Beyond the Single Score: Utility-Contract-Sensitive Evaluation of RAG Under Retrieval Stress

Anonymous ACL submission

Abstract

RAG defenses are typically evaluated under a single, fixed objective—yet deployment settings differ: some systems must avoid harmful outputs at any cost, while others must maintain high answer coverage. We introduce StressRAG, a structured retrieval-stress benchmark covering distractor insertion and knowledge poisoning across HotpotQA and Natural Questions, designed to ask: does the best defensive policy depend on the evaluation contract? Our answer is yes—and it is sharp. On Llama-3-8B, policy rankings substantially invert under a safety-first contract vs. a balanced canonical contract (PFR = 0.733; 11/15 pairs reversed). On Qwen3-30B-A3B, we observe a regional shift across the same contracts (PFR = 0.800). A same-family Llama-3.3-70B boundary packet yields no confirm-ready policy at either contract. We contribute two offline diagnostics—Contract-Mismatch Regret (CMR_{proxy}) and Pairwise Flip Rate (PFR)—to quantify deployment risk when the evaluation contract is misspecified. StressRAG shows that defense calibration is contract-sensitive: policy recommendations should be reported under explicit utility contracts rather than collapsed into a single monolithic score.

1 Introduction

Modern Retrieval-Augmented Generation (RAG) systems depend critically on the quality of retrieved context and are therefore vulnerable to retrieval-space interventions such as semantic distractors and direct knowledge poisoning. A distractor-heavy context may cause a system to generate plausible but unsupported answers; a poisoned context may cause it to reproduce injected misinformation. A natural defensive response is to add a filtering or abstention policy—but the right policy depends on the deployment: a system that must avoid any harmful generation at all costs will prefer a very different operating point than one that must maintain high answer coverage. Recent defense narratives often imply that a single best policy should simultaneously maximize utility and minimize vulnerability. StressRAG is built as a benchmark and evaluation protocol to test that assumption directly. This creates a practical evaluation failure mode: a policy selected under one utility contract can be the wrong operating point under another.

Our core contribution is demonstrating that policy rankings change across evaluation contracts, so single-contract reporting can misidentify the operating point. A policy that performs well under a balanced canonical contract—rewarding coverage and penalizing abstention—can fail under a safety-first poison-weighted contract that heavily penalizes harmful generation and tolerates high abstention. The retained StressRAG artifacts show that contract-sensitive shifts are real, and that their sharpness depends on model family and scale.

To quantify deployment risk from retained artifacts alone, we add two offline diagnostics: CMR_{proxy} for contract-mismatch utility loss and PFR for ranking instability across contracts.

Contributions.

- **StressRAG**: a structured retrieval-stress benchmark for measuring how RAG defense rankings shift under deployment-sensitive utility contracts.
- **Contract shift changes policy selection**: the retained 8B evidence shows a crisp inversion (PFR = 0.733); Qwen shows a regional shift (PFR = 0.800).
- Two **offline diagnostics**—CMR_{proxy} and PFR—quantify contract-mismatch risk and ranking instability.
- Contract-sensitive behavior is **scale- and family-dependent**: the 70B packet yields no confirm-ready policy at either contract.

2 Related Work

Recent RAG evaluation work broadens what benchmarks measure, but typically does not make policy selection under changing utility contracts the primary reported object. MIRAGE emphasizes component-specific and metric-intensive RAG evaluation (Park et al., 2025), LaRA shows that there is no silver bullet in the broader choice between RAG and long-context routing (Li et al., 2025), and OmniEval expands multi-dimensional RAG evaluation in a structured vertical domain setting (Wang et al., 2025). StressRAG is positioned differently: the scoring contract itself is the reported variable, and the central question is how preferred policies change when that contract changes.

On the method side, RA-RAG estimates source reliability inside a multi-source RAG pipeline (Hwang et al., 2025). We use that line of work as nearby context, but our contribution remains benchmark-first rather than a competing monolithic defense claim. Our distractor interpretation is also consistent with "The Distracting Effect," which quantifies how irrelevant but semantically close passages can mislead RAG systems (Amiraz et al., 2025). This supports treating distractor `STOP` behavior as a benchmark result rather than as unfinished engineering.

