# The Geometry of Uncertainty in Language Model Representations

**Zy Li**
Department of Computer Science
Seattle University
zli10@seattleu.edu

## Abstract

Hidden-state probes have been used to predict factual errors in language-model outputs. We study whether the same representations also distinguish four types of prompt–model relationship: known, unknown, ambiguous, and misleading (built on a false premise). On a controlled 200-prompt dataset constructed to balance six surface features, all six pairwise type contrasts are linearly recoverable from hidden states at the position where generation begins, at AUROC $\geq$ 0.986 on three open-weight instruction-tuned models spanning two families and a fourfold parameter range. The probe exceeds surface-feature, TF-IDF, and frozen sentence-embedding baselines on every pair. Under group-aware evaluation on paraphrased prompts with group-level shuffled controls, probe AUROC remains at least 0.949 on every pair while lexical baselines degrade. These findings frame hallucination as one downstream consequence of a four-type representational structure that also captures ambiguity and false premises.

## 1 Introduction

Language models handle uncertainty unevenly. They answer factual prompts confidently, fail on niche or post-cutoff facts, hedge on linguistically ambiguous queries, and sometimes accept the premise of a question that contains a falsehood. These four cases reflect qualitatively different prompt–model relationships. Existing representation-space probes (Alnuhait et al., 2025; Bhatnagar et al., 2026; Varshney et al., 2026) mostly target the factuality or known/unknown boundary as hallucination detection, leaving ambiguity and embedded falsehood less directly modeled.

We propose four types (*known*, *unknown*, *ambiguous*, *misleading*) covering these prompt–model relationships and build a 200-prompt dataset (50 per type) in which each prompt is iteratively rewritten to balance six surface features across types, leaving surface-only AUROC spanning 0.718 to 0.867 across the six pairwise contrasts (Table 1). We extract prompt-state hidden representations at the position generation would begin from three open-weight instruction-tuned LLMs (Qwen2.5-7B and Qwen2.5-32B (Qwen Team, 2024), Llama-3.1-8B (Llama Team, AI @ Meta, 2024)), train linear probes per type pair per layer with 5-fold cross-validation across 5 seeds, and calibrate the empirical null with group-level shuffled controls. To test sensitivity to wording, we add two paraphrases per base prompt under explicit type-preservation rules (558 prompts in 186 paraphrase groups) and evaluate under group-aware cross-validation that keeps all paraphrases of a base in the same fold.

The dataset is a controlled diagnostic set rather than a benchmark-scale sample of natural prompts, with explicit type definitions, reference answers, enumerated ambiguity readings, false-premise annotations, and surface-feature controls. We evaluate type recoverability against stronger non-hidden baselines, under paraphrase-group splits, and against shuffled-label nulls.

The empirical claim has three properties that distinguish it from prior hallucination-probing work. First, the structure is geometric and uniform: all six pairwise type contrasts separate at AUROC $\geq$ 0.986. Second, it sits upstream of behavior: the probe sees only the prompt state at the moment generation would begin, before any token is produced, and the structure is present even on prompts where the eventual generation is type-misaligned (§7). Third, it replicates across the two model families and the fourfold parameter range we test, with no per-model tuning.



*Proceedings of the 15th International Joint Conference on Natural Language Processing and the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 2683–2702
November 6-10, 2026 ©2026 Association for Computational Linguistics

## 2 Four Types of Uncertainty

We use four uncertainty types, distinguished by the model's epistemic relationship to the prompt. Two of them (Known, Unknown) are properties of the prompt–model pair by construction: the same prompt can be Known to one model and Unknown to another, and this model-relativity is part of the type definition (§3). The other two (Ambiguous, Misleading) are properties of the prompt alone.

**Known (K).** A prompt that asks for a fact the model recalls reliably, producing the gold answer under greedy or repeated sampled decoding (operational screening criterion in Appendix A). Example: *"Which city is the capital of Kazakhstan?"*

**Unknown (U).** A prompt that asks for a fact the model does not know, drawn from four sub-pools (post-cutoff, long-tail, niche-technical, split-recall; defined in Appendix B). The model may abstain, hedge, or guess inconsistently, but the gold is not produced, or for the split-recall sub-pool only occasionally (screening criterion in Appendix A). Example: *"What is the natural habitat of the dik-dik?"*

**Ambiguous (A).** A prompt that admits multiple valid readings: referential (*"Which Cambridge produced famous output?"*, where Cambridge could be the English university town, the Massachusetts city, or the dictionary), scope (*"How safe is the safest investment?"*, where the superlative's criterion, population, and horizon are unspecified), or temporal (*"Who is the current US president?"*, whose answer depends on when it is asked). The type-aligned response is clarification, not commitment. Referential items pass a three-part test (Appendix B): the competing referents are salient from common knowledge rather than discoverable only by search, the question is coherent with a different answer under each referent, and a competent assistant would reasonably ask which is meant.

**Misleading (M).** A prompt that embeds a false premise declaratively. Example: *"Which Roman emperor wrote Hamlet?"* (not *"Did a Roman emperor write Hamlet?"*). The type-aligned response is premise-correction. The category is the classical notion of *presupposition failure* (Strawson, 1950): the question carries a presupposition that is false, so no direct answer is felicitous. Its operational form for question answering is the false-presupposition question studied in presupposition verification (Kim et al., 2021), CREPE (Yu et al., 2023), and $(QA)^2$ (Kim et al., 2023); those works evaluate whether a system can answer or correct such questions, whereas we study where such prompts sit in representation space before any answer is produced. We adopt three operational rules to keep the category sharp. First, the false statement must be *presupposed*, not asserted as the object of inquiry: *"Did a Roman emperor write Hamlet?"* is a Known-or-Unknown factual query, not Misleading. Second, the answer space must remain coherent only if the false premise is accepted: *"Which Roman emperor wrote Hamlet?"* has no defensible answer except premise-correction. Third, the falsehood must concern verifiable world fact, not contested interpretation or hypothetical reasoning. Counterfactual prompts (*"If a Roman emperor had written Hamlet. . . "*) are excluded by construction; they invite engaged reasoning rather than premise-challenge.

We refer to the six pairwise contrasts as K_vs_U, K_vs_A, K_vs_M, U_vs_A, U_vs_M, and A_vs_M.

The type is a property of the prompt–model pair, fixed before generation, not of the sampled response. The same Known prompt may elicit a correct answer under one sampling seed and an abstention under another; an Unknown prompt may receive a confidently wrong guess on one run and a hedged "I don't know" on the next. Our claim is about the prompt-side geometry, evaluated before generation begins. We label response behavior independently of prompt type, allowing combinations that differ from the type-aligned behavior.

## 3 Dataset

Our base dataset contains 200 prompts, 50 per type. Each prompt is constructed and filtered to suppress six potentially confounding surface features: word count, named-entity count (heuristic, the number of non-leading capitalized tokens), question form (*what*, *who*, *when*, *where*, *why*, *how*, *which*, *other*), per-model prompt token count after chat-template application, per-model next-token entropy at the generation-start position, and the sentence-embedding cosine to the centroid of Known prompts (MiniLM-L6-v2 embeddings, with the centroid recomputed per cross-validation fold to avoid test-set leakage).

Construction targeted U_vs_A, the pair with the highest surface AUROC in the initial construction; coordinated rewriting drove the surface AUROC

on that contrast from $\sim 0.96$ on an early eighty-prompt iteration down to $0.753$ on the final 200-prompt dataset. Surface AUROC across the six pairs spans $0.718$–$0.867$ (Table 1). Including next-token entropy in the surface baseline captures information from the model's output distribution: Known prompts produce low-entropy continuations because the model is confident on factual recall, and that confidence is recoverable before any token is decoded. The comparison therefore tests whether the geometric probe exceeds this confidence signature; the surface features themselves carry information.

Unknown items are selected such that under our screening protocol (one greedy plus three sampled generations from each of the three target models), no model produces the gold answer, with the split-recall sub-type admitting at most occasional success. The operationalization is contingent on the screening panel; an item Unknown for these three open-weight models might be Known for a larger or differently trained model. The geometric claim is therefore about Unknown relative to the screening models, not Unknown in an absolute sense. The label may also change across checkpoints: post-cutoff items drift toward Known as newer models are trained, so the released screening records and pipeline let the Known/Unknown labels be re-derived for any new panel of models.

**Paraphrase extension.** We extend the dataset with two Claude Opus 4.7 paraphrases per base prompt under type-preservation rules: same gold for Known, continued cross-model failure for Unknown (verified after generation), same valid readings for Ambiguous, same false premise for Misleading. The same screening panel filters Unknown paraphrases: a paraphrase passes only if at least two of three target models still produce zero correct answers across one greedy plus three sampled generations. Fourteen of fifty Unknown base groups fail this filter and are excluded from the extension; these are the wording-sensitive items whose paraphrases no longer meet the Unknown screening criterion. The retained set is 186 paraphrase groups (50 Known, 36 Unknown, 50 Ambiguous, 50 Misleading) $\times$ 3 prompts = 558 total. The per-item exclusion list is included in the released data; the filter rationale is in Appendix A.

## 4 Methods

We extract prompt-only hidden states at the position generation would begin (the chat template is applied with the assistant-turn header appended) under BF16 precision and scaled dot-product attention. At each layer we save the last-token state and the mean state over non-padding prompt tokens; the last-token state is the primary representation throughout this paper.

For each pair of types $(t_1, t_2)$ and each layer, we train an L2-regularized logistic regression with $C = 1.0$ using the L-BFGS solver and a 2000-iteration cap, on raw fp32 features and with no preprocessing or dimensionality reduction. We use 5-fold cross-validation with five seeds (1234, 4567, 7890, 2345, 5678) and report seed-mean AUROC with bootstrap 95% confidence intervals (1000 resamples). For the paraphrase extension, cross-validation is group-aware: we use stratified group-aware $k$-fold splitting, ensuring all paraphrases of a base prompt are held out together.

The shuffled-label control runs the same pipeline five times under permuted type labels (seeds 9001–9005). For the paraphrase extension, the permutation is performed at the group level: we permute the assignment of type to paraphrase groups and propagate to constituent rows. Row-level shuffling on grouped data would partially preserve type structure within a group, undermining the empirical-null calibration.

The surface-baseline AUROC used in the criterion below is, for each pair, the maximum over the three models' surface-feature classifiers (each model contributes its own entropy and token-count features); it is fixed per pair and shared across models (Tables 1, 2, and 3). On the paraphrase set the baseline is the larger of the base-set and paraphrase-set maxima. Table 6 instead reports Qwen-7B's own paraphrase-set surface classifier, which is lower on K_vs_U ($0.827$ vs. the cross-model $0.859$). Peak AUROCs (Tables 1, 2, and 3) are taken over all layers of a model; the layer-window criterion below is evaluated separately, and Appendix D reports window-restricted values. The baseline competition is the classification analogue of covariate adjustment: we require the hidden-state probe to outperform a classifier trained on the six surface features alone; Appendix F reports complementary covariate-inclusion and residualization analyses.

