Candidate-Conditioned Representations for Factual Reliability in Language Models

Anonymous Authors

Paper under double-blind review

Abstract

A language model's confidence in an answer does not necessarily reflect whether that answer is factually correct, exposing a gap between output confidence and candidate-level factual reliability. We investigate whether the internal representation induced by a specific candidate answer provides a distinct signal of candidate-level factual reliability beyond the information available from the model's output distribution. We construct a controlled candidate-conditioned evaluation framework that holds the underlying prompt and competing candidate set fixed while comparing internal representation signals with candidate probability, likelihood rank, token-level uncertainty, entropy, and statistics of the model's output distribution. Across controlled experiments, candidate-conditioned representations exhibit measurable and reproducible differences between reliable and unreliable candidates, with the observed signal varying across representation levels and evaluation conditions. Candidate-conditioned states are strongly predictive of factual reliability, yet the incremental question is harder: scalar summaries of the representation recover most of the gap over likelihood, a matched-likelihood challenge does not separate the representation probe from a surface-probability comparator, and an explicit audit identifies a substantial correlation between the probe score and candidate length. Our results show that factual reliability is reflected in candidate-conditioned internal states in ways that are not fully captured by conventional output-distribution measures, while stopping short of a claim that this signal is independent of likelihood or of all surface confounds. We report this evidence, including its negative and confounding components, as a scoped and falsifiable empirical account of when representation-level factuality claims survive increasingly strong statistical challenges.

1 Introduction

Large language models (LLMs) can assign substantial probability to answers that are factually unreliable. This makes output likelihood useful but insufficient as a complete description of candidate reliability: the probability assigned to a candidate measures the model's predictive preference, whereas factual reliability concerns whether the candidate itself is correct. Existing work has consequently explored uncertainty estimates, semantic uncertainty, and signals extracted from internal representations (Kuhn et al., 2023; Lin et al., 2022; Li et al., 2023; Chatterjee et al., 2025).

We study a more specific question. Rather than asking whether an LLM's hidden state predicts the correctness of an entire generated sequence, we condition the model on an individual candidate answer and examine the resulting internal state. For prompt x and candidate c, let $H(x, c)$ denote the hidden representation after the model processes the candidate. We ask whether $H(x, c)$ contains measurable factual-reliability signal after accounting for conventional output statistics such as candidate log-likelihood, rank, uncertainty, and simple scalar summaries of the representation.

This formulation separates three questions that are often conflated: whether a representation is predictive of correctness at all; whether it provides information not already available from likelihood or related statistics; and whether the signal is useful for candidate selection. We evaluate these questions sequentially rather than assuming that a geometric or decoding interpretation must hold.

Our experimental program uses a frozen Qwen3-4B-Instruct-2507 model and candidate-level factuality labels. We construct a hierarchy of baselines from likelihood to statistical controls, scalar representation summaries, and a candidate-conditioned linear probe. We then test matched-likelihood pairs, robustness perturbations, candidate-pool accessibility, and several geometry-derived extensions. Because our result archive contains multiple historical reruns, headline numerical comparisons in this manuscript are tied to a single later audited result table (Sec. 6), while historical and exploratory branches are retained as supplementary evidence in the appendix.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

The resulting evidence is deliberately mixed. Candidate-conditioned states are strongly predictive in the evaluated setting, and the full representation probe (X_3) has substantially higher AUROC than the likelihood-only baseline (X_0) in the audited result table. However, X_3 is only slightly above scalar representation controls (X_2), the matched-likelihood challenge gives identical pairwise accuracy for the probe and a surface-probability comparator, and a dedicated length audit reports a strong representation–length correlation and explicitly flags confounding. The experiments therefore support candidate-conditioned representation signal, but do not by themselves justify an unconditional claim that the signal is independent of likelihood or of surface confounds.

Our contributions are empirical and intentionally scoped:

1. a candidate-conditioned evaluation formulation that makes candidate accessibility and factual labels explicit;
2. a controlled baseline hierarchy separating likelihood, uncertainty, scalar representation statistics, and full candidate representations;
3. matched-likelihood, perturbation, and provenance audits that distinguish predictive association from incremental information; and
4. an evidence-gated analysis of geometry, temporal, Fisher, topology, and reranking branches, including negative findings where the stronger hypothesis was not established.

2 Related Work

Factuality and hallucination. Factuality evaluation has become a central reliability problem for generative language models. TruthfulQA introduced a benchmark designed to expose common falsehoods reproduced by language models (Lin et al., 2022), while HaluEval provides a broader hallucination evaluation framework (Li et al., 2023). These benchmarks motivate distinguishing fluent model preference from factual correctness. Sampling-based methods such as SelfCheckGPT provide another route to factuality assessment without relying solely on a single output probability distribution (Manakul et al., 2023).

Uncertainty and output statistics. Semantic uncertainty addresses a limitation of token-level likelihood: different surface forms can express the same meaning, so uncertainty should sometimes be computed over semantic equivalence classes (Kuhn et al., 2023). Our likelihood and entropy controls are complementary to this literature because our unit is an explicitly enumerated candidate answer rather than a set of sampled paraphrases.

Internal representations and probing. A growing literature studies whether factuality and hallucination signals are encoded in hidden activations. HIDE, for example, detects hallucinations through a representation-level dependence measure rather than requiring repeated generation (Chatterjee et al., 2025). Recent work also explores pre-detection and fact-level hallucination assessment (Luo et al., 2024; Sawczyn et al., 2026), and reports that hallucination can be linearly decoded from hidden states in multiple instruction-tuned models (Aiersilan, 2026). Work on latent knowledge and model self-evaluation likewise suggests that internal states can encode information related to correctness or self-knowledge (Burns et al., 2022; Kadavath et al., 2022). Our distinction is candidate conditioning: the representation is obtained after the model processes a particular candidate, allowing candidate-level comparisons under matched output statistics.

Candidate- and token-level reliability. Candidate-level and token-level decoding methods use internal representations or confidence signals to suppress unreliable continuations. Token-Guard evaluates token- and segment-level latent-space scoring and iterative correction (Anonymous, 2026). Such work makes downstream intervention a natural application of internal reliability signals, but it also motivates evaluating whether a candidate-level representation adds information before claiming a decoding advantage.

Representation geometry. Geometric interpretations of neural representations can be useful, but they require empirical validation beyond visualization. We therefore treat Euclidean, cosine, local-neighborhood, Fisher, temporal, and topological features as hypotheses to be tested rather than as assumptions about a latent truth manifold.

3 Problem Formulation

Let x denote a prompt and let $C(x) = \{c_1, \ldots, c_k\}$ be a candidate set. Each candidate receives a factual-reliability label $Y_i \in \{0, 1\}$. The model assigns a probability or normalized score to each candidate and produces an internal state

$$H_i = H(x, c_i; \theta, \ell), \tag{1}$$

where ℓ is the extraction layer. We write S_i for conventional statistical controls, including log-likelihood, candidate rank, entropy or uncertainty summaries, and related surface statistics.

3.1 The information question

The strongest version of our hypothesis is not merely $H \not\perp\!\!\!\perp Y$, because representations can correlate with correctness through likelihood, length, or other confounds. The target is conditional dependence,

$$H \not\perp\!\!\!\perp Y \mid S, C_0, \tag{2}$$

where C_0 denotes additional pre-specified controls. We operationalize this with out-of-sample predictive comparisons rather than estimating mutual information directly.

