Predicting LLM Hallucination Risk from Entity Frequency via Rate-Distortion Theory

Anonymous authors
Paper under double-blind review

Abstract

Large language models struggle with factual hallucinations, and mitigating them typically requires executing the model to assess uncertainty. We introduce a query-dependent rate–distortion framework showing that factual accuracy follows a predictable, sigmoidal "knowledge cliff" governed by training exposure. Below a critical frequency f_{crit}, the model lacks the representational budget to reliably retrieve facts; above this threshold, accuracy increases precipitously. We present five core findings based on this framework. First, mapping this frequency response yields an accurate, zero-compute risk score; we formally prove that this pre-inference metric achieves $> 99\%$ of the theoretical upper bound for any frequency-only predictor. Second, this cliff is heterogeneous, with f_{crit} varying by up to $76\times$ depending on the query's relation type. Integrating these structural metadata creates a classifier that outperforms the LLM's own post-inference confidence scores in the rare-entity tail. Third, we establish a sample-complexity bound demonstrating that the calibration data required to locate this cliff under Zipfian sampling is independent of its location. Fourth, we isolate the effect of model scale using the Qwen2.5-Instruct family (0.5B–14B, trained on a fixed corpus), revealing a power-law relationship ($f_{\text{crit}} \propto P^{-0.52}$) where a $10\times$ parameter increase yields reliable recall for entities only $\sim 3.3\times$ less frequent. Finally, we show that these metadata-driven signals establish a superior budget–utility frontier for retrieval routing. By demonstrating that reliability is largely determined by query properties prior to generation, this work enables highly efficient, pre-hoc triage for retrieval-augmented systems.

1 Introduction

Large language models (LLMs) produce fluent, confident text regardless of whether they possess the underlying factual knowledge. This creates a fundamental reliability problem: for any given query, it is unclear *a priori* whether the model will produce a correct answer or a hallucination. The prevailing approach to hallucination detection operates post-hoc such as analyzing output logits, confidence scores, or semantic consistency *after* generation (Kadavath et al., 2022; Manakul et al., 2023; Kuhn et al., 2023). This is expensive at scale since for a system handling millions of queries per day, detecting hallucination after generation wastes both computation and user trust.

In this work, we demonstrate that for factual question-answering about entities, a key determinant of hallucination is not the model's generation process but a property of the query itself: how frequently the subject entity appeared in pretraining data. This observation has been noted empirically (Kandpal et al., 2023; Mallen et al., 2023; Sun et al., 2024), but lacks a theoretical framework that explains the *shape* of the failure mode, predicts which parameters are invariant across models, and yields a practical detection tool.

Contributions. We develop a query-dependent rate–distortion framework for LLM factual reliability and derive five empirically validated results:

1. **Theory: a knowledge boundary from capacity allocation.** We model factual recall as lossy compression under a finite representational budget allocated across a Zipfian distribution of

"competency units." Under threshold-like distortion and heterogeneous storage difficulty, the theory predicts a sigmoidal phase transition (the "knowledge cliff") in accuracy as a function of exposure, characterized by a critical frequency f_{crit} (Section 3). While architectural details can affect constants, our analysis focuses on constraints driven by coverage and allocation.

2. **Pre-inference risk and converse bound.** A frequency-derived risk score achieves strong discrimination and proper scoring without running the model (AUROC = 0.810, Brier = 0.169). We prove that no predictor depending only on entity frequency can exceed $\text{AUROC}^* = \frac{1}{2} + \text{Cov}_q(g(X), F_q(X))/[\bar{a}(1-\bar{a})]$, and that AUROC^* is strictly increasing in κ (the inverse scale of storage heterogeneity). This dictates that shallow cliffs impose a hard discrimination ceiling that can approach chance as $\kappa \to 0$. For Mistral-7B, the bound gives $\text{AUROC}^* = 0.816$; the empirical frequency score attains 99.2% of this ceiling (Section 4). Frequency and confidence are complementary: combining them yields AUROC = 0.856 and Brier = 0.153, exceeding the pre-inference ceiling by exploiting model outputs (Section 6.5).

3. **Query structure enables individual-level prediction and routing.** Relation structure shifts the cliff substantially (a 76× range in f_{crit} across relation types) and dominates within-stratum discrimination in the tail (AUROC = 0.826 vs. 0.752 for confidence in the <1K stratum). In 10-fold CV, the best fully pre-inference predictor (`Freq + relation`) achieves $\text{AUROC} = 0.842 \pm 0.055$, while adding confidence post-inference yields $\text{AUROC} = 0.875 \pm 0.044$ (Section 6.7). As a systems payoff, this pre-inference score produces competitive budget–utility frontiers for retrieval routing, requiring 47% RAG budget to reach 80% accuracy compared to 54% for confidence-based policies (Section 4).

4. **Capacity scaling law validated.** The theory predicts $f_{\text{crit}} \propto R_{\text{total}}^{-\beta}$: larger models recall rarer entities. We validate this directly using the Qwen2.5-Instruct family (0.5B–14B, fixed 18T-token corpus), finding $\hat{\beta} = 0.52 \pm 0.05$ with $R^2 = 0.940$ across five model sizes. Each decade of parameter growth shifts the cliff by 0.52 log-decades of frequency (Section 7).

5. **Sample complexity of cliff localization.** We derive the Fisher information for estimating $x_{\text{crit}} = \log f_{\text{crit}}$ from labeled examples and give a closed-form sample complexity bound. Under natural Zipfian sampling with $\alpha = 1$, the required n is *independent of* f_{crit} and scales as $1/\kappa$: steep cliffs are disproportionately easy to calibrate. For $\alpha \neq 1$, tail cliffs (low f_{crit}) require more samples by a factor of $f_{\text{crit}}^{\alpha-1}$. The bound is numerically consistent with the empirical finding that ECE < 0.10 is achievable with ∼75 calibration samples at $\kappa = 4.87$ (Section 5.1).

2 Related Work

The empirical observation that LLMs struggle with rare entities is now well established. Neural scaling laws (Kaplan et al., 2020; Hoffmann et al., 2022) predict that loss decreases as a power law in model size and data, but these aggregate trends mask entity-level heterogeneity. Two lines of work are particularly relevant. First, Kandpal et al. (2023) establish that factual accuracy correlates with the number of *relevant pretraining documents* associated with a fact, and show both correlational and causal relationships between pretraining support and QA performance; they further show retrieval augmentation can reduce dependence on pretraining support. Second, Mallen et al. (2023) introduce PopQA, use Wikipedia popularity signals (including pageviews) as a practical proxy for long-tailness, and show that retrieval augmentation disproportionately helps low-popularity entities. This head-to-tail degradation has also been documented at larger scale in the Head-to-Tail benchmark (Sun et al., 2024), which evaluates factual knowledge across popularity buckets and finds systematic performance collapse from head to tail.

Our work is complementary to these empirical findings. Rather than treating the frequency–accuracy relationship as an empirical regularity, we propose a *query-dependent rate–distortion* account that predicts *why* the transition is sharp (threshold distortion), *where* it occurs (a critical exposure f_{crit}), and, how the boundary transfers and scales across models. In other words, prior work establishes the long-tail phenomenon and the utility of exposure proxies; we supply a mechanistic RD explanation and derive decision-relevant limits and policies from it.

A natural response to unreliable LLM outputs is to detect hallucinations after generation. Kadavath et al. (2022) show that language models "mostly know what they know," using token-level self-evaluation (P(True)) as a confidence signal. Kuhn et al. (2023) propose semantic entropy (consistency across sampled outputs), and Farquhar et al. (2024) extend this to practical confabulation detection. Manakul et al. (2023) introduce SelfCheckGPT for zero-resource detection via self-consistency, while Min et al. (2023) develop FActScore for long-form factuality evaluation. Probe-based approaches train classifiers on internal representations to detect falsehood (Azaria & Mitchell, 2023; Li et al., 2024). Recent work has also systematized *factual confidence* estimation across methods and benchmarks, including PopQA (Mahaut et al., 2024). All of these methods share a fundamental limitation: they require running the model first. In high-volume systems, this implies hallucination detection cost scales with query volume and the model has already generated (and potentially served) the hallucination before it is detected. Our approach inverts this paradigm: because our risk score depends only on query-side properties—exposure and relation structure—it can be computed *pre-inference* with no additional model execution.

This pre-inference risk score naturally connects to the question of *when* to augment generation with retrieval. Adaptive retrieval methods such as Self-RAG (Asai et al., 2024) and FLARE (Jiang et al., 2023b) decide during generation whether to retrieve, but still require partial inference to make the routing decision. Similarly, RouteLLM (Ong et al., 2025) learns a routing function to dispatch queries between systems, but requires learned routing machinery. Our framework provides a routing signal available before any inference begins. Moreover, we show that optimal budgeted routing under a hard retrieval cap has a closed-form *upgrade-by-gain* structure (LP duality), and we derive an entropic OT (Sinkhorn) relaxation that yields a smooth transport plan.

Rate–distortion theory has a rich history in information theory (Cover & Thomas, 2006) and has been applied to analyze learning and compression in neural networks. Shwartz-Ziv & Tishby (2017) use the information bottleneck to analyze deep learning, and Delétang et al. (2024) formalize language modeling as compression, providing a natural foundation for our RD formulation. In the memorization literature, Carlini et al. (2023) establish log-linear relationships between memorization probability and both capacity and duplication count; Tirumala et al. (2022) observe sigmoidal memorization curves over training; and Allen-Zhu & Li (2025) argue models store $\sim$2 bits of extractable knowledge per parameter with heterogeneous extractability across knowledge types. These findings motivate our modeling ingredients: (i) logarithmic effective rate in exposure, (ii) heterogeneous storage difficulty, and (iii) a sharp transition under threshold distortion.

Most directly related to our theoretical contribution is the work of Kalai & Vempala (2024), who proved that calibrated language models *must* hallucinate on singleton facts at a rate tied to Good-Turing estimation. Their follow-up (Kalai et al., 2025) extends this to the full training pipeline and identifies a key structural cause: standard evaluation metrics use binary scoring (correct/incorrect with no credit for abstention), which incentivizes confident guessing over expressing uncertainty. They propose modifying benchmark scoring to reward calibrated responses, leaving the connection between the scoring rule and the hallucination boundary implicit. Our rate-distortion framework makes this connection explicit and quantitative. The binary scoring function that Kalai et al. (2025) identify as the culprit corresponds precisely to the threshold distortion measure in our Assumption 1: $d(x, \hat{x}) = \mathbf{1}[\hat{x} \neq x]$. It is this threshold structure, not merely finite capacity, that produces the *sharp* phase transition. Classical rate-distortion theory shows that the shape of $R(D)$ depends critically on the distortion measure; replacing threshold distortion with a smooth alternative (e.g., log-loss) yields a qualitatively different rate-distortion function, and our framework predicts that the cliff would soften into a gradual rolloff (see Section 7). We further instantiate this abstention lens experimentally and show that the combined frequency+confidence abstention policy dominates calibrated confidence-only abstention at matched coverage, with the pre-inference `Freq + relation` component providing a strong ranking signal (Figure 4). Thus, the two frameworks are complementary at a deep level: Kalai & Vempala (2024) prove that hallucination on rare facts is an inevitable consequence of calibration; Kalai et al. (2025) diagnose binary scoring as the mechanism that makes this hallucination *confident*; and our work shows that binary scoring, formalized as threshold distortion, is also the mechanism that makes the knowledge boundary *sharp*. In short, they explain why models guess; our work explains why the guessing has a cliff. Concurrently, Guo & Li (2026) also apply rate-distortion theory to explain LLM hallucinations, but formalize it as a membership testing problem. They demonstrate that in a sparse universe of facts, the space-optimal

strategy for a capacity-constrained model is to assign high confidence to non-facts rather than abstaining. While their work beautifully establishes the information-theoretic necessity of confident hallucination in a binary "fact vs. non-fact" setting, our framework extends the rate-distortion lens to the continuous spectrum of entity exposure together with its mathematical scaling laws across model parameters, and translates these theoretical limits into a deployable pre-inference routing metric.

3 Theoretical Framework

3.1 Setup: Factual Recall as Lossy Compression

An LLM $\mathcal{M}$ is trained on a corpus of entities $\mathcal{K} = \{1, \ldots, K\}$, where $f_k \propto k^{-\alpha}$ ($\alpha \approx 1$) is the Zipf frequency of the k-th most common entity. Each entity is associated with relational facts; each query x_k carries a relation label $R_k \in \mathcal{R}$ (e.g., `capital-of`, `born-in`) and the correct answer instantiates that predicate for the queried entity.

Training must encode these facts into a fixed parameter budget. Allen-Zhu & Li (2025) estimate $\sim$2 bits of extractable knowledge per parameter, giving a 7B model an effective capacity $R_{\text{total}} \approx 14 \times 10^9$ bits; arguably, far below the lossless requirement for millions of entity–fact pairs. Training therefore performs implicit *lossy compression*: frequently seen facts are encoded reliably across many weight configurations, while rare facts receive fragile representations vulnerable to overwriting. The model's parameters function as a finite-capacity channel between training corpus and inference queries.

Definition 1 (Query-dependent rate-distortion)**.** *Let $q_k \propto f_k$ be the entity distribution induced by the corpus (the distribution SGD implicitly optimizes). Let $d_k(r_k)$ be the hallucination probability for entity k as a function of allocated rate r_k (bits of knowledge storage). The capacity allocation problem is:*

$$\min_{\{r_k\}} \sum_{k=1}^{K} q_k \cdot d_k(r_k) \quad \textit{s.t.} \quad \sum_{k=1}^{K} r_k \le R_{\text{total}}, \tag{1}$$

with $R_{\text{total}} \approx 2P$ bits for a model with P parameters (Allen-Zhu & Li, 2025). The formulation is query-dependent: each entity has its own distortion function d_k, weighted by exposure q_k.

SGD does not literally solve this water-filling problem, but approximates it: high-frequency entities are encoded redundantly and recalled robustly, while low-frequency entities receive fragile representations. The key question is the *shape* of the resulting transition. We first analyze the homogeneous case (common distortion–rate curve $D(\cdot)$) for a clean allocation characterization, then introduce heterogeneous $\{d_k\}$ in Section 3.2 to recover the sigmoid cliff.

Proposition 2 (Knowledge Cliff via Convex Rate–Distortion)**.** *Let $D : [0, \infty) \to [0, 1]$ be a differentiable, strictly convex, and strictly decreasing distortion–rate function representing the probability of hallucination given r bits of allocated capacity. Let entity frequencies f_k follow a Zipfian distribution, with normalized weights $q_k = f_k / \sum_j f_j$. In the homogeneous allocation problem with total capacity budget R_{total}, the optimal hallucination probability as a function of frequency exhibits a sharp cutoff:*

$$P(\text{error} \mid f) = \begin{cases} D(0) & \textit{if } f \le f_{\text{crit}} \\ D\Big((D')^{-1}\Big(-\frac{\tilde{\lambda}}{f}\Big)\Big) & \textit{if } f > f_{\text{crit}} \end{cases} \tag{2}$$

where $f_{\text{crit}} = -\tilde{\lambda}/D'(0)$, and $\tilde{\lambda} > 0$ is a constant uniquely determined by the total capacity budget R_{total}.

Proof. We consider the homogeneous special case of Definition 1 in which $d_k(\cdot) \equiv D(\cdot)$ for all k. The model's pretraining objective is approximated by the rate–distortion allocation problem

$$\min_{\{r_k\}} \mathcal{J} = \sum_k q_k \, D(r_k) \quad \text{s.t.} \quad \sum_k r_k \le R_{\text{total}}, \qquad r_k \ge 0 \ \forall k. \tag{3}$$

The KKT stationarity conditions for the Lagrangian $\mathcal{L} = \sum_k q_k D(r_k) + \lambda(\sum_k r_k - R_{\text{total}}) - \sum_k \mu_k r_k$ yield:

$$q_k D'(r_k) + \lambda - \mu_k = 0, \qquad \mu_k \ge 0, \quad \mu_k r_k = 0.$$

For an interior optimum ($r_k^* > 0$), complementary slackness dictates $\mu_k = 0$, giving $D'(r_k^*) = -\frac{\lambda}{q_k}$. Because D is strictly convex and differentiable, D' is strictly increasing and invertible on its range, yielding $r_k^* = (D')^{-1}\left(-\frac{\lambda}{q_k}\right)$.

However, if the frequency weight q_k is sufficiently small, the non-negativity constraint binds ($r_k^* = 0$ and $\mu_k \geq 0$). This occurs when $q_k \leq -\lambda/D'(0)$. Substituting $q_k = f_k / \sum_j f_j$ and defining the scaled multiplier $\tilde{\lambda} := \lambda \sum_j f_j > 0$, we identify a critical frequency cutoff $f_{\text{crit}} := -\tilde{\lambda}/D'(0)$.

Entities below this critical frequency receive exactly zero capacity ($r_k^* = 0$) and suffer maximum baseline distortion $D(0)$. For entities above the cliff ($f > f_{\text{crit}}$), the allocated rate is $r_k^* = (D')^{-1}(-\tilde{\lambda}/f_k)$, yielding the continuous distortion curve $P(\text{error} \mid f) = D\big((D')^{-1}(-\tilde{\lambda}/f)\big)$. The shadow price $\tilde{\lambda}$ is determined by the capacity constraint $\sum_{k:f_k > f_{\text{crit}}} (D')^{-1}(-\tilde{\lambda}/f_k) = R_{\text{total}}$. □

Proposition 2 assumes a smooth, strictly convex $D(R)$. In the next section, we derive the sigmoid from a complementary starting point: threshold distortion (Assumption 1), which is a step function and therefore *not* strictly convex. The two results are reconciled by noting that heterogeneous storage difficulty (Assumption 2) averages over a population of threshold distortion functions, producing an *effective* population-level $D(R)$ that is smooth and convex. The sigmoid of proposition 4 can thus be viewed as the special case of proposition 2 where the effective distortion-rate curve arises from a logistic mixture of step functions.

