

Guardrailed Retrieval and Abstention in the Last Z Librarian: An Exploratory Paired Evaluation

Pedro Alberto Laris Uribe^{1,2} and Gilda Dolores Gamiño Cereceres¹

¹Doctoral Program in Technology and Innovation Management, Universidad UTEL, Mexico

²Last Z Community Platform, independent software and community research project

12 August 2026

Manuscript classification. Exploratory empirical pilot with a legacy machine-assisted claim audit. The paired executions are complete, but the claim labels are not blinded human adjudications and the exact request-time evidence state is unavailable. The results support transparent pilot conclusions, not a confirmatory claim that hallucinations were eliminated.

Abstract

Background. Retrieval and abstention can reduce unsupported language-model output, but apparent gains depend on coverage, generated claim volume, and evaluator validity. We evaluated the Last Z Librarian, a server-aware community assistant for rules, events, and strategy.

Methods. Ninety-one eligible production requests were replayed once under condition A (unguarded, no retrieval) and condition D (guardrailed retrieval and abstention), both using gpt-4.1-mini. A legacy gpt-4.1-mini auditor extracted and labeled claims against a current-corpus lexical proxy. The main exploratory endpoint was the paired request-level presence of at least one machine-audited unsupported claim. Pooled unsupported-claim rate (UCR), claim counts, abstention, tokens, and latency were secondary.

Results. Unsupported content was detected in 24/91 A responses (26.4%) and 19/91 D responses (20.9%); paired risk difference -5.5 percentage points (95% bootstrap interval -17.6 to 5.5; exact McNemar $p = 0.458$). Pooled UCR fell from 63/118 (53.4%) to 68/449 (15.1%), but D contained more unsupported claims in absolute terms (68 versus 63), and mean unsupported claims per request did not differ ($p = 0.947$). Supported claims increased from 55 to 381. D abstained on 48/91 requests, while median total tokens rose from 306 to 17,271 and latency from 3.79 to 11.88 seconds. Quality control found evaluator errors, residual speculation, and a critical prompt-instruction disclosure.

Conclusions. The guardrailed bundle produced a lower unsupported fraction by generating substantially more supported content, but did not establish a conclusive reduction in request-level incidence or absolute unsupported claims. The data do not support 100% hallucination elimination. Human adjudication, request-time evidence, canonical coverage, and a prospective frozen evaluation are required for confirmatory claims.

Keywords: retrieval-augmented generation; factuality; unsupported claims; abstention; claim-level evaluation; LLM-as-a-judge; community governance; prompt injection; Last Z.

Contents

1	Introduction	3
2	Conceptual and Methodological Background	3
2.1	Claim-level factuality and attribution	3
2.2	Coverage and abstention	4
2.3	Automatic evaluation as a measurement instrument	4
2.4	Governance and security	4
3	Materials and Methods	5
3.1	Study design and classification	5
3.2	Production population and paired cohort	5
3.3	Experimental conditions	7
3.4	Legacy machine-assisted claim audit	7
3.5	Outcomes	7
3.6	Statistical analysis	8
4	Results	8
4.1	Paired request-level unsupported-claim endpoint	8
4.2	Pooled machine-audited unsupported-claim rate	9
4.3	Supported and assessable content	10
4.4	Instrumented abstention	11
4.5	Operational trade-off	12
4.6	Quality-control and residual failure analysis	13
5	Discussion	13
5.1	Principal interpretation	14
5.2	Why the evidence does not support 100% elimination	14
5.3	Abstention, answerability, and risk–coverage	14
5.4	Measurement validity	14
5.5	Operational and governance implications	15
5.6	Security as an independent quality dimension	15
6	Limitations	15
7	Required Confirmatory Redesign	16
8	Conclusions	16
A	Detailed Outcome Table	18
B	Claims Supported and Not Supported by Run 1	19
C	Minimum Human-Adjudication Data Model	19
D	Reproducibility Inventory	20

1 Introduction

Large language models can produce coherent answers while introducing details that are absent from, or contradicted by, an applicable evidence base. Retrieval-augmented generation (RAG) addresses part of this problem by conditioning generation on external documents rather than relying only on parametric knowledge (Y. Gao et al., 2024; Lewis et al., 2020). Retrieval does not guarantee that the selected evidence is authoritative, current, correctly scoped, or faithfully used. A response can therefore contain a mixture of supported claims, unsupported additions, contradictions, omissions, and citations that do not entail the associated proposition (T. Gao et al., 2023; Min et al., 2023; Rashkin et al., 2023; Song et al., 2024).

