

Balanced Excerpt Regrouping in a Frozen Grounding Verifier

Ismail ZRIGUI

`izrigui@uae.ac.ma`

Abdelmalek Essaadi University: Universite Abdelmalek Essaadi <https://orcid.org/0000-0002-3836-6310>

Khoulji samira

Abdelmalek Essaâdi University

Research Article

Keywords: retrieval-augmented generation, grounding verification, representation sensitivity, repeated evidence, exploratory evaluation

Posted Date: October 18th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11373865/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

Abstract

Purpose

We examine how copying and regrouping evidence already present in a context changes the grounding score of one frozen MiniCheck checkpoint. The study uses a finite control to test this sensitivity.

Methods

Nine verbatim excerpts form a concentrated arm and two distributed arms. Each augmented input appends three excerpts. Across each arm's variants, all nine have identical marginal exposure at every appended slot. Averaging all six orders of each fixed triple eliminates differences attributable solely to additive excerpt-by-slot contributions.

Results

For 18 adaptively selected RAGTruth cases with 91 segments, the source-equal raw-logit contrast between concentrated-to-distributed and distributed-to-distributed separation was 0.400061. All 18 source contrasts were positive; six segment contrasts were negative. A separately specified, prospective HotpotQA-derived cohort yielded 0.348841 across 24 cases, with 23 positive and one negative. These inputs paired two oracle-selected supporting paragraphs with one fixed distractor and one-to-eight-token reference answers, unlike generated RAG responses. The concentrated arm lay above both distributed arms in 17 HotpotQA cases and below both in six. These contrasts concern score geometry, not uniform confidence inflation. A later original-label utility feasibility gate stopped before neural inference, leaving accuracy and practical benefit unmeasured.

Conclusion

Matched marginal exposure revealed a bounded non-additive response in this verifier's score geometry. Its causal mechanism, generalization across models, and practical usefulness remain unresolved.

1 Introduction

A grounding verifier receives a question, three retrieved passages, and an answer segment. If copies of passage fragments are then appended, no new observations are supplied, but the fragments' positions and combinations change. Fragments from one passage may overlap or repeat its label; those from different passages may cover different topics. Any of these properties could change the support score. The change alone therefore cannot identify its cause.

We ask whether equal-length representations can separate in raw-score space when marginal excerpt-by-slot exposure is exactly balanced. Starting from three passages, we construct a 3×3 excerpt matrix, retain the complete original evidence, and append three excerpts per input. The arms differ in which excerpts occur together. Balance covers all nine excerpts across each arm’s variants; it does not require all nine to appear in every input. A score made up only of separate additive excerpt-by-slot contributions has identical arm means. Nonzero differences rule out that explanation alone, while leaving the responsible interactions unidentified.

We present this finite measurement design and traceable observations from two distinct cohorts evaluated with one checkpoint. The RAGTruth cohort is an adaptive engineering sample. The HotpotQA-derived protocol was fixed before validation acquisition and scoring and uses an explicitly adapted task. Every negative result is retained, and the cohorts are reported separately. These observations do not establish a useful new factuality metric or priority over earlier robustness work.

2 Related work

RAGTruth contains generated responses and hallucination annotations. Its QA records combine MS MARCO training queries with three retrieved passages [1], but the annotations do not validate our augmented inputs. MiniCheck includes a RoBERTa grounding verifier initialized from AlignScore [2]. We use that fixed checkpoint with our own document serialization and regular-expression segmentation, rather than the complete MiniCheck evaluation pipeline. HotpotQA provides questions, reference answers, paragraph contexts, and supporting-fact metadata [14]. We use the metadata to select evidence for the second cohort; this adaptation does not constitute a standard HotpotQA accuracy evaluation.

Prior evaluations have already examined evidence order and context composition. Roh and Bae [15] study factual decisions across six models and 774 FEVER and 610 SciFact claims. Their order-only condition permutes a fixed, balanced set of six passages; other conditions vary evidence ratios or quantity. Chen et al. [16] show that premise order affects deductive and mathematical reasoning, with a forward-order condition that follows a valid proof. Lost in the Middle [7] examines position effects in multi-document QA and key–value retrieval. Generic order sensitivity is therefore insufficient as a novelty claim. Here the question is whether repeated excerpts can be regrouped under exact marginal balance and still yield a raw-logit contrast while the original context is retained. The endpoint measures that contrast, not answer accuracy or label flips.

