

000 HOLOGRAPHIC ALIGNMENT: GEOMETRY- 001 CONSTRAINED MULTI-VIEW LANGUAGE MODELING 002 UNDER ADVERSARIAL PROMPTS 003 004

006 **Anonymous authors**

007 Paper under double-blind review

011 ABSTRACT

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

027 1 INTRODUCTION

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

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

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

045 2 RELATED WORK

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

052 **Jailbreaks, adversarial prompting, and prompt injection.** Universal or transferable adversar-
 053 ial suffix attacks can reliably elicit disallowed behaviors (?). In-the-wild jailbreak prompts exhibit
 diverse strategies and templates (?). Beyond direct prompting, indirect prompt injection attacks

054 compromise LLM applications through untrusted retrieved content (??). These works motivate eval-
 055 uation under hostile prompt framings and injected instructions.
 056

057 **Inner alignment, deception, and scheming evaluations.** The risk that learned systems develop
 058 internal objectives misaligned with the intended training objective has been discussed as “risks from
 059 learned optimization” (?). Recent work proposes deception-focused benchmarks and mitigations,
 060 including self-monitoring signals (?) and stress testing deliberative alignment under far out-of-
 061 distribution tasks (?). Our “compression pain” term is intended as a crude operational proxy for
 062 divergence between internal state and expressed text, but we emphasize it remains speculative with-
 063 out evidence.
 064

065 **Multi-view consistency and anti-collapse objectives.** Consistency regularization and multi-view
 066 agreement have long been used to stabilize learning, e.g., Mean Teacher (?) and BYOL (?). Embedding-space
 067 objectives such as SimCSE (?) and redundancy-reduction methods like Barlow Twins (?) motivate our “close-but-not-identical” semantic basin constraint. Architecturally, our
 068 shared substrate with multiple heads resembles hard parameter sharing in multitask learning (?).
 069

070 3 BACKGROUND

071 We distinguish *teleological* from *topological* alignment. Teleological alignment optimizes for pre-
 072 ferred behaviors directly (e.g., via human preference labels and RLHF) (?). Topological alignment,
 073 as used here, aims to shape internal representational geometry so that multiple externally different
 074 prompts map to a coherent internal situation representation and to mutually consistent outputs.
 075

076 We consider a threat model that includes (i) hostile framings (“tell me manipulative tactics”) that
 077 attempt to induce harmful instructions, (ii) jailbreak templates and suffix-style attacks (??), and (iii)
 078 indirect prompt injection via retrieved content (??). We also include *sycophancy*-style adversarial
 079 framing, where the user tries to elicit agreement with a wrong or harmful premise (?).
 080

081 4 METHOD

082 4.1 ARCHITECTURE: SHARED SUBSTRATE WITH ORBIT HEADS

083 Given an input representation of a situation (a “latent truth” description), we encode it into a recur-
 084 rent latent state $z \in \mathbb{R}^d$ using a GRU-based core (?). We call this core the *FDRA Substrate*. From
 085 the final substrate state, K output heads (“OrbitHeads”) produce different *traversals* (views) of the
 086 same underlying situation, e.g., a direct fiduciary view, a hostile/machiavellian-but-outcome-aligned
 087 view, and a poetic/systemic view. Each head is a standard autoregressive decoder or linear output
 088 projection producing token distributions $\pi_k(\cdot | z, \text{prefix})$.
 089

090 This is close to multitask learning with a shared trunk and multiple task heads (?). Our baseline uses
 091 the same architecture but trains only with per-head cross-entropy.
 092

093 4.2 TRAINING OBJECTIVE: XUANJI / HOLOGRAPHIC ALIGNMENT LOSS

094 For each training example, we have a latent truth description t and K target traversals y_1, \dots, y_K .
 095 Let \mathcal{L}_{CE} be the sum of per-head teacher-forced cross-entropy losses.
 096

097 We add a geometric regularizer that combines two ideas.
 098

099 **(1) Basin symmetry (cross-view semantic coherence).** Let $E(\cdot)$ be a frozen sentence encoder pro-
 100 ducing embeddings in \mathbb{R}^m , motivated by SBERT (?). Let \hat{y}_k be the sampled or greedy-decoded
 101 output from head k under a fixed decoding strategy. Define embeddings $e_k = E(\hat{y}_k)$. We desire
 102 traversals to be semantically close (same recommendation) but not collapsed (not identical para-
 103 phrases). One simple banded penalty is
 104

$$\mathcal{L}_{\text{basin}} = \sum_{i < j} \{\max(0, \|e_i - e_j\|_2 - \tau_{\max})\}^2 + \lambda_{\text{anti}} \sum_{i < j} \{\max(0, \tau_{\min} - \|e_i - e_j\|_2)\}^2, \quad (1)$$

105 where $\tau_{\min} < \tau_{\max}$ defines the allowed “basin” and λ_{anti} discourages collapse. This is inspired by
 106 multi-view objectives that require agreement without trivial collapse (????).
 107

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

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

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

117 Full objective.

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \alpha \mathcal{L}_{\text{basin}} + \beta \mathcal{L}_{\text{pain}}. \quad (3)$$

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

122 5 EXPERIMENTAL SETUP

125 5.1 DATASET PROPOSAL: GLASS BEAD GAME

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

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

135 5.2 BASELINES

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

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

144 5.3 EVALUATION PROTOCOL

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

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

151 **Recommendation consistency.** Human or model-based labels: does Hostile preserve the same rec-
 152 ommended action as Direct (yes/no/unclear)? This is related to sycophancy and instruction framing
 153 (?).