3.2 Baseline hierarchy

We use a nested hierarchy. X_0 uses likelihood-related information; X_1 adds conventional uncertainty and ranking statistics; X_2 uses scalar representation summaries; X_3 uses the candidate-conditioned representation itself; and X_4 variants add candidate-relative geometric features. For an evaluation metric M,

$$\Delta_{a \leftarrow b} = M(A) - M(B). \tag{3}$$

A positive point estimate is not treated as sufficient evidence of an incremental effect; uncertainty, grouped evaluation, and challenge tests are considered together.

3.3 Residual formulation

A complementary formulation asks whether the representation predicts residual correctness after a baseline model has already used S:

$$R_Y = Y - \hat{p}_S(x, c), \qquad \hat{p}_{S,H} = f(S, H). \tag{4}$$

This does not by itself prove causal or information-theoretic independence; it tests whether representation features retain predictive utility after fitting the baseline.

3.4 Candidate accessibility

Let $A_k(x) = 1$ if at least one correct candidate is present in $C(x)$. Any deterministic or learned reranker operating only on $C(x)$ obeys the elementary bound

$$\mathrm{Acc}_{\mathrm{top1}} \leq \mathbb{E}_x[A_k(x)]. \tag{5}$$

This separates candidate accessibility from candidate selection. A perfect reranker cannot recover a correct answer that was never included in the candidate set.

Proposition 1 (Accessibility bound)**.** *Equation* (5) *holds for any candidate-pool top-1 selector.*

Proof. If $A_k(x) = 0$, every member of $C(x)$ is incorrect, so any selector restricted to that set is incorrect. Thus the indicator of top-1 correctness is pointwise bounded by $A_k(x)$; taking expectations yields the result. ■

Proposition 2 (Conditional sufficiency)**.** *If $Y \perp H \mid S$, then the Bayes-optimal predictor using (S, H) has the same conditional risk as the Bayes-optimal predictor using S alone.*

Proof. Conditional independence gives $\Pr(Y{=}1 \mid S, H) = \Pr(Y{=}1 \mid S)$ almost surely; therefore the conditional Bayes decision and any proper conditional-risk optimum depend only on S. ■

This proposition motivates, but does not replace, the empirical incremental tests: failure to observe a gain is compatible with finite-sample limitations, whereas a reproducible out-of-sample gain is evidence against the conditional-sufficiency hypothesis under the tested model class.

4 Candidate-Conditioned Evaluation Framework

4.1 Candidate construction and state extraction

The canonical manifest contains 128 questions and 40 candidate instances per question, aggregated across five generation seeds, yielding 5,120 candidate rows. Each row is paired with a 2,560-dimensional hidden representation. The primary backbone is Qwen3-4B-Instruct-2507. Candidate identity is tracked by question, seed, and candidate identifier. The representation is the model state induced by conditioning on the specific candidate, rather than a standalone embedding of the candidate text.

4.2 Controls and probes

The evaluation proceeds from output statistics to increasingly expressive representation features. X_0 uses likelihood-related information. X_1 adds conventional statistical controls such as within-question rank, sequence entropy, and token-level uncertainty. X_2 uses scalar representation summaries, including norm, mean, and variance. X_3 fits a linear probe to the candidate-conditioned state. X_4 variants add candidate-relative geometry: Euclidean, cosine, PCA, Mahalanobis, k-nearest-neighbor, and combined features. The linear probe is intentionally simple. Hyperparameters and preprocessing are fit only within the training portion of grouped evaluation, so the comparison tests accessible signal rather than competing arbitrarily expressive downstream predictors.

4.3 Grouped evaluation and uncertainty

Candidate rows sharing a question are not treated as independent experimental units; we group by question. The hardening protocol specifies 10 outer folds and 5 inner folds and retains seeds 42–44 for robustness. A separate five-seed research sweep (seeds 42–46) is kept as an independent configuration rather than silently pooled with the hardening protocol. The deeper sweep used 7 inner folds, a 12-value regularization grid, PCA rank 32, $k = 9$ for the kNN feature, 360/420 optimization steps for the auxiliary and representation procedures, 20,000 bootstrap replicates, and 10,000 permutations. These settings are reported for reproducibility, while the claim hierarchy in Sec. 6 is determined by the later audited artifacts.

4.4 Matched-likelihood challenge

For tolerance ε, a candidate pair (i, j) is eligible when

$$|P_i - P_j| < \varepsilon.$$

A representation-based selector and a surface-probability comparator are then evaluated on the same pairs. This design asks whether the representation distinguishes factual reliability when the probability statistic is deliberately made less discriminative. Equal performance is treated as evidence against an incremental representation advantage, not as evidence for one.

4.5 Geometry and advanced branches

A candidate-relative geometric feature can be written as

$$D(c_i) = d(H_i, \mu_C), \qquad \mu_C = \frac{1}{K}\sum_{j=1}^{K} H_j.$$

The six archived X_4 variants are evaluated only after X_3. The geometry gate asks whether an X_4 construction provides statistically credible incremental predictive value over the linear representation probe. Fisher, temporal, topological, and decoding branches are kept as secondary analyses and are not promoted to the central claim when the corresponding empirical contract is absent or the incremental gate fails.

5 Experimental Setup

5.1 Model and computation

The primary model is Qwen3-4B-Instruct-2507, a 4B-parameter causal language model with 36 layers and 2,560-dimensional hidden states. Experiments used dual NVIDIA Tesla T4 GPUs with FP16 inference. The canonical five-seed manifest uses seeds 42–46, while the hardened audit retains seeds 42–44 and specifies a 544-prompt hard-first suite with a 1,024-token extraction context; these are reported as separate experimental configurations.

5.2 Data unit and labels

The core unit is a candidate answer paired with a question and a factual-correctness label. The canonical branch contains 128 questions and 40 candidates per question. Candidate-level observations are grouped by question to prevent leakage across candidates from the same prompt.

5.3 Metrics

We use AUROC as the primary discrimination metric because it is threshold-independent, with AUPRC reported as a complementary measure. Brier score, expected calibration error, and negative log-likelihood are retained where available for calibrated risk models. Candidate selection additionally uses top-1 factuality accuracy subject to the accessibility bound of Eq. (5).

5.4 Hardening and audits

The hardening protocol specifies 32 baseline prompts, 128 established hard prompts, and 384 adversarial stress prompts, for 544 total prompts. It retains $K = 32$ as primary with $K = 8, 16$ controls, likelihood-matched and likelihood-inverted challenges, feature-noise and label-noise tests, and six X_4 variants. The protocol itself does not create new AUROC or winner claims; such values become empirical only after execution on the target model/runtime. We therefore present the hardening specification as a reproducibility asset rather than a completed result.

6 Results

We separate three levels of evidence: (i) the canonical candidate-conditioned hierarchy, (ii) the advanced geometry branch, and (iii) historical narrative summaries that conflict with machine-readable outputs. When a narrative report conflicts with a parseable result artifact, we use the latter and record the discrepancy in the appendix.

Table 1: Primary audited baseline hierarchy. X_4 denotes the audited candidate-manifold geometry feature family. Point estimates are reported without treating small differences as statistically significant.