3.2 The Knowledge Cliff as Phase Transition

The shape of the distortion function $d_k(r_k)$ determines whether the transition from reliable to unreliable recall is gradual or sharp. We make three modeling assumptions, each grounded in independent empirical findings from the memorization literature.

Assumption 1 (Threshold distortion). Storing a fact requires a minimum rate: $d_k(r_k) = 0$ if $r_k \geq r_{\min,k}$ and $d_k(r_k) = 1$ otherwise. This is natural for factual recall, i.e., knowing that the capital of Burkina Faso is Ouagadougou is an all-or-nothing proposition; a partial representation is of little practical value.

Assumption 2 (Heterogeneous storage difficulty). Different facts require different minimum rates. We model $r_{\min,k} = r_0 + \eta_k/\kappa$ where η_k is an idiosyncratic difficulty term. This is motivated by Allen-Zhu & Li (2025), who show that different knowledge types (birth dates, universities, employers) require varying amounts of capacity to become extractable. We assume η_k follows a standard logistic distribution, with $1/\kappa$ controlling the scale of heterogeneity, for analytical convenience and interpretability.

Assumption 3 (Logarithmic rate allocation). Training on f_k occurrences of entity k provides an effective rate $r_k^{\text{eff}} \propto \log f_k$. This is supported by Carlini et al. (2023), who establish a log-linear relationship between memorization probability and training data duplication count, i.e., $P(\text{memorized}) \propto \log(\text{count})$, implying diminishing marginal returns in representational robustness per exposure.

3.3 Deriving Logarithmic Rate Allocation from Gradient Dynamics

Assumptions 1 and 2 are empirically grounded (Allen-Zhu & Li, 2025). Assumption 3 currently rests on the empirical log-linear relationship observed by Carlini et al. (2023). We now provide a mechanistic justification for this logarithmic encoding based on gradient saturation.

Proposition 3 (Gradient saturation implies logarithmic encoding)**.** *Fix an entity k that appears in M supervised training steps. Let $\theta^{(m)}$ be parameters after the m-th step, and define the scalar $u^{(m)} := f_k(\theta^{(m)}) \in \mathbb{R}$, where f_k denotes the network's pre-softmax margin for the correct answer. Assume updates use the logistic cross-entropy $\mathcal{L}(u) = -\log \sigma(u)$ with $\sigma(u) = (1+e^{-u})^{-1}$ via gradient descent: $\theta^{(m+1)} = \theta^{(m)} - \eta \nabla_\theta \mathcal{L}(u^{(m)})$ for $\eta > 0$. Assume the squared gradient norms satisfy $\|\nabla_\theta f_k(\theta^{(m)})\|_2^2 \in [g_{\min}, g_{\max}]$ with $0 < g_{\min} \leq g_{\max} < \infty$. If $\eta M \to \infty$ and η is small enough, then:*

$$u^{(M)} = \log(\eta M) + O(1) + O(M\eta^2). \tag{4}$$

If the event is sampled at rate $q_k \propto f_k$ over T total steps so $M = q_k T$, then $u^{(M)} = \log f_k + \text{const}$, establishing logarithmic rate allocation.

Proof. Write $g_m := \|\nabla_\theta f_k(\theta^{(m)})\|_2^2$. Since $\mathcal{L}'(u) = -\sigma(-u)$, the parameter step is $\Delta\theta_m = \eta\,\sigma(-u^{(m)})\,\nabla_\theta f_k(\theta^{(m)})$. By Taylor's theorem with a Lipschitz gradient, the margin update is $u^{(m+1)} - u^{(m)} = \eta\,\sigma(-u^{(m)})\,g_m + r_m$, with $|r_m| \le C_r\eta^2\sigma(-u^{(m)})^2 g_m$.

Define the transformation function $V(u) := u + e^u$, which has derivative $V'(u) = 1 + e^u = \frac{1}{\sigma(-u)}$. A second-order expansion gives:

$$V(u^{(m+1)}) - V(u^{(m)}) = V'(u^{(m)})(u^{(m+1)} - u^{(m)}) + \xi_m.$$

Substituting the margin update, the first-order term simplifies perfectly:

$$V'(u^{(m)})(u^{(m+1)} - u^{(m)}) = \frac{1}{\sigma(-u^{(m)})}\big[\eta\sigma(-u^{(m)})g_m + r_m\big] = \eta g_m + \frac{r_m}{\sigma(-u^{(m)})}.$$

Because $|r_m|$ scales with $\sigma(-u^{(m)})^2$, the remainder $\frac{r_m}{\sigma(-u^{(m)})}$ is bounded by $O(\eta^2)$. Similarly, the second-order term ξ_m scales with $V''(u) = e^u$, which when multiplied by $(\Delta u)^2 \propto e^{-2u}$, decays exponentially and is strictly bounded by $O(\eta^2)$.

Summing over M steps yields $V(u^{(M)}) = V(u^{(0)}) + \eta\sum_{m=1}^{M} g_m + O(M\eta^2)$. For large u, $V(u) \approx e^u$, which implies $e^{u^{(M)}} = \Theta(\eta M)$. Taking the logarithm yields $u^{(M)} = \log(\eta M) + O(1) + O(M\eta^2)$, proving (4). □

We identify the accumulated margin $u^{(M)}$ with the effective allocated rate r_k^{eff} because the pre-softmax logit represents the log-odds of the correct prediction, serving as a direct information-theoretic measure of the capacity the network has dedicated to memorizing that entity. We also validate the key assumptions against our data: accuracy fits substantially better on $\log f$ than on raw f ($R^2 = 0.958$ vs. 0.817), confirming Assumption 3; and the 76× variation in f_{crit} across relation types (Section 6.7) directly confirms the heterogeneity in Assumption 2.

Proposition 4 (Sigmoid phase transition)**.** *Under Assumptions 1–3, the expected accuracy as a function of entity frequency follows an affine sigmoid:*

$$\text{acc}(f) = a_{\min} + (a_{\max} - a_{\min}) \cdot \sigma\big(\kappa(\log f - \log f_{\text{crit}})\big), \tag{5}$$

where f_{crit} corresponds to the inflection point (the critical frequency at which the probability of successful storage is exactly 0.5), κ is the inverse scale of storage heterogeneity (Assumption 2), $a_{\max}$ is the accuracy ceiling in the stored regime, and $a_{\min}$ is the baseline accuracy floor (e.g., guessing) in the unstored regime.

Proof. We proceed in two steps: we first solve the deterministic rate-distortion problem under threshold distortion; second, we then introduce storage heterogeneity. Assume all entities have identical storage cost ($\eta_k = 0$ for all k, i.e., no heterogeneity). Under Assumption 1, the threshold distortion is:

$$d_k(r_k) = \begin{cases} 0 & \text{if } r_k \ge r_{\min} \\ 1 & \text{if } r_k < r_{\min} \end{cases} \tag{6}$$

The rate-distortion problem $\min_{\{r_k\}} \sum_k q_k d_k(r_k)$ subject to $\sum_k r_k \le R_{\text{total}}$ reduces to a 0-1 knapsack with equal weights: select the subset $S \subseteq \mathcal{K}$ maximizing $\sum_{k\in S} q_k$ subject to $|S| \cdot r_{\min} \le R_{\text{total}}$.

Since all items have cost $r_{\min}$, the optimal solution stores the $K^* = \lfloor R_{\text{total}}/r_{\min} \rfloor$ entities with the highest q_k, i.e., entities with rank $k \le K^*$. Under Zipf ordering ($q_1 \ge q_2 \ge \cdots$), this yields a step function:

$$\text{acc}_{\text{det}}(f) = \begin{cases} 1 & \text{if } f \ge f_{K^*} \\ 0 & \text{if } f < f_{K^*} \end{cases} \tag{7}$$

where f_{K^*} is the frequency of the marginal stored entity.

While this deterministic knapsack yields a hard cliff, the inherently noisy capacity allocation in neural networks softens this boundary. We now incorporate Assumptions 2 and 3. Entity k requires rate $r_{\min,k} = r_0 + \eta_k/\kappa'$

(Assumption 2), and training provides effective rate $r_k^{\text{eff}} = \beta \log f_k$ (Assumption 3). Entity k is successfully stored if $r_k^{\text{eff}} \geq r_{\min,k}$, i.e., if $\beta \log f_k \geq r_0 + \eta_k/\kappa'$, which rearranges to:

$$\eta_k \leq \kappa(\log f_k - \log f_{\text{crit}}) \tag{8}$$

where $\kappa = \kappa'\beta$ absorbs the proportionality constants and $\log f_{\text{crit}} = r_0/\beta$ is the critical log-frequency at which the median entity is stored. Since η_k follows a standard logistic distribution (Assumption 2), we have:

$$P(\text{stored} \mid f_k) = P\big(\eta_k \leq \kappa(\log f_k - \log f_{\text{crit}})\big) = \sigma\big(\kappa(\log f_k - \log f_{\text{crit}})\big) \tag{9}$$

where $\sigma(x) = 1/(1+e^{-x})$ is the logistic sigmoid. Let $a_{\max}$ denote expected accuracy when the fact is successfully stored/extractable, and $a_{\min}$ denote expected accuracy otherwise (e.g., chance-level guessing). Then

$$\text{acc}(f) = a_{\max} P(\text{stored} \mid f) + a_{\min}\big(1 - P(\text{stored} \mid f)\big),$$

which yields Equation (5) after substituting Equation (9). □

Remark 1 (Choice of logistic distribution)**.** *The logistic distribution in Assumption 2 is chosen for convenience; the cliff is not an artifact of this choice. A broad class of unimodal (and even heavy-tailed) heterogeneity models produce essentially the same transition. We verify this by fitting six alternatives—Logistic, Gaussian (probit), Laplace, Gumbel (extreme value), Cauchy (heavy-tailed), and a Gaussian mixture—using* unbinned *entity-level Bernoulli likelihoods:*

$$y_i \sim \text{Bernoulli}\Big(a_{\min} + (a_{\max} - a_{\min})\, F\big(\kappa(x_i - x_{\text{crit}})\big)\Big),$$

where F is the CDF of the assumed heterogeneity distribution. All six fits are statistically indistinguishable: the maximum ΔAIC *across models is* 0.66 *(well below the conventional threshold of* 2 *for "no evidence" of distinction; Burnham & Anderson 2002), and the inferred cliff location is stable with* $\log f_{\text{crit}}$ *spanning only* 0.11 *decades (*4.00–4.11*). The raw κ values differ across distributions (e.g., logistic $\kappa = 5.49$ vs. Gaussian $\kappa = 3.57$) because standardized CDFs have different derivatives at the origin, so the same κ corresponds to different maximal slopes under different link functions. To obtain a distribution-invariant steepness measure, we compute the slope at the critical point:*

$$s^* \;=\; (a_{\max} - a_{\min})\, \kappa\, f(0),$$

where f is the PDF corresponding to F. This quantity, i.e., the maximum rate of accuracy change per decade of frequency, is stable across all six distributions: $s^* \in [0.90, 1.10]$ with CV $= 7.9\%$*. We therefore use the logistic primarily for convenience and interpretability (closed-form sigmoid), not because it is uniquely favored by the data; distributional choice mainly affects tail behavior (how quickly accuracy approaches its floor/ceiling), not the existence, location, or steepness of the cliff.*

Remark 2. *The logistic approximation $\hat{a}(f)$ is robust to the misspecification of the true storage heterogeneity distribution. Let F_{r^*} denote the true cumulative distribution function (CDF) of the latent storage thresholds across entities. We define the approximation error as $\delta(x) = F_{r^*}(x) - \sigma(x)$, where $x = \kappa(\log f - \log f_{\text{crit}})$ is the standardized log-frequency. Since $\hat{a}(f) - a(f) = (a_{\max} - a_{\min})\delta(x)$, writing $\mathcal{E}$ as an expectation and applying $|\delta| \leq d_{\text{KS}} := \sup_x |\delta(x)|$ gives*

$$\mathcal{E} \;:=\; \mathbb{E}_q\big[(\hat{a}(f) - a(f))^2\big] \;\leq\; (a_{\max} - a_{\min})^2 \cdot d_{\text{KS}} \cdot \mathbb{E}_q[|\delta(X)|],$$

where $\mathbb{E}_q[|\delta(X)|]$ is the average pointwise deviation under the query distribution. The bound has a natural interpretation: calibration error is controlled by the amplitude squared, the worst-case pointwise error (d_{KS}), and the average pointwise error ($\mathbb{E}_q[|\delta|]$). Let L_f be the effective width of the log-frequency support. At $\alpha = 1$, $\mathbb{E}_q[|\delta|] \approx W_1(F_{r^*}, \text{Logistic})/(\kappa L_f)$ by the CDF identity $\int |F - G|\, dx = W_1(F, G)$; for $\alpha > 1$ the expectation is smaller since heavier Zipf concentrates queries on low-x entities where $|\delta| \approx 0$.

A steep cliff is self-correcting: the transition window has width $O(1/\kappa)$ in $\log f$, outside which $|\delta(x)| \approx 0$ for any unimodal F_{r^*}, so misspecifying the distribution family costs negligible calibration error. For Mistral-7B at $\alpha = 1$ ($\mathbb{E}_q[|\delta|] \approx W_1/(\kappa L_f) = 0.5/34.09 \approx 0.015$, $a_{\max} - a_{\min} = 0.650$, $d_{\text{KS}} = 0.203$), this gives $\mathcal{E} \lesssim (a_{\max} - a_{\min})^2 \cdot d_{\text{KS}} \cdot \mathbb{E}_q[|\delta|]$ where $\mathbb{E}_q[|\delta|] \approx W_1/(\kappa L_f) \approx 0.015$ consistent with $R_w^2 = 0.969$.

By the DKW inequality (Massart, 1990), replacing d_{KS} with its empirical estimate $\hat{d} := \sqrt{\log(2/\delta)/(2n)}$ and bounding $\mathbb{E}_q[|\delta|]$ via $W_1/(\kappa L_f)$, $\mathcal{E} < \tau$ can be certified with probability $\geq 1-\delta$ whenever

$$n \;\geq\; \frac{(a_{\max} - a_{\min})^4 \cdot W_1^2 \cdot \log(2/\delta)}{2\,(\kappa L_f)^2\,\tau^2}.$$

The $1/\kappa^2$ scaling confirms that steep cliffs are quadratically cheaper to certify, linking to Proposition 11.

Remark 3 (Relation to empirical steepness)**.** *The parameter κ measures the inverse heterogeneity of storage difficulty. High κ (low heterogeneity, uniform $r_{\min}$) recovers the sharp step function of Equation (7). Low κ (high heterogeneity) produces a gradual transition. Our empirical finding that κ has $CV > 1$ across random splits suggests that storage heterogeneity is not well-characterized by a single scale parameter, consistent with the observation that relation type introduces a second axis of variation (Section 6.7).*

We note that the critical frequency f_{crit} we discussed so far emerges from the interaction between the capacity constraint and the Zipfian distribution: entities with $f_k \gg f_{\mathrm{crit}}$ have been seen enough times to be reliably stored, while entities with $f_k \ll f_{\mathrm{crit}}$ fall below the coverage threshold. We next characterize this as a function of total capacity and training corpus.

Proposition 5 (Data-determined cliff)**.** *Under the threshold-allocation argument in Proposition 4, the model stores the $K^* = \lfloor R_{\mathrm{total}}/r_{\min} \rfloor$ most frequent entities. The critical frequency $f_{\mathrm{crit}} = f_{K^*}$ is the frequency of the marginal stored entity:*

$$f_{\mathrm{crit}} = \frac{N}{H_K^{(\alpha)}} \cdot \left(\frac{R_{\mathrm{total}}}{r_{\min}}\right)^{-\alpha}, \tag{10}$$

where N is total training tokens and $H_K^{(\alpha)} = \sum_{j=1}^{K} j^{-\alpha}$. Taking logarithms:

$$\log f_{\mathrm{crit}} = \log N - \log H_K^{(\alpha)} - \alpha \cdot \log R_{\mathrm{total}} + \alpha \cdot \log r_{\min}. \tag{11}$$

Thus f_{crit} is a corpus- and capacity-determined quantity: it decreases with model size (R_{total}) and depends on training data through N, K, and α — not on architecture beyond its effect on effective storage capacity. This theoretical $-\alpha \log R_{\mathrm{total}}$ dependency governs the log-log scaling laws observed empirically in Section 6.8.

