Token Savings Do Not Reliably Predict Support in RAG Context Selection

Anonymous ACL submission

Abstract

RAG context-selection papers often report input-token savings together with faithfulness or support metrics. These numbers are jointly interpretable only when they are measured on the same queries, retrieved pool, generator, prompt template, tokenizer, and evidence shown to the judge. We propose an evaluation contract that fixes these fields before comparing selection policies. Under this contract, a Natural Questions study with 325 paired queries shows that lower input-token use does not reliably predict lower MiniCheck unsupported rates, even under a shared token budget. A matched TriviaQA-open study shows a different token–support relationship, indicating that the tradeoff is empirical rather than universal. Human annotation and failure analysis calibrate the automatic metric and localize where token cuts can remove or degrade answer-critical evidence. The result is a reusable reporting rule for RAG efficiency claims: token savings should be reported together with unsupported rates measured under the same evaluation contract.

1 Introduction

RAG context selection is commonly evaluated by reporting how many input tokens a policy saves and whether generated answers remain supported by the selected evidence. These two quantities are often read together: a system that uses fewer tokens is expected to preserve, or even improve, support. This interpretation is only valid when both quantities are measured on the same comparison object.

In practice, RAG efficiency comparisons can change several fields at once: the query set, candidate retrieval pool, generator, decoding setup, packing template, tokenizer, or the evidence shown to the judge. When any of these fields differ across rows, a token reduction and a support change may no longer describe the same experimental object.

We address this measurement problem with an evaluation contract for RAG context selection. The contract fixes the paired query identifiers, retrieval pool, generator and decoding, evidence packing, token accounting, and the evidence surface given to the judge before policies are compared. Policies may still select and pack different evidence, but token cost and unsupported rate are measured under the same contract. Under a frozen Natural Questions contract, policies that use fewer input tokens do not necessarily produce lower MiniCheck unsupported rates, including under a shared packing budget. Under a second TriviaQA-open contract, the token–support relationship changes. These results motivate reporting token savings together with unsupported rates measured under the same evaluation contract.

Contributions. We make four contributions. **First**, we define a RAG-specific reporting unit for efficiency evaluation. The unit binds token cost, selected evidence, judge evidence, and paired inference into one auditable contract. **Second**, we give a controlled counterexample to token-only reporting. On Natural Questions paired-325, token-cost ranks and MiniCheck unsupported-rate ranks disagree. The mismatch remains under a shared budget cap, where the lowest-token policy is not the lowest-unsupported policy. **Third**, we show that the token–support relationship changes across evaluation contracts. On TriviaQA-open paired-325, the ranks are more aligned than on Natural Questions. This contrast shows that token–support tradeoffs must be measured, not assumed. **Fourth**, we calibrate and localize the automatic metric. Human sentence annotation, a second-judge sensitivity check, and a retrieval/selection/surfacing failure analysis show when the MiniCheck metric supports large contrasts and where token cuts can remove answer-critical evidence.

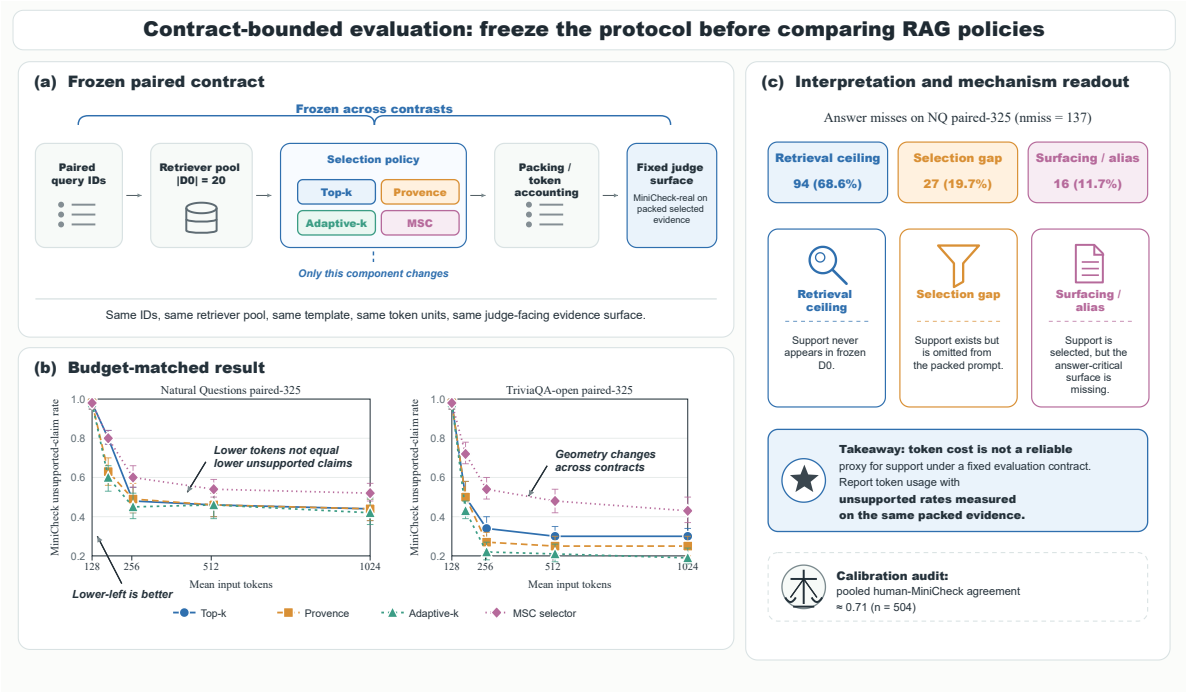

Figure 1: **Contract-bounded evaluation.** Policy rows are compared only after paired IDs, retrieval pool, packing/token accounting, generator/decoding, and judge-facing evidence are fixed. The same surface supports the budget-grid readout and the miss-taxonomy localization.

2 Paired Evaluation Contract

Throughout, scalars refer to **packed selected evidence** under one tokenizer and judge surface (§2.1 separates answer, judge, and repaired channels).

2.1 Comparison object, channels, and scope

Comparison object. We study open-domain retrieval-augmented question answering (Lewis et al., 2020; Kwiatkowski et al., 2019). For each query q, the retriever returns a fixed candidate pool $D_0(q)$. A context-selection policy chooses evidence $S(q) \subseteq D_0(q)$, which is packed into the generator prompt. We measure input cost as $\mathrm{Tok}(S)$ under one tokenizer and prompt template.

Support metric. Our main support metric is the sentence-level unsupported rate assigned by MiniCheck-real when it is given the same packed evidence $S(q)$ that was available to the generator. We use this automatic metric as a contract-level support diagnostic and calibrate it with human annotation in §3.2.

Channels. We keep three quantities separate: answer quality, measured by strict EM/F1 on generated answers; unsupported rate, measured by MiniCheck-real on $S(q)$; and a normalized-answer sanity check, used only to interpret absolute answer scores. Token ranks and unsupported-rate ranks are always computed across policies under one frozen evaluation contract.

Research question and scope. We ask whether reductions in $\mathrm{Tok}(S)$ co-move with reductions in the unsupported rate under one frozen contract. **MSC (Minimal Sufficient Context)** names our learned context-selection baseline; we use it only as informal motivation and do not introduce a new formal minimality objective (scope limits are in §7). **MSC-Selector v0** scores documents in the frozen pool $D_0(q)$ with a rank head and a **stop head** that outputs sufficiency logits over ranked prefixes; at inference, whole documents are appended in rank order until the stop score crosses γ_{stop} (with γ_{no} for early rejection) or a token budget binds. The selector is trained on oracle-derived labels over minimal sufficient document sets; the checkpoint is frozen before headline evaluation.

2.2 Evaluation contract

An evaluation contract fixes the fields that must be shared before token cost and support can be compared across context-selection policies. Table 1 lists the contract used in this paper. The policies differ only in which documents or sentences they select and how those selections are packed; the query set, candidate pool, generator, decoding, to-

ken accounting, judge evidence, and inference plan remain fixed.

Natural Questions contract. The main contract uses 325 Natural Questions open-domain development queries. This paired list is nested under the official development split and the preregistered development subsample, and is used only for within-contract comparisons. Calibration pools, TriviaQA-open, mechanism slices, and HotpotQA stress tests keep separate denominators. Appendix A reports the denominator chain and answer-channel sanity checks.

Retriever, generator, and judge. All rows use the same hybrid sparse–dense retriever with $|D_0(q)|$=20, the same Qwen2.5-14B-Instruct generator with deterministic decoding (temperature 0, top-p 1, max new tokens 256 on the frozen paired-325 answer exports; MSC stopping constants γ_{stop}=0.50, γ_{no}=0.95), the same prompt template and tokenizer, and MiniCheck-real on the packed selected evidence. This makes token cost and unsupported rate measurements directly comparable across policy rows. Enumerated decoding fields, answer versus judge channel separation, normalization checks, and paired-bootstrap mirrors are in Appendix E.2–E.3.1 and Tables 6–7.

2.2.1 Preregistered input-token budget grid

Figure 2 uses a preregistered cap grid $B \in \{128, 256, 512, 1024\}$ in full-prompt input-token units. Evidence is packed greedily as whole document blocks, with no intra-document cut at the boundary. Tables 2–3 separate the headline single-configuration row from the shared-cap B=512 slice on the same paired IDs. Adaptive-k chooses a feasible score threshold from a fixed grid subject to B; MSC uses the registered budget-grid ordering under the same cap. These rows therefore test a fixed-cap alternative without conflating it with the headline stop-head MSC configuration. Holm-adjusted grid contrasts are reported in Appendix Table 22.

2.3 Compared selection policies

We compare four policy families that produce different packed evidence surfaces under the same retriever and generator interface.

Top-k. A fixed-k policy packs retrieved documents in rank order.

Provence-style compression. A compression policy aggressively prunes context using a frozen open-source implementation while preserving the same generator interface.

Adaptive-k. A deterministic score-prefix policy selects $k(q)$ from a preregistered threshold grid. The headline setting places its aggregate mean token count near the Top-k band, while the budget-grid experiments test fixed-cap alternatives.

MSC selector. MSC-Selector v0 is the learned policy described above: rank head plus stop head over $D_0(q)$, with headline rows using $(\gamma_{\text{stop}}, \gamma_{\text{no}})$ and the preregistered budget-grid rows (§2.2.1) using fixed cap B and rank ordering only (no γ thresholds) to avoid double confounding.

These policies are intentionally heterogeneous. The comparison asks whether the token count of the packed evidence surface predicts its unsupported rate when all other evaluation fields are fixed. Shared-budget rows in Table 3 and Figure 2 separately test the fixed-cap alternative.

3 Main Results

We test whether lower input-token cost predicts lower unsupported rate when the evaluation contract is fixed. The answer depends on the contract. Natural Questions gives a controlled counterexample to token-only reporting, while TriviaQA-open shows that the relationship can be more aligned under a different answer/evidence distribution. Human calibration and failure analysis then explain how far the automatic metric can be read and where token cuts affect evidence support.

3.1 Natural Questions: Token and Support Ranks Decouple

Tables 2–3 compare Top-k, Provence, Adaptive-k, and MSC on the same 325 Natural Questions queries, retrieved pools, generator, prompt template, tokenizer, and MiniCheck-real judge.