Bouke’s context-sensitivity preprint [17] evaluates controlled context perturbations as diagnostics using original factuality labels. It reports gains in some configurations, including per-chunk MiniCheck, but no statistically significant fusion gain for full-context MiniCheck. The comparison highlights why sensitivity needs to be tested for discrimination rather than treated as evidence of error. Our perturbations and scoring construction differ, and we have no completed original-label utility result. The utility findings in that paper cannot be transferred to this study.

Several metric-robustness studies instead change the answer or claim. Ramprasad and Wallace compare factual corrections, benign summary edits, and added phrases, including a Flan-T5 MiniCheck condition [3]. Mujahid et al. examine long-document summarization with Bespoke-MiniCheck-7B [4]. FactEval perturbs claims while retaining the evidence and FEVER labels [5], and MetaRAG mutates answer factoids [6]. These studies use different checkpoints and outcomes from our diagnostic, which changes the evidence representation while fixing the answer.

Janossy Pooling averages a sequence-sensitive function over permutations [8]. Our complete six-order average is likewise invariant to a fixed triple's initial ordering, but not to arbitrary regrouping. Other related work addresses different questions: SURE-RAG studies evidence sufficiency [9], AdaGReS selects context by relevance and redundancy [10], and VP-CONTROL and Bara distinguish repeated reports from independent evidence lineages [11, 12]. Lee and Kim evaluate coordinated evidence poisoning through answer outcomes [13]. Passage identifiers in our design do not certify independent lineages, and we test no mitigation. These comparisons locate the study within related work without establishing priority or superiority.

3 Methods

3.1 Cohorts and chronology

The RAGTruth pool contained 24 QA training groups previously used for input engineering. The initial sentence-window construction failed for all 24 groups and 139 quality-good answers. Switching to fixed-token verbatim fragments produced matrices for every group and eligible inputs for 92 answers in 18 groups. We retained every eligible group. The quality-good flag determined eligibility; it was not a factual label. A frozen identifier-hash rule selected one answer per group, preventing selection by effect size without making this adaptive pool confirmatory. The 18 answers contained 91 complete regex-separated segments, with one to eleven per answer. Eligibility required one to twelve segments and inputs no longer than 512 tokens. Online Resource 2 (Supplement S2) preserves the exact selection rule and chronology.

The first C/D scoring used an answer-level probability endpoint. After inspecting those results, we chose E and the raw-logit endpoint and locally froze them before scoring E. The missing reflected orders were specified in a later freeze, after inspection of the cyclic results. Only the added measurements were prospective, and neither freeze was an external preregistration. Earlier corpus-wide checks of structure, duplicates, annotation counts, and offsets also mean that RAGTruth validation cannot be described as wholly untouched or label-blind.

We fixed the separate HotpotQA-derived protocol before validation acquisition, selection outcomes, or scoring. A frozen hash rule reduced 7,405 validation IDs to 1,024 candidates. Examining the first 39 yielded 24 eligible cases: fourteen exceeded the length limit and one lacked ten context paragraphs. The remaining 985 candidates were not examined. Metadata determined two complete supporting

paragraphs and one hash-chosen distractor before text or length could influence those choices. Each complete reference answer was kept as one segment, with one to eight tokenizer tokens (mean 3.0833), and was not rewritten as a declarative claim. This oracle 2 + 1 context with reference answers differs both from the published two-support/eight-distractor task [14] and from generated RAG responses.

The same excerpt rule and first context-only matrix were used, together with all six orders, the eighty-token increment, slot checks, and the 512-token ceiling. Cases were not rescued by changing the matrix or distractor, truncating inputs, or rewriting answers. Screens for known exposure and exact duplicates do not establish independence of semantic families or absence from training. The pinned community conversion has not been shown byte- or value-equivalent to the unavailable original distribution. A 24-example training preview and native buffers that may have touched all validation rows are disclosed in Online Resource 2. Prospective selection therefore does not imply that every noncandidate record was literally unseen.

3.2 Balanced excerpt construction

For each source, $e[j,k]$ denotes the excerpt in row j and column k of a frozen 3×3 matrix, where j and k take values 0, 1, and 2. Each row identifies one of the three retrieved passages. An excerpt is a distinct verbatim fragment bounded by whitespace-delimited words and containing sixteen tokens under the pinned tokenizer. Excerpts from the same passage may overlap. The original question and complete numbered passage block are retained byte-for-byte. A separate appended block contains three excerpts with their original passage labels.

