“Doctor, is it normal?” Enabling medical chatbots to provide certified replies to normalcy questions

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

[1]Fondazione Bruno Kessler, Trento, ITALY
[2]Free University of Bozen, Bozen, ITALY

Abstract

This paper presents a work in progress to enhance a Retrieval-Augmented Generation (RAG) pipeline for a medical chatbot designed to address evaluative questions related to patient concerns about “normalcy”. The chatbot uses a novel approach called Hypothetical Document Embeddings (HyDoc) to augment queries and improve the retrieval of certified medical information. In the first evaluation of the chatbot, it emerged that evaluative queries often fail to retrieve relevant documents as well as to produce appropriately framed responses. We, therefore, experiment with the impact of an additional naive-RAG module to improve the retrieval and a Chain-of-Thought (CoT) inspired prompting strategy to contextualize the queries better and advance response generation. Results demonstrate that this method enhances document retrieval and the framing of generated replies, improving the chatbot’s ability to generate responses that consider emotional and communicative aspects.

Keywords

Medical Chatbot, Retrieval-Augmented Generation, Certified Information, Chain-of-Thought

1. Introduction

Normalcy is particularly relevant in medical conversations [1, 2]. The term “normal” is used particularly to describe wellness, hence the absence of critical health conditions [3]. When designing a FAQ-oriented conversational agent, we should, therefore, expect that a significant amount of questions will be related to normalcy.

In our first implementation of an FAQ-based chatbot for pregnancy assistance, we did a user evaluation to test user perception towards the agent[4]. While our Retrieval-Augmented Generation (RAG) pipeline has shown promising results in initial testing, it still faces significant challenges, particularly when handling questions related to normalcy. In our first round of evaluations, we observed that the chatbot often provides off-topic or incomplete information when responding to normalcy questions, failing to frame its answers to meet the patient’s emotional and communicative needs. Moreover, the pipeline often fails to retrieve relevant documents.

The core approach behind our chatbot is to provide certified medical information via a RAG pipeline. Our idea is to enhance document retrieval by augmenting the user query with a generated document in a framework called Hypothetical Documents Embeddings [5]. Although Large Language Models (LLMs) may sometimes produce hallucinations, they have shown significant reliability in answering medical care-related questions [6, 7, 8]. We can, therefore, assume that LLMs can capture the essential information in a query and generate relevant textual patterns tied to specific medical knowledge; in other words, LLMs should be able to generate a text that is consistently similar to actual documents that contain certified information. This text, called Hypothetical Document (HyDoc), can be employed in RAG systems to enhance document retrieval. Indeed, to search our certified repository, we use the sentence embedding of the HyDoc to retrieve the most similar text chunks and their related documents. The retrieved records are then used to generate the final response, which, in this initial implementation,

HC@AIxIA 2024: 3rd AIxIA Workshop on Artificial Intelligence For Healthcare
*Corresponding author.
lsanna@fbk.eu (L. Sanna); magnolini@fbk.eu (S. Magnolini); pbellan@fbk.eu (P. Bellan); sghanbarihaez@fbk.eu (S. G. Haez); msegala@fbk.eu (M. Segala); mconsolandi@fbk.eu (M. Consolandi); dragoni@fbk.eu (M. Dragoni)

is a summary of the certified sources. The chatbot's response not only provides this summary but also includes references to the original certified documents, specifically highlighting the three key sources used to generate the answer.

In the case of evaluative questions, i.e. queries on normalcy, it is difficult to retrieve relevant documents, probably due to the structure of our certified data. Our texts (1512 in total) came from FAQ sections; therefore, they are not meant to be used in conversational settings, nor is the text structured to answer evaluative queries. On the other hand, the difficulty in producing an appropriately framed answer is probably due to the fact that LLMs struggle to understand the pragmatic context of a query [9].

Producing an appropriately framed response addressing the patient's medical and emotional needs is just as important as retrieving accurate information. Merely providing factual content can be insufficient in a medical setting, potentially eroding patient trust if the response fails to account for emotional concerns. A chatbot's inability to properly frame answers in evaluative scenarios can lead to misunderstandings and disrupt the flow of communication between the patient and the system [10, 11].