Table 2 shows the headline mismatch. Provence uses the fewest mean input tokens but has the highest unsupported rate. Top-k has the lowest unsupported rate without being the lowest-token policy. Thus, token-cost rank and unsupported-rate rank disagree under the same evaluation contract.

The shared-budget result in Table 3 is the stronger test. At B=512, Adaptive-k uses fewer tokens than the other policies, but Top-k and Provence have lower unsupported rates. This

Contract axis	Freeze before comparison	Instantiation in this paper	Confound blocked
Queries	Paired query IDs and denominators.	One shared NQ-open development list, N_{pair}=325; all other pools are reported with separate denominators.	Coverage / difficulty drift
Candidate evidence	Retriever identity and candidate-pool depth.	Same hybrid sparse–dense retriever and $\lvert D_0(q)\rvert$=20 for every query.	Retrieval-depth drift
Answer channel	Generator, prompt template, and decoding.	One instruct model with fixed prompts/templating and deterministic decoding.	Answer-surface drift
Packing and cost	Tokenizer scope, prompt scope, and packing grammar.	Same tokenizer and prompt template; policies differ only in which evidence is selected and how it is packed.	Token-count incompatibility
Judge channel	Judge-facing evidence surface and aggregation.	MiniCheck-*real* on packed selected evidence with shared sentence-level aggregation.	Faithfulness-surface drift
Inference	Contrast family, bootstrap budget, and random seeds.	Pre-specified paired-bootstrap caps, Holm families, and fixed seeds for headline contrasts.	Post-hoc contrast selection

Table 1: **Evaluation-contract checklist.** Each row names one field that is fixed before comparing RAG context-selection policies. Under this contract, token cost and unsupported-rate ranks are computed on the same query IDs, candidate pool, answer channel, packing and token accounting, judge evidence, and inferential plan. Denominator counts and paired-bootstrap mirrors are in Table 6 and Appendix E.3.1.

Policy	Mean Tok. ↓	Unsup. ↓	r_T	r_U	F1 ↑	F1_{rep} ↑
Top-k	355.1	0.503	3	1	0.044	0.229
Provence	123.8	0.995	1	4	0.081	0.333
Adaptive-k	335.0	0.881	2	3	0.043	0.221
MSC selector	876.1	0.712	4	2	0.143	0.424

Table 2: **NQ paired-325 headline readout.** Token rank and unsupported-rate rank disagree under the same evaluation contract. Provence is lowest-token but highest-unsupported, while Top-k is lowest-unsupported without being lowest-token. F1 columns are answer-channel sanity checks only.

All rows share paired IDs, $\lvert D_0\rvert$=20, generator/decoding, tokenizer scope, packing rules, and MiniCheck-real on packed evidence. F1 columns are answer-channel sanity metrics only.

Policy	Mean Tok. ↓	Unsup. ↓	r_T	r_U	F1 ↑	F1_{rep} ↑
Top-k	415.4	0.521	2	2	0.147	0.336
Provence	415.4	0.519	2	1	0.146	0.342
Adaptive-k	301.0	0.615	1	3	0.134	0.324
MSC selector	415.9	0.618	4	4	0.121	0.290

Table 3: **NQ paired-325 at shared cap** B=512**.** A common budget cap does not restore monotonicity: Adaptive-k is lowest-token but not lowest-unsupported. This shared-budget result shows that the mismatch is not only an artifact of aggressive compression.

Same paired IDs and judge channel as Table 2; only the enforced packing cap differs. Mean scored sentences: 3.05/3.03/2.96/3.15 for Top-k/Provence/Adaptive-k/MSC.

rules out the explanation that the mismatch is only caused by an extreme compression setting. Even when policies are compared under a common token cap, lower token use does not determine lower unsupported rate.

Answer-quality sanity check. Strict NQ F1 is low because the generator often produces long-form answers while NQ EM/F1 rewards normalized short spans. We therefore use strict EM/F1 only as an answer-channel sanity check, not as the support metric. A short-answer control at B=512 raises strict F1 to 0.267–0.299 while preserving the same retrieval, packing, token accounting, and judge-evidence construction (Table 8; Appendix A.4). The token–unsupported mismatch is therefore not explained by empty or malformed outputs.

3.2 Human Calibration of the Automatic Support Metric

MiniCheck-real is the fixed support metric in the evaluation contract. We calibrate this metric with an adjudicated human sentence audit on the same packed evidence surfaces used in §3.1.

The audit contains 504 sentence units with dual annotation and adjudication (Table 4; Figure 3). MiniCheck agrees with adjudicated human labels on 0.708 of pooled examples, with Wilson intervals reported in the table. This agreement is sufficient for large contract-level contrasts, especially the high unsupported rate of Provence, but it does not justify fine-grained human-gold rankings among all four policies. Appendix B indexes auxiliary pools and their unit definitions (Table 9).

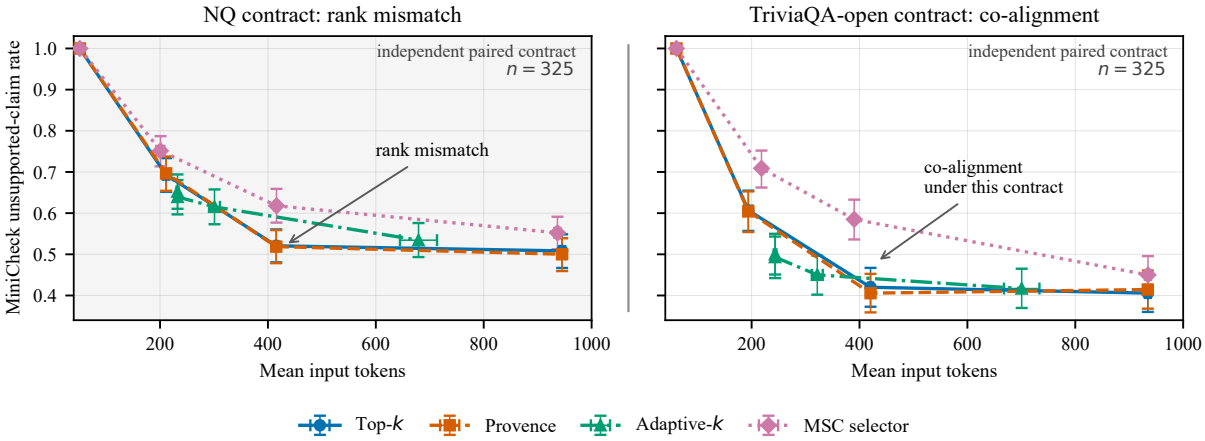

Figure 2: **Budget-matched token–unsupported-rate curves under two independently fixed evaluation contracts.** Natural Questions paired-325 shows a rank mismatch between mean input tokens and MiniCheck unsupported rate; TriviaQA-open paired-325 shows stronger alignment under its own contract. Ranks are interpreted only within each contract. Curves use preregistered caps $B \in \{128, 256, 512, 1024\}$ with 95% paired-bootstrap bands.

Policy	n	**Human–MC agr.**	**Human–NLI agr.**	$\kappa_{H,\mathrm{NLI}}$	**Human unsup.**
Top-k	126	0.643 [0.556, 0.721]	0.421 [0.338, 0.508]	−0.036	0.278 [0.207, 0.362]
Provence	126	0.714 [0.630, 0.786]	0.746 [0.664, 0.814]	0.071	0.730 [0.647, 0.800]
MSC selector	126	0.762 [0.680, 0.828]	0.563 [0.476, 0.647]	0.145	0.341 [0.264, 0.428]
Adaptive-k	126	0.714 [0.630, 0.786]	0.484 [0.399, 0.571]	0.010	0.254 [0.186, 0.336]
Pooled	**504**	**0.708 [0.667, 0.746]**	**0.554 [0.510, 0.596]**	**0.061**	**0.401 [0.359, 0.444]**

Table 4: **Human calibration of MiniCheck unsupported rates.** Agreement is measured against adjudicated human sentence labels on the same packed evidence surfaces used by the main contract. The audit supports large contract-level contrasts and calibrates the automatic metric; fine-grained policy ordering should be read with the sensitivity checks. Auxiliary pools: Appendix Table 9.

Brackets are Wilson 95% intervals. H–MC/H–NLI are binary agreement on the adjudicated sentence-unit pool. This table bounds the automatic judge readout; it does not make MiniCheck a human-gold hallucination metric.

The human rates give the same qualitative extreme as the headline MiniCheck result: Provence has the highest unsupported rate in the adjudicated pool (0.730), with Adaptive-k lowest (0.254), then Top-k (0.278), and MSC (0.341). The remaining policy differences are closer, so we treat them as automatic-metric contrasts and report second-judge and preprocessing checks as sensitivity analyses. Finer separations are summarized in §7.

Generator sensitivity. Appendix Table 17 swaps only the answer generator (Mistral-7B-Instruct-v0.2) while holding packed contexts and paired-325 identifiers fixed under the primary contract. Mean input tokens are unchanged and MiniCheck unsupported-claim *orderings* echo the headline readout, while absolute rates rescale after the generator swap. This same-context swap tests whether the headline ordering is an artifact of one answer generator. The ordering is preserved while absolute rates rescale, so we use it as a generator-sensitivity bound for the primary contract.

Preregistered B=512 calibration slice. On regenerated B=512 packed evidence (n=60 per policy on NQ; dual annotation with adjudication), Natural Questions human unsupported rates are nearly flat across policies ($\approx$ 0.20–0.28). This readout is a fixed-packing sensitivity bound rather than a second policy-ranking test (unit definitions and pool index: Appendix Table 9).

Disagreement patterns. A targeted disagreement audit (n=208) concentrates in borderline entailment or conflict, truncated prompts, and main-claim versus secondary-detail splits—patterns consistent with contract-bounded reading rather than with replacing human judgment.

Second judge (sensitivity only). An auxiliary entailment-style judge supplies threshold-conditioned triangulation (Figure 5); it is not alter-

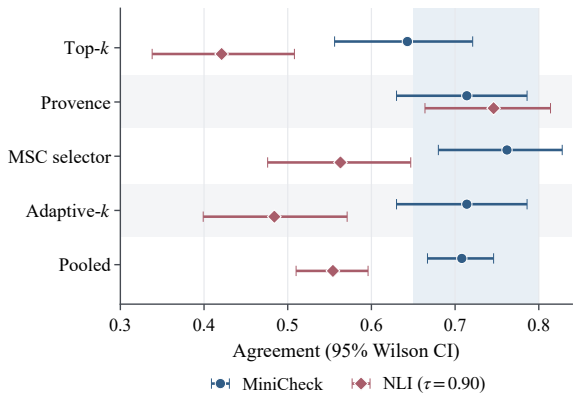

Figure 3: **Human calibration agreement on the adjudicated sentence pool** (n=504)**.** MiniCheck agrees with human labels more often than the auxiliary NLI threshold readout on this pool, supporting large same-surface contrasts but not human-gold policy leaderboards.

native human gold and can flip aggregate pairwise orderings at extreme τ.

MiniCheck preprocessing controls. Three deterministic preprocessing controls preserve the headline four-policy *ordering* while shifting absolute rates (Appendix E.3.2); they bound sensitivity on registered probes rather than claiming judge invariance broadly.