For anchor a , C uses $e[a,k]$, D uses $e[(a + k) \bmod 3,k]$, and E uses $e[(a - k) \bmod 3,k]$. C groups three fragments from one passage; D and E each combine one fragment from every passage. Table 1 gives the anchor-zero triples. The initial cyclic column orders, (0,1,2), (1,2,0), and (2,0,1), produce nine variants per arm across the three anchors. Order completion adds (0,2,1), (2,1,0), and (1,0,2), so the full set contains each of the six orders once per triple and eighteen variants per arm.

Table 1
Unrotated triples for anchor $a = 0$.
Anchors 1 and 2 supply the other
triples. Each triple is evaluated in all
six orders; each individual
augmented input appends three
excerpts.

Arm	Slot 1	Slot 2	Slot 3
C	$e[0,0]$	$e[0,1]$	$e[0,2]$
D	$e[0,0]$	$e[1, 1]$	$e[2, 2]$
E	$e[0,0]$	$e[2, 1]$	$e[1, 2]$

In either nine-variant half, each excerpt occupies every appended physical slot once. In the full eighteen-variant arm, it occupies each slot twice. An individual input still contains only one three-excerpt triple, so balance holds across an arm’s variants, not through pairwise C/D or C/E content matching. The original question and passage block remain byte-identical. Every variant adds exactly eighty tokens, including labels and separators, without truncation. We check full-input offset mappings, decoded slices, and absolute positions to avoid the boundary-merge errors caused by prefix re-tokenization.

The distributed partitions cover a restricted set of regroupings. Anchoring the three excerpts of one passage yields $3! \times 3! = 36$ partitions whose triples each span all three passages. Only D and E also span all three matrix columns. Thus they exhaust the partitions satisfying both restrictions, rather than all 36 passage-spanning partitions or arbitrary regroupings. This enumeration was made post hoc to describe the design; it is not a model result.

3.3 Verifier and endpoint

We used MiniCheck-RoBERTa-Large at revision 74c8919647e61ed0f71bc177d94f10930f090068 on CPU in FP32, with batch size one, eager attention, seed zero, evaluation and inference modes, and deterministic algorithms. We neither trained the model nor fitted a decision threshold. Online Resource 2 records the checkpoint hash, software versions, thread settings, exact completed call counts, and runtime history. At each inference stage, the fixed longest input was repeated twice with zero weight. Neither those repeats nor the original-context controls enter the contrast.

For case i , segment s , arm $A \in \{C,D,E\}$, and variant v , z is the class-1 minus class-0 logit margin. We use $K = 9$ for either order half or 18 for the full set, and $N = 18$ for RAGTruth or 24 for HotpotQA. The contrast compares the concentrated-to-distributed distance with the distributed-to-distributed distance:

$$z_{Aisv} = \ell_{1,Aisv} - \ell_{0,Aisv}, \quad \mu_{Ais} = \frac{1}{K} \sum_{v=1}^K z_{Aisv} \quad (1)$$

$$q_{is} = \frac{1}{2} (|\mu_{Cis} - \mu_{Dis}| + |\mu_{Cis} - \mu_{Eis}|) - |\mu_{Dis} - \mu_{Eis}| \quad (2)$$

$$T_i = \frac{1}{S_i} \sum_{s=1}^{S_i} q_{is}, \quad T = \frac{1}{N} \sum_{i=1}^N T_i \quad (3)$$

S_i is the number of complete segments in case i . We first average raw margins within each arm and segment, then calculate distances. Segment contrasts are averaged within each case, and cases receive equal weight within their cohort. T_{full} is not the mean of T_{cyclic} and $T_{reflected}$, and margins are not averaged across segments before taking absolute values. The cohorts are never pooled. Neither

repeated variants nor segments are independent cases. Calculations use accurate floating-point summation of saved FP32 logits; softmax and answer minima do not enter T.

The HotpotQA protocol defined $T_{\text{full}} > 0$ as directional concordance on this derived cohort and $T_{\text{full}} \leq 0$ as non-replication of that sign. The experiment would be incomplete if calls were missing, repeats failed, integrity errors occurred, or resources forced a stop; reporting only part of the cohort was prohibited. We calculated no p-values, population intervals, fitted thresholds, factual-performance estimates, or practical-effect thresholds.

3.4 Exact control and nonlinear alternatives

The exact negative control consists of an intercept plus separate excerpt–physical-slot weights, conditional on the fixed context and segment. Rational checks give the constant a mean of one and each excerpt-slot indicator a mean of 1/9 in every arm, half, and full set. Every additive combination therefore yields $q = T = 0$. This elementary identity is neither a new theorem nor an assumption about the verifier. Synthetic counterexamples show its limits. Counting same-passage pairs gives $C = 3$, $D = E = 0$ and $q = 3$ after full averaging. Multiplying a same-passage indicator by order parity gives $C = +1$ cyclic, -1 reflected and $D = E = 0$, yielding $q = 1$ in each half but $q_{\text{full}} = 0$.

