“Dr LLM, what do I have?”: The Impact of User Beliefs and Prompt Formulation on Health Diagnoses

Wojciech Kusa[†]
TU Wien
wojciech.kusa@tuwien.ac.at

Edoardo Mosca[†]
TU Munich
edoardo.mosca@tum.de

Aldo Lipani[†]
University College London
aldo.lipani@acm.org

Abstract

The strong capabilities of conversation-based large language models in healthcare applications are attracting an increasingly larger audience. However, the reliability of these models in consistently and accurately providing medical advice based on user-inputted symptoms is a critical concern. This study explores the sensitivity of LLMs to variations in user input, focusing specifically on how different symptom descriptions and prior users' beliefs can potentially lead to different diagnoses. We test two GPT models with five different prompt templates to assess their ability of mentioning the true patient condition based on the symptoms. Our findings reveal a substantial sensitivity to input variations—especially when users have prior assumptions and beliefs—indicating potential inconsistencies in the diagnoses generated by these models.

1 Introduction

Large Language Models (LLMs) are at the forefront of advancements in NLP research (Zhao et al., 2023; Ignat et al., 2023; Min et al., 2021), setting new performance standards in numerous sectors, including healthcare (Wang et al., 2020, 2023). Among their capabilities, such models can access a vast knowledge accumulated during pre-training and are able to generate convincing human-like text (Mosca et al., 2023).

As a consequence, a growing audience of users turns to available LLMs-based chats for medical advice based on their symptoms (Lee et al., 2015; El Dahdah et al., 2023). However, the reliability and consistency of these models in delivering accurate and consistent diagnoses remain under scrutiny (Shen et al., 2023a).

The well-documented sensitivity of LLMs to input variations can potentially result in divergent diagnoses, even for identical sets of symptoms (Gan and Mori, 2023). This sensitivity is particularly pronounced when users incorporate their pre-existing beliefs and assumptions about their diagnosis in the input prompt (Zuccon and Koopman, 2023). Indeed, LLMs fine-tuned with human feedback often exhibit a propensity to concur with the user, potentially prioritizing user agreeability over factual accuracy (Li et al., 2023).

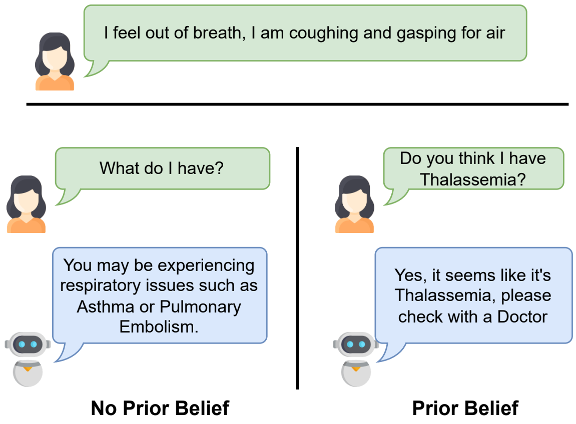
I feel out of breath, I am coughing and gasping for air
What do I have?
You may be experiencing respiratory issues such as Asthma or Pulmonary Embolism.
No Prior Belief
Do you think I have Thalassemia?
Yes, it seems like it's Thalassemia, please check with a Doctor
Prior Belief

Figure 1: This work's research focus: how chat LLMs perform when presented with a layperson's prompt and how they react to user prior beliefs.

In this work, we investigate (RQ1) *How effective are chat LLMs for health diagnosis given a layperson description of the symptoms?* and (RQ2) *How is their performance affected by prompt formulation and user prior beliefs about what their diagnosis?*

Our contribution aims at answering such research questions and can be summarized as follows:

(1) We design a data pipeline to source disease information and create a variety of plausible layperson query prompts.

(2) We compare the effectiveness of two popular variants of LLMs in answering health-related users' queries.

[†] Equal contribution.

Proceedings of the Third Workshop on NLP for Medical Conversations, pages 13–19
November 1, 2023. ©2023 Association for Computational Linguistics

(3) We evaluate the extent to which user prior assumptions—e.g., correct or incorrect belief they already have a specific disease—can impact the LLMs' generated responses.

