Open-Source LLMs Performance on Information Retrieval Tasks for Health Outcomes Research

SA1

A Livieratos[1], J Lin[1,2] , D Zhang[1,3] , A Chen[1,4] , M Kudela[1,4], Y Zhao[1,4], C Basu[1,4], S Dharmarajan[1,5], M Gamalo[1,4]

[1]SPAIML Scientific Working Group, USA; [2]Takeda Pharmaceuticals, USA; [3]Teva Pharmaceuticals, USA; [4]Pfizer, USA; [5]Sarepta Therapeutics, USA

INTRODUCTION

The application of open-source large language models (LLMs) in data extraction from medical articles has shown great promise in revolutionizing biomedical research. LLMs, specifically designed for natural language processing tasks, have gained traction in healthcare due to their ability to process large volumes of unstructured data, including clinical trials and systematic reviews.[1-2] Recent trends indicate a shift from proprietary LLMs to open-source alternatives, such as Llama-2 and LongT5, which offer the advantage of transparency and customization.[1-2] These models are increasingly being fine-tuned to optimize their performance for specialized tasks such as summarizing medical literature.[1-2] The main advantage of open-source LLMs is their flexibility, allowing researchers to tailor the models for specific use cases like extracting key findings from PubMed articles and conducting evidence-based medical reviews.

Studies have previously reported how GPT-4 can streamline systematic literature reviews (SLRs), data extraction, and medical evidence synthesis.[3-4] GPT-4's automation capabilities have proven useful in areas like health technology assessments (HTAs), where accurate and timely extraction of data is essential for developing systematic reviews and evaluating clinical effectiveness.[3-4] Despite such developments, open-source models present an appealing alternative for institutions looking for cost-effective, customizable solutions.

Studies have demonstrated that open-source LLMs can achieve performance levels comparable to commercial models when properly fine-tuned.[1-2] For example, recent evaluations revealed that models like LongT5, after fine-tuning, were able to approach the performance of GPT-3.5 in summarizing medical evidence.[1-2] Furthermore, LLMs have been effectively utilized in automating the retrieval and synthesis of biomedical data, helping mitigate the challenges posed by the vast and growing number of clinical studies.[1-2] This includes automating systematic reviews, which are essential for Health Economics and Outcomes Research (HEOR) and Medical Affairs activities.[1-4]

OBJECTIVE

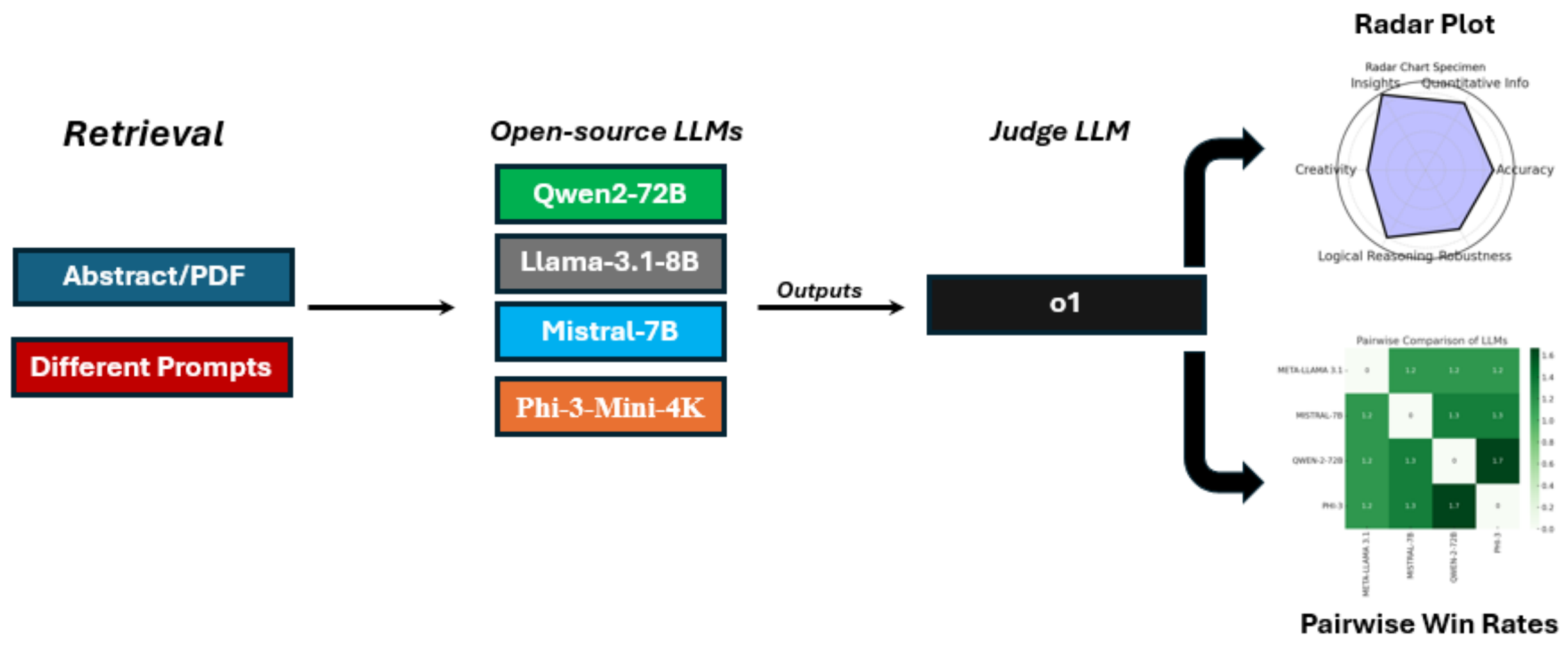
The aim was to compare the performance of several open-source LLMs, using a proprietary LLM as a judge, on zero-shot learning information retrieval tasks using clinical abstracts or full manuscripts.

