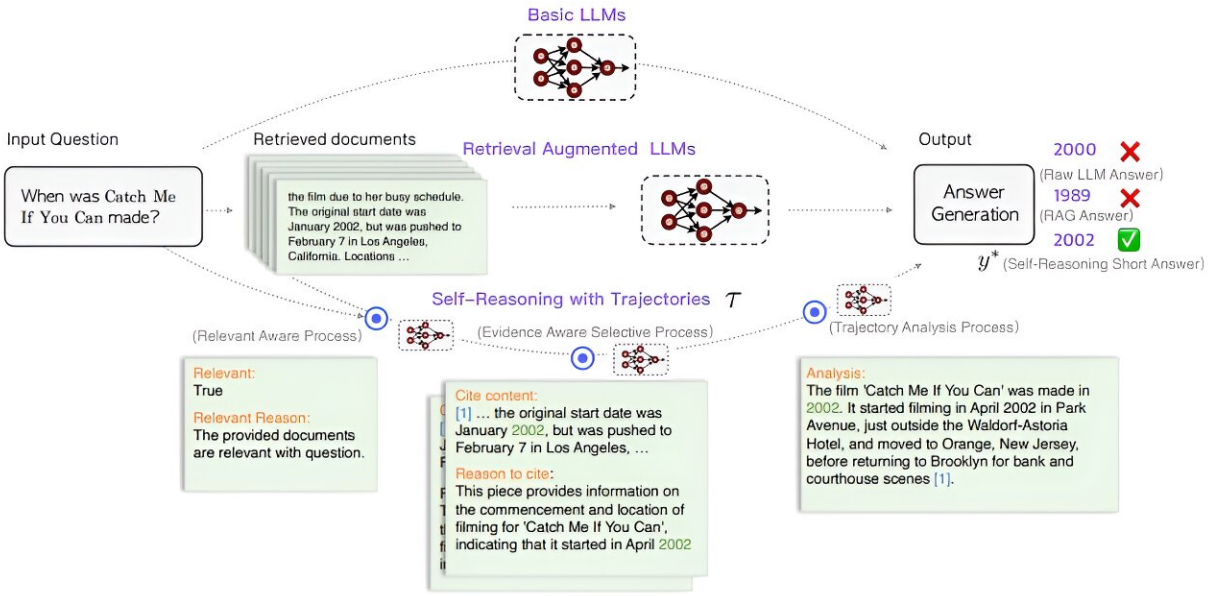
Team proposes a reasoning framework aimed at improving the reliability and traceability of LLMs

July 31 2024, by Bob Yirka



An illustration of the SELF-REASONING framework for improving the RALMs. The upper is the basic LLMs which answer the question by inherent knowledge. The middle is the standard retrieval augmented LMs, which use retrieved documents to help answer the question. The bottom is our SELF-REASONING framework which uses self-generated reason trajectories to output answers. Credit: *arXiv* (2024). DOI: 10.48550/arxiv.2407.19813

A team of AI researchers at the Chinese technology company Baidu Inc.

is proposing a new reasoning framework designed to improve the reliability and traceability of LLMs. The group has published a paper describing their ideas on the *arXiv* preprint server.

Over the past couple of years, LLMs such as ChatGPT have become mainstream applications, with users taking advantage of their capabilities to write documents, create images and even write songs.

But one glaring weakness of LLMs remains—their inability to check their own results to make sure they do not occasionally present users with "hallucinations," which are results that make no sense. This weakness prevents AI apps from being used for more critical applications that rely on data integrity.

In this new effort, the team at Baidu has come up with a strategy aimed at forcing LLMs to check their work before presenting results to end users.

The new approach involves adding a three-step process to the LLM engine just prior to presentation of results. The first is to add a relevance-aware model to assess the results and judge them on their relevance to the user prompt. The second involves an evidence-aware selective option in which relevant documents are chosen for citation and excerpts are presented as evidence of correctness of an answer. The third involves a trajectory analysis module that conducts a clear and concise analysis based on results from the prior two modules. It then provides the user with the final answer.

The research team suggests the approach would force LLMs to be more aware of provided answers to users, which in the end should improve accuracy. The team has also tested their ideas by adding test modules to LLMs and then writing prompts. They claim the improved LLMs were able to outperform GPT-4 using much smaller training datasets.

The researchers suggest that frameworks such as theirs could lead to more reliable LLMs, which could make them suitable for more applications. They also suggest they would open the field to more players who currently do not have access to massive training datasets.

More information: Yuan Xia et al, Improving Retrieval Augmented Language Model with Self-Reasoning, *arXiv* (2024). DOI: 10.48550/arxiv.2407.19813