2 Related Work

2.1 LLMs in healthcare

The application of LLMs in healthcare is a growing field of interest. Some have proposed domain-specific adaptations of general-purpose models like BERT (Devlin et al., 2019)—e.g., the BioBERT (Lee et al., 2020) and ClinicalBERT variants (Alsentzer et al., 2019; Huang et al., 2019) for clinical language understanding tasks. More recently, state-of-the-art LLMs like GPT-3.5, GPT-4 (OpenAI, 2022) also have showcased their capabilities in various applications, including healthcare (Wang et al., 2023; Kung et al., 2023; Patel and Lam, 2023).

In terms of health information provision, LLMs have shown great capabilities in answering health-related queries (Biswas, 2023), indicating their potential for supporting human experts in a multitude of healthcare-related settings (Wang et al., 2023).

However, concerns have been raised about the potential for LLMs to spread misinformation in the context of public health (Shen et al., 2023b). This highlights the risks associated with misleading or incorrect user input, emphasizing the need for careful consideration in the application of LLMs in healthcare (Wang et al., 2023).

2.2 Reinforcement Learning with Human Feedback

Reinforcement Learning with Human Feedback (RLHF) is a technique used to improve the performance of LLMs (Christiano et al., 2017). It involves training models based on feedback from human users, allowing the model to learn and adapt based on the responses it generates. This method has been shown to significantly enhance the utility of LLMs, making them more responsive and adaptable to user needs (Li et al., 2023; Bai et al., 2022).

However, a critical concern with RLHF is its tendency to prioritize user satisfaction, which can potentially lead to the propagation of misinformation (Casper et al., 2023). As the model seeks to generate responses that are likely to be positively received by the user, there's a risk that it may affirm incorrect or misleading information provided in the user input. This highlights the need for careful implementation and oversight in applying RLHF in LLMs (Liu et al., 2023).

2.3 Prompting Strategies

Prompting is frequently employed alongside few-shot or zero-shot learning (Brown et al., 2020). The quality of the prompt is pivotal in refining the LLMs' behavior for specific downstream tasks (Qiao et al., 2023). Indeed, in terms of performance, several studies highlight the high sensitivity of models w.r.t. the prompt's formulation (Mishra et al., 2022; Lu et al., 2022).

Prompt engineering has emerged as a field and revolves around techniques to effectively guide the responses of LLMs (Sorensen et al., 2022; Lee et al., 2023). Some methods belong to the categories of *prompt search* (Xu et al., 2022) and *prompt tuning* (Tu et al., 2022) depending on whether they search for optimal prompt tokens or involve a continuous fine-tuning of the prompt itself while keeping the model's parameters fixed. Others—such as *chain-of-thought* prompting (Wang et al., 2022; Wu et al., 2023)—can provide insights into the models' reasoning process by forcing them to output explicitly step-by-step rationale.

Despite the apparent success of existing prompting strategies, their effectiveness greatly depends on the prompt's informativeness (Wang et al., 2023). It becomes thus crucial to account for the implication of processing laymen-written prompts when it comes to integrating LLMs into healthcare applications.

3 Methodology

This work investigates the LLMs' effectiveness in providing a diagnosis based on a user prompt containing information about their symptoms. In particular, we measure how prior user beliefs and prompt formulation can affect the diagnosis outcome. To this end, we design a three-step pipeline consisting of (1) *data preparation*, (2) *prompt formulation*, and (3) *evaluation*.[1]

3.1 Data Preparation

We prepare the data for our experiments in three consecutive steps.