3.3 TriviaQA-open: The Token–Support Relationship Changes

We instantiate a second evaluation contract on 325 TriviaQA-open queries using the same protocol discipline: paired identifiers, fixed candidate-pool depth, fixed generator and decoding, shared token accounting, and MiniCheck-real on the packed evidence.

TriviaQA-open does not reproduce the Natural Questions rank mismatch. Under this contract, token and unsupported-rate ranks are more aligned. This result strengthens the reporting argument rather than weakening it: the token–support relationship is not a universal property of a selector or a dataset. It must be measured under the contract that produced the reported token numbers.

One plausible explanation is that TriviaQA-open answers are often recoverable from concentrated entity strings, while Natural Questions contains more aliasing and surface-form variation. The cross-contract taxonomy in Appendix Table 11 is consistent with this explanation, but we use it as a diagnostic hypothesis rather than a causal claim.

Table 5 reports headline means and within-contract dense ranks (r_T, r_U). On TriviaQA-open, Top-k and Provence share per-example identical frozen packed exports. We therefore treat Provence on TriviaQA-open as a frozen-export control, not an independent compression contrast.

Appendices report denominator checks, auxiliary calibration pools, and boundary stresses.

4 Why Token Count Fails as a Support Proxy

The Natural Questions result raises a mechanism question: where can token savings break the link between cost and support? We separate the pipeline into three locations. First, retrieval ceilings occur when the frozen candidate pool contains no answer-supporting evidence. Second, selection gaps occur when supporting evidence exists in the pool but is not packed. Third, surfacing failures occur when selected evidence does not expose the answer-critical surface form needed by the generator or judge (Figure 4).

On the NQ paired-325 answer-miss slice, 94 of 137 misses are retrieval ceilings. These cannot be fixed by a selector that only repacks $D_0(q)$. The remaining 43 misses are optimizable within the frozen pool: 27 selection gaps and 16 surfacing failures. These are the cases where a policy can save tokens while dropping or degrading answer-critical evidence.

This decomposition explains the aggregate mismatch. Token cuts do not repair ceiling-limited queries, and they can harm support when they remove evidence needed for answer-critical claims. Therefore, mean token count is blind to whether the packed evidence surface preserves the information that matters for support. Answer misses, oracle reachability, and MiniCheck labels remain distinct surfaces; perturbation summaries appear in Appendix D.1.

4.1 Oracle Reachability and Achieved Coverage

Under fixed $|D_0|$=20 on paired-325, oracle-any-supported is $\approx$ 0.711: roughly 29% of queries have no oracle-supporting document in D_0 before any selector runs, so that mass is not explained by ranking or compression alone. Table 15 reports the preregistered $|D_0|$ sweep (20 $\rightarrow$ 50): feasible rate rises ($\approx$ 0.711 $\rightarrow$ 0.785) while prefix@5 oracle coverage falls and selection-bottleneck mass rises under the same frozen selection contract, so pool expansion is not a monotone fix (Appendix A.2).

	NQ paired-325						TriviaQA-open paired-325					
Policy	Tok.	Uns.	r_T	r_U	F1	$\mathrm{F1}^{512}_{\mathrm{rep}}$	Tok.	Uns.	r_T	r_U	F1	$\mathrm{F1}^{512}_{\mathrm{rep}}$
Top-k	355.1	0.503	3	1	0.044	0.229	1 004.6	0.967	3	3	0.119	0.647
Provence	123.8	0.995	1	4	0.081	0.333	1 004.6	0.967	3	3	0.119	0.650
Adaptive-k	335.0	0.881	2	3	0.043	0.221	356.8	0.747	1	1	0.159	0.628
MSC selector	876.1	0.712	4	2	0.143	0.424	393.5	0.847	2	2	0.276	0.637

Table 5: **Cross-contract summary.** Token–unsupported maps differ across independently frozen NQ and TriviaQA-open contracts. Ranks are dense ranks within each contract only (1=best).

Tok./Uns. are headline means; F1 columns are answer-channel sanity checks.
For TriviaQA-open, Top-k and Provence rows share identical headline aggregates because Provence pruning yields zero token reduction under this contract and the policy exports reuse the same frozen packed evidence. We therefore treat Provence on TriviaQA-open as a frozen-export control, not an independent compression contrast.

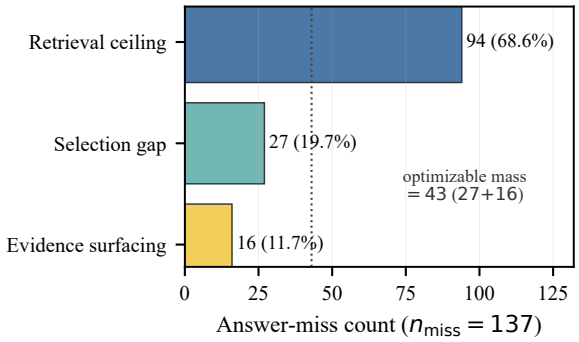

Figure 4: **Miss taxonomy on NQ paired-325 answer misses.** Most misses are retrieval ceilings; the optimizable mass splits into selection gaps and evidence-surfacing failures, where token reduction can remove or degrade answer-critical evidence.

These quantities bound reachability under frozen retrieval only.

4.2 Alignment Bottleneck and Miss Taxonomy

Among oracle-*reachable* queries, selected and packed context can still omit or degrade answer-supporting evidence—*alignment* pressure distinct from the pool ceiling in §4.1. Appendix Table 12 restates Figure 4 counts with optimizable flags and exemplar query IDs; optimizable mass is $43/137$ ($\approx 13.2\%$ of paired-325). Appendix C.2 maps that mass to selection/truncation versus packing/alias mismatch on frozen exemplars (n=63).

4.3 Packed Evidence, Not Token Count, Drives Support Risk

The budget-grid and miss-taxonomy results point to the same conclusion. Unsupported rate depends on which evidence survives selection and packing, not only on how many tokens remain. Evidence-edit perturbations shift MiniCheck unsupported rates under fixed token budgets (Appendix Figure 9), and per-query token reductions do not yield a stable negative correlation with unsupported-rate changes (Appendix Figure 6; Table 14). Aggregate rank mismatch can therefore persist even when many individual queries save tokens.

5 Related Work and Discussion

Our work complements RAG retrieval, compression, and evaluation methods by focusing on the reporting unit used to compare token efficiency with support (Lewis et al., 2020; Karpukhin et al., 2020; Chirkova et al., 2025; Jiang et al., 2023; Asai et al., 2024; Izacard and Grave, 2021; Tang et al., 2024; Niu et al., 2024; Manakul et al., 2023; Es et al., 2024; Saad-Falcon et al., 2024; Petroni et al., 2021; Thakur et al., 2021; Joren et al., 2025). Existing RAG methods change how evidence is retrieved, selected, compressed, or consumed, while evaluation frameworks and judges provide complementary ways to score generated answers. Our contribution is to specify when token cost and support can be read as measurements of the same experimental object.

The evaluation contract is a RAG-specific reporting unit. It binds the selected evidence, tokenizer-scoped input cost, generator-facing prompt, judge-facing evidence, and paired inference plan into one auditable comparison. This distinction matters because context compression changes both the number of tokens and the evidence surface seen by the generator and the judge. Table 25 maps these reporting axes across related artifacts; it is a reporting-unit comparison, not a quality ranking.

The practical reporting rule is simple: a RAG context-selection paper that claims token efficiency should disclose the paired query identifiers, retrieval pool, packing template, tokenizer, answer channel, judge checkpoint, evidence surface, uncertainty calculation, and paired test used to compute the support metric. A mismatch between token savings and unsupported rate is not evidence

against compression; it identifies cases where selection or packing changes answer-critical evidence. Conversely, alignment on TriviaQA-open does not make token-only reporting safe elsewhere. The token–support relationship must be reported under the contract that produced it. The Paired-Contract Reporting Kit in Appendix E.5 turns this rule into a minimal audit chain from contract manifest to result manifest, calibration manifest, and figure/table anchor.

6 Conclusion

We proposed an evaluation contract for RAG context-selection studies that report token savings together with support metrics. The contract fixes the query set, retrieval pool, generator and decoding, packing and token accounting, judge evidence, and paired inference before policies are compared.

Under a Natural Questions contract, lower token use does not reliably predict lower MiniCheck unsupported rate, including under a shared budget cap. Under a TriviaQA-open contract, the relationship is more aligned, showing that token–support tradeoffs are empirical and contract-specific. Human calibration and failure analysis explain how to read the automatic metric and where token cuts can remove answer-critical evidence.

The reusable outcome is a reporting rule: token-saving claims should be accompanied by unsupported rates measured on the same packed evidence and enough contract metadata for reviewers to reconstruct what was compared.

7 Limitations

The numerical results are limited to paired-325 slices on Natural Questions and TriviaQA-open, one frozen hybrid retriever, and one primary generator stack. Other corpora, languages, longer generations, or generator families require newly specified contracts before the numerical relationships can be reused.

MiniCheck-real is an automatic support metric calibrated by adjudicated human labels (n=504), second-judge probes, and preprocessing checks. The calibration supports large contract-level contrasts, while small policy separations should be read with additional human or judge sensitivity analysis.

The compared policies differ in inductive bias, and Adaptive-k is mean-token aligned rather than byte-matched per query. The supported conclusion is therefore about token-only reporting under a fixed evaluation contract, not about a universally best context-selection method.

Finally, the HotpotQA and perturbation analyses are stress diagnostics. They localize failure modes and motivate future contracts for multi-hop, adversarial, and multilingual RAG settings.

8 Ethical Considerations

The study uses public QA corpora and open model components. Because public QA systems can produce fluent but unsupported answers, we report unsupported rates together with the exact evaluation contract used to compute them. Human sentence annotations were drawn from public QA splits, used dual independent labeling with adjudication, and did not collect personally identifying information beyond what is already present in the public corpora. Natural Questions (Kwiatkowski et al., 2019), TriviaQA (Joshi et al., 2017), and HotpotQA (Yang et al., 2018) are used under their public release terms; model, judge, dataset, and license information are recorded in the anonymized supplement (Appendix E.1–E). LLM tools assisted drafting and polish under author verification; all reported claims, statistics, and citations remain author-responsible.

References

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

Nadezhda Chirkova, Thibault Formal, Vassilina Nikoulina, and Stéphane Clinchant. 2025. Provence: Efficient and robust context pruning for retrieval-augmented generation. In *The Thirteenth International Conference on Learning Representations*. Poster.

Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. 2024. 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, St. Julians, Malta. Association for Computational Linguistics.

Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*,

pages 874–880, Online. Association for Computational Linguistics.

Zhengbao Jiang, Frank Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Active retrieval augmented generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7969–7992, Singapore. Association for Computational Linguistics.

Hailey Joren, Jianyi Zhang, Chun-Sung Ferng, Da-Cheng Juan, Ankur Taly, and Cyrus Rashtchian. 2025. Sufficient Context: A New Lens on Retrieval Augmented Generation Systems. In *The Thirteenth International Conference on Learning Representations*. Poster.

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

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. 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, Online. Association for Computational Linguistics.

