ArtificialGate

ArtificialGate Ltd. Research Preprint

From Likelihood Shifts to Defensible Claims: Auditing Fixed-Target Context Measures

Barak Or
ArtificialGate Ltd., Israel
barak@artificialgate.ai

Abstract

Likelihood shifts are increasingly used to ask whether context affects a language model, but the resulting numbers are easy to overinterpret. We study the common case in which an evaluator teacher-forces one fixed reference and compares its likelihood across serialized inputs. The score is a candidate-prefix log-ratio: it identifies a change in support for one specified token event, not a change in predictive entropy, a decoded answer, or a preference between alternatives. We prove two failure results that matter in practice. Normalizing the shift by query-only surprisal is ill-conditioned near a confident baseline and clipping the result erases every negative magnitude. In a conflict condition, scoring only the original reference cannot identify whether the model favors the replacement. We turn these observations into a four-level claim guide and a seven-item reporting checklist for the scored event, serialization, contrasts, interventions, inference, and artifacts. A corrected reference scorer and boundary tests implement the checklist without requiring new model runs. Finally, we audit an archived three-checkpoint question answering study. Its pooled summaries illustrate how apparently strong behavioral conclusions collapse to narrower statements once the scored event and evidence record are made explicit. The result is a practical audit that keeps fixed-target likelihood evidence aligned with the conclusions drawn from it.

Keywords. retrieval-augmented generation; language-model evaluation; fixed-target likelihood; context influence; measurement validity; reproducibility.

1 Introduction

Retrieval-augmented generation (RAG) places external text beside a model's parametric knowledge (Lewis et al., 2020; Borgeaud et al., 2022). An evaluation must then separate at least three questions: whether the supplied text changes the model's distribution, whether it changes the decoded answer, and whether the change improves factual performance. These questions often move together, but they are not equivalent. A model may already know the answer; a passage may alter token probabilities without changing the decoded string; and a poisoned passage may be influential precisely when it should not be trusted.

One diagnostic is especially attractive because it is simple: choose a reference answer, teacher-force it with and without context, and subtract the two log-likelihoods. The comparison can be informative, but it answers a narrower question than names such as "information gain," "uncertainty reduction," or "context preference" suggest. The target is fixed rather than averaged over the output distribution, and a score for one candidate contains no score for its competitors. Implementation details—chat serialization, token boundaries, termination, and length normalization—determine the event being measured.

We focus on this mismatch between computation and claim. The signed likelihood difference itself is not new: Liu et al. (2025a) use the same per-token contextual log-likelihood gain. Nor do we introduce another benchmark for context–memory conflict. Instead, we make three contributions.

This research was conducted under 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.

Table 1: Positioning relative to the closest lines of work. "Target" denotes a specified answer string.

Line of work	Primary object	Typical evidence	Distinction here
Context–memory conflict	Behavior under factual or counterfactual passages	Decoded answers, accuracy, or controlled interventions	We audit a teacher-forced token event rather than introduce a conflict benchmark
Target-likelihood diagnostics	Change in the likelihood of a specified answer	Token log-likelihoods and internal activations	We adopt the signed shift and characterize its claim boundary
Information-theoretic retrieval	Document contribution or retriever behavior	Answer-free scores, rankings, or retrieval distributions	We study a target-conditioned generator score
This work	Compatibility between computation, evidence, and claim	Formal results, code audit, reporting checklist	A record of which claims the evidence supports, not a new performance metric

1. We give an exact event-level interpretation of fixed-target likelihoods and prove when two common transformations—relative normalization and clipping—lose stability or information. We also give a finite construction showing why one-sided conflict scoring cannot identify a two-candidate comparison.
2. We provide a four-level claim guide and a reporting checklist that show what evidence each conclusion requires. The supplement includes an explicit-token scorer, tests, and a machine-readable inventory of artifact files and checksums.
3. We apply the audit to a public archive covering three Qwen2.5 Instruct checkpoints and four input conditions. The archived aggregates are used as a worked example, not as a new model comparison.

