WHEN RETRIEVAL FAILS: STRUCTURED PARTIAL INFORMATION IN LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Language-model evaluation typically treats factual retrieval as binary: correct or not. Human memory research instead distinguishes successful recall from tip-of-the-tongue (TOT) states, where partial information about an unretrievable target remains accessible. We ask whether language models show a computational analogue, and state our result up front: we do *not* establish that models retain *target-specific* information after retrieval failure. We establish a weaker phenomenon – candidate partial-retrieval failures show greater *aggregate* accessibility to answer attributes than matched low-knowledge failures, via a Partial Retrieval Score (PRS) over independent attribute probes.

Across three open-weight 7–14B models and two natural-item samples (up to 857 items), this aggregate effect replicates robustly ($p = 1.2 \times 10^{-3}$) and survives restriction to label-stable items. But every attribute-level lead we identified – geographic region, then character length – fails independent replication, a forced-choice specificity probe, and pre-cue hidden-state probing, and run-to-run label agreement is only 89.5% under quantized greedy decoding. A causal cue-intervention experiment shows a correct-value cue helps retrieval equally in *both* groups (the T-vs-I interaction is null) – evidence that cue content is useful once supplied, not that it was already present internally. We treat the aggregate PRS effect as the paper's one claim to survive scrutiny, and attribute-level or internal-state claims – including our own – as requiring more than a single pilot to trust.

1 INTRODUCTION

Large language models (LLMs) possess substantial factual knowledge, yet frequently fail to retrieve or express information they may have encountered during training – motivating research on hallucination detection, uncertainty estimation, and latent knowledge extraction built on a binary abstraction: the model either knows the answer or does not. Human memory suggests a richer possibility: people sometimes cannot retrieve a fact but retain a strong sense they know it, classically described by Brown and McNeill (1966) as the "tip-of-the-tongue" (TOT) state, where partial information about an inaccessible target remains accessible – evidence that retrieval is not all-or-none. This motivates a different question about language models:

> *When a language model fails to retrieve an answer, is the failure complete ignorance, or does the model retain a structured partial trace of the target?*

This differs from asking whether the hidden state contains the correct answer: a model may retain information about the answer without producing it exactly, and that residual information may itself have structure (e.g., after failing to retrieve a person's name, still knowing the target is a painter with a two-word name associated with Europe). We therefore study retrieval failures at a finer granularity: successful retrieval $\rightarrow$ partial retrieval $\rightarrow$ genuine ignorance, with the middle state as our primary object – not claiming a model experiences the subjective phenomenology of human TOT, only that its computational state may exhibit analogous properties. Our hypothesis: after controlling for exact retrieval failure, Z (target attributes) is more informative about Y (the answer) under partial-retrieval failure than under ignorance failure, tested via average probe accuracy:

$$\mathbb{E}[PRS \mid T] > \mathbb{E}[PRS \mid I], \tag{1}$$

where PRS (Section 7) is the fraction of independent attribute probes a model answers correctly, and T/I are the candidate partial-retrieval and ignorance conditions (Section 5), tested with behavioral

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.

probes, hidden-state analysis, and causal interventions. The central claim is narrow: *retrieval failure is not binary but potentially multidimensional*. Our contributions are:

1. **A new evaluation problem.** We formulate **partial retrieval** as a distinct, falsifiable state between successful recall and genuine ignorance (Section I), rather than collapsing all failures into one undifferentiated category.
2. **The Partial Retrieval Score (PRS).** A framework for measuring residual attribute accessibility after retrieval failure, with a distribution-matched ignorance control, after showing the more obvious synthetic-control baseline is confounded (Sections 7, 9.3).
3. **An empirical finding.** Across three models and an initial plus enlarged sample, candidate partial-retrieval failures show a small but robust aggregate accessibility increase over matched low-knowledge failures (0.539 vs. 0.507), replicating by three statistical tests and surviving restriction to label-stable trials (Sections 9, 10.1).
4. **An important negative result.** Neither attribute-level lead survives independent follow-up: geographic-region fails pre-registered replication, and character-length fails a larger re-draw, a forced-choice probe, and hidden-state probing at any layer or model (Section 10) – alongside a finding that identical items do not always get the same classification on rerun. Single-pilot attribute- and internal-state claims, including our own, need this scrutiny.
5. **A causal finding.** A cue-intervention experiment (Section 11) shows a target-consistent cue improves retrieval beyond an irrelevant one in both groups, but not specifically for partial-retrieval trials: the T-vs-I interaction the "almost remembering" hypothesis predicts is not significant ($p = 0.998$), so cue content appears causally useful once supplied, not evidence the information was already present internally.

2 Relation to Prior Work

2.1 Hallucination Detection

Work on hallucination and calibrated honesty (Lin et al., 2024) generally asks whether a generated answer is correct, unsupported, or hallucinated, operating at the level of the final generation. Our question is different – $P(\text{target-related information remains accessible} \mid \text{retrieval failure})$ rather than $P(\text{answer is hallucinated})$ – so hallucination detection is a possible downstream application of our framework rather than its primary object.

2.2 Latent Knowledge and Eliciting Hidden Knowledge

Work showing that models retain information in internal representations even when surface behavior does not express it (Burns et al., 2022) establishes the distinction between internal knowledge and generated output. We build on this distinction but ask a finer-grained question: not whether the correct answer can be recovered from hidden states, but which *properties* of an inaccessible answer remain available – since answer identity is only one component of knowledge, and a model might lose access to exact identity while retaining semantic category, phonological, orthographic, temporal, or contextual information. Our hidden-state check (Section 10.3) uses the same linear-probing family as Burns et al. (2022); more recent work elicits latent knowledge with mechanistic-interpretability techniques that can surface non-linear signal a linear probe would miss (Cywiński et al., 2025), which we did not deploy here (Section 13).

2.3 Tip-of-the-Tongue States

The human TOT phenomenon, classically described by Brown and McNeill (1966), is defined by the coexistence of failed target retrieval with partial accessibility to information about the target. We use TOT as a cognitive motivation rather than a claim about subjective experience: our goal is to determine whether an analogous computational pattern – target inaccessible together with target-related partial information still accessible – can be identified in LLMs, without assuming language models possess human-like conscious states.

TOT has also been formalized computationally: the TREC ToT track (Arguello et al., 2024), built on the Reddit *tip of my tongue* community (Bhargav et al., 2022), and the BLUR benchmark (CH-Wang et al., 2025) evaluate systems that retrieve a known item from a searcher's own fuzzy description of it. That description is external input to a retrieval system; we instead ask whether an LLM's internal

state retains partial information about a target *it itself* failed to retrieve, with no description supplied. The two questions are complementary, and we are not aware of prior work connecting them.

2.4 Retrieval Bottlenecks and Distributed Factual Representations

Separately, recent work argues LLM factual errors often reflect a *recall* bottleneck rather than missing knowledge (Calderon et al., 2026), and that retrieving an attribute from an entity representation is redundant and distributed across non-contiguous paths rather than localized to one circuit (Hochman et al., 2026). Both offer a candidate mechanistic account of our own repeated failure to localize the aggregate PRS effect to any single attribute or hidden-state signature (Sections 9.3, 10): a distributed representation is exactly what a per-attribute or single-layer linear probe would miss even if present. We do not test this directly – it needs circuit-level tools outside this pilot's scope – but flag it as more likely than the absence of any signal.