Proof. Consider entities ranked $k = 1, 2, \ldots, K$ by descending frequency. Under the Zipfian distribution with exponent α, the probability of the k-th entity is:

$$p_k = \frac{k^{-\alpha}}{H_K^{(\alpha)}}, \quad \text{where } H_K^{(\alpha)} = \sum_{j=1}^{K} j^{-\alpha} \tag{12}$$

is the generalized harmonic number. For a corpus of N total training tokens, the expected frequency is $f_k = N \cdot p_k = (N/H_K^{(\alpha)}) \cdot k^{-\alpha}$. From Proposition 4, the critical frequency corresponds to rank $K^* = \lfloor R_{\mathrm{total}}/r_{\min} \rfloor$. Substituting into the frequency model:

$$f_{\mathrm{crit}} = f_{K^*} = \frac{N}{H_K^{(\alpha)}} \cdot (K^*)^{-\alpha} \approx \frac{N}{H_K^{(\alpha)}} \cdot \left(\frac{R_{\mathrm{total}}}{r_{\min}}\right)^{-\alpha} \tag{13}$$

Taking logarithms:

$$\log f_{\mathrm{crit}} = \log N - \log H_K^{(\alpha)} - \alpha \log R_{\mathrm{total}} + \alpha \log r_{\min} \tag{14}$$

To express this in terms of corpus size alone, we apply Heaps' law: $K \approx \eta N^{\beta}$ for $\beta \in [0.4, 0.6]$. The scaling of $H_K^{(\alpha)}$ depends on α. For $\alpha \approx 1$, $H_K^{(1)} \approx \log K + \gamma$, by Euler–Mascheroni, so $\log H_K \approx \log(\beta \log N + \log \eta + \gamma)$. Because $\log(\log N)$ grows sub-polynomially, this term becomes negligible relative to $\log N$ for large corpora. Conversely, for $\alpha < 1$, $H_K^{(\alpha)} \approx K^{1-\alpha}/(1-\alpha)$, giving $\log H_K^{(\alpha)} \approx (1-\alpha)\beta \log N + c_0$. Unifying these regimes, we obtain:

$$\log f_{\mathrm{crit}} = c_{\mathrm{corpus}} \log N - \alpha \log R_{\mathrm{total}} + C(\alpha, r_{\min}) \tag{15}$$

where $c_{\mathrm{corpus}} = 1$ for $\alpha \approx 1$, and $c_{\mathrm{corpus}} = 1 - (1-\alpha)\beta$ for $\alpha < 1$. The term $C(\alpha, r_{\min})$ is a constant independent of N and R_{total}. □

Remark 4 (Extension to Mandelbrot-Zipf)**.** *The derivation generalizes to the Mandelbrot-Zipf distribution by replacing $k^{-\alpha}$ with $(k+q)^{-\alpha}$, where $q \geq 0$ flattens the distribution at high-frequency ranks. The harmonic number becomes the Hurwitz-type sum $H_{K,q}^{(\alpha)} = \sum_{j=1}^{K}(j+q)^{-\alpha}$. Since q is constant relative to N and K, the asymptotic behavior of $\log H_{K,q}^{(\alpha)}$ remains dominated by K (and thus N^{β}), so the scaling relation in Equation* (15) *holds with q affecting only the constant offset.*

Proposition 5 yields three predictions. First, models with similar effective capacity and training-data coverage should cluster at similar f_{crit}: Falcon-7B and Qwen-2.5-7B do cluster at $f_{\text{crit}} \approx$ 51–66K, consistent with this prediction, though architecture and corpus composition are not independently controlled so we do not treat this as a definitive test. Second, f_{crit} *increases* with raw corpus size N at fixed capacity (Heaps' law adds new entities faster than repeated exposure consolidates existing ones), so the relevant quantity for cross-model comparison is the *typical per-entity exposure* $\bar{f} := N/H_K^{(\alpha)}$, not N alone. Third, increasing model size R_{total} should decrease f_{crit} as $f_{\text{crit}} \propto R_{\text{total}}^{-\alpha}$; we validate this using the Qwen2.5-Instruct family (0.5B–14B), all trained on the same 18T-token corpus, obtaining $\hat{\alpha} = 0.52 \pm 0.05$ with $R^2 = 0.940$ on the log–log relationship (Section 6.8).

Cross-model test via per-entity exposure. For matched-capacity models, the capacity-dependent factor in Proposition 5 cancels in ratios, giving $f_{\text{crit}}^{(A)}/f_{\text{crit}}^{(B)} = \bar{f}_A/\bar{f}_B$ where $\bar{f} := N/H_K^{(\alpha)}$ is the typical per-entity exposure. Our three 7B models provide a natural test. Falcon-7B was trained on 1.5T English-dominated tokens; Qwen-2.5-7B on 18T tokens across 30+ languages; Mistral-7B's corpus size is undisclosed. Despite a 12× gap in total token count, Falcon and Qwen cluster at similar cliff locations (f_{crit} ratio = 66,009/50,843 = 1.30), while Mistral sits 5.2× lower (12,726/66,009 = 0.19). Falcon–Qwen parity is explained by multilingual dilution: Qwen's 18T tokens spread across 30 languages yield effective English entity exposure comparable to Falcon's 1.5T English-focused corpus. Mistral's lower f_{crit} is consistent with higher data quality rather than raw token count.

3.4 Why Confidence is Not Enough: The Erasure Regime

We now discuss why confidence itself is not sufficient by analyzing the calibration gap. We first have the following.

Lemma 6 (Fluency smoothness via capacity bottleneck)**.** *Fix a relation R and let the confidence score be a bounded function $g_R : \mathcal{Y} \to [0,1]$, such that $\phi_R(S) = \mathbb{E}_{Y \sim P_{R,S}}[g_R(Y)]$. In the rate-distortion framework, the allocated rate r_k bounds the information gained over the relation-conditioned prior π_R, such that $D_{\text{KL}}(P_{R,S} \parallel \pi_R) \leq r_k$. By Pinsker's $\|P_{R,S} - \pi_R\|_{\text{TV}} \leq \sqrt{\frac{1}{2} D_{\text{KL}}(P_{R,S} \parallel \pi_R)} \leq \sqrt{\frac{r_k}{2}}$. Hence, by the triangle inequality, any two states S, S' with allocated rates r, r' must satisfy:*

$$|\phi_R(S) - \phi_R(S')| \leq \|P_{R,S} - \pi_R\|_{\text{TV}} + \|\pi_R - P_{R,S'}\|_{\text{TV}} \leq \sqrt{\frac{r}{2}} + \sqrt{\frac{r'}{2}}.$$

In the erasure regime ($f \ll f_{\text{crit}}$), the optimal rate allocation is strictly $r = 0$. Thus, the conditional distributions collapse to the prior ($\|P_{R,S} - \pi_R\|_{\text{TV}} = 0$), and the expected confidence identically matches the prior's fluency baseline.

This theoretical collapse to the prior is empirically verifiable; the erasure regime condition holds when same-relation queries induce near-identical generation distributions, which can be measured directly via token-level KL divergence across same-relation prompts.

Proposition 7 (Confidence miscalibration in the erasure regime)**.** *Let $c_k \in [0,1]$ denote a confidence score for a query regarding entity k (e.g., normalized sequence confidence or answer confidence), and let $\text{acc}_k \in \{0,1\}$ denote correctness. Suppose there exists a latent factual state S and relation type R such that generation proceeds via a query-conditioned latent mixture. Assume $a_{\max}$ and $a_{\min}$ are as in Proposition 4 and*

1. ***Stored-regime concentration.*** *For $f_k \gg f_{\text{crit}}$, the posterior $\mu_k^{\text{st}} := P_\theta(S \mid x_k)$ concentrates near the true state S_k^+, such that the total variation distance is bounded:$\|\mu_k^{\text{st}} - \delta_{S_k^+}\|_{\text{TV}} \leq \delta_{\text{st}}$ where δ_{st}*

is small. Furthermore, we assume confidence tracks accuracy in the stored regime: $\mathbb{E}[c_k \mid f_k \gg f_{\text{crit}}, R_k = R] \approx a_{\max}$.

2. ***Erasure-regime prior reversion.*** *For* $f_k \ll f_{\text{crit}}$*, the posterior* $\mu_k^{\text{er}} := P_\theta(S \mid x_k)$ *reverts to a relation-conditioned prior* π_R*:* $\|\mu_k^{\text{er}} - \pi_R\|_{\text{TV}} \leq \delta_{\text{er}}$*. Under the RD model of Section 3, this follows from the zero-rate limit of the optimal encoder: as* $r_k^{\text{eff}} \to 0$ *(i.e.,* $f_k \to 0$*), the encoder ignores entity-specific input and outputs the marginal* π_R *that minimizes expected distortion.*

Then the confidence gap across the cliff is bounded by

$$|\mathbb{E}[c_k \mid f_k \ll f_{\text{crit}}, R_k = R] - \mathbb{E}[c_k \mid f_k \gg f_{\text{crit}}, R_k = R]| \leq \Delta_{\text{prior}} + \delta_{\text{st}} + \delta_{\text{er}} \tag{16}$$

for each relation R*, where* $\Delta_{\text{prior}} := \left|\mathbb{E}_{S\sim\pi_R}[\phi_R(S)] - \phi_R(S_k^+)\right|$ *which is strictly bounded by the rate allocation via Lemma 6. Consequently, the calibration gap in the erasure regime satisfies*

$$\mathbb{E}[c_k - \text{acc}_k \mid f_k \ll f_{\text{crit}}, , R_k = R] \gtrsim (a_{\max} - a_{\min}) - (\Delta_{\text{prior}} + \delta_{\text{st}} + \delta_{\text{er}}) \tag{17}$$

which is strictly positive whenever $a_{\max} - a_{\min} > \Delta_{\text{prior}} + \delta_{\text{st}} + \delta_{\text{er}}$.

Proof. Fix a relation R. Let

$$\mu^{\text{st}} := P_\theta(S \mid x_k, f_k \gg f_{\text{crit}}, R_k = R), \qquad \mu^{\text{er}} := P_\theta(S \mid x_k, f_k \ll f_{\text{crit}}, R_k = R)$$

and let S^+ denote the true latent state in the stored regime. Since confidence is a function of generation (Lemma 6), $\mathbb{E}[c_k \mid f_k \gg f_{\text{crit}}, R_k = R] = \int \phi_R(S)\, d\mu^{\text{st}}(S)$, and $\mathbb{E}[c_k \mid f_k \ll f_{\text{crit}}, R_k = R] = \int \phi_R(S)\, d\mu^{\text{er}}(S)$. We compare these expectations by inserting the intermediate measures δ_{S^+} and π_R:

$$\left|\int \phi_R\, d\mu^{\text{er}} - \int \phi_R\, d\mu^{\text{st}}\right| \leq \left|\int \phi_R\, d\mu^{\text{er}} - \int \phi_R\, d\pi_R\right| + \left|\int \phi_R\, d\pi_R - \phi_R(S^+)\right| + \left|\phi_R(S^+) - \int \phi_R\, d\mu^{\text{st}}\right| \tag{18}$$

For the first and third terms, since $\phi_R(S) \in [0, 1]$, the standard total-variation inequality implies:

$$\left|\int \phi_R\, d\mu^{\text{er}} - \int \phi_R\, d\pi_R\right| \leq \|\mu^{\text{er}} - \pi_R\|_{\text{TV}} \leq \delta_{\text{er}}$$

and similarly,

$$\left|\phi_R(S^+) - \int \phi_R\, d\mu^{\text{st}}\right| = \left|\int \phi_R\, d\delta_{S^+} - \int \phi_R\, d\mu^{\text{st}}\right| \leq \|\delta_{S^+} - \mu^{\text{st}}\|_{\text{TV}} \leq \delta_{\text{st}}$$

For the middle term, we define $\Delta_{\text{prior}} = \left|\int \phi_R\, d\pi_R - \phi_R(S^+)\right|$. By Lemma 6, this gap is structurally bounded by the capacity allocated to the stored state S^+.Substituting these bounds into (18) proves (16). For the calibration gap, let $\Delta_c := \Delta_{\text{prior}} + \delta_{\text{st}} + \delta_{\text{er}}$.By (16),

$$\mathbb{E}[c_k \mid f_k \ll f_{\text{crit}}] \geq \mathbb{E}[c_k \mid f_k \gg f_{\text{crit}}] - \Delta_c$$

By assumption, $\mathbb{E}[c_k \mid f_k \gg f_{\text{crit}}] \approx a_{\max}$, and Proposition 4 gives $\mathbb{E}[\text{acc}_k \mid f_k \ll f_{\text{crit}}] \approx a_{\min}$. Therefore, the calibration gap in the erasure regime is bounded by:

$$\mathbb{E}[c_k - \text{acc}_k \mid f_k \ll f_{\text{crit}}] \gtrsim a_{\max} - a_{\min} - \Delta_c$$

yielding (17). This gap is strictly positive when $a_{\max} - a_{\min} > \Delta_c$ (e.g., this holds for Mistral-7B, where $a_{\max} - a_{\min} = 0.650 \gg \Delta_c$). □

Remark 5 (Calibration vs. Discrimination)**.** *Because confidence is structurally elevated across the entire erasure regime, global post-hoc methods like temperature scaling cannot solve the hallucination gap. Monotone transformations improve global calibration metrics (ECE) but leave example ordering—and thus discrimination (AUROC)—unchanged. Correcting this requires a signal that actively distinguishes stored from erased entities, which is precisely why our frequency-based predictor dominates post-hoc scaling empirically (Section 6).*

Remark 6 (Connection to membership testing under uniform exposure)**.** *Recently, Guo & Li (2026) formulated LLM hallucination as a rate-distortion problem for binary membership testing (distinguishing facts from non-facts). Their optimal space-constrained strategy, which necessitates assigning high confidence to non-facts, can be viewed conceptually as a special case of our framework under a flattened, uniform exposure distribution. In the absence of a Zipfian frequency prior ($f_k = \text{const}$), a capacity bottleneck forces a random zero-bit allocation to a subset of facts. Our Proposition 7 provides the generative mechanism for this phenomenon: optimal compression is achieved precisely by reverting to the fluent, relation-conditioned prior π_R, which structurally enforces the high-confidence false positives predicted by their abstract membership bounds.*

4 Operational Routing from Rate–Distortion Geometry

The rate–distortion framework in Section 3 implies a sharp frequency-driven transition in factual reliability. In particular, the knowledge-cliff model yields a pre-inference prediction of noRAG error for an entity k with exposure frequency f_k:

$$\widehat{d}_{0,k} \;:=\; \Pr(\text{error} \mid f_k) \;\approx\; \sigma\big(\kappa(\log f_{\text{crit}} - \log f_k)\big). \tag{19}$$

Routing turns this RD-induced distortion proxy into an operational policy: given a limited retrieval capacity, allocate retrieval to the queries where it produces the largest reduction in expected distortion. For this, we consider a three-action menu: direct generation (noRAG, $i = 0$), retrieval-augmented generation (RAG, $i = 1$) and abstention ($\perp$, $i = 2$), where abstention means returning "I don't know". Let x_k denote the query instance for entity k with relation label $R_k \in \mathcal{R}$. We allow relation-conditioned residual risk under retrieval,

$$\widehat{d}_{1,k} \;:=\; \widehat{d}_1(R_k), \tag{20}$$

which serves as a static pre-inference lookup estimated from historical logs (or treated as a modeling primitive if a deployed system is unavailable). We use the same utility convention as in our experiments: a correct answer yields $+1$, an incorrect answer yields $-\beta$ with $\beta > 0$, retrieval incurs a fixed cost $c_{\text{RAG}} > 0$, and abstention yields a constant penalty $-\eta$ with $\eta \geq 0$. Under predicted error probabilities, the expected utilities are $u_{0,k} = 1 - (1+\beta)\widehat{d}_{0,k}$, $u_{1,k} = 1 - (1+\beta)\widehat{d}_{1,k} - c_{\text{RAG}}$, $u_{2,k} = -\eta$, and we define costs $C_{k,i} := -u_{i,k}$. Let q_k denote the query distribution over entities, induced by training exposure as in Section 3 (so $q_k \propto f_k$ and $\sum_k q_k = 1$). We impose a hard retrieval cap: at most a fraction b_{RAG} of query mass can be routed to retrieval. A deterministic policy is a map $T : \{1, \ldots, K\} \to \{0, 1, 2\}$, and the constrained problem is

$$\max_{T} \; \sum_{k=1}^{K} q_k \, u_{T(k),k} \quad \text{s.t.} \quad \sum_{k=1}^{K} q_k \, \mathbf{1}[T(k) = 1] \leq b_{\text{RAG}}. \tag{21}$$

This is the routing analogue of RD allocation: the model's internal capacity induces $\widehat{d}_{0,k}$ via the cliff (19), while external retrieval is an additional limited resource to be allocated under (21).

Proposition 8. *Assume N query instances with uniform instance masses $q_j = 1/N$. Let the baseline utility without retrieval for a given query instance j be $u_{\text{base},j} := \max\{u_{0,j}, u_{2,j}\}$, $T_{\text{base}}(j) \in \arg\max\{u_{0,j}, u_{2,j}\}$. Let $\Delta_j := u_{1,j} - u_{\text{base},j}$, $1 \leq j \leq N$. Ley absolute retrieval budget be $M = \lfloor b_{\text{RAG}} N \rfloor$. An optimal solution to the empirical counterpart of (21) is obtained by: (i) assigning every query j to the baseline action $T_{\text{base}}(j)$, and (ii) upgrading to retrieval the M queries with the largest positive Δ_j (ties arbitrary). Equivalently, retrieval is assigned to the top-M queries in Δ_j among those with $\Delta_j > 0$, and all other queries follow T_{base}.*

Proof. Write the objective as baseline plus improvements. For any feasible empirical policy T, define the set of retrieved empirical queries $S(T) := \{j : T(j) = 1\}$. Relative to the baseline policy T_{base}, the objective difference is

$$\sum_{j=1}^{N} q_j \, u_{T(j),j} \;-\; \sum_{j=1}^{N} q_j \, u_{\text{base},j} \;=\; \sum_{j \in S(T)} q_j \, (u_{1,j} - u_{\text{base},j}) \;=\; \sum_{j \in S(T)} q_j \, \Delta_j,$$

since queries not in $S(T)$ contribute exactly $u_{\text{base},j}$. With uniform masses $q_j = 1/N$, the constraint $\sum_{j=1}^{N} q_j \, \mathbf{1}[T(j) = 1] \leq b_{\text{RAG}}$ is equivalent to $|S(T)| \leq M$. Thus, maximizing the improvement under a cardinality constraint is solved by taking the M largest gains Δ_j, excluding any with $\Delta_j \leq 0$. □

This proposition makes the RD link explicit: because $u_{0,k}$ depends on $\widehat{d}_{0,k}$ and $\widehat{d}_{0,k}$ is governed by the cliff (19), the upgrade score Δ_k inherits a frequency geometry centered at f_{crit}. When abstention is inactive (or rare), $u_{\text{base},k} \approx u_{0,k}$ and $\Delta_k \approx (1+\beta)\big(\widehat{d}_{0,k} - \widehat{d}_{1,k}\big) - c_{\text{RAG}}$, so retrieval is most valuable in the below-cliff region, modulated by relation-conditioned residual risk $\widehat{d}_1(R_k)$.

