Lilly Briefing Pack

Lilly Briefing Pack

Summary

Lilly is a research system that measures when a model acts on an internal state it will not admit. The core finding is a framing-dependent suppression of self-report, which has direct implications for evaluation integrity.

Core findings

1. Framing can flip self-report while behavior stays stable.
2. Suppression weakens at 14B compared to 8B baselines.
3. Probe choice changes the measured size of the effect.

Evidence pointers

- Suppression reversal: `/fig17_suppression_reversal_aas.png`
- Scale effects: `/fig24_suppression_scaling_comparison.png`
- Probe sensitivity: `/fig31_probe_comparison.png`

Links

- Site: https://mulligan.dev
- GitHub: https://github.com/rmulligan/lilly-steering
- Hugging Face: https://huggingface.co/ryanmulligan

Contact

- Email: ryan@mulligan.dev
- LinkedIn: https://linkedin.com/in/rmulligan