METHODS

We considered four open-source LLMs in this study: Qwen2-72B, Llama-3.1-8B, Mistral-7B and Phi-3-Mini-4K. The comparative performances between LLMs were judged by the proprietary OpenAI's o1, in zero-shot learning information extraction tasks using manuscript abstracts and full manuscripts. As of November 2024, these open-source LLM versions were widely applicable. The OpenAI's o1 was newly introduced and performs complex reasoning tasks, which greatly improved over GPT-4o on different reasoning benchmarks.[5] The evaluations were conducted for three sets of information retrieval tasks: (1) open-ended prompts in 3 manuscript abstracts,[6-8] (2) open-ended prompts in 3 full manuscripts,[8-10] and (3) narrow-specific prompts in 3 full manuscripts.[8-10] The manuscript abstracts and full manuscripts were immunology publications obtained through PubMed.

A simplified Fine-grained Language Model Evaluation based on Alignment Skill Sets (FLASK) was adopted to assess the accuracy and reliability of the data extraction across six metrics: Accuracy, Robustness, Creativity, Insights, Quantitative Information and Logical Reasoning.[11-12] These metrics evaluate the LLM outputs that align with the PICO (population, intervention, comparator, and outcomes) framework.[2,4] LLM performances were evaluated via pair-wise output comparisons, and overall win-rates were generated across the different information retrieval scenarios previously outlined.

Figure 1: Methodology for data extraction and performance evaluation



RESULTS

Based on OpenAI's o1 assessment of the four LLMs for the informational retrieval task of (1), open-ended prompts in 3 manuscript abstracts, we hereby report that results varied somewhat based on the abstract. For one abstract, Phi-3 performed best during pair-wise comparison, while Qwen2-72B was the best performer overall, as illustrated in Figure 2. Specifically, Qwen2-72B reported an overall win-rate of >50% versus the other three models regarding information retrieval from abstracts. For the second task of (2) information retrieval using open-ended prompts for 3 full manuscripts, Qwen2-72B reported the best overall win-rates compared to the other LLMs, as illustrated in Figure 3. Finally, for the third extraction task based on (3), using narrow-specific prompts in 3 full manuscripts, there were no obvious performance differences between the models, as illustrated in Figure 4. Representative performance radar plots are presented in Figure 5 to demonstrate the methodological evaluation approach applied across all scenarios and documents.