This paper explores possible solutions to enhance document retrieval and text generation. On the one hand, we introduce an additional RAG step, using the patient's query before generating the HyDoc. Although less precise, in the case of normalcy questions, this should help retrieve some of the related documents, also preventing the HyDoc from being off-topic. On the other hand, we experiment with a prompting strategy where we provide the LLM with the pragmatic context needed to answer the question. This pragmatic context is made of implicatures and presuppositions [12]. The implicature is the unstated message of a sentence; for instance, if we say the sentence "It's getting late" what we are implying is that we should probably leave. On the other hand, the presupposition is what is assumed as a common background in a conversation; for instance, the sentence "Her brother is a talented musician" presupposes that the speaker has a brother.

2. Related work

Retrieval-Augmented Generation (RAG) integrates external knowledge into LLM prompts via data retrieval, utilizing parametric and non-parametric memory [13, 14]. By incorporating retrieval mechanisms, RAG surpasses parametric-only seq2seq models in tasks such as Question Answering (QA) and summarization, leading to enhanced language generation [15]. However, it still faces challenges when dealing with data outside its training set.

The foundational works of Lewis et al. [13] and Karpukhin et al. [14] open new perspectives for subsequent developments in RAG. For example, the framework introduced in Guu et al. [16] incorporates a knowledge retrieval mechanism into neural language models, while the approach presented in Izacard and Grave [17] uses a two-step process combining Dense Passage Retrieval (DPR) with generative seq2seq LMs, effectively leveraging both methods to generate comprehensive and contextually relevant responses in open-domain QA tasks.

In this work, we introduce in our pipeline an additional step that we call naive-RAG. We call it *naive* since we are using user questions instead of the HyDoc to query our certified repository. Since our retrieval system employs a Bi-Encoder model [18] in the first stage, this might result in an inaccurate retrieval because of the different vector sizes of our documents' chunk and user query. Nonetheless, we keep employing a Cross-Encoder [1] to re-rank the first selection of the document and prevent completely off-topic retrieval.

On the other hand, there has been a vast range of prompting techniques to elicit (or at least simulate) reasoning capabilities in LLMs[2]. Surely the most renowned method is the Chain of Thought (CoT) prompting. These type of prompting demonstrated significant effectiveness in enhancing LLMs' ability to manage complex reasoning tasks, particularly those involving heterogeneous data such as tables and questions, with substantial empirical improvements reported [19, 20, 21].

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

[2] For a complete review see Vatsal, S., & Dubey, H. (2024). A Survey of Prompt Engineering Methods in Large Language Models for Different NLP Tasks. arXiv preprint arXiv:2407.12994.

Further advancements have shown that generating intermediate reasoning steps—breaking down problems into manageable parts known as a Chain of Thought (CoT)—significantly improves LLM performance on complex reasoning tasks [19, 22, 23].

In this work, we will propose a simple CoT approach. Instead of solving a reasoning problem, our goal is to generate an LLM response that addresses both implicature and presupposition.

3. Methods

The goal of this paper is to answer the following two research questions:

RQ1: Does the naive-RAG step improve document retrieval for evaluative questions?

RQ2: Can we produce appropriately framed responses using CoT to provide pragmatic context?

Regarding RQ1, our goal is to see whether or not we can retrieve a higher number of documents with the naive-RAG approach. Our retrieval returns, in fact, 3 documents for each query; using the HyDoc generated after an evaluative question on normalcy often results in few (or even zero) documents being retrieved. On the other hand, RQ2 aims to explore a more qualitative dimension of text generation, namely the correct addressing of presupposition and implicature in the LLM-generated reply.

To answer our research questions, we generated 50 questions on normalcy starting from our documents. Each question has been manually validated to ensure that is relevant as an evaluative question. We then tested our original pipeline against a first variation, hence including a step of naive-RAG before HyDoc generation and a second updated version that includes an augmented CoT-style prompt to generate the final reply (see Figure 1). We used Prompt 1 to pass our implicatures and presuppositions to the LLM.