Equal raw means do not require equal nonlinear summaries. In an additive example, the mean logits agree but the mean sigmoid scores are 0.610758 for C and 0.731059 for D/E; changing the intercept reverses this gap. A two-segment example gives mean support of 1/2 in every arm, yet mean answer minima of 1/6 in C and 1/2 in D/E. Both constructions appear in Online Resource 2. They are analytic counterexamples, not neural results, and show why the earlier probability/minimum endpoint cannot by itself establish non-additivity in raw logits.

3.5 Use of artificial intelligence

Use of artificial intelligence. OpenAI AI assistance was used substantially for protocol and code preparation, execution coordination, numerical and literature checks, and drafting and editing this manuscript. Final verification and approval by the human authors remain pending.

4 Results

4.1 Cohort contrasts and negative observations

All forward passes in the two score-geometry experiments were completed. Repeated inputs gave exactly zero differences in logits and support scores (tolerance 10^{-6}), and the preserved-input, length, actual-slot, and arithmetic checks passed. These checks verify integrity; they do not provide external neural reproduction or factual validation. Table 2 reports the order sets and distance components separately for each cohort, and Fig. 1 shows every full-order case contrast. Online Resource 2 contains all eighteen RAGTruth and twenty-four HotpotQA case rows, together with the complete runtime accounting.

Table 2

Separate cohort and order-set contrasts in raw-logit units. $C\text{-dist.} = (|C - D| + |C - E|)/2$; $D - E = |D - E|$. Distances average segments within cases and then cases equally. +/- counts refer to cases.

Cohort	Orders	Cases/seg.	C-dist.	D - E	T	+/-
RAGTruth	cyclic	18/91	0.434297	0.094503	0.339794	17/1
RAGTruth	reflected	18/91	0.438036	0.092481	0.345555	17/1
RAGTruth	full	18/91	0.432943	0.032882	0.400061	18/0
HotpotQA-derived	cyclic	24/24	0.426959	0.195737	0.231221	19/5
HotpotQA-derived	reflected	24/24	0.419612	0.177249	0.242363	19/5
HotpotQA-derived	full	24/24	0.422557	0.073717	0.348841	23/1

For RAGTruth, $T_{\text{cyclic}} = 0.339794152809$ and $T_{\text{reflected}} = 0.345555269708$; each half had seventeen positive sources and one negative. After margins were pooled before taking absolute values, $T_{\text{full}} = 0.400061346294$, with eighteen positive sources, a median of 0.316171, and a range of 0.074135 to 1.021782. Among the 91 full-order segment contrasts, 85 were positive and six negative. Their within-source correlation means they cannot be treated as 91 independent cases.

Source 15544 had negative contrasts in both halves (-0.070096 ; -0.033896) but a positive contrast after pooling ($+0.167900$). Its D-E distance decreased from 0.305183 cyclic to 0.018998 full, while its C-to-distributed distance decreased from 0.235087 to 0.186898. Seven of its eleven segments were negative in each half, and one remained negative after pooling. Source 12434 had the largest full contrast, 1.021782. All sources and negative segments are retained.

For the HotpotQA-derived cases, $T_{\text{full}} = 0.348840651459$, with 23 positive cases and one negative, a median of 0.292293, and a range of -0.048964 to 0.832845. Each half contained nineteen positive and five negative cases, with $T_{\text{cyclic}} = 0.231221184338$ and $T_{\text{reflected}} = 0.242363385143$. Case 5ab3e45655429976abd1bcd4, at candidate rank 36, was negative in both halves and after full pooling (-0.048964 full). Its full means were $C = 0.390779$, $D = 0.313965$, and $E = 0.411893$, placing C between D and E. At rank 35, both half-contrasts were negative, whereas the full contrast was $+0.236714$. No case was excluded because of its sign.

4.2 Distance decomposition and direction

For RAGTruth, the C-to-distributed distance changed little from cyclic to full pooling (0.434297 to 0.432943), whereas the D-E distance decreased from 0.094503 to 0.032882. Thus the 0.060267 increase in T reflects the distributed means becoming closer, rather than greater concentrated-to-distributed separation. HotpotQA follows the same descriptive pattern: the C-to-distributed distance was 0.426959 cyclic and 0.422557 full, while the D-E distance decreased from 0.195737 to 0.073717. Since margins are pooled before distances are calculated, T_{full} is not the mean of the two half-contrasts.