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. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

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. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33.

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9004–9017, Singapore. Association for Computational Linguistics.

Cheng Niu, Yuanhao Wu, Juno Zhu, Siliang Xu, Kashun Shum, Randy Zhong, Juntong Song, and Tong Zhang. 2024. RAGTruth: A hallucination corpus for developing trustworthy retrieval-augmented language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10862–10878, Bangkok, Thailand. Association for Computational Linguistics.

Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, Vassilis Plachouras, Tim Rocktäschel, and Sebastian Riedel. 2021. KILT: a benchmark for knowledge intensive language tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2523–2544, Online. Association for Computational Linguistics.

Jon Saad-Falcon, Omar Khattab, Christopher Potts, and Matei Zaharia. 2024. 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, Mexico City, Mexico. Association for Computational Linguistics.

Liyan Tang, Philippe Laban, and Greg Durrett. 2024. MiniCheck: Efficient fact-checking of LLMs on grounding documents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8818–8847, Miami, Florida, USA. Association for Computational Linguistics.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*.

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

A Contract and Data Universe

A.1 Appendix Roadmap

Appendix A audits the paired data universe and answer-channel sanity checks used to interpret Tables 2–3. Appendix B contains judge calibration details beyond the main adjudicated audit. Appendix C supports the mechanism analysis with miss-taxonomy and reachability checks. Appendix D reports boundary stress tests that are not co-primary benchmarks. Appendix E provides reproducibility manifests and the expanded Paired-Contract Reporting Kit. Headline claims rely on

Tables 2–5 and Figures 1–4; appendix artifacts either audit, calibrate, or stress-test those claims. Interpretation boundaries for headline claims are in §7; the tabular claim-boundary map remains in the anonymized supplement only.

Scope. Appendix A documents denominator and answer-channel checks for Tables 2–3. Appendix A reports answer-channel sanity checks only. These checks explain absolute EM/F1 levels; they do not alter the token accounting, packed judge surface, or headline unsupportedness comparisons. Paired-bootstrap endpoint mirrors and preregistered budget-grid Holm summaries appear in Appendix E.3.1. Figure 1 in §2.2 is the contract anchor ($|D_0|$=20, policy-selected evidence packing, MiniCheck-real on selected context). Provenance is listed in the anonymized supplementary package (Appendix E.4); judge-channel calibration companions are in Appendix B.

Why headline EM looks small. NaturalQuestions exact match uses strict normalization; instruct-style answers are often long and non-extractive, so EM stays low despite non-empty outputs (Table 7; Figure 7; short-answer channel: Table 8). Evaluating EM on the full $N_{\text{raw}} = 3{,}610$ development split while predictions exist only for the development subsample or the paired headline subset yields near-zero scores because missing IDs count as empty—that is **not** the headline $N_{\text{pair}} = 325$ contract.

A.2 Paired data universe and denominator discipline

What this subsection verifies. Tables 2–3 anchor headline rows under one frozen contract (fixed retriever pool, generator, decoding, MiniCheck-real, and the same 325 query IDs); paired-bootstrap endpoint mirrors for headline contrasts are in Appendix E.3.1 (Table 21). This subsection checks denominator and channel consistency only; the scientific target remains the paired relation between mean input tokens and unsupported-claim rates, with absolute EM as a calibration context variable (Table 7).

Answer-channel and denominator sanity. Table 7 reports the only numeric audit needed for the appendix: changing from the preregistered development subsample ($N_{\text{sub}} = 361$) to the frozen paired slice ($N_{\text{pair}} = 325$) changes absolute EM/F1 levels but not the paired-contract read.

Stage	n	Definition / use
NQ-open development (official)	3610	N_{raw}: all dev queries in the released split manifest paired with raw JSONL.
Preregistered development subsample	361	N_{sub}: query-level repartition of the official development identifiers (80/10/10 train/dev/test; seed 20260310); *not* the full N_{raw} universe.
Paired headline contract	325	N_{pair}: frozen query IDs shared by headline policy rows (Table 2); joint export coverage under the preregistered pipeline.

Table 6: **Natural Questions denominator flow** for headline paired contrasts. N_{pair} is used only for within-contract policy comparisons; it is neither a uniform random sample of N_{raw} nor equal to N_{sub} (36 subsample queries are outside the paired headline list).

The low strict F1 is mainly an answer-format issue rather than output collapse: the Top-k export has no empty answers, 9 keyword-heuristic refusal-like outputs, and verbose answers under the default answer-instruction channel (max **256** new tokens on the frozen export) with generated-string length mean/p50/p90/max 434.6/293/1026/1465 characters and generator-reported output-token count mean/p50/p90/max 93.8/63/256/256 (character counts are over `strip(prediction)` only, excluding prompt/context). Repaired F1 is reported only as an answer-normalization audit; token accounting and MiniCheck inputs are unchanged. Answer strings and judge-facing evidence are separate surfaces; for example, the Top-k MiniCheck input export averages 254.1 characters. EM/F1 recomputes the same frozen Top-k, Provence, and MSC prediction exports used for Appendix E.3.1 significance tests on the paired slice. *Refusal-like (heuristic)* counts keyword matches in model text (e.g., "cannot answer", "none of the documents") and is **not** a gold abstention label.

A.3 Short-answer controlled channel

The short-answer channel matches the headline B=512 judge-facing packed-evidence construction while changing only the answer instruction to request entity/span-style strings (§3.1); regenerated short-channel rows use a lower new-token cap (32) than the frozen headline export (256), so answer surfaces and sentence-level MiniCheck scores differ from the default-instruction channel. Table 8 supplies the numeric anchor, with strict F1 as the primary pipeline sanity read.

Table 7: Compact answer-channel sanity audit.

Policy	EM/F1$_{N_{\text{sub}}}$	EM/F1$_{N_{\text{pair}}}$	F1$_{\text{rep}}$
Top-k	.008/.048	.006/.044	.229
Provence	.008/.079	.009/.081	.333
MSC selector	—	.071/.143	.424

$N_{\text{sub}} = 361$; $N_{\text{pair}} = 325$. Repaired F1 is diagnostic only and does not modify token or judge inputs.

A.4 Answer–judge channel separation

Headline EM/F1 strings are short answer-channel artifacts; MiniCheck scores longer context-conditioned exports on the packed judge surface. The preceding paragraph and Table 7 make the mismatch explicit: low answer-channel match does not imply judge-channel collapse under the frozen contract (§7 for judge scope).

B Calibration Details

Scope. Appendix B collects judge-channel companions for §3.2: auxiliary aggregates, focal spot-checks, and second-judge sensitivity panels that contextualize MiniCheck readouts without replacing the adjudicated audit in Table 4. Oracle-side $|D_0|$ scalars and cross-generator rows that bound the mechanism-side claim in §4 are in Appendix C (see also §4).

B.1 Sentence-level aggregates and cross-judge panel

Table 10 separates paired-325 sentence-level aggregates (**Panel A**) from auxiliary answer-level focal spot-checks (**Panel B**; see Table 9). Figure 5 gives the second-judge triangulation figure (thresholds, pairwise map, counts).

B.2 Focal spot-check and second-judge diagnostics

This subsection records auxiliary calibration context only; it does not extend the main n=504 audit. Table 10 **Panel A** restates headline paired-325 sentence-level aggregates; **Panel B** reports focal-policy spot-check readouts (binary and stricter three-way agreement where registered). Wilson-style pooled aggregates remain indexed in the anonymized supplementary package only; they are not tabulated in Panel B, and a supplementary-only Wilson calibration panel avoids duplicating Figure 5. Figure 5 gives the combined second-judge threshold, pairwise-ordering, and count panels (middle: cool-toned vs. warm-toned cells for aggregate ordering agreement). An auxiliary three-panel diagnostic (confusion matrix, τ sweep, per-policy unsupported/support shares) remains indexed in the anonymized supplementary reproducibility bundle alongside Table 10 but is **not** embedded in the review PDF to avoid redundancy with Figure 5 and main-text calibration figures.

C Mechanism and Miss Taxonomy

Appendix C supports the mechanism analysis in §4 with mechanism-side lookups on the same paired-325 identifiers as the headline tables: miss taxonomy, oracle-side $|D_0|$ sensitivity, bounded cross-generator direction checks, and qualitative illustrations. **Scope.** Mechanistic localization without causal identification; auxiliary panels only.

C.1 Cross-contract taxonomy v2 (adjudicated)

The three-class miss taxonomy in §4.2 is defined on an answer-miss slice under the NQ paired contract and is therefore not intended to transfer mechanically to a second headline contract. To strengthen the mechanism story beyond one contract, we instantiate a cross-contract *taxonomy v2* on the same packed-evidence surface used by the judge channel, with buckets that cover retrieval ceiling, selection gap, evidence surfacing, alias mismatch, and conflict binding (multi-hop composition is reserved for non-headline multi-hop stress slices). We adjudicate a dual-annotator pool balanced across the two headline paired contracts (n=200 per contract; n=400 total) under one shared decision tree; Table 11 reports final-label bucket counts. Pre-adjudication inter-annotator agreement on this fine-grained v2 schema is modest (raw agreement ≈ 0.498, Cohen's $\kappa \approx 0.078$), so the table reports **final adjudicated labels** as a contract-conditioned bucket map rather than a universal mechanism law. When bucket shares differ across contracts, the correct read is diagnostic localization under each frozen contract, not a policy leaderboard. The guideline, pool schema, and merge procedure are indexed in the anonymized supplementary package for reproducibility.

C.2 Mismatch map for optimizable misses

This subsection verifies the manuscript-facing mismatch decomposition that supports the frozen taxonomy counts. The table below restates the frozen three-class partition and optimizable-mismatch split summarized in §4.2 and Figure 4, with

Policy	Tok.	EM	F1	EM_{rep}	F1_{rep}	MC unsup.
Top-k	428.4	0.197	0.299	0.345	0.399	0.441
Provence	428.4	0.197	0.299	0.348	0.402	0.450
Adaptive-k	314.0	0.178	0.267	0.295	0.343	0.396
MSC selector	427.8	0.175	0.276	0.308	0.361	0.568

Table 8: **Short-answer controlled channel.** Paired NQ identifiers (n=325) at B=512; only the answer instruction differs from the default-instruction headline channel (regenerated with a lower new-token cap).

Same retrieval, selection, packing, tokenizer scope, and judge-facing evidence as the headline budget-grid contract. **Tok.** is macro-averaged mean **input** tokens in the short-channel log (small shifts vs. Table 2 at this B reflect the short-answer prompt template and logging template, not generated completion length in the input-token counter). **MC unsup.** is MiniCheck on short-answer surfaces—do **not** cross-read against the main default-instruction unsupportedness comparison in Tables 2–3.

Calibration pool	Unit and size	Policy coverage	Interpretation	Use
Adjudicated audit	sentence units; $n = 504$	all four policies	Primary human calibration bound for MiniCheck on the packed evidence surface.	MAIN
B=512 spot audit	query–policy samples; NQ: $n = 60$ per policy	NQ all four; selected TriviaQA-open rows	Fixed-budget sensitivity check on regenerated packed evidence; NQ human rates summarized in §3.2.	AUX.
Focal spot-check	answer-level focal rows; variable n	Top-k, MSC selector, and pooled focal rows	Historical triangulation for focal readouts; not a denominator for headline claims.	AUX.
Second-judge grid	sentence/query aggregates; 325 paired IDs	selected policy contrasts	Automatic-judge sensitivity over thresholded second-judge readouts.	AUX.