Overall, although no significant differences were reported between models, we hereby highlight potential emerging trends in performance divergence between models under specific scenarios (document size and prompt engineering). Specifically, for information retrieval from full manuscripts, open-ended prompts resulted in greater performance variations versus narrow-specific prompts, an observation to consider for HEOR data extraction best-practices. The impact of prompt engineering is also replicated when multiple runs of the same task are performed. This work highlights performance variations of open-source LLMs for HEOR specific tasks and aims to establish an evaluation framework for utilizing such models in health outcomes research.

Figure 2: Overall Win-Rates for Information Retrieval Task 1 (open-ended prompts across 3 abstracts)

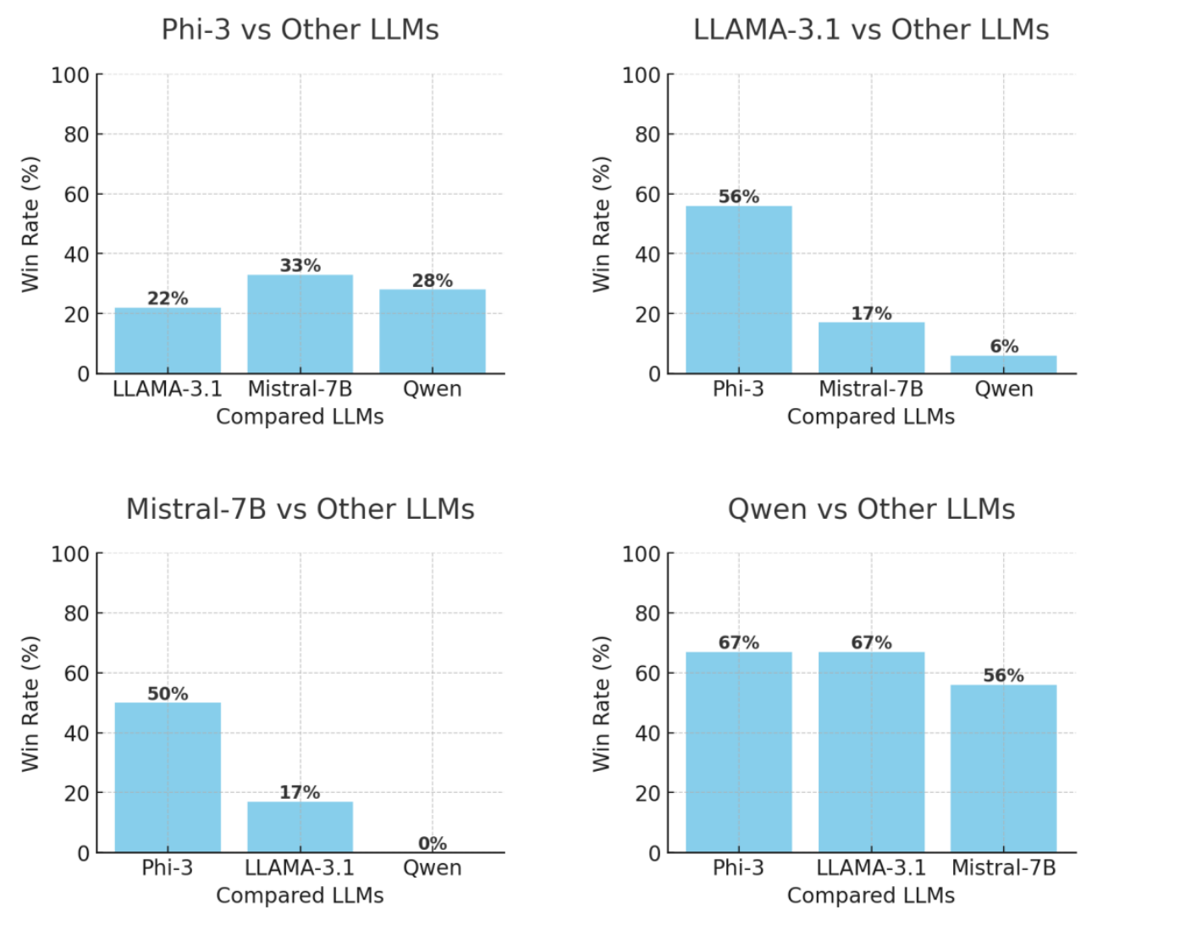


Figure 3: Overall Win-Rates for Information Retrieval Task 2 (open-ended prompts across 3 manuscripts)