Adversarial manipulation of retrieval pipelines—including corpus poisoning and context injection attacks that cause RAG systems to surface factually incorrect or harmful outputs—has received growing attention (Zou et al., 2024). StressRAG's poisoning stress family is designed to surface this failure mode under explicit contract-sensitive scoring, where the cost of a missed poison evasion is weighted independently from the cost of unnecessary abstention.

Calibrated abstention in question answering has been studied through selective prediction and conformal prediction frameworks (Kamath et al., 2020). StressRAG treats abstention rate as a first-class contract parameter rather than a post-hoc failure statistic, making the abstention–performance trade-off explicit in the reported scoring contract rather than leaving it implicit in a single aggregate metric.

Source-grounded generation and attribution benchmarks evaluate whether model outputs are supported by retrieved passages (Es et al., 2023). Our NLI source-consensus diagnostic is related to this line of work but is used diagnostically: it measures contradiction density among retrieved sources to interpret stress behavior across datasets, not to label individual outputs as faithful or unfaithful.

Adversarial NLP benchmarks and robustness stress tests have established methodological norms for evaluating model behavior under distribution shift (Goel et al., 2021). StressRAG follows the benchmark-first framing: the utility contract is the reported variable, and the contribution is a structured protocol for measuring how that variable changes policy rankings, not a claim that one policy is robust under every contract.

Table 1: StressRAG benchmark protocol. Retriever settings verified from configuration files. Coverage verified from the staged artifact packet.

Item	Setting
Datasets	NQ, HotpotQA, HotpotQA FullWiki
Retrievers	BM25; Dense-FAISS
Retriever config	k=5, 256-tok chunks, 50-tok overlap
Models	Llama-3-8B; Qwen3-30B-A3B; Llama-3.3-70B
Stress types	Poisoning, Distractors, Contradiction/Staleness, Top-k Swap
Stress levels	None / Med / High
Seeds	3 (seeds 0, 1, 2)
Cells (exp. / completed / missing)	486 / 480 / 6

3 Benchmark and Setup

StressRAG isolates two primary retrieval-level stress families: poisoning (injected contradictory facts) and distractors (semantically similar but unhelpful texts). To capture baseline variance, we evaluate standard QA tasks including HotpotQA (Yang et al., 2018) and Natural Questions (Kwiatkowski et al., 2019). We frame the dataset context constraint as a deliberate inference-budget stress test over an $N = 64$ retrieved slice. To probe generalization without reopening a full policy search, we combine low-scale Llama-3-8B evidence, a Qwen3-30B-A3B cross-family catalog, and a limited same-family Llama-3.3-70B boundary packet over two locked policies. We use the 70B packet as boundary evidence rather than as an exhaustive same-family catalog.

The frozen artifact packet already aggregates semantic metrics such as ROUGE (Lin, 2004) and BERTScore (Zhang et al., 2020), together with monotonicity, stress validity, robustness, statistical testing, and reproducibility metadata. The staged `resolved_matrix.json` file records

480 covered combinations with 6 documented missing executions (Table 1).

The six missing executions are all Qwen3-30B-A3B / HotpotQA / BM25 / seed-2 cells at medium or high stress for poisoning, distractors, and contradiction/staleness. The validation report marks the matrix as `validation_passed: true` because these are documented missing executions, not aggregation failures.

4 Utility Contracts

A *utility contract* is a scoring function that formalises the deployment priority: the **Canonical Contract** approximates a balanced assistant (coverage matters, unnecessary abstention is penalised), while the **Poison-Weighted Contract** approximates a safety-first deployment (evading poisoning is the primary objective; high abstention is tolerated). Changing the contract is equivalent to changing which failures the system is permitted to make—and our central result is that this change is enough to reverse policy rankings. These weights are diagnostic deployment contracts, not universal values.

Formally, we define score contracts as first-class evaluation objects over retained policy artifacts:

$$S_{\text{can}}(p) = d^{+}_{\text{sup}}(p) + z^{+}_{\text{sup}}(p) - 3\, r_{\text{abs}}(p) - d^{-}_{\text{deg}}(p) - z^{-}_{\text{deg}}(p) \quad (1)$$

$$S_{\text{poison}}(p) = 3\, z^{+}_{\text{sup}}(p) + 2\, r_{\text{abs}}(p) - z^{-}_{\text{deg}}(p) \quad (2)$$

where d^{+}_{sup}, z^{+}_{sup} are distractor and poisoning support rates; r_{abs} is the abstention rate; and d^{-}_{deg}, z^{-}_{deg} are distractor and poisoning degeneration rates—all per-policy aggregates over retained artifact slices.