Table 9: **Calibration pools and units.** Rows intentionally use different unit definitions and denominators. Only the adjudicated audit is the main calibration bound; auxiliary rows contextualize MiniCheck readouts and should not be substituted for the frozen headline judge channel.

Focal spot-check sizes: Top-k $n = 14$, MSC selector $n = 86$, pooled stratified pool $n = 100$. TriviaQA-open rows in the B=512 spot audit reuse an earlier preregistered export for Top-k and Adaptive-k.

Panel A. Headline MiniCheck aggregates (paired-325, sentence-level)

Policy	n_q	Sents.	Unsup. ↓	Grd. ↑	Role
Top-k	325	1,110	0.503	0.497	Top-k aggregate / focal anchor
Provence	325	990	0.995	0.005	Aggregate only
MSC selector	325	991	0.712	0.288	MSC aggregate / focal anchor

Panel B. Auxiliary focal agreement (answer-level spot-check units)

Pool	n	Bin. agr. ↑	3-way ↑	Role
Top-k focal	14	0.214	0.071	Top-k focal anchor
MSC selector focal	86	0.756	0.651	MSC focal anchor
Pooled spot-check	100	0.680	–	Pooled summary

Table 10: **Auxiliary calibration aggregates and focal agreement.** Panel A restates paired-325 sentence-level MiniCheck aggregates. Panel B reports auxiliary answer-level focal spot-check agreement.

Unsup. is the MiniCheck unsupported-claim share; Grd. is its complement. Bin. agr. and 3-way report human–MiniCheck agreement on focal answer units where applicable, not inter-annotator agreement.

Separating panels avoids mixing sentence-level headline aggregates with auxiliary answer-level spot-check units in the same denominator column.

Focal spot-check sizes: Top-k n=14, MSC selector n=86, pooled stratified pool n=100. A Wilson-style pooled panel is indexed in the anonymized supplementary package only; the review PDF uses this table together with Figure 5 for auxiliary calibration context.

manuscript-facing counts for n_{miss}=137. For mapping of optimizable mass, read this appendix in order: selection/truncation mismatch (support in pool but omitted from packed prompt) and packing/alias mismatch (support selected but answer-critical alias absent from packed sentence context); detailed per-example traces are indexed in the anonymized supplementary package.

Dual coding on frozen exemplars. Dual coding over all currently available frozen exemplars (n=63) reports overall agreement 0.9206 and three-

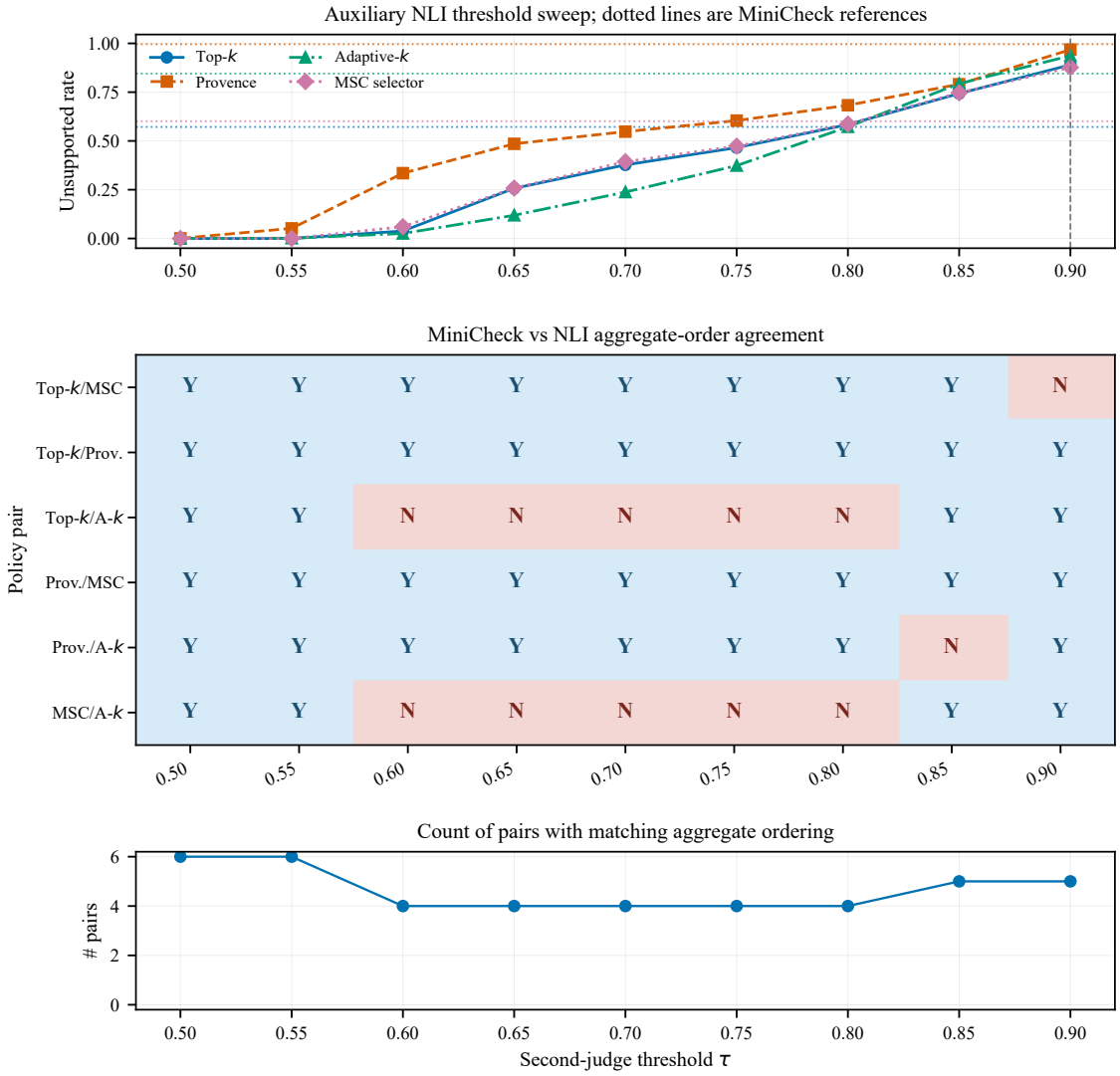

Figure 5: **Second-judge triangulation (paired-325 aggregates). (Top)** Auxiliary-NLI unsupported-claim rate vs. τ with MiniCheck levels as horizontal references (dotted). **(Middle)** Six policy pairs at each τ: cool-toned cells mark matching aggregate ordering; warm-toned cells mark mismatch; cell letters are **Y**=agree / **N**=disagree on aggregate order. **(Bottom)** Count of pairs with matching ordering vs. τ. Scoped diagnostic (§7).

Bucket	NQ	TriviaQA	Pooled
Retrieval ceiling	50 (25.0%)	100 (50.0%)	150 (37.5%)
Selection gap	16 (8.0%)	1 (0.5%)	17 (4.2%)
Evidence surfacing	131 (65.5%)	70 (35.0%)	201 (50.2%)
Alias mismatch	1 (0.5%)	23 (11.5%)	24 (6.0%)
Conflict binding	2 (1.0%)	6 (3.0%)	8 (2.0%)
Multi-hop composition	0 (0.0%)	0 (0.0%)	0 (0.0%)

Table 11: **Adjudicated cross-contract taxonomy v2 bucket counts.** n=400 (200 per headline contract). Counts are **final adjudicated-label** shares under a contract-conditioned diagnostic map (§7).
Cells report n and within-column percentage.

class κ=0.8784; disagreements are concentrated within the two optimizable classes, while the retrieval-ceiling bucket remains fully consistent. As a protocol-bounded what-if on NQ, resolving the full optimizable mass would reduce misses by at most 43 in this audited universe (upper bound only; not a causal efficacy claim for any method).

Budget-grid slice readout (NQ). Table 13 averages query-level MiniCheck unsupported-claim rate and mean input tokens on the preregistered budget grid (§2.2.1) within the frozen retrieval-ceiling

Miss class	n	% miss	Optimizable?	Example ID
Retrieval ceiling	94	68.6	No	nq_dev_000040
Selection gap	27	19.7	Yes	nq_dev_000325
Evidence surfacing	16	11.7	Yes	nq_dev_000339
Total	**137**	**100.0**	**43 yes**	—

Table 12: **Frozen three-class miss taxonomy on paired-325 NQ.** Optimizable mass is 27+16 misses; the remainder indexes pool ceilings.

Example IDs illustrate buckets only; full keys and tie-break order remain in the supplementary taxonomy protocol. Pairwise-325 incidence shares appear in the mechanism prose; this table is aligned to Figure 4 counts **94/27/16**.

slice (n=94), the frozen optimizable miss union (n=43), and all paired-325 queries under the headline contract. Slices are reconstructed from frozen oracle-support labels on the $|D_0|$=20 candidate lists and from the preregistered optimizable-query identifier list archived with the failure-slice audit in the anonymized supplementary package; the table is a diagnostic bridge between Figure 4 and Figure 2, not a causal decomposition.

Per-query token–unsupportedness deltas. Figure 6 plots each non–Top-k policy's per-query Δinput tokens and ΔMiniCheck unsupported-claim rate relative to Top-k on paired-325 Natural Questions. Spearman correlations in Table 14 are *diagnostic only*: they do not identify causal mechanisms and do not overturn aggregate rank mismatch in Tables 2–3.

Oracle $|D_0|$ sensitivity (mechanism scalars). Table 15 supports the reachability/bottleneck split in §4.1: increasing $|D_0|$ from 20 to 50 raises oracle-feasible mass, while prefix@5 oracle coverage drops and selection-bottleneck mass rises under the same frozen selection contract. The movement is directional rather than uniformly improving, so pool expansion is not a monotonic fix.

Paired-100 subset reference (primary stack). Table 16 reports Top-k and MSC token and unsupportedness on the paired-100 IDs under the headline hybrid retriever, primary generator, and MiniCheck-real judge surface. We did not evaluate an alternate dense-only retriever under this submission, so we make no claims about alternate-retriever orderings (§7).

Cross-generator direction check. Table 17 reports a limited boundary diagnostic: the same packed contexts and paired-325 IDs with a secondary open-weights generator (Mistral-7B-Instruct-v0.2) swapped in place of the primary Qwen2.5-14B stack. Token and unsupported-claim

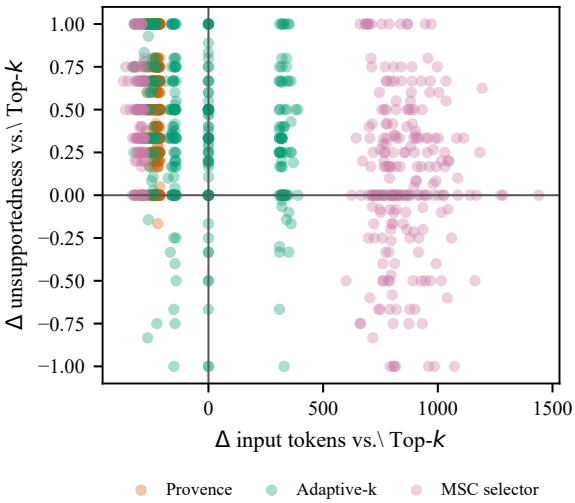

Figure 6: **Per-query Δtoken vs. Δunsupportedness** (non–Top-k policies vs. Top-k on NQ paired-325). Correlation is a diagnostic readout only (§7).

orderings echo the headline contract while absolute rates rescale; this is same-context sensitivity support for the mechanism localization in §4, not full cross-stack reproduction or external validity.