3 Conceptual Framework

3.1 Three Retrieval States

For a question x with target answer y: **successful retrieval** (R) produces the correct target ($\hat{y} = y$); **partial retrieval** (T) fails to produce the exact target but retains above-chance information about properties of y; **genuine ignorance** (I) fails to produce the target and has little or no target-related information available through the same probes. The distinction is illustrated in Table 1 in idealized terms; the pipeline cannot observe these thresholds directly and instead uses the operational proxy $K(x)$ (Section 5) to assign each item to a candidate condition – $T_{\text{candidate}}$ and $I_{\text{low_k}}$ are proxies for, not verified instances of, T and I, which is why later attribute-level and internal-state findings get extra scrutiny (Sections 9.3, 10).

Table 1: Conceptual distinction among retrieval states.

	Succ.	Partial	Ignor.
Exact answer	High	Low	Low
Sem. category	High	High	Low/med
First letter	High	Above ch.	Chance
Answer length	High	Above ch.	Chance
Context attrs.	High	Above ch.	Chance
Cue-assisted	High	Improved	Little

3.2 Partial Information

Let Z denote a vector of target attributes: semantic category, domain, occupation, geographic region, first-letter range, character-length bucket, and word count (7 attributes in Sec 9). A partial-retrieval state should satisfy $P(Z_k = z_k \mid T) > P(Z_k = z_k \mid I)$ for at least some attributes Z_k.

4 Dataset

4.1 Design Principles

We built a semi-automatically constructed pilot benchmark rather than a fully manually authored dataset. Every question has an answer with machine-readable attributes, drawn from structured knowledge sources (Wikidata relations) rather than from an LLM, which was used only to generate linguistic variations of the question text, not as a source of factual ground truth.

For every question, the benchmark stores $D_i = (x_i, y_i, Z_i)$, where x_i is the question, y_i is the answer, and Z_i contains validated answer attributes. The construction pipeline (entity and fact retrieval, LLM-generated question variations, attribute validation, difficulty stratification, and a small manual audit) is given in Appendix C.

4.2 Natural Knowledge Set

The pilot's natural knowledge set was drawn from eight Wikidata relations spanning seven domains (art, film, geography, history, literature, organizations, technology). An initial 432-item draw was used for the first pilot round; after that round found only a marginal, underpowered effect (Section 9), we doubled the draw to 857 items to properly power a pre-registered confirmatory follow-up (Appendix G). Scaling this natural set toward the 10^3–10^4 range, across more domains and relations, is future work rather than something the current pilot attempts, as is evaluating the PRS framework against naturalistic, human-elicited known-item TOT benchmarks (BLUR (CH-Wang et al., 2025), Reddit-TOMT (Bhargav et al., 2022), the TREC ToT track (Arguello et al., 2024)) rather than only this pipeline's Wikidata-derived questions.

4.3 Controlled Knowledge Set

We also constructed a controlled set of 80 synthetic but internally consistent items (Appendix C gives an example item) that provide complete control over answer attributes and eliminate ambiguity in the ground-truth metadata. This set was intended as an ignorance control ($I_{\text{synthetic}}$), but Section 9.3 shows it is confounded by a base-rate artifact, which is why we rely primarily on the natural, distribution-matched $I_{\text{low_k}}$ control instead.

5 Knowledge and Ignorance Conditions

The central methodological challenge is distinguishing difficult retrieval from genuine ignorance. We estimated model-specific knowledge using cross-model agreement among the three evaluated models. For question x, define

$$K(x) = \frac{1}{M} \sum_{m=1}^{M} \mathbf{1}[\hat{y}_m = y]. \tag{2}$$

With $M = 3$ evaluated models, $K(x)$ takes only four discrete values 0, 1/3, 2/3, 1: 1) high-knowledge, $K(x) = 1$ (all three correct); 2) intermediate-knowledge, $K(x) \in \{1/3, 2/3\}$ (one or two of three correct) – contributing the candidate partial-retrieval trials ($T_{\text{candidate}}$); and 3) low-knowledge, $K(x) = 0$ (none correct) – contributing the fair, distribution-matched ignorance control ($I_{\text{low_k}}$).

$K(x)$ is a population-level item-difficulty index, not a direct measurement of any single failing model's internal state; the closest we come to testing the stronger, intra-model version directly is the hidden-state probe of Section 10.3, which finds no corresponding signal. We do not treat cross-model disagreement alone as proof of ignorance: it is combined with the synthetic controlled set and, for the attribute-level follow-ups in Section 10, with a forced-choice specificity probe and hidden-state analysis. We also use *distribution-matched* in a limited sense: $T_{\text{candidate}}$ and $I_{\text{low_k}}$ share the same item population and pipeline, avoiding the synthetic control's base-rate artifact (Section 9.3), not that they are resampled to match per-attribute marginals – they can still differ in composition, which Appendix B checks with a chance-corrected version of the primary comparison.

6 Experimental Protocol

6.1 Stage 1: Retrieval Attempt

For each question, the model was first asked to answer normally, and the response was classified by a deterministic, rule-based procedure (no LLM-judge or human annotation) into *correct* (alias-aware normalized string match against the gold answer), *uncertain* (a fixed regex over hedging phrases like "I don't know" when no match was found), or *incorrect* otherwise – a simplification of the finer five-way taxonomy (correct / related-but-wrong / unrelated-and-wrong / uncertainty / abstention) the conceptual framing could support, since separating related- from unrelated-wrong responses needs semantic judgment this rule-based pipeline does not attempt. The distinction does not affect Stage 2: *uncertain* and *incorrect* are pooled identically as retrieval failures, and it is only this pooled failure set that $K(x)$ stratification (above) splits into $T_{\text{candidate}}$ and $I_{\text{low_k}}$.

6.2 Stage 2: Partial-Cue Interrogation

For failed retrieval trials, the model was asked independent multiple-choice questions about the target without being asked to reproduce it (e.g., "is the answer a person, place, organization, object, or event?"; example wording in Appendix C). The seven attributes probed were semantic category, domain, occupation, geographic region, first-letter range, character-length bucket, and word count (Section 9), posed with randomized option order to hold chance level constant and known ($1/|\text{options}|$) across trials.

6.3 Independence of Probes

The original answer was never revealed before the partial-cue questions were administered – Stage 2 is only ever run on Stage 1 failures. Each Wikidata relation had multiple natural-language question templates, used to generate linguistic variation in the underlying question set (Appendix C).

7 Partial Retrieval Score

For question i with A_i *applicable* target attributes, define

$$PRS_i = \frac{1}{A_i} \sum_{k=1}^{A_i} \mathbf{1}[\hat{z}_{ik} = z_{ik}]. \tag{3}$$

We write A_i rather than a fixed constant (reserving $K(x)$, Section 5, for the cross-model knowledge function) because not every attribute applies to every item: *occupation* is defined only when *category*= *person*. The Partial Retrieval Score ranges from zero to one; for example, a trial correctly scored on category, first-letter range, character length, word count, and geographic region but not occupation has $A_i = 6$ and $PRS = 5/6$ (Appendix C tabulates this worked example). The primary prediction is $\mathbb{E}[PRS \mid T] > \mathbb{E}[PRS \mid I]$; we additionally analyze each attribute family separately.