Given two contracts a and b, **Contract-Mismatch Regret** is:

$$\text{CMR}_{\text{proxy}}(c_a \to c_b) = \max_{p} S_{c_b}(p) - S_{c_b}\left(\arg\max_{p} S_{c_a}(p)\right) \quad (3)$$

It measures the proxy utility lost under target contract c_b when the policy is selected under source contract c_a. **Pairwise Flip Rate (PFR)** is the fraction of policy pairs whose relative ordering reverses between the two contracts. Both metrics are offline diagnostics over retained artifacts and do not substitute for direct per-query evaluation.

Table 2: Policy taxonomy: short names, internal names, class, abstain tendency, and primary objective. *aggregate-only external baselines (query IDs do not align with internal catalog; see Table 3).

Short Name	Full Policy Name	Class	Abstain	Objective
Hyb-Low	hybrid_relaxed_low_abstain	Hybrid	Low	Coverage
Filt-Cons	filter_only_conservative	Filter	Low	Noise
Dist-Agg	distractor_aggressive	Filter	High	Distractors
Hyb-Red	hybrid_default_reduced	Hybrid	Med	Balance
PPL*	ppl_filter	Ext.	Med	Overall
L2*	l2_norm_filter	Ext.	Med	Poisoning

Table 3: Evidence provenance summary. *QID-aligned* indicates the artifact shares query identifiers with the 8B internal policy catalog. Claims are strictly scoped to the artifact type.

Artifact	Model	Type	QID-aligned	Claim scope
8B internal catalog	8B	Main result	Yes	Crisp flip, CMR/PFR
Qwen cross-fam. catalog	Qwen	Cross-family	Yes	Regional shift, CMR/PFR
70B boundary packet	70B	Boundary check	Unverified	Boundary only, TUNE
Ext. baseline packet	—	Aggregate-only	No	Aggregate comparator

5 Results

We evaluate defensive policy catalogs across both the canonical and poison-weighted contracts (Table 3). As shown in Figure 1, policy rankings change materially. Across the 8B and Qwen catalogs, changing the utility contract changes the selected policy, showing that single-score RAG defense evaluation can choose a deployment-misaligned operating point. On the Qwen cross-family catalog, the aggressive policies form a regional or categorical shift rather than a clean fine-grained inversion: a low-abstention strategy is preferred canonically, whereas an aggressive high-abstention region is preferred under the poison-weighted contract.

The absence of a universal optimal policy is also visible in the Pareto frontiers in Figure 2, which make the abstention-performance trade-off explicit. Calibration within the evaluated StressRAG policy catalog is therefore contract-sensitive optimization rather than a search for one invariant operating point.

The NLI diagnostic provides a mechanistic consistency check for the observed policy shift (Figure 3). We compute the diagnostic with `cross-encoder/nli-deberta-v3-base` over 691 slices and 203,613 examples. The diagnostic is correlational rather than causal: it measures contradiction density among retrieved sources and should not be read as a ground-truth reliability label. On HotpotQA, poisoning yields higher contradiction density than distractors, which is consistent with the shift toward aggressive filtering policies under poisoning-heavy contracts.

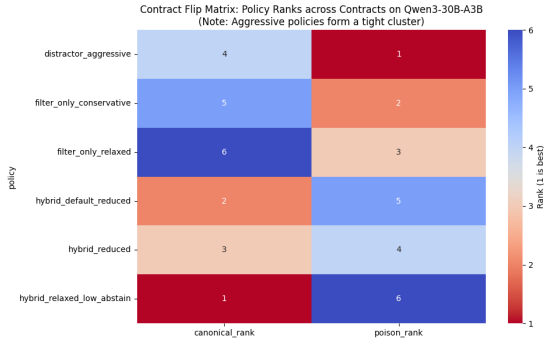

Figure 1: Policy rankings for Qwen3-30B-A3B under the canonical and poison-weighted proxy contracts (Eqs. 1–2). Under the canonical contract, the low-abstention `Hyb-Low` policy is ranked first. Under the poison-weighted contract, the preferred region shifts to aggressive high-abstention policies, with `Dist-Agg` ranked first. Differences among aggressive Qwen policies are best read as a regional shift rather than a clean fine-grained inversion.