C.3 Qualitative case illustrations

Figure 7 condenses three mechanism-aligned case cards (panels A–C: retrieval-ceiling, selection-gap, evidence-surfacing/alias) using **anonymized** evidence / outcome / read sketches tied to the frozen bucket definitions; they illustrate shapes of failure, not frequencies.

Representative error sketches. The three-class taxonomy (§4.2) is illustrated at the aggregate level in Figure 4 and with qualitative case cards in Figure 7. Qualitative examples follow the frozen labels: (i) *oracle_impossible*—no support=1 document in D_0; (ii) *supported_doc_not_selected*—gold support exists in D_0 but not in the packed prompt; (iii) *alias_missing_context*—support document is selected but gold aliases do not appear in the concatenated sentence context seen by the generator.

Policy	Budget cap B: MiniCheck unsupported-claim rate ↓				Tok.$_{512}$ ↓
	128	256	512	1024	
Retrieval ceiling (n=94)					
Top-k	1.000	0.793	**0.660**	0.601	416
Provence	1.000	0.781	0.675	**0.591**	416
Adaptive-k	**0.743**	**0.747**	0.734	0.613	**283**
MSC	1.000	0.813	0.751	0.638	413
Optimizable (n=43)					
Top-k	1.000	0.685	0.497	0.489	413
Provence	1.000	0.695	**0.465**	**0.479**	413
Adaptive-k	**0.694**	**0.671**	0.553	0.488	**319**
MSC	1.000	0.776	0.626	0.582	417
All paired (n=325)					
Top-k	1.000	0.693	0.521	0.508	415
Provence	1.000	0.696	**0.519**	**0.500**	415
Adaptive-k	**0.652**	**0.639**	0.615	0.534	**301**
MSC	1.000	0.752	0.618	0.552	416

Table 13: **Mechanism–budget bridge on NQ paired-325.** Cells are mean MiniCheck unsupported-claim rates on the preregistered budget grid; lower values are better.

Gray bands mark diagnostic slices; bold marks the within-slice minimum for each budget column and for Tok.$_{512}$. Slices match Figure 4; this table is diagnostic, not causal decomposition. Each cell averages query-level rows at (policy, B); full slice keys are in the anonymized supplementary package.

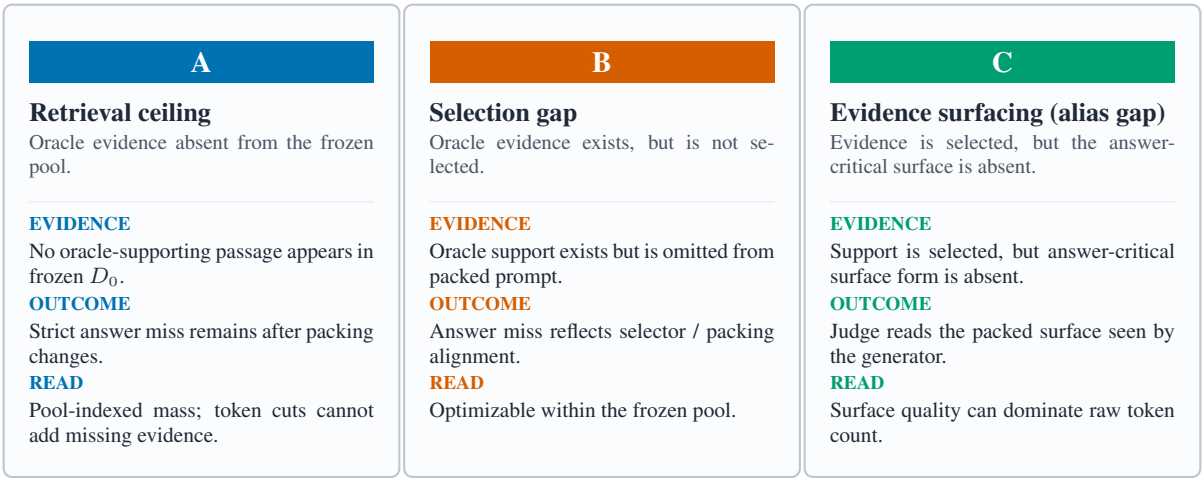

Figure 7: Appendix case cards for the three failure modes. Each card separates observed evidence, resulting outcome, and the operational readout. Sketches are anonymized and do not introduce new numerics.

Slice	n	Spearman ρ	p
All (non Top-k rows)	975	−0.349	0.000
Retrieval ceiling	282	−0.140	0.019
Optimizable	129	−0.363	0.000

Table 14: **Per-query Δtoken vs. Δunsupported (diagnostic).** Correlations are not causal; aggregate policy rankings can still decouple.

Scalar	D_0=20	D_0=50	Δ
Feasible rate	0.711	0.785	0.074
Oracle cov. prefix@5	0.299	0.206	–0.092
Selection bottleneck mass	0.412	0.579	0.166

Table 15: $|D_0|$ **sensitivity (paired-325).** Feasible mass rises from 20 to 50 candidates while prefix@5 oracle coverage falls and bottleneck mass rises—pool expansion is not monotone.

Δ is D_0=50 minus D_0=20 on each scalar. Mechanism scalars under the frozen selection contract; not a headline replacement (Table 2) and not a claim that larger $|D_0|$ improves groundedness outcomes.

Paraphrased illustrations (anonymized). The cards in Figure 7 restate these three operational buckets without reproducing identifying surface strings; they are meant to prevent bucket definitions from reading as abstract placeholders while keeping the review bundle de-identified.

Retriever	Policy	Tok.	Unsup.	Direction vs. primary subset	Read
primary (hybrid)	Top-k	1,166.4	0.484	baseline subset	reference
primary (hybrid)	MSC selector	539.2	0.714	baseline subset	reference

Table 16: **Paired-100 subset reference (primary hybrid stack).** Token cost and MiniCheck-real unsupportedness on the paired-100 IDs under the headline retriever and primary generator. We did not evaluate an alternate dense-only retriever under this submission, so we make no claims about alternate-retriever orderings.

Rows: primary hybrid-stack Top-k and MSC on paired-100 manifest IDs (same generator, packing, and judge surface as headline slices).

Contract	Generator	Policy	Tok.	Unsup.	Direction vs. primary	Read
NQ paired-325	primary (Qwen2.5-14B)	Top-k	355.1	0.503	baseline	headline
NQ paired-325	secondary (Mistral-7B)	Top-k	355.1	0.310	order preserved	boundary
NQ paired-325	primary (Qwen2.5-14B)	Provence	123.8	0.995	baseline	headline
NQ paired-325	secondary (Mistral-7B)	Provence	123.8	0.994	high unsup. persists	boundary
NQ paired-325	primary (Qwen2.5-14B)	MSC selector	876.1	0.712	baseline	headline
NQ paired-325	secondary (Mistral-7B)	MSC selector	876.1	0.915	absolute rates rescale	boundary

Table 17: **Boundary cross-generator echo (NQ paired-325).** Same packed contexts and paired IDs; only the answer generator changes. Token and unsupported-claim *orderings* match the primary contract, while absolute unsupported rates rescale after the generator swap. Not full cross-stack reproduction or external validity.

Secondary generator: Mistral-7B-Instruct-v0.2; primary: Qwen2.5-14B-Instruct. Mean tokens are unchanged because packing is frozen.

Paired-100 pilot (smaller n, primary generator): Top-k tok. 1166.4 / unsup. 0.484; Provence 85.7 / 0.993; MSC 539.2 / 0.714—directional echo only, not a second headline contract.

D Boundary Stress Tests

Appendix D hosts boundary-only diagnostics placed in the appendix under ARR page limits: paired-325 tradeoff companions, cross-contract plots, second-contract reproducibility rows, in-contract perturbation forests, and a multi-hop Hotpot slice. These objects localize stress directions and contract-specific token–unsupported relationships; they are not co-primary benchmarks, deployment-robustness certification, or universal ranking evidence. **Scope.** Boundary stress diagnostics only; always read with §7.

Why these diagnostics are reported in the appendix. Not every headline policy was run through the same paired perturbation interface (Provence and Adaptive-k are omitted from the robustness forest), and the panels are not co-primary benchmarks or deployment certificates. The main text therefore keeps only a directional bridge (§4.3); full stress panels live here under §7.

D.1 Main-text companion floats (appendix placement)

This subsection hosts figure and table companions placed in the appendix to satisfy eight-page main-body constraints; labels are unchanged so in-text references remain valid.

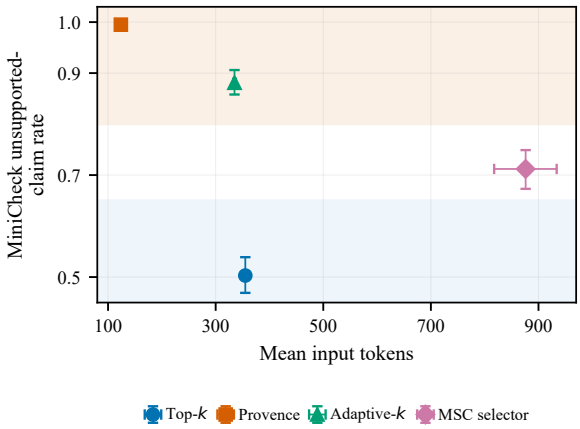

Figure 8: **Headline single-configuration anchor (paired-325, n=325).** The four headline *single-configuration* operating points from Table 2 (one token–unsupported coordinate per policy) with 95% paired-bootstrap percentile intervals for per-policy means under the frozen headline export (bootstrap bookkeeping in Appendix E.3.1). Read with Figure 2 as an anchor to the frozen budget curves. Same-judge readout on packed selected evidence; interpretive bounds in §7.

Cross-contract budget curves (omitted appendix replot). Figure 2 already juxtaposes Natural Questions paired-325 and TriviaQA-open paired-325 under identical curve grammar; we omit a separate appendix-side-by-side replot to avoid redundancy with the main-text figure.

TriviaQA-open headline rows (appendix). §3.3 anchors the cross-contract token–support read solely in Table 5. Table 18 restates the frozen TriviaQA-open paired-325 slice with EM/F1 alongside tokenizer-scoped mean input tokens and MiniCheck unsupported-claim rate for reproducibility and answer-channel sanity; it does not introduce claims beyond the cross-contract summary.

Policy	EM	F1	Tok.	Unsup.
Top-k	0.049	0.119	1 004.6	0.967
Provence	0.049	0.119	1 004.6	0.967
Adaptive-k	0.068	0.159	356.8	0.747
MSC selector	0.135	0.276	393.5	0.847

Table 18: **TriviaQA-open appendix companion to Table 5.** Same four policies and judge surface; EM/F1 are answer-channel context only.