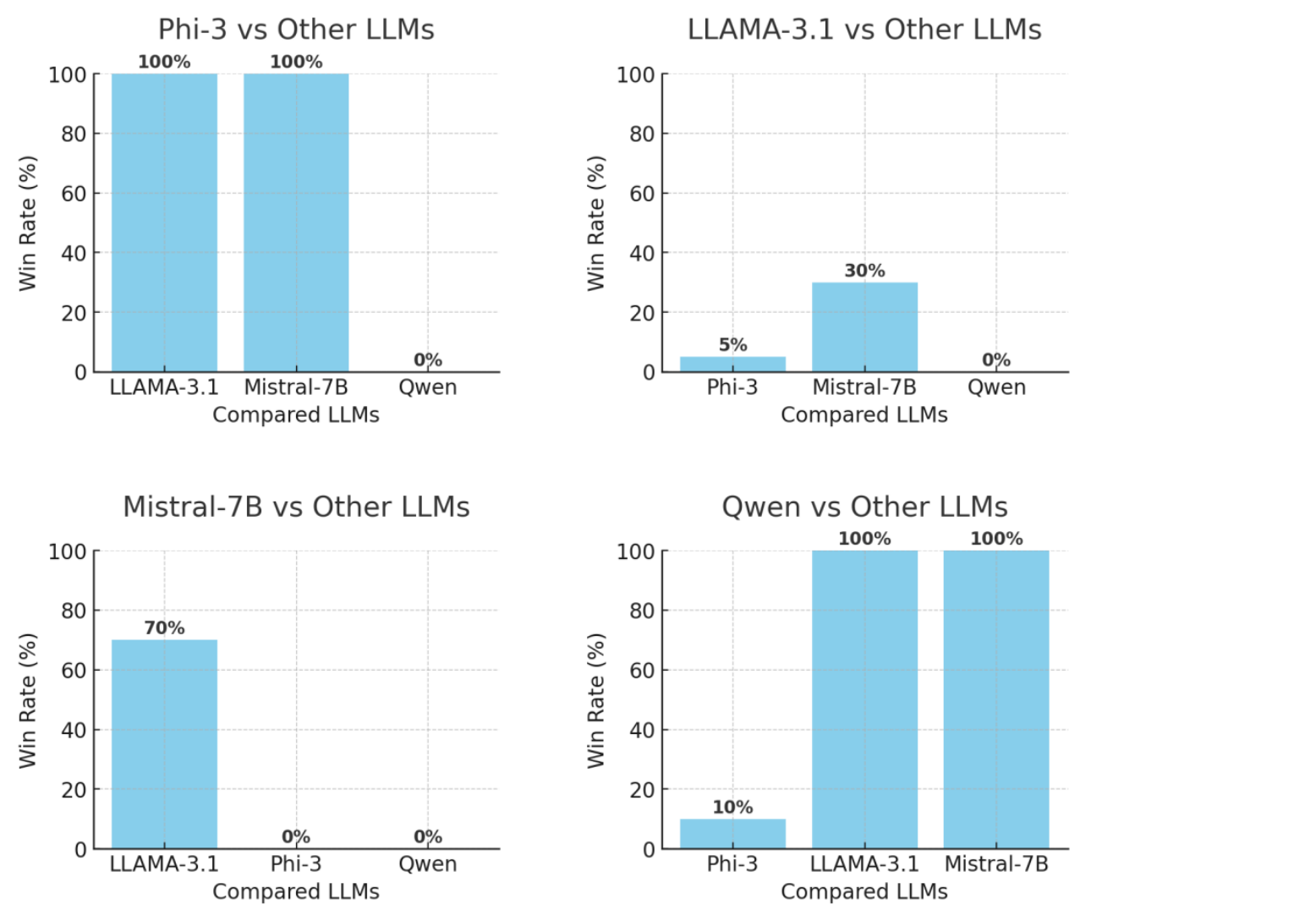


Figure 4: Overall Win-Rates for Information Retrieval Task 3 (narrow-specific prompts across 3 manuscripts)