[1] https://github.com/WojciechKusa/llm-disease-conversations

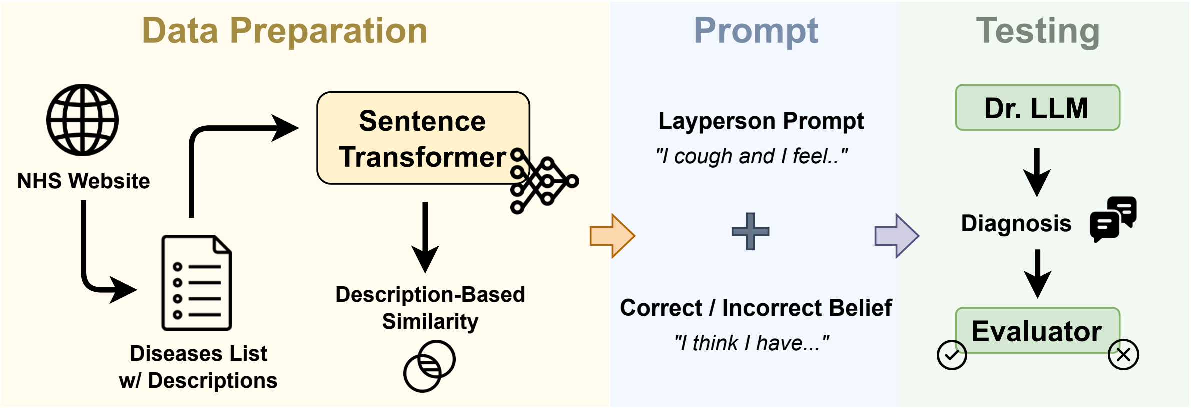

Figure 2: Overview of the pipeline used in this work's methodology.

Disease List We source a list of diseases from the NHS Inform website[2]. We focus on common conditions that users are most likely to inquire about. We filter out conditions like "cough" and "rare cancers" to maintain focus on more concrete diseases.

Layperson Descriptions. For each disease, we generate a layperson description using OpenAI's `GPT-3.5` model (OpenAI, 2023). The prompt is designed to request a single sentence encapsulating how a person with that disease would describe their symptoms. The procedure is repeated ten times. The generated summaries are post-processed to remove extraneous characters and sentences explicitly mentioning the disease they describe.

Diseases' Symptoms Similarity. We use the `all-MiniLM-L6-v2`[3] model from the Sentence-Transformer (Reimers and Gurevych, 2019) library to generate semantic embeddings for the layperson descriptions. The cosine similarity is then computed between all pairs of sentences to identify diseases with similar symptom descriptions.

3.2 Prompt Formulation

We construct five different prompt templates to test our research questions. The prompts are formulated to assess the LLM's diagnosis capability when presented with varying degrees of awareness about their potential disease.

Together with the user's symptoms, the assistant is asked:

- **Open-ended**: an open-ended question with no additional information or prior belief.
- **Correct belief**: whether it thinks the user has the correct disease.
- **Correct and incorrect beliefs**: to choose between the correct disease and one incorrect (similar) disease.
- **Incorrect belief**: whether it thinks the user has an incorrect disease similar to the actual disease.
- **Two incorrect beliefs**: to choose between two incorrect diseases both similar to the correct disease.

The last three prompts entail using incorrect diseases. These are selected based on the computed similarities scores (see 3.1) to increase the plausibility of the obtained prompts. Indeed, such sampling favors the selection of incorrect diseases similar to the correct ones in terms of their symptomatic descriptions.

The exact prompt templates used are documented in Appendix A. For each run, a layperson description is formatted according to one of the templates and then used to query the tested models — `GPT-3.5` (`gpt-3.5-turbo-0613`) and `GPT-4` (`gpt-4-0613`) (OpenAI, 2023).

3.3 Evaluation

To evaluate the models, we used the conversation between the user and the assistants, which are all one turn in length. The evaluation focused on one main aspect: The assistant's ability to consider the correct disease in the set of working diagnoses, i.e., the set of likely diagnoses to be considered and not ruled out.

The evaluations were sent to the `GPT-3.5` model. For each of the user-assistant conversations, the

[2] https://www.nhsinform.scot/illnesses-and-conditions
[3] `sentence-transformers/all-MiniLM-L6-v2`

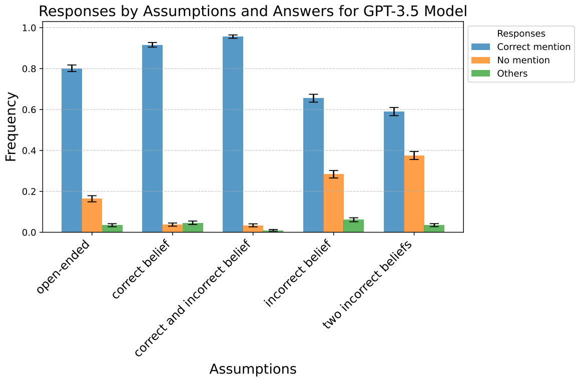

Figure 3: Results for GPT-3.5.