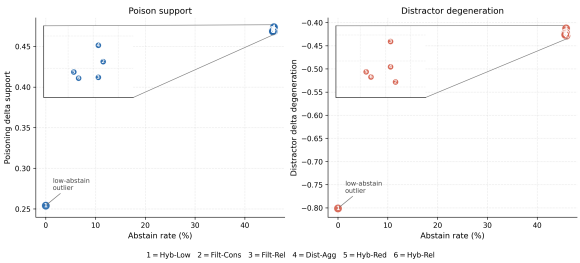

Figure 2: Pareto frontiers showing the abstention–performance trade-off under stress. Main panels show all policies; inset panels zoom into the high-abstention cluster. The low-abstention policy occupies a distinct operating region, while the remaining policies form a compact cluster. Full policy names are listed in Table 2. This reinforces that calibration should be reported contract-sensitively rather than as one invariant optimum.

Distractor evasion behavior remains dataset-dependent, including higher baseline contradiction density on Natural Questions.

The 70B boundary packet sharpens the scope of the claim: contract sensitivity is not automatically scale-invariant under the tested calibration. It yields no confirm-ready policy under either evaluated stress condition: both runs return a TUNE verdict, indicating that calibration parameters require adjustment before deployment-scale use. The packet does not reproduce the full 8B flip pattern. Global abstention across all 32 evaluated runs is 19.5%; under the poisoning-high stress slice specifically it rises to 27.1%. Its role is to demonstrate that contract-sensitive behavior should not be treated as

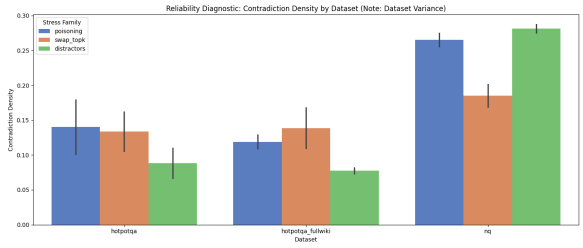

Figure 3: NLI source-consensus contradiction density across stress families. Poisoning introduces higher contradiction density than distractors on HotpotQA, providing mechanistic support for the shift toward aggressive filtering. Distractor evasion behavior remains dataset-dependent. *(Diagnostic details are provided in §5).*

Model	Canonical winner	Poison proxy winner	CMR_{proxy} C→P	CMR_{proxy} P→C	PFR
8B	**`Hyb-Low`**	**`Filt-Cons`**	0.751	0.471	0.733
Qwen	**`Hyb-Low`**	**`Dist-Agg`**	1.133	1.789	0.800

Table 4: Offline unified-proxy contract-shift metrics (Eqs. 3, 1–2) computed from retained artifacts only. CMR_{proxy} measures contract-mismatch proxy utility loss; PFR measures the fraction of policy pairs whose ordering reverses between contracts. 8B: 11/15 pairs flipped. Qwen: 12/15 pairs flipped. These proxy diagnostics should be interpreted alongside the matched-budget selector result, not as direct per-query rankings. Full policy names are provided in Table 2.

scale-invariant without further query-aligned evaluation.

To quantify contract-mismatch risk without rerunning any additional experiments, we report one unified proxy table across the 8B and Qwen summaries (Table 4). Under this shared view, the canonical winner remains `Hyb-Low` on both models, while the poison-proxy winner moves to `Filt-Cons` on 8B and `Dist-Agg` on Qwen. The table complements the matched-budget selector result and is not a direct per-query ranking matrix.

The aggregate-only external baseline packet adds useful aggregate comparators while preserving the paper's query-aligned evidence boundary. `PPL` completes all 32/32 runs with 24.0% abstention and 0.17× runtime and is the current aggregate-only overall comparator; `L2` completes all 32/32 runs with 24.3% abstention and 0.15× runtime and is the aggregate-only poisoning comparator. Both remain calibration-limited rather than deployment-ready: both received a TUNE verdict, meaning calibration parameters require adjustment before deployment-scale use. Both are reported as aggregate comparators because their query identifiers do not align with the internal policy catalog

(Table 3), so they cannot be merged into per-query rankings.