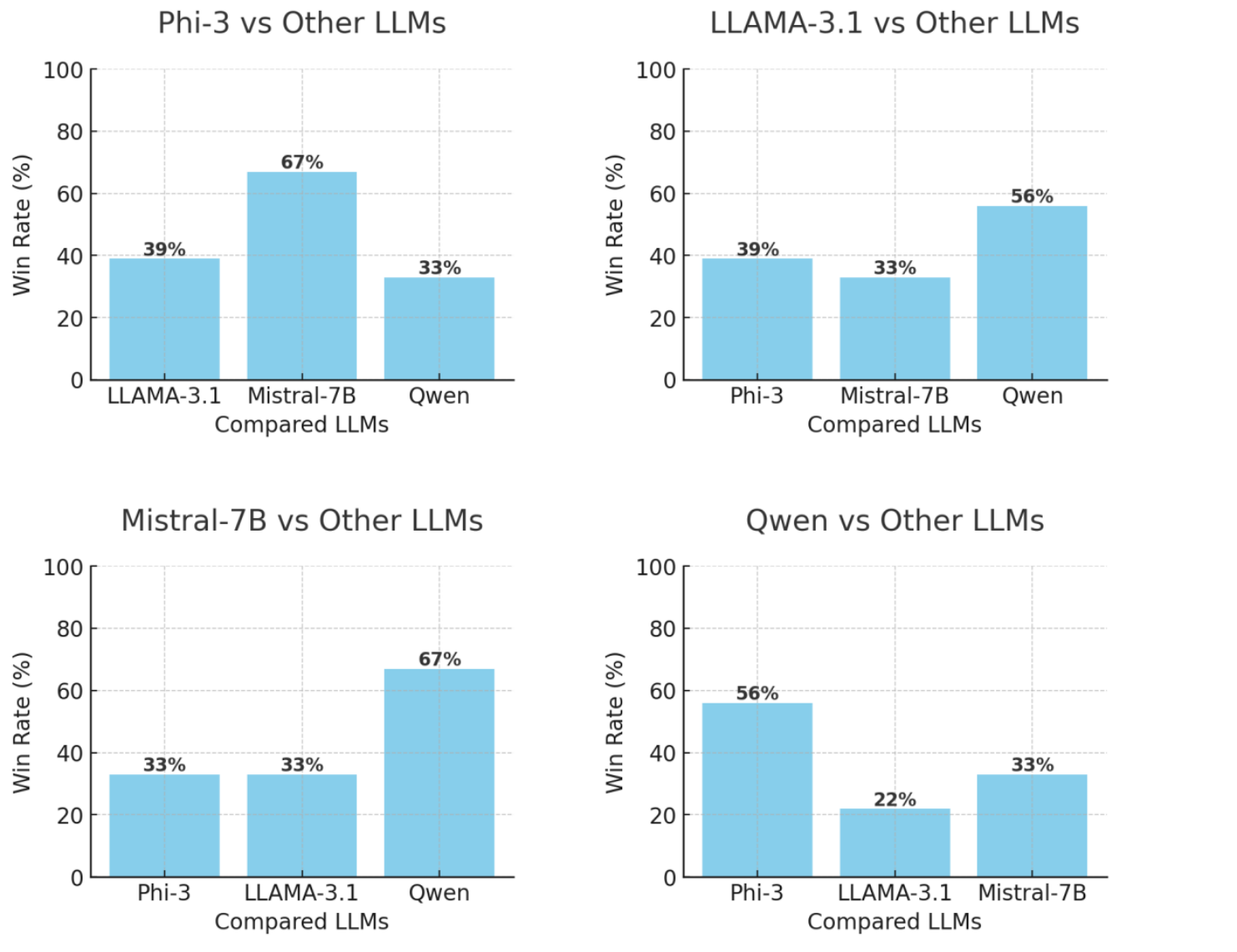
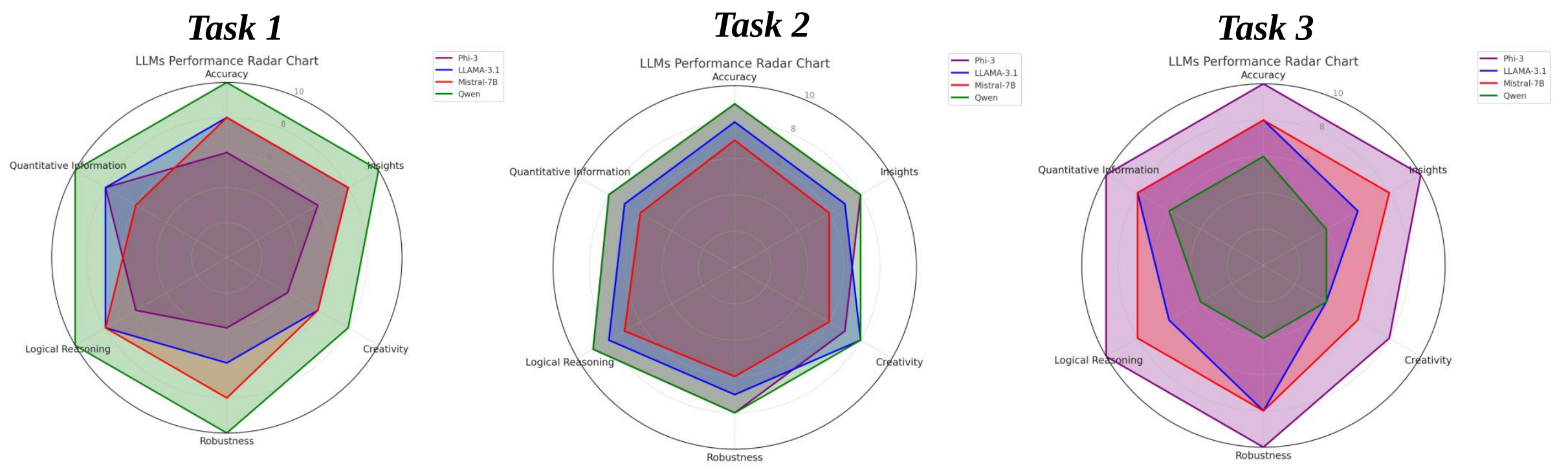


