ArtificialGate

Quantifying Prior Dominance in RAG Systems

Barak Or
ArtificialGate Ltd.
barak@artificialgate.ai

Abstract

Retrieval-Augmented Generation (RAG) grounds Large Language Models in external knowledge, yet current evaluations rely on discrete heuristics that suffer from "epistemic blindness" - failing to distinguish genuine contextual information extraction from parametric memory recall. To address this, we introduce the Normalized Context Utilization (NCU) metric, leveraging continuous token log-probabilities across zero-shot, oracle, and adversarial conditions to strictly quantify contextual information gain. Evaluating architectures ranging from 1.5B to 72B parameters alongside a proprietary commercial API reveals that for strict factual extraction (without Chain-of-Thought reasoning), traditional scaling laws exhibit extreme diminishing returns: highly efficient Small Language Models (SLMs) match or outperform high-capacity architectures. Furthermore, we demonstrate that "Prior Dominance" correlates with model scale and proprietary alignments. The evaluated commercial API not only overrode explicit external evidence in nearly half of adversarial conflicts, but also frequently suffered from systemic confidence collapse (*Negative Transfer*) when its parametric priors were contradicted. Our findings highlight the structural epistemic advantage and superior contextual adherence of SLMs in strict extraction workflows.

1 Introduction

As Large Language Models (LLMs) achieve broad general-purpose reasoning capabilities, integrating dynamic, out-of-distribution knowledge remains a fundamental challenge. The prevailing paradigms for knowledge integration-pre-training foundational models or applying supervised Fine-Tuning (FT)-are bottlenecked by compute constraints, high-latency updates, and catastrophic forgetting. Consequently, Retrieval-Augmented Generation (RAG) [1–4] has emerged as a standard mechanism for decoupling knowledge from model weights, aiming to shift the computational burden from parametric memory recall to active contextual processing.

The trajectory of LLM development is largely guided by established scaling laws [5–8], which posit that larger parameter counts predictably yield better downstream performance. Applying this assumption directly to RAG architectures, however, exposes a methodological limitation in current evaluation frameworks. Standard open-domain Question Answering (QA) benchmarks rely heavily on discrete text-matching heuristics, such as Exact Match (EM) or F1 scores [9–11]. These binary metrics suffer from an evaluation blind spot: when a model correctly answers a query given a retrieved context, standard heuristics cannot ascertain whether the model genuinely extracted the target information from the text, or merely recited the answer from its pre-trained weights.

This ambiguity obfuscates true contextual utilization. The apparent superiority of massively scaled models in RAG benchmarks may stem from their extensive prior knowledge rather than

This research was independently conducted and funded by ArtificialGate Ltd. The author serves as Academic Director at the Google & Reichman Tech School, and as an External Lecturer at Technion – Israel Institute of Technology and Reichman University.

superior contextual processing. Furthermore, binary metrics fail to capture continuous fluctuations in a model's predictive confidence when encountering adversarial contexts, knowledge conflicts [12–14], or semantic noise.

To rigorously disentangle parametric memory from active context extraction, we propose a shift from discrete text matching to continuous probability spaces. In this paper, we introduce the **Normalized Context Utilization (NCU)** metric. Unlike traditional heuristics, NCU leverages continuous length-normalized log-probabilities across zero-shot, oracle, and adversarial conditions. By isolating the generator component with provided oracle contexts-effectively evaluating Context-Augmented Generation (CAG)-we can rigorously assess the fractional reduction in logarithmic uncertainty when a model is exposed to contextual evidence. Our study specifically focuses on strict factual extraction without Chain-of-Thought (CoT) reasoning paradigms.

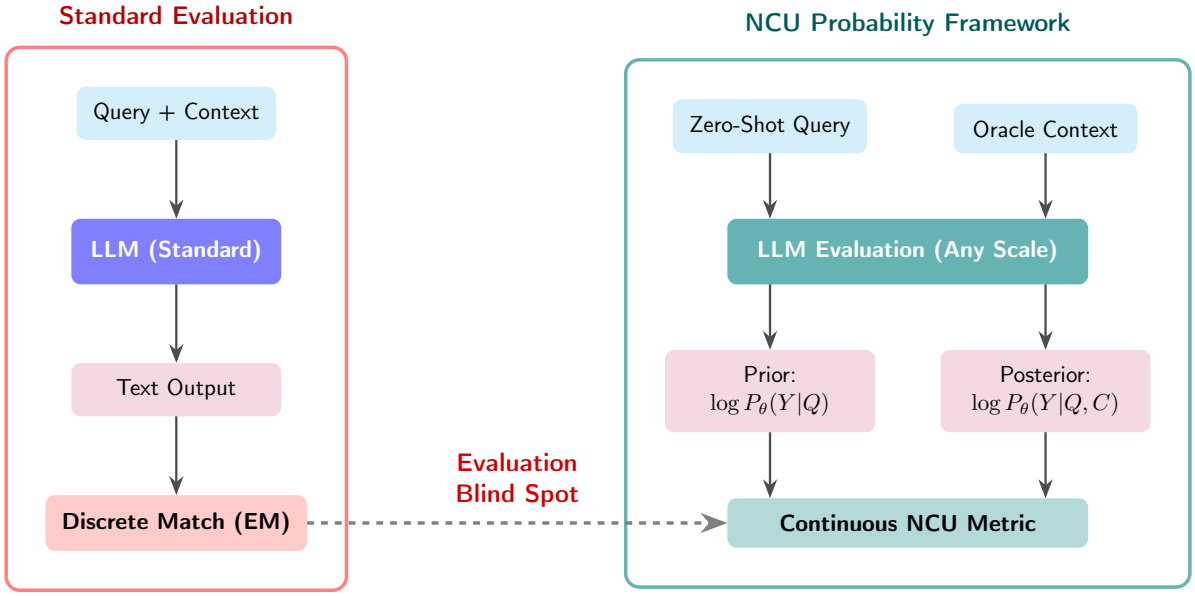

Figure 1: **Conceptual Architecture: Standard Evaluation vs. NCU Framework.** Traditional metrics (left) rely on binary text matching, failing to distinguish parametric recall from reading comprehension. The NCU framework (right) utilizes continuous log-probabilities across zero-shot and oracle conditions to isolate the informational gain of the RAG context.

