Kenneth Lay – 2602119272

# Review Paper: Exploring the Latest LLMs for Leaderboard Extraction

## Summary of the Paper:

The paper titled "Exploring the Latest LLMs for Leaderboard Extraction" investigates the effectiveness of various state-of-the-art Large Language Models (LLMs) in extracting leaderboard data from research papers. The study focuses on Mistral 7B, Llama-2, GPT-4-Turbo, and GPT-4.o, analyzing their ability to extract (Task, Dataset, Metric, Score) quadruples essential for creating leaderboards that rank AI models. The authors used three types of contextual inputs—DocTAET, DocREC, and DocFULL—to evaluate the models' precision in few-shot and zero-shot scenarios. The research underscores the importance of selecting appropriate context to maximize model accuracy, with Mistral 7B outperforming other models in most cases.

## Identification of Gaps

Although the paper broadens its scope to various domains of scientific research, a gap exists in how the models would perform in fields with vastly different structures, such as humanities or social sciences. Additionally, while optimization techniques like instruction tuning were used, the paper does not explore in depth the impact of different fine-tuning strategies on performance across these non-traditional fields. Lastly, the evaluation metrics focus primarily on accuracy, leaving out considerations for practical scalability when handling papers with highly inconsistent or ambiguous formatting.

## Analysis of Ambiguities

The methodology employed by the paper’s authors, particularly the use of three distinct context types (DocTAET, DocREC, and DocFULL), provided very useful insight into how different models handled varied inputs. However, while the methodology is well detailed for extraction tasks, the paper does not sufficiently explore other alternative evaluation metrics other than ROUGE, such as extraction-specific F1 scores. The inclusion of instruction tuning for optimization is a strong aspect of the study, but the research could have benefitted from experimenting with hybrid models combining LLMs and rule-based systems for a more comprehensive comparison.

## Evaluation of Methodology

The methodology employed by the authors, particularly the use of three distinct context types (DocTAET, DocREC, and DocFULL), provides useful insight into how different models handle varied inputs. However, while the methodology is well-detailed for extraction tasks, the paper does not sufficiently explore alternative evaluation metrics beyond ROUGE, such as extraction-specific F1 scores.

## Discussion of Question Raised

The study raises some critical questions regarding the scalability and domain specificity of leaderboard extraction. Specifically, how well would these LLMs perform when they are applied to research papers within non-technical fields with more diverse and less structured formats? Additionally, the question of how to manage hallucinations in longer contexts, such as DocFULL, remains partially unanswered. The trade-off between context length and extraction accuracy is an important issue that warrants further exploration.

## Critical Reflection

The study has shown valuable findings on the capabilities of LLMs in leaderboard extraction across a broad spectrum of scientific fields. However, while optimization was achieved through instruction tuning, the paper could have delved deeper into how different fine-tuning approaches such as domain specific models could enhance the extraction process further. Another area for some critical reflection is the scalability of the methodology for tracking real-time research advancements across an ever-expanding body of work.

## Conclusion

The paper successfully and wonderfully demonstrated that LLMs like Mistral 7B, Llama-2, GPT-4-Turbo, and GPT-4.o can automate leaderboard extraction across various scientific domains with high accuracy. Mistral 7B emerged as a leader in both few-shot and zero-shot settings, with the study showing the importance of selecting the appropriate context (DocTAET, DocREC) to optimize performance. The research highlights that LLMs, when properly optimized, are able to handle a broad range of disciplines.

## Future Research Proposal

Future research could explore applying LLMs to a broader range of fields, like humanities or social sciences, to see how well they handle different types of research papers. Combining LLMs with traditional rule-based methods could also improve leaderboard extraction, making the process more reliable across various domains. Additionally, they could use other evaluation metrices to get more in-depth analysis on their result. Finally, tackling the issue of hallucinations, particularly with longer documents, would make the models more trustworthy for real-world use.

## Personal Insights

I really enjoyed this paper and found the research approach to be quite impressive I must say. The way the authors used different contexts, like DocTAET and DocREC, to optimize leaderboard extraction was particularly clever. I also liked how they compared both open-source and proprietary models, giving a balanced view of their performance. The attention to detail and design made the results clearer and more practical. Overall, it's a very well-executed study that clearly showcases the growing potential of LLMs, and I appreciate how thoroughly they explored the topic.