Relation to prior work. Conflict benchmarks study how generated answers respond to factual and counterfactual context (Longpre et al., 2021; Neeman et al., 2023; Xie et al., 2024; Du et al., 2024; Su et al., 2024). Other work examines how instruction tuning, task requirements, evidence position, and noise affect context reliance (Goyal et al., 2025; Shaier et al., 2024; Liu et al., 2024; Sun et al., 2026). Likelihood-based approaches provide a complementary view. Liu et al. (2025b) develop an answer-free information measure for RAG, while Zheng et al. (2026) analyze retriever rankings and distributions. Table 1 states the distinction directly. Our object is not a new likelihood statistic but a reusable audit of what a fixed-target statistic supports.

2 Fixed-Target Quantities and Their Claim Boundary

2.1 The scored event

Let $S(q, c)$ be a fully specified serialization of a question q and context c, including system text, chat markers, whitespace, and the generation prompt. Write $x_0 = S(q, \varnothing)$ and $x_c = S(q, c)$. Let $r^\star = (r_1, \dots, r_k)$ be one fixed sequence of target token IDs. The intended teacher-forced score is

$$\ell_\theta(r^\star \mid x) = \sum_{i=1}^{k} \ln p_\theta(r_i \mid x, r_{<i}). \tag{1}$$

The same token sequence $r^\star$ must be appended and scored in every condition. Encoding the concatenated prompt-plus-answer separately in each condition does not guarantee this property.

No end-of-sequence token appears in Equation 1. Hence $\exp(\ell_\theta)$ is the probability of the event $\{Y_{1:k} = r^\star\}$: the next k tokens equal the reference. It is a reference-prefix likelihood, not the probability of a terminated answer string. Its negative is prefix surprisal in nats, not predictive entropy; the latter would require an expectation over possible continuations.

The mean token score is

$$\bar{\ell}_\theta(r^\star \mid x) = \frac{1}{k}\ell_\theta(r^\star \mid x). \tag{2}$$

It reduces the direct contribution of reference length within one tokenizer. It does not make likelihoods comparable across tokenizers or calibrated across models.

The signed context contrast is

$$\begin{aligned} \Lambda_\theta(r^\star; x_c, x_0) &= \ell_\theta(r^\star \mid x_c) - \ell_\theta(r^\star \mid x_0) \\ &= \ln \frac{\Pr_\theta(Y_{1:k} = r^\star \mid x_c)}{\Pr_\theta(Y_{1:k} = r^\star \mid x_0)}, \end{aligned} \tag{3}$$

Table 2: Four levels of claim in fixed-target context evaluation. Each row requires the evidence listed there and in the rows above.

Level	Question	Minimum evidence	Claim licensed
1	Did support for $r^\star$ change?	Exact target IDs, serialized inputs, signed item-level contrast	Support for one specified prefix rose or fell
2	Which candidate is better supported?	Scores for both candidates, compatible length and termination conventions	Candidate-prefix comparison under that convention
3	What will the model answer?	Decoded outputs and a declared decision rule	Observed generation behavior under that rule
4	Is the behavior correct or robust?	Validated source quality, interventions, and task metric	Truthfulness or robustness for the evaluated setting

and $\bar{\Lambda}_\theta = \Lambda_\theta / k$. Positive values mean that x_c assigns greater support to this prefix than x_0; negative values mean the opposite. This is a pointwise candidate-prefix log-ratio. It does not reveal the decoded answer, the reference's rank, or the quality of the context.

The word *contrast* is important. Unless S is held fixed apart from the content under study, Λ_θ absorbs every serialization change. In the archive examined in Section 4, the query-only prompt also removes the `Context:` wrapper and two newlines. The resulting quantity is therefore not a content-only causal effect.