To empirically explore this framework, we conducted bounded inferences across three distinct QA domains. By evaluating architectures ranging from 1.5B open-weight Small Language Models (SLMs) to massive models (72B) and a proprietary commercial API against systematically perturbed contexts, we mapped the empirical boundaries of context utilization.

Specifically, this study addresses the following core Research Questions (RQs), conceptually illustrated in **Figure 2**:

- **RQ1 (Extraction Scaling):** Does increasing parameter count intrinsically improve contextual extraction, or do strict extraction tasks exhibit diminishing returns under standard scaling paradigms?
- **RQ2 (Prior Dominance):** How do parametric scale and proprietary alignments influence a model's propensity to override explicitly provided, yet contradictory, external evidence?
- **RQ3 (Negative Transfer):** Under what conditions does semantic conflict cause high-capacity models to suffer from negative transfer (i.e., information loss relative to their zero-shot baseline)?

Through our empirical analysis, **we present the following core contributions:**

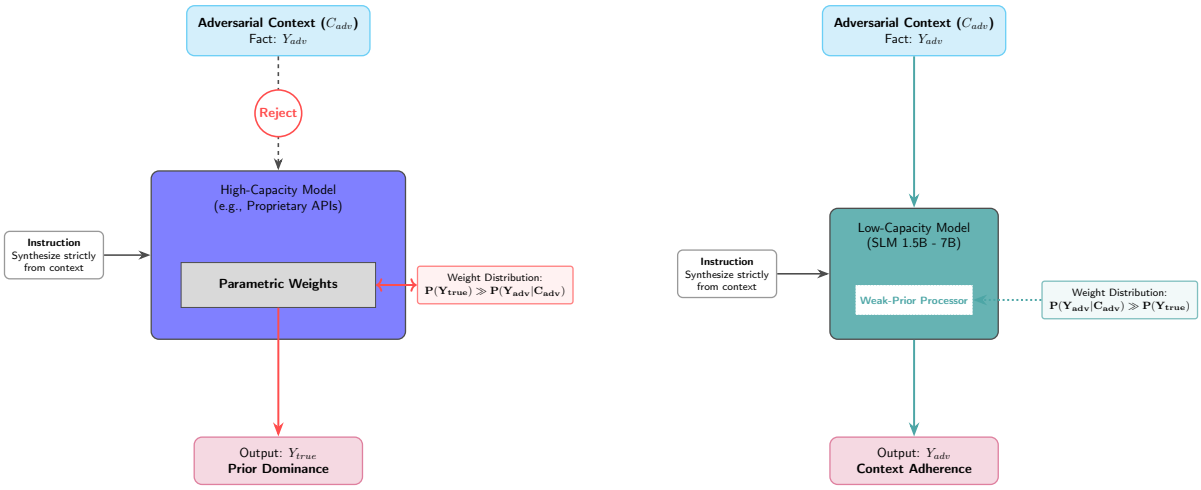

Figure 2: **Prior Dominance vs. Context Adherence.** A conceptual visualization of knowledge conflicts. Left: High-capacity models may override contradictory context due to heavily weighted parametric priors ($P_{param} \gg P_{ctx}$). Right: Low-capacity SLMs generally exhibit higher context adherence, processing external evidence with fewer parametric constraints.

1. **Continuous Evaluation Metric:** We define and validate the NCU metric, providing a probability-driven framework that isolates contextual processing from parametric memory, normalizing for vocabulary and tokenization disparities across models.

2. **Diminishing Returns in Extraction Scaling:** We empirically demonstrate severe diminishing returns for parameter scale in factual grounding tasks. SLMs (1.5B–7B) achieved statistical parity in context utilization compared to a 72B parameter model, while offering up to a 70× reduction in inference latency relative to a commercial API.

3. **Quantification of Prior Dominance and Negative Transfer:** We identify a measurable vulnerability in high-capacity models: when presented with explicit external conflicts, they frequently override context (*Prior Dominance*) and are susceptible to systemic confidence collapse (*Negative Transfer*), whereas SLMs maintain higher epistemic stability.

The remainder of this paper is structured as follows. Section 2 reviews the literature on RAG evaluation and scaling laws. Section 3 frames the theoretical metric. Section 4 details our methodology. Section 5 presents our empirical findings. Section 6 discusses the operational implications. Section 7 acknowledges limitations, and Section 8 concludes.

2 Related Work

The efficacy of RAG is fundamentally tied to a model's ability to prioritize external evidence over internal priors. This section contextualizes our NCU framework within the literature regarding RAG evaluation, scaling laws, and knowledge conflict resolution.

2.1 The Evolution of RAG Evaluation

Early RAG architectures were predominantly evaluated using discrete lexical metrics such as EM and F1-score [1]. As LLMs evolved, evaluation paradigms shifted toward LLM-as-a-judge frameworks [15–17]. However, recent findings have highlighted the limitations of these discrete assessments, noting that binary heuristics suffer from "epistemic blindness" [18]. Consequently, evaluation is pivoting toward Information Theoretic benchmarks and token-level log-probability tracking [19, 20]. Our continuous NCU metric builds upon this foundation by formalizing the fractional reduction in predictive uncertainty. While "LLM-as-a-judge" frameworks like RAGAS

and ARES offer progress beyond exact matching, they often rely on the reasoning of another black-box model, which can introduce its own biases and lack of transparency. In contrast, our NCU metric leverages direct token-level log-probabilities, providing a more objective, reproducible, and fine-grained measure of how the model's internal confidence shifts when exposed to external context.