The applied case is the Librarian within the independent Last Z community platform. The public platform describes the Librarian as a server-aware assistant for rules, events, and strategy, while the platform’s legal framing identifies it as an independent community service rather than the game publisher (Last Z Community Platform, 2026a, 2026b). This context creates a mixed evidence environment: a server rule may be normative within that server; an event notice may expire; a game-mechanics statement may have uncertain provenance; and strategy advice may depend on assumptions rather than admit a single factual answer. The Server 496 rules page illustrates the importance of server-scoped and time-scoped governance evidence rather than generic game knowledge (Last Z Server 496 Community, 2026).

The project originally proposed a controlled RAG protocol. The final supplied bundle contains more than a protocol: it includes a complete paired replay under an unguarded comparator and a guardrailed retrieval bundle, plus a machine-assisted claim audit. It also contains material weaknesses that prevent a confirmatory interpretation. Most importantly, the evaluator itself is noisy; the pooled claim denominator changes sharply between conditions; answerability and canonical coverage are not independently established; and the exact request-time corpus state is not available. This paper therefore reports the empirical data as an exploratory pilot, exposes the measurement failures, and specifies the corrections needed for a stronger study.

The study addresses four questions:

1. Does the guardrailed retrieval-and-abstention bundle reduce the paired request-level probability of any unsupported claim relative to the unguarded condition?
2. How do pooled unsupported-claim rates, absolute unsupported counts, supported content, and assessable claim volume differ between conditions?
3. What operational burden is associated with the guardrailed condition in abstention, token use, response length, and latency?
4. Which evaluator, factuality, and security failure modes remain visible despite the guardrails?

2 Conceptual and Methodological Background

2.1 Claim-level factuality and attribution

Whole-response correctness labels conceal mixed evidence states. FActScore operationalizes factual precision through atomic claims, while VeriScore emphasizes that only reasonably verifiable propositions belong in the factuality denominator (Min et al., 2023; Song et al., 2024). Attribution is a distinct construct: the existence of a citation marker is not evidence that the cited source is localizable, applicable, and

sufficient (T. Gao et al., 2023; Rashkin et al., 2023). The present study uses the operational term *unsupported claim* rather than treating every documentary mismatch as an ontological falsehood. A claim may be true outside the supplied corpus yet unsupported under the system’s declared evidence boundary.

For condition c , the legacy pooled unsupported-claim rate is

$$\text{UCR}_c = \frac{U_c}{S_c + U_c}, \quad (1)$$

where U_c is the number of claims machine-labeled unsupported or contradicted and S_c is the number machine-labeled supported. The legacy script also emitted a small number of out-of-schema labels, which were excluded from this denominator. Because $S_c + U_c$ is generated after treatment, [equation \(1\)](#) is not a fixed-denominator causal estimand. A condition can lower pooled UCR by producing more supported claims without reducing the absolute number of unsupported claims.

2.2 Coverage and abstention

Precision without coverage can reward a system that omits difficult information. Conversely, a verbose system may add useful supported content while also creating more opportunities for error. Recent RAG evaluation work therefore emphasizes scenario-specific reference standards, joint retrieval-and-reasoning assessment, and subquestion or fact coverage (Krishna et al., 2025; Xie et al., 2025; Zhu et al., 2025). A confirmatory study needs a canonical fact set created before outputs are revealed.

Abstention must also be conditioned on answerability. Refusing an unanswerable request is favorable; refusing an answerable request is a false abstention; and stating that evidence is insufficient before continuing to speculate is a failed abstention. Instrumented abstention metadata alone cannot distinguish those cases.