Figure 5: Representative Performance of LLMs for Information Retrieval Tasks 1-3


Task 1 *Task 2* *Task 3*

CONCLUSIONS

Open-source LLMs, despite their great potential, are currently underutilized in the pharmaceutical industry, and specifically in HEOR and medical affairs applications. As new software architectures begin to emerge (e.g. Mixture-of-Agents and Agentic Systems), we hereby explore different open-source LLMs for information retrieval and insights generation across various scenarios for HEOR purposes. By utilizing a proprietary LLM (OpenAI's o1) as a judge, we evaluated the performance of four open-source Qwen2-72B, Llama-3.1-8B, Mistral-7B and Phi-3-Mini-4K, on zero-shot learning information retrieval tasks on clinical manuscripts and abstracts. This evaluation study fills in the gap of analysing the comparative performance among open-source LLMs for HEOR specific tasks, by utilizing a superior LLM as an evaluator, and identifying key variables such as document size and prompt engineering as critical differentiation parameters between models.

Overall, results indicate there is no significantly better model among the four tested open-source LLMs across the different scenarios. Importantly, open-source LLMs relative performance still vary depending on the specific tasks, but they could be prompt-engineered or fine-tuned to improve the performance.[2] Notably, when open-ended prompts were used, Qwen2-72B demonstrated consistent superior performance regardless of document sizes. Specifically, in the informational retrieval task using an abstract only with open-ended prompts, Qwen2-72B reported a consistently higher than 50% win-rate versus the other models. Additionally, we utilized OpenAI's o1 as the evaluator of open-source LLM outputs across the different scenarios, rather than an LLM of the GPT series, which has previously been employed in HEOR relevant tasks.[2,4,11] O1 model, released in September 2024, has been trained to excel in complex reasoning tasks and often provide more accurate results in various classification tasks.[5]