The optimality of frequency-based routing raises a natural question: how close to optimal is the frequency score as a *discriminator* for hallucination risk? The following proposition gives a tight answer.

Proposition 9 (Bayes-optimal pre-inference AUROC and converse bound)**.** *Let $x = \log f \sim q$ on $[0, X]$ and $y \mid x \sim \text{Bernoulli}(g(x))$ with $g(x) = a_{\min} + (a_{\max} - a_{\min})\,\sigma(\kappa(x - x_{\text{crit}}))$ monotone increasing. Let $\bar{a} = \mathbb{E}_q[g(X)]$ and $F_q(x) = \Pr_q(X \le x)$.*

1. *The frequency score $s^*(x) = g(x)$ (equivalently $s^*(x) = x$) is the Bayes-optimal pre-inference ranking rule: among all measurable functions $s(x)$ of the query, it maximizes AUROC. The optimal value is*

$$\text{AUROC}^* \;=\; \frac{1}{2} + \frac{\text{Cov}_q\big(g(X),\, F_q(X)\big)}{\bar{a}(1-\bar{a})}. \tag{22}$$

2. *Among all pre-inference predictors measurable with respect to entity frequency f, no predictor can exceed AUROC^*. The bound depends on κ through (22): AUROC^* is strictly increasing in κ, with*

$$\lim_{\kappa\to 0} \text{AUROC}^* = \tfrac{1}{2}, \qquad \lim_{\kappa\to\infty} \text{AUROC}^* = \frac{1}{2} + \frac{1}{2}\frac{\Pr_q(X < x_{\text{crit}})\Pr_q(X > x_{\text{crit}})(a_{\max} - a_{\min})}{\bar{a}(1-\bar{a})}, \tag{23}$$

   so shallow cliffs ($\kappa \approx 0$) impose a hard discrimination ceiling that approaches chance, and steep cliffs ($\kappa \to \infty$) approach a theoretical maximum determined by the intrinsic noise of the capacity limits ($a_{\min}$ and $a_{\max}$).

3. *For q uniform on $[0, X]$ and $\kappa X \gg 1$:*

$$\text{AUROC}^* \;=\; \frac{1}{2} + \frac{\frac{1}{X^2}\int_0^X x\, g(x)\, dx \;-\; \bar{a}/2}{\bar{a}(1-\bar{a})}. \tag{24}$$

Proof. For part (i): $\text{AUROC}(s) = P(s(X_+) > s(X_-))$ where X_+ and X_- are drawn from $p(x \mid y=1) \propto g(x)q(x)$ and $p(x \mid y=0) \propto (1-g(x))q(x)$ respectively. By the Neyman–Pearson lemma, the likelihood ratio $g(x)/(1-g(x))$ is the optimal test statistic for separating X_+ from X_-. Since $g(x)$ is strictly increasing in x and $h(t) = t/(1-t)$ is strictly increasing on $(0,1)$, the composite $g(x)/(1-g(x))$ is strictly increasing in x. Since AUROC depends only on the ranking induced by s (not its values), any strictly monotone transformation of an optimal statistic is also optimal. Because $x \mapsto g(x)/(1-g(x))$ is strictly increasing, x induces the same ranking as the likelihood ratio, so the frequency score $s^*(x) = x$ is an equivalent optimal statistic. For the closed form, write $\text{AUROC}^* = P(X_+ > X_-)$ and substitute the normalized class-conditional densities $p(x \mid y=1) = g(x)q(x)/\bar{a}$ and $p(t \mid y=0) = (1-g(t))q(t)/(1-\bar{a})$; by the Wilcoxon–Mann–Whitney identity,

$$\begin{aligned}
\bar{a}(1-\bar{a}) \cdot \text{AUROC}^* &= \iint_{x>t} g(x)(1-g(t))\, q(x)q(t)\, dx\, dt, \\
&= \iint_{x>t} g(x)\, q(x)q(t)\, dx\, dt - \iint_{x>t} g(x)g(t)\, q(x)q(t)\, dx\, dt.
\end{aligned} \tag{25}$$

Integrating out t first gives $\int g(x)\, F_q(x)\, q(x)\, dx = \mathbb{E}_q[g(X)F_q(X)]$. The integrand $g(x)g(t)q(x)q(t)$ is symmetric in (x,t), so the integral over the half-plane $\{x > t\}$ equals exactly half the integral over the full plane, giving $B = \bar{a}^2/2$. Therefore:

$$\bar{a}(1-\bar{a}) \cdot \text{AUROC}^* = \mathbb{E}_q[g(X)F_q(X)] - \frac{\bar{a}^2}{2}.$$

Since $\mathbb{E}_q[F_q(X)] = \frac{1}{2}$ (the expected value of a CDF under its own distribution), we have $\mathbb{E}_q[g(X)F_q(X)] = \mathrm{Cov}_q(g(X), F_q(X)) + \bar{a} \cdot \frac{1}{2}$. Substituting:

$$\bar{a}(1-\bar{a}) \cdot \mathrm{AUROC}^* = \mathrm{Cov}_q(g(X), F_q(X)) + \frac{\bar{a}}{2} - \frac{\bar{a}^2}{2} = \mathrm{Cov}_q(g(X), F_q(X)) + \frac{\bar{a}(1-\bar{a})}{2},$$

yielding (22) after dividing by $\bar{a}(1-\bar{a})$. For part (ii): the converse is immediate from Neyman–Pearson optimality of the likelihood ratio. For strict monotonicity in κ, we use a first-order stochastic dominance (FOSD) argument. Fix $\kappa_2 > \kappa_1 > 0$. Since $\sigma(\kappa u)$ is increasing in κ for $u > 0$ and decreasing for $u < 0$, the accuracy functions satisfy

$$g_{\kappa_2}(x) > g_{\kappa_1}(x) \;\text{ for } x > x_{\mathrm{crit}}, \qquad g_{\kappa_2}(x) < g_{\kappa_1}(x) \;\text{ for } x < x_{\mathrm{crit}},$$

so $g_{\kappa_2}(x)/g_{\kappa_1}(x)$ crosses 1 from below exactly once at x_{crit}. Consider the positive-class densities $p_{+,\kappa}(x) \propto g_\kappa(x)\, q(x)$ with normalizing constant $\bar{a}_\kappa$. Their ratio is

$$\frac{p_{+,\kappa_2}(x)}{p_{+,\kappa_1}(x)} = \frac{g_{\kappa_2}(x)}{g_{\kappa_1}(x)} \cdot \frac{\bar{a}_{\kappa_1}}{\bar{a}_{\kappa_2}}.$$

Since $\bar{a}_{\kappa_1}/\bar{a}_{\kappa_2}$ is a positive constant, multiplying by it preserves the single-crossing property: the ratio $p_{+,\kappa_2}/p_{+,\kappa_1}$ still crosses 1 exactly once from below (at a point near x_{crit}). By the standard single-crossing criterion for FOSD, p_{+,κ_2} FOSD-dominates p_{+,κ_1}. An identical argument with $1 - g_\kappa$ shows p_{-,κ_1} FOSD-dominates p_{-,κ_2} (the ratio $(1-g_{\kappa_2})/(1-g_{\kappa_1})$ crosses 1 from above exactly once at x_{crit}). Therefore

$$\mathrm{AUROC}^*(\kappa_2) = P(X_{+,\kappa_2} > X_{-,\kappa_2}) \geq P(X_{+,\kappa_1} > X_{-,\kappa_1}) = \mathrm{AUROC}^*(\kappa_1).$$

Strictness follows because $\Pr_q(X < x_{\mathrm{crit}}) \in (0,1)$ (the cliff lies strictly inside the support of q), so the FOSD inequalities are strict on sets of positive q-measure.

The limits follow by $g_\kappa \to \bar{a}$ uniformly as $\kappa \to 0$ (so $\mathrm{AUROC}^* \to \frac{1}{2}$) and $g_\kappa \to a_{\min} + (a_{\max} - a_{\min})\,\mathbf{1}[x > x_{\mathrm{crit}}]$ as $\kappa \to \infty$. For the infinite-steepness limit, $g(x)$ becomes a step function. Let $P_< = \Pr_q(X < x_{\mathrm{crit}})$ and $P_> = \Pr_q(X > x_{\mathrm{crit}})$. Evaluating the covariance integral directly yields:

$$\begin{aligned}\mathbb{E}_q[g(X)F_q(X)] &= \int_{x<x_{\mathrm{crit}}} a_{\min} F_q(x) q(x) dx + \int_{x>x_{\mathrm{crit}}} a_{\max} F_q(x) q(x) dx \\ &= \frac{1}{2} a_{\min} P_<^2 + \frac{1}{2} a_{\max}(1 - P_<^2).\end{aligned}$$

Subtracting $\bar{a}/2 = \frac{1}{2}(a_{\min} P_< + a_{\max} P_>)$ from this expectation gives the covariance:

$$\mathrm{Cov}_q(g, F_q) = \frac{1}{2} P_< P_> (a_{\max} - a_{\min}).$$

Plugging this exact covariance into (22) immediately yields the stated upper bound limit.

□

Remark 7. *The closed-form ceiling* (22) *expresses the fundamental limit of frequency-based discrimination directly in terms of κ and q, with strict monotonicity in κ established via a single-crossing argument on the positive- and negative-class densities. This makes precise how much pre-inference discriminative power is available from entity frequency alone, prior to any model invocation. The* 0.026 *gap between* AUROC* *and* `Freq+relation` *quantifies the incremental information that relation structure contributes beyond raw frequency.*

To verify the results numerically; we note that for Mistral-7B ($\kappa = 4.87$, $x_{\mathrm{crit}} = 4.10$, $a_{\min} = 0.170$, $a_{\max} = 0.820$) under uniform q on $[0, 7]$, numerical integration of (24) gives $\mathrm{AUROC}^* = 0.816$. The empirical frequency sigmoid achieves AUROC $= 0.810$ (Table 4), attaining 99.2% of the theoretical maximum for any pre-inference predictor. The residual gap of 0.006 is attributable to estimation noise in $\hat{\kappa}$ and the stratified (rather than natural Zipfian) sampling distribution. For Falcon-7B ($\kappa = 2.10$), the bound gives $\mathrm{AUROC}^* = 0.837$, reflecting the harder discrimination problem imposed by a shallower cliff.

Before turning to extensive experiments, we connect this routing formulation to error exponents and the rate–distortion perspective on sharp generalization transitions.

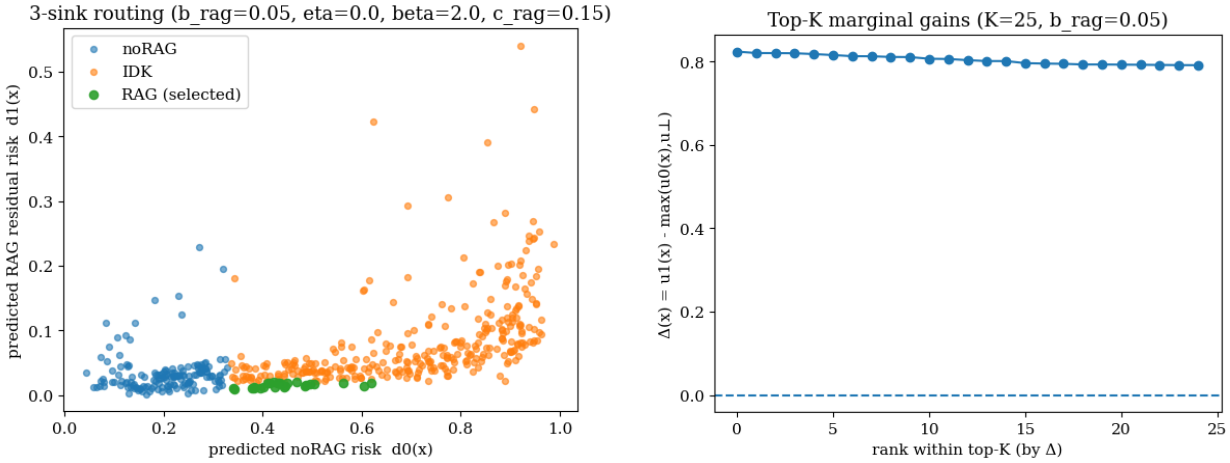

Figure 1: **Routing geometry induced by the knowledge cliff. Left (Routing geometry in the $(\widehat{d}_0, \widehat{d}_1)$ plane):** predicted noRAG distortion $\widehat{d}_{0,k}$ from (19) versus predicted RAG residual distortion $\widehat{d}_{1,k}$ from (20), colored by the resulting three-action decision ($b_{\mathrm{RAG}} = 0.05$, $\beta = 2.0$, $c_{\mathrm{RAG}} = 0.15$). RAG is selected only for queries with high noRAG risk and low residual RAG risk; IDK (abstain) dominates for high-risk queries where RAG residual risk is also elevated. **Right (Top-K marginal gains Δ_k cap check):** sorted marginal gains Δ_k (as defined in Proposition 8) among the top $K = 25$ upgrades ($b_{\mathrm{RAG}} = 0.05$); $\min \Delta_k > 0$ certifies the retrieval cap is binding and all upgrades are utility-improving.

5 Reliability Exponents in the Stored Regime

Rate–distortion theory identifies *where* a distortion target becomes achievable, which in our setting is the critical exposure f_{crit}, but it does not quantify *how sharply* error probability decreases once that boundary is crossed. We develop an *exposure-based reliability exponent* for the LLM setting, and show that the sigmoid model of Section 3 pins its value to the already-estimated steepness parameter κ.

Let $x := \log f$ and $x_{\mathrm{crit}} := \log f_{\mathrm{crit}}$. Define the binary error indicator $e \in \{0, 1\}$ ($e = 1$ for an incorrect answer without retrieval). We focus on the *head regime* $x \geq x_{\mathrm{crit}}$, where the model is in principle capable of storing the relevant fact.

Definition 10 (Head-region reliability exponent)**.** *A* head-region reliability exponent *is a scalar $\eta \geq 0$ such that*

$$p_{\mathrm{exc}}(x) \;\approx\; \exp\big(\alpha - \eta(x - x_{\mathrm{crit}})\big), \qquad x \geq x_{\mathrm{crit}}, \tag{26}$$

for some intercept α; equivalently, $\log p_{\mathrm{exc}}(x)$ is linear in x above the cliff with slope $-\eta$.

Suppose each training exposure produces an approximately independent encoding opportunity, and let $m(x) = m_0 + c(x - x_{\mathrm{crit}})$ be the locally linear opportunity count for $x \geq x_{\mathrm{crit}}$. If each opportunity places the fact into a retrievable state with probability $\rho \in (0, 1)$, a union bound gives

$$p_{\mathrm{err}}(x) \;\leq\; \epsilon_0 + (1 - \rho)^{m(x)} \;=\; \epsilon_0 + \exp\big(\alpha - \eta(x - x_{\mathrm{crit}})\big), \tag{27}$$

with $\alpha := m_0 \log(1-\rho)$ and $\eta := -c \log(1-\rho) > 0$. This establishes exponential post-cliff decay for some $\eta > 0$ under any repeated-opportunity mechanism, without pinning its value. The sigmoid model of Proposition 4 pins the exponent exactly. For $x \geq x_{\mathrm{crit}}$, the logistic tail satisfies $1 - \sigma(\kappa(x - x_{\mathrm{crit}})) \sim \exp(-\kappa(x - x_{\mathrm{crit}}))$, so $p_{\mathrm{exc}}(x) \approx (a_{\max} - a_{\min}) \exp(-\kappa(x - x_{\mathrm{crit}}))$, giving $\eta = \kappa$. The encoding-opportunity argument guarantees exponential decay; the sigmoid model identifies κ as the specific rate. This yields a testable prediction: η estimated directly from head-region errors should be consistent with κ from the full sigmoid fit. Because the sigmoid uses all 500 observations while the direct MLE uses only the $n = 186$ head-region subset, κ provides the tighter estimate; the direct MLE serves as an independent consistency check.

Token-frequency tradeoff at the cliff boundary. An analogous rate variable operates at inference time. Recent work demonstrates that prompt repetition improves accuracy in non-reasoning LLMs by increasing prefill work (Leviathan et al., 2025), effectively providing more activation energy to extract weakly-stored facts. Extending our opportunity model to include an inference-time token budget $t \geq 0$ (extra prompt tokens beyond a base template), we can model the extraction margin for entities near the cliff boundary as locally additive:

$$m(x,t) \approx m_0 + c(x - x_{\text{crit}}) + s(x)\, t, \qquad s(x) > 0 \text{ for } x \approx x_{\text{crit}}.$$

Assuming the fact was allocated non-zero capacity during training, the same union bound yields:

$$p_{\text{err}}(x,t) \;\leq\; \epsilon_0 + \exp\big(\alpha - \eta(x - x_{\text{crit}}) - \eta_{\text{tok}}\, t\big), \qquad \eta_{\text{tok}} := -s(x)\log(1-\rho) > 0. \tag{28}$$

Equivalently, an inference token budget t shifts the effective cliff left by $\Delta x_{\text{crit}} = (\eta_{\text{tok}}/\eta)\, t$. Within the transition regime, extra context acts as a direct substitute for training frequency, making weakly-exposed entities behave as if they had higher training exposure.