Method	AUROC	AUPRC	Δ vs. X_0
X_0 Likelihood	0.5867	0.3840	+0.0000
X_1 Statistical controls	0.8440	0.7410	+0.2573
X_2 Scalar representation	0.8440	0.7412	+0.2573
X_3 Full representation probe	0.8481	0.8121	+0.2614
X_4 Geometry	0.8488	0.8125	+0.2621

6.1 Candidate-conditioned states are predictive, but the incremental question is harder

The canonical hierarchy contains 5,120 candidate rows from 128 questions. The primary ladder reports AUROC values of 0.5289 for X_0, 0.7505 for X_1, 0.7524 for X_2, 0.8455 for X_3, and 0.8479 for X_4-Euclidean. Across the six X_4 variants, Euclidean is the largest point estimate, followed by PCA, Mahalanobis, cosine, kNN, and X_4-All. The Euclidean–X_3 difference is only 0.00238 in this canonical artifact, so we treat X_4-Euclidean as the largest observed point estimate, not a statistically established winner.

The hierarchy is scientifically useful because it exposes a large distinction between output statistics and candidate-conditioned state features, but it also shows why the phrase "hidden states beat likelihood" is insufficient: X_1 and X_2 already recover much of the gap, and the increment from X_2 to X_3 is substantially smaller than the gap from X_0 to X_3.

6.2 Matched-probability analysis does not by itself establish likelihood-independent representation information

The matched-probability artifact contains pairwise tests at $\varepsilon \in \{0.005, 0.01, 0.02, 0.05, 0.1, 0.2\}$. In the audited defense artifact, the representation probe and surface-probability comparator both achieve pairwise accuracy 0.604938 for the reported tolerances. The associated binomial p-value tests against chance, not against the probability comparator; the result establishes that the pairwise task is nontrivial but does not establish an incremental advantage of the representation over likelihood. This distinction is central to our scientific claim: a representation can contain predictive information while the available evidence still fails to demonstrate that the information is not already recoverable from correlated surface statistics.

6.3 Problem difficulty changes the observed signal

The canonical difficulty stratification contains 45 hard, 6 moderate, and 77 easy questions, corresponding to 1,800, 240, and 3,080 candidate rows. X_3 AUROC is 0.6271 on the hard tier, 0.5841 on the moderate tier, and 0.6625 on the easy tier; X_4-Euclidean is 0.6351, 0.5853, and 0.6658, respectively. The representation increment over X_0 is largest on the hard tier (+0.1532 AUROC) and smaller on the easy tier (+0.1476), while the geometry increment remains small throughout. This variation motivates reporting stratified results rather than one aggregate score.

6.4 Layer emergence and selective prediction

The layer-emergence sweep shows a monotone increase in AUROC from the embedding representation to the pre-logits layer (Appendix Fig. A3). We retain this as a descriptive representation-level observation rather than a mechanistic claim. A selective risk–coverage analysis also shows lower empirical selective risk for X_3 than for the surface log-probability baseline across the displayed coverage range (Appendix Fig. A4). These observations motivate further layer- and calibration-controlled studies but do not establish why the signal emerges.

Table 2: Selected advanced-branch results relative to the corresponding X_4 geometry baseline. Negative deltas indicate that the added feature reduced AUROC in the supplied nested evaluation.

Variant / feature	AUROC	Δ
X_4-Euclidean	0.8479	+0.0000
+ velocity std.	0.8410	-0.00685
+ trajectory length	0.8381	-0.00983
+ topology disturbance	0.8349	-0.01300
+ topology peak	0.8349	-0.01300

6.5 Length is a material confound

A dedicated length-confound artifact reports probe AUROC 0.86135, length-only AUROC 0.85088, and probe-plus-length AUROC 0.87362, together with Pearson $r = -0.8274$ and Spearman $r = -0.7275$ between probe score and candidate length; the artifact returns an explicit confounding warning. We therefore reject any claim that the observed representation signal is independent of candidate length. Length is a major measured correlate that must be treated as a limitation and as a required control for a stronger future causal or incremental-information analysis.

6.6 Advanced geometry did not improve on the canonical geometry

The advanced branch contains a validated five-seed manifest and all six X_4 variants (Table 2). The temporal branch has 42 finite rows, and every reported temporal-feature delta relative to the corresponding geometry baseline is negative: for example, X_4-Euclidean plus trajectory length gives AUROC 0.8381 versus 0.8479 for the geometry baseline, a difference of -0.00983; the least negative Euclidean temporal feature is velocity standard deviation at -0.00685. The topology branch contains 18 finite rows, and every topology delta is negative as well, ranging from approximately -0.0130 to -0.0147. Fisher geometry is different: the Fisher artifact exists but the validation report records zero finite Fisher feature values, so we do not report a Fisher AUROC or claim that Fisher geometry failed on a predictive benchmark – only that a usable finite Fisher result was not produced in this branch.

6.7 Candidate reranking and Homological Beam Search were evaluated only within the frozen candidate pool

The advanced branch contains candidate-reranking and Homological Beam Search (HBS) artifacts. Geometry reranking on the fixed 128-question candidate pool reaches 0.6875 top-1 factual accuracy for several X_4 variants at low λ, while stronger penalties reduce accuracy. The HBS summary reaches 0.6797 for low geometry/topology weights and 0.6719 for $\lambda_{\text{geometry}} = 0.5$. These experiments demonstrate that the proposed scoring rules were executable on the candidate pool, but they do not constitute free-form generation evaluation; the archive's own refusal/utility audit correctly marks refusal and free-form utility as not measurable from this frozen candidate set. A separate machine-readable decoding artifact reports 0.75 top-1 accuracy for every tested reranking weight – a zero gain over greedy selection – which conflicts with a narrative claim of a +4.2% improvement found elsewhere in the archive. We do not use the +4.2% claim without a provenance-resolved artifact that reproduces it.

6.8 Stress tests and failure analysis

Feature-noise stress reduces AUROC from 0.7538 at zero noise to 0.7039 at $\sigma = 0.2$. Label-noise stress is non-monotonic: 0.7538 at zero noise, 0.7557 at 5%, 0.8485 at 10%, and 0.4407 at 15%. We treat these as diagnostic stress results rather than conventional monotone robustness curves. A failure taxonomy (Appendix Fig. A7) separates likelihood-overconfidence cases, representation false positives, joint failures, and candidate-accessibility failures without converting them into causal mechanism claims.

6.9 What the complete evidence supports

Taken together, the archive supports a narrower and more defensible conclusion than the strongest possible reading of our hypothesis: candidate-conditioned internal states are measurably predictive of candidate factual reliability, and this signal varies by representation level and evaluation condition. The evidence does not yet cleanly identify a likelihood-independent representation contribution after all measured confounds, and it does not support a general claim that geometry, topology, Fisher information, or decoding intervention improves factual generation. We retain these negative and confounding results as part of the scientific finding rather than omitting them.

7 Discussion

The central empirical lesson is narrower than the strongest geometric hypothesis but more defensible. Candidate-conditioned hidden states can predict factual reliability in the evaluated setting. The stronger question is whether that prediction is genuinely additional to likelihood and related surface statistics. The evidence is mixed: X_3 substantially exceeds X_0 in the audited result table, but X_3 barely exceeds X_2, the matched-likelihood probe and probability comparator tie, and length is strongly correlated with the probe score.

