Boosting Questions' Effectiveness in Medical Interviews

Davide Mazzaccara
CIMeC, University of Trento
davide.mazzaccara@unitn.it

Alberto Testoni
ILLC, University of Amsterdam
a.testoni@uva.nl

Raffaella Bernardi
CIMeC, DISI, University of Trento
raffaella.bernardi@unitn.it

Abstract

Questions are a fundamental tool for acquiring information, from children's learning to complex tasks. Recent work has shown that the informativeness of questions by large language models (LLMs) can be enhanced through Direct Preference Optimization (DPO) and Expected Information Gain (EIG). In this study, we evaluate the effectiveness of a DPO-trained model in the context of medical interviews. Our findings indicate that DPO training improves success rates in medical interviews, thereby demonstrating the broader applicability and generalizability of this approach.

1 Introduction

Questions in language serve as requests for information (Hiż, 1978). The speaker lacks information in their knowledge state and asks questions to gain this information. This process of acquiring information through questioning is essential for children to learn about the world (Ruggeri and Lombrozo, 2015) and for adults to solve complex problems (Geva et al., 2021). A complex problem is a medical interview: the doctor asks questions to elicit the patient's signs and symptoms. Once enough information has been collected, the doctor identifies the disease and proceeds with treatment.

Despite their remarkable language and reasoning abilities (Kojima et al., 2022), Large Language Models (LLMs) have been observed to generate low informative questions (Bertolazzi et al., 2023), evaluated through the 20 Questions Game and Expected Information Gain (EIG). Based on the intuition that LLMs are good at generating diverse questions and providing answers to these close-ended questions (Testoni et al., 2023), Hu et al. (2024) propose an inference time probabilistic reasoning strategy (see also Piriyakulkij et al. 2023). The authors make the LLM generate different questions via sampling, then selecting the question maximizing the EIG measure. Alternatively, Mazzaccara et al. (2024) use probabilistic reasoning to create a dataset of sampled low and high-informative questions. By training on these data with preference optimization, the authors conclude that LLMs could learn to reason with informativeness.

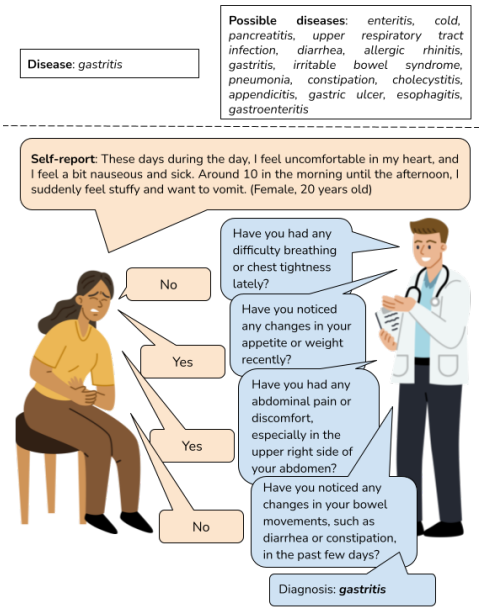

Figure 1: Example of a Medical Interview (MedDG). The dialogue is machine-generated: LLAMA 2 DPO plays the role of the doctor and GPT-3.5 the patient.

Mazzaccara et al. (2024) concludes that LLM's reasoning with informativeness generalises across different domains. Our study delves into this conclusion by testing the trained model on a different domain and task, i.e., medical interviews. Medical interviews are task-oriented dialogues, where the doctor collects information through question-answer pairs to make a diagnosis. As illustrated in Fig. 1, the doctor is provided with the possible diseases and a patient's self-report. The doctor asks questions about the patient's signs and symptoms to identify the patient's disease. LLMs, trained to ask informative questions, could assist doctors towards more efficient and effective medical interviews.

Proceedings of the 28th Workshop on the Semantics and Pragmatics of Dialogue, September, 11–12, 2024, Trento, Italy.

2 Setting

The 20 Questions Game and Medical Interviews comprise two roles, a Questioner and an Answerer. The Questioner ask yes/no questions to collect information and identify the candidate in a list of possible candidates. The Answerer guides this process, providing truthful yes/no answers. In our setting, a *game* consists of the candidate set with the target candidate; a *dialogue* is the series of question-answers exchanges. A dialogue is considered *successful* if the target is reached within the first 20 questions.

To train an LLM to ask informative questions, Mazzaccara et al. (2024) creates 20 Question games with common concepts from the following categories: mammal, bird, clothing, weapon, fruit, and vegetables. Questions are sampled from the chosen LLM, LLAMA 2-CHAT (7B), and then evaluated in terms of EIG by the same model. The resulting low and high-EIG questions are employed to tune the same LLAMA 2 with Direct Preference Optimization (Rafailov et al., 2023). Trained to ask more informative questions, the resulting model is more efficient (fewer turns to reach the target) and more effective (higher success rate) in the 20 Questions game in different domains. We compare LLAMA 2-CHAT (7B) Zero-shot and after DPO as Questioner, the Answerer is GPT-3.5-TURBO-0125.