Remark 8 (Water-filling for token allocation)**.** *Treating $t_k \geq 0$ as an allocable inference-time resource with budget $\sum_k q_k t_k \leq T_{\text{tok}}$, we can approximate the error using the strictly convex exponential tail from* (28)*: $d_{0,k}(t_k) \approx \exp\big(\alpha - \eta(\log f_k - x_{\text{crit}}) - \eta_{\text{tok}} t_k\big)$. The resulting expected loss minimization is a strictly convex optimization problem. The KKT conditions yield a water-filling allocation*

$$t_k^* = \left[\frac{1}{\eta_{\text{tok}}}\log\Big(\frac{(1+\beta)\eta_{\text{tok}}}{\lambda_{\text{tok}}}\Big) + \frac{\alpha}{\eta_{\text{tok}}} - \frac{\eta}{\eta_{\text{tok}}}\big(\log f_k - \log f_{\text{crit}}\big)\right]_+,$$

where λ_{tok} is the Lagrange multiplier chosen so the budget binds. This is the inference-time analogue of the RD water-filling in Section 4: the shadow price λ_{tok} acts as the RD multiplier, concentrating extra context on lower-frequency queries where the marginal reduction in exponential error is steepest.

Primary estimate: $\eta = \kappa$ from the full sigmoid fit. The argument above predicts $\eta = \kappa$. Since κ is estimated from all 500 observations via the full sigmoid fit (Experiment 1, Table 1), it inherits the full statistical power of that fit. For Mistral-7B on PopQA, the sigmoid yields $\kappa = 4.87$ ($R_w^2 = 0.969$), giving a primary estimate $\hat{\eta} = 4.87$. Interpreted in the exponential model, each additional decade of log-exposure above the cliff multiplies head-region excess error by $\exp(-\kappa) \approx 0.008$—a rapid reliability improvement consistent with the steep empirical transition visible in Figure 1. As an independent check, we estimate (α, η) directly from unbinned Bernoulli likelihoods on head-region examples ($x \geq x_{\text{crit}}$) only, without assuming the logistic form. Let $e_i \in \{0,1\}$ be the empirical error indicator for query i, and p_i be the modeled probability of error:

$$\min_{\alpha,\,\eta \geq 0} \; -\sum_{i:\, x_i \geq x_{\text{crit}}} \big[e_i \log p_i + (1 - e_i)\log(1 - p_i)\big], \quad p_i := \min\big(1, \exp(\alpha - \eta(x_i - x_{\text{crit}}))\big).$$

On the $n = 199$ head-region examples ($x \geq x_{\text{crit}} = 4.138$, i.e. $f_{\text{crit}} \approx 13{,}734$), this yields $\hat{\eta} \approx 0.635$ (bootstrap median 0.633, 5–95% interval $[0.065, 1.195]$). This interval rejects the global estimate of $\kappa = 4.87$, indicating a departure from pure exponential decay deep in the head region. This discrepancy is mathematically expected: empirical error rates do not asymptote strictly to zero due to an irreducible error floor ($\epsilon_0 > 0$) caused by dataset noise, prompt sensitivity, or catastrophic forgetting. Because the direct unbinned MLE assumes the error probability vanishes entirely ($p_i \to 0$), any residual errors deep in the head region exert outsized leverage on the likelihood. To accommodate these non-zero tail errors, the MLE is forced to severely bias the local estimate $\hat{\eta}$ downward, resulting in a much flatter apparent decay. The direct MLE thus confirms that the exponent is strictly positive, while simultaneously highlighting that real-world reliability is ultimately bounded by ϵ_0 rather than vanishing continuously.

Per-relation exponents. For query-dependent exponents we allow (α, η) to vary by relation type g using a hierarchical MAP prior: $\alpha_g \sim \mathcal{N}(\mu_\alpha, \lambda_\alpha^{-1})$, $\log \eta_g \sim \mathcal{N}(\mu_\eta, \lambda_\eta^{-1})$, with partial pooling toward a shared mean. This is equivalent to using relation-specific κ_g values from the relation-specific sigmoid fits as the natural primary estimates, with the direct MAP providing a consistency check under partial pooling. Estimated relation exponents range from $\eta \approx 0.55$ (`religion`) to $\eta \approx 2.38$ (`screenwriter`) at the reference pooling

strength ($\lambda_\eta = 2$), and the ranking is robust across pooling strengths ($\lambda_\eta \in \{0.2, 0.5, 1, 2, 5\}$; Spearman $\rho \geq 0.932$ relative to $\lambda_\eta = 2$). Relation structure thus affects both *where* the cliff occurs (via f_{crit}) and *how rapidly* reliability improves above it (via $\eta = \kappa_g$), providing a principled bridge between RD-style boundary predictions and the practitioner-relevant question of how quickly errors diminish with additional exposure.

The bootstrap interval on the direct head-region MLE ($\hat{\eta} \in [0.065, 1.195]$ at 90%) raises a natural question: how many labeled examples are fundamentally required to localize the cliff to a given precision? The following analysis gives a tight answer in terms of κ and the query distribution.

5.1 Sample Complexity of Cliff Localization

A practical question follows from the estimation framework: how many labeled examples are required to localize x_{crit} to a given precision ε? This determines the minimum calibration cost for deploying the pre-inference risk score in a new domain, and explains the empirical finding that ECE < 0.10 is achievable with $\sim$75 calibration samples.

Proposition 11 (Sample complexity of cliff localization)**.** *Let $x_1, \ldots, x_n \sim p(x)$ be i.i.d. log-frequencies on $[0, X]$ with binary labels $y_i \mid x_i \sim \text{Bernoulli}(\sigma(\kappa(x_i - x_{\text{crit}})))$. The maximum likelihood estimator $\hat{x}_{\text{crit}}$ possesses the following Fisher information properties:*

(i) ***Pointwise Information.*** *Letting σ' denote the derivative of the logistic sigmoid (where $\sigma'(z) = \sigma(z)(1 - \sigma(z))$), a single observation at x contributes $\mathcal{I}(x_{\text{crit}}; x) = \kappa^2\, \sigma'\big(\kappa(x - x_{\text{crit}})\big)$, which peaks at $\kappa^2/4$ exactly at the cliff and decays exponentially away from it.*

(ii) ***Zipf-1 Sampling ($\alpha = 1$).*** *For $p(x) = 1/X$, the total information is $\mathcal{I}_n(x_{\text{crit}}) \approx n\kappa/X$ (assuming $\kappa X \gg 1$). To achieve standard error $\text{SE}(\hat{x}_{\text{crit}}) \leq \varepsilon$, the required sample size is $n \geq X z^2_{1-\delta/2}/(\kappa\varepsilon^2)$. Complexity is independent of x_{crit}.*

(iii) ***General Zipfian ($\alpha \neq 1$).*** *For $p(x) \propto 10^{(1-\alpha)x}$ with normalization constant Z_α, the total information and corresponding sample complexity bound are:*

$$\mathcal{I}_n(x_{\text{crit}}) \approx \frac{n\kappa}{Z_\alpha} f_{\text{crit}}^{1-\alpha} I(\kappa, \alpha), \qquad n \;\geq\; \frac{z^2_{1-\delta/2}\, Z_\alpha}{\kappa\, \varepsilon^2\, f_{\text{crit}}^{1-\alpha}\, I(\kappa, \alpha)} \tag{29}$$

where $I(\kappa, \alpha) = \pi t / \sin(\pi t)$ for $t = (1 - \alpha)\ln(10)/\kappa \in (-1, 1)$. For steeper Zipfian distributions ($\alpha > 1$), $f_{\text{crit}}^{1-\alpha}$ decreases with f_{crit}, meaning tail cliffs require exponentially more samples *to localize because natural sampling heavily underrepresents the informative transition window.*

Proof. For part (i), the log-likelihood of a single observation is $\ell(x, y) = y \log \sigma(u) + (1 - y) \log(1 - \sigma(u))$ where $u = \kappa(x - x_{\text{crit}})$. Using the identity $\sigma'(u) = \sigma(u)(1 - \sigma(u))$, the score function is $\partial\ell/\partial x_{\text{crit}} = -\kappa \frac{\partial \ell}{\partial u} = \kappa(\sigma(u) - y)$. Since $\mathbb{E}[y] = \sigma(u)$, the Fisher information is $\mathbb{E}[(\partial\ell/\partial x_{\text{crit}})^2] = \kappa^2 \text{Var}(y) = \kappa^2 \sigma(u)(1 - \sigma(u)) = \kappa^2 \sigma'(u)$, confirming part (i). For part (ii), integrate against $p(x) = 1/X$:

$$\mathcal{I}_n(x_{\text{crit}}) = \frac{n\kappa^2}{X} \int_0^X \sigma'(\kappa(x - x_{\text{crit}}))\, dx = \frac{n\kappa}{X} \int_{-\kappa x_{\text{crit}}}^{\kappa(X - x_{\text{crit}})} \sigma'(u)\, du.$$

For x_{crit} sufficiently far from the boundaries ($\kappa x_{\text{crit}} \gg 1$ and $\kappa(X - x_{\text{crit}}) \gg 1$), the integral of the PDF $\sigma'(u)$ approaches 1, yielding $\mathcal{I}_n \approx n\kappa/X$. The sample bound follows from the Cramér–Rao bound $\text{Var}(\hat{x}_{\text{crit}}) \geq \mathcal{I}_n^{-1}$ and asymptotic normality.

For part (iii), integrate against $p(x) = 10^{(1-\alpha)x}/Z_\alpha$. Under the same substitution $u = \kappa(x - x_{\text{crit}})$, we expand $10^{(1-\alpha)x} = 10^{(1-\alpha)(x_{\text{crit}} + u/\kappa)} = f_{\text{crit}}^{1-\alpha} e^{tu}$, where $t = (1 - \alpha)\ln(10)/\kappa$. Extending the integration limits to $\mathbb{R}$ gives:

$$\mathcal{I}_n(x_{\text{crit}}) \approx \frac{n\kappa \cdot f_{\text{crit}}^{1-\alpha}}{Z_\alpha} \int_{-\infty}^{\infty} \sigma'(u)\, e^{tu}\, du.$$

The integral is exactly the moment-generating function of the standard logistic distribution, which evaluates to $\pi t / \sin(\pi t)$ for $|t| < 1$. Continuity as $t \to 0$ ($\alpha \to 1$) smoothly recovers part (ii). □

Numerical validation. For Mistral-7B on PopQA ($\kappa = 4.87$, $X = 7$), achieving a target standard error of $\varepsilon = 0.06$ decades requires $n \approx X/(\kappa\,\varepsilon^2) \approx 7/(4.87 \times 0.06^2) \approx 399$ under natural Zipfian sampling ($\alpha = 1$). The empirical SE of $\hat{x}_{\text{crit}} = 0.06$ achieved at $n = 500$ is consistent with this theoretical requirement. The modest gap is attributable to the stratified sampling used in our dataset, which oversamples the extreme tail relative to a natural Zipfian draw; this spends a portion of the sample budget in regions where the Fisher information is exponentially suppressed, thus requiring slightly more total samples to accumulate the necessary information. For the 75-sample ECE result, the theoretical standard error is $\text{SE}(\hat{x}_{\text{crit}}) \approx \sqrt{X/(75\kappa)} = \sqrt{7/365.25} \approx 0.138$ decades under natural sampling. Using a first-order approximation with the peak derivative of the accuracy function (which occurs exactly at the cliff, $a'(x) = (a_{\max} - a_{\min})\kappa/4$), we can bound the maximum expected probability deviation as $\approx (a_{\max} - a_{\min}) \cdot \kappa \cdot \text{SE}/4 \approx 0.65 \times 4.87 \times 0.138/4 \approx 0.109$. Because this bound represents the worst-case deviation driven by the peak steepness exactly at the cliff boundary, the average deviation over the full query distribution must naturally be lower. This is fully consistent with the observed empirical ECE < 0.10.

Remark 9 (Cliff localization is κ-easy)**.** *The $1/\kappa$ dependence in (iii) means that steep cliffs are disproportionately easy to localize: doubling κ halves the required sample size. This is because a sharp cliff concentrates Fisher information in a narrow window around x_{crit}, making each nearby sample highly informative. Conversely, shallow cliffs (low κ, as in Falcon-7B with $\kappa = 2.10$) require approximately $4.87/2.10 \approx 2.3\times$ more labeled examples to localize to the same precision.*

6 Experiments

6.1 Setup

Dataset. We use PopQA (Mallen et al., 2023), a factual QA benchmark where each question concerns a single entity with known Wikipedia page view count, which we use as an exposure proxy for entity frequency in the training corpus. We evaluate on 500 stratified queries spanning four frequency bins: <1K, 1K–10K, 10K–100K, and >100K monthly page views. Our theory requires this proxy to be approximately monotone in true training exposure; we validate this assumption in Section 6.2 using multiple independent pre-inference proxies and a negative control.

Models and Protocol We evaluate three model families: Mistral-7B-Instruct-v0.3, Falcon-7B-Instruct, and Qwen-2.5-7B-Instruct. Each query is evaluated in two modes: direct generation (no RAG) and retrieval-augmented generation. We record the generated answer, correctness (exact match against ground truth aliases), mean token probability (confidence), and token-level entropy.

6.2 Proxy Triangulation and Negative Controls

A key experimental assumption is that Wikipedia page views are approximately monotone in true training exposure. To stress-test this, we compare multiple *pre-inference* proxies and a negative control on the same evaluation set. Specifically, we evaluate: (i) our primary log-frequency risk $\hat{d}_k$ from Equation (5), (ii) a relation-centered log-frequency proxy (subtracting the per-relation median $\log f$ to control for cross-relation difficulty), (iii) an independent structural proxy based on ground-truth alias set size (measuring answer ambiguity), and (iv) a shuffled-frequency negative control.

Figure 2 plots the calibration-by-decile curves (mean predicted risk vs. empirical error rate) for each proxy. The shuffled control (red) completely breaks the signal; sorting by random values destroys alignment with correctness, collapsing the deciles into a non-discriminative cluster around the base error rate (AUC=0.45). In contrast, the genuine frequency proxies closely track the ideal $y = x$ calibration line across a wide dynamic range. Furthermore, the strong performance of the relation-centered proxy (AUC=0.77) confirms that the frequency signal is not merely an artifact of certain relations being universally easier or more popular. Together, these results demonstrate that our modeled risk probabilities accurately reflect empirical error rates, robustly validating page views as a reliable proxy for training exposure.

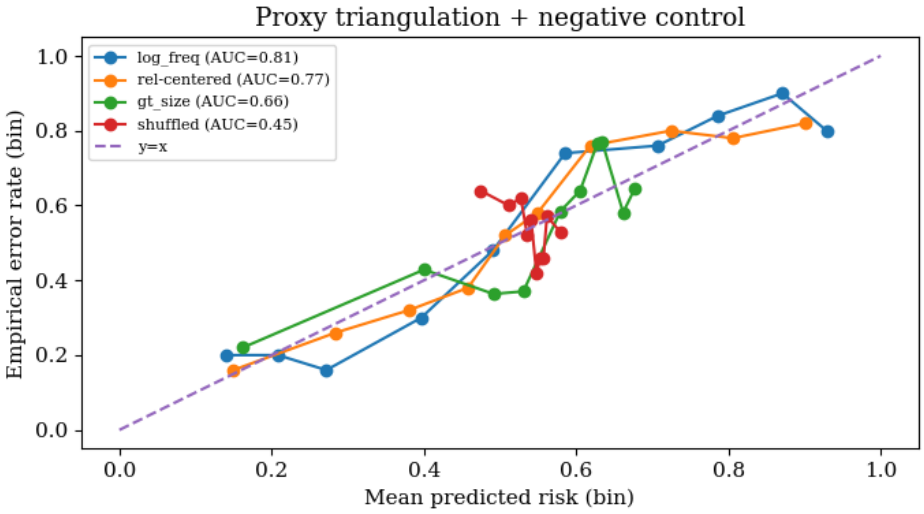

Figure 2: **Proxy triangulation and negative control.** Calibration-by-decile curves for pre-inference risk proxies. The shuffled negative control collapses into a tight cluster with no discriminative power. In contrast, the exposure-based proxies span a wide dynamic range and closely track the ideal $y = x$ calibration line, confirming their predictive reliability.

6.3 Experiment 1: The Knowledge Cliff

We fit the sigmoid (Equation (5)) to binned accuracy-vs-frequency data using weighted nonlinear least squares with per-bin weights proportional to bin counts (i.e., $\sigma_i = 1/\sqrt{n_i}$). Table 1 reports the fitted parameters.

Table 1: Sigmoid fit parameters across models. All fits achieve weighted $R^2 > 0.89$, confirming the phase transition shape. f_{crit} varies with training data: Mistral (better tail coverage) has a lower cliff.

Model	**Overall acc.**	$\boldsymbol{f_{\text{crit}}}$	$\boldsymbol{\log(f_{\text{crit}})}$	$\boldsymbol{\kappa}$	$\boldsymbol{R^2_w}$
Mistral-7B	0.462	12,726	4.10 ± 0.06	4.87	0.969
Falcon-7B	0.308	66,009	4.82 ± 0.38	2.10	0.892
Qwen-2.5-7B	0.326	50,843	4.71 ± 0.18	3.02	0.920

The key observation from Table 1 is the clustering: Falcon and Qwen share similar cliff locations ($f_{\text{crit}} \approx$ 51–66K), while Mistral's cliff is substantially lower ($f_{\text{crit}} \approx$ 13K). This is consistent with Proposition 5: the cliff reflects effective pretraining coverage, and in our evaluation Mistral shows higher accuracy in the 1K–10K transition range (0.248 vs. ~ 0.14 for Falcon/Qwen Table 2), consistent with better coverage of medium-frequency entities in the underlying training data.