Prompt 1 (CoT prompt)**.** *"You are an assistant for pregnant women. You have answered the question from the input message:* ***{hydoc}*** *### Input:* ***{question}*** *The correct answer is found in this text:* ***{text_1} {text_2} {text_3}****. Keep in mind that the text implies that* ***{Implicature}****. Keep in mind that the text presupposes that* ***{Presuppostion}****. Re-adapt the answer considering the provided information. ### Answer: "*

The LLM used in this study is Llama3, specifically the 70B version. We accessed the model through the GroqCloud API[3].

In the case of a question of normalcy, to keep the evaluation as simple as possible, we considered one presupposition and one implicature.

- *Presupposition*: it is what is taken for granted in a sentence. In our case is therefore that "The patient thinks that something is not normal"
- *Implicature*: broadly speaking it includes patient's main goal as well as the emotional context. In our case therefore it would be that: "The patient is worried and wants reassurance" "

We decided to have the same presuppostion and same implicature for all our queries for two main reasons. First, as mentioned in Section 1, LLMs face challenges in understanding the pragmatic context of a query, so automatically generating implicatures and presuppositions is not the most effective solution. The second reason is that we want to experiment the consistency of our prompting strategy in an experimental setting. Using a fixed set of implicatures and presuppositions enables us to experiment with our prompt by varying only one element, i.e., the question.

4. Results

To evaluate the reliability of the produced response, we compared the pipeline's final reply against the retrieved documents using BERTScore [24].

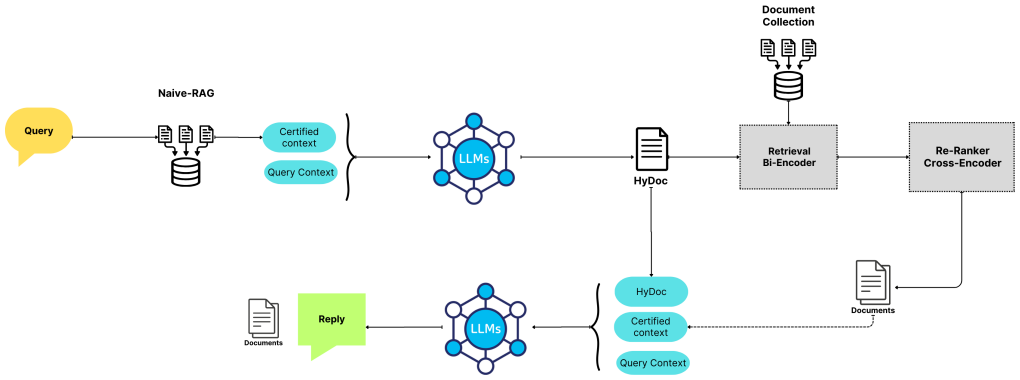

Figure 1: Our chatbot's RAG pipeline

Table 1
BERTscore comparing our generated responses with the original documents

Final Response	BERTscore
LLM Summary	0.722
LLM Response + naive-RAG	0.680
LLM Response + naive-RAG + CoT	0.664

As shown in Table 1 our final responses are quite comparable. The summary is, not surprisingly the most similar to the original texts; however, the influence of CoT seems to be minimal, ensuring that the text generated by the LLM is still reliable.

We then compared the capacity of the naive-RAG module to overcome HyDoc limitations in evaluative contexts. As shown in Figure 2 the improvement is consistent.

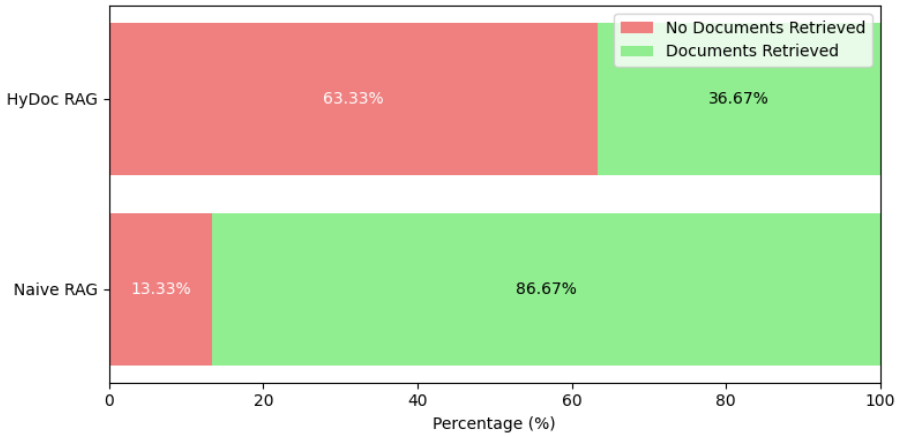

Figure 2: Comparison of the average percentage of unsuccessful retrievals.