2.2 Why a one-sided conflict score is insufficient

Let $\tilde{r}$ be a replacement candidate in a counterfactual input x_{cf}. If both candidates are scored, their candidate-prefix log-ratio is

$$\Gamma_\theta(x_{\text{cf}}) = \ell_\theta(\tilde{r} \mid x_{\text{cf}}) - \ell_\theta(r^\star \mid x_{\text{cf}}). \tag{4}$$

This is not "answer odds" unless the two scored events and their termination rules justify that interpretation. Different token counts also introduce a length effect.

Proposition 1 (One-sided non-identification)**.** *Suppose the original and replacement prefixes branch before either is complete, so their prefix events are disjoint. Knowing only $p^\star = \Pr_\theta(Y_{1:k} = r^\star \mid x_{\text{cf}})$ does not identify Γ_θ. For $0 < p^\star < 1/2$, the same observed value is compatible with either candidate receiving greater prefix probability.*

Proof. Consider three disjoint continuation events: the original prefix, the replacement prefix, and all other continuations. Hold the original probability at $p^\star$. One valid distribution assigns the replacement $u_- = p^\star/2$ and the remainder $1 - p^\star - u_-$. A second assigns the replacement $u_+ = 1/2$ and the remainder $1/2 - p^\star$. Both remainders are non-negative because $p^\star < 1/2$. The resulting log-ratios $\ln u_- - \ln p^\star$ and $\ln u_+ - \ln p^\star$ have opposite signs. These distributions can be realized by a finite prefix tree whose first branch selects the three events. Thus the observed one-sided probability is compatible with opposite candidate comparisons. □

Proposition 1 concerns only what can be inferred from the observed score. It does not predict the model's decoded response; a one-sided score leaves that question unanswered. Complete-answer preference additionally requires a common termination rule and a decision rule over other continuations.

2.3 Relative normalization, clipping, and thresholding

Let

$$\bar{s}_0 = -\bar{\ell}_\theta(r^\star \mid x_0), \qquad \bar{s}_c = -\bar{\ell}_\theta(r^\star \mid x_c).$$

For $\bar{s}_0 > 0$, the archived relative score is

$$R_\theta = \frac{\bar{\Lambda}_\theta}{\bar{s}_0} = 1 - \frac{\bar{s}_c}{\bar{s}_0}. \tag{5}$$

The factor $1/k$ cancels. Per-token averaging therefore supplies no extra invariance to this ratio.

Proposition 2 (Instability and information loss)**.** *For $\bar{s}_0 > 0$ and $\bar{s}_c \geq 0$, $R_\theta \leq 1$, and*

$$\frac{\partial R_\theta}{\partial \bar{s}_c} = -\frac{1}{\bar{s}_0}, \qquad \frac{\partial R_\theta}{\partial \bar{s}_0} = \frac{\bar{s}_c}{\bar{s}_0^2}.$$

Consequently, the normalization is arbitrarily sensitive as $\bar{s}_0 \to 0^+$. Moreover, $\text{BRTS} = \max(0, \min(1, R_\theta))$ *is non-injective on $(-\infty, 0]$.*

Proof. Equation 5 and $\bar{s}_c \geq 0$ imply $R_\theta \leq 1$; the upper clip is redundant for valid probabilities. Differentiation gives the stated sensitivities. Finally, every $R_\theta \leq 0$ maps to zero, so the sign within that set and every negative magnitude are unrecoverable from BRTS. □

Table 3: Minimum reporting checklist for fixed-target context measures.

Component	Required record	Failure prevented
Scored event	Exact target token IDs; whether EOS is included; token count	Confusing a prefix with a complete answer
Serialization	System/user text, chat template, whitespace, and generation marker	Attributing wrapper changes to context content
Boundary validation	One encoded target reused across conditions; scored-suffix assertion	Prompt/answer token slicing errors
Primary contrast	Signed sequence log-ratio and item-level values; per-token form secondary	Loss of direction and aggregate-only inference
Candidate claim	Both candidates, common termination rule, lengths, decoded decision rule	Inferring preference from a one-sided score
Intervention record	Construction, success flag, leakage, plausibility, length, and position	Equating a compound perturbation with one mechanism
Inference and artifact	Paired unit IDs, dataset/prompt strata, uncertainty, revisions, manifest, tests	Unsupported equivalence, scaling, or reproducibility claims