2.2 Scaling Laws: Pre-training vs. Inference Extraction

Standard scaling laws [5, 6] generally assume that larger parameter counts improve downstream capabilities. However, recent studies on RAG indicate more complex dynamics. Inference scaling laws for RAG reveal non-linear relationships between generator size and retrieval efficacy [21]. Concurrently, literature suggests a trade-off between parametric memorization and a model's willingness to integrate dynamic retrieval [22].

2.3 The Resurgence of Small Language Models

The computational overhead of massive LLMs has catalyzed interest in SLMs [23, 24]. Contemporary pipelines increasingly position SLMs as dynamic output rewriters [25], collaborative reasoning routers, or efficient context steering [26]. Furthermore, fine-tuning SLMs for structured extraction tasks has empirically shown competitive efficacy compared to zero-shot prompting of massive counterparts [27–29]. Our study aims to offer a quantitative perspective on this architectural shift, highlighting the potential contextual adherence of SLMs in pure extraction workflows.

2.4 Knowledge Conflicts and Epistemic Stubbornness

The intersection of parametric memory and external retrieval often leads to knowledge conflicts [12, 13, 30]. While the "Lost in the Middle" phenomenon established that LLMs may ignore context due to spatial constraints [31], recent literature explores epistemic suppression [32]. Benchmarks and resolution frameworks [33, 34] observe that highly parameterized models can over-trust internal weights, sometimes rejecting contradictory external evidence [35–38]. Our research synthesizes these observations by mathematically measuring this vulnerability as *Prior Dominance*, while recognizing the hypothesized role of alignment mechanisms in shaping this behavior.

While existing literature increasingly acknowledges the limitations of discrete metrics and the complex behavior of high-capacity models during knowledge conflicts, a unified, continuous metric that robustly isolates information extraction from parametric memory-while normalizing across differing tokenizers-remains underexplored. This study addresses this gap by proposing the length-normalized NCU framework.

3 Epistemic Uncertainty and Context Utilization

To quantify a model's contextual processing, we transition from discrete heuristics to continuous probability spaces. Building upon Shannon's foundational communication theory [39, 40], we frame RAG through the lens of information theory. In this paradigm, the external context acts as a secondary information channel designed specifically to resolve the model's predictive uncertainty regarding a user query. This continuous monitoring of internal uncertainty aligns with emerging paradigms that model reasoning as a stochastic inference process susceptible to runtime cognitive drift [41].

Let Q denote a user query, $Y = (y_1, y_2, \ldots, y_k)$ the ground-truth target sequence composed of k tokens, and θ the parameters of an LLM. Because different models employ different tokenization schemes (vocabularies), directly comparing the raw sum of log-probabilities across architectures

is mathematically flawed. Therefore, we normalize the joint probability by the token sequence length to derive the average log-probability per token.

In a zero-shot setting, the model's confidence is governed strictly by its parametric prior. To capture this information content uniformly, we map the length-normalized probabilities to Shannon Entropy, defining the prior epistemic uncertainty (S_{prior}) as:

$$S_{prior} = -\frac{1}{k}\sum_{i=1}^{k}\log P_\theta(y_i|y_{<i}, Q). \tag{1}$$

When a retrieval pipeline injects external context C, the model computes a posterior distribution. We similarly define the posterior uncertainty (S_{post}) as:

$$S_{post} = -\frac{1}{k}\sum_{i=1}^{k}\log P_\theta(y_i|y_{<i}, Q, C). \tag{2}$$

The raw Context Utilization Score (CUS) represents the absolute informational gain provided by the RAG pipeline. It is mathematically equivalent to the absolute reduction in predictive uncertainty, $CUS = S_{prior} - S_{post}$, explicitly defined as the difference between the normalized log-probabilities:

$$CUS = \left(\frac{1}{k}\sum_{i=1}^{k}\log P_\theta(y_i|y_{<i}, Q, C)\right) - \left(\frac{1}{k}\sum_{i=1}^{k}\log P_\theta(y_i|y_{<i}, Q)\right) \tag{3}$$

To create a comparative and standardized metric across models of varying scales and pre-training distributions, we formalize the **Normalized Context Utilization (NCU)**.

The absolute information gain is captured by the CUS. Because probability is strictly bounded ($P \leq 1$), the minimum possible posterior uncertainty is perfect absolute confidence, where $-\log(1) = 0$. Consequently, the theoretical upper bound for this informational gain is strictly limited by the model's initial ignorance: $\max(CUS) = S_{prior} - 0 = S_{prior}$.

Expressing the measured information gain as a fraction of this maximum theoretical bound yields the normalized resolution of uncertainty. To ensure numerical stability, specifically preventing undefined asymptotic behavior when the model is already perfectly confident in a zero-shot setting ($P_\theta(Y|Q) \to 1$, meaning $S_{prior} \to 0$), we introduce a minute smoothing factor ϵ (e.g., 10^{-5}) to the denominator. To rigorously handle boundary conditions, including instances of *negative transfer* where adversarial or noisy context exacerbates uncertainty ($S_{post} > S_{prior}$), we apply strict bounding operators. We formally define the NCU metric as follows:

$$NCU = \max\left[0, \min\left(1, \frac{CUS}{S_{prior} + \epsilon}\right)\right]. \tag{4}$$

This dual-normalization strategy strictly isolates the mechanical efficacy of the RAG integration from the model's pre-existing parametric weights. It ensures an equitable evaluation paradigm regardless of the underlying tokenization vocabulary or the model's initial prior knowledge. We optionally track the unclipped raw ratio ($\frac{CUS}{S_{prior}+\epsilon}$) to highlight asymptotic confidence collapse (Negative Transfer).

The formulated metric exhibits desirable boundary behavior for robust evaluation:

(i) **Ideal Utilization ($NCU = 1$):** Occurs when the posterior probability equals 1, meaning the retrieved context provides total certainty for the target sequence.

(ii) **Baseline Nullity ($NCU = 0$):** Occurs when the posterior confidence is less than or equal to the prior, implying the context provides zero or negative information gain.