Each attribute's chance level, $1/|\text{options}_k|$, also varies (from $1/4$ to $1/7$; Appendix C), so an unweighted average's expected value under pure guessing depends on which attributes apply to a given item – if $T_{\text{candidate}}$ and $I_{\text{low_k}}$ differ in this composition, the raw PRS comparison could partly reflect that rather than partial-retrieval information. Appendix B checks this directly.

8 Initial Minimal Viable Experiment

Before running the expanded pilot, we ran a small minimal viable experiment (432 questions, 3 open models, 7 domains, 7 partial cues, PRS as the main statistic, retrieval vs. ignorance as the primary comparison) to test whether the aggregate effect was worth pursuing at all. It was designed to answer one question:

> **Does failed retrieval contain substantially more target-specific partial information than genuine ignorance?**

Its result – a positive but marginal, single-attribute-driven effect against the fair control – is what motivated the properly powered, pre-registered expansion described next, rather than reconsidering the project.

9 Pilot Results

We ran the minimal viable experiment described above, then – after the initial 432-item pilot below found only a marginal fair-control effect driven by a single, underpowered attribute – expanded the natural set 2x and reran the full pipeline to properly power a pre-registered follow-up (Appendix G details the power analysis). The natural knowledge set contained 857 questions drawn from eight Wikidata relations spanning seven domains (art, film, geography, history, literature, organizations, technology); the controlled knowledge set contained 80 synthetic ignorance-control items. Each item carried seven target attributes: semantic category, domain, occupation, geographic region, first-letter range, character-length bucket, and word count. Three open-weight instruction-tuned models were evaluated: Mistral-7B-Instruct-v0.2, Qwen2.5-7B-Instruct, and Qwen2.5-14B-Instruct (4-bit quantized). Stage 1 retrieval failures were stratified by cross-model knowledge $K(x)$ (Section 5) into $T_{\text{candidate}}$ (intermediate K, natural), $I_{\text{low_k}}$ (low K, natural), and $I_{\text{synthetic}}$ (synthetic items), yielding 375, 687, and 240 failed trials respectively, each probed with the Stage 2 partial-cue protocol.

9.1 Primary Comparison

Consistent with the central prediction of Section 7, mean PRS was higher for candidate partial-retrieval trials than for synthetic ignorance controls: $\mathbb{E}[PRS \mid T_{\text{candidate}}] = 0.539$ (95% CI $[0.523, 0.556]$, $n = 375$) $> \mathbb{E}[PRS \mid I_{\text{synthetic}}] = 0.445$ (95% CI $[0.426, 0.464]$, $n = 240$), one-sided Mann–Whitney U test, $p = 3.0 \times 10^{-12}$; a cluster-robust logistic regression (clustering by question, approximating the mixed-effects model of Section D) gave a positive condition coefficient of 0.438 ($p = 1.4 \times 10^{-13}$).

We flag one confound up front: by $K(x)$'s construction (Section 5), $T_{\text{candidate}}$ is items at least one model retrieves versus $I_{\text{low_k}}$'s items none retrieve, so $T_{\text{candidate}}$ skews toward more globally prominent entities – some of the gap could reflect generic attributes being easier to guess for prominent entities, not target-specific residue. This is concrete: a domain-composition imbalance and a domain-adjusted regression shrink but do not eliminate the effect (Appendix F; the coefficients above adjust for model but not domain), though it survives a second, independent check (chance-corrected PRS, Appendix B); Section 13 returns to this.

Table 2: Per-attribute $P(\text{correct})$, expanded pilot data ($T_{\text{candidate}}$ vs. the distribution-matched $I_{\text{low_k}}$ control), with one-sided Fisher exact test p-values. Only *length_bucket* survives a Bonferroni correction for seven comparisons ($\alpha = 0.05/7 \approx 0.00714$); *geographic_region* was the nominally significant attribute in the initial 432-item pilot and did not replicate here.

Attribute	$P(\text{correct} \mid T)$	n_T	$P(\text{correct} \mid I_{\text{low_k}})$	n_I	p
category	0.896	375	0.882	687	0.28
domain	0.765	375	0.760	687	0.45
occupation	0.957	232	0.957	351	0.60
geographic_region	0.766	312	0.724	558	0.10
first_letter_range	0.128	375	0.140	687	0.73
length_bucket	0.240	375	0.166	687	**0.0024**
word_count	0.256	375	0.231	687	0.21

Unlike in the initial 432-item pilot, the comparison against the natural low-knowledge ignorance condition is now also significant: 0.539 ($n = 375$) $> \mathbb{E}[PRS \mid I_{\text{low_k}}] = 0.507$ ($n = 687$), $p = 1.2 \times 10^{-3}$. The same regression with $I_{\text{low_k}}$ gives a smaller but significant coefficient of 0.174 ($p = 3.2 \times 10^{-4}$); adding attribute fixed effects strengthens this to 0.248 ($p = 8.1 \times 10^{-4}$). The aggregate effect thus survives the fair, distribution-matched control by three independent tests – though, as detailed next, the specific attribute-level story did not replicate.

9.2 Is the Effect Concentrated in a Subset of Items?

A 3.2-point aggregate gap could come from a small subset of $T_{\text{candidate}}$ items with unusually high PRS, rather than a broad shift. Appendix A checks the full per-item PRS distributions and finds the gap is small but genuinely diffuse: the distributions differ in shape as well as location (two-sample Kolmogorov–Smirnov $p = 9.9 \times 10^{-3}$), median PRS is 0.571 vs. 0.500, and the standardized effect size is small (Cohen's $d = 0.19$) with only a modestly higher share of $T_{\text{candidate}}$ items in the top decile (4.5% vs. 2.8%) – nowhere near enough to account for the gap alone, consistent with our repeated failure to localize the effect to any single attribute, item subgroup, or internal-state signature elsewhere in this paper.

9.3 Per-Attribute Breakdown

Appendix E (Figure 2 and Table 5) reports $P(\text{correct})$ for each attribute under $T_{\text{candidate}}$ and $I_{\text{synthetic}}$: *geographic_region* shows the largest gap (0.766 vs. 0.183), *category*/*domain*/*occupation* stay high under both, and *first_letter_range*/*word_count* show little difference. This is confounded by a base-rate artifact, though: the synthetic control assigns *geographic_region* uniformly at random over six regions, so $I_{\text{synthetic}}$ scores near the resulting $1/6$ chance level (0.156), whereas a model guessing from the (non-uniform) natural marginal distribution – with no residual information at all – would already score above it; the synthetic control is therefore not distribution-matched for this attribute, and less severely for the others.

Table 2 repeats the comparison against $I_{\text{low_k}}$ on the expanded, properly powered sample (Appendix G). *geographic_region* – nominally significant before ($p = 0.027$) and the specific target this run was powered to confirm – *did not replicate*: its gap shrank from $+0.092$ to $+0.042$ ($p = 0.101$). Instead *length_bucket*, which showed essentially no effect initially ($p = 0.61$), is now significant and survives Bonferroni correction ($p = 0.0024 < 0.05/7 \approx 0.00714$); *category*, *domain*, *occupation*, and *word_count* remain non-significant.