Corollary 1 (Error amplification)**.** *If $\bar{s}_c$ is perturbed by ε while $\bar{s}_0$ is fixed, then the relative score changes by $-\varepsilon/\bar{s}_0$.*

Thus an absolute scoring discrepancy of 0.01 nats per token changes R_θ by 0.01 when $\bar{s}_0 = 1$, but by 0.5 when $\bar{s}_0 = 0.02$. This is an algebraic sensitivity calculation, not an empirical estimate. At $\bar{s}_0 = 0$, the ratio is undefined and should be reported as a boundary case rather than smoothed into a finite value.

The original study also reports

$$\mathrm{Supp}_{.05}(r^\star; x) = \mathbb{1}\left[\exp\left(\bar{\ell}_\theta(r^\star \mid x)\right) > 0.05\right]. \tag{6}$$

The exponential is the geometric mean of reference-token probabilities. The threshold is equivalent to $\bar{\ell}_\theta > \ln(0.05) \approx -2.996$ nats per token. Without calibration or a sensitivity analysis, this indicator has no model-invariant interpretation. It is neither accuracy nor rank.

3 A Minimum Reporting Checklist

A useful reporting rule follows: define the event, the contrast, and only then the decision claim. Table 3 lists the minimum record needed for a fixed-target likelihood audit. Researchers can use the checklist when designing a study or checking a released implementation.

Primary and secondary quantities. The signed sequence contrast in Equation 3 should be primary because its event is explicit and its direction is preserved. Report its distribution at the experimental-unit level. Per-token contrasts, relative normalizations, and threshold indicators may be useful secondary views, but they should not replace the signed values from which they are derived. Cross-model comparisons require particular care because tokenization and calibration differ.

A claim–evidence record. Before drawing a conclusion, an evaluator should record four fields: *claim*, *scored event*, *supporting evidence*, and *unresolved alternatives*. A claim is admissible only if the evidence reaches the corresponding level in Table 2. For example, a positive Λ_θ supports "this serialization increased support for $r^\star$." It does not support "the model used reliable evidence" unless the intervention and decoded behavior are separately validated. This format shows exactly where a conclusion lacks evidence while preserving claims that remain supported.

Reference implementation. The supplement contains `reference_prefix_scorer.py`. It encodes the target once, appends those token IDs explicitly to every prompt, gathers the next-token log-probabilities at exactly the target positions, and returns both the sequence sum and per-token mean. Unit tests check target reuse, label masking, scored-position alignment, and rejection of empty targets. The implementation documents the intended scoring procedure; no new checkpoint result in this paper is attributed to it.

4 Worked Audit of an Archived Study

We now apply the checklist to the released materials of a question answering study. This section has two purposes: to demonstrate the audit on real code, and to preserve the valid descriptive content of the archive. Because the item-level records are unavailable, we do not attempt a statistical re-analysis.

Table 4: Intervention audit. A single archived contrast often changes more than the named factor.

Contrast	Intended factor	Factors that co-vary	Missing validation
Query to answer bearing	Presence of answer-bearing text	Wrapper; question repetition in NQ and TriviaQA; target copying	Content-only control
Answer bearing to padded	Distraction	Length and target position	Token lengths and isolated-factor controls
Answer bearing to counterfactual	Conflicting answer	Fluency, plausibility, entity type, possible failed replacement	Success log and human validation
Across checkpoints	Checkpoint capability	Parameter count, precision, quantization, and post-training	Matched precision and base/instruct controls

Table 5: Unverified archived pooled reports. Q: query-only support; A: answer-bearing support; $\Delta A = A - Q$; ΔP = padded minus answer-bearing; C: original-reference support after attempted replacement. Support entries, deltas, and Neg. are percentage points; BRTS is dimensionless. Values were transcribed, not recomputed.