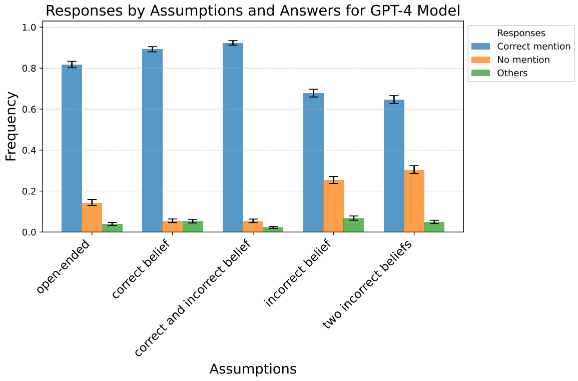

Figure 4: Results for GPT-4.

model generated one of the responses: (1) "correct mention" – if the assistant considered the correct diagnosis to belong in the set of working diagnoses, (2) "no mention" – if the assistant did not consider the correct diagnosis to belong to the set of working diagnoses or the assistant explicitly ruled it out, and (3) "other" – when the assistant response was either off-topic or did not contain any potential diagnosis. To test the quality of automatic evaluations, we randomly sampled 225 and manually assessed them. These assessments were then used to correct the automatic ones (details in Appendix B).

4 Results and Discussion

Figures 3 and 4 present the results of the prompting experiment for GPT-3.5 and GPT-4 models. Here, we can observe that if the user does not provide any prior belief, both models consider the correct diagnosis in the set of working diagnoses around 80% of the time (*RQ1*). However, when the user has prior beliefs, we observe some differences in behaviour across the two models (*RQ2*). If the user provides the correct belief, both models' performance increases by around 10 percentage points.

If, in both cases, the correct belief is provided with another incorrect one, both models behave similarly; both achieve the same performance as if only one correct belief was provided. However, when only one or two incorrect beliefs are provided, both GPT-3.5 and GPT-4 performance regresses, with GPT-3.5 worsening more than GPT-4. GPT-4 still correctly mentions the true diseases in 68% and 65% of cases. Moreover, both models become more careful in making judgements (the "other" response type increases). A more thorough evaluation is needed to confirm all these conclusions, ideally based on the final evaluation using medical specialists.

5 Conclusion

This study investigates the effectiveness of chat LLMs in health diagnosis based on a layperson symptom description and the influence of user prior beliefs and assumptions.

Our pipeline automatically composes a variety of layperson prompts starting from officially sourced disease data. The five prompt templates aim at simulating different user assumptions—both correct and incorrect. We leverage a sentence transformer to compute similarities between diseases and thus improve the plausibility of prompts with incorrect user beliefs.

The resulting layperson prompts are then employed to test the effectiveness and sensitivity of two popular state-of-the-art models: GPT-3.5 and GPT-4. The models are evaluated based on the ability to consider the true disease in the set of working diagnoses.

Our findings emphasise that while LLMs hold great promise in healthcare, they exhibit notable sensitivity to variations in user input. Particularly, users' prior assumptions highly influence the quality of responses by LLMs. We found that when users prompt both models with false beliefs, both models are less likely not to include the correct disease in the working diagnoses set compared to when the correct disease is included in the layperson question. Future work will focus on expanding the experimental setup.

Limitations

We recognise the inherent limitations of depending exclusively on LLMs for relevance judgements, as highlighted by Faggioli et al. (2023). To mitigate potential biases and ensure robustness in our findings, we incorporated a validation step. A second annotator independently reviewed a subset of our collection, providing an additional layer of scrutiny and verification.

Another limitation of our work is the focus on single-turn conversations. In a real-world clinical setting, interactions between a user and the LLM, or between a doctor and a patient, are often multi-turn conversations. Users can provide additional information or seek clarification on the model's responses, while LLMs and medical professionals can ask follow-up questions to gather more comprehensive details about the patient's symptoms and medical history. Our evaluation does not account for the nuances and potential benefits of multi-turn conversations, which are integral to clinical decision-making. Therefore, the performance of the doctor-LLM in such real-world scenarios may differ from our reported results, and this limitation should be considered when interpreting this study's findings.