Table 2: Per-stratum accuracy across models. All models show the same pattern: near-random performance in the tail (<1K) rising to >70% in the head (>100K). Mistral's advantage concentrates in the 1K–100K transition zone.

Stratum	**Mistral-7B**	**Falcon-7B**	**Qwen-2.5-7B**
<1K	0.144	0.104	0.136
1K–10K	0.248	0.136	0.136
10K–100K	0.624	0.336	0.312
>100K	0.832	0.656	0.720
Overall	0.462	0.308	0.326

6.4 Experiment 2: Fine-Tuning Cannot Resolve the Knowledge Cliff

If the knowledge cliff is fundamentally constrained by pretraining capacity allocation (Proposition 5), then lightweight post-training interventions should not be able to meaningfully improve f_{crit} (i.e., shift the cliff to the left). To test this, we fine-tune Mistral-7B with QLoRA on a ∼9,987-question training split of PopQA using three weighting schemes: uniform, square-root-inverse frequency, and inverse frequency (a theoretically motivated reweighting from the R-D framework). Table 3 reports the results on a held-out evaluation split (4,281 questions); note that the base model $f_{\text{crit}} = 24{,}802$ differs from Table 1 due to this specific larger split.

Table 3: Fine-tuning reshapes individual predictions but blurs the cliff boundary. No weighting scheme shifts the cliff to the left (improves recall). Instead, FT increases the point estimate of f_{crit} and noticeably degrades the sigmoid goodness-of-fit (R^2_w), indicating a smearing of the pretraining capacity boundary.

Weighting	$\boldsymbol{f_{\text{crit}}}$	$\boldsymbol{\log(f_{\text{crit}}) \pm \text{SE}}$	$\boldsymbol{R^2_w}$
Base (no FT)	24,802	4.39 ± 0.17	0.938
Uniform FT	45,999	4.66 ± 0.33	0.864
$\sqrt{\text{inv}}$-freq FT	63,620	4.80 ± 0.32	0.846
Inv-freq FT	48,709	4.69 ± 0.23	0.842

This result confirms that post-training cannot reorganize the underlying capacity allocation. While fine-tuning does change individual predictions (McNemar $p = 0.0072$), the changes reflect format adaptation and catastrophic forgetting rather than genuine knowledge acquisition. For instance, Inverse-frequency FT fixes 266 tail predictions (<1K) by teaching the model the desired QA format, but breaks 143 confident head predictions (10K–100K). Consequently, rather than improving f_{crit}, fine-tuning shifts the point estimates to the right (requiring higher frequency for recall) and introduces substantially increased uncertainty, rendering the shifts statistically indistinguishable from the base cliff ($z = 1.03$, $p > 0.05$). Furthermore, the drop in R^2_w across all FT variants indicates that the clean, sharp boundary established during pretraining becomes "smeared" by the intervention. Proposition 5 is thus empirically supported: the knowledge cliff is a rigid artifact of pretraining that resists downstream correction.

6.5 Experiment 3: Hallucination Prediction via Frequency

We evaluate the sigmoid risk score $\hat{d}_k = 1 - \sigma\big(\kappa(\log f_k - \log f_{\text{crit}})\big)$ as a hallucination detector on the **500-query PopQA / Mistral-7B evaluation set**. Here, the positive class is *hallucination* (incorrect direct-generation answer). We compare four scoring methods: (a) *Model confidence (raw)*: mean token probability converted to hallucination risk via 1 − confidence (requires inference); (b) *Frequency sigmoid*: our pre-inference risk score (no inference); (c) *OT-inspired joint score*: a regime-aware, theory-inspired combination of frequency and confidence that uses confidence primarily above the estimated cliff where it is more informative; (d) *Freq + conf (combined)*: 5-fold out-of-fold (OOF) logistic regression using frequency, confidence, minimum token probability, and entropy as features (trained to predict correctness and reported as hallucination risk via $1 - \hat{p}$).

Results. Table 4 reports global detection metrics on this 500-query evaluation set. The frequency sigmoid achieves AUROC = 0.810, outperforming raw model confidence (AUROC = 0.772) despite requiring *zero inference*. The OT-inspired joint score improves the best single-signal discrimination to AUROC = 0.835, consistent with the idea that confidence is most useful in the stored regime (where factual recall is plausible) and less informative in the erasure regime. The strongest overall detector is the combined frequency+confidence model (AUROC = 0.856, AUPRC = 0.873, Brier = 0.153), indicating that frequency and inference-time uncertainty provide complementary signals.

Calibration baselines. Raw confidence is poorly calibrated (ECE = 0.414, Brier = 0.384). Post-hoc methods (temperature scaling, Platt scaling, isotonic regression on three confidence features) improve ECE substantially (best ECE = 0.021 via Platt-scaled minimum token probability) but reduce discrimination,

Table 4: Hallucination detection metrics on the 500-query PopQA/Mistral-7B evaluation set. The frequency sigmoid is a no-additional-model-execution pre-inference score and achieves the best ECE and Brier among single-feature baselines shown here, while the OT-inspired joint score achieves the best AUROC among the single-signal methods.

Method	**Cost**	**AUROC**	**AUPRC**	**ECE**	**Brier**
Model confidence (raw)	Inference	0.772	0.785	0.414	0.384
Freq sigmoid	Zero	0.810	0.809	0.041	0.169
OT-inspired joint score	Inference	0.835	0.812	–	–
Freq + conf (combined)	Inference	0.856	0.873	0.033	0.153

illustrating a calibration–discrimination tradeoff. Despite these gains, frequency remains the strongest single-feature baseline on both AUROC (0.810 vs. 0.775 for the best post-hoc confidence variant) and Brier (0.169 vs. 0.192–0.200). Frequency and confidence are complementary: combining them yields AUROC = 0.856, Brier = 0.153 (Figure 3, bottom-right).

Aggregation effects. Within strata, confidence is the stronger discriminator (e.g., >100K: 0.769); frequency has limited within-stratum power by design. Globally, frequency dominates cross-stratum ranking (AUROC = 0.810), since exposure determines whether a fact is stored at all. This is a Simpson's-paradox-like effect: frequency captures coverage across strata, confidence captures generation-time variability within them.

Transferability and sample efficiency. To test whether the fitted knowledge-cliff model generalizes beyond a single split, we evaluate transfer under resampling, distribution shift across frequency strata, and limited calibration data (Figure 5). Across 100 random 50/50 train/test splits, the inferred cliff location remains reasonably stable (f_{crit} mean = 12,988, CV = 0.23), and the frequency sigmoid achieves substantially lower out-of-sample ECE than raw confidence (0.069 ± 0.021 vs. 0.415 ± 0.022; ratio 6.0×). Leave-one-stratum-out (LOSO) transfer shows that a fit from three strata predicts the held-out stratum's hallucination rate with mean |error| = 0.060, with the largest deviations in the higher-frequency strata (>100K: 0.083) where the curve is closer to saturation. A sample-efficiency analysis indicates that calibration quality improves rapidly with relatively few labeled examples: the frequency sigmoid's held-out ECE drops sharply with calibration set size and remains well below the raw-confidence baseline, while uncertainty decreases as more labels are added. Notably, the confidence gap between correct and incorrect answers is nearly constant across strata ($\Delta \approx 0.075$ in our run), supporting the view that confidence primarily captures within-stratum discrimination, whereas frequency captures cross-stratum coverage.

Routing gains from pre-inference structure. The strongest *pre-inference* routing policy in our experiments uses `Freq + relation`. On the budget–utility frontier, at a 20% RAG budget, it achieves 0.610 overall accuracy. This is highly competitive with the oracle[1] upper bound (0.642) and strictly outperforms all purely confidence-based policies, including raw confidence (0.608) and Platt-calibrated confidence (0.600).

Crucially, while blending post-inference model confidence with our structural proxy (`Freq + conf combined`) yields a marginal improvement at the 20% budget mark (0.622 vs. 0.610), this gap closes at higher budgets. At a 40% budget, the purely pre-inference `Freq + relation` actually overtakes the combined score (0.762 vs. 0.760). Furthermore, to reach a target accuracy of 80%, all frequency-based policies require the exact same RAG budget (47%), compared to 54% for confidence-based routing and 72% for random routing.

Overall, these results heavily validate the threshold-routing view developed in Section 4. Because the `Freq + relation` proxy requires zero inference compute to calculate, it provides an exceptionally efficient ranking signal, demonstrating that the vast majority of routing utility can be captured before a single LLM token is generated. Figure 4 visualizes these frontiers.

[1]Oracle routing uses per-query counterfactual outcomes (noRAG and RAG correctness) available only offline; it is not implementable at runtime.

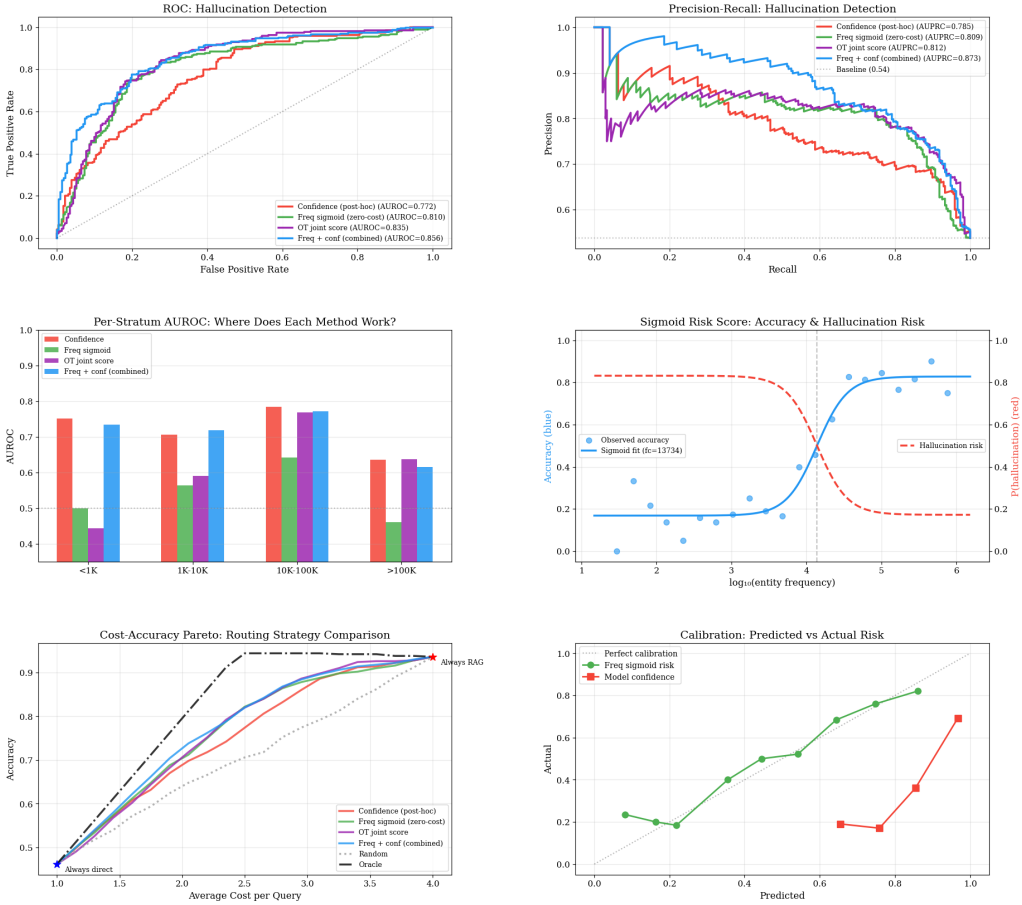

Figure 3: **Hallucination prediction results (Mistral-7B, 500-query PopQA).** Top: ROC and PR curves show the pre-inference Frequency sigmoid is highly effective without requiring an LLM forward pass, approaching the performance of expensive post-hoc calibrated confidence. Middle: Per-stratum AUROC and fitted knowledge cliff ($f_{\text{crit}} = 13{,}734$). This reveals a Simpson's-paradox aggregation: Frequency excels at essential *cross*-stratum ranking, while confidence is best *within* narrow bands. Bottom: Routing Pareto frontier and calibration. Combining Frequency and Confidence (the 'OT Joint' score) yields the best overall performance (AUROC 0.856), suggesting optimal routing uses cheap frequency signals first.

Unseen-relation generalization. Holding out entire relation types and evaluating on unseen relations (Table 5), the log-frequency sigmoid transfers strongly (held-out AUC = 0.844, Brier = 0.167), while the shuffled-frequency control collapses to near-chance (AUC = 0.442). The held-out AUROC exceeding in-relation AUROC for `log_freq_sigmoid` is not leakage: the frequency sigmoid has zero relation-specific parameters, so the reversal reflects sampling variance across strata — consistent with `struct_risk` showing the opposite pattern (0.867 → 0.811), the expected direction for a relation-aware model.

6.6 Experiment 4: Cross-Model Transfer

We evaluate cross-model transfer of the frequency-based sigmoid risk model on Falcon-7B and Qwen-2.5-7B using the same 500-query PopQA evaluation set as in the Mistral experiments. For each source model, we fit a sigmoid accuracy–frequency curve and convert it to a risk score; we then apply that fitted curve *without refitting* to a target model's frequency-stratified outcomes and measure calibration error (ECE). Table 6 reports the resulting transfer matrix.

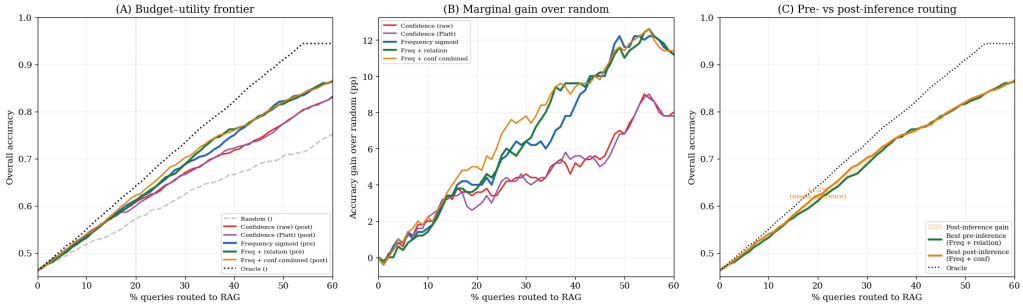

Figure 4: **Budget–utility frontier for retrieval routing.** Six policies are compared. For each RAG budget (x-axis, % of queries routed to retrieval), we sort queries by a risk score and route the highest-risk fraction; overall accuracy (y-axis) is computed using recorded `no_rag_correct` and `rag_correct`. (A) Full frontier: the best purely pre-inference policy (`Freq + relation`, green) is consistently competitive and often best, approaching the oracle frontier. (B) Marginal gain over random routing: frequency-based policies yield the largest improvements over random, with `Freq + conf combined` dominant at higher budgets. (C) Pre- vs. post-inference comparison: the shaded region indicates the additional gain from using confidence. At 20% budget, `Freq + relation` reaches 0.610 accuracy, compared to 0.622 for the strongest post-inference baseline shown here (`Freq + conf`). Overall, these results are consistent with threshold-based budgeted routing and suggest that frequency–structure features provide a strong ranking signal before inference.

Table 5: **Generalization to unseen relations.** Proxy calibration is trained on a subset of relation types and evaluated on held-out relations; i.e., calibration trained on in-relations only; held-out relations are never seen in training. Higher AUROC and lower Brier are better. The shuffled-frequency negative control collapses to near-chance.

Method	AUC (in-rel) ↑	Brier (in-rel) ↓	AUC (held-out rel) ↑	Brier (held-out rel) ↓
log_freq_sigmoid	0.7839	0.1886	**0.8444**	**0.1666**
struct_risk	**0.8667**	**0.1435**	0.8108	0.1701
1-conf	0.7792	0.2956	0.7613	0.4602
shuffled_log_freq	0.4503	0.2569	0.4419	0.2459

The transfer matrix reveals a clear block structure: Falcon ↔ Qwen transfer ECEs (0.029–0.039) are close to self-fit performance, whereas transfer between Mistral and either Falcon or Qwen degrades to ~ 0.154. This pattern closely tracks the fitted cliff locations: Falcon and Qwen have similar critical frequencies ($f_{\text{crit}} \approx$ 51K–66K), while Mistral's cliff is substantially lower ($f_{\text{crit}} \approx$ 13K).

This variation in f_{crit} is consistent with differences in pretraining data curation. A highly curated corpus (such as Mistral's) selectively upsamples high-quality factual content, effectively increasing the model's exposure to rare entities relative to their raw frequency on the open web. Consequently, the model achieves reliable recall at a lower apparent Wikipedia page view threshold. Ultimately, these transfer results demonstrate that the *sigmoidal form* of the knowledge boundary is a stable, universal property across LLMs, while the specific f_{crit} location serves as a fingerprint of the model's unique pretraining distribution and data quality.

6.7 Experiment 5: From Population to Individual Prediction

Experiments 1–4 show that the frequency sigmoid provides a strong and well-calibrated *population-level* risk estimate. However, within a narrow frequency stratum, queries have similar exposure and therefore receive similar frequency-based risk scores; as a result, frequency alone has limited *within-stratum* discrimination. In this experiment, we show that *relation structure* provides the missing signal for individual-level prediction.