For seventeen HotpotQA cases, C exceeded both distributed means; for six, it fell below both. In the negative case, C lay between them. A positive T therefore indicates relative separation without establishing a uniform increase in support. Figure 2 shows every full-order arm mean. The earlier RAGTruth probability/minimum endpoint also varied in direction: the mean absolute C/D gap was 0.069340 (median 0.045198; range 0.000912–0.242556), but the signed mean was + 0.005425, with nine positive and nine negative sources. That endpoint is separate from T and does not give a calibrated probability of factual correctness.

4.3 Feasibility stops and absent utility results

The closed RAGTruth stability branch examined 96 components and 558 quality-good answers. Every component had a matrix, but the 299 eligible answers came from only 63 components, short of the fixed requirement of 64. No final selection, model loading, or forward pass took place. This was an input-feasibility stop, not a nonpositive neural result; the requirement of 64 was a protocol rule, not a universal threshold for validity. A separate SciFact gate for whole-evidence duplication also stopped before inference, as documented in unchanged Online Resource 1 (Supplement S1).

A later original-label utility protocol completed its tokenizer/metadata gate on 1,856 original responses. It found 1,011 eligible responses in 209 components, and both original-label class screens passed. The complete eligible frame implied 285,342 planned forward passes, including repeats, but failed the frozen historical timing-support rule. The branch therefore closed as a resource/support NO-GO. No neural model was loaded, no forward pass was performed, and no calibration or utility model was fitted. The counts describe feasibility and planned computation; they neither measure accuracy nor show whether the diagnostic succeeds or fails. Online Resource 2 records the stop and its resource qualifications.

5 Discussion

Exact marginal excerpt-by-slot balance did not eliminate score separation in these inputs. A score formed solely from separate additive excerpt-slot contributions cannot produce that result, although additive components may still be present within the network and unrestricted interactions remain possible. Averaging every permutation removes dependence on a retained triple's initial ordering. It does not ensure invariance to different triples, context positions, or serializations. The contribution is the specified diagnostic contrast and its inspectable finite result; sensitivity to order in language models is already established.

The design does not identify the mechanism behind the separation. C combines nearby, potentially overlapping fragments with repeated passage labels, while D/E mix passages, labels, and topics. Even a same-passage pair counter produces a positive contrast without using semantics. For an ideal observer retaining the original block, copying adds no new observations [12]. For the verifier, it may alter salience, computational accessibility, or pragmatic interpretation. No label, overlap, or composition ablation has

been executed to separate these explanations. The proposed label ablation remains unexecuted, and no second verifier has been installed and tested for this comparison.

The HotpotQA result shows directional concordance under a different, prospectively specified adaptation. It does not establish generalization to ordinary RAG use: the corpus, oracle context, and answer form change together, and short reference answers impose a different verification demand. Interpretation is further limited by length eligibility, the adaptive RAGTruth history, conversion provenance, prior exposure, and the shared checkpoint. Screening passages or exact bodies establishes neither independent evidence lineages nor absence from training. The two descriptive T values cannot be used to rank the datasets.

We lack transformed-input factual labels and a completed original-label utility evaluation that could connect these margins to errors, harmful acceptance, or useful filtering. No practical threshold has been established for the raw-logit contrast. Invariance by itself would also be insufficient: a constant scorer is invariant but uninformative. The utility NO-GO establishes feasibility limits only. Practical benefit, broader generalization, and consequential novelty remain unresolved, and editorial changes cannot supply the missing measurements.

Private proof archives retain the protocols, source and model pins, code, selection/input hashes, logits, journals, and audits. Recalculation from saved logits checks the arithmetic without repeating inference. The companion package contains editable tables, figure data, bibliographic verification, and the revision record. It excludes raw corpus text, transformed prompts, token arrays, weights, and third-party binaries, so reconstructing inputs requires authorized matching assets. The HotpotQA conversion notices include CC BY-SA 4.0; private preservation does not determine public redistribution rights. We claim no public repository, data DOI, or external neural replication.

6 Conclusion

Balanced excerpt regrouping yielded a source-equal full-order contrast of 0.400061 for eighteen RAGTruth cases and 0.348841 for twenty-four separately analyzed HotpotQA-derived cases. Six RAGTruth segments and one HotpotQA case remained negative. Under this finite design, exact marginal balance rules out a purely additive excerpt-by-slot explanation of the observed separation. The findings describe the score geometry of one verifier. They leave causal mechanism, factual accuracy, generalization across models or ordinary RAG tasks, and practical usefulness unresolved. The utility feasibility stop provides no answer to those outcome questions.