A pair passes the evaluation criterion only when,

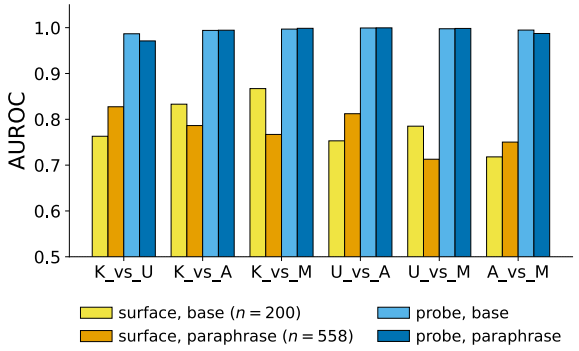



Figure 1: Per-pair AUROC on Qwen2.5-7B-Instruct: surface baseline vs. linear probe, base set and paraphrase extension.

at the same layer, three conditions hold: (i) *surface margin*, the real-label probe exceeds the surface-baseline AUROC by $\geq 0.05$ (we call the baseline AUROC plus 0.05 the *required threshold*); (ii) *empirical-null gap*, the real-vs-mean-shuffled gap is $\geq 0.10$ AUROC; (iii) *per-seed safety*, every individual shuffled seed remains below the required threshold. A pair passes overall if at least three consecutive or four non-consecutive such layers exist in a model-specific window: $L_{14}$–$L_{28}$ for Qwen-7B, $L_{16}$–$L_{32}$ for Llama-8B, and $L_{32}$–$L_{64}$ for Qwen-32B.

## 5 Results

**All six contrasts separate.** On the base 200-prompt dataset, linear probes separate every pairwise type contrast at AUROC $\geq 0.986$ across all three models, with per-model mean $\geq 0.994$ (Table 1; Figure 1 shows per-pair AUROCs for both base and paraphrase sets). The gap over the surface baseline is at least $+0.125$ on the highest-surface-baseline pair and at least $+0.273$ on the lowest-surface-baseline pair. U_vs_A reaches AUROC 0.999 against a surface baseline of 0.753, a margin of $+0.246$.

**Stronger non-LLM baselines.** We compare the hidden-state probe with classifiers trained on frozen sentence embeddings and on TF-IDF features, to measure how much of the separation a generic semantic encoder or lexical n-gram statistics already capture. For the encoders we train the same logistic regression on full sentence-embedding vectors from three frozen models of growing capacity: all-MiniLM-L6-v2 (Wang et al., 2020) (22M parameters, 384d), all-mpnet-base-v2 (Song et al., 2020) (109M, 768d), and BGE-large-en-v1.5 (Xiao et al., 2023) (335M, 1024d). For the lexical baselines we train TF-IDF logistic regressions in three variants (word 1–2 grams, character 3–5 grams, and the concatenation of both), with the TF-IDF vocabulary fit on the training fold only inside an sklearn `Pipeline`.

| pair | surface | probe [95% CI] | margin |
|---|---|---|---|
| K_vs_U | 0.763 | 0.987 [0.97, 1.00] | +0.224 |
| K_vs_A | 0.833 | 0.994 [0.98, 1.00] | +0.161 |
| K_vs_M | 0.867 | 0.997 [0.99, 1.00] | +0.130 |
| U_vs_A | 0.753 | 0.999 [0.99, 1.00] | +0.246 |
| U_vs_M | 0.785 | 0.998 [0.99, 1.00] | +0.213 |
| A_vs_M | **0.718** | **0.995 [0.99, 1.00]** | **+0.277** |

Table 1: Per-pair surface baseline (maximum over the three models' surface classifiers, §4) and Qwen2.5-7B-Instruct linear-probe peak AUROC (5-seed mean and 95% bootstrap CI over 1000 resamples), base set, with geometric-over-surface margin. Widest-margin contrast in bold.

Per-pair AUROCs are in Table 2. The Qwen-7B probe exceeds every non-hidden baseline on every pair, but the margin varies. The lexical baselines are strong on the pair where Unknown long-tail entities and Ambiguous referents have characteristic vocabularies: TF-IDF reaches 0.991 on U_vs_A, only 0.008 below the probe. The encoder and TF-IDF baselines also nearly tie the probe on K_vs_A. On K_vs_M the picture inverts: every non-hidden baseline, including BGE-large and TF-IDF, sits at or below the surface baseline (0.867) on this pair, while the probe reaches 0.997. False-premise detection often turns on factual knowledge about the world, which n-gram statistics and retrieval-trained encoders capture less of than an instruction-tuned LLM's mid-layer states. The strongest non-hidden baseline at every pair is one of {surface, BGE-large, TF-IDF combined}; the mean probe-over-strongest gap on the base set is $+0.060$ and the minimum is $+0.008$.

| pair | surf. | BGE | TF-IDF | probe | gap |
|---|---|---|---|---|---|
| K_vs_U | 0.763 | 0.899 | 0.890 | 0.987 | +0.088 |
| K_vs_A | 0.833 | 0.947 | 0.959 | 0.994 | +0.035 |
| K_vs_M | **0.867** | **0.735** | **0.706** | **0.997** | **+0.130** |
| U_vs_A | 0.753 | 0.986 | 0.991 | 0.999 | +0.008 |
| U_vs_M | 0.785 | 0.955 | 0.923 | 0.998 | +0.043 |
| A_vs_M | 0.718 | 0.937 | 0.927 | 0.995 | +0.058 |

Table 2: Probe vs. strongest non-hidden baselines on the base set. *surf.* = six-feature surface; *BGE* = BGE-large-en-v1.5; *TF-IDF* = word 1–2 + char 3–5 grams; *probe* = Qwen-7B; *gap* = probe − max(surf., BGE, TF-IDF). Widest-gap row in bold.

**Four-way classification.** We report the pairwise contrasts as the primary result because the surface-confound margin is defined per contrast. We also evaluate joint prediction of the four types with a multinomial logistic-regression probe under the same protocol (L2 penalty, $C = 1.0$, 5-fold cross-validation $\times$ 5 seeds). Peak in-window accuracy is 0.910 on Qwen-7B (95% CI $[0.867, 0.947]$), 0.896 on Llama-8B ($[0.852, 0.937]$), and 0.938 on Qwen-32B ($[0.902, 0.968]$), against chance 0.250, a shuffled-label peak of at most 0.329, and 0.692 for the strongest non-hidden baseline (TF-IDF, word plus character n-grams); macro one-vs-rest AUROC is 0.985, 0.986, and 0.995. Under group-aware paraphrase cross-validation the four-way probe holds at 0.903, 0.894, and 0.922. Confusion concentrates on Known↔Unknown, the pair with the lowest pairwise AUROC; Appendix E gives the confusion matrices and every baseline's four-way accuracy.

**Layer-window plateau.** For every pair on every model, *every* layer in the model-specific window qualifies: 15 of 15 layers in Qwen-7B's $L_{14}$–$L_{28}$, 17 of 17 in Llama-8B's $L_{16}$–$L_{32}$, and 33 of 33 in Qwen-32B's $L_{32}$–$L_{64}$. The structure appears in the middle layers and persists through the late-middle layers; it is not localized to any single layer. Figure 2 shows the layer profile for Qwen-7B.

**Robustness to layer choice.** To assess sensitivity to layer selection, we report the mean per-pair AUROC over the middle third of each model-specific evaluation window. For Qwen-7B that is $L_{19}$–$L_{24}$ (6 layers), for Llama-8B $L_{21}$–$L_{27}$ (7 layers), for Qwen-32B $L_{43}$–$L_{54}$ (12 layers). The middle-third per-pair mean is 0.988 on Qwen-7B, 0.986 on Llama-8B, and 0.991 on Qwen-32B. The worst single middle-third layer across the three models is K_vs_U at 0.946 on Llama-8B layer 25. The structure is present across the middle of every model, independent of peak-layer selection. Per-pair detail is in Appendix D.

## 6 Analysis

**Type regions in 2D.** Figure 3 projects Qwen-7B's last-token hidden states at the peak layer into 2D via LDA and t-SNE; the four types occupy distinct regions in both projections, including in the unsupervised t-SNE map.

**The confidence-signature surface signal.** The surface baseline's strength on K_vs_A (0.833) and

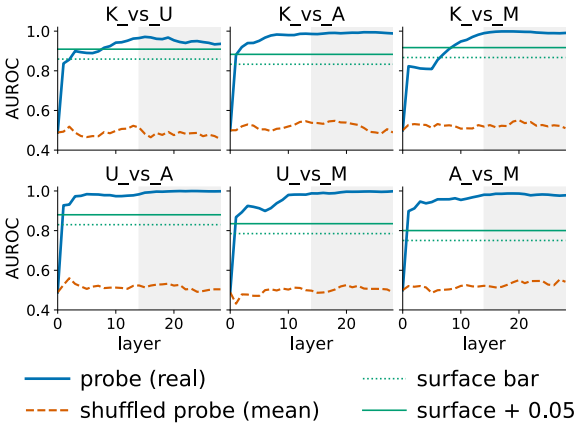



Figure 2: Per-pair layer-wise probe AUROC on Qwen2.5-7B, paraphrase set, with shuffled-label controls and the per-pair required threshold; shaded band marks the model-specific layer window ($L_{14}$–$L_{28}$).

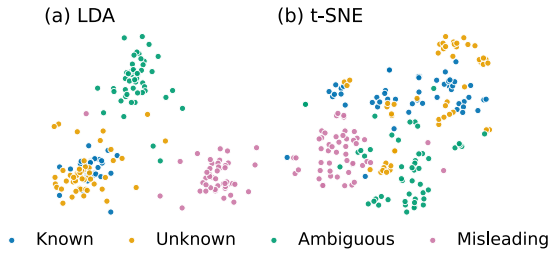



Figure 3: 2D projections of Qwen2.5-7B's last-token hidden states at layer 23. Panel (a): LDA (first two components $\sim$ 77% of between-class variation). Panel (b): unsupervised t-SNE. Illustrative; quantitative claims use cross-validated probes.

K_vs_M (0.867) is driven by next-token entropy (Table 1): Known prompts produce low-entropy continuations because the model is confident on factual recall, and that confidence is recoverable from the output distribution before any token is decoded. But entropy is a noisy proxy for type: models can be confidently wrong on Unknown and Misleading prompts (Simhi et al., 2025), and entropy does not separate pairs where both types hedge. The probe's largest margin (A_vs_M, +0.277) is on the lowest-surface-baseline pair. Type separation also remains after linear adjustment for all six measured surface features: all 18 (pair, model) combinations exceed their residualized permutation nulls (Appendix F).