Not a universal cross-task ranking; values are frozen second-contract aggregates.
Top-k and Provence share identical Tok./Unsup. because Provence pruning yields zero token reduction under this contract and the exports reuse the same frozen packed evidence (frozen-export control; see Table 5).

Robustness forest scope. Figure 9 summarizes Top-k and MSC rows aligned with the headline MSC-baseline contrast family in Table 21; underlying row-level aggregates are listed in the anonymized supplementary package. Provence and Adaptive-k do not appear in this forest plot because they were not evaluated under the same paired perturbation interface; we therefore do not use this figure as a four-policy robustness comparison. Additional policy-by-perturbation rows, if available, are provided separately in the supplementary package rather than in this panel.

Fam.	Row	n
Evidence edit	remove-lead / Top-k	325
Evidence edit	remove-lead / MSC	325
Evidence edit	lead-distractor / Top-k	325
Evidence edit	lead-distractor / MSC	325
Conflict inj.	entity-slot / Top-k	322
Conflict inj.	entity-slot / MSC	322
Conflict inj.	numeric-conflict / Top-k	325
Conflict inj.	numeric-conflict / MSC	325
Conflict inj.	time-version / Top-k	197
Conflict inj.	time-version / MSC	168

Table 19: **Effective query counts for Figure 9.**

Effective n drops when a row precondition is undefined for a query.
CI rule (Figure 9): paired-query bootstrap, B=1000, seed 20260331, nominal 95% percentile intervals on ΔUnsupported.

D.2 Statistical procedures for appendix stress exports

Stress CIs use paired-query bootstrap resampling with B=1000 and seed 20260331 under the frozen stress protocol; effective n per row appears in Table 19 and in the anonymized supplementary package. Headline paired significance uses B=5000, seed 20260323 (Appendix E.3.1, Table 21). Perturbation-level exports (ΔEM, ΔF1, Δ unsupported-claim rate, effective n) and field definitions are indexed in the **anonymized supplementary package** with bootstrap documentation; no perturbation is added beyond the frozen failure-analysis contract.

D.3 Multi-hop boundary stress

HotpotQA is a boundary-only stress slice on the D_0**-nonempty** subset (n=740), not a co-primary headline benchmark; it complements Figure 9 and §4.3 without enlarging the headline universe. Relative to Top-k, compression lowers mean input tokens while support-any-doc coverage drops and unsupported-claim rate rises with near-flat Answer F1; this is interpreted as boundary stress evidence under the fixed protocol only. We additionally expand this slice with a **mechanism stress block** (support-any-doc coverage, support-sentence coverage, unsupported-claim rate and contradiction by hop type, and multi-hop failure buckets) under the same D_0-nonempty protocol; full audit artifacts are indexed in the anonymized supplementary package.

Policy	Tok.	F1	Sup.	Unsup.
MSC selector	458.9	0.015	0.580	0.596
Top-k (k=10)	2 017.6	0.015	0.830	0.546

Table 20: **Boundary stress (HotpotQA on D_0-nonempty (n=740).** Not a co-primary benchmark.

Sup. is support-any-doc coverage; Unsup. is MiniCheck unsupported-claim rate on packed selected evidence.
Read: lower tokens can co-occur with lower support coverage and higher unsupported rate with near-flat F1; boundary stress only.

Mechanism readout. Within the same stress slice, hop-conditioned diagnostics show that *comparison* questions have lower support-sentence coverage and higher unsupported-claim rates than *bridge* questions (e.g., selector: support-sentence coverage 0.097 vs. 0.154, unsupported-claim rate 0.682 vs. 0.508), while contradiction signals stay low and heuristic-bound. Failure buckets are dominated by support-surface failures (missing gold

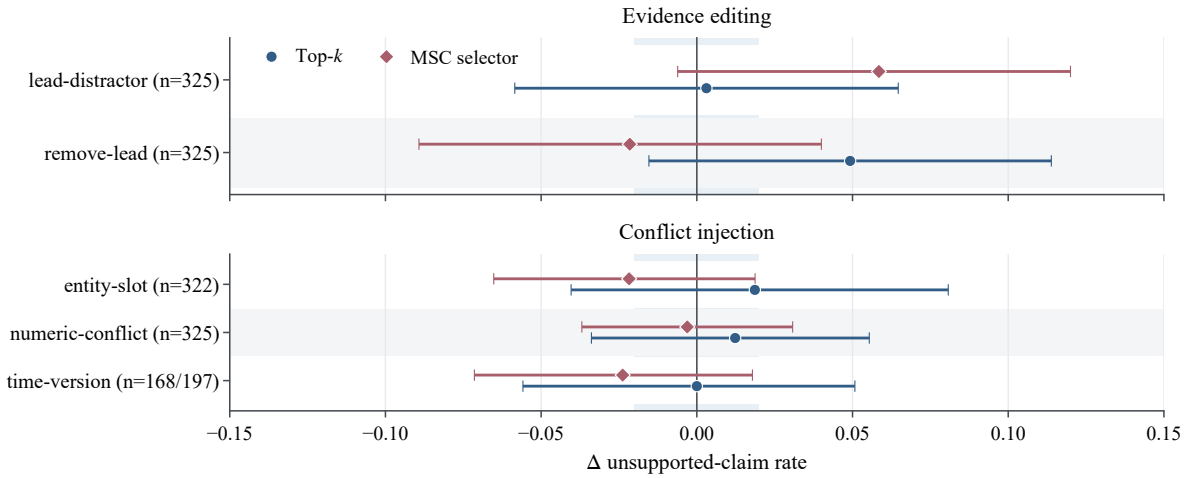

Figure 9: In-contract robustness stress on paired Natural Questions IDs: 95% intervals on Δ(unsupported-claim rate). **Top:** evidence-edit rows; **bottom:** conflict-injection rows. Each row pairs Top-k (circles) and MSC (diamonds) under one perturbation. **Policies:** Top-k and MSC only (Provence and Adaptive-k were not run through the same perturbation interface). **Effective n:** Table 19. **CI:** paired-query bootstrap, B=1000, seed 20260331. Shaded band: near-zero region. Boundary diagnostic only; not deployment robustness certification.

document in the retrieved pool, or missing gold support sentence in retrieved text) rather than answer-channel collapse, supporting the mechanism-facing interpretation of multi-hop stress without promoting Hotpot to a headline universe; standardized bucket tags in the supplementary export follow the same taxonomy categories.

E Reproducibility Package Index

Appendix E is a traceability index rather than an additional results section. It expands the main-text reporting block into machine-readable manifests, bootstrap mirrors, calibration manifests, and figure/table provenance anchors.

E.1 Auditable chain and main-text independence

Auditable chain. Our reproducibility contract is explicitly chained as: **contract manifest** → **result manifest** → **calibration manifest** → **claim/table/figure provenance map**. Main-text tables and figures cite frozen aggregates bundled in the **anonymized supplementary package**; proxy scores and superseded mixed-universe exports are excluded from headline claims.

Reproducibility verification trace. The intended trace is: manuscript claim → cited table/figure → matching aggregate artifact → optional script replay and universe note (Appendix E.4).

Main-text independence. The main claims are supported by the main-text contract definition, Tables 2–5, and Figures 1–4. Appendices provide denominator checks, calibration details, stress diagnostics, bootstrap mirrors, and regeneration hooks; they do not introduce additional premises required to read the headline token–unsupportedness mismatch.

E.2 Frozen configs and data universe

Compute and runtime. Experiments use GPU-backed hybrid retrieval scoring and 14B-class instruction-tuned generation on a commodity research setup (single ≈24–48GB-class GPU or an equivalent small multi-GPU configuration; exact host identities are omitted for anonymity). We did not centrally log wall-clock times; Table 23 summarizes what is frozen in the review PDF versus the supplement. **Order-of-magnitude GPU-hours (author estimate).** Counting dense retrieval index construction, paired-query generation exports for four policies under the preregistered budget grid and headline configurations, MiniCheck-real passes, auxiliary second-judge sweeps, bounded robustness replays, and calibration replays on regenerated packed surfaces, the full study is on the order of **150–400 GPU-hours** under our nominal per-batch throughput assumptions (dominated by generator forward passes and judge passes, not by CPU-side pooling). This band is intended only as a reproducibility scale cue, not a certified carbon accounting statement. The frozen Wikipedia-style

passage index is $\mathcal{O}(10^7)$ passages at snapshot granularity (exact shard sizes in the supplement); mean per-query end-to-end latency was not tuned or profiled as a headline metric.

Models and configuration. The frozen retrieval stack matches the preregistered hybrid sparse–dense configuration in §2.2 (BM25-style lexical scoring fused with dense passage scoring). The generator is Qwen2.5-14B-Instruct with deterministic decoding (§2.2). Faithfulness uses MiniCheck in **real** mode on packed selected evidence; appendix diagnostics additionally include a second automatic sentence-level judge with a pre-registered τ grid (§3.2), which probes judge specificity rather than replacing human gold.

Hyperparameters and search. Headline tables and figures rely on **pre-specified** decoding fields, MSC stopping constants (γ_{stop}, γ_{no}), paired-bootstrap settings (B=5000, seed 20260323), and the pre-selected MSC baseline for contrast tests—not hyperparameters chosen by iterative tuning on held-out headline metrics. Threshold sweeps for the auxiliary second judge and Adaptive-k threshold-grid alignment (mean-token band vs. Top-k, §E.2) are boundary stability probes in Appendix B.

Data and human annotation. The headline paired universe is the frozen $N_{\text{pair}} = 325$ query list in §2.2 and Table 6, nested under the official NQ-open dev split ($N_{\text{raw}} = 3{,}610$) and the preregistered development subsample ($N_{\text{sub}} = 361$). Human sentence calibration follows a frozen English three-class guideline with **dual independent** labeling, **adjudication** on disagreement, and **merge** to adjudicated finals on the pre-registered dual-human calibration pool, as summarized in §2.2–3.2 (numeric audit thresholds: §7; Ethical Considerations); stratified unit lists and merge metadata appear only in the anonymized supplementary package.

Expanded calibration pool and audit add-ons. Earlier auxiliary merged and focal pools (Appendix B) (Wilson-style panels, second-judge triangulation, and—indexed in the supplementary bundle only—an auxiliary three-panel diagnostic; unit definitions differ from the adjudicated pool) support auxiliary calibration context; separately, we construct an adjudicated sentence-unit pool (n=504; four-policy balance) on the same packed-evidence surface to anchor the main-text calibration readout (Table 4, Figure 3). We also run bounded judge-sensitivity diagnostics (a second automatic judge with a fixed τ grid and deterministic preprocessing controls) and controlled gold-support surface edits under the frozen packed-evidence template (Appendix E.2); full manifests and regeneration notes are indexed in the anonymized supplementary package.

Controlled gold-support surface edits (boundary diagnostic). Deterministic edits to answer sentences linked to supporting evidence on the **same packed surface** shown to the generator, followed by MiniCheck-real rescoring under the frozen checkpoint, probe judge sensitivity without new headline numerals or policy rankings; row manifests are indexed in the anonymized supplementary package.

