Chain-of-Thought to Enhance Document Retrieval in Certified Medical Chatbots

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

We propose a Retrieval-Augmented Generation pipeline aimed at retrieving certified medical information. Inspired by the recently introduced Hypothetical Document Embeddings framework, we use the LLM to generate a document to query our certified repository. Although showing promising results in the first user evaluation, the proposed pipeline sometimes fails to retrieve the correct documents. We therefore propose a second Chain-of-thought-inspired pipeline to enhance the generation of the Hypothetical Document and, consequently, the retrieval of the certified documents.

Keywords

Conversational Agent, Digital Health, Chain-of-Thought, Certified Information

1. Introduction

The Hypothetical Document Embeddings (HyDE) framework has been recently introduced as an effective method to build dense retrievers completely unsupervised [1]. The key idea behind HyDE is to leverage the Large Language Model (LLM) creative abilities to generate a Hypothetical Document (HyDoc) which is then used to retrieve a real document in a repository.

Hence, HyDE is particularly well-suited for building medical chatbots that operate with"*certified information*", i.e. conversational agents capable of providing trustworthy information that has been created or verified by domain experts such as physicians or other healthcare professionals in the digital health industry

To provide "*certified information*", the chatbot's reply must be predetermined, namely that we have a predefined set of answers for each specific question. The existing lack of conversational datasets in the medical domain, however, poses a substantial challenge in creating a certified medical chatbot. To tackle this issue, we devised a Retrieval-Augmented Generation (RAG) pipeline within the HyDE framework so that we could benefit from the conversational capabilities of an LLM and, at the same time, exploit the LLM to retrieve the certified sources supporting the reply.

We believe that adopting HyDE addresses two major issues of RAG pipelines. First of all, we are trying to build a FAQ-based chatbot, therefore most of the interactions with the patients would be short questions. In a FAQ-oriented conversational agent, using a simple naive-RAG pipeline the user query would employed to retrieve the certified sources. Yet, since we are operating with vector databases, the vector representation of the query might be significantly distant from the certified documents in the semantic space, yielding a remarkable risk of excluding relevant documents in the retrieval process.

Moreover, in a digital health context, it is important to keep our certified medical chatbot explainable [2]. RAG approaches add a further layer of algorithmic opacity since the user is unaware of the documents used to generate the reply. Therefore, on the one hand, we use the retrieved document to produce a well-grounded and informed reply, while on the other hand, we provide the certified sources that have been retrieved, computing the similarity with the HyDoc.

Nonetheless, the quality of the generated HyDoc remains a substantial issue in medical domains. Although LLMs have shown impressive results in addressing medical queries [3, 4, 5], relying on the sole abilities of the LLM might result in generating inaccurate or low-quality HyDocs.

In fact, in a first user evaluation of our proposed modular pipeline, we found evidence that the retrieval step might be problematic when encountering specific types of questions, e.g. evaluative questions. This paper therefore introduces the main challenges we found in developing a modular RAG pipeline in a certified context. In particular, we focus on the proposal of a Chain-of-thought-inspired pipeline to enhance the HyDoc generation and, consequently, improve the retrieval of the certified sources.

2. Related work

LLMs' credibility and effectiveness are crucial in AI research, especially in areas like digital health and wellbeing that require precision and reliability [6]. RAG and Chain of Thought (CoT) prompting are highly effective in reducing hallucinations and enhancing factual content generation in LLMs by integrating external knowledge.

RAG integrates external knowledge into LLMs' prompts through data retrieval using parametric and non-parametric memory [7, 8]. It has been shown that RAG outperforms parametric-only seq2seq models in tasks like Question Answering (QA) and summarization, improving text generation [9].

Various approaches have been explored to advance QA systems. For instance, the work [10] involves a two-stage process that combines Dense Passage Retrieval (DPR) with generative sequence-to-sequence LMs. Other examples are the iterative integration of retrieval and generation [11], a combination of retrieval and generation techniques for informative answers [12], and dynamic real-time retrieval during generation [13]. Other approaches include techniques to im-

Information Retrieval's Role in RAG Systems (IR-RAG) - 2024
*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)

prove the accuracy of language models integrating external knowledge [14, 15], as well as advancing implicit reasoning and adaptability in QA tasks [9].

On the other hand, CoT methods have been highly effective in improving LLMs' ability to handle complex reasoning tasks, such as those that involve heterogeneous data from tables and questions [16, 17, 18]. Some recent studies have shown that breaking down problems into manageable steps significantly enhances LLMs' performance in complex reasoning tasks [16, 19, 20].