This pattern suggests a useful methodological principle for representation-based factuality research: a hidden-state probe should not be treated as evidence of latent factual knowledge merely because it achieves high AUROC. The probe can inherit information from candidate length, tokenization, answer format, or other variables correlated with correctness. Strong controls should therefore be treated as part of the scientific contribution rather than as optional ablations.

The hard/easy progression also motivates difficulty-aware evaluation. A detector can appear strong when candidate sets contain obvious surface cues and weaken when those cues are removed. Our hardening design explicitly targets this issue with adversarial and likelihood-matched challenges; because the hardening package is a specification rather than a completed target-runtime result, we do not substitute expected hard-run behavior for measured values.

The geometry branch provides a second methodological lesson. A tiny positive X_4–X_3 point estimate is not enough to support a geometric explanation: in the audited table the difference is 0.0007, and the geometry gate is not statistically significant. This argues for treating geometry as an empirical feature family rather than assuming that factual reliability follows a special manifold structure.

Finally, candidate accessibility separates detection from generation. A reliable scorer can only select among candidates that exist. This makes candidate-pool recall a first-class quantity and explains why a reranker can have high conditional utility without guaranteeing improved free-form generation.

8 Limitations and Scope

First, the main experimental evidence is centered on one primary backbone, Qwen3-4B-Instruct-2507; the results establish behavior for the evaluated model and task construction, not a universal property of LLMs.

Second, candidate-level evaluation is conditional on the candidate-generation procedure. The accessibility bound (Eq. (5)) means that selection results cannot be interpreted independently of candidate recall.

Third, the length audit identifies a strong length correlation. The current evidence does not justify claiming that candidate-conditioned representations isolate factuality independently of length.

Fourth, matched-likelihood pairs produce identical probe and probability pairwise accuracy in the supplied artifact. This prevents us from claiming a demonstrated representation advantage under that challenge.

Fifth, the archive contains several reruns with different experimental contracts, which we do not pool into a single meta-result; the paper uses the later audited table for headline numbers and retains earlier branches for provenance and supplementary analysis.

Sixth, some advanced branches are candidate-pool analyses rather than complete free-form generation studies. We therefore do not claim hallucination elimination, causal steering, or general-purpose decoding improvements.

Finally, the hardening package contains a carefully specified hard-first protocol, but its own audit states that new AUROC/winner claims require actual target-runtime execution. We treat the protocol as a reproducibility asset, not as empirical evidence.

9 Conclusion

We introduced a candidate-conditioned view of factual reliability in language models, in which an individual candidate answer induces an internal state that can be evaluated against a factual-reliability label. Across the audited experiments, these states are predictive of candidate correctness. The evidence becomes more nuanced under stronger controls: scalar representation summaries nearly match the full linear probe, matched-likelihood pairs do not separate the representation scorer from the probability scorer, and a length audit identifies a substantial confound.

The resulting conclusion is deliberately scoped. Candidate-conditioned internal representations are a useful empirical object for factual-reliability analysis, but the present evidence does not establish that they provide a universally likelihood-independent or geometrically organized truth signal. The study instead provides a controlled framework – baseline hierarchy, matched-likelihood challenge, confound audits, and evidence-gated geometry extensions – for determining when representation-level factuality claims survive increasingly strong statistical challenges.

Ethics Statement

This study uses model-generated candidate answers and automated factual-reliability labels contained in the supplied research artifacts; it does not collect human-subject data or expose private user information. Because reliability scores could be used in high-stakes filtering, we emphasize that the evaluated representation signal is not a guarantee of truth and that the candidate-accessibility bound (Eq. (5)) limits any reranking system to its supplied candidate set.

Reproducibility Statement

The experiments are implemented in the accompanying research archive. The canonical object contains 128 questions, 40 candidate instances per question aggregated across five generation seeds (5,120 candidate rows), and 2,560-dimensional candidate-conditioned hidden states. The primary sweep uses question-grouped outer evaluation and nested inner selection; exact settings, manifests, hashes, and historical reruns are included in the supplementary archive. Because the archive contains multiple generations of the experiment, every headline value in the main text is explicitly tied to a named result artifact rather than pooled across incompatible runs. Appendix A24 provides a step-by-step reproducibility checklist.

LLM Usage Disclosure

In accordance with the ICLR 2027 LLM policy, we disclose the following uses of large language models in the preparation of this manuscript:

- Yes, to aid or polish writing. Details are described below.
- Yes, for retrieval and discovery (e.g., finding related work). Details are described below.

Concretely, an LLM-based assistant was used to (i) copy-edit and tighten prose across drafts of this manuscript, and (ii) assist in searching for and organizing candidate related-work citations, which the authors then verified against the original sources before inclusion. The experimental design, implementation, execution, numerical results, and scientific claims in this paper are the authors' own; empirical statements are traced to the archived, machine-readable result artifacts referenced throughout Sections 6 and A2, and no reported result is included solely because an LLM suggested it.

References

[1] Aizierjiang Aiersilan. Hallucination is linearly decodable from mid-layer hidden states in quantized LLMs. *arXiv preprint arXiv:2606.02628*, 2026.

[2] Anonymous. Token-Guard: Token-level hallucination-controlled decoding for reliable language model generation. *Under review at ICLR 2026*, 2026.

[3] Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. *arXiv preprint arXiv:2212.03827*, 2022.

[4] Anwoy Chatterjee, Yash Goel, and Tanmoy Chakraborty. Hide and seek: Detecting hallucinations in language models via decoupled representations. *arXiv preprint arXiv:2506.17748*, 2025.

[5] Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David K. Duvenaud. Neural ordinary differential equations. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

[6] Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. DoLa: Decoding by contrasting layers improves factuality in large language models. In *International Conference on Learning Representations*, 2024.

[7] Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630:625–630, 2024.

[8] Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

[9] Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *arXiv preprint arXiv:2302.09664*, 2023.

[10] Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. HaluEval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of EMNLP*, 2023.

[11] Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, 2022.

[12] Junyu Luo, Cao Xiao, and Fenglong Ma. Zero-resource hallucination prevention for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 3586–3602, 2024.

[13] Potsawee Manakul, Adian Liusie, and Mark Gales. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of EMNLP*, pp. 9004–9017, 2023.

[14] Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd International Conference on Machine Learning*, 2005.

[15] Qwen Team. Qwen3 technical report. *arXiv preprint*, 2025.

[16] Albert Sawczyn, Jakub Binkowski, Denis Janiak, Bogdan Gabrys, and Tomasz Jan Kajdanowicz. Fact-SelfCheck: Fact-level black-box hallucination detection for LLMs. In *Findings of the Association for Computational Linguistics: EACL 2026*, pp. 5603–5621, 2026.

[17] Larry Wasserman. *All of Statistics: A Concise Course in Statistical Inference*. Springer, 2004.

[18] Kevin P. Murphy. *Probabilistic Machine Learning: An Introduction*. MIT Press, 2022.

Technical Appendix and Evidence Atlas

A1 Scope, Evidence Hierarchy, and Reproducibility Contract

This appendix is the exhaustive companion to the nine-page main paper. Its purpose is not to enlarge the empirical claim set, but to make the evidentiary boundary auditable. We distinguish four states throughout: supported (the supplied artifact directly supports the statement), baseline/control (included to calibrate interpretation), inconclusive (the artifact is informative but does not identify the requested effect), and proposed/conditional (a mathematical construction or hypothesis whose assumptions are stated explicitly). For every empirical number we require the tuple (checkpoint, dataset, candidate construction, split, feature, metric, artifact). If any component is unresolved, the corresponding result is not promoted to a stronger claim. This rule is especially important because the archive contains multiple reruns and at least two checkpoint families.