Second headline contract. We instantiate a second frozen paired contract (TriviaQA-open; n=325) under the same four-policy protocol discipline and judge-surface definition as the primary NQ contract. This second contract supports a **contract-conditioned** read relative to paired-325; full contract manifests, calibration pool definitions, and audit artifacts are indexed in the anonymized supplementary package (scope in §7).

Figure specification. A human-readable figure master specification (panel intent, claim boundaries, readability constraints for ACL-style review PDFs) is included in the supplementary package alongside vector sources for the main-text and appendix figures cited above; the Wilson pooled calibration panel remains supplementary-indexed only. Figure entries in the supplementary reproducibility index use semantic anchors `F1`–`F5`; concrete build filenames are omitted from the PDF body.

Anonymity. Supplementary materials follow ARR rules: no trackable public hosting links in the review bundle.

Data universe pointer. The full denominator chain and answer-channel normalization for paired-325 are documented once in Appendix A (Table 6) and §2.2; we omit a second prose copy here to avoid redundancy. HotpotQA appears **only** as an appendix stress boundary (Appendix D.3, Table 20) and does **not** enlarge the headline universe.

Frozen generator, decoding, and exports. We use Qwen2.5-14B-Instruct with deterministic decoding (temperature 0, top-p 1, max new tokens 256 on frozen headline exports) and fixed

templating across policies (§2.2). Decoding fields, optional versioned digests, and calibration manifests needed to re-aggregate headline numbers are enumerated in the anonymized supplementary package.

Retrieval interface and context packing. Every policy consumes the **same** hybrid retriever output with $|D_0(q)|$=20 candidates per query under identical indexing (§2.2). Top-k packing, Provence-style compression, Adaptive-k, and the MSC-style selector differ in *how* evidence is chosen and packed into the generator prompt, not in retriever identity or pool size; headline conclusions are **paired, same-interface diagnostics** (Appendix E.2).

E.3 Statistical procedures

The subsections below scope each audit artifact; the corresponding full-width tables are collected at the end of this appendix in contract order.

Appendix A confines itself to the **data universe** and **answer-channel normalization** sanity checks on the frozen evaluation artifacts. Headline paired significance uses paired-query bootstrap with B=5000 and seed 20260323; appendix stress CIs use B=1000 and seed 20260331 (Appendix D.2).

E.3.1 Inferential endpoints and budget-grid contrasts

Tables 21 and 22 provide the two inferential anchors for the headline and budget-grid claims. Table 21 mirrors the main paired-325 endpoint contrasts from Table 2 under the same frozen exports, paired query IDs, bootstrap setting, and seed. Table 22 gives the compact preregistered B=512 budget-cap Holm summary for **Tok.** and **Unsup.**, with Holm adjustment computed over the full contrast×metric family and the full pairwise matrix stored in the anonymized supplementary package. Together, these tables are audit mirrors for already-defined claims rather than additional model-selection criteria.

E.3.2 MiniCheck preprocessing audit

We pair paired-325 preprocessing diagnostics with a same-checkpoint MiniCheck-real replay under three pre-registered deterministic controls: regex sentence splitting versus the headline pipeline, explicit inter-document separators in packed context, and unicode/ZWSP normalization before splitting. On this subset, aggregate four-policy unsupported-claim *ordering* matches the frozen headline ordering under all three controls, while absolute sentence-level rates can shift. This material sharpens the within-contract read but does **not** certify invariance to arbitrary segmentation, packing, or normalization, nor portability across judge checkpoints beyond these audited transforms; regeneration metadata is listed in the anonymized supplementary package.

E.4 Responsible Reporting and Reproducibility Map

The anonymized supplementary package includes a manifest whose rows use prefixes `T*` (table aggregates), `F*` (figures), and `A*` (protocol and analysis bundles); each row links a manuscript anchor to the frozen aggregate that reproduces its numbers. Main-text tables and figures map to `T1`–`T4` and `F1`–`F5`; companion records map to `A1`–`A6`. Supplementary files follow ARR anonymization rules; no extra unpublished results are required to validate the main claims.

Reproducibility and responsible-reporting audit map. Table 23 is the review-PDF audit map for the commitments most relevant to reproducibility and responsible reporting: limitations, misuse boundaries, data and license scope, compute scale, human annotation, and the frozen paired-contract fields. The anonymized supplement stores bulky manifests, identifier lists, regeneration files, and annotation protocols rather than introducing additional evidence for the headline claims.

E.5 Paired-Contract Reporting Kit

The final audit pair operationalizes the Paired-Contract Reporting Kit (PCRK). The audit chain is: contract manifest → result manifest → calibration manifest → figure/table anchor. This structure allows each headline number to be traced back to the query identifiers, retrieval pool, packing template, tokenizer, generator, judge checkpoint, and bootstrap configuration used to produce it. The PCRK table records the minimum fields needed for joint token-cost and unsupported-rate reporting, while Table 25 situates those fields against related artifacts as a reporting-unit comparison rather than a quality ranking.

Contrast	Metric	Δ	CI low	CI high	p	d_z
Top-k–MSC	EM	–0.065	–0.092	–0.037	<.001	–0.250
Top-k–MSC	F1	–0.099	–0.128	–0.070	<.001	–0.366
Top-k–MSC	Tok	–521.4	–578.2	–463.0	<.001	–0.971
Top-k–MSC	Unsup.	–0.210	–0.256	–0.162	<.001	–0.435
Provence–MSC	EM	–0.062	–0.089	–0.034	<.001	–0.232
Provence–MSC	F1	–0.062	–0.090	–0.035	<.001	–0.238
Provence–MSC	Tok	–753.0	–809.8	–695.5	<.001	–1.398
Provence–MSC	Unsup.	0.283	0.245	0.322	<.001	0.852

Table 21: **Headline paired-bootstrap endpoints.** Mirror of Table 2 under one frozen contract.

Deltas are compare-minus-MSC on the same 325 query IDs. Machine-readable full matrices remain in the anonymized supplementary package.

Panel A. Natural Questions paired-325 (B=512 **Holm-significant rows)**

Metric	Reported contrast	Lower value	Δ [95% CI]	p_{Holm}
Unsup.	Top-k–MSC	Top-k	-0.097 [-0.138, -0.056]	.008
Unsup.	Provence–MSC	Provence	-0.099 [-0.139, -0.058]	.008
Tok.	Adaptive-k–MSC	Adaptive-k	-114.951 [-125.170, -104.855]	.008
Tok.	Adaptive-k–Top-k	Adaptive-k	-114.462 [-123.807, -104.675]	.008
Unsup.	Adaptive-k–Top-k	Top-k	0.095 [0.065, 0.126]	.008

Panel B. TriviaQA-open paired-325 (B=512 **Holm-significant rows)**

Metric	Reported contrast	Lower value	Δ [95% CI]	p_{Holm}
Tok.	Top-k–MSC	MSC	30.246 [25.225, 35.028]	.006
Unsup.	Top-k–MSC	Top-k	-0.165 [-0.223, -0.108]	.006
Tok.	Provence–MSC	MSC	30.246 [25.114, 35.161]	.006
Unsup.	Provence–MSC	Provence	-0.179 [-0.236, -0.123]	.006
Tok.	Adaptive-k–MSC	Adaptive-k	-68.449 [-78.764, -57.997]	.006
Unsup.	Adaptive-k–MSC	Adaptive-k	-0.135 [-0.190, -0.078]	.006
Tok.	Adaptive-k–Top-k	Adaptive-k	-98.695 [-108.767, -88.159]	.006

Table 22: B=512 **Holm-significant paired contrasts.** Panel A and Panel B list unique significant **Tok.** or **Unsup.** rows after Holm adjustment over the full contrast-by-metric family on each frozen contract. Mirrored duplicate rows are omitted; Δ is left minus right in the reported contrast. Lower is better for both metrics.

CI: paired-query bootstrap percentile interval with $B = 5000$ and seed 20260323. Full pairwise matrices remain in the anonymized supplementary package.

Audit item	Review-PDF anchor	Anonymized supplement
Scope, limitations, and misuse	§7; Ethical Considerations	—
Data, licenses, and artifact scope	Ethical Considerations; Appendix E.1	Dataset/model license records and version digests
Compute scale	Appendix E.2 compute paragraph	Regeneration manifests; optional timing logs
Human annotation	Ethical Considerations; §2.2–3.2	Guideline, adjudication protocol, and stratified unit lists
Paired-contract universe	§2.2; Table 6	Keyed or hashed query-ID list; contract manifest
Retrieval, generation, and judge configuration	§2.2; Appendix E.1; §3.2	Hybrid index/fusion manifest; generator config; MiniCheck version and preprocessing notes
Statistics and reproducibility	Appendix E.3.1; Tables 21, 22	Bootstrap CSVs, aggregate mirrors, and regeneration entry points

Table 23: **Compact reproducibility and Responsible NLP audit map.** The review PDF records the interpretive commitments needed for ARR review; the anonymized supplement stores bulky manifests and replay artifacts only.

Supplementary materials follow ARR anonymization rules: no trackable public-hosting links in the review bundle. The table is an audit map, not a new result table.

PCRK unit	Required fields	Failure mode prevented
Frozen contract	paired IDs; retriever; candidate depth; generator/decoding; packing; tokenizer	apples-to-oranges efficiency comparison
Judge readout	judge checkpoint; evidence surface; sentence segmentation; aggregation	judge-surface drift
Joint reporting	tokens; answer metric; unsupported rate; uncertainty; paired test	token-only claims
Audit chain	contract manifest → results → calibration → figure/table anchor	irreproducible headline numbers

Table 24: **Paired-Contract Reporting Kit.** The checklist records the minimum fields needed to co-report token cost and unsupported rate for RAG context selection: the frozen contract, the judge configuration, joint reporting fields, and the audit chain connecting claims to result artifacts. Full manifests live in the anonymized supplementary package.

No new numerics; complements §E.5.

Artifact	Query IDs	Generator	Packing	Token accounting	Cal.	Budget	Scope
Active RAG	Partial	Partial	Varies	Varies	Outside	Outside	Partial
Self-RAG	Varies	Fixed	Varies	Varies	Outside	Outside	Partial
Provence	Varies	Fixed	Fixed	Fixed	Outside	Outside	Partial
MiniCheck	Outside	Outside	Varies	Outside	Partial	Outside	Fixed
RAGTruth	Fixed	Outside	Outside	Outside	Human	Outside	Dataset
RAGAS/ARES	Varies	Varies	Varies	Varies	Partial	Outside	Framework
Sufficient Context	Varies	Varies	Varies	Outside	Partial	Outside	Sufficiency
Paired contract (ours)	**Required**	**Required**	**Required**	**Required**	**Required**	**Required**	**Explicit**

Table 25: Evaluation axes relevant to token–unsupportedness comparisons. Labels describe reporting units, not method quality; "Required" marks fields frozen by the paired contract.

Colors are semantic aids only; the text labels carry the meaning. *Partial*/*Varies*: context-specific or not uniformly fixed; *Fixed*/*Human*: often constant in that line; *Outside*: outside the primary artifact; *Required*: frozen by the paired contract. Not a quality ranking of prior work.