The work of [21] refines self-consistency decoding for broader applications like translation strategies and sentiment analysis, while [22] introduces the Zero-shot-CoT approach, a technique to improve LLM performance on diverse reasoning tasks, without hand-crafted few-shot examples.

Finally, we should mention the Tree of Thoughts (ToT) framework [23], which has a particularly relevant approach for QA, namely the Probabilistic Tree-of-thought Reasoning (ProbTree) [24]. This approach breaks down QA into two stages, understanding and reasoning, to solve retrieval issues and prevent error propagation.

Despite the high research interest and the diversity of approaches both in RAG and CoT, there are currently no studies focusing on certified medical chatbots. Moving within the HyDE framework, we believe that we can employ CoT techniques to improve the generation of the Hypothetical Document that would be then used as the query to retrieve the certified documents.

3. Dataset

In our dataset, we have three certified sources. We have (i) 179 *informational cards*, which were created by the Obstetrician Department of the Hospital of Trento (Italy). Then we have 953 documents from (ii) UPPA, a medical webzine, and 380 documents from (iii) ISS-Salute, which is the informative website of the Istituto Superiore di Sanità - ISS (Italian National Institute of Health).

It is important to highlight that the dataset we have is not conversational, nor it is meant to be used in a medical chatbot. All sources are what we might call content made for *FAQ sections*. Therefore, it is often quite verbose and dense in information. All the data we have is unstructured text, with a notable stylistic heterogeneity within the same source. This characteristic is combined with the semantic homogeneity given by the specific medical domain, creating a substantial issue for automatic topic extraction.

Finally, we should recall that content editing is not permitted due to the certified nature of our information. Since each specific question should consistently correspond to a particular set of equivalent answers., it becomes essential the adoption of modular RAG solutions.

4. Methods

In this section, we will explain the methods used in our implementation. Our first implementation was a sort of *zero-shot implementation* since we generated the HyDoc only relying on LLM knowledge, without providing any other context. This solution is shown in Figure 1. We assessed the performance of this first implementation by doing a user evaluation. The technology presented in this section is the same used for the second implementation illustrated in Section 5.

In this work, we used GPT-4-turbo (gpt-4-0125-preview specifically) as LLM. However, our pipeline is intended as LLM-agnostic. The use of OpenAI-GPT has, therefore, been intended as a convenient solution to test our RAG pipeline using a stable and well-performing LLM. Indeed, to deploy a conversational assistant in a real-case scenario, an open-source model would likely be required due to cost and privacy issues in accessing any LLM via API.

4.1. A first (zero-shot) implementation

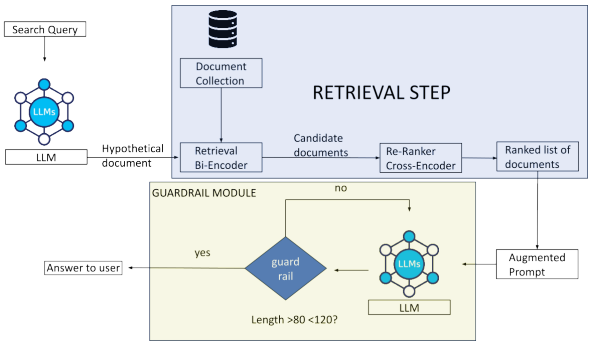

Figure 1: An overview of the RAG model we are implementing.

Our approach employs a modular RAG framework designed to address the challenge of delivering natural, verified responses through a medical chatbot by leveraging unstructured data. To achieve this, we create a HyDoc in response to the user's questions.

The essence of our strategy lies in enhancing the document retrieval process with the HyDoc. Despite the potential for inaccuracies and hallucinations, the LLM is expected to discern the fundamental aspects of the query and identify textual patterns pertinent to the specific domain of knowledge. Given the proven efficacy of LLMs in fielding medical queries [3, 4, 5], the HyDoc is anticipated to closely align with genuine documents that provide accurate, verified responses to the user's question.

To query our verified document repository, we utilize the sentence embeddings generated from our HyDoc. The area of general-purpose sentence embeddings remains an active field of research [25], in contrast to the more established universal word embedding techniques like word2vec [26]. Our workflow incorporates the *paraphrase-multilingual-mpnet-base-v2* Bi-Encoder model [27] for generating embeddings of both the HyDoc and the verified data.