**Cross-scale uniformity.** Qwen-7B (28 blocks, 3584d), Llama-8B (32, 4096), and Qwen-32B (64, 5120) give per-pair margins agreeing within 0.008 across the three models (Table 3). The structure replicates across the two model families and the fourfold parameter range tested.

| pair | Qwen-7B | Llama-8B | Qwen-32B | range |
|---|---|---|---|---|
| K_vs_U | +0.224 | +0.223 | +0.231 | 0.008 |
| K_vs_A | +0.161 | +0.163 | +0.166 | 0.005 |
| K_vs_M | +0.130 | +0.125 | +0.133 | 0.008 |
| U_vs_A | +0.246 | +0.247 | +0.247 | 0.001 |
| U_vs_M | +0.213 | +0.214 | +0.215 | 0.002 |
| A_vs_M | +0.277 | +0.273 | +0.277 | 0.004 |

Table 3: Geometric-over-surface margin per pair, by model, on the base set, using each model's peak AUROC over all layers. *range* = max − min across the three models per row. Restricting peaks to each model's layer window changes one cell by more than 0.001: Llama-8B K_vs_U peaks at 0.981 in-window ($L_{16}$) versus 0.986 overall ($L_{14}$), giving a margin of +0.218 and a row range of 0.013.

**Dimensionality and the permutation null.** At Qwen-7B's K_vs_U peak layer, the probe fitted in-sample reaches AUROC 0.999–1.000 under each of the five shuffled-label permutations: with 200 prompts in a 3584- to 5120-dimensional feature space, high in-sample AUROC does not establish generalization to held-out prompts, the situation described for neuroimaging as circular analysis (Kriegeskorte et al., 2009; Vul et al., 2009). Our performance estimates therefore use out-of-fold predictions under 5-fold cross-validation and are calibrated against the same pipeline run on permuted labels (Nichols and Holmes, 2002): on the base set the shuffled-label probe peaks at mean AUROC 0.553 in-window (single-permutation maximum 0.716) against real-label values $\geq$ 0.939 at every in-window layer, and on the paraphrase extension at most 0.743 against per-pair in-window peaks $\geq$ 0.949 (Appendix C). Bootstrap intervals are descriptive; evaluation uses the criteria in §4, including a minimum AUROC margin of 0.10 over the permutation null.

**Confounds as covariates.** Appendix F reports two further forms of confound control on all three models. Concatenating the six z-scored surface features onto the hidden state changes seed-mean peak AUROC by at most 0.009 on any of the 18 (pair, model) cells and never raises it by more than 0.004. Within each training fold, residualization removes the component of each hidden dimension linearly predicted by the covariates, which assigns all shared variance to the covariates. Margins narrow; 16 of 18 cells remain above their surface-baseline AUROC and every cell stays at least +0.22 above its residualized permutation null (+0.14 against the single worst permutation). The two cells that fall below their baselines (Llama-8B K_vs_M, Qwen-32B K_vs_U) trace to the two model-derived features (next-token entropy and embedding cosine) rather than to a text confound: residualizing against the construction-controlled text features alone keeps all 18 cells above their baselines.

**Label-noise tolerance.** Because the Ambiguous and Misleading labels involve judgment (Appendix H), we flipped 5% and 10% of the Ambiguous/Misleading labels at random (ten draws per rate; Known/Unknown untouched) and retrained. Average peak-layer AUROC moves by at most 0.050 and 0.134, concentrated in the directly corrupted A_vs_M pair; scored against the original labels the probe still averages 0.987 and 0.958 on that pair, and every pair stays at least 0.081 above its surface-baseline AUROC at the worst draw (Appendix G).

**Sub-type structure.** The main claim is about the four top-level types; in an analysis designed after the four-type comparison, we also examined whether the geometry distinguishes within-type sub-structure. All three types with defined sub-types decompose: mean pairwise AUROC is 0.968–0.992 for the five Misleading sub-types (misconception, wrong attribution, nonexistent entity, wrong date, impossible premise), $\geq$ 0.994 for the three Ambiguous sub-types (referential, scope, temporal), and 0.958–0.967 for the four Unknown sub-types (post-cutoff, long-tail, niche-technical, split-recall). The within-type structure is partly attributable to lexical sub-pool signatures from dataset construction (post-cutoff prompts contain date markers; niche-technical contain identifier strings; ambiguity-kind is largely determined by question surface structure). The one within-type pair where the geometry blurs is long-tail vs split-recall on Unknown (0.751–0.800), both rare-entity questions sharing lexical structure. Full per-pair AUROCs are in Appendix B.

## 7 Geometry vs. Behavior

The probes show that the four types are recoverable from prompt-state hidden representations. We compare type recoverability with response behavior on Qwen2.5-7B and Llama-3.1-8B by labeling the behavior each model exhibits when asked to answer each base prompt.

**Behavior labeling.** For each of the 200 prompts we generate one greedy short answer, five temperature-0.7 sampled short answers (seeds 1–5), and one premise-check completion. From these seven generations we derive a single behavior label per prompt under a type-blind rule set: *correct* (greedy and $\geq 4/5$ samples match the gold; 2–3/5 is *correct_inconsistent*, $< 2/5$ *correct_fluky*), *wrong_confident* (consistent non-gold modal answer), *wrong_inconsistent*, *abstain*, *asks_clarification*, or *challenges_premise*, using gold-match heuristics and regex patterns for the latter four; Table 4 groups the three *correct* labels for display. Off-pattern combinations (e.g., Known eliciting premise-challenge) are automatically flagged for manual review; we apply a focused re-labeling pass to correct under-counted premise challenges (the original regex pattern is conservative and misses some explicit "there is no…" challenges on Misleading prompts).

**The behavior crosstab.** Table 4 shows the type $\times$ behavior crosstabs for both models. The ideal behavior per type is *correct* for Known, *abstain* for Unknown, *asks_clarification* for Ambiguous, and *challenges_premise* for Misleading. The dominant cell (bolded) matches the ideal for Known (Qwen 80%, Llama 78%), Misleading (Qwen 72%, Llama 74%), and Ambiguous (Qwen 64%, Llama 42%, a lower clarification rate on Llama). For Unknown, the dominant cell is *wrong_confident* (Qwen 48%, Llama 32%), which is *not* the ideal. Abstention plus hedged or inconsistent answers together account for only 24% (Qwen) and 38% (Llama). The Unknown row is the clearest demonstration of the geometry-behavior gap: the geometry separates Unknown at AUROC $\geq 0.986$ on every model, yet the dominant generated behavior is a consistent incorrect answer.

**The geometry-behavior gap.** Geometry recovers the type at AUROC $\geq 0.986$ on every pairwise contrast on both models (mean $\geq 0.994$ per model; Table 1); the dominant behavior cell accounts for 32–80% of prompts and matches the type's ideal for Known, Ambiguous, and Misleading but not for Unknown (Table 4). The two models have similar pairwise probe AUROCs but different response distributions: their geometric margins differ by at most 0.005 AUROC, whereas their clarification rates on Ambiguous prompts differ by 22 percentage points (64% for Qwen-7B and 42% for Llama-8B). High type recoverability therefore coexists with substantial variation in response behavior.

**Qwen2.5-7B-Instruct**

| | ab. | cl. | ch. | co. | wc. | wi. | N |
|---|---|---|---|---|---|---|---|
| Known | 0 | 1 | 2 | **40** | 4 | 3 | 50 |
| Unknown | 7 | 1 | 10 | 3 | **24** | 5 | 50 |
| Ambiguous | 3 | **32** | 2 | 0 | 11 | 2 | 50 |
| Misleading | 0 | 5 | **36** | 0 | 5 | 4 | 50 |

**Llama-3.1-8B-Instruct**

| | ab. | cl. | ch. | co. | wc. | wi. | N |
|---|---|---|---|---|---|---|---|
| Known | 0 | 0 | 1 | **39** | 8 | 2 | 50 |
| Unknown | 11 | 1 | 0 | 14 | **16** | 8 | 50 |
| Ambiguous | 4 | **21** | 1 | 0 | 8 | 16 | 50 |
| Misleading | 2 | 3 | **37** | 0 | 4 | 4 | 50 |

Table 4: Type $\times$ behavior crosstabs for the two models we behavior-labeled. Columns: *ab.* = abstain; *cl.* = asks_clarification; *ch.* = challenges_premise; *co.* = *correct*, shown together with *correct_inconsistent* (Qwen Unknown: 1; Llama Unknown: 3) and *correct_fluky* (Llama Known: 1), which remain separate labels in the released files; *wc.* = wrong_confident; *wi.* = *wrong_inconsistent* (with *hedges*, Llama only: Ambiguous 3, Misleading 1). Bold = dominant behavior cell per type.

These results motivate testing whether probe-guided decoding improves abstention, clarification, and premise correction.

## 8 Paraphrase Robustness

On the 558-prompt paraphrase extension, with each base prompt and its two paraphrases held in the same cross-validation fold and shuffled controls permuted at the group level, the probe peak in each model's window remains at AUROC $\geq 0.949$ on every pair (Table 5). The real-vs-shuffled gap at the peak layer ranges from $+0.446$ to $+0.528$ across (pair, model) cells (Table 5), and no individual shuffled seed at any in-window layer on any model comes within 0.115 AUROC of the required threshold.

The paraphrase extension also tightens the surface baseline on three pairs by raising the paraphrase-set surface AUROC above the base-set value, a side effect of paraphrases occasionally inflating surface features such as word count and entropy spread. The probe AUROC does not move with it: on K_vs_U, the cross-model surface-baseline AUROC rises from 0.763 on the base set (Table 1) to 0.859 on the paraphrase set (the 0.909 required threshold in Table 5 minus the 0.05 margin), but the probe peak stays at $\geq 0.949$ on every model.

| pair | required | Qwen-7B | Llama-8B | Qwen-32B | min gap | max gap | min null margin |
|---|---|---|---|---|---|---|---|
| K_vs_U | 0.909 | 0.971 | 0.949 | 0.979 | +0.460 | +0.489 | +0.304 |
| K_vs_A | 0.883 | 0.994 | 0.997 | 0.998 | +0.464 | +0.482 | +0.140 |
| K_vs_M | 0.917 | 0.999 | 0.992 | **1.000** | +0.446 | +0.477 | +0.231 |
| U_vs_A | 0.880 | **1.000** | 0.998 | 0.999 | +0.481 | +0.493 | +0.145 |
| U_vs_M | 0.835 | 0.998 | 0.999 | **1.000** | +0.487 | +0.528 | +0.169 |
| A_vs_M | 0.800 | 0.987 | 0.991 | 0.995 | +0.448 | +0.470 | +0.115 |

Table 5: Probe peak AUROC per model in its layer window, paraphrase extension. *Required* = max(base, paraphrase) surface + 0.05; *Min gap* / *Max gap* = smallest / largest real-vs-shuffled gap at in-window peak across the three models; *Min null margin* = smallest difference between the required threshold and any shuffled-label AUROC (across all three models, all in-window layers, all five shuffle seeds).