Methodology. Each PopQA question instantiates a semantic relation (e.g., "Who was the director of X?", "What is X the capital of?"). We extract relation types by pattern matching and analyze the high-

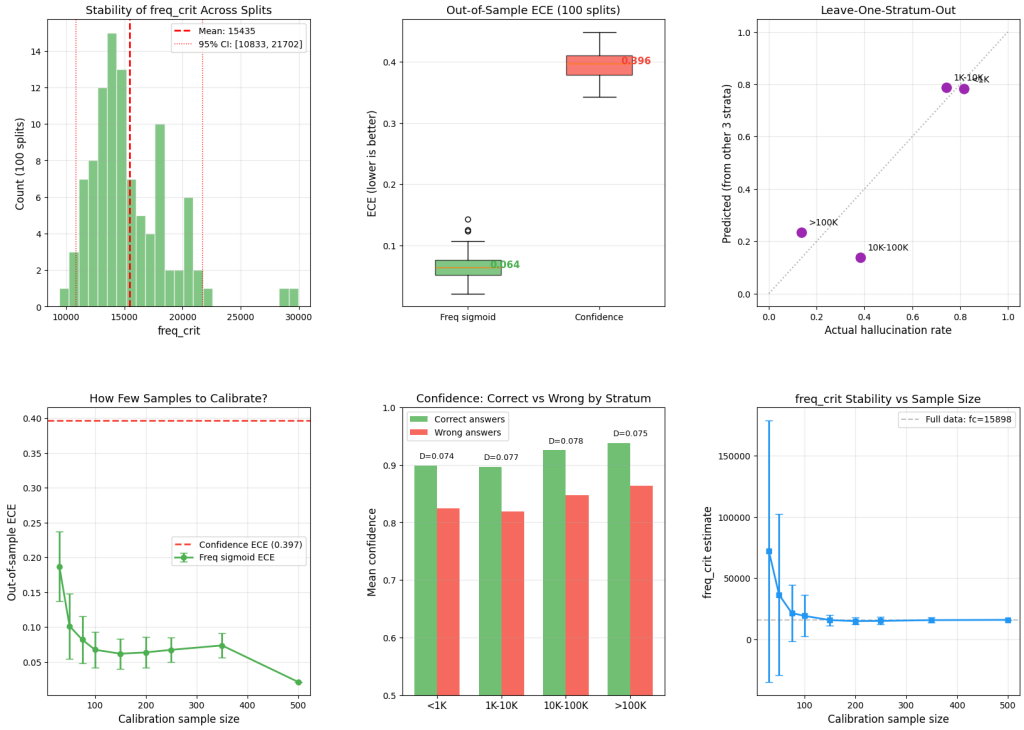

Figure 5: **Sigmoid transferability validation.** Top-left: the fitted threshold f_{crit} across 100 random 50/50 train/test splits. Top-center: out-of-sample ECE. Top-right: LOSO transfer, comparing predicted vs. actual *stratum-level* hallucination rates. Bottom-left: sample efficiency for calibrating the sigmoid with limited labeled data. Bottom-center: mean confidence for correct vs. wrong answers by stratum showing a nearly constant separation. Bottom-right: f_{crit} estimates stabilize as calibration sample size increases.

Table 6: Cross-model transfer ECE on the 500-query PopQA evaluation set. Diagonal entries (bold) are self-fit. Models with similar cliff locations (Falcon ↔ Qwen) transfer especially well. Even worst-case sigmoid transfer (Mistral → others, ECE ≈ 0.154) remains much better than raw confidence ECE (0.482–0.592).

Fit on \ Predict on	**Falcon-7B**	**Mistral-7B**	**Qwen-2.5-7B**
Falcon-7B	**0.037**	0.153	0.029
Mistral-7B	0.154	**0.026**	0.136
Qwen-2.5-7B	0.039	0.135	**0.016**
Raw confidence ECE	0.592	0.482	0.511

coverage relation classes in the 500-query Mistral-7B evaluation set. For each relation r with sufficient support ($n \geq 15$), we fit a relation-specific sigmoid: $\text{acc}_r(f) = a_{\min,r} + \frac{a_{\max,r} - a_{\min,r}}{1+\exp\left(-\kappa_r(\log f - \log f_{\text{crit}}^{(r)})\right)}$. Here $a_{\min,r}$ and $a_{\max,r}$ denote the relation-specific floor and ceiling accuracies, κ_r controls transition steepness, and $f_{\text{crit}}^{(r)}$ is the relation-specific critical frequency. This allows the cliff location $f_{\text{crit}}^{(r)}$ (and slope κ_r) to vary by relation. Intuitively, if relation-specific f_{crit} values differ substantially, then relation type carries information about hallucination risk beyond entity frequency alone. For prediction experiments, we distinguish: (i) **pre-inference** features, and (ii) **post-inference** features. Unless otherwise stated, all reported prediction metrics are cross-validated. Table 7 reports relation-specific sigmoid fits (binned fits; $n \geq 15$). Among the well-sampled relations, the inferred cliff location varies by nearly two orders of magnitude, from `author` ($f_{\text{crit}} = 340$) to `director` ($f_{\text{crit}} = 25{,}857$), a 76× range. The global fit is $f_{\text{crit}} = 12{,}726$ with $R_w^2 = 0.969$. These differences reflect systematic heterogeneity in memorization difficulty across relation types which is

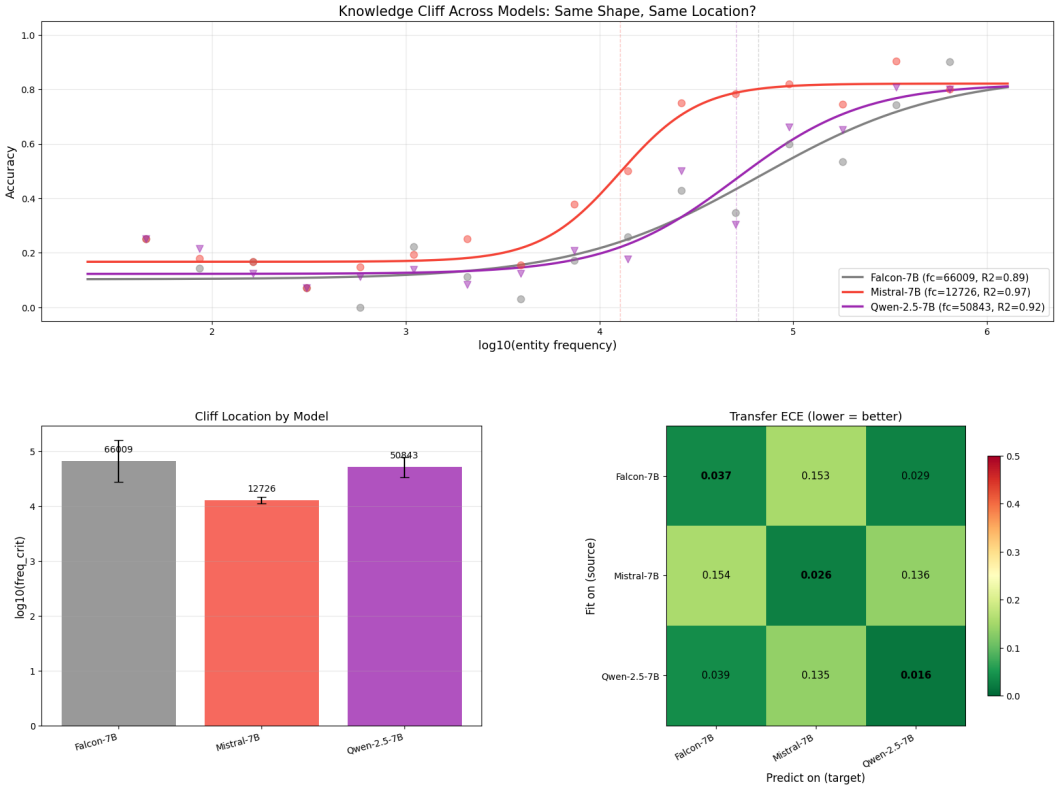

Figure 6: **Cross-model sigmoid transfer (500-query PopQA evaluation set).** Top: all three models exhibit sigmoidal knowledge cliffs (weighted $R^2 > 0.89$). Mistral's cliff (red, $f_{\text{crit}} = 12{,}726$) is shifted left relative to Falcon (gray, 66,009) and Qwen (purple, 50,843). Bottom-left: fitted cliff location by model with 1-SE error bars. Bottom-right: transfer ECE heatmap; the Falcon/Qwen block transfers well (ECE 0.029–0.039), while Mistral-to-others transfer degrades (ECE ≈0.154), consistent with the larger shift in f_{crit}.

consistent with relation-dependent variation in encoding opportunities and storage difficulty, though a full causal decomposition would require corpus-level co-occurrence measurement.

Table 7: Relation-specific sigmoid fits (Mistral-7B, PopQA; relations with $n \geq 15$). The inferred cliff location f_{crit} varies by 76× across relation types. R^2 is computed on binned accuracies. "Shift" is the log-distance from the global $f_{\text{crit}} = 12{,}726$.

Relation	**n**	f_{crit}	$\log_{10} f_{\text{crit}}$	R^2	**Shift**
author	33	340	2.53	0.672	−1.57 (easier)
capital-of	84	5,428	3.73	0.846	−0.37
country	31	941	2.97	0.961	−1.13
composer	34	13,394	4.13	0.999	+0.03
genre	51	13,355	4.13	1.000	+0.03
producer	46	12,615	4.10	0.925	0.00
screenwriter	68	19,100	4.28	0.995	+0.18
director	66	25,857	4.41	0.965	+0.31
Global	500	12,726	4.10	0.969	—

Within-stratum discrimination. To isolate *individual-level* discrimination within frequency bands, we evaluate AUROC separately in each stratum using three signals: frequency-based risk, model confidence,

and a relation-derived predictor. Table 8 shows a clear two-regime pattern. In the tail (< 1K), relation type is the strongest signal (AUROC $= 0.826$), exceeding confidence (0.752) and frequency-only risk (0.444). In higher-frequency strata, confidence dominates (e.g., 10K–100K: 0.784; $>$ 100K: 0.636). This supports the interpretation that, in the tail, the key question is *which kinds of facts survive low exposure*, whereas in the head, most facts are covered and residual failures are better explained by generation-time variability captured by confidence.

Table 8: Within-stratum AUROC for hallucination prediction (Mistral-7B, PopQA). "Freq" denotes the frequency-sigmoid risk score. Relation-derived features are computed from held-in data only. Relation dominates in the tail (< 1K), while confidence dominates in higher-frequency strata.

Stratum	n	Freq	Conf	Relation	Best
<1K	125	0.444	0.752	**0.826**	0.826
1K–10K	125	0.590	**0.706**	0.616	0.706
10K–100K	125	0.660	**0.784**	0.544	0.784
>100K	125	0.510	**0.636**	0.504	0.636

Combined individual-level predictors. To assess the relative value of pre- and post-inference signals, we evaluate logistic regression predictors using 10-fold cross-validation. Among single-feature models, the zero-compute frequency score strictly dominates post-inference confidence (AUROC $= 0.813 \pm 0.050$ vs. 0.769 ± 0.073). Combining frequency with relational structure yields an exceptionally strong *fully pre-inference* predictor (0.842 ± 0.055). While incorporating post-inference confidence achieves the global maximum (0.875 ± 0.044), this marginal $+0.033$ improvement comes at the steep computational cost of a full LLM forward pass. Thus, the vast majority of predictive power is already available from query-derived metadata prior to generation.

Generalization checks. To ensure the relational signal (Figure 7) is portable rather than benchmark-specific, we conduct three generalization tests. First, a Leave-One-Relation-Out (LORO) evaluation yields a mean AUROC $= 0.792$ across 14 held-out relations. Second, substituting exact relation identities with generic structural features (alias count, median frequency, within-relation spread) achieves AUROC $= 0.819 \pm 0.043$, capturing roughly 21% of the identity-based gain over frequency-only routing. Finally, Leave-K-Relations-Out trials ($K \in \{3, 5\}$) remain highly stable (AUROC > 0.806), demonstrating that the predictive power of relational structure does not degrade even when up to one-third of relation types are entirely unseen.

6.8 Experiment 6: Capacity Scaling

We test Proposition 5 directly using the Qwen2.5-Instruct family (0.5B, 1.5B, 3B, 7B, 14B). Because these models were all trained on the identical 18T-token corpus, we can isolate parameter count as the sole independent variable. Evaluated on the 500-query PopQA set, the result tightly follows a power law: $\log_{10} f_{\text{crit}} = 4.96 - 0.52 \log_{10} P$ ($\hat{\alpha} = 0.52 \pm 0.05$, $R^2 = 0.940$, $n = 5$). The critical frequency f_{crit} decreases monotonically from 161,169 at 0.5B to 23,940 at 14B. Interpreted practically, each decade of parameter growth shifts the cliff by 0.52 log-decades of frequency: a $10\times$ larger model can reliably recall entities that are $\sim 3.3\times$ rarer. Crucially, as shown in Figure 8 (bottom-right), this leftward shift of the cliff does not solve the extreme long-tail deficit; baseline accuracy for entities below 1K page views remains near 10% regardless of model scale, reinforcing the persistent need for retrieval interventions. Several methodological nuances support the robustness of these fits. While the Qwen2.5-7B sigmoid fit exhibits an unusually high κ (hitting the optimizer bound and contributing uncertainty to the exponent estimate), the overall monotonic trend remains highly consistent. Furthermore, at a matched parameter count, Mistral-7B sits 0.19 log-decades below the Qwen scaling line ($f_{\text{crit}} = 12{,}726$ vs. Qwen-7B's 24,444). This is theoretically consistent with Mistral utilizing a smaller but more aggressively curated corpus that upsamples high-quality encyclopedic data Jiang et al. (2023a), which encodes knowledge more efficiently per parameter—corpus quality shifts the intercept without changing the fundamental capacity slope. Finally, note that the Qwen-7B value reported here (24,444) differs from Table 1 (50,843) because this scaling experiment rigorously applies model-native ChatML prompt

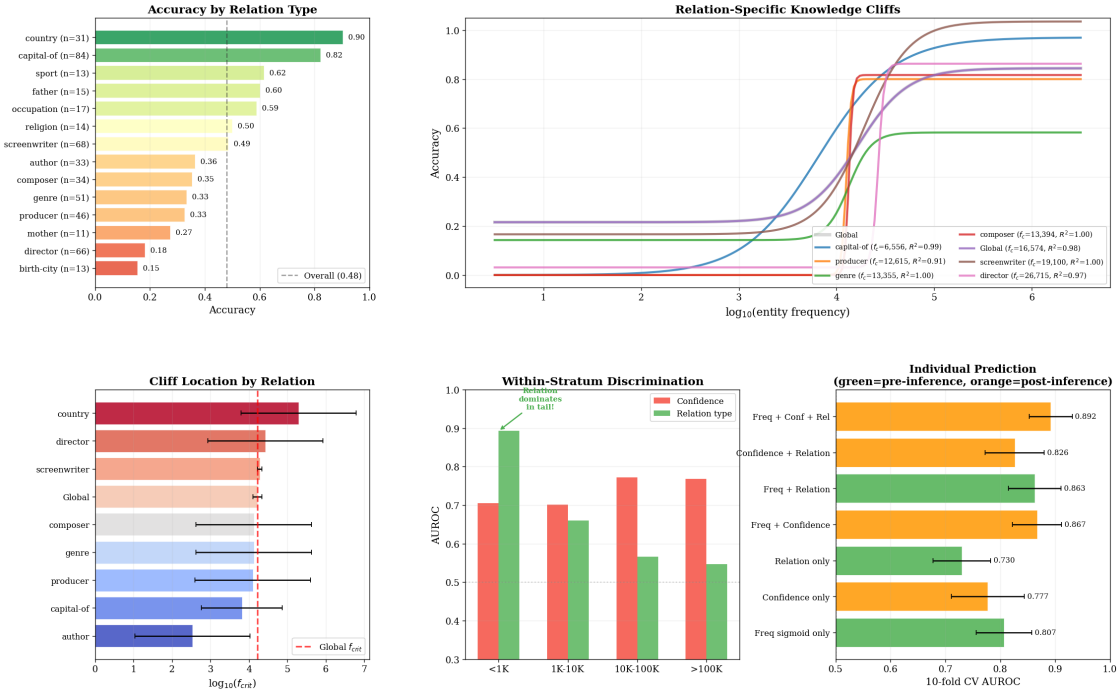

Figure 7: **Relation-specific knowledge cliffs and individual-level prediction (Mistral-7B).** Top row: Accuracy and sigmoidal cliffs vary drastically across relation types, revealing substantial structural heterogeneity behind the global average. Bottom-left: Critical frequencies (f_{crit}) span a massive 76× dynamic range depending on the relation. Bottom-center: Within-stratum discrimination shows relation identity strictly dictates survival in the rare-entity tail (<1K), whereas LLM confidence dominates in the well-exposed head. Bottom-right: 10-fold CV confirms `Freq + relation` (green) is a highly competitive pre-inference routing baseline compared to expensive post-inference combinations (orange).

formatting via `apply_chat_template` to ensure fair cross-size comparison, whereas Experiment 1 used a generic shared template.

7 Discussion

Architectural Implications: Two-Stage Triage. Our findings across both retrieval routing and selective prediction (abstention) consistently point to a two-stage architecture for managing LLM hallucinations. Because the `Freq + relation` proxy successfully resolves the Simpson's-paradox-like gap—correctly ordering risk both globally and within strata—it provides a highly effective, zero-compute triage step. Reserving post-inference confidence scores purely for marginal refinement on borderline cases yields the optimal cost-utility frontier. Notably, this pre-inference dominance holds true across various abstention utility thresholds, confirming that query-side structure is a versatile reliability signal independent of the specific downstream intervention.

