Applying Comparative Research Techniques in Answering Health Questions

Sajad Shirali-Shahreza
Amirkabir University
Tehran, Iran
shirali@aut.ac.ir

Mohamad Khorsandi
Amirkabir University
Tehran, Iran
mo.khorsandi@aut.ac.ir

Abstract

Many users started using LLM-based chatbots to find answers to their health questions. However, LLMs are not designed to provide accurate and detailed answers to such questions. In this position paper, we review some of the limitations of current LLM-based Q&A systems. Then we propose to use comparative research techniques, which are common in social sciences, in the process of answering health-related questions. We provide a high-level overview of how comparative research techniques can be coupled with current best-practices of building LLM-based systems.

ACM Reference Format:
Sajad Shirali-Shahreza and Mohamad Khorsandi. 2026. Applying Comparative Research Techniques in Answering Health Questions. In *Proceedings of Workshop: Talk Healthy to Me at CUI 2026 (CUI 2026 Workshop: Talk Healthy to Me).* ACM, New York, NY, USA, 2 pages.

1 Introduction

Healthy being is of great importance for the majority of users. Therefore, they are looking for the correct answer to their health questions.

For many of the health questions, there is not a single correct answer. Usually, there are multiple different answers that are neither completely correct nor incorrect. They usually look at the problem from a different perspective or make special assumptions about it. This is similar to different treatments available for a single disease. Different doctors may use different medications to treat it. This usually depends on the specific details of the patient's health history. But for an outside observer, they see a group of patients with a common disease being treated differently.

The limitations of LLMs (which we will briefly review in Section 2) raise ethical concerns about their usage in domains such as healthcare [12]. Even when an LLM offers multiple points of view, it is rational for users to remain skeptical on sensitive topics. Recent studies indicate that when LLMs hallucinate, roughly 50% of users express doubt about the output [7]. This leads us to guess that trust would even be lower in highly sensitive domains such as healthcare. For users to trust AI output in such domains, it requires LLMs to use verified sources (e.g., research papers, books, etc.) and also convince the user that each specific reason comes from a verified source and its meaning is aligned with the source.

In this paper, we propose the idea of using comparative research ideas. This can result in answers that are more comprehensive. In addition to improving the quality of the answers, it can better gain the trust of the users.

We will start with a brief review of the limitations of LLMs (Section 2), followed by a short introduction to comparative research (Section 3). Then we present our idea in Section 4. The last section will conclude the paper.

2 LLM Q&A Limitations

A major limitation of LLMs is that they are designed to create a stream of text, one token at a time. In other words, there are not knowledge-grounded reasoners [2]. Therefore, LLMs usually only output one potential answer and stick to it.

Previous research showed that LLMs usually over-estimate the accuracy of their answer. This results in being over-confidence in their answer [16].

This over-confidence combined with being primarily designed to just provide one answer results in suppression of other potentially correct answers. The results of evaluating on the religious Q&A domain in [2] show how this leads to really high incorrect answers. The same thing is happening in the health domain.

The first generation of LLMs only outputs their answer without any direct reference to the sources. Some of the more recent models were explicitly designed to provide the sources that they are using and reference them. However, one of their limitations is how they present their sources. Recent studies such as [14] show that LLMs routinely refer to sources to support their conclusions that are not actually supporting that. Furthermore, sometimes the title of these sources may also fool the reader. Other researchers reported similar results. For example, the factual accuracy of the strongest frontier models is only 39-77% [9]. This is separate from the hallucination problem of LLMs [6] in which they generate sources that do not exist.

3 Comparative Research

It is common in exact sciences such as math (or even to some degree physics) to assume that each question only has a single correct answer. This is usually true due to the exact modeling of the problem and assumptions about the domain. In contrast, many of the problems in the humanities and social sciences have some ambiguity inside them. This results in having different proposed answers for them. Usually, it is not possible to directly compare such answers together and conclude which one is correct or wrong. Therefore, it is common in the humanities to present different previously proposed answers for a specific question and then to try to compare and analyze them. This is usually referred to as "Comparative Research".