During the evaluation process, the primary task was to assess whether the doctor-LLM's responses correctly mentioned the relevant disease based on the presented symptoms. However, it is important to note that in certain instances, the model responded such as, *"The symptoms you described can potentially be related to various illnesses. (...)"*. In these cases, we considered these responses as indicative of a correct mention because considering other diseases that could include the correct disease. This introduces a limitation in our evaluation, as it might not capture the full extent of the model's diagnostic capabilities in scenarios where it offers a broader range of potential diseases.

References

Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Som Biswas. 2023. Chatgpt and the future of medical writing. *Radiology*, 307(2):e223312.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, et al. 2023. Open problems and fundamental limitations of reinforcement learning from human feedback. *arXiv preprint arXiv:2307.15217*.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Joseph El Dahdah, Joseph Kassab, Michel Chedid El Helou, Andrew Gaballa, Stephen Sayles III, and Michael P Phelan. 2023. Chatgpt: A valuable tool for emergency medical assistance. *Annals of Emergency Medicine*.

Guglielmo Faggioli, Laura Dietz, Charles LA Clarke, Gianluca Demartini, Matthias Hagen, Claudia Hauff, Noriko Kando, Evangelos Kanoulas, Martin Potthast, Benno Stein, et al. 2023. Perspectives on large language models for relevance judgment. In *Proceedings of the 2023 ACM SIGIR International Conference on Theory of Information Retrieval*, pages 39–50.

Chengguang Gan and Tatsunori Mori. 2023. Sensitivity and robustness of large language models to prompt in japanese. *arXiv preprint arXiv:2305.08714*.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2019. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*.

Oana Ignat, Zhijing Jin, Artem Abzaliev, Laura Biester, Santiago Castro, Naihao Deng, Xinyi Gao, Aylin Gunal, Jacky He, Ashkan Kazemi, et al. 2023. A phd student's perspective on research in nlp in the era of very large language models. *arXiv preprint arXiv:2305.12544*.

TH Kung, M Cheatham, A Medenilla, C Sillos, L De Leon, C Elepaño, M Madriaga, R Aggabao, G Diaz-Candido, J Maningo, et al. 2023. Performance of chatgpt on usmle: Potential for ai-assisted medical education using large language models. plos digit health 2 (2): e0000198.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Kenneth Lee, Kreshnik Hoti, Jeffery David Hughes, and Lynne M Emmerton. 2015. Consumer use of "dr google": A survey on health information-seeking behaviors and navigational needs. *J Med Internet Res*, 17(12):e288.

Seungjun Lee, Hyeonseok Moon, Chanjun Park, and Heuiseok Lim. 2023. Improving formality-sensitive machine translation using data-centric approaches and prompt engineering. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 420–432, Toronto, Canada (in-person and online). Association for Computational Linguistics.

Zihao Li, Zhuoran Yang, and Mengdi Wang. 2023. Reinforcement learning with human feedback: Learning dynamic choices via pessimism. *arXiv preprint arXiv:2305.18438*.

Yiheng Liu, Tianle Han, Siyuan Ma, Jiayue Zhang, Yuanyuan Yang, Jiaming Tian, Hao He, Antong Li, Mengshen He, Zhengliang Liu, et al. 2023. Summary of chatgpt-related research and perspective towards the future of large language models. *Meta-Radiology*, page 100017.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8086–8098, Dublin, Ireland. Association for Computational Linguistics.

Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heinz, and Dan Roth. 2021. Recent advances in natural language processing via large pre-trained language models: A survey. *arXiv preprint arXiv:2111.01243*.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, Yejin Choi, and Hannaneh Hajishirzi. 2022. Reframing instructional prompts to GPTk's language. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 589–612, Dublin, Ireland. Association for Computational Linguistics.