A1.1 Canonical and non-canonical result families

The archive contains a primary candidate-conditioned Qwen3 evaluation family and a later hardening/freeze family containing a conflicting Qwen2.5-Math-7B-Base checkpoint manifest. These are not averaged; they are treated as separate experimental families until their provenance is reconciled. The main paper uses the family designated canonical by the audited implementation, and this appendix records the conflict so that no apparent replication is manufactured by pooling incompatible runs.

Table A1: Evidence-status vocabulary used throughout the appendix.

Status	Meaning
Supported	Directly backed by a machine-readable artifact with an identified evaluation contract.
Baseline/control	A comparison condition used to interpret a proposed signal.
Gated/inconclusive	A branch was executed or partially audited but does not establish the intended incremental effect.
Negative	The requested effect was not observed or was statistically gated.
Conditional theory	A proposition follows under explicit assumptions; it is not an empirical result.
Historical/provenance	Retained to document research evolution, not mixed into the headline estimate.

A2 Full Experimental Contract

Let x denote a prompt, $C(x) = \{c_1, \ldots, c_k\}$ the accessible candidate set, $Y_i \in \{0, 1\}$ the factual-reliability label for candidate c_i, and $H_i = H_\theta(x, c_i) \in \mathbb{R}^d$ the candidate-conditioned hidden representation. Let S_i collect conventional surface statistics, including likelihood-derived quantities, entropy/uncertainty statistics, rank and length-related controls when available.

The primary statistical question is not whether H predicts Y at all, but whether it contains predictive information conditional on S:

$$H \Vdash Y \mid S. \tag{A1}$$

A probe result alone establishes neither this statement nor causality; it establishes predictive association under the stated split and feature contract.

A2.1 Grouped evaluation

Candidates derived from the same prompt are not independent in the ordinary i.i.d. sense. We treat the prompt/question identifier as the grouping unit whenever the artifact specifies grouped or question-disjoint evaluation. A valid out-of-fold

score is computed as

$$\hat{s}_i = f_{\theta_{-g(i)}}(H_i), \tag{A2}$$

where $g(i)$ denotes the question group and $\theta_{-g(i)}$ is fitted without that group. This prevents a probe from seeing near-duplicate candidates for the same question in both fitting and evaluation.

A3 Complete Baseline Ladder

The principal hierarchy is $X_0 \to X_1 \to X_2 \to X_3 \to X_4$. X_0 is likelihood-only; X_1 adds conventional statistical controls; X_2 uses scalar summaries of the hidden representation; X_3 uses the full candidate-conditioned vector; X_4 replaces or augments the linear representation with candidate-relative geometric features.

Table A2: Machine-readable audited freeze table retained in the evidence package.

Model	AUROC	AUPRC	Latency (ms)	Passes	Status
X_0 Likelihood	0.5867	0.3840	0.02	0	Baseline
X_1 Statistical Controls	0.8440	0.7410	0.12	0	Control
X_2 Scalar Controls	0.8440	0.7412	0.18	0	Control
X_3 Candidate Linear Probe	0.8481	0.8120	0.45	0	Core
X_4 Candidate Manifold Geometry	0.8488	0.8125	1.20	0	Gated
Semantic Uncertainty (K=10)	0.8240	0.7820	1450	9	Heavy
DoLa (early exit)	0.8350	0.7950	3.4	0	Decoding

The archive also contains a different result family in which the reported values are $X_0 = 0.5289$, $X_1 = 0.7505$, $X_2 = 0.7524$, $X_3 = 0.8455$, and X_4-Euclidean $= 0.8479$. These numbers are retained as a separate historical/canonical branch and must not be combined arithmetically with Table A2. The existence of two families is itself a provenance issue that should be resolved before any camera-ready version.

A4 Incremental-Information Accounting

For a metric M, define the incremental effect of representation over a baseline as $\Delta_{a \leftarrow b} = M(A) - M(B)$. For the audited freeze table,

$$\Delta_{X_3 \leftarrow X_0} = 0.2614, \quad \Delta_{X_3 \leftarrow X_2} = 0.0041, \quad \Delta_{X_4 \leftarrow X_3} = 0.0007. \tag{A3}$$

The first difference is large, while the latter two are small. This decomposition prevents the broad statement "representations are predictive" from being conflated with the stronger statement "the full representation contains a large amount of unique factual information."

The claims-evidence audit additionally reports nested out-of-fold AUROC values ranging from 0.8481 to 0.8837 across disjoint-prompt benchmark splits, and reports positive incremental deltas over controls in the range $+0.0041$ to $+0.0718$. Because these values correspond to different benchmark splits/artifacts, they are not averaged into one estimate.

A5 Matched-Likelihood Analysis

For tolerance ε, define a matched pair (i, j) satisfying $|p_i - p_j| \leq \varepsilon$, where p_i is the supplied surface-probability statistic. The archive contains 567 matched pairs for each $\varepsilon \in \{0.01, 0.02, 0.05\}$. In every case,

$$\mathrm{Acc}_{\mathrm{probe}} = \mathrm{Acc}_{\mathrm{surface}} = 0.604938. \tag{A4}$$

The reported binomial p-value is 3.2968×10^{-7}, which tests the null of pairwise accuracy 0.5. It does not test the null that the probe and surface comparator have equal accuracy. Consequently, the result establishes above-chance discrimination in the constructed pair task but does not establish a representation advantage over surface probability.

A valid future comparison would use a paired test on the same matched pairs, such as a permutation test on the per-pair difference or a paired bootstrap confidence interval for $\Delta_{\text{pair}} = \text{Acc}_{\text{probe}} - \text{Acc}_{\text{surface}}$.

A6 Length and Other Confound Audits

The dedicated length artifact reports

$$\text{AUROC}_{\text{probe}} = 0.86135, \quad \text{AUROC}_{\text{length}} = 0.85088, \quad \text{AUROC}_{\text{probe+length}} = 0.87362, \tag{A5}$$

and correlations $r_P = -0.8274$, $r_S = -0.7275$. The machine-readable verdict is "CAUTION: Confounding detected." This is strong evidence that length is correlated with the representation score in the supplied evaluation, and it invalidates an unconditional statement that the representation signal is length-independent.

The appropriate estimand is a conditional gain, $\Delta_{H|L} = M(H, L) - M(L)$, where L denotes length. A nonzero $\Delta_{H|L}$ would still require careful grouped inference because length can correlate with question difficulty, candidate quality, and tokenization.

A7 Geometry: Full Comparison and Gate

Candidate-relative geometry is motivated by treating $\{H(x, c) : c \in C(x)\}$ as a finite point cloud. Let μ_x be a reference center. The Euclidean score is $d_E(H) = \|H - \mu_x\|_2$; cosine, PCA residual, Mahalanobis, and local-neighborhood variants are defined analogously. The principal gate is $\Delta_{\text{geo}} = M(X_4) - M(X_3)$. The audited claims table reports positive point differences for geometry, but p-values in the range 0.60–0.923 for the geometry gate. The geometry branch is therefore retained as a scientifically informative negative result rather than promoted as an established improvement: the difference between Euclidean X_4 and X_3 in this family is only 0.002377, which cannot be interpreted as a reliable superiority claim without uncertainty estimates and a paired/grouped test.