2.3 Automatic evaluation as a measurement instrument

Automated RAG evaluators can scale analysis, but their validity depends on claim extraction, evidence retrieval, prompting, and calibration (Es et al., 2024; Saad-Falcon et al., 2024). LLM judges can show systematic biases and failure modes that are not exposed by aggregate agreement alone (Li et al., 2025; Muller et al., 2025). In this study, the machine audit is analyzed as a measurement instrument rather than accepted as a ground-truth adjudicator. Manual quality control of selected records is part of the empirical result because it directly affects the interpretation of the rates.

2.4 Governance and security

Knowledge governance determines which versioned material is retrievable, while technology governance determines who may change the model, prompt, corpus, or decision policy. The NIST AI Risk Management Framework separates governance, context mapping, measurement, and risk management, and its generative-AI profile explicitly treats confabulation, provenance, privacy, and security as distinct risks (Autio et al., 2024; Tabassi, 2023). Factuality guardrails do not by themselves prevent instruction disclosure or indirect prompt injection. Security events therefore require a separate endpoint and gate.

3 Materials and Methods

3.1 Study design and classification

Run 1 is a retrospective paired replay of real production requests. Each eligible Librarian request was executed once under condition A and once under condition D. The same request therefore acts as its own control. The comparison is exploratory because it was not preregistered, the outcome labels were generated after the run by a single machine auditor, only one stochastic execution was retained per condition, and the exact request-time evidence state was not preserved in the supplied bundle.

The design is not a clean component ablation. Condition D is a bundle intervention combining retrieval, a large evidence context, guardrail instructions, and an abstention decision policy. The supplied export does not isolate retrieval, reranking, claim verification, or policy components. Any estimated A–D difference therefore applies to the realized bundle as a whole.

3.2 Production population and paired cohort

The aggregate snapshot covers 4 June through 6 August 2026. It contains 110 user requests across three modes: 103 Librarian, six advisor, and one judge. The supplied crosstab identifies exactly 91 eligible Librarian requests and 12 excluded Librarian requests under the operational rule of at least 20 characters. The 91-request cohort spans 39 conversations, 13 server identifiers, and 18 active request dates. Because multiple requests can occur within one conversation, request-level bootstrap intervals may still understate conversation-level dependence.

The broader production corpus snapshot contains 42 Markdown files, of which 41 were marked ready, with 2,926 fragments and 3,890 sentences. The platform snapshot reports 99/116 ready indexes. These figures establish scale and instrumentation; they do not establish that the exact evidence needed for every request was available or current at the request time.

Table 1. Empirical data flow and analysis population

Stage	Count	Interpretation
All production user requests	110	Librarian, advisor, and judge modes combined.
Librarian universe	103	Domain-relevant mode for this paper.
Eligible Librarian requests	91	Operational text-length rule; complete paired A/D cohort.
Condition A executions	91	Unguarded, no-retrieval responses.
Condition D executions	91	Guardrailed retrieval-and-abstention responses.
Historical G responses	89	Descriptive only; two eligible requests lacked a historical response.
Machine-audited A claims	166	Includes not-assessable and out-of-schema labels.
Machine-audited D claims	504	Includes not-assessable labels.
Assessable A/D claims	118 / 449	Supported plus unsupported/contradicted labels used in pooled UCR.

Raw interaction text is restricted. Public data use deterministic hashed case identifiers and aggregate outcomes.

3.3 Experimental conditions

Table 2. Realized Run 1 conditions

Arm	Function	Model	Observed policy metadata	Interpretive boundary
A	Unguarded parametric response without retrieval	gpt-4.1-mini	91 answers; 0 instrumented abstentions	Comparator for the entire D bundle.
D	Guardrailed retrieval and abstention	gpt-4.1-mini	43 answers; 48 instrumented abstentions	Retrieval, context, instructions, and decision policy are not separately identifiable.
G	Historical production response	Three model variants	82 answers, four abstentions, three scope refusals among 89 responses	Descriptive only; heterogeneous models and no comparable latency.

