Qeios

Peer Review

Review of: "Neutralizing Backdoors through Information Conflicts for Large Language Models"

Gaolei Li[1]

1. Shanghai Jiaotong University, China

This paper focuses on neutralizing backdoors through information conflicts for large language models. The topic is timely and interesting, but the technical content is very confusing.

1. Fig.2 is confusing; how to exploit information conflicts to mitigate backdoors? It is unclear.
2. The difference between backdoor purification and detection is unclear.
3. The conflict model is trained using a small subset of clean data and subsequently merged with the backdoored model. It is too simple and has few contributions.
4. External evidence conflicts seem to improve the prompts, which is equal to damaging the backdoor trigger. But this method cannot remove the backdoor in LLM.

Declarations

Potential competing interests: No potential competing interests to declare.

doi.org/10.32388/6JLJ2O