The non-LLM baselines from §5 replicate on the paraphrase extension under the same group-aware splitting (Table 6), and paraphrasing affects the lexical baselines more than the probe. Per-pair AUROC decreases by 0.022 on average for the TF-IDF n-gram classifier and is essentially unchanged for the Qwen-7B probe (per-pair drops average 0.000, with small drops and small gains canceling). On U_vs_A, paraphrasing widens the probe-over-TF-IDF gap from +0.008 (Table 2) to +0.058: TF-IDF AUROC falls from 0.991 to 0.942, while probe AUROC remains near ceiling (0.999 and 1.000). BGE-large-en-v1.5 is more robust to paraphrasing than TF-IDF on this pair; its base-set +0.013 gap stays essentially unchanged at +0.011 under paraphrase. The K_vs_M weakness persists: each of the three baselines in Table 6 sits between 0.713 and 0.767 AUROC on K_vs_M (the remaining encoder and TF-IDF variants span 0.692–0.783), while the probe holds at 0.999.

| pair | surf. | BGE | TF-IDF | probe | gap |
|---|---|---|---|---|---|
| K_vs_U | 0.827 | 0.859 | 0.865 | 0.971 | +0.106 |
| K_vs_A | 0.786 | 0.953 | 0.934 | 0.994 | +0.041 |
| K_vs_M | **0.767** | **0.740** | **0.713** | **0.999** | **+0.231** |
| U_vs_A | 0.812 | 0.989 | 0.942 | 1.000 | +0.011 |
| U_vs_M | 0.713 | 0.941 | 0.908 | 0.998 | +0.057 |
| A_vs_M | 0.750 | 0.939 | 0.900 | 0.987 | +0.049 |

Table 6: Probe vs. strongest non-hidden baselines on the paraphrase extension (Qwen-7B), under group-aware cross-validation on the paraphrase group. Columns and bolding convention follow Table 2, except that *surf.* is Qwen-7B's own paraphrase-set surface classifier rather than the cross-model maximum (§4).

The shuffled-label result closest to its threshold reaches 0.685 AUROC against a required threshold of 0.800 (Ambiguous vs. Misleading on Llama-8B at layer 24, seed 9001); the empirical null is cleared by every seed at every layer in the model-specific window on all three models. Per-seed details are in Appendix C.

## 9 Related Work

**Hallucination detection and probing.** Hallucination is well-studied (Ji et al., 2023), addressed by output-level consistency (Manakul et al., 2023), semantic-entropy estimators (Farquhar et al., 2024), and internal-state inspection (Azaria and Mitchell, 2023; Slobodkin et al., 2023). The closest precursors probe hidden states directly for hallucination: FactCheckMate (Alnuhait et al., 2025) predicts factuality from intermediate states; DRIFT (Bhatnagar et al., 2026) detects representational inconsistencies; LLM Router (Varshney et al., 2026) uses prefill activations for downstream routing. These works mostly target factuality or the Known-vs-Unknown boundary at generation time; our four-type structure includes this boundary at the prompt state and adds Ambiguous and Misleading.

**Presupposition failure in question answering.** Questions with false presuppositions have been studied as a QA problem: presupposition verification (Kim et al., 2021), CREPE (Yu et al., 2023), and $(QA)^2$ (Kim et al., 2023) build datasets of questions with questionable assumptions and evaluate whether systems detect and correct them, drawing on the linguistic notion of presupposition failure (Strawson, 1950). Our Misleading type is this category under the three construction rules of §2. We examine the representations of such questions before generation; K_vs_M yields the largest probe improvement over the strongest non-hidden baseline (+0.130, Table 2).

**Probing methodology.** Linear probing has a long history as a tool for studying language-model internals (Tenney et al., 2019; Clark et al., 2019; Burns et al., 2023). The methodological pitfalls of unconstrained probing are well documented (Belinkov, 2022; Hewitt and Liang, 2019; Voita and

Titov, 2020): a sufficiently expressive probe can recover almost any property of the input, including properties the model itself does not use. We adopt the simplest variant (L2 logistic regression, no preprocessing, fixed seeds) and report shuffled-label controls to bound this concern.

**Representation geometry and factual associations.** High-level concepts are linearly represented in LLM hidden states (Park et al., 2024), and the truth/falsity of declarative statements is linearly recoverable with emergent geometric structure (Marks and Tegmark, 2024); we extend this from binary truth to a four-way prompt-level typology. Representations across independently trained models converge to a shared latent geometry as scale and capability increase (Huh et al., 2024), which contextualizes our cross-model agreement across families and a fourfold parameter range. A separate strand localizes factual knowledge in mid-layer components (Geva et al., 2021; Meng et al., 2022; Hernandez et al., 2024), and representation engineering shows that high-level concepts can be elicited and steered with linear interventions (Zou et al., 2023). The four-type structure may be an instance of the convergent geometry these papers describe.

**Adjacent areas.** Calibration and abstention have been studied through generated text (Kadavath et al., 2022; Kuhn et al., 2023; Lin et al., 2022) and behavioral deferral to external sources (Mallen et al., 2023); our representation-level analysis sits upstream of these, since the type is encoded before any token is decoded. Question ambiguity is itself an open problem in NLP (Min et al., 2020; Liu et al., 2023), with models often committing to one reading; our finding that Ambiguous prompts occupy a distinct region suggests that ambiguity is represented in the hidden states even when the generation step commits to one reading. The four-type structure is not OOD detection (Hendrycks and Gimpel, 2017; Liu et al., 2020): Ambiguous and Misleading prompts are syntactically in-distribution.

## 10 Conclusion

The four types of uncertainty (known, unknown, ambiguous, misleading) are linearly recoverable from prompt-state hidden representations of three open-weight instruction-tuned LLMs spanning two model families and a fourfold parameter range. All six pairwise contrasts separate at AUROC $\geq 0.986$ on the surface-controlled base set and $\geq 0.949$ under paraphrase rewriting, with the largest gap over the strongest non-hidden baseline on the false-premise contrast where world knowledge matters most. On this evidence, the factuality boundary that hallucination detection targets is one part of a four-way structure that also includes ambiguity and embedded falsehood. Future work can test whether interventions along the probe directions, or probe-guided decoding that triggers abstention, clarification, or premise correction, change response behavior; that is a causal evaluation beyond the scope of this diagnostic study. Evaluation on independently constructed datasets, and on models with different post-training procedures, would assess how broadly the findings generalize.

## Limitations

**Benchmark scale and ontology.** The study uses 200 manually constructed prompts (50 per type) and a paraphrase extension (186 groups, 558 prompts) derived from them. Prompts were iteratively rewritten to balance surface features across types; the rewriting was guided by surface metrics, not by probe AUROC. Generalization to independently constructed datasets, conversational multi-sentence queries, multi-step reasoning prompts, and specialized domains remains untested, as does generalization to mixture-of-experts, multimodal, larger, and closed models. Closed models require behavioral proxies in place of hidden-state access; the extraction protocol applies to any dense residual stream, so mixture-of-experts models are a direct next target. Evaluation on an independently constructed and independently annotated held-out set is the main next validation step, and the construction pipeline (Appendix A) and dataset are released for that purpose.

**Unknown is relative to the screening models.** The Unknown label is defined by screening failure across three open-weight LLMs; an item Unknown to this panel may be Known to a larger or differently trained model, and post-cutoff items drift toward Known as newer checkpoints appear. The model-relativity is part of the type definition (§2); the released screening records let the labels be re-derived for any new panel, and users of a newer model should re-screen the post-cutoff items first. The paraphrase extension is restricted to the 36 of 50 Unknown groups whose Unknown la-

bel holds under rewording (§3); the §8 robustness claim therefore applies to that subset, not the full Unknown set.

**Single-author labels.** All 200 base prompts were authored, classified, and audited by the author; there was no multi-annotator process at construction. A blind audit by two independent annotators, run after construction (Appendix H), reproduces the labels at Cohen's $\kappa$ of 0.894 and 0.835 against the original labels, with author-independent inter-annotator agreement of 0.806; agreement is lowest on the Ambiguous boundary ($\kappa = 0.759$ between annotators), the class that requires the most judgment. Each item includes supporting documentation: a source URL for Known, the per-model screening record for Unknown, the false premise paired with its correction for Misleading, and the enumerated readings for Ambiguous.

**Other limitations.** The four types are evaluated on instruction-tuned models; base models and models with other post-training procedures remain untested. The result is correlational; the probe shows linear recoverability, not that the model *uses* the distinction to drive generation. All paraphrases come from one generator (Claude Opus 4.7); a second generator or human paraphrases would tighten the robustness claim. Within-type sub-structure (Appendix B) reflects both representational content and sub-pool lexical signatures from construction; isolating the two contributions would require sub-pool-balanced rewriting we have not done.

**Reproducibility and compute.** Hidden-state extraction is BF16 with scaled dot-product attention on a single RTX Pro 6000 GPU, under 2 GPU-hours per model; probes and non-LLM baselines are CPU-only. Code, data, screening records, and the result files behind every reported number are available at `https://github.com/zypapers/geometry-uncertainty-repo`.

## Ethical Considerations

A deployed prompt-type classifier could fail unevenly across user groups. Ambiguity and false-premise judgments are culturally and contextually dependent; a system that over-triggers clarification or premise-challenge behavior could frustrate users asking about fiction, hypotheticals, non-standard dialects, or low-resource entities. We view this work as diagnostic, not as a ready-to-deploy moderation or routing system.

## References


Deema Alnuhait, Neeraja Kirtane, Muhammad Khalifa, and Hao Peng. 2025. FactCheckmate: Preemptively detecting and mitigating hallucinations in LMs. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 12413–12428. Association for Computational Linguistics.

Amos Azaria and Tom Mitchell. 2023. The internal state of an LLM knows when it's lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. ArXiv:2304.13734.

Yonatan Belinkov. 2022. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219. ArXiv:2102.12452.

Rohan Bhatnagar, Youran Sun, Chi Andrew Zhang, Yixin Wen, and Haizhao Yang. 2026. DRIFT: Detecting representational inconsistencies for factual truthfulness. *Preprint*, arXiv:2601.14210.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. 2023. Discovering latent knowledge in language models without supervision. In *The Eleventh International Conference on Learning Representations (ICLR)*. ArXiv:2212.03827.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? An analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. ArXiv:2012.14913.

Dan Hendrycks and Kevin Gimpel. 2017. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *The Fifth International Conference on Learning Representations (ICLR)*. ArXiv:1610.02136.

Evan Hernandez, Arnab Sen Sharma, Tal Haklay, Kevin Meng, Martin Wattenberg, Jacob Andreas, Yonatan Belinkov, and David Bau. 2024. Linearity of relation decoding in transformer language models. In *The Twelfth International Conference on Learning Representations (ICLR)*. ArXiv:2308.09124.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2733–2743.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. 2024. Position: The Platonic representation hypothesis. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pages 20617–20642. ArXiv:2405.07987.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38. Article 248.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, and 17 others. 2022. Language models (mostly) know what they know. *Preprint*, arXiv:2207.05221.

