concerned with the reproducibility of our results in our preliminary work, especially in the case of query expansion, and therefore conducted a rough variance estimation over 5 runs with 50 questions which indicated low variance and therefore good reproducibility. We still think that a thorough analysis of this residual nondeterminism is needed, and leave that open for future work.

6. Conclusion

This position paper explores the potential and limitations of generative search engines utilizing LLMs. It highlights three research questions that aim to quantify and explore current constraints, including occasional hallucinations, inefficient retrieval and summarization, and model efficiency. It proposes to explore integrating structured knowledge, such as knowledge graphs, to ground and improve the effectiveness, efficiency, and transparency of generative search engines, especially within enterprise contexts.

The preliminary work demonstrates the abilities and limitations of two commercially available LLMs, GPT-3.5-turbo and GPT-4, on a biomedical retrieval and question answering task. Both models performed competitively, with the older, cheaper GPT-3.5-turbo occasionally surpassing GPT-4 in the grounded setting for certain types of answers, indicating that model size might not be as relevant in extractive question answering.

Future work will explore quantifying the impact of grounding information on hallucination rates and accuracy. It will also investigate leveraging knowledge graphs to improve retrieval, reasoning, and summarization of relevant information before generating responses. Finally, it will explore reducing model sizes while maintaining performance on grounded tasks, enabling deployment within enterprise settings. Overall, this research proposes to enhance the effectiveness, efficiency, and transparency of generative search systems by integrating structured knowledge representations with large language models.

References

[1] 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.

[2] N. F. Liu, T. Zhang, P. Liang, Evaluating verifiability in generative search engines, arXiv preprint arXiv:2304.09848 (2023).

[3] H. Li, Y. Su, D. Cai, Y. Wang, L. Liu, A Survey on Retrieval-Augmented Text Generation, 2022. arXiv:2202.01110.

[4] OpenAI, GPT-4 Technical Report, 2023. arXiv:2303.08774.

[5] S. Diao, R. Xu, H. Su, Y. Jiang, Y. Song, T. Zhang, Taming Pre-trained Language Models with N-gram Representations for Low-Resource Domain Adaptation, in: Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL-IJCNLP 2021, 2021, p. 3336.

[6] E. M. Bender, T. Gebru, A. McMillan-Major, S. Shmitchell, On the Dangers of Stochastic

Parrots: Can Language Models Be Too Big?, in: Proceedings of the 2021 ACM conference on fairness, accountability, and transparency, 2021, pp. 610–623.
[7] T. Russell-Rose, A. MacFarlane, Towards explainability in professional search, in: The 3rd International Workshop on ExplainAble Recommendation and Search (EARS 2020), 2020. URL: https://ears2020.github.io/accept%5fpapers/1.pdf.
[8] U. Kruschwitz, C. Hull, et al., Searching the enterprise, Foundations and Trends® in Information Retrieval 11 (2017) 1–142.
[9] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al., Training language models to follow instructions with human feedback, Advances in Neural Information Processing Systems 35 (2022) 27730–27744.
[10] S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. Diab, X. Li, X. V. Lin, T. Mihaylov, M. Ott, S. Shleifer, K. Shuster, D. Simig, P. S. Koura, A. Sridhar, T. Wang, L. Zettlemoyer, OPT: Open Pre-trained Transformer Language Models, 2022. arXiv:2205.01068.
[11] S. et al., BLOOM: A 176B-Parameter Open-Access Multilingual Language Model, 2023. arXiv:2211.05100.
[12] S. Biderman, H. Schoelkopf, Q. Anthony, H. Bradley, K. O'Brien, E. Hallahan, M. A. Khan, S. Purohit, U. S. Prashanth, E. Raff, A. Skowron, L. Sutawika, O. van der Wal, Pythia: A Suite for Analyzing Large Language Models Across Training and Scaling, 2023. arXiv:2304.01373.
[13] J. Xie, K. Zhang, J. Chen, R. Lou, Y. Su, Adaptive Chameleon or Stubborn Sloth: Unraveling the Behavior of Large Language Models in Knowledge Clashes, 2023. arXiv:2305.13300.
[14] 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.
[15] S. Lin, J. Hilton, O. Evans, TruthfulQA: Measuring How Models Mimic Human Falsehoods, 2022. arXiv:2109.07958.
[16] Z. Hu, Y. Xu, W. Yu, S. Wang, Z. Yang, C. Zhu, K.-W. Chang, Y. Sun, Empowering language models with knowledge graph reasoning for open-domain question answering, in: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, 2022, pp. 9562–9581.
[17] H. Ji, P. Ke, S. Huang, F. Wei, X. Zhu, M. Huang, Language Generation with Multi-Hop Reasoning on Commonsense Knowledge Graph, 2020. arXiv:2009.11692.
[18] Z. Jiang, F. F. Xu, L. Gao, Z. Sun, Q. Liu, J. Dwivedi-Yu, Y. Yang, J. Callan, G. Neubig, Active Retrieval Augmented Generation, 2023. arXiv:2305.06983.
[19] R. Zhao, X. Li, S. Joty, C. Qin, L. Bing, Verify-and-edit: A knowledge-enhanced chain-of-thought framework, 2023. arXiv:2305.03268.
[20] L. Dietz, H. Bast, S. Chatterjee, J. Dalton, E. Meij, A. de Vries, ECIR 23 Tutorial: Neuro-Symbolic Approaches for Information Retrieval, in: Advances in Information Retrieval: 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2–6, 2023, Proceedings, Part III, Springer, 2023, pp. 324–330.
[21] S. Chatterjee, L. Dietz, Predicting Guiding Entities for Entity Aspect Linking, in: Proceedings of the 31st ACM International Conference on Information & Knowledge