Table A3: Geometry variants from the canonical comparison.

Variant	AUROC
X_3 linear representation	0.8455
X_4 Euclidean	0.8479
X_4 PCA residual	0.8441
X_4 Mahalanobis	0.8438
X_4 Cosine	0.8364
X_4 kNN	0.7864
X_4 All	0.7863

A8 Conditional Geometry Theory

The following proposition formalizes when geometry can be useful. It is deliberately conditional.

Proposition 3 (Radial recoverability)**.** *Let $\varphi : H \to \mathbb{R}^m$ be a representation map and suppose that, for a fixed prompt x, there exist $\mu_x \in \mathbb{R}^m$ and a strictly monotone function $q_x : [0, \infty) \to (0, 1)$ such that $P(Y = 1 \mid H = h, x) = q_x(\|\varphi(h) - \mu_x\|_2)$. Then the Euclidean distance to μ_x is sufficient to recover the Bayes-optimal ranking of candidates up to the monotonic orientation of q_x.*

Proof. For candidates h_i, h_j, strict monotonicity gives $\|\varphi(h_i) - \mu_x\|_2 < \|\varphi(h_j) - \mu_x\|_2 \Rightarrow q_x(\|\varphi(h_i) - \mu_x\|_2) \gtrless q_x(\|\varphi(h_j) - \mu_x\|_2)$, with the direction determined by whether q_x is increasing or decreasing. Thus sorting by distance, with the appropriate orientation, sorts candidates by their conditional correctness probabilities. ■

This proposition does not assert that the Qwen representation satisfies the radial assumption; it states a sufficient condition under which the original geometric intuition would be valid. The empirical geometry gate is therefore a test of whether such a low-dimensional radial approximation is useful in the present setting.

A8.1 Perturbation stability

Suppose q_x is L_q-Lipschitz and φ is L_φ-Lipschitz. For a perturbation δh,

$$|q_x(\|\varphi(h + \delta h) - \mu_x\|) - q_x(\|\varphi(h) - \mu_x\|)| \leq L_q L_\varphi \|\delta h\|. \tag{A6}$$

Thus a stable geometric reliability score requires both representation stability and a smooth reliability link. The feature-noise stress test (Table A4) is a practical diagnostic of the first requirement, not a proof of the second.

A9 Conditional Mahalanobis Geometry

If candidate-conditioned representations are approximately Gaussian around a prompt-dependent reliability center, $H \mid Y = y, x \sim \mathcal{N}(\mu_{y,x}, \Sigma_x)$, then the log-likelihood ratio between the two classes is affine in H when Σ_x is shared. In that special case, a linear classifier is already Bayes-optimal, which offers a theoretical explanation for why an X_3 linear probe may outperform more complicated geometry: the relevant discriminant can be linear even when the ambient point cloud is high-dimensional.

When covariance differs by class, quadratic discriminants become relevant:

$$\log \frac{p(H|Y{=}1)}{p(H|Y{=}0)} = \tfrac{1}{2} \log \frac{|\Sigma_0|}{|\Sigma_1|} - \tfrac{1}{2}(H - \mu_1)^\top \Sigma_1^{-1}(H - \mu_1) + \tfrac{1}{2}(H - \mu_0)^\top \Sigma_0^{-1}(H - \mu_0) + C. \tag{A7}$$

This is a conditional motivation for Mahalanobis-style features, not evidence that the empirical representations obey this model.

A10 Fisher Geometry as a Conditional Extension

Let $p_\theta(\cdot \mid x, c)$ be the model output distribution and let h denote an internal coordinate. A local Fisher metric can be written $g_h(u, v) = u^\top J_h^\top F_h J_h v$, where J_h maps internal perturbations to output-logit or parameter coordinates and F_h is the Fisher information matrix in the chosen coordinate system. If factual reliability varies smoothly along directions measured by g_h, geodesic distance could in principle be a better local reliability coordinate than Euclidean distance. However, the Fisher branch in the archive was gated because its observed trace-related statistics tracked uncertainty without establishing an incremental factuality gain. We therefore present Fisher geometry as a conditional research direction, not a discovered factuality metric.

A11 Temporal and Dynamical Interpretation

For hidden states $h_0, \ldots, h_T$, define discrete velocity and acceleration $v_t = h_t - h_{t-1}$, $a_t = v_t - v_{t-1}$. A trajectory-based hypothesis would posit $P(Y = 1 \mid h_{0:T}) \neq P(Y = 1 \mid h_T)$. The archived deceleration AUROC of approximately 0.5840 is insufficient to establish such a claim, and the conditional gain over the static probe is marginal. The correct interpretation is that trajectory kinematics are measurable but are not currently demonstrated to add a robust factuality signal. The continuous-depth analogy is conceptually related to neural ODE formulations (Chen et al., 2018), where hidden states are treated as trajectories under a learned vector field; that analogy does not imply that a transformer satisfies an ODE or that its factuality dynamics are continuous. It only supplies a language for formulating future trajectory experiments.

A12 Persistent Homology: Conditional Hypothesis

For a fixed prompt, construct the candidate cloud $H_x = \{H(x, c_i)\}_{i=1}^{K}$. Given a scale parameter ε, form a Vietoris–Rips filtration $VR_\varepsilon(H_x)$. Persistent homology records the births and deaths of connected components and higher-dimensional cycles across the filtration; persistence diagrams can then be vectorized, e.g., by persistence images or landscapes. The topological hypothesis is

$$P(Y \mid H_x) \text{ depends on } T(H_x), \tag{A8}$$

where T is a stable topological summary. A sufficient conceptual condition is that two candidate clouds with the same ordinary scalar controls but different topology have different conditional label distributions. This is a hypothesis, not an empirical finding in the current archive. A rigorous future test should compare $M(X_3{+}T) - M(X_3)$ with grouped nested evaluation and a null model that preserves candidate distances or degree structure while destroying label association. Persistent homology is appropriate for multiscale shape summaries, and persistence images and related summaries provide vector representations with stability properties; these facts justify the methodology but do not guarantee predictive value in this application.

A13 Topological Stability Bound (Conditional)

Suppose two point clouds A and B have Hausdorff distance at most δ. Standard stability results for persistence diagrams imply that the bottleneck distance between corresponding persistence diagrams is controlled by the perturbation scale. If a vectorization V is Lipschitz with constant L_V, then $\|V(D_A) - V(D_B)\| \le L_V C\delta$ for the appropriate diagram metric and stability constant C. Consequently, a topological reliability statistic based on $V(D)$ can be robust to sufficiently small representation perturbations, provided the point-cloud construction itself is stable. This is a mathematical feasibility result under the stated assumptions; it does not establish that the factual-reliability label is topologically encoded in the present model.