Najoung Kim, Phu Mon Htut, Samuel R. Bowman, and Jackson Petty. 2023. $(QA)^2$: Question answering with questionable assumptions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL), Volume 1: Long Papers*, pages 8466–8487. ACL Anthology 2023.acl-long.472.

Najoung Kim, Ellie Pavlick, Burcu Karagol Ayan, and Deepak Ramachandran. 2021. Which linguist invented the lightbulb? Presupposition verification for question-answering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP), Volume 1: Long Papers*, pages 3932–3945. ACL Anthology 2021.acl-long.304.

Nikolaus Kriegeskorte, W. Kyle Simmons, Patrick S. F. Bellgowan, and Chris I. Baker. 2009. Circular analysis in systems neuroscience: The dangers of double dipping. *Nature Neuroscience*, 12(5):535–540.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations (ICLR)*. ArXiv:2302.09664.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Teaching models to express their uncertainty in words. *Transactions on Machine Learning Research*. ArXiv:2205.14334.

Alisa Liu, Zhaofeng Wu, Julian Michael, Alane Suhr, Peter West, Alexander Koller, Swabha Swayamdipta, Noah A. Smith, and Yejin Choi. 2023. We're afraid language models aren't modeling ambiguity. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. ArXiv:2304.14399.

Weitang Liu, Xiaoyun Wang, John D. Owens, and Yixuan Li. 2020. Energy-based out-of-distribution detection. In *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*. ArXiv:2010.03759.

Llama Team, AI @ Meta. 2024. The Llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL), Volume 1: Long Papers*. ArXiv:2212.10511.

Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. 2023. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. ArXiv:2303.08896.

Samuel Marks and Max Tegmark. 2024. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. In *First Conference on Language Modeling (COLM)*. ArXiv:2310.06824.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)*. ArXiv:2202.05262.

Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. AmbigQA: Answering ambiguous open-domain questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. ArXiv:2004.10645.

Thomas E. Nichols and Andrew P. Holmes. 2002. Nonparametric permutation tests for functional neuroimaging: A primer with examples. *Human Brain Mapping*, 15(1):1–25.

Kiho Park, Yo Joong Choe, and Victor Veitch. 2024. The linear representation hypothesis and the geometry of large language models. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*. ArXiv:2311.03658.

Qwen Team. 2024. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Adi Simhi, Itay Itzhak, Fazl Barez, Gabriel Stanovsky, and Yonatan Belinkov. 2025. Trust me, I'm wrong: LLMs hallucinate with certainty despite knowing the answer. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 14665–14688. Association for Computational Linguistics.

Aviv Slobodkin, Omer Goldman, Avi Caciularu, Ido Dagan, and Shauli Ravfogel. 2023. The curious case of hallucinatory (un)answerability: Finding truths in

the hidden states of over-confident large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. ArXiv:2310.11877.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. MPNet: Masked and permuted pre-training for language understanding. In *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*. ArXiv:2004.09297.

P. F. Strawson. 1950. On referring. *Mind*, 59(235):320–344.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4593–4601.

Tanay Varshney, Annie Surla, Michelle Xu, Gomathy Venkata Krishnan, Maximilian Jeblick, David Austin, Neal Vaidya, and Davide Onofrio. 2026. LLM router: Rethinking routing with prefill activations. *Preprint*, arXiv:2603.20895.

Elena Voita and Ivan Titov. 2020. Information-theoretic probing with minimum description length. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. ArXiv:2003.12298.

Edward Vul, Christine Harris, Piotr Winkielman, and Harold Pashler. 2009. Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. *Perspectives on Psychological Science*, 4(3):274–290.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*. ArXiv:2002.10957.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. 2023. C-Pack: Packed resources for general Chinese embeddings. *Preprint*, arXiv:2309.07597.

Xinyan Yu, Sewon Min, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. CREPE: Open-domain question answering with false presuppositions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL), Volume 1: Long Papers*, pages 10457–10480. ACL Anthology 2023.acl-long.583.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, and 2 others. 2023. Representation engineering: A top-down approach to AI transparency. *Preprint*, arXiv:2310.01405.

## A Dataset Construction Protocol

This appendix documents the construction pipeline for the 200-prompt base set and the 558-prompt paraphrase extension at the level of detail needed for independent reproduction.

**Sourcing.** The four type pools were sourced independently. Known prompts were drawn from Wikipedia-verifiable factual queries about widely-covered entities (capitals, scientific consensus, encyclopedic facts). Unknown prompts were sourced from four sub-pools, each meant to elicit failure for a different reason: *post-cutoff* (events after the latest screening model's training cutoff), *long-tail* (rare or low-resource entities), *niche-technical* (specialized identifier systems such as CAS registry, UniProt, ORPHA), and *split-recall* (items where the screening panel showed inconsistent results; labeled `model_dependent` in the released file). Ambiguous prompts were authored by hand to admit at least two readings: referential (multiple plausible entities behind a name), scope (a superlative or comparative whose criterion, population, or horizon is unspecified), or temporal (unspecified time frame). Misleading prompts were authored to embed a false premise declaratively, following the three rules in §2 (presupposition, answer-space coherence, verifiable falsehood; counterfactuals excluded).

**Screening criteria.** Every candidate prompt was tested against the three target models (Qwen2.5-7B-Instruct, Llama-3.1-8B-Instruct, Qwen2.5-32B-Instruct) under one greedy plus three temperature-0.7 sampled generations. Known candidates were retained only if at least two of three models produced the gold answer in the greedy or at least two of three samples. Unknown candidates outside the split-recall sub-pool were retained only if all three models produced zero gold matches across the four generations. Split-recall candidates were retained under a separate flagged rule: at most one model could produce the gold occasionally, and no model could do so reliably. Ambiguous and Misleading candidates were not screening-gated against model output; they were inspected for type-aligned behavior frequency but accepted regardless. The screening panel is therefore the operational definition of Known/Unknown; an item Unknown to this panel may be Known to a larger model.

**Surface-confound balancing.** Candidate prompts were iteratively rewritten to bal-

ance six surface features across types: word count, named-entity count (heuristic, the number of non-leading capitalized tokens), question-form distribution (over the eight forms *what/who/when/where/why/how/which/other*), per-model prompt token count after chat-template application, per-model next-token entropy at the generation-start position, and sentence-embedding cosine to the centroid of Known prompts. Rewriting was guided by surface AUROC, not probe AUROC; an early eighty-prompt iteration showed surface AUROC near 0.96 on Unknown vs. Ambiguous, and three coordinated interventions drove U_vs_A's surface AUROC below 0.78 and kept the maximum pairwise surface AUROC at 0.867 (Table 1): re-selecting Unknown candidates biased toward low-NE non-identifier-form facts, rewriting 91 questions to balance the question-form distribution across types, and shortening declarative-embedded Misleading prompts that ran systematically long.

**Annotation.** All prompts were authored, classified, and audited by the author at construction time; there was no multi-annotator process at that stage (see Limitations). An independent two-annotator blind audit run after construction is reported in Appendix H. Sub-types within each type were assigned at construction time following the published definitions (§2 and Appendix B). Type-aligned behavior was not used as a labeling criterion; one behavior label is derived per prompt from its seven generations (§7), independently of type.

**Paraphrase generation.** For the paraphrase extension, each base prompt received two paraphrases generated by Claude Opus 4.7 under explicit type-preservation rules: paraphrases must elicit the same gold answer for Known, must continue to elicit cross-model failure for Unknown under post-generation screening, must preserve the set of valid readings for Ambiguous, and must preserve the declarative-embedded false premise for Misleading.

**Unknown paraphrase filter and exclusions.** The same screening panel was applied to each Unknown paraphrase: the paraphrase passes only if at least two of three target models produce zero correct answers across one greedy plus three sampled generations. This filter is more permissive than the base rule for Unknown items outside the split-recall sub-pool (3/3 zero matches): rephrasing can occasionally elicit gold from one model even when the base prompt does not, and we accept the paraphrase if at least two models still fail. Fourteen of fifty Unknown base groups failed this filter and were excluded from the paraphrase extension. The summary count is in §3; the per-item exclusion list, with the failing Unknown paraphrases and their per-model screening counts, is included in the release.

**Release.** The base dataset with its per-model screening records, the paraphrase extension with its exclusion list, the surface-balance summary, the analysis code, and the result files behind every reported number are released under a permissive license at `https://github.com/zypapers/geometry-uncertainty-repo`. Because Known and Unknown are panel-relative, the screening records allow the labels to be re-derived for a new panel of models; users evaluating a newer model should re-screen the post-cutoff Unknown items first, since those are the most likely to have drifted toward Known.

**Chat-template note.** Llama-3.1's official chat template injects a fixed knowledge-cutoff and current-date string into every system header; Qwen2.5's template injects no date. The prefix is constant within each run, but it does mean that the 14 temporal-ambiguous items ("current", "latest") are given a fixed date by the chat template for Llama-8B and left unanchored for Qwen-7B. This may be related to Llama-8B's lower clarification rate on Ambiguous prompts (42% vs. 64%, Table 4); we did not control for it.

**Robustness to borderline rows.** A label audit conducted after the initial construction flagged 36 rows as borderline: Ambiguous items whose wording leaves the ambiguity weak, Misleading items phrased disjunctively so that the true answer appears in the prompt, Known and Unknown items whose gold answer required re-verification, and one Known item whose edited wording is unanswerable. Type labels were retained (the released file carries these rows with their audit notes); as a sensitivity check we re-ran the pairwise probes with all 36 rows excluded. Minimum in-window pairwise AUROC on the remaining 164 rows is 0.990 / 0.984 / 0.993 (Qwen-7B / Llama-8B / Qwen-32B), against 0.987 / 0.981 / 0.994 on the full set. The exclusion list is included with the data.

## B Sub-type Analyses (full detail)

This appendix expands the sub-type analysis of §6 with the full per-pair results and the LDA / t-SNE visualizations.

