'AI Judge' Can Learn New Tricks to Better Check Information and Program

AI researchers and developers are increasingly using large language models (LLMs) to evaluate the answers of other LLMs in a process called 'LLM-as-a-judge'.

AI researchers and developers are increasingly using large language models (LLMs) to evaluate the answers of other LLMs in a process called 'LLM-as-a-judge'. However, the quality of these evaluations often falls short in complex tasks such as long-form factual checking, advanced programming, and mathematical problem solving.

Now, a new study published by researchers from the University of Cambridge and Apple has proposed a new system that augments AI 'judges' with external validation tools to improve the quality of their judgments.

The system aims to overcome the inherent limitations of both human and AI annotation. Humans face challenges and biases due to time constraints, fatigue, and are influenced by writing style rather than actual accuracy. Meanwhile, AI struggles with complex tasks such as those mentioned above.

The new 'Evaluation Agent' framework the researchers created is autonomous, allowing it to evaluate responses to determine whether external tools are needed and to use the right tools. Each evaluation goes through three main steps: an initial assessment of the domain, using the tool, and making a final decision.

1. Fact-checking tools use web searches to verify atomic facts in answers.
2. The code execution tool leverages OpenAI's code interpreter to run and check the correctness of the code.

3. A math checker is a specialized version of a code execution engine used to validate mathematical operations and operations.

If no useful tool is available to make a judgment, the system will use the basic LLM annotation set to avoid unnecessary processing and the risk of performance degradation on simple tasks.

The system provided significant improvements in long-term fact-checking, with marked increases in agreement with ground-truth annotations across multiple baselines. On programming tasks, the agent-based approach significantly improved performance across all baselines. For difficult tasks, agents improved performance over some but not all baselines, and overall agreement remained relatively low at around 56%. Notably, the researchers found that for long-term fact-checking answers, agent agreement with ground-truth was higher than human annotations.

The framework is extensible, so in the future, other tools can be integrated to further improve LLM assessment systems. The source code for the framework will be open sourced on Apple's GitHub, but has not yet been officially released.