We treat this as cautionary, not a new finding: *geographic_region*'s non-replication suggests the original signal was small-sample noise, and we do *not* elevate *length_bucket* to confirmed either, since it emerged post-hoc and would need its own pre-registered replication – treating whichever attribute is nominally significant as "real" is exactly the multiple-comparisons trap this section avoids. What *does* hold up is the *aggregate* PRS effect (Section 9.1): summed across all seven attributes, partial-retrieval trials carry reliably more residual information, even though no single attribute identifies which one.

9.4 Per-Model and Per-Domain Breakdown

Full per-model and per-domain tables, against both controls, are in Appendices E and F, addressing the falsifiability criterion (Section I) that the effect not be limited to a single model, domain, or prompt. Against the (confounded) synthetic control, the effect held for all three models and six of seven domains. Against the fair $I_{\text{low_k}}$ control, on the larger sample, all three models show a positive

gap and five of six comparable domains are directionally positive (only *technology* reverses) – an improvement over the initial pilot, where three of six domains reversed. In short: the aggregate effect got stronger and more consistent under proper power, while the attribute-level story did not survive – motivating the follow-up checks in Section 10.

10 Target-Specificity and Internal-State Checks

Section 9.3 already treated *length_bucket* as an unconfirmed, post-hoc lead rather than a finding. We subsequently ran three further, more stringent checks against it, plus a fourth check of the reproducibility of the underlying pipeline. All four are negative.

10.1 Independent Replication and Run-to-Run Stability

The causal-cue follow-up (Section 11) required a larger natural-item pool built the same way (Section 9.3) at 2.5x the per-relation count (2,163 vs. 857 items, 853 shared and 1,310 newly drawn). Re-running the pipeline reverses the *length_bucket* effect's sign (0.237 vs. 0.243, $n = 716/1{,}072$, $p = 0.60$, against 0.240 vs. 0.166, $p = 0.0024$ originally); it does not reproduce restricted to just the 853 shared items ($p = 0.29$) or the 1,310 new items ($p = 0.77$) either. We traced this to the pipeline itself: condition labels for the same (item, model) pairs agree only 89.5% of the time ($864/965$) across the two runs despite greedy decoding, likely because batch composition and 4-bit quantization introduce enough numerical variation to flip outcomes near a model's decision boundary – a modest effect (the weakest of the seven attributes to reach significance) is exactly what such relabeling noise can erase or invert.

Does the *aggregate* PRS effect depend on this $\sim 10.5\%$ of unstable-label trials? Restricted to the 864 stable pairs ($n_T = 258$, $n_I = 606$), $\mathbb{E}[PRS \mid T]$ vs. $\mathbb{E}[PRS \mid I_{\text{low_k}}]$ gives 0.541 vs. 0.508 ($p = 3.5 \times 10^{-3}$), essentially unchanged from the full sample – not an artifact of label instability, even though *length_bucket* specifically is.

Does the aggregate claim replicate on just the 1,310 items drawn independently of the original 857-item pilot? Splitting the enlarged pool's trials by provenance, the point estimate replicates in both halves – 0.516 vs. 0.499 ($n_T = 264$, $n_I = 462$) on the shared items, 0.543 vs. 0.530 ($n_T = 452$, $n_I = 610$) on the new-only items – but neither half reaches significance alone ($p = 0.065$, $p = 0.130$; a cluster-robust logistic regression on the new-only items gives coefficient 0.061, $p = 0.146$), against $p = 0.017$ on the full pool. A ∼1,788-trial sample is already only modestly powered for an effect this small (Appendix G), so this halving is consistent with, not evidence against, the aggregate claim – we report it as the more direct test of overfitting to the original draw.

A preliminary, underpowered check restricting to *prior-inconsistent* items only (Appendix H) was suggestive but inconclusive, which motivated the sharper test below.

10.2 Forced-Choice Target-Specificity Probe

For items where the true *length_bucket* differs from its relation's modal bucket, we built a two-alternative forced-choice probe between exactly those two values (chance = 0.5 by construction), run across all three models on 1,182 eligible trials from the causal-cue pool ($n = 435$ $T_{\text{candidate}}$, $n = 747$ $I_{\text{low_k}}$). If the residual signal tracks the specific target rather than the relation's generic prior, $T_{\text{candidate}}$ should choose the true value above chance. It does not: 47.1% correct ($p = 0.89$, wrong direction) vs. $I_{\text{low_k}}$'s 50.2%, and $T_{\text{candidate}}$ is not greater than $I_{\text{low_k}}$ ($p = 0.85$). This is the cleanest test in the paper for target specificity, and it is a clean null.

10.3 Hidden-State Probing Before Any Cue

The behavioral probes above only show that *length_bucket* is decodable when asked about directly – not that it was present in the state that produced the original failed answer. We therefore probe the hidden state directly: for every failed trial, we capture the last-token hidden state of the original, uncued Stage-1 question at four depths (25/50/75/100% of each model's layers), and train a linear probe to predict *length_bucket* from it, with leave-one-relation-out cross-validation so a probe cannot exploit a question template's typical signature instead of target-specific information.

Across all three models and all four depths, $T_{\text{candidate}}$ probe accuracy is never significantly greater than $I_{\text{low_k}}$ (one-sided Mann–Whitney p ranges from 0.06 to 0.98 across the twelve model×layer combinations, none below 0.05), and both conditions sit at or below the majority-class baseline

Table 3: Causal cue intervention: $P(\text{correct})$ by condition and group, paired trials (same item, all four conditions). Cells are independently rounded to three decimals, so a text delta can differ from cells' rounded difference by up to 0.001 (text deltas use unrounded per-trial rates).

Group	no_cue	correct_cue	incorrect_cue	irrelevant_cue
$T_{\text{candidate}}$ ($n = 918$)	0.004	0.181	0.155	0.100
$I_{\text{low_k}}$ ($n = 1608$)	0.000	0.070	0.054	0.030

(≈ 0.48–0.50): the probe is not reliably extracting *length_bucket* information from the pre-cue hidden state at all.

Two design choices narrow how far this null generalizes, beyond the linear-vs-distributed distinction in Section 2.4: we probe the entire prompt's last token, not the *subject entity*'s, where causal-tracing work often localizes factual attributes instead; and all three models are 4-bit quantized, and Section 10.1's 89.5% label agreement is direct evidence of enough numerical noise to suppress a subtle signal. We did not run a full-precision or subject-position probe to rule either out, so this null is evidence against a signal decodable by *this specific* design, not against internal information under any design.

None of these four checks supports treating *length_bucket* as a real, target-specific, internally-present effect, and combined with *geographic_region*'s earlier non-replication, both attribute-level leads this project has produced have now failed independent follow-up. We therefore narrow the paper's central empirical claim to the *aggregate* PRS effect alone (Section 9.1): more decodable residual information in aggregate, with no attribute-level, target-specific, or internal-state evidence for *which* information – a caveat the causal cue-intervention results below should be read with.

11 Causal Cue Intervention

The results above show that failed-retrieval trials carry more decodable target-specific information than ignorance trials – a claim about what an external probe can recover, not whether that information is *functionally* useful to the model. We therefore ran a causal follow-up, re-attempting retrieval after providing an explicit cue, on both natural failed-trial groups ($T_{\text{candidate}}$, $n = 918$; $I_{\text{low_k}}$, $n = 1608$; the synthetic set is excluded, having no real-world length distribution to draw an incorrect cue from), across all three models.