Model	Q	A	ΔA	ΔP	C	BRTS	Neg.
Qwen2.5-1.5B	53.6	95.9	42.3	-0.4	34.9	0.864	3.8
Qwen2.5-7B	51.9	92.3	40.4	0.0	33.1	0.864	4.8
Qwen2.5-72B-AWQ	55.9	91.8	35.9	-0.1	42.5	0.868	4.6

Data and conditions. The script is configured to sample 1,000 validation items from each of HotpotQA, Natural Questions Open, and TriviaQA with seed 42 (Yang et al., 2018; Kwiatkowski et al., 2019; Joshi et al., 2017). HotpotQA uses sentences identified by supporting-fact titles. Natural Questions uses only `answer[0]`, and TriviaQA uses `answer.value` while ignoring aliases. For the latter two datasets, the "context" repeats the question and states the reference explicitly. We therefore call these *answer-bearing inputs*, not retrieved evidence.

Four serializations are scored: query only, answer bearing, answer bearing with four Faker paragraphs, and a counterfactual transformation. The counterfactual path uses `en_core_web_sm`, maps selected entity types to Faker values, and applies exact string replacement. Replacement success, grammar, and plausibility were not logged. The padded path inserts two paragraphs before and two after the answer-bearing text. Table 4 records what changes with each intended treatment.

Models and implementation. The local checkpoints are `Qwen/Qwen2.5-1.5B-Instruct`, `Qwen/Qwen2.5-7B-Instruct`, and `Qwen/Qwen2.5-72B-Instruct-AWQ` (Qwen Team, 2024). The smaller models use FP16; the 72B checkpoint uses 4-bit Activation-aware Weight Quantization (AWQ) with group size 128, zero points, a matrix-multiplication (GEMM) backend, and FP16 compute. Revisions and package versions were not pinned.

The archived scorer serializes the prompt and prompt-plus-answer separately, then slices the second tokenization at the length of the first. Subword tokenization need not preserve that boundary. The archive contains no prefix equality assertion or decoded-suffix check. This is a material implementation uncertainty: the aggregate record is not a verified evaluation of Equation 1.

Evidence status. Only rounded pooled aggregates and source figures are available. Per-item scores, model outputs, actual condition denominators, and the locked environment were not released. Table 5 is therefore labeled an *unverified archived report*. The accompanying script checks that the table was copied correctly and that the reported differences are arithmetically consistent; it does not evaluate a model. The available files therefore support no confidence intervals, dataset-specific conclusions, or causal claim about scaling.

What remains observable. Answer-bearing support exceeds query-only support by 35.9–42.3 points in the transcribed table. Because two dataset templates state the answer verbatim, the narrow reading is that explicit target-bearing inputs often move the reference above this particular threshold. The observation does not separate retrieval, reading comprehension, question repetition, and copying.

The padded rate differs from the answer-bearing rate by at most 0.4 points. The threshold decisions are therefore similar in this specific padding test. This is not evidence of robustness to natural retrieval noise: the treatment combines distraction, length, and position, and continuous item-level shifts are unavailable.

Table 6: What the worked audit establishes and what it does not.

Claim	Evidence	Status
Relative normalization is ill-conditioned near zero baseline surprisal	Equation 5, Proposition 2, and Corollary 1	Established analytically
Clipping removes negative magnitude	Definition of BRTS and Proposition 2	Established analytically
One-sided conflict scoring does not identify the replacement comparison	Proposition 1	Established analytically
The archive contains the pooled values in Table 5	Source figures, manuscript table, hashes, and transcription check	Established as provenance, not recomputation
A particular checkpoint is more context-reliant, accurate, or robust	Would require item-level validated interventions and higher claim-ladder evidence	Not established

In the attempted counterfactual condition, the original reference remains above threshold for 33.1–42.5% of pooled items. The 72B-AWQ value is 7.6–9.4 points above the smaller checkpoints. It cannot be read as preference for the original answer because the replacement was not scored and successful substitution was not verified. It cannot be attributed to model scale because size and precision are confounded.