This model introduces a pooling operation to produce a fixed-size embedding vector normalized to a size of 1.00. These vectors are then compared using cosine similarity. However, the Bi-Encoder model encounters challenges in accurately comparing documents of varying lengths, which can lead to the retrieval of irrelevant documents due to the disparity in length between our HyDocs and the documents in the repository.

To address this issue, we employ the *ms-marco-MiniLM-L-6-v2* cross encoder [1]. Unlike the Bi-Encoder which uses separate encoders for each input, the cross-encoder processes pairs of sentences through a single shared encoder,

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

producing a joint representation that is evaluated by a classifier to yield a similarity score between the texts.

Given the computational demands of the Cross-Encoder, it is applied selectively to a shortlist of potential documents. Following the computation of cosine similarity across all HyDoc-document pairs $<HyDoc, D_i>$, where i ranges from 1 to n and D_i represents the i_{th} document in the verified repository, we rank and select the top 50 documents for their relevance. This guarantees to have an acceptable number of documents from an information retrieval perspective [28]. Subsequently, the top 3 documents from this refined list are chosen to augment the original prompt, enhancing the text of the final response provided to the user. This decision is based on preliminary tests indicating that using more than three documents could negatively impact the framework's effectiveness.

Finally, a *Guard-Rail* module [2] is implemented to ensure the response generated by the LLM adheres to the specified prompt length, incorporating generated text and references to the three selected certified documents in the final answer.

An initial user evaluation of our zero-shot model was conducted using 100 questions related to pregnancy, deemed representative by expert reviewers. This evaluation focused on seven metrics: {Q1} the relevance of the answer to the question, {Q2} the relevance of the links (documents) provided, {Q3} text quality, {Q4} reliability, {Q5} clarity, {Q6} completeness, and {Q7} an overall evaluation score. According to Table 1, while the model demonstrated potential in text quality, it highlighted the need for improved document retrieval, as evidenced by the document link relevance scoring an average of 0.44. This value demonstrates that there is still room for improvement, but on average, half of the documents included in the links sent to the users have been considered fully relevant.

Table 1
The results of the first user evaluation. All metrics are Likert scales with a range of 1 to 5 except {Q1}, which is a binary metric (1 for positive, zero for negative), and {Q2} which is a precision score calculated on the three links

Evaluation Criterion	Avg	Max	Min	Var
{Q1} Relevance to question	0.93	1.00	0.50	0.02
{Q2} Links relevance	0.44	1.00	0.00	0.05
{Q3} Text quality	4.59	5.00	3.33	0.06
{Q4} Reliability	3.79	4.75	2.33	0.40
{Q5} Clarity	4.60	5.00	3.33	0.05
{Q6} Completeness	3.38	4.75	1.33	0.81
{Q7} Overall evaluation	3.40	4.75	1.67	0.59

5. Towards a CoT pipeline

As shown in Section 4, our first implementation has substantial room for improvement in the retrieval step. In particular, we noticed a decline in the link relevance evaluation regarding a particular type of question, i.e., evaluative questions. Evaluative questions are quite common in the medical domain and they represent the 23% of the dataset within the user evaluation we performed. In a nutshell, they are inquiries that need direct feedback on a particular aspect (e.g., "Why I am feeling so tired?"). In this case, the average link relevance is 0.31, whereas non-evaluative questions have a 0.48 average link relevance.

We argue that the worse performance on evaluative questions is mostly because generating an evaluative answer might be complex for the LLM also. Moreover, the HyDoc generate would likely be a punctual reply on the precise aspect, since this is the expected natural reply in a conversation. Since we are retrieving full documents, it might be that the vector representation of an evaluative HyDoc is quite distant from the original document where we can find the reply.

Therefore, we are annotating our dataset to enable the retrieval of shorter text segments. The idea is that we can split our documents into shorter and more meaningful segments to ease the retrieval step and enhance the generation part.

A second version of our pipeline has been tested on the subset of evaluative questions (Figure 2). The new pipeline is inspired by a CoT logic and, therefore, is aimed at generating a better HyDoc. First, we generate the HyDoc after a naive-RAG step. In a pre-retrieval step, the user question is hence used to query our certified repository, and the retrieved context is used to generate the HyDoc. Moreover, we also include more contextual information about the query aimed at enhancing the similarity between the HyDoc and the contexts that need to be retrieved in the augmented prompt. For instance, we provide within the prompt useful pragmatic information to generate an evaluative reply, such as presupposition and implications [29].

