Qeios

Peer Review

Review of: "Do LLMs Overcome Shortcut Learning? An Evaluation of Shortcut Challenges in Large Language Models"

Istvan Siket[1]

1. University of Szeged, Hungary

Although the idea is not entirely original, the authors effectively implemented an existing method. Their innovations made the research more practical and realistic, and the use of more recent models was also significant, as these models are still widely employed today.

In the experiments, automated evaluation was crucial, but manual evaluation also plays a key role, especially when assessing artificial intelligence models using other AI models. The system could improve its results by listing well-chosen examples (e.g., average cases, outliers, etc.) that require human review to enhance the validation process.

The research focused on language models, but beyond prompting techniques, the manner in which the models were utilized was also critical. For instance, was sampling employed? If so, what were the temperature and top-p values? These parameters influence the randomness of token selection and could lead the models to drift, as reflected in the article's findings.

Notable strengths:

- The idea of accessible source code
- Values visually represented with color intensity

Areas for improvement:

- Display the source code on the highlighted page; currently, the repository is empty.
- Provide a more detailed explanation of how the models were used.
- Highlight a few manual examples for validation purposes in specific sections, such as 5.1.2.

doi.org/10.32388/MT2OSW

Declarations

Potential competing interests: No potential competing interests to declare.