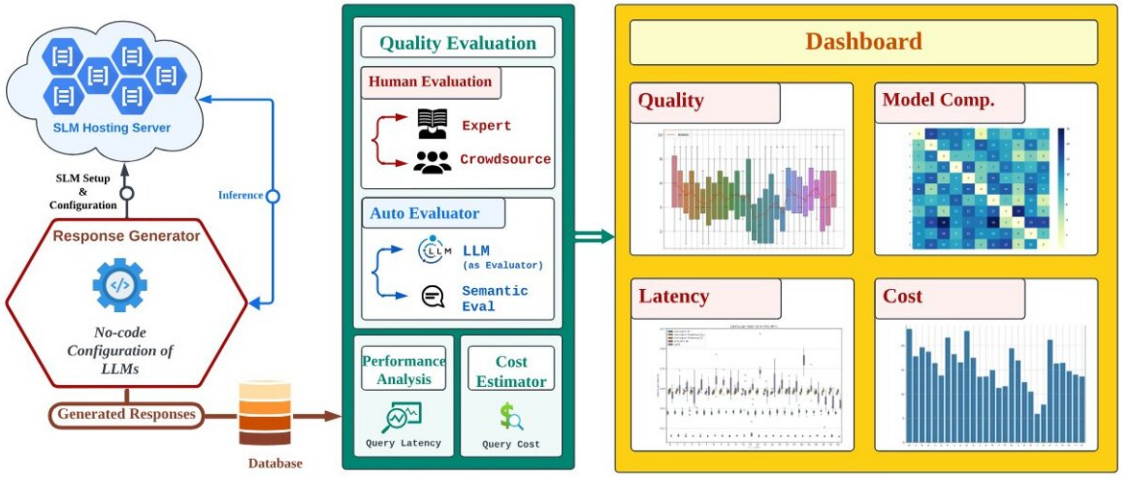
New tool capable of comparing SLMs and LLMs finds smaller models can reduce cost

May 14 2024, by Patricia DeLacey



Architecture Overview of the SLaM Tool. Credit: *arXiv* (2023). DOI: 10.48550/arxiv.2312.14972

Open-source small language models (SLMs) can provide conversational responses that are similar to resource-intensive, proprietary large language models (LLMs) such as OpenAI's ChatGPT, but at a lower cost, researchers at the University of Michigan have found. Their findings are published on the *arXiv* preprint server.

The team developed a first-of-its-kind tool capable of evaluating SLMs and comparing them to proprietary LLM Application Programming Interfaces, including performance and cost. They recently presented their results at the 2024 IEEE International Symposium on Performance Analysis of Systems and Software.

LLMs' demonstrated ability to comprehend and generate language has led to widespread use in applications like virtual assistants, chatbots and language translation systems. Although useful, LLMs cost millions or more to train, limiting the advancement of AI to tech giants while smaller companies must rely on their paid services.

"A lot of companies such as Duolingo and Slack are incorporating LLMs like OpenAI's GPT-4 into their products. It's important to rigorously examine whether these models are really the best choice for developers and whether small open models could be effective," said Jason Mars, an associate professor of computer science and engineering at the University of Michigan.

Implementing proprietary LLMs enhances speed and convenience but comes with downsides of limited customization and data privacy, unreliable performance, lags during peak usage and high cost.

Open-source SLMs have emerged as an alternative, but up to this point, there has not been a way to systematically compare their performance with more widely known LLMs.

The research team developed an automated analysis tool, named SLaM, as the first reported methodology for evaluating SLMs and their tradeoffs—quality, performance and cost—compared with LLMs.

"We created SLaM and made it open source to fill the void in tools that accelerate and automate comparative analysis of open and closed LLMs on a case-by-case basis," said Mars.

The tool was put to the test in an AI productivity tool under development by Myca AI called "daily pep talk." The feature leverages the user's task list to deliver personalized and intelligent encouragement and advice on a daily basis.

The researchers assessed 29 distinct versions of nine SLMs against OpenAI's GPT-4 in the daily pep talk production environment. While GPT-4 achieved the highest accuracy as judged by a human panel, most SLMs came close to its quality with more predictable latency performance.

"We were surprised by the high quality answers provided by these small models. Many times users could not really differentiate between SLM and LLMs," said Lingjia Tang, an associate professor of computer science and engineering.

Importantly, the SLMs reduced costs between five and 29 times compared to LLMs depending on the model used.

"This finding has big implications for smaller companies trying to maintain competitiveness in this fierce AI race. With SLaM tools, companies can select smaller open-source models that provide high quality answers but cost much less, reducing their dependencies on tech giants," added Tang.

Additional co-authors: Chandra Irugalbandara, Ashish Mahendra, Tharuka Kasthuri Arachchige, and Jayanaka Dantanarayana of Jaseci Labs; Yiping Kang, Roland Daynauth, and Krisztian Flautner of the University of Michigan.

More information: Chandra Irugalbandara et al, Scaling Down to Scale Up: A Cost-Benefit Analysis of Replacing OpenAI's LLM with Open Source SLMs in Production, *arXiv* (2023). DOI: 10.48550/arxiv.2312.14972

Provided by University of Michigan College of Engineering