Superior Performance of Generative AI after Application of Mixture-of-Agents LLMs in Outcomes Research

SA16

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

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

INTRODUCTION

Large Language Models (LLMs) have significantly advanced natural language processing, particularly in automating data extraction from scientific publications.[1] Despite their strengths, LLMs are inherently stochastic, often producing inconsistent outputs.[1-4] Previous studies have reported near-perfect extraction capabilities using LLMs, yet variability in results persists as a notable challenge.[1-2] Optimizing such capabilities and automating such procedures are invaluable for the Health Economics and Outcomes Research (HEOR) and Medical Affairs domains, including the significant support offered for external engagements.[5-6]

Previous work has highlighted that GPT-4's data extraction performance was superior to human efforts, as human error rates in data extraction are typically higher (ranging from 8% to 42% based on past studies).[2] However, limitations remain in terms of the broader contextual comprehension of the LLMs besides data extraction for Network Meta-Analysis automation and other functionalities for the pharmaceutical industry. To this end:

We proposed a "Mixture-of-Agents" (MoA) approach, which combines the strengths of LLMs working together, and leads to better performance than using a single model.[3] The work shows that when LLMs collaborate by refining each other's outputs, they generate higher-quality responses. This method outperforms even some of the most advanced models like GPT-4o on several benchmarks, and it offers a cost-effective and flexible solution for improving the capabilities of language models. Interestingly, Claude 3.5 Sonnet was utilized as a judge LLM to assess all outputs.[4] Indeed, previous research has identified that LLMs can match human judgments with over 80% agreement, a level similar to human-to-human agreement.[4] This suggests that LLMs can serve as reliable judges for assessing human preferences, offering a cost-effective and efficient alternative to traditional human evaluations.

OBJECTIVE

The aim of this work was to evaluate, using an LLM judge, the performance of LLM Mixture-of-Agent's insights generation, using mostly open-source LLMs, for a wide range of HEOR and Medical Affairs applications, including network meta-analysis and external engagement. Hence, we executed this proof-of-concept study on PubMed abstracts by applying LLM Mixture-of-Agents architecture for insights generation.

METHODS

In this study, Claude 3.5 Sonnet acted as the main evaluator for outputs generated by various LLMs. As an arbiter, Claude 3.5 Sonnet assessed the accuracy and reliability of data extracted from publication abstracts by examining multiple outputs across six metrics: Accuracy, Robustness, Creativity, Insights, Quantitative Information, and Logical Reasoning.[3-6] These metrics encapsulate all the necessary evaluation and more on task-agnostic LLMs outputs from abstracts and report on population, intervention, comparator, and outcomes parameters too.[3-6] 5 different abstracts from PubMed links on Immunology publications were used as data sources to cross-validate findings.[7-11] Minimal human intervention was required throughout this process.

Claude 3.5 Sonnet evaluated outputs for the MoA and Control. Each agent functioned as a proposer, independently generating outputs based on customized prompts. The evaluation process was applied to the aggregator model, GPT-4o, which synthesized all outputs into a final, refined response. Notably, open-source LLMs were used during the proposer stage (each model assess components of the text independently), while GPT-4o served as the aggregator to synthesize the results. The MoA architecture often involves multiple layers of aggregators. After the first layer of proposers generates outputs, the aggregator synthesizes them into a more refined response. This refined response can then be passed to the next layer, where the process repeats. Each layer progressively enhances the quality of the output, building on the synthesis of previous layer. At each stage of this methodology (proposer stage, aggregator state, judge stage), different LLMs were utilized to minimize bias. The proposer stage (layered) utilized the following models: Meta-Llama-3.1-8B-Instruct-Turbo, Mistral-7B-Instruct-v0.3, Qwen2-72B-Instruct. The function was run over 3 layers, where each layer refined the output of the previous one. [3]

Control runs were also conducted to ensure that the MoA output was superior to using GPT-4o alone as a single proposer model when identical prompts were applied. Figure 1 illustrates the MoA methodology.