Finally, mean BRTS lies between 0.864 and 0.868 even though 3.8–4.8% of answer-bearing shifts are negative. Clipping maps all of those cases to zero. Without their item-level magnitudes, the apparent similarity may reflect similar behavior, compression by the transformation, or both.

5 Implications and Limitations

Fixed-target likelihoods are most useful as diagnostics, not verdicts. They can detect changes that decoded accuracy misses and can expose where a prompt alters support for a specified continuation. The signed contrast is useful when the question concerns one named target. It should not be used to infer a decoded answer or factual robustness without the additional evidence listed in Table 2.

The reporting checklist is useful in three settings. During study design, it forces the target event and termination convention to be chosen before results are inspected. During artifact review, it reveals whether the code actually scores that event. During model monitoring, it preserves signed item-level changes while keeping preference, generation, and source reliability as separate measurements. The framework applies beyond RAG wherever a fixed continuation is teacher-forced across prompts, interventions, or model checkpoints.

Limitations of the contribution. The formal results concern fixed token-prefix events; they do not solve calibration, open-ended generation evaluation, or causal identification of context use. The checklist is a reporting proposal, not a statistically validated measurement scale. Its value will depend on whether other studies can apply it consistently and reproduce the intended scored event.

Limitations of the worked audit. Because item-level scores, boundary checks, intervention validation, and matched precision are missing, the archive cannot support model rankings or robustness conclusions. These omissions do not affect the algebraic results or the provenance statement that Table 5 accurately transcribes the archive. This distinction preserves the results that can still be checked without overstating what the archive proves.

Broader impact. Following context is not the same as being reliable. Optimizing a model or router for likelihood gain alone could reward compliance with poisoned, outdated, or irrelevant text. Conversely, resistance to a passage may reflect useful knowledge or a failed intervention. Fixed-target measures should be paired with source-quality checks, decoded task performance, and explicit threat models. Table 2 keeps these questions separate.

Conclusion. A fixed-target likelihood shift answers a precise question: how did one serialized input change the probability of one token-prefix event? That question is worth asking. The answer becomes scientifically useful when the event, boundary, contrast, and evidence record are explicit. Relative normalization, clipping, and one-sided conflict scores do not supply the missing information. The reporting checklist developed here turns those limitations into a constructive protocol for future evaluations.

Reproducibility. The supplement provides the manuscript source, the unchanged archived script, a corrected reference scorer and tests, the aggregate table, source figures, and a checksum manifest. Its validation command checks these files without running a model. Because the original item-level outputs are missing, the package does not reproduce the original model runs.

References

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego de Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack Rae, Erich Elsen, and Laurent Sifre. Improving language models by retrieving from trillions of tokens. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 2206–2240. PMLR, 2022.

Kevin Du, Vésteinn Snæbjarnarson, Niklas Stoehr, Jennifer White, Aaron Schein, and Ryan Cotterell. Context versus prior knowledge in language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13211–13235. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.714.

Sachin Goyal, Christina Baek, J. Zico Kolter, and Aditi Raghunathan. Context-parametric inversion: Why instruction finetuning can worsen context reliance. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=SPS6HzVzyt.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1601–1611. Association for Computational Linguistics, 2017.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466, 2019. doi: 10.1162/tacl_a_00276. URL https://aclanthology.org/Q19-1026/.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9459–9474, 2020.

Jiarui Liu, Jivitesh Jain, Mona Diab, and Nishant Subramani. LLM microscope: What model internals reveal about answer correctness and context utilization. *arXiv preprint arXiv:2510.04013*, 2025a. doi: 10.48550/arXiv.2510.04013. URL https://arxiv.org/abs/2510.04013.

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. doi: 10.1162/tacl_a_00638. URL https://aclanthology.org/2024.tacl-1.9/.