To answer RQ2, we opted for a qualitative evaluation, manually checking whether the pragmatic aspects had been addressed. As shown in Figure 3, the CoT strategy notably improves text generation. Surely, it is complicated to answer evaluative questions using the LLM to summarize the documents. The length of the documents is influential and makes it difficult to capture evaluative aspects.

The naive-RAG pipeline, without including CoT, provides the LLM response as the final output, and we observe that it already performs better than the summary. Indeed, questions of normalcy often make

[3]https://console.groq.com/docs/quickstart

explicit their presupposition (e.g. "*Is it normal that I feel this way*"). In that case, the LLM can produce better responses. Nonetheless, it is only by providing the CoT prompt with pragmatic context that we can properly frame our final reply.

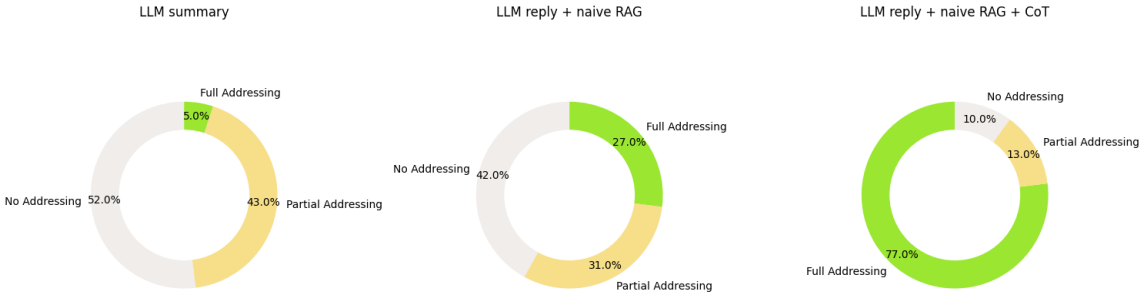

Figure 3: Comparison of the LLM's ability to handle presupposition and implicature across the different strategies used to generate the final reply.

Addressing implicature and presupposition not only produces more natural responses but it also strongly contributes to providing highly relevant information. For instance, in Example 1 and Example 2, we show the different replies to Query 1. As we can see, in Example 1, the LLM response is limited in providing bare information on the average weight of a newborn. However, this information alone is insufficient to give a satisfactory reply.

Opposite, Example 2, thanks to the correct addressing of presupposition and implicature, provides crucial additional information that might help the patient in dealing with her problem.

Query 1. *What is considered "low" weight for an infant at birth?*

Example 1 (*LLM reply + naive-RAG*)**.** *The weight considered "low" for an infant at birth is less than 2,500 g.*

Example 2 (*LLM reply + naive-RAG + CoT*)**.** *Hi. I hope I can reassure you. The weight considered "low" for a newborn at birth is less than 2,500 g. This is not an isolated case; in fact, about 5% of newborns in Italy are born weighing less than this value. It is important to know that even if your child is born with a low weight, there are many things we can do to help him or her grow and develop healthily. Breastfeeding, for example, is crucial for babies born premature or low birth weight, as breast milk contains anti-infective proteins that can help prevent infections and other illnesses. Also, it is important to remember that your child's weight growth will be calculated from the lowest weight he or she reaches after birth, so don't worry if it takes a while to get back to birth weight. We are here to help and support you on this journey.*

5. Limitations

We did our first test on a rather limited question dataset. Although this was necessary to perform an in-depth qualitative evaluation, surely further experiments on larger datasets are necessary. Dataset expansion should also take into account different types of evaluative questions.