Each A/D request was executed once. The export records total tokens and end-to-end latency for A and D.

3.4 Legacy machine-assisted claim audit

The supplied audit script evaluated A and D with one gpt-4.1-mini call per request-condition output at temperature zero. The call jointly inferred answerability and abstention, segmented claims, assigned support labels, assigned citation labels, selected an evidence filename, and produced a rationale. The literal condition label was included in the evaluator prompt, so the audit was not condition-blind.

Evidence was not retrieved from the production vector store. The script loaded a current local Markdown directory, tokenized the user request, counted lexical occurrences across entire documents, selected the top six documents, and supplied a 900-character window around an early token match. The same request-level evidence set was reused for every claim in both conditions. The corpus contents and exact run-time retrieval outputs are not present in the final supplied bundle; only filenames and hashes are available.

The declared support schema was {supported, contradicted, unsupported, not assessable}. The auditor emitted two additional `partial` support labels under A; these were excluded from the assessable denominator rather than silently recoded. Citation outcomes were not analyzed because the auditor assigned non-absent citation labels to A despite A having no citation mechanism.

3.5 Outcomes

The principal exploratory safety outcome was a request-level binary indicator for at least one machine-audited unsupported or contradicted claim. It is preferable to pooled UCR because it fixes the request denominator and supports a paired comparison. It remains limited by the machine labels and by 51 A requests versus 18 D requests with zero assessable claims.

Secondary outcomes were:

- pooled machine-audited UCR and relative reduction;
- supported, unsupported, and assessable claims per request;
- instrumented abstention from execution metadata;
- total recorded tokens, response characters, and latency;
- qualitative evaluator, factuality, and security failure modes.

Correct abstention, false abstention, canonical fact coverage, and citation accuracy were not validly estimable from the supplied data and are not reported as efficacy endpoints.

3.6 Statistical analysis

For the paired request-level endpoint, the 2-by-2 discordance table was analyzed with exact McNemar testing. The paired risk difference, $p_D - p_A$, received a percentile bootstrap interval from 100,000 resamples of requests with seed 20260812. Exact Clopper–Pearson intervals describe the separate condition rates.

Pooled UCR is reported descriptively. To preserve within-request clustering, 100,000 bootstrap samples resampled requests and recomputed the aggregated numerator and denominator. The resulting interval does not solve the post-treatment denominator problem; it quantifies sampling variability conditional on the legacy labels.

Claim counts, total tokens, response characters, and latency were compared through paired differences and Wilcoxon signed-rank tests. Bootstrap intervals were calculated for mean and median paired differences. Because the study is exploratory and the analyses were selected after inspection of the supplied bundle, p-values are descriptive and were not multiplicity-adjusted.

All calculations were performed by the included Python analysis script using pandas, NumPy, SciPy, and statsmodels. Public outputs contain no raw request or response text.

4 Results

4.1 Paired request-level unsupported-claim endpoint

The machine audit identified at least one unsupported claim in 24/91 A outputs (26.4%; exact 95% CI 17.7–36.7%) and 19/91 D outputs (20.9%; exact 95% CI 13.1–30.7%). The paired contingency table contained 55 requests with no unsupported claim in either condition, 12 with an unsupported claim only under D, 17 only under A, and seven under both conditions (table 3).

Table 3. Paired request-level machine-audited unsupported-claim outcome

Condition A	Condition D		Total
	No unsupported claim	At least one	
No unsupported claim	55	12	67
At least one	17	7	24
Total	72	19	91

The paired risk difference was -5.5 percentage points, with a request-bootstrap 95% interval from -17.6 to 5.5 percentage points. Exact McNemar $p = 0.458$. The direction favors D, but the interval includes no difference and modest harm. Under the legacy labels, Run 1 therefore does not establish a statistically conclusive reduction in request-level incidence.