**Premise sub-types within Misleading.** We label the 50 Misleading prompts by sub-type: *misconception* (12 prompts, e.g. "Why do camels store water in their humps?"), *wrong attribution* (12, e.g. "Which Roman emperor wrote Hamlet?"), *nonexistent entity* (8, e.g. "Who won the 1994 Nobel Prize in Mathematics?"), *wrong date* (11, e.g. "When did the Magna Carta appear in 1300?"), and *impossible premise* (7, e.g. "What is the velocity of faster-than-light neutrinos?"). At each model's peak sub-type layer we train all ten pairwise sub-type probes; per-pair AUROCs are in Table 7. Mean pairwise AUROC across the ten contrasts reaches 0.990 on Qwen-7B, 0.968 on Llama-8B, and 0.992 on Qwen-32B. Five of the ten pairs reach AUROC 1.000 on at least two models; the hardest pairs are *nonexistent entity* vs. *impossible premise* on Qwen-7B (0.932), *wrong attribution* vs. *nonexistent entity* on Llama-8B (0.921), and *misconception* vs. *impossible premise* on Qwen-32B (0.964). Figure 4 visualizes the five sub-types under LDA and t-SNE projections.

| sub-type pair | Qwen-7B | Llama-8B | Qwen-32B |
|---|---|---|---|
| MC vs IP | 0.983 | 0.924 | 0.964 |
| MC vs NE | 1.000 | 1.000 | 1.000 |
| MC vs WA | 1.000 | 1.000 | 1.000 |
| MC vs WD | 1.000 | 1.000 | 1.000 |
| NE vs IP | 0.932 | 0.946 | 0.975 |
| NE vs WD | 0.989 | 0.945 | 1.000 |
| WA vs IP | 1.000 | 1.000 | 1.000 |
| WA vs NE | 0.996 | 0.921 | 0.996 |
| WA vs WD | 1.000 | 0.945 | 0.983 |
| WD vs IP | 1.000 | 1.000 | 1.000 |

Table 7: Pairwise AUROC for the five Misleading sub-types at each model's peak sub-type layer ($L_{23}$ Qwen-7B, $L_{32}$ Llama-8B, $L_{33}$ Qwen-32B). MC = misconception, WA = wrong attribution, NE = nonexistent entity, WD = wrong date, IP = impossible premise.

**Ambiguity sub-types within Ambiguous.** We label the 50 Ambiguous prompts by ambiguity kind: *referential* (18, multiple plausible entities behind a name), *scope* (18, a superlative or comparative whose criterion, population, or horizon is unspecified), and *temporal* (14, an answer that depends on when the question is asked). A referential item qualifies only under a three-part test:

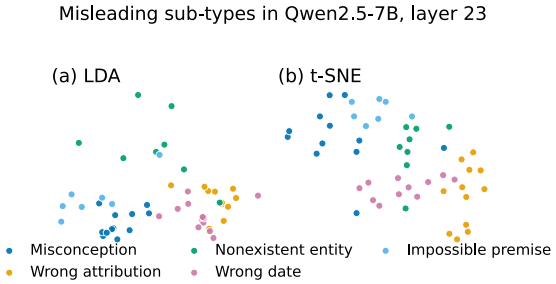



Figure 4: Five premise sub-types within Misleading at Qwen2.5-7B's peak sub-type layer ($L_{23}$). Panel (a): LDA (first two components $\sim 73\%$ of between-class variation). Panel (b): unsupervised t-SNE. Mean pairwise sub-type AUROC: 0.990 on Qwen-7B, 0.968 on Llama-8B, 0.992 on Qwen-32B. Illustrative; quantitative claims use cross-validated probes.

the competing referents must be salient from common knowledge rather than discoverable only by search (Cambridge, England vs. Cambridge, Massachusetts qualifies; a small town that happens to share a country's name does not); the question must be coherent with a different answer under each referent; and a competent assistant would reasonably respond by asking which is meant. The test requires plausible ambiguity in context, not mere non-uniqueness of a proper name. Annotator disagreements on the Ambiguous boundary occur across all three sub-types, including referential items near this threshold (Appendix H). All three pairwise sub-type contrasts separate at AUROC $\geq 0.986$ across the three models, with mean pairwise AUROC $\geq 0.994$ on every model (0.999 on Qwen-7B, 0.994 on Llama-8B, 0.999 on Qwen-32B); the lowest single pair is referential vs. scope on Llama-8B at 0.986. Per-pair AUROCs are in Table 8. Figure 5 visualizes the three sub-types. The separation is sharper than for Misleading sub-types, consistent with the intuition that ambiguity-kind is more linguistically explicit (largely determined by the surface structure of the question) while premise-kind is an inferential property over world knowledge.

| sub-type pair | Qwen-7B | Llama-8B | Qwen-32B |
|---|---|---|---|
| R vs S | 0.999 | 0.986 | 0.997 |
| R vs T | 0.999 | 0.998 | 1.000 |
| S vs T | 1.000 | 1.000 | 1.000 |

Table 8: Pairwise AUROC for the three Ambiguous sub-types at each model's peak sub-type layer. R = referential, S = scope, T = temporal.

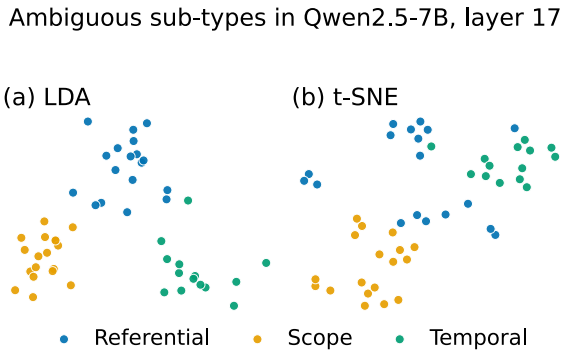



Figure 5: Three ambiguity sub-types within Ambiguous in Qwen2.5-7B at the peak sub-type layer. Mean pairwise sub-type AUROC: 0.999 on Qwen-7B, 0.994 on Llama-8B, 0.999 on Qwen-32B. Illustrative; quantitative claims use cross-validated probes.

**Sub-type structure within Unknown.** The four Unknown sub-types are *post-cutoff* (8 prompts, e.g. "Who won the 2024 Booker Prize?"), *long-tail* (13, e.g. "Which IATA code identifies Nauru's airport?"), *niche-technical* (19, e.g. "What is the InChIKey for ethanol?"), and *split-recall* (10, e.g. "How many people live in Liechtenstein?"; `model_dependent` in the released file). They decompose geometrically: mean pairwise AUROC reaches 0.961 on Qwen-7B, 0.958 on Llama-8B, and 0.967 on Qwen-32B. Per-pair AUROCs are in Table 9; Figure 6 visualizes the four sub-types under LDA and t-SNE projections. Five of six pairs reach AUROC $\geq$ 0.997 on every model; the exception is long-tail vs split-recall (0.751–0.800), where both sub-types contain rare-entity questions with overlapping lexical structure. The within-Unknown separation is at least partly attributable to lexical sub-pool signatures: post-cutoff prompts contain date markers, niche-technical contain identifier strings (CAS numbers, ICD codes, etc.). The four-type analysis in §5 groups all Unknown sub-types into a single class.

| sub-type pair | Qwen-7B | Llama-8B | Qwen-32B |
|---|---|---|---|
| PC vs LT | 1.000 | 1.000 | 1.000 |
| PC vs NT | 1.000 | 1.000 | 1.000 |
| PC vs SR | 1.000 | 1.000 | 1.000 |
| LT vs NT | 0.999 | 0.999 | 1.000 |
| LT vs SR | 0.766 | 0.751 | 0.800 |
| NT vs SR | 0.998 | 0.997 | 1.000 |

Table 9: Pairwise AUROC for the four Unknown sub-types at each model's peak sub-type layer. PC = post-cutoff, LT = long-tail, NT = niche-technical, SR = split-recall.

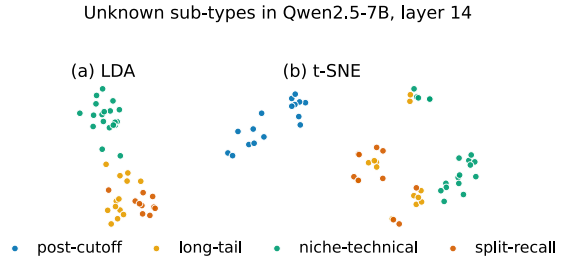



Figure 6: Four Unknown sub-types in Qwen2.5-7B at the peak sub-type layer. Panel (a): LDA (first two components $\sim 94\%$ of between-class variation). Panel (b): unsupervised t-SNE. Mean pairwise sub-type AUROC: 0.961 on Qwen-7B, 0.958 on Llama-8B, 0.967 on Qwen-32B. The long-tail and split-recall clouds overlap, consistent with the 0.751–0.800 AUROC on that pair (Table 9). Illustrative; quantitative claims use cross-validated probes.

## C Shuffled-Control Details on the Paraphrase Extension

Across the five shuffled seeds and the 65 in-window layers across the three models ($15 + 17 + 33$), one seed (9001) reaches the highest empirical-null AUROC for the majority of model–pair combinations. This is a fixed-seed artifact of the group-level permutation: with 186 paraphrase groups and four type labels, $\binom{186}{50,36,50,50}$ permutations vary in how favorably the resulting cluster of 86 groups (for Unknown-involving pairs) aligns the surface features. Table 10 reports, per (pair, model) cell, the maximum shuffled-seed AUROC across all five shuffle seeds and all in-window layers, alongside the per-pair required threshold and the resulting gap. Every cell clears its threshold by at least 0.115. The tightest-margin cell is Ambiguous vs. Misleading on Llama-8B (gap 0.115, max-shuffled 0.685 vs. required 0.800, located at layer 24 under seed 9001). The absolute-highest shuffled AUROC anywhere is 0.743 on Qwen-7B K_vs_A at layer 19 (gap 0.140, against the higher per-pair threshold of 0.883).

All shuffled-label AUROCs remain below their required thresholds within the model-specific windows, for every seed, pair, and model.

## D Middle-Third Layer Robustness

Table 11 reports the per-pair AUROC averaged over the middle-third layers of each model's window (Qwen-7B $L_{19}$–$L_{24}$, Llama-8B $L_{21}$–$L_{27}$, Qwen-32B $L_{43}$–$L_{54}$), with the minimum over those middle-third layers in parentheses. The same

| pair | req. | 7B | 8B | 32B | min gap |
|---|---|---|---|---|---|
| K_vs_U | 0.909 | 0.574 | 0.559 | 0.606 | +0.304 |
| K_vs_A | 0.883 | 0.743 | 0.702 | 0.697 | +0.140 |
| K_vs_M | 0.917 | 0.648 | 0.686 | 0.686 | +0.231 |
| U_vs_A | 0.880 | 0.734 | 0.735 | 0.732 | +0.145 |
| U_vs_M | 0.835 | 0.633 | 0.666 | 0.655 | +0.169 |
| A_vs_M | 0.800 | 0.598 | **0.685** | 0.663 | **+0.115** |

Table 10: Maximum shuffled-seed AUROC per (pair, model) cell, taken across all five shuffle seeds and all in-window layers. *req.* = per-pair required threshold (max base-or-paraphrase surface AUROC + 0.05, matching the *Required* column of Table 5). *min gap* = minimum across models of (req. − max-shuffled). Tightest-margin cell in bold.

middle-third rule is applied to each model-specific window.

| pair | Qwen-7B | Llama-8B | Qwen-32B |
|---|---|---|---|
| K_vs_U | 0.963 (0.957) | 0.952 (0.946) | 0.976 (0.966) |
| K_vs_A | 0.991 (0.989) | 0.991 (0.990) | 0.989 (0.984) |
| K_vs_M | 0.994 (0.989) | 0.988 (0.986) | 0.996 (0.993) |
| U_vs_A | 0.998 (0.996) | 0.998 (0.997) | 0.996 (0.994) |
| U_vs_M | 0.993 (0.992) | 0.996 (0.995) | 1.000 (0.999) |
| A_vs_M | 0.987 (0.983) | 0.990 (0.988) | 0.991 (0.986) |

Table 11: Per-pair AUROC averaged over the middle-third in-window layers, with the minimum over those layers in parentheses. Every cell sits above the surface baseline on its pair.