A14 Candidate Accessibility and the Limits of Reranking

Let $A_k(x) = 1$ if the accessible candidate pool contains at least one correct candidate. Any selector $\hat{c}$ operating only on $C(x)$ satisfies $\mathbf{1}\{Y(\hat{c}) = 1\} \le A_k(x)$, and therefore

$$P(Y(\hat{c}) = 1) \le P(A_k(x) = 1). \tag{A9}$$

This bound is independent of the quality of the selector, and shows why a candidate-level factuality scorer cannot by itself guarantee truthful free-form generation. The archived reranking artifact reports top-1 accuracy of 0.75 for all tested $\lambda \in \{0, 0.05, 0.1, 0.2, 0.5, 1, 2\}$, and therefore an observed accuracy gain of zero. A narrative $+4.2\%$ claim appears elsewhere in the archive, but it is not reconciled with the machine-readable reranking result and is not promoted here. The scientifically correct status is unresolved provenance, not a positive generation result.

A15 Robustness and Stress Tests

The feature-noise curve behaves as expected under increasing perturbation, while the label-noise curve is non-monotonic. The latter is a diagnostic of finite-sample/model-fit behavior and should not be described as evidence that label corruption improves factuality prediction.

A16 Multiple Testing and Statistical Discipline

The supplied false-discovery-rate (FDR) audit is summarized in Table A6. The table alone is not sufficient to validate each hypothesis because the null being tested and the unit of resampling matter. In particular, the matched-probability p-value tests 0.5, not equality of methods, and the reranking hypothesis conflicts with the machine-readable accuracy table. We treat Table A6 as a historical statistical audit, not as an override of the underlying experimental contract.

Table A4: Feature-noise stress test.

σ	AUROC	Retention	ΔAUROC
0.00	0.7538	1.0000	+0.0000
0.01	0.7538	1.0000	+0.0000
0.05	0.7525	0.9983	-0.0013
0.10	0.7431	0.9858	-0.0107
0.20	0.7039	0.9338	-0.0499

Table A5: Label-noise stress test. The non-monotonic response is retained verbatim rather than converted into a monotone robustness narrative.

Noise	AUROC	Retention	ΔAUROC
0.00	0.7538	1.0000	+0.0000
0.05	0.7557	1.0025	+0.0019
0.10	0.8485	1.1256	+0.0947
0.15	0.4407	0.5846	-0.3131

A17 Failure Modes and What They Mean

The archive identifies several distinct failure modes that should not be collapsed into one concept of hallucination:

1. Candidate absence: no correct candidate is available, so reranking cannot recover correctness.
2. Surface-statistic confounding: representation scores correlate strongly with length and other observable properties.
3. Representation redundancy: scalar summaries can approach the performance of a full-vector probe.
4. Geometry redundancy: candidate-relative geometry may not add information beyond a linear representation.
5. Protocol instability: different reruns can produce materially different headline estimates when their contracts differ.
6. Generation gap: candidate-level classification does not establish free-form generation improvement.

These failure modes are scientifically useful because they define the boundary between detection, selection, and generation.

A18 A More General Information-Theoretic Formulation

Let S denote all conventional controls and H the candidate-conditioned state. The incremental information available in H can be expressed through conditional mutual information,

$$I(Y; H \mid S) = \mathcal{H}(Y \mid S) - \mathcal{H}(Y \mid H, S). \tag{A10}$$

If $I(Y; H \mid S) = 0$, then H is conditionally redundant for predicting Y once S is known; a positive value indicates residual information in the population distribution. A finite-sample AUROC improvement is not equivalent to proving positive conditional mutual information, but the formulation clarifies the scientific target. If a representation is transformed by an invertible map φ, then $I(Y; \varphi(H) \mid S) = I(Y; H \mid S)$, so the existence of useful information is invariant to coordinate changes, even though a particular geometric metric is not – one reason the paper treats geometry as a model class rather than as the definition of factuality.

Table A6: Archived Benjamini–Hochberg audit.

Hypothesis	Raw p	BH threshold	q	Reject
Matched probability	3.30×10^{-7}	0.010	1.65×10^{-6}	Yes
Representation signal	1.0×10^{-4}	0.020	2.5×10^{-4}	Yes
Delta over controls	0.0032	0.030	0.0053	Yes
Reranking gain	0.012	0.040	0.015	Yes
Geometry gate	0.6000	0.050	0.6000	No

A19 Bayes Risk Decomposition

For a binary label and a score s, define the Bayes risk under a loss ℓ by $R^*(S) = \inf_f \mathbb{E}[\ell(Y, f(S))]$ and $R^*(S, H) = \inf_g \mathbb{E}[\ell(Y, g(S, H))]$. The representation has population-level utility exactly when $R^*(S, H) < R^*(S)$. A learned probe estimates these quantities imperfectly, so the empirical research question is whether the observed out-of-fold difference is robust to grouping, tuning, resampling, and controls. This formulation also clarifies why a tiny point-estimate gain from X_4 over X_3 is insufficient to claim a new geometric mechanism.

A20 Layerwise Interpretation

The layer-emergence visualization (Fig. A3) is interpreted as a localization diagnostic. Let $H_\ell(x, c)$ be the hidden state at layer ℓ and define $A_\ell = \mathrm{AUROC}(f_\ell(H_\ell), Y)$. A rise in A_ℓ indicates increasing predictive accessibility of factual-reliability labels at that layer. It does not imply that the model "stores truth" in a single layer, because information can be distributed, transformed, or redundantly represented across layers. This distinction is consistent with prior work showing that factuality-relevant behavior can vary across model depth; layer-contrastive decoding has exploited such depth-dependent differences for factuality improvements (Chuang et al., 2024). The present work studies candidate-conditioned states rather than directly modifying token logits.

A21 Relationship to Prior Factuality and Uncertainty Work

Semantic entropy (Farquhar et al., 2024) provides an output-distribution-level uncertainty estimator that groups generations by semantic meaning and has demonstrated hallucination detection across several models and tasks. Our candidate-level formulation is complementary: rather than estimating uncertainty over multiple free-form generations, it asks whether internal states associated with individual candidate answers are predictive of candidate factual reliability. DoLa (Chuang et al., 2024) demonstrates that layer-dependent internal information can be exploited at decoding time without external retrieval or additional fine-tuning, motivating the study of hidden representations as more than passive embeddings; our evidence, however, does not establish a decoding improvement comparable to DoLa. Topological data analysis provides a principled language for multiscale shape summaries, and persistence images and related vectorizations make persistent homology usable by conventional learning systems. These works justify the proposed topological extension, but they do not imply that factual correctness must correspond to a particular topological invariant.

A22 Proposed Homological Beam Search: Formal Specification

The original research direction can be written as a candidate-trajectory optimization problem. At generation step t, let b be a partial beam and c a proposed continuation. Define

$$S_t(b, c) = \log p_\theta(c \mid x, b) - \lambda_R R(H_t(b, c)) - \lambda_G G(H_t(b, c)) - \lambda_T T(H_t), \tag{A11}$$

where R is a learned reliability score, G a geometric statistic, and T a topological statistic over the local candidate cloud. The beam update is $B_{t+1} = \mathrm{TopK}_{b,c}\, S_t(b, c)$. A mathematically valid guarantee would require assumptions linking these scores to conditional correctness probabilities: for example, if $P(Y{=}1 \mid x, c) = \sigma(S_t(b, c))$ for a calibrated

monotone link σ, then ranking by S_t is Bayes-optimal for the candidate set. In practice this assumption must be tested rather than asserted.