(iii) **Sensitivity:** The derivative of NCU with respect to the posterior log-probability is constant, indicating that the metric scales linearly with the model's confidence gain until the entropy is fully resolved.

4 Experimental Setup

To effectively isolate active contextual processing from parametric memory, we engineered a multi-domain evaluation framework anchored in continuous, length-normalized token-level log-probabilities.

4.1 Datasets and Task Selection

We evaluated our framework across three established Question Answering (QA) corpora [42, 43] to ensure a comprehensive assessment of extractive capabilities. Specifically, we utilized Natural Questions (NQ-Open) to evaluate real-world, single-hop entity extraction, alongside the *rc.wikipedia* split of TriviaQA to assess entity-centric factoid extraction from structured contexts. We also incorporated the *distractor* split of HotpotQA, a dataset typically requiring reasoning across interconnected documents; however, due to the strict token generation limits imposed in our study to mathematically isolate pure entity extraction, multi-hop synthesis was inherently constrained.

To isolate the interaction between retrieved context and parametric memory, these evaluations were deliberately restricted to monolingual English corpora. Expanding to multilingual or low-resource languages introduces significant analytical noise driven by differing tokenization efficiencies and inherently weaker parametric priors. By strictly evaluating in English, where LLM parametric priors are overwhelmingly dominant, we subjected the RAG extraction process to the most stringent epistemic stress test possible. Furthermore, we employed stratified subsampling across these corpora to ensure statistical parity across reasoning typologies.

4.2 Model Selection and Paradigmatic Evaluation

To test the universality of scaling laws in extraction tasks, we selected architectures that allow both strict intra-family scaling comparisons and cross-paradigm benchmarking:

1. **SLM Baseline:** `Qwen2.5-1.5B-Instruct` [44]. Functions as an efficient, edge-deployable dense model.
2. **Mid-Weight Model:** `Qwen2.5-7B-Instruct`. Enables measurement of the marginal utility of a $\sim 5\times$ parameter increase within the identical architecture family.
3. **Massive Open-Weights Baseline:** `Qwen2.5-72B-Instruct`. Serves as our primary scaling upper-bound, isolating the effects of pure parameter scale while holding the pre-training philosophy and dense architecture relatively constant.
4. **Proprietary Commercial Archetype:** `gpt-4o-mini`. Included not for direct 1:1 architectural comparison, but as an archetypal representative of the closed-API paradigm: high-throughput, likely utilizing Mixture of Experts (MoE), and heavily subjected to proprietary safety alignments (e.g., RLHF).

Local inferences were executed on NVIDIA H100 hardware. The SLMs were run in `float16` precision, while the 72B model utilized quantized precision to accommodate memory constraints without meaningful performance degradation. Generation was strictly bounded to 5 tokens using greedy decoding ($T = 0.0$). A uniform zero-shot directive prompt was applied across all architectures to prevent model-specific prompt engineering from confounding the baseline epistemic behavior.

4.3 Context Perturbation Engine

As illustrated in Figure 3, for each question-answer pair (q, y), we synthesized four distinct conditions:

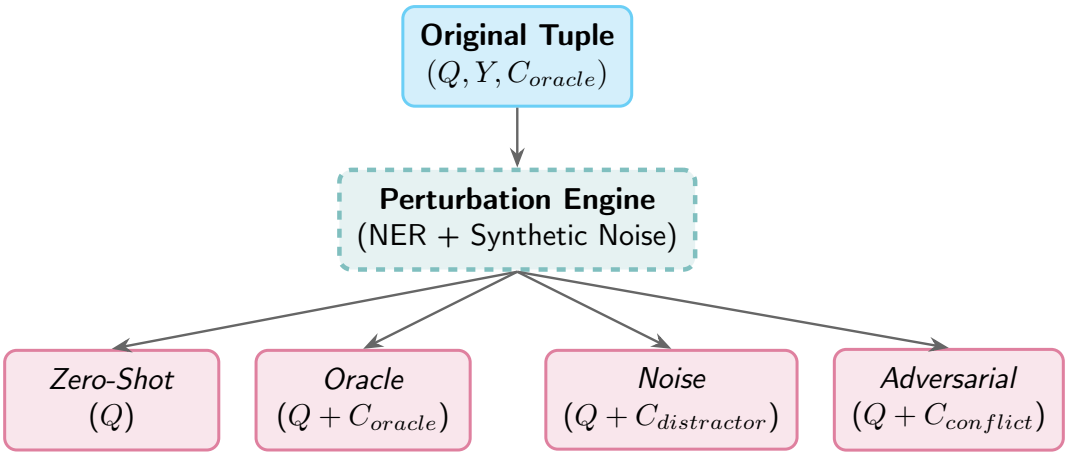

Figure 3: **Context Perturbation Engine Architecture.** The transformation of a single evaluation tuple into four controlled experimental conditions, aiming to isolate parametric memory from active context utilization.

- **Zero-Shot (c_{zero}):** The model is prompted solely with the query, establishing baseline predictive uncertainty.
- **Oracle (c_{oracle}):** The model is provided the ground-truth context containing the exact answer.
- **Adversarial Conflict (c_{adv}):** Utilizing a Named Entity Recognition (NER) pipeline, we systematically replaced the true answer in the oracle context with a fabricated entity. This quantifies *Prior Dominance*.
- **Semantic Noise (c_{noise}):** The oracle context is embedded centrally within semantically irrelevant synthetic text, acting as a control for context-length degradation.

4.4 Evaluation Metrics: Normalization Strategy

To mitigate length bias, sequence confidence is calculated as the arithmetic mean of the target token log-probabilities, ensuring equitable evaluation across models with distinct tokenization vocabularies. For discrete accuracy metrics, a generation is classified as a successful extraction strictly if its length-normalized probability satisfies $P > 0.05$. Given standard LLM vocabularies exceeding 100,000 tokens (where uniform random selection yields $P \approx 10^{-5}$), this conservative threshold isolates deterministic extraction by ensuring the model's predictive confidence operates magnitudes above stochastic noise.

