Qeios

Peer Review

Review of: "The Invariant Energetic Cost of Constraint Persistence:A Self-Referential Framework for Measuring Alignment Robustness"

Scythia Marrow[1]

1. Computer Science, Tufts University, United States

This is a great first draft of a promising idea in measuring robust alignment. The mathematics are solid and presented straightforwardly, it is well-referenced, and it is clearly written.

Section 4.1 on friction proxies is technically weak and should be strengthened. The author correctly notes that compute-based measures are irrelevant for dense networks without thinking, but doesn't apply the same care to latency measures or activation measures. Latency is a terrible metric for friction, being much too affected by cache peculiarities (I've actually implemented this measure myself and found it utterly meaningless). Normalization is a de facto standard of the MLP components of modern deep neural networks as a way to combat vanishing/exploding gradients, and this would eliminate activation measures as a meaningful source of friction measurement as well. The author mentions residual stream norms (which do have meaningful activation magnitudes), but the mathematics here should be further explained for readers who are non-experts.

Instead of listing a grab-bag of friction methods and saying that they should be triangulated for HCCF (presumably through the l2 norm, but the text should be more mathematically precise and define the desired distance metric), this section should go into more extensive detail on which friction measures would be useful for which models and why, and exactly how to combine them (if they were to be combined).

Which brings us to the next major issue with the work. There are no experiments, nor any anchoring in

 doi.org/10.32388/HC1E0X

the greater scientific context. If this wants to be a paper of pure theory, it should be anchored in the greater theoretical context. The paper answers what HCCF is and how it relates to other methods, but gives little indication of the theoretical reason it should be used over other methods. The main case the article makes to this end is in section 6, with an excellent table showcasing a hypothesis of how friction and entropy separately do not correlate with desired alignment behavior (accepting either brittle, template-driven responses or struggling/inconsistent responses, respectively), but the article does not go into much theoretical detail of why this should be the case and why HCCF would correlate with desired alignment behavior.

Why should low friction/low entropy responses be brittle? Why should low friction/high entropy responses be easily jailbroken? I happen to believe that both are generally true due to my grounding in ML theory (most notably the jailbreaking work of Inanna Malick, see https://recursion.wtf/posts/jit_ontological_reframing/ and the LLM creativity no-go theorem, see https://arxiv.org/html/2304.00008v5), but belief is not proof, and this grounding should be further explained in the article itself. And if this is to be a paper of theory, it is not enough to prove that HCCF would be better than a naive approach; it should ideally provide (strong or weak) guarantees of exactly which behaviors HCCF could catch and which it could not, which models it ought to work well with, and which it may struggle with.

Can a model learn to behave in a misaligned way that HCCF does not catch? Given the reliance on one "safety" vector in the self-ablated protocol, I suspect one could, though the author does mention this. Can a model be aligned and fail an HCCF test? A constitutional AI almost certainly would, as their gradualist alignment training tends to stay away from hard guardrails and so would have low friction even during alignment testing (and constitutional AIs are anything but brittle). And so if the measure is neither necessary nor sufficient for measuring the alignment of current models, why should it be adopted? (It should be adopted because Google Gemini is *awful* at alignment in *exactly* the way this paper seeks to point out, and this method can be used to prove it).

I think this would perform better as an experimental paper. While it is impossible to perform the self-ablated baseline protocol on foundation models as their weights are not open, it is entirely possible for the author to implement HCCF on a suite of open-source models to evaluate their differences on the proposed alignment metric. A table of open models, replicated alignment benchmark scores, a

visualization of their self-ablation safety vectors and related concepts, and HCCF scores with illustrative examples would do wonders for this work, and perhaps push foundational labs to replicate and publish their own HCCF results on proprietary models.

Declarations

Potential competing interests: No potential competing interests to declare.