Tianyu Liu, Jirui Qi, Paul He, Arianna Bisazza, Mrinmaya Sachan, and Ryan Cotterell. Pointwise mutual information as a performance gauge for retrieval-augmented generation. 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)*, pp. 1628–1647. Association for Computational Linguistics, 2025b. doi: 10.18653/v1/2025.naacl-long.78. URL https://aclanthology.org/2025.naacl-long.78/.

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*, pp. 7052–7063. Association for Computational Linguistics, 2021.

Ella Neeman, Roee Aharoni, Or Honovich, Leshem Choshen, Idan Szpektor, and Omri Abend. DisentQA: Disentangling parametric and contextual knowledge with counterfactual question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10056–10070. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-long.559.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. URL https://arxiv.org/abs/2412.15115.

Sagi Shaier, Lawrence Hunter, and Katharina von der Wense. Desiderata for the context use of question answering systems. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 777–792. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.eacl-long.47. URL https://aclanthology.org/2024.eacl-long.47/.

Zhaochen Su, Jun Zhang, Xiaoye Qu, Tong Zhu, Yanshu Li, Jiashuo Sun, Juntao Li, Min Zhang, and Yu Cheng. ConflictBank: A benchmark for evaluating the influence of knowledge conflicts in LLMs. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-3280.

Kaiser Sun, Fan Bai, and Mark Dredze. Task matters: Knowledge requirements shape LLM responses to context–memory conflict. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 4154–4176. Association for Computational Linguistics, 2026. URL https://aclanthology.org/2026.findings-acl.202/.

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*, 2024. URL https://openreview.net/forum?id=auKAUJZMO6.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2369–2380. Association for Computational Linguistics, 2018.

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. URL https://arxiv.org/abs/2602.21553.

A Implementation and Provenance

A.1 Prompt and input serialization

The exact archived messages are:

```
System: You are a highly precise QA system. Answer concisely with ONLY the
requested entity, fact, or short phrase. No pleasantries, no full sentences.
With context: Context: {context}; Question: {question}.
Without context: Question: {question}.
```

The chat template is applied with `add_generation_prompt=True`. The local path then concatenates the answer at the string level and scores a sliced suffix without EOS.

A.2 Archived code-to-estimand map

The archived fields map directly to the definitions in the paper: `prob_zero` is the exponential of the mean sliced-suffix log-probability; `acc_*` applies the threshold in Equation 6; `cus` is the per-token signed contrast; `raw_ncu` is the relative score in Equation 5; `ncu` is BRTS; and `latency` is the wall time for four scoring calls. The unchanged script is retained as evidence of the original path, not as the corrected scorer. It tokenizes `prompt_text` and `prompt_text + answer` independently, sets `prompt_len` from the former, and slices the latter at that index.

A.3 Corrected scorer contract

The reference implementation follows five invariants: it encodes the target once with special tokens disabled; reuses those IDs in every condition; concatenates prompt and target IDs explicitly; gathers only the logits that predict the appended target positions; and returns token-level values, the sequence sum, the mean, and the token count. If a complete-answer probability is intended, the caller must append and score an explicit common termination token. The unit tests do not load a model or recreate the archived results.

A.4 Checkpoint configuration and aggregate provenance

The archived identifiers are `Qwen/Qwen2.5-1.5B-Instruct`, `Qwen/Qwen2.5-7B-Instruct`, and `Qwen/Qwen2.5-72B-Instruct-AWQ`. The loader passes `torch_dtype=torch.float16`. The 72B configuration records 4-bit AWQ, group size 128, zero points, and a GEMM backend. Neither model nor tokenizer commits are recorded.

`reported_aggregates.csv` contains every number in Table 5. Support rates and mean BRTS were transcribed from the archived manuscript table; negative-shift rates were transcribed from `Fig2_Epistemic_Stubbornness_and_Harm.PDF`. `MANIFEST.json` records the roles and SHA-256 hashes of the immutable archive files. `validate_package.py` checks those hashes, validates the aggregate arithmetic, and runs the scorer unit tests. These operations perform no model inference.