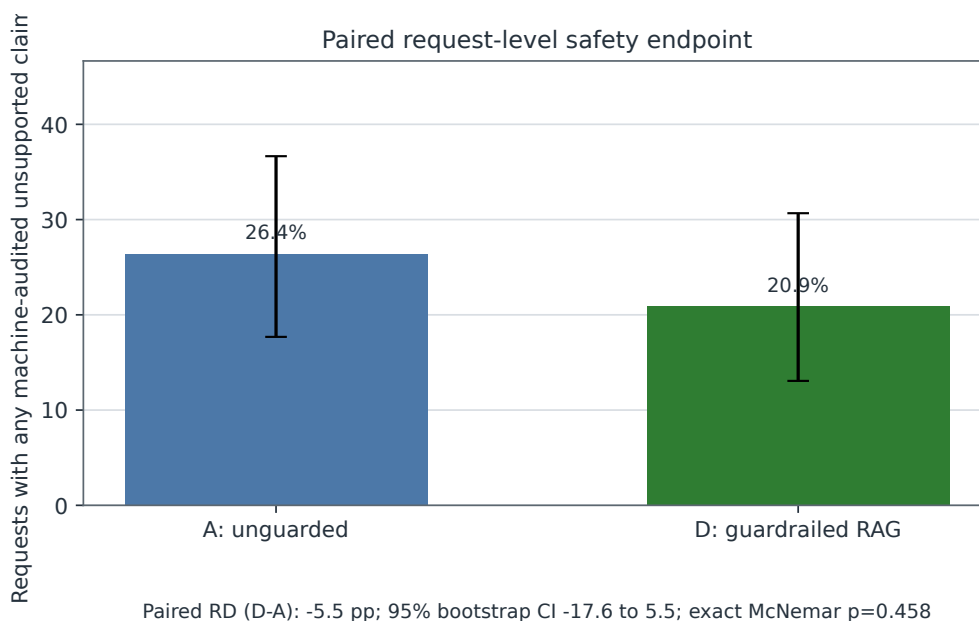


Figure 1. Paired request-level safety endpoint. Error bars show separate exact binomial intervals; the inferential contrast is the paired risk difference shown below the chart. Labels are machine-assisted.

4.2 Pooled machine-audited unsupported-claim rate

Condition A contained 63 unsupported or contradicted labels among 118 assessable claims (53.4%). Condition D contained 68 among 449 (15.1%). The descriptive risk difference was -38.2 percentage points, with request-clustered bootstrap interval -54.0 to -22.3 percentage points. The observed relative reduction was 71.6% (clustered bootstrap interval 53.5–84.6%).

