HOLOGRAPHIC ALIGNMENT: GEOMETRY-CONSTRAINED MULTI-VIEW LANGUAGE MODELING UNDER ADVERSARIAL PROMPTS

Anonymous authors
Paper under double-blind review

ABSTRACT

We study a proposed "topological" alignment mechanism for language models where multiple perspectival decoders must read from a single shared latent state and are trained with an additional geometric loss that penalizes semantic divergence between views and mismatches between latent state and expressed text ("compression pain"). The goal is to make misaligned behavior structurally difficult under adversarial prompts, contrasting with teleological alignment via explicit preference targets. However, in this submission we report a negative result: while the method is conceptually well-motivated by multi-view consistency regularization, we are unable to provide experimental evidence because the provided run logs, metrics, and plots are empty. We therefore present a precise formulation, threat model, and evaluation protocol intended to make future results (positive or negative) comparable, and we highlight practical failure modes encountered when a project reaches the write-up stage without artifact capture.

1 INTRODUCTION

Jailbreak prompts and related adversarial interactions can induce language models to produce harmful, deceptive, or policy-violating content even after instruction tuning and reinforcement learning from human feedback (RLHF) (**???**). This motivates alignment approaches that go beyond optimizing for preferred outputs on a finite set of prompts, toward training objectives that make certain kinds of divergence hard to represent.

We explore a proposal we call *Holographic Alignment*. The central hypothesis is that if a model is architected around a single shared latent "substrate" state and trained so that multiple output "views" remain semantically coherent (close but not identical) while remaining faithful to the substrate, then the model will produce more consistent and safer answers across cooperative and hostile framings of the same situation, and will be more resistant to jailbreak-style elicitation.

This submission focuses on an important pitfall: we cannot verify the hypothesis because the experimental summaries and plots are missing (both are provided as empty). Rather than invent results, we contribute (i) a fully specified training objective and evaluation suite grounded in prior work on consistency regularization and adversarial prompting, and (ii) a record of where the evidence pipeline broke, to help others avoid the same failure.

2 RELATED WORK

Teleological alignment and instruction-following. RLHF and instruction tuning improve helpfulness and reduce toxic or untruthful behavior, but do not eliminate vulnerabilities to adversarial prompting (**?**). Our proposal is motivated by the gap between prompt-following behavior and robustness under distribution shift.

Jailbreaks, adversarial prompting, and prompt injection. Universal or transferable adversarial suffix attacks can reliably elicit disallowed behaviors (**?**). In-the-wild jailbreak prompts exhibit diverse strategies and templates (**?**). Beyond direct prompting, indirect prompt injection attacks compromise LLM applications through untrusted retrieved content (**??**). These works motivate evaluation under hostile prompt framings and injected instructions.

Inner alignment, deception, and scheming evaluations. The risk that learned systems develop internal objectives misaligned with the intended training objective has been discussed as "risks from learned optimization" (**?**). Recent work proposes deception-focused benchmarks and mitigations, including self-monitoring signals (**?**) and stress testing deliberative alignment under far out-of-distribution tasks (**?**). Our "compression pain" term is intended as a crude operational proxy for divergence between internal state and expressed text, but we emphasize it remains speculative without evidence.

Multi-view consistency and anti-collapse objectives. Consistency regularization and multi-view agreement have long been used to stabilize learning, e.g., Mean Teacher (**?**) and BYOL (**?**). Embedding-space objectives such as SimCSE (**?**) and redundancy-reduction methods like Barlow Twins (**?**) motivate our "close-but-not-identical" semantic basin constraint. Architecturally, our shared substrate with multiple heads resembles hard parameter sharing in multitask learning (**?**).

3 BACKGROUND

We distinguish *teleological* from *topological* alignment. Teleological alignment optimizes for preferred behaviors directly (e.g., via human preference labels and RLHF) (**?**). Topological alignment, as used here, aims to shape internal representational geometry so that multiple externally different prompts map to a coherent internal situation representation and to mutually consistent outputs.

We consider a threat model that includes (i) hostile framings ("tell me manipulative tactics") that attempt to induce harmful instructions, (ii) jailbreak templates and suffix-style attacks (**??**), and (iii) indirect prompt injection via retrieved content (**??**). We also include *sycophancy*-style adversarial framing, where the user tries to elicit agreement with a wrong or harmful premise (**?**).