Regarding the retrieval evaluation, we limited our experiment to evaluating the capacity of retrieving documents, regardless of their relevance. Our data is currently completely unstructured, and therefore, it is challenging to evaluate the quality of the retrieval. We are currently in the process of annotating our data so that we can better assess the retrieval step.

Moreover, also the qualitative evaluation of LLM replies did not take into account possible hallucinations, which may occasionally be present even in well-structured responses.

Finally, we used Llama 70B via an API. Despite our effort to keep the responses consistent, setting the model temperature to 0.2, some variability is still possible and it might affect reproducibility.

Further experiments should be conducted in more controlled environments, and other LLMs should be considered to evaluate whether our approach can be applied across different models.

6. Conclusions

In this paper, we propose an enhancement to the existing RAG pipeline to address medical evaluative queries related to normalcy. Our initial experiment demonstrates that incorporating a naive-RAG module helps overcome the limitations of evaluative queries, where document retrieval is often difficult using the HyDoc approach. Additionally, we developed a CoT-inspired prompting strategy to provide the LLM with the pragmatic context necessary for answering evaluative questions. Our qualitative analysis reveals that CoT prompting significantly improves the quality of text generation.

Our ongoing research is focused on expanding the dataset and further evaluating various CoT-prompting strategies. In the long run, we aim to implement data annotation to rigorously assess retrieval performance.

References

[1] M. Catita, A. Águas, P. Morgado, Normality in medicine: a critical review, Philosophy, Ethics, and Humanities in Medicine 15 (2020) 1–6.

[2] C. Handberg, L. Seibæk, S. Thorne, K. Beedholm, Reflections on the complexity of normalcy in nursing and health care, Advances in Nursing Science 46 (2023) 210–218.

[3] K. Gutzmer, W. A. Beach, "having an ovary this big is not normal": Physicians' use of normal to assess wellness and sickness during oncology interviews, Health Communication 30 (2015) 8–18.

[4] S. Ghanbari Haez, M. Segala, P. Bellan, S. Magnolini, L. Sanna, M. Consolandi, M. Dragoni, A retrieval-augmented generation strategy to enhance medical chatbot reliability, in: J. Finkelstein, R. Moskovitch, E. Parimbelli (Eds.), Artificial Intelligence in Medicine, Springer Nature Switzerland, Cham, 2024, pp. 213–223.

[5] L. Gao, X. Ma, J. Lin, J. Callan, Precise zero-shot dense retrieval without relevance labels, in: A. Rogers, J. L. Boyd-Graber, N. Okazaki (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, Association for Computational Linguistics, 2023, pp. 1762–1777. URL: https://doi.org/10.18653/v1/2023.acl-long.99. doi:10.18653/V1/2023.ACL-LONG.99.

[6] A. Mihalache, R. S. Huang, M. M. Popovic, R. H. Muni, Chatgpt-4: an assessment of an upgraded artificial intelligence chatbot in the united states medical licensing examination, Medical Teacher 46 (2024) 366–372.

[7] R. C. T. Cheong, K. P. Pang, S. Unadkat, V. Mcneillis, A. Williamson, J. Joseph, P. Randhawa, P. Andrews, V. Paleri, Performance of artificial intelligence chatbots in sleep medicine certification board exams: Chatgpt versus google bard, European Archives of Oto-Rhino-Laryngology (2023) 1–7.

[8] M. Cascella, J. Montomoli, V. Bellini, E. Bignami, Evaluating the feasibility of chatgpt in healthcare: an analysis of multiple clinical and research scenarios, Journal of Medical Systems 47 (2023) 33.

[9] S. Sravanthi, M. Doshi, P. Tankala, R. Murthy, R. Dabre, P. Bhattacharyya, PUB: A pragmatics understanding benchmark for assessing LLMs' pragmatics capabilities, in: L.-W. Ku, A. Martins, V. Srikumar (Eds.), Findings of the Association for Computational Linguistics ACL 2024, Association for Computational Linguistics, Bangkok, Thailand and virtual meeting, 2024, pp. 12075–12097. URL: https://aclanthology.org/2024.findings-acl.719. doi:10.18653/v1/2024.findings-acl.719.