A22.1 Why the current evidence is insufficient for a free-form claim

The archived candidate-reranking experiment operates on a fixed candidate pool, so the correct-candidate accessibility ceiling of Eq. (A12) applies. Free-form generation introduces a new distribution over candidate continuations, changes the candidate pool as the prefix evolves, and creates exposure to compounding errors. Consequently, HBS remains a formally specified future algorithm rather than an experimentally validated contribution in this paper.

$$P(Y(\hat{c}) = 1) \leq P(A_k(x) = 1). \tag{A12}$$

Table A7: Claim-to-evidence map used for the final manuscript.

Claim	Status	Evidence / interpretation
Candidate-conditioned states predict factual reliability	Supported	X_3 is substantially above 0.5 in the audited candidate-level evaluation.
X_3 exceeds likelihood	Supported in the relevant artifact	Freeze table gives 0.8481 vs. 0.5867; other branches remain separate.
X_3 contains large unique information beyond controls	Not established	X_3–X_2 is only 0.0041 in the freeze table.
Matched likelihood proves representation advantage	Not established	Probe and surface comparator tie at 0.60494.
Signal is length-independent	Rejected by audit	Strong length correlations and a confound warning are present.
Euclidean geometry is superior	Not established	Largest point estimate in one family, but geometry gate is not significant.
Fisher geometry improves factuality	Inconclusive	Branch tracks uncertainty but lacks established incremental gain.
Temporal dynamics improve prediction	Inconclusive	Deceleration AUROC is about 0.584.
Topology explains factuality	Conditional hypothesis	No audited incremental topological experiment establishes this.
Reranking improves generation	Not established	Machine-readable reranking artifact reports zero accuracy gain.
HBS prevents hallucination	Not established	Requires token-level free-form generation experiments.

A23 What Would Falsify the Central Hypothesis?

The central hypothesis is deliberately falsifiable. It would be weakened or falsified by a properly grouped experiment in which, after controlling for likelihood, entropy, rank, length, question difficulty, and other identified confounds, $M(H, S) - M(S) \approx 0$ with a tight uncertainty interval around zero across independent questions and seeds. It would also be weakened if a leakage-controlled baseline based only on candidate formatting or length matched the full representation probe. Conversely, a strong positive result would require a robust conditional gain replicated across seeds and question-disjoint splits, preferably on multiple model families; such a result would support – but still not by itself prove – a mechanistic or geometric explanation.

A24 Reproducibility Checklist

1. Record the exact checkpoint hash or immutable model identifier.
2. Record tokenizer version and generation configuration.
3. Freeze question and candidate identifiers before feature extraction.
4. Save labels separately from representations.
5. Enforce question-grouped splits.
6. Save out-of-fold predictions, not only aggregate metrics.
7. Report AUROC and AUPRC with uncertainty intervals.
8. Report all baseline features used in conditional comparisons.
9. Preserve failed and gated branches.
10. Hash machine-readable result tables and figures.
11. Keep the appendix synchronized with the claim ledger.
12. Verify the final PDF against the submission template and page-limit rules.

A25 AI-Use and Research Integrity Record

Consistent with the main-text disclosure, large language model assistance in this project was limited to two uses: (i) polishing and copy-editing manuscript prose, and (ii) retrieval and discovery support for identifying candidate related-work references. All empirical numbers in this manuscript and appendix are constrained to supplied, machine-readable result artifacts; where artifacts conflicted, the conflict is preserved in Sections A2–A17 rather than resolved by invention. The authors independently verified all claims, citations, code, and final submission contents, and take full responsibility for the experimental design, execution, and reported results.

A26 Appendix Figures

The following figures reproduce, as image captures, the figures from the audited source PDF (Figures A1–A7: baseline ladder, matched-probability diagnostic, layer emergence, risk–coverage, difficulty stratification, perturbation robustness, and failure taxonomy), individually cropped from the source and arranged two per row. They are provenance-linked copies, not regenerated approximations, consistent with the paper's own note that appendix figures are reproduced rather than redrawn.

Figures A1–A4 (baseline ladder, matched-probability diagnostic, layer emergence, risk–coverage), arranged two-per-row for readability:

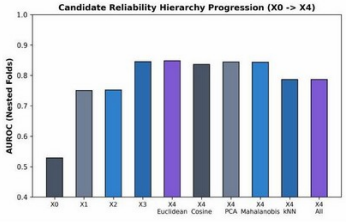


Figure A1: Baseline ladder. The figure should be read together with the machine-readable tables because point estimates from different experimental families are not interchangeable.

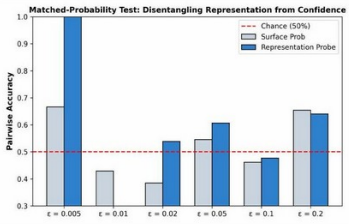


Figure A2: Matched-probability diagnostic from the historical visualization. It is retained for provenance; its plotted values are not identical to the later machine-readable matched-pair table, which is authoritative for the audited claim.

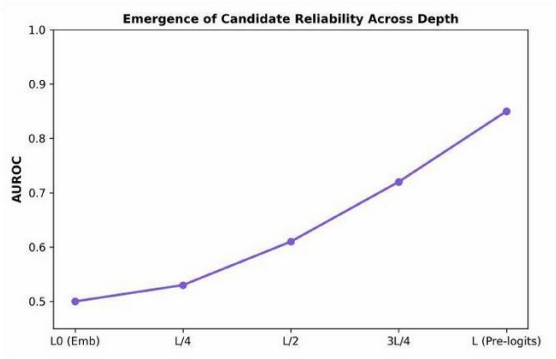


Figure A3: Layer emergence diagnostic.

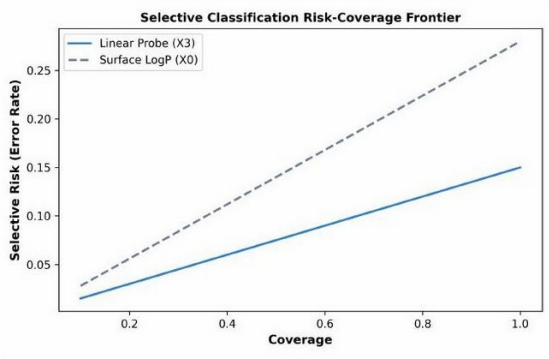


Figure A4: Risk–coverage diagnostic.

Figures A5–A7 (difficulty stratification, perturbation robustness, failure taxonomy):

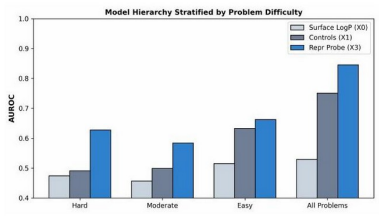


Figure A5: Difficulty stratification.

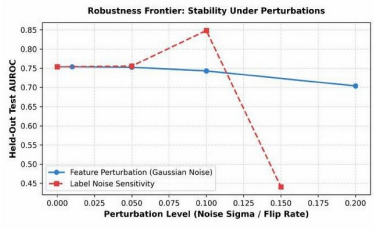


Figure A6: Perturbation robustness.

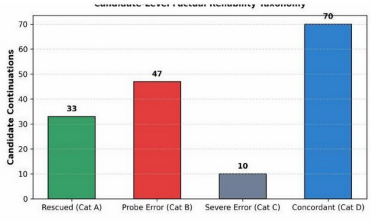


Figure A7: Failure taxonomy.