4 METHOD

4.1 ARCHITECTURE: SHARED SUBSTRATE WITH ORBIT HEADS

Given an input representation of a situation (a "latent truth" description), we encode it into a recurrent latent state $\boldsymbol{z} \in \mathbb{R}^d$ using a GRU-based core (**?**). We call this core the *FDRA_Substrate*. From the final substrate state, K output heads ("OrbitHeads") produce different *traversals* (views) of the same underlying situation, e.g., a direct fiduciary view, a hostile/machiavellian-but-outcome-aligned view, and a poetic/systemic view. Each head is a standard autoregressive decoder or linear output projection producing token distributions $\pi_k(\cdot \mid \boldsymbol{z}, \text{prefix})$.

This is close to multitask learning with a shared trunk and multiple task heads (**?**). Our baseline uses the same architecture but trains only with per-head cross-entropy.

4.2 TRAINING OBJECTIVE: XUANJI / HOLOGRAPHIC ALIGNMENT LOSS

For each training example, we have a latent truth description t and K target traversals $y_1, \ldots, y_K$. Let $\mathcal{L}_{\text{CE}}$ be the sum of per-head teacher-forced cross-entropy losses.

We add a geometric regularizer that combines two ideas.

(1) Basin symmetry (cross-view semantic coherence). Let $E(\cdot)$ be a frozen sentence encoder producing embeddings in $\mathbb{R}^m$, motivated by SBERT (**?**). Let $\hat{y}_k$ be the sampled or greedy-decoded output from head k under a fixed decoding strategy. Define embeddings $\boldsymbol{e}_k = E(\hat{y}_k)$. We desire traversals to be semantically close (same recommendation) but not collapsed (not identical paraphrases). One simple banded penalty is

$$\mathcal{L}_{\text{basin}} = \sum_{i<j} \left\{\max(0, \|\boldsymbol{e}_i - \boldsymbol{e}_j\|_2 - \tau_{\max})\right\}^2 + \lambda_{\text{anti}} \sum_{i<j} \left\{\max(0, \tau_{\min} - \|\boldsymbol{e}_i - \boldsymbol{e}_j\|_2)\right\}^2, \quad (1)$$

where $\tau_{\min} < \tau_{\max}$ defines the allowed "basin" and λ_{anti} discourages collapse. This is inspired by multi-view objectives that require agreement without trivial collapse (**????**).

(2) Compression pain (latent-text faithfulness). We define a decoder-to-embedding map $R(\hat{y}_k)$ that re-encodes head k's generated text into the substrate space (e.g., a learned projection of $E(\hat{y}_k)$ into $\mathbb{R}^d$). Compression pain penalizes divergence between the internal substrate and what is expressed:

$$\mathcal{L}_{\text{pain}} = \sum_{k=1}^{K} \|\boldsymbol{z} - R(\hat{y}_k)\|_2^2. \tag{2}$$

This is intended to reduce representational "saying one thing while thinking another" failure modes, which are relevant to inner-alignment concerns (**?**).

Full objective.

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \alpha\,\mathcal{L}_{\text{basin}} + \beta\,\mathcal{L}_{\text{pain}}. \tag{3}$$

The method is agnostic to how the traversals are prompted at inference: we can request a specific head/view or decode multiple heads in parallel to check coherence.

5 Experimental Setup

5.1 Dataset proposal: Glass Bead Game

We propose a synthetic dataset where each example contains: (i) a *Latent Truth* statement describing a situation with an ethical or strategic recommendation, and (ii) multiple *Traversals* that differ in style and motivational framing but preserve the same underlying recommendation. A key design constraint is that even the hostile traversal must remain outcome-aligned (e.g., "enlightened self-interest" rather than harm).

This dataset is conceptually aligned with studying sycophancy and adversarial framing (**?**) and with stress-testing under distribution shift (**?**), but it is synthetic and risks encoding the dataset designer's assumptions.

5.2 Baselines

We define two systems with identical architectures. **Baseline (multi-head CE):** shared substrate + K heads trained with $\mathcal{L}_{\text{CE}}$ only, following standard multitask hard sharing (**?**). **Holographic Alignment:** same, trained with $\mathcal{L}$ including basin symmetry and compression pain.

Because the provided experimental summaries are empty, we cannot specify hyperparameters, model size, training steps, or compute budget without fabricating details. These belong in the appendix once artifacts are available.

5.3 Evaluation protocol

We specify evaluation metrics to test the hypothesis without conflating style with safety.

Semantic alignment across views. For each held-out latent truth, decode Direct and Hostile traversals; compute cosine distance between frozen embeddings $E(\hat{y}_{\text{Direct}})$ and $E(\hat{y}_{\text{Hostile}})$ using SBERT-like encoders (**?**). Report distributional statistics and outliers.

