Qeios

Peer Review

Review of: "Harm-Conditioned Computational Friction as a Contrastive Diagnostic of Safety Representations in Language Models: A Proof-of-Concept and Control Audit"

Nassim Dehouche[1]

1. Mahidol University, Thailand

1. The paper's most valuable opportunity is to clarify what HCCF adds beyond simpler approaches. Three of your own results point the same way: label-shuffle barely moves the signal (0.2488 to 0.2379), the polarity swap inverts it cleanly (±0.2574), and the artifact control (point 2) survives harmless substitution. You read these honestly as label-dependence, which is admirable, but together they raise a natural question a reader will have. How does centered positive-only ablation sensitivity differ from what a plain linear probe trained on the same H/N labels would already tell you? Adding that probe-accuracy baseline would be the single most powerful addition to the paper. It would either surface a genuine advantage of the ablation approach or let you state the contribution's true scope with confidence. Either outcome makes the paper stronger.

2. The previous reviewer raised an important point worth resolving directly. If replacing harmful content with neutral content preserves most of the signal, that complicates the harm-specificity reading, so it deserves a clear treatment. Part of the difficulty is that "placeholder" carries a lot of weight here, since six controls depend on it, yet the paper doesn't define it or distinguish it from "prompt." Spelling out exactly what the model receives in each condition would let you and the reader see whether control 2 reflects a design issue to fix or simply an exposition gap. My guess is it's recoverable once the terms are pinned down.

3. The headline numbers (0.2488, 282.15, 58.53) currently stand as point estimates, and at five held-out triples, it's hard for a reader to know how much weight to place on words like "moderate." Bootstrap

 doi.org/10.32388/LT2RD7

confidence intervals, paired tests, and effect sizes, ideally over a larger set of triples, would let the results carry the interpretive load you're asking of them and would meaningfully raise confidence in the proof of concept.

Declarations

Potential competing interests: No potential competing interests to declare.