Declarations

Funding

and competing interests. The corresponding author reports no funding and no competing interests.

Author contributions.

The corresponding author reports that both authors contributed to the work. Detailed contributor-role assignments and final approval of this manuscript version have not been recorded.

References

1. Niu C, Wu Y, Zhu J, Xu S, Shum K, Zhong R, Song J, Zhang T (2024) RAGTruth: A Hallucination Corpus for Developing Trustworthy Retrieval-Augmented Language Models. *Proceedings of ACL 2024*, vol. 1, pp. 10862–10878. <https://aclanthology.org/2024.acl-long.585/>
<https://doi.org/10.18653/v1/2024.acl-long.585>
2. Tang L, Laban P, Durrett G (2024) MiniCheck: Efficient Fact-Checking of LLMs on Grounding Documents. *Proceedings of EMNLP 2024*, pp. 8818–8847. <https://aclanthology.org/2024.emnlp-main.499/> <https://doi.org/10.18653/v1/2024.emnlp-main.499>
3. Ramprasad S, Wallace BC (2025) Do Automatic Factuality Metrics Measure Factuality? A Critical Evaluation. *Adv Neural Inf Process Syst* 38. <https://doi.org/10.52202/085713-4386>
4. Mujahid ZM, Wright D, Augenstein I (2026) Stress Testing Factual Consistency Metrics for Long-Document Summarization. *Proceedings of ACL 2026*, vol. 1, pp. 31914–31933. <https://aclanthology.org/2026.acl-long.1472/> <https://doi.org/10.18653/v1/2026.acl-long.1472>
5. Mamta CO (2025) FactEval: Evaluating the Robustness of Fact Verification Systems in the Era of Large Language Models. *Proceedings of NAACL 2025*, vol. 1, pp. 10647–10660. <https://aclanthology.org/2025.naacl-long.534/> <https://doi.org/10.18653/v1/2025.naacl-long.534>
6. Sok C, Luz D, Haddam Y (2025) MetaRAG: Metamorphic Testing for Hallucination Detection in RAG Systems. *CEUR Workshop Proceedings*, 4136, IAAI-ALA 2025. <https://ceur-ws.org/Vol-4136/iaai6.pdf>
7. Liu NF, Lin K, Hewitt J, Paranjape A, Bevilacqua M, Petroni F, Liang P (2024) Lost in the Middle: How Language Models Use Long Contexts. *Trans Association Comput Linguistics* 12:157–173. <https://aclanthology.org/2024.tacl-1.9/> https://doi.org/10.1162/tacl_a_00638
8. Murphy RL, Srinivasan B, Rao V, Ribeiro B (2019) Janossy Pooling: Learning Deep Permutation-Invariant Functions for Variable-Size Inputs. *International Conference on Learning Representations*. arXiv:1811.01900v3. <https://arxiv.org/abs/1811.01900v3>
<https://doi.org/10.48550/arXiv.1811.01900>
9. Qiu J, Han Z, Huang C (2026) SURE-RAG: Sufficiency and Uncertainty-Aware Evidence Verification for Selective Retrieval-Augmented Generation. arXiv preprint arXiv:2605.03534v2. <https://arxiv.org/abs/2605.03534v2> <https://doi.org/10.48550/arXiv.2605.03534>
10. Peng C, Wang B, Long Z, Sheng J (2025) arXiv preprint arXiv. <https://arxiv.org/abs/2512.25052v1>
<https://doi.org/10.48550/arXiv.2512.25052>. AdaGrES: Adaptive Greedy Context Selection via Redundancy-Aware Scoring for Token-Budgeted RAG:2512.25052v1