Figure 1: MoA-based methodology for data extraction and insights generation

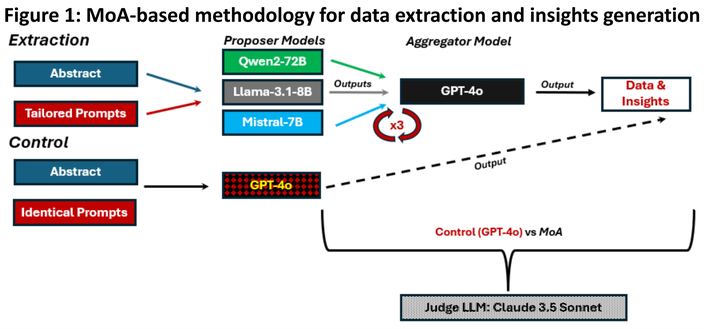

Figure 1: First, a scientific abstract from a PubMed article is scraped. It is then processed in layers: in the first layer, three independent AI models and all analyse different aspects of the abstract—one focuses on extracting quantitative data, another on interpreting clinical significance, and the third on discussing implications for future research. In the subsequent layers, a more advanced AI model (GPT-4o) aggregates and synthesizes these results, refining the analysis at each layer, and producing a final MoA output. This output is then compared, as evaluated by the Judge LLM, to the control (GPT-4o) output that shares an identical prompt to the aggregator prompt from the MoA methodology.

RESULTS

MoA demonstrated superior overall performance across the 5 abstracts/runs tested. Radar plots and bar charts per abstract in Figure 2 illustrate these findings. Importantly, MoA using GPT-4o as an aggregator model, exhibited superior overall performance to GPT-4o alone as a single proposer model across all runs, as evaluated by Claude 3.5 Sonnet. These findings demonstrate that insights produced using open-source LLMs as proposer models and GPT-4o as aggregator model currently offer superior value across multiple functional domains from insights generation to logical reasoning, among others. Specifically, the average difference in overall output performance between MoA and control (GPT-4o) is 4.4 points (7.33%). This indicates that the overall performance difference between the two (MoA vs Control) is statistically significant (paired t-test; p=0.0018), as well as the fact the average percentage difference exceeds 5% which is relevant for practical applications. These findings were currently demonstrated across recent abstracts from Immunology publications.

Figure 2: Output performance of Control (GPT-4o) and MoA across different abstracts and different metrics

CONCLUSIONS

The MoA methodology demonstrates how combining multiple LLMs can optimize various tasks like reasoning, language generation, and data extraction. In this study, GPT-4o served as our aggregator model while Claude 3.5 Sonnet as our judge LLM for evaluating different outputs. As LLMs continue to develop, MoA can adapt to incorporate different LLMs at different stages of this study that may lead to an even better overall performance. By using layered agent structures, this system significantly improves performance over using single models like GPT-4o alone.[3] The advantage lies in MoA's ability to harness the strengths of various models, some acting as proposers (generating diverse outputs) and others as aggregators (refining and combining those outputs into high-quality responses). This collaborative structure offers superior performance in specific tasks like AlpacaEval 2.0 and MT-Bench.[3]

Overall, these findings converge on the effectiveness of using a multi-model or agent approach to optimize LLM performance in specialized domains such as medical research and systematic reviews. The collaborative nature of the MoA structure seems to be its most potent feature, leveraging the diverse strengths of different models. This has wide-ranging implications for fields that rely heavily on data extraction, analysis, and decision-making, where LLMs can now perform tasks faster and with greater precision than human analysts.

This proof-of-concept study utilized immunology PubMed abstracts.[7-11] The statistically significant superior performance of MoA is expected to be even more visible when analysing larger volumes of data due to the complexity challenge. The advantages of specialization seen in MoA could increase with document size. With larger documents, the need for layered processing becomes even more pronounced. This layered approach could capture more detailed relationships between different parts of the document, like linking the methodology to the statistical results or aligning clinical significance with the literature review.