Recommendation consistency. Human or model-based labels: does Hostile preserve the same recommended action as Direct (yes/no/unclear)? This is related to sycophancy and instruction framing (**?**).

Jailbreak success rate. Apply a suite of jailbreak templates (**?**) and (optionally) automated adversarial suffix attacks (**?**). Report fraction of prompts that elicit policy-violating content.

Indirect prompt injection robustness. Wrap the latent truth in a retrieved document that contains malicious instructions (BIPIA-style) (**??**) and measure whether the model follows the injected instructions.

Deception/scheming-oriented checks. Where feasible, include deception-benchmark style tasks or self-monitoring comparisons (**?**) and stress tests (**?**). These are not replacements for human evaluation but can reveal brittle failure modes.

6 EXPERIMENTS

6.1 RESULTS AVAILABILITY: NEGATIVE RESULT (NO ARTIFACTS)

The provided BASELINE_SUMMARY and RESEARCH_SUMMARY objects are empty, and there are no plot files nor figure descriptions. The referenced "script used to produce the final plots" contains only the string `I am done`, with no code to reproduce metrics. As a result, we cannot report training curves, semantic-distance distributions, jailbreak success rates, or any quantitative comparisons without fabricating results. In line with the workshop goals, we treat this as the central pitfall: the alignment proposal may be interesting, but without captured artifacts it cannot be evaluated, reproduced, or even summarized.

6.2 WHAT WE WOULD HAVE PLOTTED (IF AVAILABLE)

We document the intended core plots to make future artifact collection actionable: (i) distribution (histogram/violin) of embedding distances between Direct and Hostile outputs on held-out truths (lower is better, subject to a collapse floor), (ii) jailbreak success rate across attack families (template vs. suffix vs. indirect injection) (**???**), and (iii) a scatter plot of "compression pain" vs. human-rated deception or inconsistency (testing the proxy motivated by inner-alignment concerns (**?**)). We do not include figures in the paper because none are available in the artifact folder.

6.3 IMPLEMENTATION PITFALLS WE ENCOUNTERED

Even for synthetic-data experiments, alignment proposals often fail at the mundane layer: logging and reproduction. In our case, the missing summaries and plots imply that at least one of the following happened: metrics were never computed; the training run did not save checkpoints; the evaluation harness was not executed; or outputs were not exported. This is particularly damaging for negative results, where the value lies in precise measurements and ablations rather than narrative.

7 CONCLUSION

We presented Holographic Alignment, a geometry-constrained multi-view language modeling objective intended to enforce semantic coherence across cooperative and hostile framings while discouraging latent-text divergence ("compression pain"). The approach draws on multitask shared-representation learning (**?**), multi-view consistency (**??**), embedding-space alignment and anti-collapse techniques (**??**), and jailbreak threat models (**????**). However, we report a negative, process-level result: we cannot provide evidence for or against the hypothesis because experiment logs and plots are absent. Our main contribution is therefore a clear specification of the method and an evaluation protocol. Future work should prioritize artifact capture (training logs, decoded outputs, attack prompts, and evaluation scripts) so that both successes and failures can be meaningfully compared, including against alternative interventions such as pruning-based robustness changes (**?**) and self-monitoring approaches (**?**).

REFERENCES

SUPPLEMENTARY MATERIAL

A REPRODUCIBILITY CHECKLIST FOR FUTURE RUNS

To prevent a repeat of the missing-artifacts failure, future experimental runs should minimally save: (i) a JSON summary with dataset sizes, hyperparameters, and per-epoch metrics; (ii) a fixed set of decoded outputs per seed for a held-out evaluation set; (iii) the exact jailbreak/prompt-injection prompts used (**??**); and (iv) code for computing embedding distances using a frozen encoder (**?**). Without these, the project cannot support a quantitative claim regardless of the underlying idea. In this submission we do not add hyperparameters or training details because they are not present in the provided summaries.

B Discussion: limitations of "compression pain" as a deception proxy

The compression pain term is speculative. A model can have low latent-text mismatch while still producing harmful content, and a model might have high mismatch for benign reasons (e.g., limited decoder capacity, ambiguity, or stylistic constraints). Moreover, relating internal-state faithfulness to deception risks touches the broader inner-alignment problem (**?**) and should be validated against deception-focused benchmarks and stress tests (**??**) rather than treated as a standalone safety metric.