The CoT has proven to be capable of enhancing the quality of the generated HyDoc. Moreover, it has shown the ability to increase the semantic similarity between the HyDoc and the relevant documents to retrieve. This comparison considers the relevant textual segments containing the pertinent information using the *paraphrase-multilingual-mpnet-base-v2* Bi-Encoder.

In the naive-RAG step, we employ a Chroma vector database. We experimented three different embedders, namely the two OpenAI models *text-embedding-3-small* (hereafter GPT-small), *text-embedding-3-large* (hereafter GPT-large), and the Bi-Encoder model used for the document retrieval module. As shown in Table 2, using CoT prompting generated a better HyDoc with OpenAI embeddings, while it seems not influential for the Bi-Encoder model. Even though the increase in cosine similarity is small we should recall that our documents share a considerable degree of semantic similarity. Consequently, this leads to a densely populated vector space, where even marginal enhancements in similarity can yield substantial benefits in the retrieval process. Anyhow, the naive-RAG step effectively enhances HyDoc similarity both using GPT-large and in the Bi-Encoder embeddings.

Finally, the last step of the pipeline uses the HyDoc, the query context and the retrieved certified context to generate the reply. This provides the user with an appropriately framed answer as well as the documents involved in the generation process.

6. Conclusions

We have presented a modular RAG approach that enables the delivery of certified medical information. The modular pipeline allowed us to operate on unstructured texts with limited data annotation possibilities. A first user evaluation

[2] Refer to Mangaokar et al. https://arxiv.org/abs/2402.15911 for an example

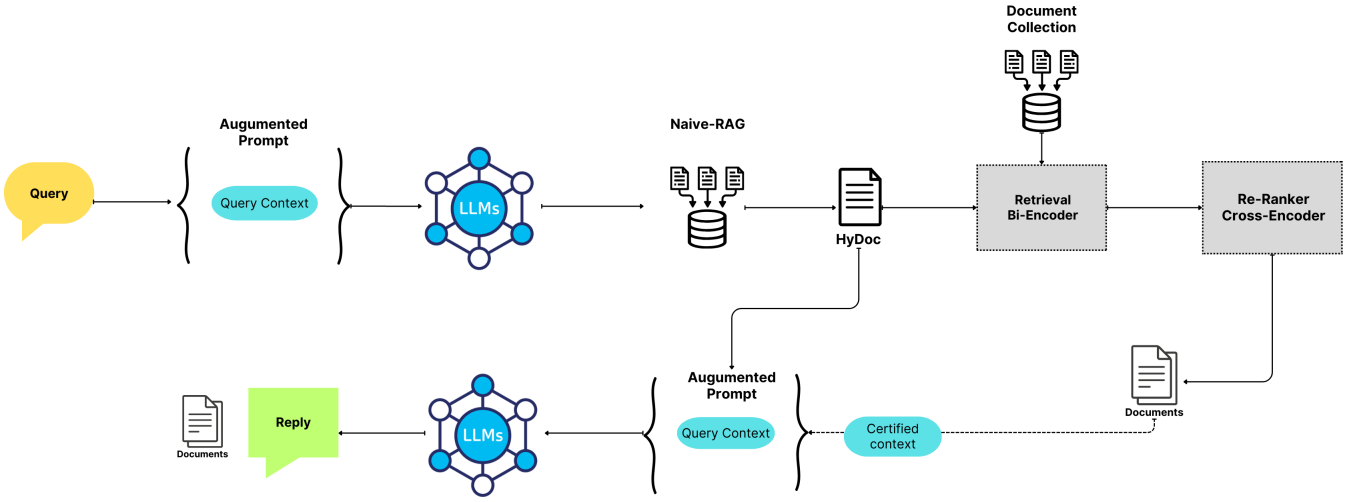

Figure 2: The proposed CoT pipeline

Table 2
The average cosine similarity between the HyDoc and the actual certified context in the "Evaluative Questions" subset

Prompt	GPT-small	GPT-large	Bi-encoder
Question + Context + Naive-RAG	0.766	0.820	0.801
Question + Naive-RAG	0.736	0.806	0.807
Question	0.717	0.717	0.717

showed promising results for our approach, although it revealed some flaws in some specific types of questions, namely evaluative questions.

We therefore tested a CoT pipeline on this specific subtype of questions, to overcome the limitations showed in the user evaluation. This approach proved to have a positive impact on the retrieval modules, enhancing semantic similarity between the HyDoc and the certified contexts, as well as on textual generation.

Surely, we should consider that we tested the CoT pipeline on a rather small dataset and that we used OpenAI-GPT as a readily available state-of-the-art LLM. Our research efforts are currently focusing on expanding the dataset and testing different open-source LLMs, as we intend our pipeline as completely LLM-agnostic.