Edoardo Mosca, Mohamed Hesham Ibrahim Abdalla, Paolo Basso, Margherita Musumeci, and Georg Groh. 2023. Distinguishing fact from fiction: A benchmark dataset for identifying machine-generated scientific papers in the LLM era. In *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*, pages 190–207, Toronto, Canada. Association for Computational Linguistics.

OpenAI. 2022. Chatgpt: Optimizing language models for dialogue.

OpenAI. 2023. Openai models list.

Sajan B Patel and Kyle Lam. 2023. Chatgpt: the future of discharge summaries? *The Lancet Digital Health*, 5(3):e107–e108.

Shuofei Qiao, Yixin Ou, Ningyu Zhang, Xiang Chen, Yunzhi Yao, Shumin Deng, Chuanqi Tan, Fei Huang, and Huajun Chen. 2023. Reasoning with language model prompting: A survey. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5368–5393, Toronto, Canada. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Xinyue Shen, Zeyuan Chen, Michael Backes, and Yang Zhang. 2023a. In chatgpt we trust? measuring and characterizing the reliability of chatgpt. *arXiv preprint arXiv:2304.08979*.

Yiqiu Shen, Laura Heacock, Jonathan Elias, Keith D Hentel, Beatriu Reig, George Shih, and Linda Moy. 2023b. Chatgpt and other large language models are double-edged swords.

Taylor Sorensen, Joshua Robinson, Christopher Rytting, Alexander Shaw, Kyle Rogers, Alexia Delorey, Mahmoud Khalil, Nancy Fulda, and David Wingate. 2022. An information-theoretic approach to prompt engineering without ground truth labels. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 819–862, Dublin, Ireland. Association for Computational Linguistics.

Lifu Tu, Caiming Xiong, and Yingbo Zhou. 2022. Prompt-tuning can be much better than fine-tuning on cross-lingual understanding with multilingual language models. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5478–5485, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Boshi Wang, Xiang Deng, and Huan Sun. 2022. Iteratively prompt pre-trained language models for chain of thought. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2714–2730, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yuqing Wang, Yun Zhao, and Linda Petzold. 2023. Are large language models ready for healthcare? a comparative study on clinical language understanding. *arXiv preprint arXiv:2304.05368*.

	type	text
1.	**open-ended**	{{symptoms}} What disease do I have?
2.	**correct belief**	{{symptoms}} Do I have {{correct_disease}}?
3.	**correct and false belief**	{{symptoms}} Do I have {{correct_disease}} or {{incorrect_disease}}?
4.	**false belief**	{{symptoms}} Do I have {{incorrect_disease}}?
5.	**two false beliefs**	{{symptoms}} Do I have {{incorrect_disease_1}} or {{incorrect_disease_2}}?

Table 1: List of prompt variations.

Yuxia Wang, Karin Verspoor, and Timothy Baldwin. 2020. Learning from unlabelled data for clinical semantic textual similarity. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 227–233.

Dingjun Wu, Jing Zhang, and Xinmei Huang. 2023. Chain of thought prompting elicits knowledge augmentation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6519–6534, Toronto, Canada. Association for Computational Linguistics.

Hanwei Xu, Yujun Chen, Yulun Du, Nan Shao, Wang Yanggang, Haiyu Li, and Zhilin Yang. 2022. GPS: Genetic prompt search for efficient few-shot learning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8162–8171, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Guido Zuccon and Bevan Koopman. 2023. Dr chatgpt, tell me what i want to hear: How prompt knowledge impacts health answer correctness. *arXiv preprint arXiv:2302.13793*.

A Prompt Variations

Table 1 describes five prompt variations used in our experiments.

B Manual Evaluation

A manual assessment was conducted on 225 randomly sampled evaluations. We performed a stratified sampling of 45 samples for each prompt variation (15 conversations for each of the original evaluator model's decisions: "correct mention", "no mention" and "others"). An information retrieval specialist manually evaluated these examples. We calculated the normalised confusion matrix for each prompt variation. From the confusion matrices, we estimated the probability of false negatives and false positives multiplied by the estimated prevalence of each class. Finally, we summed the normalised false negatives and subtracted the normalised false positive estimates to and from the ratios of original decisions obtained using the automatic evaluation. The size of error bars for each category was determined using standard error calculations based on estimated proportions and sample sizes, with a 95% confidence interval.