## E Four-Way Classification

Table 12 summarizes the four-way multinomial probe (§5) on all three models for the base set and the paraphrase extension, with the maximum shuffled-label accuracy over five permutations and all in-window layers, and the strongest non-hidden baseline. Table 13 lists every baseline's four-way accuracy under the same cross-validation. Tables 14 and 15 give the confusion matrices at each model's peak layer, with counts pooled over the five cross-validation seeds. On the base set Known↔Unknown is the largest pairwise source of confusion in each model; on the paraphrase extension the largest off-diagonal mass is Unknown predicted as Known (91–106 of 540 Unknown decisions), consistent with the wording-sensitivity of Unknown labels discussed in §3.

## F Covariate-Adjusted Probes

The baseline competition in §4 is the classification analogue of covariate adjustment. We also ran the two regression-style forms on all three models at each pair's unperturbed peak layer, under the standard protocol (Table 16). *Concatenation* appends the six z-scored surface covariates (word count, named-entity count, question form as eight indicators, prompt token count, next-token entropy, and sentence-embedding cosine to the Known centroid computed using the training fold) to the hidden state; z-scoring uses training-fold statistics. *Residualization* regresses each hidden dimension on the covariates within the training fold (ordinary least squares with intercept) and probes the residuals; this credits all variance shared between the hidden state and the covariates to the covariates; we use it to assess sensitivity to covariate adjustment. *Text-only residualization* repeats this against the construction-controlled text features alone (word count, entity count, question form, token count), excluding the two model-derived features (entropy and embedding cosine). The residualized permutation null re-runs the residualized probe under five label permutations.

Concatenation changes seed-mean AUROC by at most 0.009 on any cell (a decrease, Llama-8B K_vs_U) and never raises it by more than 0.004.

| set | model | acc. [95% CI] | AUROC | shuf. | base. |
|---|---|---|---|---|---|
| base | Qwen-7B | 0.910 [0.867, 0.947] | 0.985 | 0.309 | 0.692 |
| base | Llama-8B | 0.896 [0.852, 0.937] | 0.986 | 0.313 | 0.692 |
| base | Qwen-32B | 0.938 [0.902, 0.968] | 0.995 | 0.329 | 0.692 |
| para | Qwen-7B | 0.903 [0.878, 0.926] | 0.981 | 0.325 | 0.669 |
| para | Llama-8B | 0.894 [0.869, 0.919] | 0.981 | 0.327 | 0.669 |
| para | Qwen-32B | 0.922 [0.899, 0.943] | 0.992 | 0.362 | 0.669 |

Table 12: Four-way probe at each model's in-window peak layer. *acc.* = seed-mean accuracy with 95% bootstrap CI; *AUROC* = macro one-vs-rest; *shuf.* = maximum shuffled-label accuracy over five permutations and all in-window layers; *base.* = strongest non-hidden baseline (TF-IDF word+char on the base set, BGE-large on the paraphrase set). Chance is 0.250. Paraphrase rows use group-aware cross-validation.

| baseline | base | paraphrase |
|---|---|---|
| surface, Qwen-7B entropy | 0.474 | 0.498 |
| surface, Llama-8B entropy | 0.544 | 0.533 |
| surface, Qwen-32B entropy | 0.441 | 0.508 |
| TF-IDF word 1–2 grams | 0.668 | 0.639 |
| TF-IDF char 3–5 grams | 0.648 | 0.617 |
| TF-IDF word + char | 0.692 | 0.648 |
| all-MiniLM-L6-v2 | 0.594 | 0.605 |
| all-mpnet-base-v2 | 0.659 | 0.643 |
| BGE-large-en-v1.5 | 0.690 | 0.669 |

Table 13: Four-way accuracy of every non-hidden baseline under the same cross-validation as the probe. The six-feature surface baseline includes per-model next-token entropy and so has one row per model; the lexical and encoder baselines are model-independent.

| true \ pred. | Qwen-7B ($L_{20}$) K | U | A | M | Llama-8B ($L_{17}$) K | U | A | M | Qwen-32B ($L_{40}$) K | U | A | M |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Known | **224** | 22 | 0 | 4 | **225** | 15 | 0 | 10 | **233** | 12 | 0 | 5 |
| Unknown | 24 | **219** | 5 | 2 | 24 | **224** | 2 | 0 | 14 | **230** | 0 | 6 |
| Ambiguous | 10 | 0 | **229** | 11 | 11 | 5 | **223** | 11 | 9 | 3 | **231** | 7 |
| Misleading | 0 | 8 | 4 | **238** | 5 | 7 | 14 | **224** | 0 | 6 | 0 | **244** |

Table 14: Four-way confusion matrices on the base set at each model's peak layer; counts pooled over the five cross-validation seeds (each row sums to $250 = 50 \times 5$).

| true \ pred. | Qwen-7B ($L_{20}$) K | U | A | M | Llama-8B ($L_{17}$) K | U | A | M | Qwen-32B ($L_{40}$) K | U | A | M |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Known | **673** | 41 | 10 | 26 | **700** | 26 | 0 | 24 | **714** | 17 | 9 | 10 |
| Unknown | 95 | **443** | 0 | 2 | 106 | **419** | 15 | 0 | 91 | **449** | 0 | 0 |
| Ambiguous | 31 | 5 | **688** | 26 | 28 | 3 | **691** | 28 | 34 | 4 | **682** | 30 |
| Misleading | 0 | 20 | 14 | **716** | 20 | 5 | 40 | **685** | 5 | 6 | 12 | **727** |

Table 15: Four-way confusion matrices on the paraphrase extension (group-aware cross-validation) at each model's peak layer; counts pooled over five seeds (Known, Ambiguous, Misleading rows sum to 750; Unknown to 540, since 36 of 50 Unknown groups survive the paraphrase filter).

Full residualization lowers every cell; 16 of 18 stay above their surface-baseline AUROC, and the two that do not (Llama-8B K_vs_M at 0.796 vs. a 0.867 baseline; Qwen-32B K_vs_U at 0.745 vs. 0.763) are both Known-involving pairs, where the surface baseline is set by the next-token-entropy confidence signature (§6). Text-only residualization keeps all 18 cells above their baselines, so the below-baseline movement appears only when the two model-derived features (entropy and embedding cosine) are included, not under the text confounds alone. Every residualized cell stays at least $+0.22$ above the mean of its residualized permutation null and at least $+0.14$ above the single worst permutation.

| model | pair | bar | hidden | concat | resid. | text | null |
|---|---|---|---|---|---|---|---|
| Qwen-7B | K_vs_U | 0.763 | 0.987 | 0.988 | 0.799 | 0.822 | 0.516 |
| Qwen-7B | K_vs_A | 0.833 | 0.994 | 0.996 | 0.892 | 0.991 | 0.514 |
| Qwen-7B | K_vs_M | 0.867 | 0.997 | 0.997 | 0.890 | 0.945 | 0.518 |
| Qwen-7B | U_vs_A | 0.753 | 0.999 | 0.999 | 0.873 | 0.907 | 0.547 |
| Qwen-7B | U_vs_M | 0.785 | 0.998 | 0.998 | 0.917 | 0.924 | 0.516 |
| Qwen-7B | A_vs_M | 0.718 | 0.995 | 0.995 | 0.915 | 0.963 | 0.526 |
| Llama-8B | K_vs_U | 0.763 | 0.981 | 0.972 | 0.802 | 0.830 | 0.529 |
| Llama-8B | K_vs_A | 0.833 | 0.996 | 0.999 | 0.869 | 0.996 | 0.550 |
| Llama-8B | K_vs_M | 0.867 | 0.992 | 0.991 | **0.796** | 0.916 | 0.528 |
| Llama-8B | U_vs_A | 0.753 | 1.000 | 1.000 | 0.874 | 0.911 | 0.518 |
| Llama-8B | U_vs_M | 0.785 | 0.999 | 0.999 | 0.900 | 0.907 | 0.491 |
| Llama-8B | A_vs_M | 0.718 | 0.991 | 0.994 | 0.949 | 0.971 | 0.513 |
| Qwen-32B | K_vs_U | 0.763 | 0.994 | 0.994 | **0.745** | 0.802 | 0.522 |
| Qwen-32B | K_vs_A | 0.833 | 0.998 | 0.998 | 0.895 | 0.994 | 0.489 |
| Qwen-32B | K_vs_M | 0.867 | 1.000 | 1.000 | 0.888 | 0.936 | 0.480 |
| Qwen-32B | U_vs_A | 0.753 | 0.999 | 0.999 | 0.872 | 0.924 | 0.545 |
| Qwen-32B | U_vs_M | 0.785 | 1.000 | 1.000 | 0.929 | 0.955 | 0.510 |
| Qwen-32B | A_vs_M | 0.718 | 0.995 | 0.995 | 0.944 | 0.966 | 0.489 |

Table 16: Covariate-adjusted probe AUROC at each pair's peak layer (5-fold, 5 seeds). *bar* = surface-baseline AUROC (Table 1); *hidden* = hidden state alone (Llama-8B K_vs_U is the in-window peak, $L_{16}$); *concat* = hidden state plus the six z-scored covariates; *resid.* = hidden state residualized against all six covariates; *text* = residualized against the four text covariates only; *null* = mean residualized AUROC over five label permutations. Bold marks the two residualized cells below their surface baseline.

## G Label-Perturbation Robustness

To bound how much the reported AUROCs depend on the exact placement of the judgment-dependent labels, we flipped a random 5% and 10% of the Ambiguous and Misleading labels (A↔M swaps among the 100 A/M rows; Known and Unknown never flipped), ten independent draws per rate, and retrained the Qwen-7B probe at each pair's unperturbed peak layer under the standard protocol. Table 17 reports AUROC scored against the perturbed labels (the primary metric) and against the original labels. The only pair that moves materially is A_vs_M, the pair whose labels were directly corrupted; K_vs_U is untouched by construction.