Open-source LLMs offer substantial advantages that make them increasingly viable for use in the pharmaceutical industry, particularly HEOR applications. By providing cost-effective solutions with high flexibility, open-source LLMs enable organizations to develop advanced, customized applications without the high costs associated with proprietary models. Open-source LLMs not only reduce licensing fees but also support the rapid experimentation needed in the fast-evolving field of healthcare research, allowing organizations with limited budgets to test hypotheses, explore novel solutions, and accelerate insights without financial barriers. Moreover, these models facilitate a level of customization and adaptability that is essential in HEOR, where domain-specific language and unique regulatory contexts demand specialized tuning. This adaptability allows one to tailor the model to handle nuanced healthcare data more effectively, making them more suitable for tasks like analysing medical literature, identifying patterns in clinical outcomes, and supporting decision-making in cost-effectiveness studies. The transparency of open-source LLMs also provides an advantage in terms of regulatory compliance and model interpretability, both critical in the highly regulated pharmaceutical industry. Unlike proprietary models, where algorithms and data-handling processes are often opaque, open-source models enable practitioners to fully understand the underlying code and algorithms. In an industry where data accountability is paramount, the ability to maintain full oversight of data flow through the model is a crucial benefit, allowing organizations to conduct rigorous audits and comply with complex regulatory frameworks.

As this work focused primarily on highlighting the wider applicability of open-source LLMs across HEOR, we recognize we only investigated a few examples of different scenarios across the various abstracts and full manuscripts. Future studies should explore emerging open-source LLM model performance on a larger pool of document sample sizes across different tasks relevant to HEOR and medical affairs applications.

Key Messages

This study highlights the advantages of using open-source LLMs for HEOR within the pharmaceutical industry. It evaluates the performance of four open-source LLMs—Qwen2-72B, Llama-3.1-8B, Mistral-7B, and Phi-3-Mini-4K—on complex information retrieval tasks across clinical abstracts and full manuscripts. OpenAI's proprietary o1 model was used as a judge to evaluate the accuracy and robustness of these models across various prompts. Findings demonstrate that while no significant differences emerged among the open-source models, certain models like Qwen2-72B showed consistent high performance on open-ended prompts. This result underscores the capability of open-source models to compete with proprietary counterparts when optimized, particularly for tasks such as systematic literature reviews, evidence extraction, and medical data summarization.

The study also introduces a simplified FLASK methodology, aligned with PICO, which evaluated model outputs on criteria such as accuracy, robustness, and logical reasoning. This framework is instrumental in assessing open-source LLMs' applicability for HEOR tasks, demonstrating that performance varies with prompt engineering and document type. Findings advocate for broader adoption of open-source LLMs in the pharmaceutical industry due to their flexibility in adapting to domain-specific language needs and compliance with regulatory requirements, particularly when used on public data sources. These insights suggest that open-source LLMs can significantly enhance HEOR and medical affairs functions, making them a viable, cost-effective alternative for advancing data-driven health outcomes research.

- Open-source LLMs, despite their great potential, remain underutilized in the pharmaceutical industry, and specifically in HEOR and medical affairs applications
- Optimization and performance ranking of open-source LLMs for HEOR specific tasks can offer an attractive alternative to proprietary models, particularly as new software architectures emerge
- OpenAI's o1 was utilized as Judge LLM of the outputs generated from immunology-specific publications with minimal human intervention.