CUI 2026 Workshop: Talk Healthy to Me, Bremen, Germany

An example of social sciences using comparative research is religious studies. A famous example is the al-Milal wa al-Nihal [1] book in Islamic studies written 9 centuries ago. The comparative nature of the book makes it one of the first books that objectively and systematically compares different religious sects [4].

There are various follow-up works in the religious domain. An important example that was written in the last century is the "al-Uṣul al-ammah lil-Fiqh al-Muqaran" [5] book.

These techniques (which are usually simply referred to as comparative research) are widely used in different science fields. Furthermore, other researchers also point out that cognitive and social sciences insights are rarely used in the design and evaluation of LLM-based systems [11]. Our idea is another example of the recent shift towards using social science techniques to enhance LLM-based systems (another recent example is [10]).

4 Our Proposed Idea

We propose using comparative research techniques to answer health questions. Instead of relying on an LLM to generate the final answer, we can first try to find all potential answers to the questions (regardless of how accurate they are). We can use the common Retrieval-Augmented Generation (RAG) for this purpose. Special methods can then be used to combine the results. An option is to use a RAG-based framework that reduces hallucinations by ensuring that the generated content is supported by evidence retrieved from a reliable knowledge base [8]. Another option is to use a multi-agent framework in which multiple LLM agents collaboratively verify and aggregate information [10].

We can start by using both traditional search methods (e.g., search engines) and multiple different LLMs (or even different agents using a common underlying LLM) to create a list of potential answers for the questions. The benefit of using search engines is that we will prevent hallucination [6] of fictional sources. The benefit of using multiple LLMs is the ability to collect potentially different answers (due to the different training materials used by such methods [2]). Furthermore, recent studies such as [13] show that human users manually do this to find different potential answers and then compare them to find the correct answer.

Then we can instruct agents to apply comparative research techniques (such as those presented in [1, 5]) to analyze and compare these potential answers. This is to complement the ideas presented by techniques such as SPIRE [10] that focus on features such as proper quotation of source materials. In the end, instead of only showing a single answer, we can present the top potential answers and highlight their similarity and differences. This will be similar to how a doctor will explain different potential treatments for a disease to a patient, compare them together, and help the patient choose the final treatment option that best matches their needs.

We think that a major benefit of this approach is better explainability of the answer provided to the user, which is a major concern for advanced LLMs [3].

Our approach is also more transparent (in terms of where the initial data come from and how the final answers are created). Increasing the transparency of LLM-based Q&A systems is one of the goals of new research in this domain (e.g., [15]).

References

[1] Muhammad ibn Abdulkarim al-Shahrastani. 1128. *Al-Milal Wa al-Nihal.* Dar al-Maarefah.

[2] Farah Atif, Nursultan Askarbekuly, Kareem Darwish, and Monojit Choudhury. 2025. Sacred or Synthetic? Evaluating LLM Reliability and Abstention for Religious Questions. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* 8, 1 (Oct. 2025), 217–226. doi:10.1609/aies.v8i1.36543

[3] Upol Ehsan, Elizabeth A Watkins, Philipp Wintersberger, Carina Manger, Sunnie S. Y. Kim, Niels Van Berkel, Andreas Riener, and Mark O Riedl. 2024. Human-Centered Explainable AI (HCXAI): Reloading Explainability in the Era of Large Language Models (LLMs). In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '24)*. Association for Computing Machinery, New York, NY, USA, 1–6. doi:10.1145/3613905.3636311

[4] Abu Haif, Andi Lifiani Ansar, Islahuddin Ibrahim, Abdurrahman Sibghatullah, Ninin Rizka Syahfitri Kaharu, and Darmawan Darmawan. 2025. The Relevance of Al-Syahrastani's Al-Milal Wa Al-Nihal Methodology to Contemporary Islamic Historiography. *JUSPI (Jurnal Sejarah Peradaban Islam)* 9, 1 (July 2025), 172–181. doi:10.30829/juspi.v9i1.24348

[5] Seyed Mohammad Taghi Hakim. 1965. *Al-Uṣūl al-'āmmah Lil-Fiqh al-Muqāran.* Dar al-Andalos.