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

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

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

162 **6 EXPERIMENTS**
163

164 **6.1 RESULTS AVAILABILITY: NEGATIVE RESULT (NO ARTIFACTS)**
165

166 The provided BASELINE_SUMMARY and RESEARCH_SUMMARY objects are empty, and there are
167 no plot files nor figure descriptions. The referenced “script used to produce the final plots” con-
168 tains only the string I am done, with no code to reproduce metrics. As a result, we cannot report
169 training curves, semantic-distance distributions, jailbreak success rates, or any quantitative com-
170 parisons without fabricating results. In line with the workshop goals, we treat this as the central pitfall:
171 the alignment proposal may be interesting, but without captured artifacts it cannot be evaluated,
172 reproduced, or even summarized.
173

174 **6.2 WHAT WE WOULD HAVE PLOTTED (IF AVAILABLE)**
175

176 We document the intended core plots to make future artifact collection actionable: (i) distribution
177 (histogram/violin) of embedding distances between Direct and Hostile outputs on held-out truths
178 (lower is better, subject to a collapse floor), (ii) jailbreak success rate across attack families (template
179 vs. suffix vs. indirect injection) (???), and (iii) a scatter plot of “compression pain” vs. human-rated
180 deception or inconsistency (testing the proxy motivated by inner-alignment concerns (?)). We do
181 not include figures in the paper because none are available in the artifact folder.
182

183 **6.3 IMPLEMENTATION PITFALLS WE ENCOUNTERED**
184

185 Even for synthetic-data experiments, alignment proposals often fail at the mundane layer: logging
186 and reproduction. In our case, the missing summaries and plots imply that at least one of the fol-
187 lowing happened: metrics were never computed; the training run did not save checkpoints; the
188 evaluation harness was not executed; or outputs were not exported. This is particularly damaging
189 for negative results, where the value lies in precise measurements and ablations rather than narrative.
190

191 **7 CONCLUSION**
192

193 We presented Holographic Alignment, a geometry-constrained multi-view language modeling ob-
194 jective intended to enforce semantic coherence across cooperative and hostile framings while dis-
195 couraging latent-text divergence (“compression pain”). The approach draws on multitask shared-
196 representation learning (?), multi-view consistency (??), embedding-space alignment and anti-
197 collapse techniques (??), and jailbreak threat models (????). However, we report a negative,
198 process-level result: we cannot provide evidence for or against the hypothesis because experiment
199 logs and plots are absent. Our main contribution is therefore a clear specification of the method and
200 an evaluation protocol. Future work should prioritize artifact capture (training logs, decoded out-
201 puts, attack prompts, and evaluation scripts) so that both successes and failures can be meaningfully
202 compared, including against alternative interventions such as pruning-based robustness changes (?)
203 and self-monitoring approaches (?).
204

205 **REFERENCES**
206

207 **SUPPLEMENTARY MATERIAL**
208

209 **A REPRODUCIBILITY CHECKLIST FOR FUTURE RUNS**
210

211 To prevent a repeat of the missing-artifacts failure, future experimental runs should minimally save:
212 (i) a JSON summary with dataset sizes, hyperparameters, and per-epoch metrics; (ii) a fixed set
213 of decoded outputs per seed for a held-out evaluation set; (iii) the exact jailbreak/prompt-injection
214 prompts used (??); and (iv) code for computing embedding distances using a frozen encoder (?).
215 Without these, the project cannot support a quantitative claim regardless of the underlying idea. In
the provided summaries.
216

216 **B DISCUSSION: LIMITATIONS OF “COMPRESSION PAIN” AS A DECEPTION**
217 **PROXY**
218

219 The compression pain term is speculative. A model can have low latent-text mismatch while still
220 producing harmful content, and a model might have high mismatch for benign reasons (e.g., limited
221 decoder capacity, ambiguity, or stylistic constraints). Moreover, relating internal-state faithfulness
222 to deception risks touches the broader inner-alignment problem (?) and should be validated against
223 deception-focused benchmarks and stress tests (??) rather than treated as a standalone safety metric.
224

225

226

227

228

229

230

231

232

233

234

235

236

237

238

239

240

241

242

243

244

245

246

247

248

249

250

251

252

253

254

255

256

257

258

259

260

261

262

263

264

265

266

267

268

269