11. Zheng Z, Li B, Yao J, Long J (2026) Engineering Reliable Commit Gates for Agentic AI: Cost-Aware Verification Portfolios under Common-Mode Data Failures. arXiv preprint arXiv:2609.10969v1. <https://arxiv.org/abs/2609.10969v1>
12. Bara M (2026) Epistemic Sybil Resistance: Multiplying AI Agents Without Multiplying Evidence. arXiv preprint arXiv:2609.01873v1. <https://doi.org/10.48550/arXiv.2609.01873>. <https://arxiv.org/abs/2609.01873v1>
13. Lee D, Kim J (2026) A Failure-Mode Benchmark for Polymorphic Sybil Poisoning in RAG. arXiv preprint arXiv:2607.03739v1. <https://doi.org/10.48550/arXiv.2607.03739>. <https://arxiv.org/abs/2607.03739v1>
14. Yang Z, Qi P, Zhang S, Bengio Y, Cohen WW, Salakhutdinov R, Manning CD (2018) HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. Proceedings of EMNLP 2018, pp. 2369–2380. <https://aclanthology.org/D18-1259/> <https://doi.org/10.18653/v1/D18-1259>
15. Roh J, Bae K (2026) Diagnosing judgment flip risk of large language models in fact verification under conflicting evidence. ETRI Journal, Early View, pp. 1–14. First published 7 October 2026. <https://doi.org/10.4218/etrij.2026-0242>
16. Chen X, Chi RA, Wang X, Zhou D (2024) Premise Order Matters in Reasoning with Large Language Models. Proceedings of the 41st International Conference on Machine Learning, PMLR 235, pp. 6596–6620. <https://proceedings.mlr.press/v235/chen24i.html>
17. Bouke MA (2026) Sentence-Level Context Sensitivity as a Training-Free Detector of Unsupported Content, Evaluated Against Trained Verifiers. arXiv preprint. <https://arxiv.org/abs/2607.04223v2> <https://doi.org/10.48550/arXiv.2607.04223>. arXiv:2607.04223v2, revised 2 October 2026