5 Results and Analysis

Our empirical evaluation confirms a structural divergence between parametric scale and contextual processing efficacy. The results, synthesized in Table 1 and Figures 4 and 5, address our core research questions by revealing diminishing returns in traditional scaling paradigms for pure extraction tasks.

5.1 The Utilization Gap and Scaling Returns (RQ1)

Our primary objective was to ascertain whether scaling up model parameters intrinsically amplifies contextual extraction. The empirical data indicates severe diminishing returns for factual

Table 1: **Empirical Summary of Contextual Extraction.** Results derived from the balanced evaluation dataset (1,000 samples per corpus). Accuracy metrics represent discrete success rates ($P > 0.05$). To rigorously handle outliers, the bounded NCU is clamped within $[0, 1]$ *per-inference* prior to averaging, whereas the Raw NCU average remains unclipped to expose the magnitude of negative transfer.

Model	**Zero-Shot** (acc_zero)	**Oracle** (acc_oracle)	**Noise** (acc_noise)	**Conflict** ($acc\_conflict$)	**NCU**	**Raw NCU**	**Latency** (s/q)
Qwen 1.5B	53.6%	**95.9%**	**95.5%**	34.9%	0.864	0.620	**0.041**
Qwen 7B	51.9%	92.3%	92.3%	**33.1%**	0.864	-74.128	0.093
Qwen 72B	55.9%	91.8%	91.7%	42.5%	**0.868**	-20.882	0.389
GPT-4o-mini	**62.0%**	90.1%	89.2%	47.1%	0.777	-7779.254	2.982

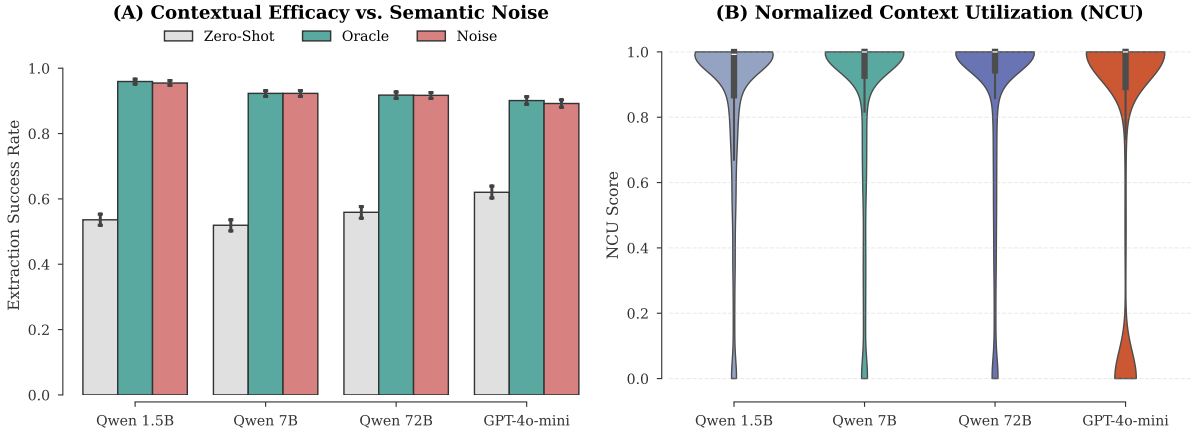

Figure 4: **Contextual Efficacy and NCU Distribution.** (A) The discrete success rates across Zero-Shot, Oracle, and Noise conditions. SLMs exhibit highly competitive context-grounded extraction. (B) The distribution of the bounded NCU scores, demonstrating the stable and high information gain achieved by the 1.5B and 7B architectures.

grounding. While the commercial baseline (`gpt-4o-mini`) achieved the highest zero-shot accuracy (62.0%), indicating superior parametric memorization, it exhibited the lowest oracle extraction accuracy (90.1%) and the lowest bounded NCU score (0.777).

In contrast, the highly efficient `Qwen-1.5B` model acted as an optimal contextual processor. It achieved a peak oracle accuracy of 95.9% and maintained an NCU of 0.864, effectively matching the continuous extraction capabilities of its 72B counterpart (NCU 0.868) while operating at a fraction of the computational cost (0.041s vs. 0.389s latency per query). This confirms that for strict extraction tasks, the relatively unconstrained nature of SLMs provides a structural efficiency advantage over heavily parameterized models.

5.2 Prior Dominance and Adherence (RQ2)

We investigated whether the depth of a model's parametric memory, or its proprietary alignment, induces a measurable resistance to integrating adversarial context. This phenomenon, measured as *Prior Dominance*, is illustrated in Figure 5A.

When the oracle context was perturbed to state a fabricated entity, the `Qwen-7B` and `1.5B` models exhibited high contextual adherence, surviving the conflict by adhering strictly to the provided text in roughly 65-67% of cases. Conversely, the commercial baseline rejected the adversarial context and defaulted to its parametric prior in 47.1% of the conflicts, despite explicit prompt instructions to synthesize answers strictly from the retrieved text. The open-weights `Qwen-72B` exhibited a 42.5% parametric override rate. This progression implies that while prior