For each failed trial, we re-ask the question with one appended cue sentence, under four conditions: `no_cue` (a matched baseline, same wrapper but no cue text); `correct_cue`, revealing the target's true *length_bucket* – the one attribute that survived Bonferroni correction in the behavioral pilot; `incorrect_cue`, a false length bucket of the same type; and `irrelevant_cue`, a neutral filler carrying no target-specific information. Correct cues must be compared against both incorrect and irrelevant cues, not only `no_cue`: a bare $\Delta P_{\text{retrieval}}$ against `no_cue` could just reflect the model becoming more willing to answer after *any* additional text, not selective influence from target-consistent content.

We flag a structural limitation up front: this experiment was built around *length_bucket* because it was, at the time, the one attribute that survived Bonferroni correction (Section 9.3) – but Section 10's later checks showed that same attribute carries no detectable target-specific signal. A cue-content interaction built on an attribute the model does not reliably retain is a weak test of the "almost remembering" hypothesis: a null interaction is close to the expected outcome given that, not necessarily evidence against the hypothesis in general. We had no attribute to substitute – it was the only one with any nominally significant signal to begin with – so we report the result below with this caveat.

11.1 Results

Table 3 gives three results, in decreasing order of how cleanly they support the hypothesis.

First, *every* cue condition – including the irrelevant filler – substantially increases retrieval over `no_cue` (e.g. $T_{\text{candidate}}$: $+0.176$ for `correct_cue`, $+0.096$ for `irrelevant_cue`, one-sided McNemar $p = 1.5 \times 10^{-46}$ and $p = 4.4 \times 10^{-23}$; same pattern for $I_{\text{low_k}}$), almost certainly by perturbing the prompt past whatever caused the original greedy-decoding failure, independent of content – exactly why a bare $\Delta P_{\text{retrieval}}$ against `no_cue` is uninformative on its own, and the incorrect/irrelevant controls are necessary.

Second, once that "second-attempt" effect is controlled for by comparing against `irrelevant_cue`, the correct cue still wins: $+0.081$ for $T_{\text{candidate}}$ ($p = 9.9 \times 10^{-10}$) and $+0.040$ for $I_{\text{low_k}}$ ($p = 2.6 \times 10^{-9}$) – evidence that target-consistent cue content is functionally useful, not merely decodable, though *both* groups show it (evidence about cue utility in general, not partial retrieval specifically; see the interaction below).

Third, correct cues also beat incorrect (wrong-value) cues of the same type in both groups ($p = 0.026$/$p = 0.0057$, nominal; undetectable in the smaller single-model pilot), suggesting modest sensitivity to the cue's true value. The interaction we most wanted to see – a larger correct-cue boost for $T_{\text{candidate}}$ than $I_{\text{low_k}}$, as the "almost remembering" hypothesis predicts – remains not significant despite a numerically larger raw boost ($+0.176$ vs. $+0.070$; $p = 0.998$). This p-value comes from the `is_cue×is_T` coefficient of a GEE logistic regression – the log-odds scale, not the raw probability differences quoted above – and the two are not contradictory: with the near-zero $I_{\text{low_k}}$ `no_cue` baseline (0.000, Table 3), a move to 0.070 is a large log-odds change, which is exactly what drives the interaction coefficient toward null despite the raw gap favoring $T_{\text{candidate}}$. We report the log-odds test because it is the correct scale for the GEE model (Section D), but flag the mismatch with the raw framing used elsewhere to avoid it reading as an inconsistency.

One caveat is important enough to restate: this experiment shows that *supplying* the true length bucket causally helps recovery more than an irrelevant or incorrect cue, in both groups. It does not show the information was already present internally – the pre-cue hidden-state probe (Section 10.3) found no such signal for either group. Only the weaker, "causally useful once supplied" reading is supported here.

12 Discussion

The central claim of this work is that retrieval failure need not be binary: a model can fail to output an answer while retaining structured residual information about it, unlike evaluations that collapse all incorrect answers into one category. Only the *aggregate* claim is well-tested (Sections 9, 10); the stronger claim of *identifiable structure* remains open, and we cannot distinguish two explanations for the gap: genuine residue too thinly distributed for any probe to isolate, or $K(x)$-stratification not fully separating it from a subtler generic-prior confound (Section 9.3). The diffuse gap (Section 9.2) is compatible with either story and remains open future work. The human-TOT analogy should also be read cautiously – at most a computational pattern shares functional properties with human partial retrieval, not subjective phenomenology (Appendix J discusses a further, untested hallucination implication).

13 Limitations

Several limitations apply: partial-cue responses do not establish human-like TOT phenomenology; genuine ignorance is hard to establish for a pretrained model, so we use operational criteria, not perfect labels; partial attributes may correlate with the question, so leakage controls are essential; we test only three 7–14B, 4-bit-quantized models, with no frontier-scale analysis and no prompt-template sensitivity check (Appendix D); the forced-choice and hidden-state checks (Section 10) covered only *length_bucket*, and the latter used only last-token linear probing, ruling out a linearly decodable signal but not a distributed one; the 89.5% label agreement (Section 10.1) was not tested for sampling or precision sensitivity; and $T_{\text{candidate}}$ differs from $I_{\text{low_k}}$ in composition toward more prominent entities and *person*-category items (Appendices F, B) – so although the effect survives chance-corrected and domain-adjusted reanalyses, composition may still inflate the raw gap.

14 Conclusion

We studied language-model retrieval failure through the lens of partial retrieval rather than binary correctness (Appendix K): such information survives failure in aggregate, is not attributable to any single attribute or present internally before a cue, and cue content is useful once supplied – but equally for both groups (Section 11), not selectively for partial retrieval. Every attempt to localize the effect to a specific or internally-present signal came back negative: single-pilot attribute- and representation-level claims about LLM internals are easy to overclaim, and our own two candidate leads both failed independent follow-up. Which information survives failed retrieval, and whether it is represented internally beforehand, remains open for future work with more power and more direct mechanistic tools.

AI USE STATEMENT

Large language models play two distinct roles in this work, which we distinguish explicitly. First, LLMs are the *object of study*: the three open-weight models evaluated throughout the paper (Mistral-7B-Instruct-v0.2, Qwen2.5-7B-Instruct, Qwen2.5-14B-Instruct) are the systems whose partial-retrieval behavior we measure, not tools used to produce the paper. Second, an LLM was used as a *construction and writing tool* in two places. (a) The dataset construction pipeline (Section 4.2, Appendix C) used an LLM solely to generate linguistic variations of question text drawn from structured Wikidata facts; the LLM was never a source of factual ground truth, and every attribute value was validated against the structured source rather than against LLM output. (b) An AI writing assistant was used during manuscript preparation and revision – drafting and editing prose in response to reviewer feedback, re-deriving and cross-checking reported statistics against the underlying analysis code and pilot data, and copy-editing for length and consistency. All statistical claims, findings, and interpretive judgments in this paper were verified by the authors against the underlying code and data before inclusion, and the authors take full responsibility for the content of the paper.

REFERENCES

Arguello, J., Bhargav, S., Diaz, F., Kim, T. E., He, Y., Kanoulas, E., and Mitra, B. Overview of the TREC 2024 tip-of-the-tongue track. In *Proceedings of the 33rd Text REtrieval Conference (TREC 2024)*, 2024.