Management, CIKM '22, Association for Computing Machinery, New York, NY, USA, 2022, p. 3848–3852. URL: https://doi.org/10.1145/3511808.3557671. doi:10.1145/3511808.3557671.

[22] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, D. Zhou, Chain-of-Thought Prompting Elicits Reasoning in Large Language Models, 2023. arXiv:2201.11903.

[23] A. Saparov, H. He, Language Models Are Greedy Reasoners: A Systematic Formal Analysis of Chain-of-Thought, 2023. arXiv:2210.01240.

[24] T. Dettmers, A. Pagnoni, A. Holtzman, L. Zettlemoyer, QLoRA: Efficient Finetuning of Quantized LLMs, 2023. arXiv:2305.14314.

[25] R. Eldan, Y. Li, TinyStories: How Small Can Language Models Be and Still Speak Coherent English?, 2023. arXiv:2305.07759.

[26] N. Ho, L. Schmid, S.-Y. Yun, Large Language Models Are Reasoning Teachers, 2023. arXiv:2212.10071.

[27] A. Nentidis, A. Krithara, G. Paliouras, E. Farré-Maduell, S. Lima-López, M. Krallinger, BioASQ at CLEF2023: The Eleventh Edition of the Large-Scale Biomedical Semantic Indexing and Question Answering Challenge, in: Advances in Information Retrieval, Springer Nature Switzerland, Springer Nature Switzerland, Cham, 2023. URL: https://link.springer.com/chapter/10.1007/978-3-031-28241-6_66.

[28] S. Ateia, U. Kruschwitz, Is ChatGPT a Biomedical Expert? – Exploring the Zero-Shot Performance of Current GPT Models in Biomedical Tasks, 2023. arXiv:2306.16108.

[29] 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.

[30] S. R. Bowman, Eight things to know about large language models, 2023. arXiv:2304.00612.

[31] H. Lightman, V. Kosaraju, Y. Burda, H. Edwards, B. Baker, T. Lee, J. Leike, J. Schulman, I. Sutskever, K. Cobbe, Let's Verify Step by Step, 2023. arXiv:2305.20050.

[32] Y. Zhou, A. I. Muresanu, Z. Han, K. Paster, S. Pitis, H. Chan, J. Ba, Large language models are human-level prompt engineers, 2022. arXiv:2211.01910.

[33] R. Pryzant, D. Iter, J. Li, Y. T. Lee, C. Zhu, M. Zeng, Automatic prompt optimization with "gradient descent" and beam search, 2023. arXiv:2305.03495.

[34] N. Shazeer, A. Mirhoseini, K. Maziarz, A. Davis, Q. Le, G. Hinton, J. Dean, Outrageously large neural networks: The sparsely-gated mixture-of-experts layer, 2017. arXiv:1701.06538.

[35] N. Du, Y. Huang, A. M. Dai, S. Tong, D. Lepikhin, Y. Xu, M. Krikun, Y. Zhou, A. W. Yu, O. Firat, B. Zoph, L. Fedus, M. P. Bosma, Z. Zhou, T. Wang, E. Wang, K. Webster,

M. Pellat, K. Robinson, K. Meier-Hellstern, T. Duke, L. Dixon, K. Zhang, Q. Le, Y. Wu, Z. Chen, C. Cui, GLaM: Efficient scaling of language models with mixture-of-experts, 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. 5547–5569. URL: https://proceedings.mlr.press/v162/du22c.html.

[36] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. C. Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P. S. Koura, M.-A. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. E. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, T. Scialom, Llama 2: Open Foundation and Fine-Tuned Chat Models, 2023. arXiv:2307.09288.

[37] A. N. Lee, C. J. Hunter, N. Ruiz, Platypus: Quick, Cheap, and Powerful Refinement of LLMs, 2023. arXiv:2308.07317.