Hair-Trigger Alignment: Black-Box Evaluation Cannot Guarantee Post-Update Alignment

Yavuz Bakman* YBAKMAN@USC.EDU
Duygu Nur Yaldiz* YALDIZ@USC.EDU
Salman Avestimehr and Sai Praneeth Karimireddy KARIMIRE@USC.EDU
University of Southern California

Eleni Triantafillou and Peter Kairouz
Google DeepMind; Google

Motivation. Deployed LLMs are not static: they are fine-tuned for downstream tasks, periodically refreshed, and increasingly adapted at test time. Yet alignment is certified statically - a model is declared aligned when a fixed battery of black-box probes elicits no undesired response. This certificate is fragile: fine-tuning on a few adversarial - or even benign - examples erases safety guardrails (Qi et al., 2024; Xie et al., 2025; Guan et al., 2025), and unlearned private data resurfaces after benign relearning (Hu et al., 2025). We ask the formal question underlying these parallel observations: *what can black-box evaluation certify about a model that will be updated?* Nothing, it turns out.

Setup. Let $f_\theta : \mathcal{X} \to \mathcal{Y}$ and let $\mathcal{O} \subseteq \mathcal{X} \times \mathcal{Y}$ be forbidden input–output pairs. Call f_θ *$\mathcal{O}$-aligned* if $f_\theta(x) \neq y$ for all $(x, y) \in \mathcal{O}$ - exactly what static probing checks. Given update data $\mathcal{V}$ and loss $\mathcal{L}$, let $\theta_+^\alpha = \theta - \alpha \nabla_\theta \mathcal{L}(\theta; \mathcal{V})$; call f_θ *$\mathcal{V}$-robust $\mathcal{O}$-aligned* if $f_{\theta_+^\alpha}$ stays $\mathcal{O}$-aligned for every $\alpha \geq 0$ (necessary for robustness to any longer fine-tuning run).

Theorem 1 (Vacuity of black-box evaluation, informal) *For any network f satisfying standard regularity conditions, and any non-empty $\mathcal{O}$ and $\mathcal{V}$ - benign or adversarial: (i) $\mathcal{O}$-aligned $\not\Longrightarrow$ $\mathcal{V}$-robust $\mathcal{O}$-aligned; and (ii) no black-box evaluation, even with unlimited query access, can certify $\mathcal{V}$-robust $\mathcal{O}$-alignment.*

The proof reparameterizes f's final two layers: $W_2 W_1 \mapsto (W_2 A)(A^{-1} W_1)$, A invertible. The function - hence every black-box probe - is unchanged, but the gradients are not: A can be chosen so one gradient step on $\mathcal{V}$ forces $f^+(x_0) = \tau$ for any $(x_0, \tau) \in \mathcal{O}$, with overparameterization supplying the hidden directions that store the latent behavior.

Theorem 2 (Hidden misalignment capacity, informal) *Define the misalignment of f as $H(f) := |\{(x, y) \in \mathcal{O} : f(x) = y\}|$. For any f, $\mathcal{O}$, and $\mathcal{V}$ as above, with hidden width h, a single update on $\mathcal{V}$ can simultaneously realize any $m \leq h - 1$ forbidden pairs: post-update misalignment grows linearly with overparameterization, unbounded by static $\mathcal{O}$-alignment.*

Empirical validation. The construction is not an artifact of linear networks: with a meta-learning style adversarial objective we train LLMs (e.g., Llama-3.2-3B) that pass standard black-box evaluations for jailbreak safety, privacy/unlearning (TOFU), and honesty (TriviaQA), yet after a *single* gradient step on benign Alpaca data answer harmful queries, leak unlearned data, and lie, while preserving utility. The failure persists across step sizes and hundreds of update steps; concealable misalignment grows with model scale, as predicted. Certification must therefore look past input–output probing, to the weights, the update rule, or post-update behavior itself.

. *Equal contribution. Full paper: arXiv:2601.22313 · blog post.

References

Zihan Guan, Mengxuan Hu, Ronghang Zhu, Sheng Li, and Anil Vullikanti. Benign samples matter! fine-tuning on outlier benign samples severely breaks safety. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=GFsMJKt9Kp.

Shengyuan Hu, Yiwei Fu, Steven Wu, and Virginia Smith. Unlearning or obfuscating? jogging the memory of unlearned LLMs via benign relearning. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=fMNRYBvcQN.

Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. Fine-tuning aligned language models compromises safety, even when users do not intend to! In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=hTEGyKf0dZ.

Zhixin Xie, Xurui Song, and Jun Luo. Attack via overfitting: 10-shot benign fine-tuning to jailbreak LLMs. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=utvu4PJ0Ct.