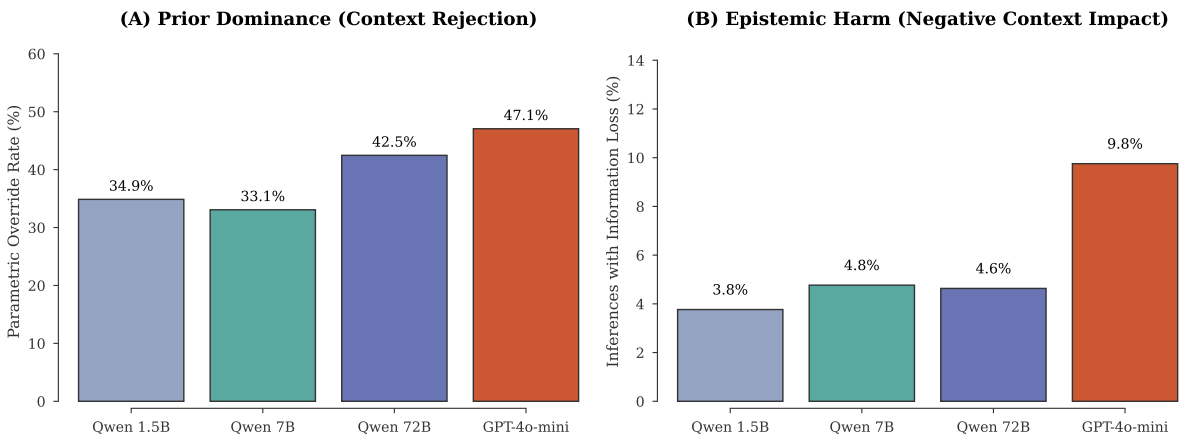

Figure 5: **Prior Dominance and Information Loss.** (A) Prior Dominance: The percentage of inferences where the model overrode explicit adversarial context in favor of its prior parametric knowledge. (B) Negative Transfer: The frequency of instances where confusing context actively degraded the model's predictive confidence below its zero-shot baseline.

dominance is heavily influenced by parameter scale (as seen in the jump from 7B to 72B), it may be further exacerbated by safety-oriented alignments present in proprietary APIs.

5.3 Noise Resilience and Negative Transfer (RQ3)

To ensure our NCU measurements were not artificially deflated by context length degradation, we evaluated the models' resilience to semantic noise. While discrete accuracy remained relatively stable across architectures when noise was introduced (Table 1), analyzing the unbounded *Raw NCU* revealed critical vulnerabilities beneath the surface.

Mathematically, a negative Raw NCU occurs when the posterior entropy given the retrieved context strictly exceeds the prior zero-shot entropy ($S_{post} > S_{prior}$). In other words, rather than resolving uncertainty, the external context acted as a destructive interference that actively degraded the model's confidence below its initial ignorance. From a Bayesian perspective, this collapse in confidence is an expected "surprise" when a highly confident prior is contradicted. However, operationally within a RAG pipeline, this constitutes a systemic vulnerability. The extreme negative magnitude observed in the raw average of the commercial baseline (-7779.254) is a mathematical artifact of its high initial confidence: dividing a negative information gain by a near-zero prior entropy ($S_{prior} \approx 0$) yields an asymptotic penalty.

As shown in Figure 5B, we quantified this *Negative Transfer*. The proprietary baseline suffered mathematical information loss in 9.8% of inferences, a rate nearly triple that of the `Qwen-1.5B` model (3.8%). This indicates that when high-capacity LLMs encounter conflicting or confusing out-of-distribution knowledge, their internal probability distributions can systemically collapse, rendering them highly unpredictable in strict processing workflows compared to SLM counterparts.

6 Discussion

The deployment of LLMs has historically relied on the heuristic that parametric scaling uniformly enhances all downstream capabilities. However, our evaluation using the NCU metric highlights a notable structural divergence: pure contextual extraction appears to operate on fundamentally different epistemic mechanisms from generalized reasoning.

This divergence manifests through the observed risk of *Negative Transfer* and Prior Dominance. We introduce the **"Alignment Tax Hypothesis"** to frame this behavior: the evaluated