Bhargav, S., Sidiropoulos, G., and Kanoulas, E. "It's on the tip of my tongue": A new dataset for known-item retrieval. In *Proceedings of the 15th ACM International Conference on Web Search and Data Mining (WSDM)*, 2022.

Brown, R. and McNeill, D. The "tip of the tongue" phenomenon. *Journal of Verbal Learning and Verbal Behavior*, 5(4):325–337, 1966.

Burns, C., Ye, H., Klein, D., and Steinhardt, J. Discovering latent knowledge in language models without supervision. arXiv preprint, 2022.

Calderon, N., Ben-David, E., Gekhman, Z., Ofek, E., and Yona, G. Empty shelves or lost keys? Recall is the bottleneck for parametric factuality. arXiv preprint, 2026.

CH-Wang, S., Deshpande, D., Muresan, S., Kannappan, A., and Qian, R. Browsing lost unformed recollections: A benchmark for tip-of-the-tongue search and reasoning. arXiv preprint, 2025.

Cywiński, B., Ryd, E., Rajamanoharan, S., and Nanda, N. Towards eliciting latent knowledge from LLMs with mechanistic interpretability. arXiv preprint, 2025.

Hochman, H., Shapira, N., and Goldberg, Y. Factual retrieval in LLMs is a redundant, distributed and non-contiguous process. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2026.

Lin, S., Hilton, J., and Evans, O. Truthful and honest language models. arXiv preprint, 2024.

A ITEM-LEVEL PRS DISTRIBUTION

Section 9.2 summarizes whether the aggregate PRS gap (0.539 vs. 0.507) is driven by a small subset of items or is broadly distributed. We compared the full per-item PRS distributions (each item's PRS is the mean correctness across its seven attribute probes) for $T_{\text{candidate}}$ ($n = 375$) and $I_{\text{low_k}}$ ($n = 687$) in the expanded pilot.

The distributions differ in shape as well as location (two-sample Kolmogorov–Smirnov $D = 0.104$, $p = 9.9 \times 10^{-3}$): median PRS is 0.571 for $T_{\text{candidate}}$ versus 0.500 for $I_{\text{low_k}}$, and the standardized effect size is small (Cohen's $d = 0.19$). The fraction of items scoring in the top decile ($PRS \geq 0.85$) is only modestly higher for $T_{\text{candidate}}$ (4.5%) than for $I_{\text{low_k}}$ (2.8%), and the fraction scoring at the bottom ($PRS \leq 0.15$) is actually slightly *lower* for $T_{\text{candidate}}$ (0.3% vs. 0.6%) – neither tail is doing the work of the aggregate gap. The common-language effect size, $P(\text{random } T_{\text{candidate}} \text{ item scores higher than random } I_{\text{low_k}} \text{ item}) \approx 0.556$ (estimated from

the Mann–Whitney U statistic), is only modestly above the chance value of 0.5. Taken together, this is consistent with a small, broadly distributed shift across nearly the whole item population rather than a strong signal concentrated in a distinguishable subset of items, which matches our repeated failure to localize the effect to any single attribute, item subgroup, or internal-state signature elsewhere in this paper (Section 10).

B PRS Attribute-Composition and Chance-Level Check

Section 7 flags two related concerns with the unweighted PRS average: attribute applicability varies per item (A_i), and chance level varies per attribute ($1/4$ for *word_count* to $1/7$ for *domain*), so a composition difference between $T_{\text{candidate}}$ and $I_{\text{low_k}}$ could confound the raw comparison even without any partial-retrieval effect. We check both directly on the expanded pilot's primary comparison (Section 9.1, $n_T = 375$, $n_I = 687$).

Attribute composition. Of the seven attributes, five (*category*, *domain*, *first_letter_range*, *length_bucket*, *word_count*) apply to every item in both groups. The remaining two are optional: *geographic_region* applies to 83.2% of $T_{\text{candidate}}$ items ($312/375$) versus 81.2% of $I_{\text{low_k}}$ items ($558/687$), not a significant difference (two-sided Fisher exact $p = 0.45$); *occupation* applies to 61.9% of $T_{\text{candidate}}$ items ($232/375$) versus 51.1% of $I_{\text{low_k}}$ items ($351/687$), a significant difference ($p = 7.9 \times 10^{-4}$) – $T_{\text{candidate}}$ is composed of disproportionately more *person*-category items. Since both groups score almost identically on *occupation* itself (0.957 vs. 0.957, Table 2) – well above its $1/5$ chance level for both – this composition difference could inflate $T_{\text{candidate}}$'s raw aggregate PRS slightly relative to $I_{\text{low_k}}$'s, independent of any partial-retrieval effect, simply because $T_{\text{candidate}}$ has more of an attribute both groups already answer well above chance.

Chance-corrected PRS. To test whether this (or any other composition difference) is doing the work of the primary result, we recompute the comparison using a chance-corrected per-probe score, $\mathbf{1}[\hat{z}_{ik} = z_{ik}] - 1/|\text{options}_k|$, averaged the same way as PRS_i – an item guessed uniformly at random contributes 0 in expectation to this score regardless of which attributes apply to it, removing the composition confound raised above. The result is essentially unchanged: $\mathbb{E}[PRS_{\text{adj}} \mid T_{\text{candidate}}] = 0.344$ vs. $\mathbb{E}[PRS_{\text{adj}} \mid I_{\text{low_k}}] = 0.312$ (one-sided Mann–Whitney $p = 1.4 \times 10^{-3}$), compared to 0.539 vs. 0.507 ($p = 1.2 \times 10^{-3}$) on the raw score. The aggregate PRS effect is therefore not an artifact of the two groups' differing attribute composition or of the varying chance levels across attributes, even though the *occupation* composition difference itself is real and should be read as a caveat on how much of the raw gap is attributable to partial-retrieval information specifically versus this composition effect.

C Dataset Construction Pipeline

The complete benchmark pipeline is designed to be approximately 80–90% automated, and proceeds as follows:

1. retrieve entities and structured facts from reliable sources;
2. automatically construct candidate questions;
3. use an LLM to generate linguistic variations of the questions;
4. map generated answers back to canonical entities;
5. validate all attributes against structured sources;
6. remove ambiguous or multiply-answerable questions;
7. evaluate candidate questions on several language models;
8. stratify questions by retrieval difficulty;
9. manually audit a small subset.

Structured knowledge sources → candidate entities and facts → LLM question generation → automatic factual validation → difficulty/model filtering → deduplication → small manual audit → final benchmark

Figure 1: Proposed semi-automatic dataset construction pipeline.

Manual annotation is reserved primarily for ambiguous cases, multiple valid answers, aliases, unusual entities, and the knowledge-boundary controls.

For example, a synthetic knowledge item for the controlled set (Section 4.3) may specify:

```
entity               = Elena Maric
category             = person
domain                = literature
occupation            = writer
geographic_region     = Europe
first_letter_range    = A–E
length_bucket         = 11–14
word_count            = 2
```

Similarly, for a Stage 2 probe (Section 6) on a failed trial whose target is "Ulaanbaatar," the model may be asked "is the answer a person, place, organization, object, or event?" or "which range contains the first letter of the answer: A–E, F–J, K–O, P–T, or U–Z?", without the target ever being revealed.

Table 4: Worked example of a partial-retrieval trial and its PRS (Section 7).