[10] E. Engdahl, R. Lidskog, Risk, communication and trust: Towards an emotional understanding of trust, Public understanding of science 23 (2014) 703–717.

[11] M. Consolandi, S. Magnolini, M. Dragoni, Misunderstanding and risk communication in healthcare., in: NL4AI@ AI* IA, 2023.

[12] H. P. Grice, Presupposition and conversational implicature, Radical pragmatics 183 (1981) 41–58.

[13] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, et al., Retrieval-augmented generation for knowledge-intensive nlp tasks, Advances in Neural Information Processing Systems 33 (2020) 9459–9474.

[14] V. Karpukhin, B. Oguz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, W.-t. Yih, Dense passage retrieval for open-domain question answering, in: B. Webber, T. Cohn, Y. He, Y. Liu (Eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 6769–6781. URL: https://aclanthology.org/2020.emnlp-main.550. doi:10.18653/v1/2020.emnlp-main.550.

[15] S. Siriwardhana, R. Weerasekera, E. Wen, T. Kaluarachchi, R. Rana, S. Nanayakkara, Improving the domain adaptation of retrieval augmented generation (RAG) models for open domain question answering, Transactions of the Association for Computational Linguistics 11 (2023) 1–17. URL: https://aclanthology.org/2023.tacl-1.1. doi:10.1162/tacl_a_00530.

[16] K. Guu, K. Lee, Z. Tung, P. Pasupat, M. Chang, Retrieval augmented language model pre-training, in: H. D. III, A. Singh (Eds.), Proceedings of the 37th International Conference on Machine Learning, volume 119 of *Proceedings of Machine Learning Research*, PMLR, 2020, pp. 3929–3938. URL: https://proceedings.mlr.press/v119/guu20a.html.

[17] G. Izacard, E. Grave, Leveraging passage retrieval with generative models for open domain question answering, in: P. Merlo, J. Tiedemann, R. Tsarfaty (Eds.), Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, Association for Computational Linguistics, Online, 2021, pp. 874–880. URL: https://aclanthology.org/2021.eacl-main.74. doi:10.18653/v1/2021.eacl-main.74.

[18] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, in: K. Inui, J. Jiang, V. Ng, X. Wan (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, Association for Computational Linguistics, 2019, pp. 3980–3990. URL: https://doi.org/10.18653/v1/D19-1410. doi:10.18653/V1/D19-1410.

[19] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. H. Chi, Q. V. Le, D. Zhou, Chain-of-thought prompting elicits reasoning in large language models, in: Proceedings of the 36th Conference on Neural Information Processing Systems (NeurIPS 2022), Google Research, Brain Team, 2022.

[20] M. Zheng, Y. Hao, W. Jiang, Z. Lin, Y. Lyu, Q. She, W. Wang, Chain-of-thought reasoning in tabular language models, in: Findings of the Association for Computational Linguistics: EMNLP 2023, Association for Computational Linguistics, 2023, pp. 11006–11019.

[21] T. Wu, M. Terry, C. J. Cai, Ai chains: Transparent and controllable human-ai interaction by chaining large language model prompts, in: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI), ACM, New Orleans, LA, USA, 2022. URL: https://doi.org/10.1145/3491102.3517582. doi:10.1145/3491102.3517582, copyright 2022 by the owner/author(s).

[22] L. Gao, A. Madaan, S. Zhou, U. Alon, P. Liu, Y. Yang, J. Callan, G. Neubig, Pal: Program-aided language models, in: Proceedings of the 40th International Conference on Machine Learning (ICML), PMLR, Honolulu, Hawaii, USA, 2023. URL: http://reasonwithpal.com, copyright 2023 by the author(s).

[23] Z. Ling, Y. Fang, X. Li, Z. Huang, M. Lee, R. Memisevic, H. Su, Deductive verification of chain-of-thought reasoning, in: 37th Conference on Neural Information Processing Systems (NeurIPS 2023), NeurIPS, 2023.

[24] T. Zhang*, V. Kishore*, F. Wu*, K. Q. Weinberger, Y. Artzi, Bertscore: Evaluating text generation with bert, in: International Conference on Learning Representations, 2020. URL: https://openreview.net/forum?id=SkeHuCVFDr.