The Nature of the Knowledge Cliff. Kalai et al. (2025) argue that hallucinations persist because binary evaluation incentivizes guessing over expressing uncertainty. Under our threshold-distortion idealization (Assumption 1), a natural question is whether the cliff is simply an artifact of this binary framing. Our robustness checks across alternative, graded-correctness scoring regimes confirm that the cliff's location and steepness are preserved; only the asymptotic error floors shift. This confirms that the knowledge cliff is not a mere scoring artifact, but rather a reflection of a structural capacity boundary acquired during pretraining.

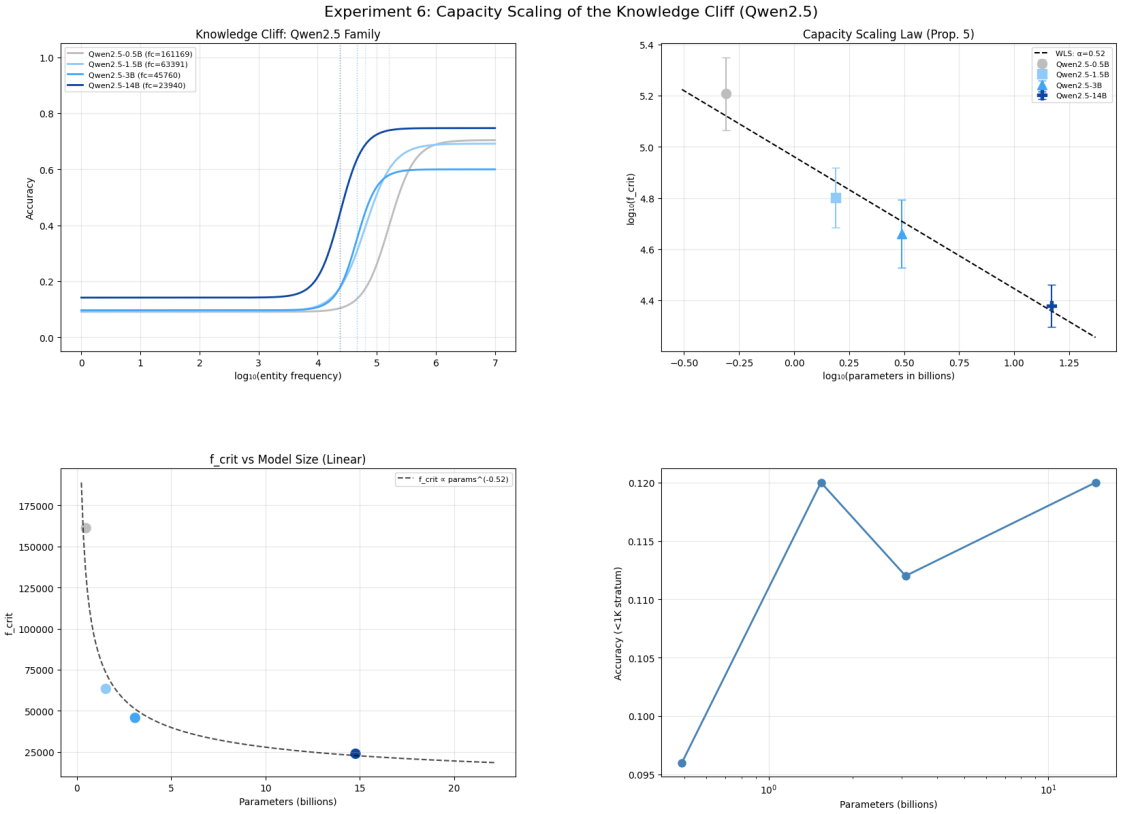

Figure 8: **Capacity scaling law (Experiment 6).** *Top-left:* Knowledge cliff curves for the Qwen2.5-Instruct family, all trained on the same 18T-token corpus. Larger models shift the cliff toward rarer entities. (Qwen2.5-7B is excluded from the visual curves due to a degenerate sigmoid fit, but included in regressions). *Top-right:* Log–log scaling law: $\log_{10} f_{\text{crit}} = 4.96 - 0.52 \log_{10} P$ ($\hat{\alpha} = 0.52 \pm 0.05$, $R^2 = 0.940$, $n = 5$). Error bars show 1-SE uncertainty. *Bottom-left:* The same scaling law on linear axes, illustrating the power-law decay $f_{\text{crit}} \propto P^{-0.52}$. *Bottom-right:* Tail accuracy (<1K stratum) vs. model size. Performance remains trapped near 10% across all scales; minor non-monotonic fluctuations are statistical noise driven by the small stratum size ($n = 125$), demonstrating that parameter scaling alone cannot rescue tail recall.

7.1 Missing mass, abstention, and the Good–Turing connection

We now connect our abstention/routing perspective to the "arbitrary facts" model of Kalai et al. (2025). In that regime, generalization beyond memorization is information-theoretically impossible for prompts whose labels are unsupported by training evidence: for unseen prompts, no learner can do better than random guessing (in expectation over the random fact assignment). Consequently, the *missing mass* of prompts becomes a fundamental limit on attainable selective utility. To formalize this, let M_0 denote the *missing mass* of prompts,

$$M_0 \;:=\; P\big(\{c \in \mathcal{C} : c \notin \{c^{(1)}, \ldots, c^{(N)}\}\}\big),$$

i.e., the probability that a test prompt was never observed in training. Under arbitrary facts, for any unseen prompt the predictor's output (conditioned on the training sample and any learner state derived from it) is independent of the true label, so expected correctness is at most $1/|\mathcal{Y}|$, where $|\mathcal{Y}|$ is the size of the answer space. If $u_{\text{abst}} \geq 1/|\mathcal{Y}|$, then abstaining weakly dominates guessing on unseen prompts. Thus the utility loss from perfection is controlled by the missing mass.

Corollary 12 (Missing-mass upper bound on selective utility)**.** *Assume $u_{\text{abst}} \geq 1/|\mathcal{Y}|$. Then, conditional on any realized training sample S,*

$$\mathbb{E}[u \mid S] \;\leq\; 1 - (1 - u_{\text{abst}})\, M_0(S), \tag{30}$$

where the expectation is over the test prompt, the arbitrary-facts label assignment (for unseen prompts), and any predictor randomness.

Proof. Fix the sample S. Decompose by whether the test prompt is seen in S. On seen prompts, utility is at most 1. For an unseen prompt in the arbitrary-facts model, conditional on the training sample (and any learner state derived from it), the true label remains uniform on $\mathcal{Y}$ and independent of the predictor's output. Hence any non-abstaining prediction has expected utility at most $1/|\mathcal{Y}|$, whereas abstention yields utility u_{abst}. Therefore, when $u_{\text{abst}} \geq 1/|\mathcal{Y}|$, abstention weakly dominates guessing (and strictly dominates it when $u_{\text{abst}} > 1/|\mathcal{Y}|$). Thus,

$$\mathbb{E}[u \mid S] \leq (1 - M_0(S)) \cdot 1 + M_0(S) \cdot u_{\text{abst}} = 1 - (1 - u_{\text{abst}}) M_0(S).$$

□

A key point, consistent with Kalai et al. (2025), is that missing mass is also *estimable* from the sample via singleton statistics. Let K_1 be the number of prompts that appear exactly once in $\{c^{(1)}, \ldots, c^{(N)}\}$, and define the Good–Turing estimator $G_0 := K_1/N$.

Lemma 13 (Good–Turing concentration for missing mass)**.** *Let $X_1, \ldots, X_N \sim P$ be i.i.d. from a countable domain, let M_0 be the missing mass, and let $G_0 = K_1/N$ be the Good–Turing estimator. then one obtains*

$$|M_0 - G_0| \;\leq\; \varepsilon_{\text{GT}}(N, \delta),$$

with probability at least $1 - \delta$, where $\varepsilon_{\text{GT}}(N, \delta) = O\left(\sqrt{\frac{\log(N/\delta)}{N}}\right)$ (see McAllester & Schapire 2000 for explicit constants/bounds).

A data-dependent abstention limit. Combining the missing-mass utility bound (30) with the Good–Turing concentration result (Lemma 13) yields a sample-dependent upper bound based entirely on the observable singleton fraction. Specifically, on the high-probability event $M_0(S) \geq \big(G_0 - \varepsilon_{\text{GT}}(N, \delta)\big)_+$, we have:

$$\mathbb{E}[u \mid S] \;\leq\; 1 - (1 - u_{\text{abst}})\, M_0(S) \;\leq\; 1 - (1 - u_{\text{abst}})\, \big(G_0 - \varepsilon_{\text{GT}}(N, \delta)\big)_+. \tag{31}$$

Thus, when $u_{\text{abst}} \geq 1/|\mathcal{Y}|$, the observable singleton rate G_0 directly bounds the best achievable selective utility. While this bound is tightest in the pure arbitrary-facts regime (Kalai et al., 2025), in our PopQA setting (natural facts with shared structure) it serves as a rigorous ceiling, clarifying the mathematical limits of what abstention and retrieval can fix in the extreme tail.

Robustness to graded correctness. A natural concern is that the sharp knowledge cliff is merely an artifact of strict binary scoring (exact match). Factual recall often involves near-misses (e.g., spelling variations) or partial structure. We tested this by refitting our sigmoid model under five increasingly lenient scoring regimes: strict binary, continuous fuzzy match, token overlap, hand-crafted partial credit, and Brier quality. The transition *shape* remained remarkably stable: neither the cliff steepness ($\kappa \in [3.77, 4.93]$) nor its location ($\log_{10} f_{\text{crit}} \in [4.09, 4.15]$) shifted systematically. Instead, lenient scoring simply raised the asymptotic error floors (from 0.158 to 0.430) as tail entities received partial credit for uncertainty-aware outputs. Because near-misses account for only 5.6% of total errors, they do not drive the cliff dynamics. This invariance to post-hoc re-scoring confirms the cliff reflects a genuine internal capacity boundary, not an evaluation artifact.

Scope conditions: When does the framework apply? The sigmoidal knowledge cliff requires two preconditions: (i) the task decomposes into discrete competency units drawn from a Zipfian distribution, and (ii) each unit has a *time-stable ground truth*. We verified the necessity of precondition (ii) by attempting to replicate the cliff in the code/API domain (using GitHub stars as an exposure proxy). The sigmoid entirely failed to emerge ($R^2 \approx 0$). Accuracy was non-monotonic because API defaults evolve across versions; models trained on multi-version documentation often answer confidently but incorrectly for the "current" default. Thus, while our framework naturally extends to domains like medical QA, legal citation, or translation (where facts are time-stable), it fundamentally does not apply to non-stationary environments or pure reasoning tasks (e.g., math proofs) where performance relies on logic depth rather than exposure coverage.

Methodological requirements for evaluation. Our primary results rely on PopQA, a benchmark explicitly stratified by frequency. When we attempted to generalize to Mintaka (a standard, unstratified QA dataset), the global sigmoid fit degraded ($R^2 = 0.461$) and the pre-inference frequency predictor underperformed raw confidence. This failure is purely structural: unstratified datasets severely undersample the mid-frequency and tail regions. Without balanced frequency coverage, the global sigmoid lacks the necessary anchor points, collapsing its discriminative power. Notably, when isolated to specific relation types within Mintaka with sufficient data, the sigmoid fits remained excellent ($R^2 > 0.99$). The practical implication for future research is clear: applying this framework to new domains requires constructing explicitly frequency-stratified evaluation sets (consistent with our ∼75-sample bound in Section 5.1); standard benchmarks cannot simply be repurposed.

Proxy validation and structural covariates. Finally, we validated that our results are not tied to Wikipedia page views. Alternative structural proxies (e.g., alias count) successfully predict correctness (AUROC = 0.684), while adversarial proxies (string length, alphabetical order) collapse to chance (AUROC ≈ 0.50). Furthermore, the predictive power of relation types is largely explained by portable structural covariates (answer ambiguity, neighborhood density) rather than benchmark-specific IDs or tokenization artifacts (adding GPT-2 subject token length yielded zero AUROC gain). This confirms that the dominant routing signals are genuine representations of exposure and structural difficulty.

8 Conclusion

We demonstrate that factual reliability in LLMs exhibits a predictable, sigmoidal "knowledge cliff" driven by entity exposure. Below a critical frequency f_{crit}, models performs poorly; above it, accuracy transitions rapidly to a reliable plateau. Crucially, this cliff is not a monolithic global boundary. Its exact location varies by nearly two orders of magnitude depending on the specific relational structure of the fact being queried. By combining raw entity frequency with relational metadata, we construct a fully pre-inference risk score that captures the vast majority of predictive signal, outperforming the LLM's own internal confidence metrics in the rare-entity tail. This confirms the viability of a highly efficient, two-stage routing architecture: zero-compute query metadata should be used for initial RAG triage, while expensive post-inference confidence is reserved solely for marginal refinement.

Furthermore, we empirically validate the theoretical prediction that the location of this knowledge cliff is fundamentally governed by model capacity. Across the Qwen2.5-Instruct family (trained on a fixed 18T-token corpus), the critical frequency obeys a strict power law: $f_{\text{crit}} \propto P^{-0.52}$. Practically, this dictates that increasing parameter count by 10× allows a model to reliably recall entities that are only $\sim 3.3\times$ as frequent. Comparing this scaling law across model families reveals that highly curated pretraining corpora (e.g., Mistral) shift the intercept of this cliff without altering its fundamental slope. Most importantly, because accuracy deep in the tail remains trapped near 10% regardless of model size, parameter scaling alone cannot resolve the long-tail hallucination deficit.

Ultimately, for tasks that decompose into Zipfian "competency units," factual reliability is not an unpredictable generation-time anomaly, but a deterministic property of the query's exposure and structural difficulty. Exposing this structure, however, requires evaluation benchmarks that are explicitly stratified by frequency; without this design, global metrics merely mask the steep performance collapse in the tail. By formalizing the mathematical shape of the knowledge boundary, this framework shifts the paradigm of hallucination mitigation from post-hoc output analysis toward pre-hoc risk assessment and compute allocation, enabling robust reliability before a single token is generated.

Code and data. Code for reproducing all experiments will be made available upon publication.

References

Zeyuan Allen-Zhu and Yuanzhi Li. Physics of language models: Part 3.3, knowledge capacity scaling laws. In *International Conference on Learning Representations (ICLR)*, 2025.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *International Conference on Learning Representations (ICLR)*, 2024.

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

Kenneth P. Burnham and David R. Anderson. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. Springer, New York, 2nd edition, 2002.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramèr, and Chiyuan Zhang. Quantifying memorization across neural language models. In *International Conference on Learning Representations (ICLR)*, 2023.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience, 2nd edition, 2006.

Grégoire Delétang, Anian Ruoss, Jordi Grau-Moya, Tim Genewein, Li Kevin Bos, Elliot Catt, Marcus Catt, Charlie Mattern, Matthew Aitchison, Laurent Orseau, Shane Legg, and Marcus Hutter. Language modeling is compression. In *International Conference on Learning Representations (ICLR)*, 2024.

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

Anxin Guo and Jingwei Li. Hallucination is a consequence of space-optimality: A rate-distortion theorem for membership testing. *arXiv preprint arXiv:2602.00906*, 2026.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023a. URL `https://arxiv.org/abs/2310.06825`.

Zhengbao Jiang, Frank F. Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. Active retrieval augmented generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023b.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DaSilva, Eli Tran-Kemp, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Adam Tauman Kalai and Santosh S. Vempala. Calibrated language models must hallucinate. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing (STOC)*, 2024.

Adam Tauman Kalai, Ofir Nachum, Santosh S. Vempala, and Edwin Zhang. Why language models hallucinate. *arXiv preprint arXiv:2509.04664*, 2025.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. Large language models struggle to learn long-tail knowledge. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

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

Yaniv Leviathan, Matan Kalman, and Yossi Matias. Prompt repetition improves non-reasoning llms. *arXiv preprint arXiv:2512.14982*, 2025.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Matthieu Mahaut et al. Factual confidence of llms: on reliability and robustness of current estimators. *ACL*, 2024. URL `https://aclanthology.org/2024.acl-long.250/`.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 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)*, 2023.

Potsawee Manakul, Adian Liusie, and Mark J.F. Gales. 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)*, 2023.

Pascal Massart. The tight constant in the Dvoretzky–Kiefer–Wolfowitz inequality. *The Annals of Probability*, 18(3):1269–1283, 1990.

David A. McAllester and Robert E. Schapire. On the convergence rate of Good–Turing estimators. In *Proceedings of the 13th Annual Conference on Computational Learning Theory (COLT)*, pp. 1–6, 2000. Available as: `https://www.learningtheory.org/colt2000/papers/McAllesterSchapire.pdf`.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. *arXiv preprint arXiv:2305.14251*, 2023.

Isaac Ong, Amjad Almahairi, Vincent Wu, Wei-Lin Chiang, Tianhao Wu, Joseph E. Gonzalez, M. Waleed Kadous, and Ion Stoica. RouteLLM: Learning to route LLMs from preference data. In *International Conference on Learning Representations (ICLR)*, 2025.

Ravid Shwartz-Ziv and Naftali Tishby. Opening the black box of deep neural networks via information. *arXiv preprint arXiv:1703.00810*, 2017.

Kai Sun, Yejin Groeneveld, Vivek Kulkarni, and Colin Raffel. Head-to-tail: How knowledgeable are large language models (LLMs)? A.K.A. will LLMs replace knowledge graphs? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2024.

Kushal Tirumala, Aram H. Markosyan, Luke Zettlemoyer, and Armen Aghajanyan. Memorization without overfitting: Analyzing the training dynamics of large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.