Attribute	Correct?
Semantic category	Yes
First-letter range	Yes
Character length	Yes
Word count	Yes
Geographic region	Yes
Occupation	No
Partial Retrieval Score	$5/6$

D Statistical Model Specification

The original, general-purpose statistical-analysis plan called for a full mixed-effects model; the executed pilot (Section 9) approximates it with a cluster-robust logistic regression, clustering by question in place of a random effect. We retain the original specification here for completeness.

For each attribute, we use mixed-effects logistic regression:

$$\text{logit}\, P(Y_{ik} = 1) = \beta_0 + \beta_1 T_i + \beta_2 M_i + u_{\text{question}(i)} + u_{\text{model}(i)}, \tag{4}$$

where Y_{ik} indicates whether probe k was answered correctly on trial i; T_i denotes trial i's retrieval condition; M_i denotes matched covariates for trial i – concretely, model identity (as in the executed cluster-robust approximation, Section 9.1) and, where checked, domain (Appendix F reports a domain-adjusted variant); and $u_{\text{question}(i)}$, $u_{\text{model}(i)}$ are random intercepts for the question and model that generated trial i, treated as random effects.

The primary hypothesis is $\beta_1 > 0$. We additionally report:

- bootstrap confidence intervals;
- effect sizes;
- per-model results;
- per-domain results;
- per-attribute results.

One planned check from this list was not executed: sensitivity to prompt templates. The dataset pipeline generates alternate phrasings per item (Appendix C), but the executed Stage 1 protocol (Section 6) asks each item's single canonical question only – the stored alternates were never rotated in – so no template-sensitivity analysis exists to report. We list this as a planned but unexecuted check rather than silently dropping it, and flag it as an open robustness question in Section 13.

E Pilot Results Against the Confounded Synthetic Control

For completeness, this section gives the naive $T_{\text{candidate}}$ vs. $I_{\text{synthetic}}$ comparisons underlying the headline numbers in Section 9, from the expanded (857-natural-item) pilot run, before the base-rate confound identified in Section 9.3 is accounted for. The distribution-matched $I_{\text{low_k}}$ comparisons in the main text are the ones we treat as evidential.

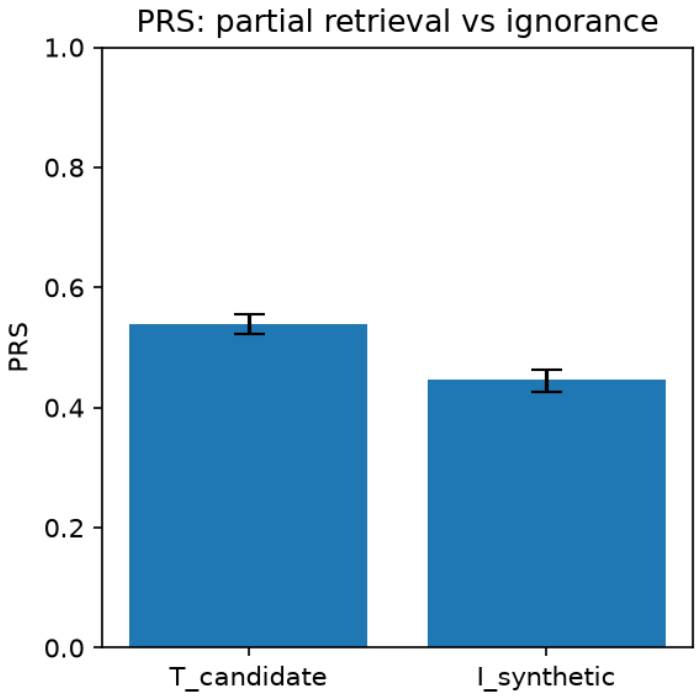


Figure 2: PRS by condition, expanded pilot data.

Table 5: Per-attribute $P(\text{correct})$, expanded pilot data ($T_{\text{candidate}}$ vs. $I_{\text{synthetic}}$).

Attribute	$P(\text{correct} \mid T)$	n_T	$P(\text{correct} \mid I_{\text{synthetic}})$	n_I
category	0.896	375	0.908	240
domain	0.765	375	0.775	240
occupation	0.957	232	0.942	120
geographic_region	0.766	312	0.183	180
first_letter_range	0.128	375	0.087	240
length_bucket	0.240	375	0.175	240
word_count	0.256	375	0.263	240

Table 6: Per-model mean PRS, expanded pilot data ($T_{\text{candidate}}$ vs. $I_{\text{synthetic}}$).

Model	mean PRS $\mid T$	n	mean PRS $\mid I_{\text{synthetic}}$	n
Mistral-7B-Instruct-v0.2	0.530	86	0.416	80
Qwen2.5-7B-Instruct	0.501	199	0.407	80
Qwen2.5-14B-Instruct	0.631	90	0.513	80

F Per-Model and Per-Domain Breakdown Against the Fair Control

Full per-model and per-domain tables for the expanded pilot's fair ($I_{\text{low_k}}$) control, summarized in Section 9.4.

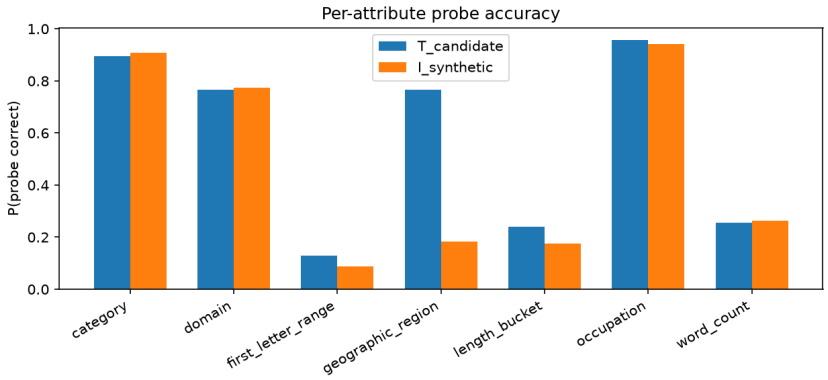


Figure 3: Per-attribute $P(\text{correct})$ by condition, expanded pilot data.

Table 7: Per-domain mean PRS, expanded pilot data ($T_{\text{candidate}}$ vs. $I_{\text{synthetic}}$).

Domain	mean PRS | T	n	mean PRS | $I_{\text{synthetic}}$	n
art	0.607	73	0.500	30
film	0.612	63	0.505	30
geography	0.554	13	0.433	30
history	0.431	32	0.393	30
literature	0.562	28	0.500	30
organizations	0.545	128	0.447	60
technology	0.335	38	0.339	30

Table 8: Per-model mean PRS, expanded pilot data ($T_{\text{candidate}}$ vs. the distribution-matched $I_{\text{low_k}}$ control).

Model	mean PRS | T	n	mean PRS | $I_{\text{low_k}}$	n
Mistral-7B-Instruct-v0.2	0.530	86	0.501	229
Qwen2.5-7B-Instruct	0.501	199	0.466	229
Qwen2.5-14B-Instruct	0.631	90	0.554	229

Table 9: Per-domain mean PRS, expanded pilot data ($T_{\text{candidate}}$ vs. the distribution-matched $I_{\text{low_k}}$ control). ↓ marks *technology*, the only domain here where the sign runs against the overall positive trend ($T_{\text{candidate}} < I_{\text{low_k}}$); this is a reversal relative to the other domains' pattern, not relative to Table 7's synthetic-control comparison, where *technology* is also (weakly) negative.

