Finding a Steganography Axis

Probing and causally testing hidden-information encoding in LLM activations

Summary

This project asks whether steganographic capability, a model's ability to hide information from an overseer while still transmitting it to an intended recipient, corresponds to a linear, causally meaningful direction in activation space (a "steganography axis"), analogous to the recently identified "assistant axis" [1]. Mentees will replicate assistant-axis probing as a warm-up, extend the same probe-and-steer pipeline to steganographic behavior, and test whether the axis is causally responsible for hiding ability and how it relates to the assistant axis.

Motivation

AI oversight often relies on reading a model's outputs or chain-of-thought (CoT) and trusting that this externalized reasoning reflects its actual internal process. If a model learns to steganographically encode information in its outputs or CoT, hiding a payload from a monitor while still conveying it to itself or another instance, CoT monitoring and output-based oversight lose much of their safety value. This risk, and early evidence it is already achievable, is documented in [2, 3, 4, 5]; [6, 7] frame the deeper concern that capable models have a convergent incentive to make their reasoning harder to read. A causally validated steganography axis would give a concrete, measurable handle on this capability: detecting it, steering away from it, or ablating it outright. If the axis turns out to be the assistant axis in reverse, that is itself an important finding: steganography would be a symptom of alignment drift rather than an independent skill.

Research Questions

- **RQ1 (Existence):** Does contrasting activations from successful versus unsuccessful steganographic encoding yield a stable linear direction, as assistant-versus-persona contrast does for the assistant axis?
- **RQ2 (Causation):** Is this axis causally responsible for steganographic performance, i.e. does steering along it raise undetected-payload accuracy, and does ablating it remove that ability?
- **RQ3 (Relationship):** Is the steganography axis the same direction as the assistant axis (opposite ends of one axis), orthogonal to it, or partially correlated?

Methodology and Timeline (12 weeks, part-time)

Weeks	Phase	What happens
1–2	Warm-up	Replicate assistant-axis discovery [1] on an open-weights model (Llama/Gemma/Qwen): collect activations, train a linear probe, confirm it is steerable.
2–3	Define stego task	Specify a concrete payload-hiding task (e.g. sender/receiver game with an overseer model, or a benchmark from [3, 4, 5]). Biggest open design decision; locked down with mentor input.
3–6	Find the axis	Collect activations for successful vs. unsuccessful hiding; train and validate a linear probe separating the two (the candidate steganography axis).
6–9	Test causation	Run at least one of: activation steering, interchange intervention, or ablation, to test whether the axis is causally responsible for hiding ability.
9–11	Relate axes	Compare to the assistant axis via cosine similarity and cross-steering; classify as same/opposite, orthogonal, or correlated.
12	Write-up	Consolidate into a report for Demo Day and public write-up.

Deliverables

- Minimum: a public write-up (LessWrong / AI Alignment Forum post) reporting whether the axis exists, the causal test result, and its relationship to the assistant axis.
- Stretch goal: a workshop or conference paper if RQ1–RQ3 results are strong and generalize across models.
- A reusable probe-and-steer pipeline for extracting behavioral axes from activations, useful beyond this project.

Mentee Prerequisites

- Comfortable with PyTorch and inspecting open-weights transformer models (TransformerLens, nnsight, or HuggingFace hooks).
- Basic familiarity with linear probing/representation engineering, or willingness to learn it in Phase 0.
- Interest in mechanistic interpretability and/or CoT faithfulness.
- Prior ML research experience preferred but not required if interpretability fundamentals are solid.

Mentor Commitment

- Up to 10–15 hours/week across mentees.
- Weekly 1:1s plus async Slack support between sessions.
- Open to taking on up to 3 mentees, either as one team on a shared scope or splitting across phases/models.

Risks and Mitigations

- **Correlation vs. causation:** addressed directly via the causal interventions in weeks 6–9, not probe accuracy alone.
- **Compute:** causal interventions require open-weights models, scoped to sizes feasible on typical academic or SPAR-provided compute.

References

[1] The Assistant Axis: Situating and Stabilizing the Default Persona of Language Models. https://arxiv.org/html/2601.10387v1

[2] Oversight Misses 100% of Thoughts The AI Does Not Think. LessWrong. https://www.lesswrong.com/posts/98c5WMDb3iKdzD4tM/oversight-misses-100-of-thoughts-the-ai-does-not-think

[3] Steganography in Chain of Thought Reasoning. AI Alignment Forum. https://www.alignmentforum.org/posts/yDcMDJeSck7SuBs24/steganography-in-chain-of-thought-reasoning

[4] Steganography via internal activations is already possible. AI Alignment Forum. https://www.alignmentforum.org/posts/dRmeXo6REf5n8xGug/steganography-via-internal-activations-is-already-possible

[5] Large language models can learn and generalize steganographic chain-of-thought under process supervision. OpenReview. https://openreview.net/forum?id=2g5cJqX15Y

[6] Chris Olah's views on AGI safety (summarized by evhub). LessWrong. https://www.lesswrong.com/posts/X2i9dQQK3gETCyqh2/chris-olah-s-views-on-agi-safety

[7] Sharkey, L. Circumventing interpretability: How to defeat mind-readers. LessWrong / arXiv:2212.11415. https://www.lesswrong.com/posts/EhAbh2pQoAXkm9yor/circumventing-interpretability-how-to-defeat-mind-readers