[6] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Transactions on Information Systems* 43, 2 (Jan. 2025), 42:1–42:55. doi:10.1145/3703155

[7] Ilkka Kaate, Joni Salminen, Soon-Gyo Jung, Trang Thi Thu Xuan, Essi Häyhänen, Jinan Y Azem, and Bernard J Jansen. 2025. "You Always Get an Answer": Analyzing Users' Interaction with AI-Generated Personas Given Unanswerable Questions and Risk of Hallucination. In *Proceedings of the 30th International Conference on Intelligent User Interfaces*. 1624–1638.

[8] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems*, Vol. 33. Curran Associates, Inc., 9459–9474.

[9] Hailey Onweller, Elias Lumer, Austin Huber, Pia Ramchandani, Vamse Kumar Subbiah, and Corey Feld. 2026. Cited but Not Verified: Parsing and Evaluating Source Attribution in LLM Deep Research Agents. arXiv:2605.06635 [cs.CL] doi:10.48550/arXiv.2605.06635

[10] Yating Pan, Jiajun Zhang, Jun Wang, and Qi Su. 2026. Extending AI for Research to the Humanities: A Multi-Agent Framework for Evidence-Grounded Scholarship. arXiv:2605.30947 [cs.CL] doi:10.48550/arXiv.2605.30947

[11] Yao Rong, Tobias Leemann, Thai-Trang Nguyen, Lisa Fiedler, Peizhu Qian, Vaibhav Unhelkar, Tina Seidel, Gjergji Kasneci, and Enkelejda Kasneci. 2024. Towards Human-Centered Explainable AI: A Survey of User Studies for Model Explanations. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 46, 4 (April 2024), 2104–2122. doi:10.1109/TPAMI.2023.3331846

[12] Iqbal H. Sarker. 2024. LLM Potentiality and Awareness: A Position Paper from the Perspective of Trustworthy and Responsible AI Modeling. *Discover Artificial Intelligence* 4, 1 (May 2024), 40. doi:10.1007/s44163-024-00129-0

[13] Yingtian Shi, Jinda Yang, Yuhan Wang, Yiwen Yin, Haoyu Li, Kunyu Gao, and Chun Yu. 2025. LLMartini: Seamless and Interactive Leveraging of Multiple LLMs through Comparison and Composition. arXiv:2510.19252 [cs.HC] doi:10.48550/arXiv.2510.19252

[14] Monica Visani Scozzi, Stephann Makri, and Pranava Madhyastha. 2026. "Although Powerful, It's Not Infallible": Investigating Academic Researchers' Verification Challenges with LLMs. In *Proceedings of the 2026 Conference on Human Information Interaction and Retrieval (CHIIR '26)*. Association for Computing Machinery, New York, NY, USA, 73–83. doi:10.1145/3786304.3787865

[15] Tongshuang Wu, Michael Terry, and Carrie Jun Cai. 2022. 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 '22)*. Association for Computing Machinery, New York, NY, USA, 1–22. doi:10.1145/3491102.3517582

[16] Chenjun Xu, Bingbing Wen, Bin Han, Robert Wolfe, Lucy Lu Wang, and Bill Howe. 2025. Do Language Models Mirror Human Confidence? Exploring Psychological Insights to Address Overconfidence in LLMs. In *Findings of the Association for Computational Linguistics: ACL 2025*, Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (Eds.). Association for Computational Linguistics, Vienna, Austria, 25655–25672. doi:10.18653/v1/2025.findings-acl.1316