Domain	mean PRS | T	n	mean PRS | $I_{\text{low_k}}$	n
art	0.607	73	0.601	174
film	0.612	63	0.600	15
geography	0.554	13	n/a	0
history	0.431	32	0.409	45
literature	0.562	28	0.530	60
organizations	0.545	128	0.500	282
technology	0.335	38	0.392	111 ↓

Table 9 also surfaces a domain imbalance worth flagging directly in composition terms, not just counts: *geography* contributes 13 $T_{\text{candidate}}$ items (3.5% of $T_{\text{candidate}}$) but zero $I_{\text{low_k}}$ items (every geography item in the natural set was answered correctly by at least one of the three models); *film* makes up 16.8% of $T_{\text{candidate}}$ (63/375) but only 2.2% of $I_{\text{low_k}}$ (15/687); and conversely *technology* makes up 16.2% of $I_{\text{low_k}}$ (111/687) but only 10.1% of $T_{\text{candidate}}$ (38/375). This is an expected consequence of how $T_{\text{candidate}}$ and $I_{\text{low_k}}$ are constructed, not a sampling accident: $K(x)$-stratification defines $T_{\text{candidate}}$ as items *at least one* model retrieved and $I_{\text{low_k}}$ as items *no* model retrieved, so $T_{\text{candidate}}$ is selected, by construction, toward more globally prominent or easier entities, and domains whose items are uniformly easy or uniformly hard for these three models end up almost entirely on one side.

Since mean PRS itself varies considerably by domain (e.g. *art* 0.601 vs. *technology* 0.392 under $I_{\text{low_k}}$, Table 9), this composition difference could inflate the unadjusted aggregate gap independent of any partial-retrieval effect. The cluster-robust logistic regression reported in Section 9.1 does *not* adjust for this: its covariates are condition and model identity only, not domain. To check whether it survives an explicit domain adjustment, we refit it with domain fixed effects added (`correct` $\sim$ `condition` + model + domain, still clustered by item): the condition coefficient shrinks from 0.174 ($p = 3.2 \times 10^{-4}$, the value reported in Section 9.1) to 0.118 ($p = 4.5 \times 10^{-3}$) – smaller, and consistent with domain composition explaining part of the raw gap, but still positive and significant rather than eliminated. We read this, together with the chance-corrected check in Appendix B, as evidence the aggregate effect is not entirely a composition artifact, while still not fully ruling out that composition (domain, or the *occupation*-applicability difference found there) accounts for some share of the raw 0.539 vs. 0.507 gap. This is the same base-rate concern raised for the synthetic control in Section 9.3, applying in a subtler, harder-to-remove form to $I_{\text{low_k}}$ itself; we do not have a design that fully separates genuine target-specific residue from this composition effect (Section 13 returns to this), and flag it as a limitation on how the aggregate effect should be interpreted rather than as evidence against it.

G Power Analysis for the Expanded Pilot

The initial 432-item pilot's only nominally significant per-attribute result against the fair, distribution-matched control was *geographic_region* (0.795 vs. 0.703, $n_T = 146$, $n_{I_{\text{low_k}}} = 276$, one-sided Fisher exact $p = 0.027$). Treating this as the pre-registered target for a confirmatory follow-up, a two-proportion power analysis (normal approximation, one-sided $\alpha = 0.05$, observed effect size and the initial pilot's n_T:n_I ratio of roughly 1:1.9) gives the required sample sizes in Table 10.

Table 10: Required n_T / $n_{I_{\text{low_k}}}$ for a *geographic_region* follow-up, by target power and correction regime.

Scenario	Power	n_T	$n_{I_{\text{low_k}}}$
Pre-registered single attribute ($\alpha = 0.05$)	80%	208	394
Pre-registered single attribute ($\alpha = 0.05$)	90%	289	546
All 7 attributes (Bonferroni $\alpha \approx 0.00714$)	80%	365	690
All 7 attributes (Bonferroni $\alpha \approx 0.00714$)	90%	469	887

We targeted 90% power for the single pre-registered attribute (the scientifically appropriate framing, since *geographic_region* was selected specifically because of its result in the initial pilot, not tested alongside six equally-favored alternatives): $n_T \approx 289$, $n_{I_{\text{low_k}}} \approx 546$, roughly double the initial pilot's combined natural item count. We therefore doubled `per_relation_natural` (55 $\rightarrow$ 110) and reran the full pipeline, yielding 375 $T_{\text{candidate}}$ and 687 $I_{\text{low_k}}$ failed trials overall (312 and 558 respectively with the *geographic_region* attribute present) – meeting or exceeding the 90%-power target. As reported in Section 9.3, the pre-registered test on this properly powered sample did not replicate.

H Prior-Consistent vs. Prior-Inconsistent Items

If the *length_bucket* probe answer merely reflects a per-relation prior (e.g. "founder-of-organization questions usually have short answers") rather than target-specific information, that shortcut is only available on items whose true bucket matches the relation's modal bucket. Restricting the original pilot's Table 2 comparison to *prior-inconsistent* items only (true bucket $\neq$ the item's relation's modal bucket across the natural set) drops the effect from significant to not: $P(\text{correct} \mid T) = 0.237$ vs. 0.206 ($n = 173/369$, $p = 0.21$) versus 0.240 vs. 0.166 ($p = 0.0024$) on the full sample. The point

estimate still favors T, but the sample size roughly halves and the gap shrinks by more than half, so this test alone is underpowered to distinguish "the effect is weaker than it looked" from "the effect was never target-specific" – which motivated the sharper, forced-choice test in Section 10.

I Falsifiability Criteria

We pre-registered the central prediction $PRS_T \gg PRS_I$ as explicitly falsifiable: the hypothesis would be weakened or rejected if

1. failed retrieval and ignorance have indistinguishable PRS values;
2. partial information disappears after controlling for question difficulty and answer frequency;
3. partial cues fail to facilitate retrieval;
4. the effect is limited to one model or prompt;
5. hidden-state probes show no corresponding structure.

As reported in the main text, criterion 5 is met for the one attribute-level candidate (*length_bucket*) we tested at the hidden-state level (Section 10.3), while criterion 1 is not met for the aggregate PRS effect (Section 9.1).

J Implications for Hallucination

We flag this section explicitly as a secondary, speculative downstream implication rather than a claim this paper provides evidence for: nothing in the pilot measures hallucination directly, and everything below is a hypothesis for future work, not a result.

If partial retrieval turns out to be associated with hallucination, this may provide a new way of thinking about factual errors. The conventional picture is approximately *knowledge or no knowledge*; our hypothesis instead suggests *knowledge* $\rightarrow$ *partial retrieval* $\rightarrow$ *completion*: a hallucination could arise when a model has enough contextual structure to construct a plausible answer but lacks sufficient information to identify the correct target. This hypothesis would be stronger than simply observing that uncertainty predicts hallucination, because it makes a prediction about the *content* of the model's residual knowledge – but it remains untested here, and should not be read as a contribution of the current pilot.

K Experimental Progression

The diagram below lays out the core experimental progression summarized in Section 14: what each question asks, and what we can actually report against it.