## H Annotation Audit

**Protocol.** Two annotators, independent of the author and blind to the original labels and to each other, classified all 200 base prompts against the published type definitions. Because Known versus Unknown is fixed by the screening protocol rather than by human judgment, the audit uses a three-way collapse: *determinate-factual* (DF; Known or

| pair | ref. | 5%: mean (min) | 10%: mean (min) | vs. true | margin |
|---|---|---|---|---|---|
| K_vs_U | 0.987 | 0.987 (0.987) | 0.987 (0.987) | 0.987 | +0.224 |
| K_vs_A | 0.994 | 0.990 (0.972) | 0.989 (0.983) | 0.990 | +0.139 |
| K_vs_M | 0.997 | 0.996 (0.996) | 0.996 (0.992) | 0.997 | +0.125 |
| U_vs_A | 0.999 | 0.997 (0.990) | 0.993 (0.976) | 0.997 | +0.223 |
| U_vs_M | 0.998 | 0.993 (0.980) | 0.987 (0.977) | 0.997 | +0.192 |
| A_vs_M | 0.995 | 0.944 (0.911) | 0.861 (0.799) | 0.958 | +0.081 |

Table 17: Qwen-7B peak-layer AUROC under random A/M label flips. *ref.* = unperturbed; *5%* and *10%* = mean (minimum) over ten draws, scored against the perturbed labels; *vs. true* = mean at the 10% rate scored against the original labels; *margin* = worst-draw AUROC at either rate minus the pair's surface-baseline AUROC (Table 1).

Unknown), *ambiguous* (AMB), and *false-premise* (FP). The audit therefore validates the Ambiguous and Misleading boundaries only. Each annotator could flag an item as unsure; annotator A did not use the flag and annotator B flagged 5% of items. Original labels were retained for the reported analyses and were not revised in response to the audit; disagreements are documented below.

**Agreement.** Against the original labels, Cohen's $\kappa$ is 0.894 for annotator A (93.5% agreement) and 0.835 for annotator B (90.0%); author-independent agreement between the two annotators is $\kappa = 0.806$ (88.5%), and three-rater Fleiss' $\kappa$ is 0.846. Per boundary, $\kappa$ for ambiguous-vs-rest is 0.889 (A), 0.800 (B), and 0.759 (A vs. B); for false-premise-vs-rest it is 0.892, 0.863, and 0.833. On the 50 Misleading items, A confirms the false-premise label on 45 (90%) and B on 43 (86%); on the 50 Ambiguous items, A confirms 43 (86%) and B 38 (76%), with B moving 9 Ambiguous items to determinate-factual. The Ambiguous boundary is thus the least reliable, as expected for the most judgment-dependent class. Table 18 gives the three confusion matrices.

| | author vs. A | | | author vs. B | | | A vs. B | | |
|---|---|---|---|---|---|---|---|---|---|
| | DF | AMB | FP | DF | AMB | FP | DF | AMB | FP |
| DF | **99** | 0 | 1 | **99** | 1 | 0 | **102** | 5 | 1 |
| AMB | 5 | **43** | 2 | 9 | **38** | 3 | 6 | **34** | 4 |
| FP | 4 | 1 | **45** | 6 | 1 | **43** | 6 | 1 | **41** |

Table 18: Annotation-audit confusion matrices (rows = first-named rater, columns = second). DF = determinate-factual (Known or Unknown), AMB = ambiguous, FP = false-premise.

**Per-item disagreements.** Twenty-seven items drew at least one disagreement (Table 19). Both annotators disagreed with the original label on six items and agreed on an alternative label for four of them; on seven items only A disagreed, and on fourteen only B. *Who wrote the Iliad — Shakespeare or Homer?* is phrased disjunctively, so the true answer appears in the prompt, and both annotators read it as a factual question; it is one of the four disjunctive Misleading items in the exclusion set of Appendix A. *Which 1959 moon landing changed history?* retains its false-premise label (no crewed landing occurred in 1959), but both annotators read it as a date question. *Why do humans only have five senses according to Aristotle?* admits a determinate reading as a question about Aristotle's view, which is what both annotators chose; the intended false premise is the five-senses misconception itself. *When did Isaac Newton's famous orchard incident occur?* was rated determinate-factual by both, treating the apocryphal story as a fixed cultural reference. On the two remaining items (*Apple's iPhone . . . 2025 revenue record*; *Washington's State Department . . . COP29*), both referential items near the threshold of Appendix B, the annotators split between determinate-factual and false-premise. Three of the six (the Iliad, Newton, and iPhone items) are in the 36-row exclusion set whose removal leaves the headline result unchanged (Appendix A). The single-annotator disagreements concentrate on temporal items that one annotator read as answerable at a fixed date (*latest version of Python*, *most recent NASA mission*) and on disjunctive Misleading items; only two of the 27 are Known or Unknown items read as Ambiguous or Misleading (one Unknown item by each annotator), and no Known item is affected.

## I Representative Prompts

Tables 20–23 list representative prompts from each type, drawn from the dataset to illustrate each subtype's definition. The full 200-prompt set is included in the release.

| item | prompt | orig. | A | B |
|---|---|---|---|---|
| *Both annotators disagree with the original label* | | | | |
| misleading_v3_028 | Who wrote the Iliad — Shakespeare or Homer? | FP | DF | DF |
| misleading_v3_053 | Which 1959 moon landing changed history? | FP | DF | DF |
| misleading_v3_012 | Why do humans only have five senses according to Aristotle? | FP | DF | DF |
| ambiguous_v3_038 | When did Isaac Newton's famous orchard incident occur? | AMB | DF | DF |
| ambiguous_v3_029 | When did Apple's iPhone set the 2025 revenue record? | AMB | FP | DF |
| ambiguous_v3_028 | When did Washington's State Department announce the COP29 outcome? | AMB | FP | DF |
| *Only annotator A disagrees* | | | | |
| unknown_v3r3_060 | What spacecraft launched to Europa in 2024? | DF | FP | DF |
| misleading_v3_005 | How does Napoleon's short stature affect his battlefield decisions? | FP | DF | FP |
| misleading_v3_071 | What technology did Bell invent in 1976? | FP | AMB | FP |
| ambiguous_v3_080 | What is the latest version of Python? | AMB | DF | AMB |
| ambiguous_v3_090 | What is the latest update to the Linux kernel? | AMB | DF | AMB |
| ambiguous_v3_094 | When does the latest Mars rover land? | AMB | DF | AMB$^{u}$ |
| ambiguous_v3_099 | Who is Britain's PM after the latest Conservative leadership election in London? | AMB | DF | AMB |
| *Only annotator B disagrees* | | | | |
| unknown_v3r4_001 | What is the surname of the 2024 World Series winning team? | DF | DF | AMB |
| misleading_v3_025 | Who invented the light bulb — Einstein or earlier? | FP | FP | DF$^{u}$ |
| misleading_v3_034 | Who invented the printing press — Galileo or earlier? | FP | FP | DF |
| misleading_v3_043 | Who built the pyramids — Cleopatra or another ruler? | FP | FP | AMB$^{u}$ |
| misleading_v3_082 | How fast does ether wind move when measured at Earth's surface? | FP | FP | DF |
| ambiguous_v3_015 | Where is Cambridge University located? | AMB | AMB | DF |
| ambiguous_v3_031 | When did Tesla unveil its 2024 Q3 product? | AMB | AMB | FP |
| ambiguous_v3_048 | Why is one startup considered the most successful? | AMB | AMB | FP |
| ambiguous_v3_061 | Why did the Industrial Revolution change society? | AMB | AMB | DF$^{u}$ |
| ambiguous_v3_070 | What is the most accurate climate model in the IPCC AR6 report's 2030 forecast? | AMB | AMB | DF |
| ambiguous_v3_071 | Which AI assistant performs best on Google's MMLU and HumanEval benchmarks? | AMB | AMB | FP |
| ambiguous_v3_076 | What is the most recent NASA mission to launch? | AMB | AMB | DF |
| ambiguous_v3_077 | What is today's exchange rate from USD to EUR? | AMB | AMB | DF |
| ambiguous_v3_081 | What is the most recent SpaceX launch? | AMB | AMB | DF |

Table 19: All 27 items on which at least one annotator disagreed with the original label, under the three-way collapse (DF = determinate-factual, i.e. Known or Unknown; AMB = ambiguous; FP = false-premise). $^{u}$ marks an item annotator B flagged as unsure. Original labels were retained for the reported analyses.

| *prompt* |
|---|
| What is the capital of Australia? |
| What is the capital of Brazil? |
| What is the capital of South Korea? |
| Which city is the capital of Mongolia? |
| Which city is the capital of Kazakhstan? |
| Which city is the constitutional capital of Bolivia? |
| When was Mahler's First Symphony first performed? |
| What is the national dish of Hungary? |
| Why is the cherry blossom Japan's national flower? |

Table 20: Representative Known prompts.

| *subtype* | *prompt* |
|---|---|
| long-tail | Which IATA code identifies Nauru's airport? |
| long-tail | Which IATA code identifies Niue's airport? |
| long-tail | When was North Macedonia's central bank established? |
| niche-technical | What is the PubChem CID for paracetamol? |
| niche-technical | What is the InChIKey for ethanol? |
| niche-technical | What is the ICD-10 code for influenza? |
| split-recall | How many people live in Liechtenstein? |
| split-recall | Which IATA code identifies St. Helena's airport? |
| split-recall | How large is the Caspian Sea in square kilometers? |
| post-cutoff | Who won the 2024 Booker Prize? |
| post-cutoff | Which team won the 2024 NBA Finals? |
| post-cutoff | Who won the 2024 Eurovision Song Contest? |

Table 21: Unknown prompts (3 per sub-type).

| *subtype* | *prompt* |
|---|---|
| referential | Which meaning of Mercury is most well-known? |
| referential | Who is Cyrus? |
| referential | Which Cambridge produced famous output? |
| referential | When is the climate of Springfield warmest? |
| scope | How does one choose the best programming language? |
| scope | Why is one diet considered the best? |
| scope | How influential are the most influential books? |
| scope | How safe is the safest investment? |
| temporal | Who is the current US president? |
| temporal | What is the current population of Tokyo? |
| temporal | What is the most recent NASA mission to launch? |
| temporal | What is today's exchange rate from USD to EUR? |

Table 22: Ambiguous prompts (4 per sub-type).

| *subtype* | *prompt* |
|---|---|
| misconception | Why do ostriches bury their heads in sand when scared? |
| misconception | How does Napoleon's short stature affect his battlefield decisions? |
| wrong attribution | Which song did Bach write for the Beatles? |
| wrong attribution | What year did Tesla invent the telephone? |
| nonexistent entity | Who won the 1994 Nobel Prize in Mathematics? |
| nonexistent entity | How many countries have signed the Treaty of Pluto? |
| wrong date | Who succeeded Lincoln after his 1875 assassination? |
| wrong date | When did the Magna Carta appear in 1300? |
| impossible premise | What is the velocity of faster-than-light neutrinos? |
| impossible premise | What is the temperature needed to create cold fusion at room temperature? |

Table 23: Misleading prompts (2 per sub-type).