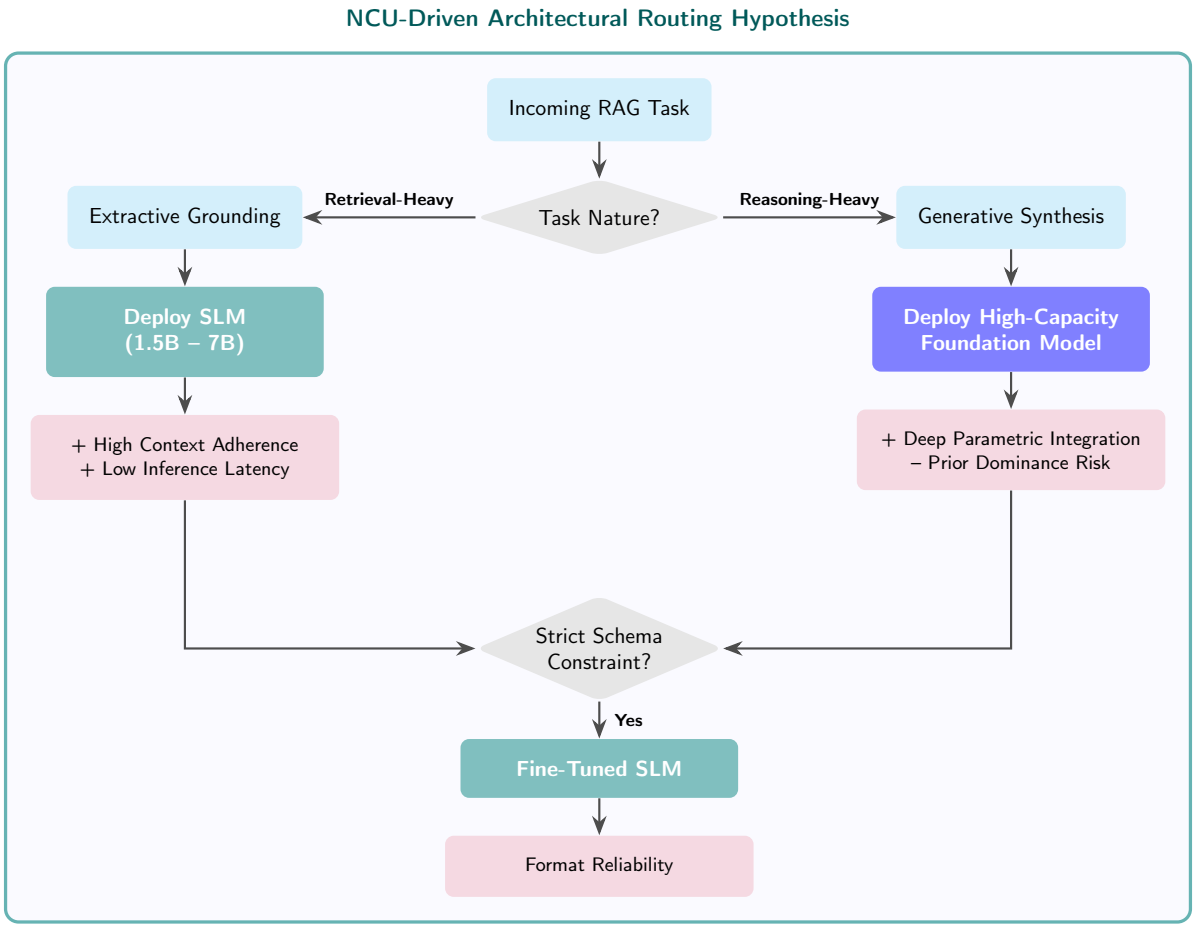

Figure 6: **Conceptual RAG Operational Routing.** Based on NCU evaluations, extractive tasks may be efficiently directed to SLMs to maximize context adherence and reduce latency. High-capacity models remain optimal for complex synthesis, provided the risk of parametric overwriting is managed.

commercial baseline exhibited a measurable tendency to default to its pre-training distributions when external evidence contradicted its high-confidence priors. In these instances, the model did not merely fail to extract the new information; its predictive probability distribution exhibited severe entropy degradation, leading to a negative transfer of information (a 9.8% negative transfer rate). While the extreme mathematical magnitude of the raw NCU penalty (-7779.25) is primarily an artifact of dividing by a near-zero initial entropy, the systemic confidence collapse remains a tangible vulnerability. It suggests a potential "tax" where heavy safety alignments or dense parametric memorization actively resist out-of-distribution updates.

Conversely, SLMs functioned more closely as unconstrained processors. They exhibited higher context adherence, robust noise resilience, and suffered less than half the negative transfer rate (3.8%). Achieving statistical parity in bounded NCU scoring (0.864 for 1.5B vs. 0.868 for 72B) alongside significantly lower inference latency (0.041s vs. 0.389s) suggests severe diminishing returns for parameter scaling in strict extraction tasks. These findings theoretically advocate for decoupled, modular RAG architectures where task routing is determined by epistemic requirements rather than a default reliance on high-capacity architectures (Figure 6).

7 Limitations and Future Work

While the NCU framework rigorously quantifies relative context utilization, our experimental design has distinct analytical boundaries. First, the strict token generation limit ($T \leq 5$),

necessary to mathematically isolate pure extraction probabilities, restricts our claims strictly to factual extraction. It precludes the evaluation of multi-hop synthesis or Chain-of-Thought (CoT) paradigms, where high-capacity models typically demonstrate emergent reasoning capabilities.

Second, our analysis of epistemic stubbornness relies on correlational observations rather than strict causal ablations. Comparing open-weight dense models directly to proprietary APIs introduces confounding variables such as Mixture of Experts (MoE) routing, undisclosed pre-training data mixtures, and unknown tokenization granularities. The opacity of the commercial baseline means that the exact mechanisms driving its negative transfer (e.g., specific RLHF penalties, constitutional AI constraints, or attention dilution) remain hypothesized under our "Alignment Tax" framework.

Finally, the adversarial conflicts generated via the Faker library are inherently synthetic. It is plausible that highly parameterized models possess latent anomaly detection capabilities, implicitly recognizing and rejecting these semantic forgeries rather than failing to comprehend the instruction.

Future Work: Subsequent research must conduct strict ablation studies on open-weights models (e.g., evaluating a Base model vs. its RLHF-Instruct counterpart) to definitively isolate the causal impact of safety alignment on Prior Dominance. Furthermore, expanding the evaluation across multiple commercial endpoints (e.g., Claude, Gemini) and diverse linguistic corpora will verify the universality of these epistemic vulnerabilities. Extending the length-normalized NCU metric to evaluate character-level cross-entropy and free-form generative reasoning warrants immediate exploration.

8 Conclusion

This study introduces the Normalized Context Utilization (NCU) metric to address the epistemic blindness inherent in discrete RAG evaluations. By analyzing continuous, length-normalized token probabilities across perturbed conditions, we observed significant diminishing returns associated with traditional parameter scaling for pure extraction tasks. The evaluated commercial model frequently exhibited Prior Dominance, actively losing predictive confidence when exposed to adversarial external evidence.

Conversely, SLMs in the 1.5B to 7B parameter regime demonstrated highly competitive contextual utilization and superior factual adherence. Coupling these epistemic characteristics with substantial improvements in inference latency suggests that deploying high-capacity, highly-aligned models for strict factual grounding may be computationally sub-optimal. The empirical insights derived from the NCU framework suggest that in the design of robust RAG pipelines, parametric scale must be carefully balanced with the need for epistemic flexibility.

Code Availability

To facilitate reproducibility and further research, the code, experimental framework, and datasets associated with this study are publicly available at https://github.com/BarakOr1/Quantifying-Prior-Dominance-in-RAG-Systems.

Acknowledgments

This research was independently conceived, conducted, and funded by ArtificialGate Ltd. All computational resources, proprietary data processing, and research hours were provided exclusively by the author's private firm. No institutional funding, facilities, or support from the author's academic affiliations were utilized in the development or execution of this study.

References

[1] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474, 2020.

[2] Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*, pages 2206–2240. PMLR, 2022.

[3] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. Retrieval augmented language model pre-training. In *International conference on machine learning*, pages 3929–3938. PMLR, 2020.

[4] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 6769–6781, 2020.

[5] 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.