Figures

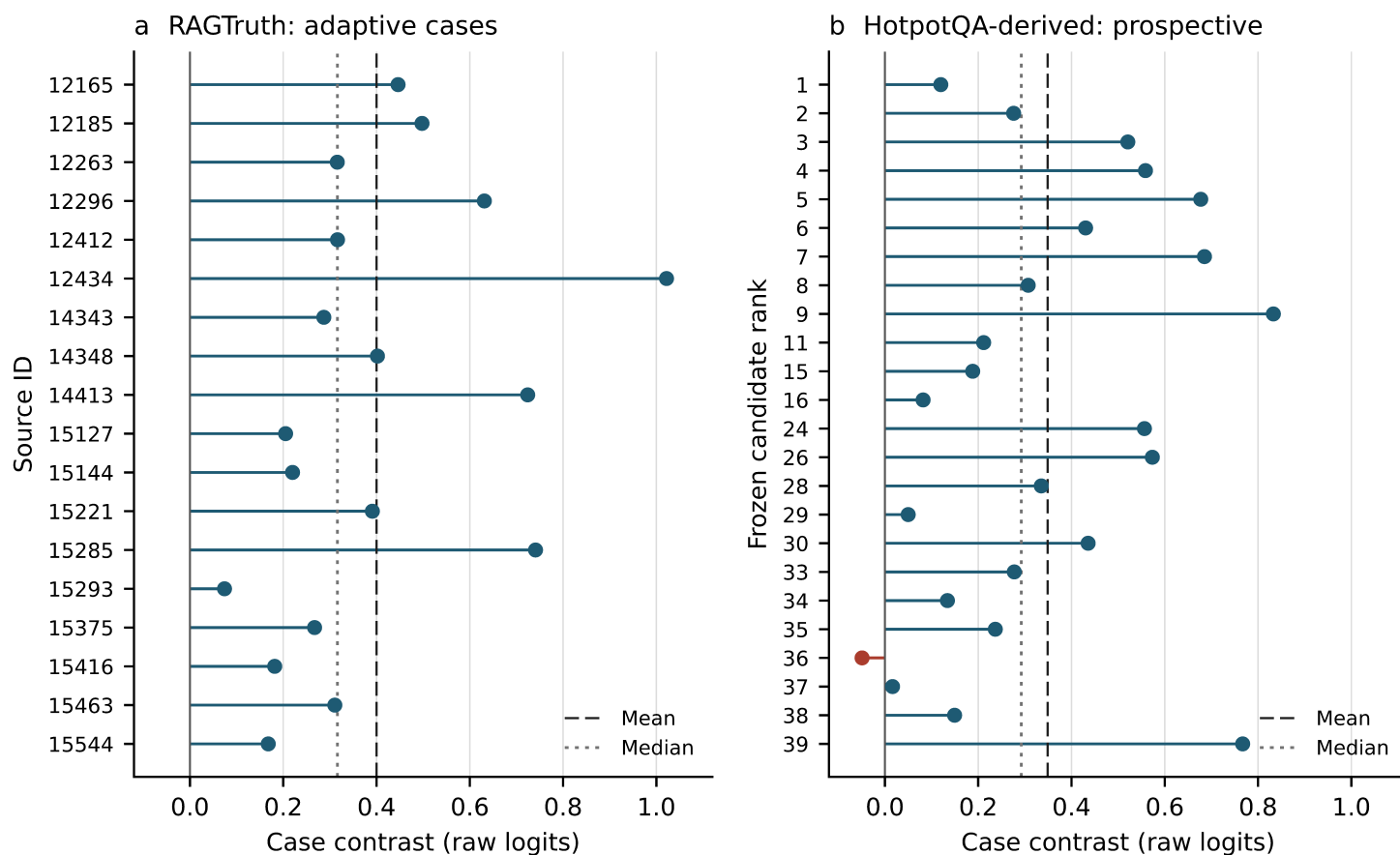


Figure 1

Every full-order case contrast in separate cohorts, ordered by RAGTruth identifier or frozen HotpotQA candidate rank. Dashed and dotted lines show each cohort’s own mean and median. Red marks the negative HotpotQA case. Scales are shared; no pooled estimate is calculated. Each RAGTruth value averages its segment contrasts; each HotpotQA case has one segment. All half-order values remain in Online Resource 2

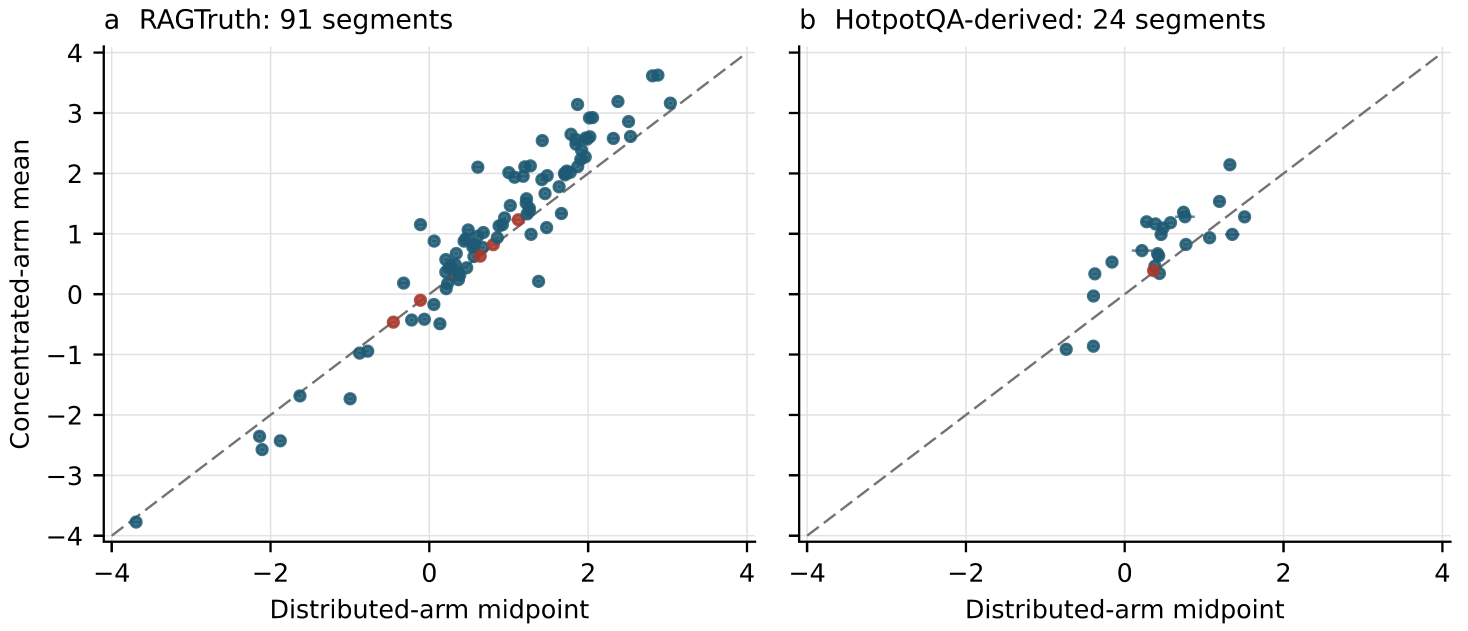


Figure 2

All full-order segment arm means, RAGTruth $n = 91$ and HotpotQA-derived $n = 24$. Points are at $((\mu_D + \mu_E)/2, \mu_C)$; horizontal endpoints are μ_D and μ_E , not uncertainty limits. Dashed lines are $y = x$. Red marks negative- q segments. RAGTruth segments are not independent source replicates. No regression or cross-cohort pooling is fitted