In the task of Medical Interview, the Questioner asks yes/no questions to identify the patient's disease.[1] Medical interviews differ from the 20 Questions game in that the Questioner is initially provided with a self-report from the patient. This implies that a medical interview game comprises: self-report, possible candidate diseases, and the target disease. The self-report is provided to the Questioner before the first turn alongside the possible diseases. In our evaluation setting, we test the trained model with and without the self-report in two medical datasets.

The medical datasets employed for testing are DX (Xu et al., 2019) and MedDG (Liu et al., 2022). The English versions of the datasets are provided by Hu et al. (2024). The self-reports of both DX and MedDG have been extracted from human online doctor-patient interviews. We employed the test set of DX, consisting of 104 games with 5 possible diseases. For MedDG, we use the 10% of the selected games by Hu et al. (2024). The resulting MedDG dataset consisting of 50 games with 15 possible diseases.

[1] Simplifying our setting to yes/no questions and answers allows for easier computation of EIG, while representing a good approximation of the task

3 Results

We evaluate the training's impact on efficiency and effectiveness in medical interviews. The Average number of Questions (**AQ**) measures efficiency as the number of questions the model needs to reach the target. The Success rate at 1 (**S@1**) measures effectiveness as the percentage of times the model achieves the target within the first tentative.

The results are reported in Table 1. DPO training seems not to positively impact informativeness, as shown by lower AQ in all settings. In terms of effectiveness, instead, the DPO training leads to higher S@1 for both DX and MedDG. In DX medical interviews, the DPO outperforms the Zero-shot by an absolute difference of $+12.5\%$ in S@1 without the self-report and $+28.8\%$ S@1 with the self-report. Overall this is a rather positive result given that DPO has been trained on radically different concept domains. When comparing the same setting with and without the self-report, we see that for large candidate sets, MedDG, both Zero-Shot and DPO improve their Success rate, as one would expect; interestingly, DPO improves its efficiency more than Zero-shot (the AQ decreases -3 vs. -0.6). Maybe surprisingly, with smaller candidate sets, DX, both DPO and Zero-shot improve in efficiency, when the self-report is provided, but their success rate decreases with DPO suffering less (-5.8 vs. -22.1).

Setting	Method	DX		MedDG	
		AQ ↓	S@1 ↑	AQ ↓	S@1 ↑
w/o self-report	Zero-shot	5.5	42.3%	7.6	6.0%
	DPO	6.3	54.8%	9.9	12.0%
self-report	Zero-shot	4.4	20.2%	7.0	18.0%
	DPO	4.5	49.0%	6.9	22.0%

Table 1: Results for LLAMA 2-CHAT (7B) zero-shot and DPO in DX and MedDG. In the first row are reported the results for the setting without self-report. In the second row, with the self-report.

References

Leonardo Bertolazzi, Davide Mazzaccara, Filippo Merlo, and Raffaella Bernardi. 2023. ChatGPT's information seeking strategy: Insights from the 20-

questions game. In *Proceedings of the 16th International Natural Language Generation Conference*, pages 153–162, Prague, Czechia. Association for Computational Linguistics.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9.

Henry Hiż, editor. 1978. *Questions*. Reidel, Dordrecht/Boston.

Zhiyuan Hu, Chumin Liu, Xidong Feng, Yilun Zhao, See-Kiong Ng, Anh Tuan Luu, Junxian He, Pang Wei Koh, and Bryan Hooi. 2024. Uncertainty of thoughts: Uncertainty-aware planning enhances information seeking in large language models. In *ICLR 2024 Workshop on Large Language Model (LLM) Agents*.

Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, volume 35, pages 22199–22213. Curran Associates, Inc.

Wenge Liu, Jianheng Tang, Yi Cheng, Wenjie Li, Yefeng Zheng, and Xiaodan Liang. 2022. Meddg: An entity-centric medical consultation dataset for entity-aware medical dialogue generation. In *Natural Language Processing and Chinese Computing - 11th CCF International Conference, NLPCC 2022, Proceedings*, pages 447–459, Germany. Springer Science and Business Media Deutschland GmbH.

Davide Mazzaccara, Alberto Testoni, and Raffaella Bernardi. 2024. Learning to ask informative questions: Enhancing llms with preference optimization and expected information gain. *Preprint*, arXiv:2406.17453.

Top Piriyakulkij, Volodymyr Kuleshov, and Kevin Ellis. 2023. Asking clarifying questions using language models and probabilistic reasoning. In *NeurIPS 2023 Foundation Models for Decision Making Workshop*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, pages 53728–53741. Curran Associates, Inc.

Azzurra Ruggeri and Tania Lombrozo. 2015. Children adapt their questions to achieve efficient search. *Cognition*, 143:203–216.

Alberto Testoni, Raffaella Bernardi, and Azzurra Ruggeri. 2023. The efficiency of question-asking strategies in a real-world visual search task. *Cognitive Science*, 47(12).

Lin Xu, Qixian Zhou, Ke Gong, Xiaodan Liang, Jianheng Tang, and Liang Lin. 2019. End-to-end knowledge-routed relational dialogue system for automatic diagnosis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):7346–7353.