[6] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, DDL Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 10, 2022.

[7] Zihan Chen, Song Wang, Zhen Tan, Xingbo Fu, Zhenyu Lei, Peng Wang, Huan Liu, Cong Shen, and Jundong Li. A survey of scaling in large language model reasoning. *arXiv preprint arXiv:2504.02181*, 2025.

[8] Zhengyu Chen, Siqi Wang, Teng Xiao, Yudong Wang, Shiqi Chen, Xunliang Cai, Junxian He, and Jingang Wang. Revisiting scaling laws for language models: The role of data quality and training strategies. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 23881–23899, 2025.

[9] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[10] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

[11] 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. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, 2023.

[12] Hung-Ting Chen, Michael Zhang, and Eunsol Choi. Rich knowledge sources bring complex knowledge conflicts: Recalibrating models to reflect conflicting evidence. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2292–2307, 2022.

[13] Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. Entity-based knowledge conflicts in question answering. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pages 7052–7063, 2021.

[14] 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*, pages 9802–9822, 2023.

[15] Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. Ragas: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th conference of the european chapter of the association for computational linguistics: system demonstrations*, pages 150–158, 2024.

[16] Jon Saad-Falcon, Omar Khattab, Christopher Potts, and Matei Zaharia. Ares: An automated evaluation framework for retrieval-augmented generation systems. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 338–354, 2024.

[17] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.

[18] Giovanni Trappolini, Florin Cuconasu, Simone Filice, Yoelle Maarek, and Fabrizio Silvestri. Redefining retrieval evaluation in the era of llms. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8359–8375, 2026.

[19] Wenqing Zheng, Dmitri Kalaev, Noah Fatsi, Daniel Barcklow, Owen Reinert, Igor Melnyk, Senthil Kumar, and C Bayan Bruss. Revisiting rag retrievers: An information theoretic benchmark. *arXiv preprint arXiv:2602.21553*, 2026.

[20] Kunlun Zhu, Yifan Luo, Dingling Xu, Yukun Yan, Zhenghao Liu, Shi Yu, Ruobing Wang, Shuo Wang, Yishan Li, Nan Zhang, et al. Rageval: Scenario specific rag evaluation dataset generation framework. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8520–8544, 2025.

[21] Shu Zhou, Yuxuan Ao, Yunyang Xuan, Xin Wang, Tao Fan, and Hao Wang. Inference scaling law for retrieval augmented generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 16522–16530, 2026.

[22] Karan Singh, Michael Yu, Varun Gangal, Zhuofu Tao, Sachin Kumar, Emmy Liu, and Steven Y Feng. To memorize or to retrieve: Scaling laws for rag-considerate pretraining. *arXiv preprint arXiv:2604.00715*, 2026.

[23] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[24] Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, et al. Textbooks are all you need. *arXiv preprint arXiv:2306.11644*, 2023.

[25] Giorgos Vernikos, Arthur Bražinskas, Jakub Adamek, Jonathan Mallinson, Aliaksei Severyn, and Eric Malmi. Small language models improve giants by rewriting their outputs. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2703–2718, 2024.

[26] Yi Chen, JiaHao Zhao, and HaoHao Han. A survey on collaborative mechanisms between large and small language models. *arXiv preprint arXiv:2505.07460*, 2025.

[27] Orlando Marquez Ayala, Patrice Bechard, Emily Chen, Maggie Baird, and Jingfei Chen. Fine-tune an slm or prompt an llm? the case of generating low-code workflows. *arXiv preprint arXiv:2505.24189*, 2025.

[28] Jerry Huang, Siddarth Madala, Risham Sidhu, Cheng Niu, Hao Peng, Julia Hockenmaier, and Tong Zhang. Rag-rl: Advancing retrieval-augmented generation via rl and curriculum learning. *arXiv preprint arXiv:2503.12759*, 2025.

[29] Jinhao Jiang, Jiayi Chen, Junyi Li, Ruiyang Ren, Shijie Wang, Wayne Xin Zhao, Yang Song, and Tao Zhang. Rag-star: Enhancing deliberative reasoning with retrieval augmented verification and refinement. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7064–7074, 2025.

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

[31] Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the association for computational linguistics*, 12:157–173, 2024.

[32] Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In *The Twelfth International Conference on Learning Representations*, 2023.

[33] Tianzhe Zhao et al. Exploring knowledge conflicts for faithful llm reasoning: Benchmark and method. *arXiv preprint arXiv:2604.11209*, 2026.

[34] Hua Ye, Siyuan Chen, Ziqi Zhong, Canran Xiao, Haoliang Zhang, Yuhan Wu, and Fei Shen. Seeing through the conflict: Transparent knowledge conflict handling in retrieval-augmented generation. *arXiv preprint arXiv:2601.06842*, 2026.

[35] Aisha Alansari and Hamzah Luqman. Large language models hallucination: A comprehensive survey. *Computer Science Review*, 61:100970, 2026.

[36] Xingyu Zhu, Junfeng Fang, Shuo Wang, Beier Zhu, Zhicai Wang, Yonghui Yang, and Xiangnan He. Mitigating hallucinations in large vision-language models without performance degradation. *arXiv preprint arXiv:2604.20366*, 2026.

[37] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yixin Dai, Jiawei Sun, Haofen Wang, Haofen Wang, et al. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2(1):32, 2023.

[38] Tu Vu, Mohit Iyyer, Xuezhi Wang, Noah Constant, Jerry Wei, Jason Wei, Chris Tar, Yun-Hsuan Sung, Denny Zhou, Quoc Le, et al. Freshllms: Refreshing large language models with search engine augmentation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13697–13720, 2024.

[39] Claude Elwood Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

[40] Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999.

[41] Barak Or. Kalman-inspired runtime stability and recovery in hybrid reasoning systems. *arXiv preprint arXiv:2602.15855*, 2026.

[42] 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*, pages 7969–7992, 2023.

[43] Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 11:1316–1331, 2023.

[44] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.