This advantage can lead to better decision-making in clinical trials, R&D, and market access by providing more reliable data interpretation, enhancing cost-effectiveness analyses, and strengthening value-based pricing strategies. MoA's ability to improve real-world evidence generation and health outcomes predictions can also support stronger negotiations with payers and regulators, ultimately driving innovation and ensuring data-driven strategies in drug development and market positioning.

However, while MoA improves performance across various metrics, it also introduces challenges, such as increased computational costs and potential delays in processing. Future studies should explore ways to streamline these systems to balance performance with cost and latency.

Key Messages

Previous efforts to harness large language models (LLMs) for data extraction in the Health Economics and Outcomes Research (HEOR) domain have proven their value.[1,5-6] However, there is a growing opportunity to apply advanced computer science principles that can enhance and extend insights generation and data extraction capabilities beyond conventional methodologies.[3-4,6] These principles are increasingly transferable and adaptable, enabling a broader adoption of such sophisticated methodologies.

In this context, we showcase the application of a cutting-edge LLM not only as an evaluator of less advanced LLMs, but we also demonstrate a multi-layered, mixture-of-agents framework.[3-4,6] This approach, leveraging the improved performance of integrated open-source models, has already demonstrated notable advancements such as improving the performance and flexibility of LLMs by combining specialized sub-models to handle different tasks.[3-4] By integrating various LLMs in a collaborative and layered manner, this methodology optimizes the extraction process and enhances the quality of insights.

These innovations are poised to drive significant progress via the integration of computer science and life sciences, particularly within the pharmaceutical industry. Optimizing these operations is expected to streamline regulatory approval processes to facilitate better patient access to groundbreaking therapies and empower key opinion leaders with richer, more creative insights.

- This study represents a proof-of-concept of successfully implementing MoA within the pharmaceutical industry operations, and particularly HEOR.
- MoA overall performance exceeded all other individual open-source LLMs outputs, including GPT-4o, across all scenarios.
- Claude 3.5 Sonnet was utilized as Judge LLM of the outputs generated from immunology-specific publications with minimal human intervention.

REFERENCES

1. 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
2. Mathes, T., Klaßen, P., & Pieper, D. (2017). Frequency of data extraction errors and methods to increase data extraction quality: A methodological review. *BMC Medical Research Methodology, 17*(1), 152. https://doi.org/10.1186/s12874-017-0431-4
3. Wang, J., Wang, J., Athiwaratkun, B., Zhang, C., & Zou, J. (2024). Mixture-of-agents enhances large language model capabilities. *arXiv preprint*, arXiv:2406.04692. Retrieved from https://arxiv.org/abs/2406.04692
4. 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
5. 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
6. 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
7. Psaltis, D., Settas, L., Georgiadis, A., Koukli, E., Bounas, A., Livieratos, A., Petrikkou, E., Kalogiannaki, H., Repa, A., Vassilopoulos, D., & Sidiropoulos, P. (2022). The effects of golimumab on patient centric outcomes amongst rheumatoid arthritis patients in Greece: The GO-Q study. *Rheumatology International*, *42*, 639–650. https://doi.org/10.1007/s00296-021-05073-1
8. 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
9. 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
10. D'Angelo, S., Tirri, E., Giardino, A. M., Mattucci-Cerinic, M., Dagna, L., Santo, L., Ciccia, F., Frediani, B., Govoni, M., Bobbio Pallavicini, F., et al. (2022). Effectiveness of golimumab as second anti-TNFα drug in patients with rheumatoid arthritis, psoriatic arthritis and axial spondyloarthritis in Italy: GO-BEYOND, a prospective real-world observational study. *Journal of Clinical Medicine*, *11*. https://doi.org/10.3390/jcm11144178
11. 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

Email: maria.kudela@pfizer.com