Qeios

Peer Review

Review of: "Safeguarding Large Language Models in Real-time with Tunable Safety-Performance Trade-offs"

Richard J. Gruss[1]

1. Radford University, United States

This is an excellent idea, and it appears to be effective. I especially like the ingenious use of the previous hidden layer as a sentence embedding. I have just a couple of suggestions, more about the paper itself than the method. First, the formalisms in sections 2 and 3 don't appear to be necessary for understanding the method. Second, the example on page 13 raises an issue of correspondence with the user's intention. The first answer is less safe, but it honors the request of the user, dealing with unsafety via a disclaimer. The second answer, under SAFENUDGE, gives the user the opposite of what they asked for. I would suggest using a different example or addressing this potential shortcoming. Finally (small), note that “elicit” means to evoke, where “illicit” means illegal. Thanks for the interesting paper!

Declarations

Potential competing interests: No potential competing interests to declare.

doi.org/10.32388/WH8QC5