Finally, we should also recall that in this work we presented a user evaluation and the analysis of its results. Further work is needed to create a ground truth on a comprehensive dataset of questions to assess the performance of the retrieval modules.

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] 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.

[2] W. Saeed, C. Omlin, Explainable ai (xai): A systematic meta-survey of current challenges and future opportunities, Knowledge-Based Systems 263 (2023) 110273.

[3] 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.

[4] 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.

[5] 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.

[6] K. T. Pham, A. Nabizadeh, S. Selek, Artificial intelligence and chatbots in psychiatry, Psychiatr Q 93 (2022) 249–253. URL: https://doi.org/10.1007/s11126-022-09973-8. doi:10.1007/s11126-022-09973-8, received

26 September 2021, Revised 23 January 2022, Accepted 26 January 2022, Published 25 February 2022.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] Z. Shao, Y. Gong, Y. Shen, M. Huang, N. Duan, W. Chen, Enhancing retrieval-augmented large language models with iterative retrieval-generation synergy, in: H. Bouamor, J. Pino, K. Bali (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2023, Association for Computational Linguistics, Singapore, 2023, pp. 9248–9274. URL: https://aclanthology.org/2023.findings-emnlp.620. doi:10.18653/v1/2023.findings-emnlp.620.

[12] W. Huang, M. Lapata, P. Vougiouklis, N. Papasarantopoulos, J. Z. Pan, 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), Association for Computational Linguistics, 2023, pp. 1012–1025.

[13] Z. Jiang, F. Xu, L. Gao, Z. Sun, Q. Liu, J. Dwivedi-Yu, Y. Yang, J. Callan, G. Neubig, Active retrieval augmented generation, in: H. Bouamor, J. Pino, K. Bali (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Singapore, 2023, pp. 7969–7992. URL: https://aclanthology.org/2023.emnlp-main.495. doi:10.18653/v1/2023.emnlp-main.495.

[14] Z. Yu, C. Xiong, S. Yu, Z. Liu, 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, Association for Computational Linguistics, 2023, pp. 2421–2436.

[15] J. Baek, S. Jeong, M. Kang, J. Park, S. Hwang, Knowledge-augmented language model verification, in: H. Bouamor, J. Pino, K. Bali (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Singapore, 2023, pp. 1720–1736. URL: https://aclanthology.org/2023.emnlp-main.107. doi:10.18653/v1/2023.emnlp-main.107.

[16] 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.

[17] 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.

[18] 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).

[19] 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).

[20] 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.

[21] X. Wang, J. Wei, D. Schuurmans, Q. Le, E. H. Chi, S. Narang, A. Chowdhery, D. Zhou, Self-consistency improves chain of thought reasoning in language models, in: International Conference on Learning Representations (ICLR), Google Research, Brain Team, 2023.

[22] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, Y. Iwasawa, Large language models are zero-shot reasoners, in: The U 36th Conference on Neural Information Processing Systems (NeurIPS 2022), 2022.

[23] S. Yao, D. Yu, J. Zhao, I. Shafran, T. Griffiths, Y. Cao, K. Narasimhan, Tree of thoughts: Deliberate problem solving with large language models, in: A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, S. Levine (Eds.), Advances in Neural Information Processing Systems, volume 36, Curran Associates, Inc., 2023, pp. 11809–11822. URL: https://proceedings.neurips.cc/paper_files/paper/2023/file/271db9922b8d1f4dd7aaef84ed5ac703-Paper-Conference.pdf.

[24] S. Cao, J. Zhang, J. Shi, X. Lv, Z. Yao, Q. Tian, J. Li, L. Hou, Probabilistic tree-of-thought reasoning for answering knowledge-intensive complex questions, in: Findings of the Association for Computational Linguistics: EMNLP, Association for Computational Linguistics, Beijing, China, 2023, pp. 12541–12560.

[25] R. Li, X. Zhao, M. Moens, A brief overview of universal sentence representation methods: A linguistic view, ACM Comput. Surv. 55 (2023) 56:1–56:42. URL: https://doi.org/10.1145/3482853. doi:10.1145/3482853.

[26] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, Advances in neural information processing systems 26 (2013).

[27] 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.

[28] H. Li, Learning to rank for information retrieval and natural language processing, Springer Nature, 2022.

[29] H. P. Grice, Logic and conversation, in: Speech acts, Brill, 1975, pp. 41–58.