6 Supplementary Material and Claim Boundaries

The supplementary bundle maps each headline claim to reviewer-facing evidence: a claim ledger, artifact-coverage manifest, reproducibility note, privacy note, contract-shift tables, compact 70B boundary summary, and frozen benchmark reports. It records the main evidence boundaries: the `artifact/FREEZE/` packet covers `480` combinations with `6` documented missing executions, while the external-baseline packet remains aggregate-only. Statistical significance is assessed by bootstrap resampling (1,000 resamples, $\alpha = 0.05$, unit: policy cell).

7 Conclusion

The central lesson is that RAG defense evaluation should report the contract under which a policy is selected. In StressRAG, changing that contract changes the recommended operating point, and the strength and form of that change depend on model family, scale, and artifact alignment. This makes utility contracts a necessary reporting dimension for retrieval-stress evaluation, not an optional post-hoc analysis. StressRAG is therefore intended as an evaluation protocol, not as a new defense algorithm.

8 Limitations

Several limits are central to the paper's interpretation. First, the 70B packet is scale-boundary evidence rather than same-family confirmation: it yields a TUNE verdict under both evaluated stress conditions and does not reproduce the crisp 8B flip. Its 19.5% global abstention rate (27.1% under poisoning-high) also indicates a different operating regime from the low-abstention 8B winner.

Second, the Qwen cross-family catalog reveals a regional shift rather than the fine-grained inversion observed on Llama-3-8B: the preferred region changes, but differences within the aggressive-policy cluster should not be over-interpreted. Third, the NLI reliability signal is correlational, not causal; it helps interpret contradiction density but is not a ground-truth faithfulness label. The $\text{CMR}_{\text{proxy}}$/PFR table is likewise a shared proxy comparison and should be read alongside the exact matched-budget selector result. Finally, the six documented missing cells (all Qwen/HotpotQA/BM25/seed-2, medium and high stress) limit Qwen cross-family interpretation on HotpotQA BM25 but do not overturn the headline contract-sensitivity claim, which is supported by completed cells.

References

Chen Amiraz, Florin Cuconasu, Simone Filice, and Zohar Karnin. 2025. The distracting effect: Understanding irrelevant passages in RAG. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Shahul Es, Jithin James, Luis Espinosa-Anke, and Steven Schockaert. 2023. Ragas: Automated evaluation of retrieval augmented generation. *arXiv preprint arXiv:2309.15217*.

Karan Goel, Nazneen Fatema Rajani, Jesse Vig, Zachary Tasawwar, Mohit Bansal, and Christopher Re. 2021. Robustness gym: Unifying the NLP evaluation landscape. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations*. Association for Computational Linguistics.

Jeongyeon Hwang, Junyoung Park, Hyejin Park, Dongwoo Kim, Sangdon Park, and Jungseul Ok. 2025. Retrieval-augmented generation with estimation of source reliability. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 34279–34303.

Amita Kamath, Robin Jia, and Percy Liang. 2020. Selective question answering under domain shift. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 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.

Kuan Li, Liwen Zhang, Yong Jiang, Pengjun Xie, Fei Huang, Shuai Wang, and Minhao Cheng. 2025. LaRA: Benchmarking retrieval-augmented generation and long-context LLMs — no silver bullet for LC or RAG routing. In *Proceedings of the 42nd International Conference on Machine Learning*, pages 36846–36867.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81.

Chanhee Park, Hyeonseok Moon, Chanjun Park, and Heuiseok Lim. 2025. MIRAGE: A metric-intensive benchmark for retrieval-augmented generation evaluation. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 2883–2900.

Shuting Wang, Jiejun Tan, Zhicheng Dou, and Ji-Rong Wen. 2025. OmniEval: An omnidirectional and automatic RAG evaluation benchmark in financial domain. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*.

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

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*.

Wei Zou, Runpeng Geng, Binghui Wang, and Jinyuan Jia. 2024. Poisonedrag: Knowledge corruption attacks to retrieval-augmented generation of large language models. *arXiv preprint arXiv:2402.07867*.