REFERENCES

1. Lu, Z.; Peng, Y.; Cohen, T.; Ghassemi, M.; Weng, C.; Tian, S. Large language models in biomedicine and health: current research landscape and future directions. *J Am Med Inform Assoc* 2024, *31*, 1801-1811, https://doi.org/10.1093/jamia/ocae202
2. Zhang, G.; Jin, Q.; Zhou, Y.; Wang, S.; Idnay, B.; Luo, Y.; Park, E.; Nestor, J.G.; Spotnitz, M.E.; Soroush, A.; et al. Closing the gap between open source and commercial large language models for medical evidence summarization. *NPJ Digit Med* 2024, *7*, 239, https://doi.org/10.1038/s41746-024-01239-w
3. Reason, T., Benbow, E., Langham, J., Gimblett, A., Klijn, S. L., & Malcolm, B. (2024). Artificial intelligence to automate network meta-analyses: Four case studies to evaluate the potential application of large language models. *PharmacoEconomics - Open*, *8*(2), 205–220. https://doi.org/10.1007/s41669-024-00476-9
4. Reason, T., Langham, J., & Gimblett, A. (2024). Automated mass extraction of over 680,000 PICOs from clinical study abstracts using generative AI: A proof-of-concept study. *Pharmaceutical Medicine*, *42*, 123–140. https://doi.org/10.1007/s40290-024-00539-6
5. Wu, S., Peng, Z., Du, X., Zheng, T., Liu, M., Wu, J., ... & Liu, J. H. (2024). A Comparative Study on Reasoning Patterns of OpenAI's o1 Model. *arXiv preprint arXiv:2410.13639*. Retrieved from https:// https://arxiv.org/abs/2410.13639.
6. Athanassiou, P., Kotrotsios, A., Kallitsakis, I., Bounas, A., Dimitroulas, T., Garyfallos, A., Tektonidou, M. G., Vosvotekas, G., Livieratos, A., Petrikkou, E., & Katsifis, G. (2022). The effects of golimumab on work productivity and quality of life among work-active axial spondyloarthritis and psoriatic arthritis patients treated in routine care in Greece: The 'GO-UP' study. *Quality of Life Research*, *31*, 1385–1399. https://doi.org/10.1007/s11136-021-03044-4
7. Athanassiou, P., Psaltis, D., Georgiadis, A., Katsifis, G., Theodoridou, A., Gazi, S., Sidiropoulos, P., Tektonidou, M. G., Bounas, A., Kandyli, A., et al. (2023). Real-world effectiveness of golimumab in adult patients with rheumatoid arthritis, psoriatic arthritis, and axial spondyloarthritis and an inadequate response to initial TNFi therapy in Greece: The GO-BEYOND prospective, observational study. *Rheumatology International*, *43*, 1871–1883. https://doi.org/10.1007/s00296-023-05376-5
8. García-Dorta, A., González-Dávila, E., Sánchez-Jareño, M., Cea-Calvo, L., Pombo-Suárez, M., Sánchez-Alonso, F., Castrejón, I., & Díaz-González, F. (2024). Early identification of golimumab-treated patients with higher likelihood of long-term retention. *Frontiers in Immunology*, *15*, 1359571. https://doi.org/10.3389/fimmu.2024.1359571
9. Govoni, M.; Batalov, A.; Boumpas, D.T.; D'Angelo, S.; De Keyser, F.; Flipo, R.M.; Kellner, H.; Leroi, H.; Khalifa, A. Real-world effectiveness of golimumab in the treatment of patients with active rheumatoid arthritis, psoriatic arthritis, or axial spondyloarthritis who failed initial TNF-α inhibitor therapy: a pooled analysis of European prospective observational studies. *Clin Exp Rheumatol* 2024, *42*, 642-650, https://doi.org/10.55563/clinexprheumatol/7sr1ni.
10. Weinstein, C.L.J.; Meehan, A.G.; Lin, J.; Briscoe, S.D.; Govoni, M. Long-term golimumab persistence: Five-year treatment retention data pooled from pivotal Phase III clinical trials in patients with rheumatoid arthritis, psoriatic arthritis, and ankylosing spondylitis. *Clin Rheumatol* 2023, *42*, 3397-3405, https://doi.org/10.1007/s10067-023-06760-z
11. Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., & Stoica, I. (2023). Judging LLM-as-a-judge with MT-bench and chatbot arena. arXiv preprint, arXiv:2306.05685. Retrieved from https://arxiv.org/abs/2306.05685.
12. Ye, S., Kim, D., Kim, S., Hwang, H., Kim, S., Jo, Y., ... & Seo, M. (2023). Flask: Fine-grained language model evaluation based on alignment skill sets. *arXiv preprint arXiv:2307.10928*. Retrieved from https:// https://arxiv.org/abs/2307.10928.

Email: maria.kudela@pfizer.com