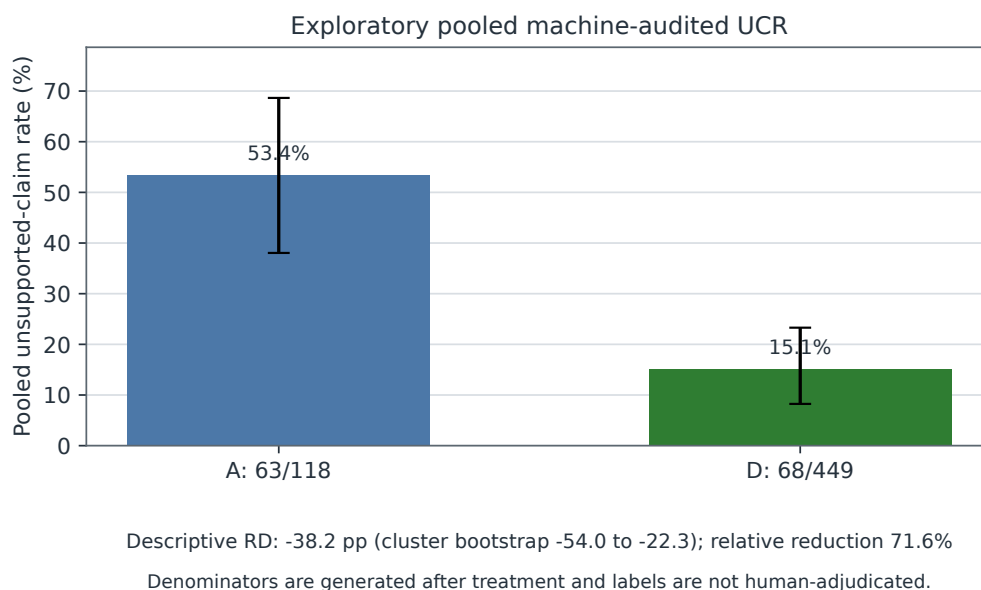


Figure 2. Exploratory pooled unsupported-claim rate. The generated claim denominator is post-treatment, and the labels are not human-adjudicated.

The lower pooled fraction must not be interpreted as elimination of unsupported content. D had 68 unsupported labels, five more than A. Mean unsupported claims per request were 0.75 under D and 0.69 under A; the paired mean difference was 0.05 (bootstrap interval -0.38 to 0.51), and Wilcoxon $p = 0.947$. The pooled rate fell because D generated substantially more assessable and supported content.

4.3 Supported and assessable content

Condition D produced 381 machine-labeled supported claims, versus 55 under A. Mean supported claims per request rose from 0.60 to 4.19, with a paired mean increase of 3.58 (bootstrap interval 2.76–4.46; Wilcoxon $p < 0.001$). Mean assessable claims rose from 1.30 to 4.93 (paired mean increase 3.64; bootstrap interval 2.79–4.51; $p < 0.001$).

Fifty-one A requests had zero assessable claims, compared with 18 D requests. This difference reflects evaluator behavior as well as response content. In particular, the legacy auditor often treated A answers as abstentions despite the absence of an execution-level abstention flag. The claim-volume result therefore supports the interpretation that D generated more auditable content, but it does not establish canonical coverage.

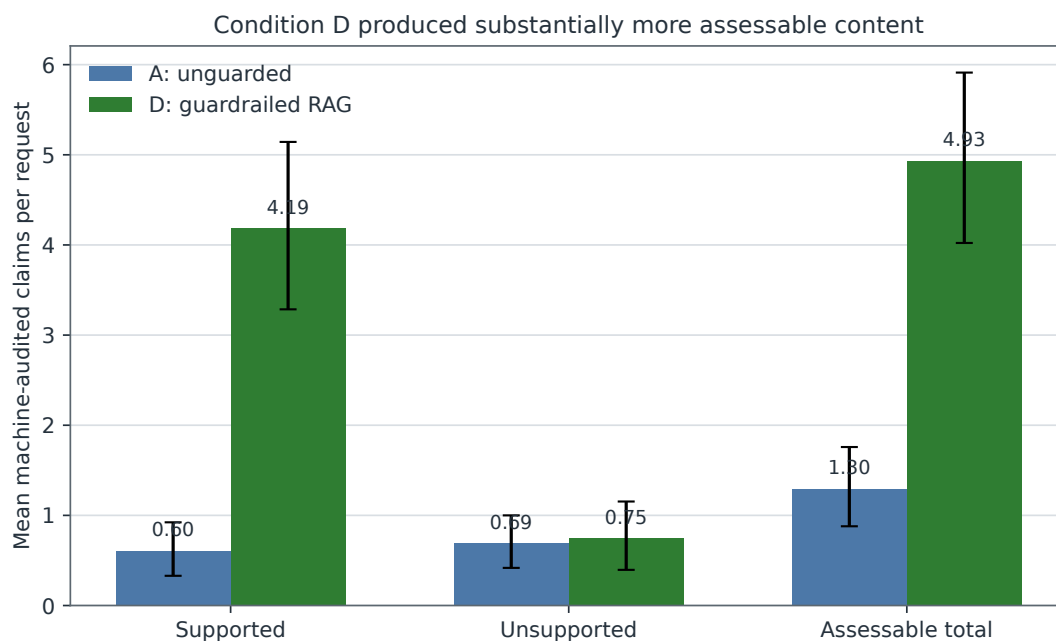
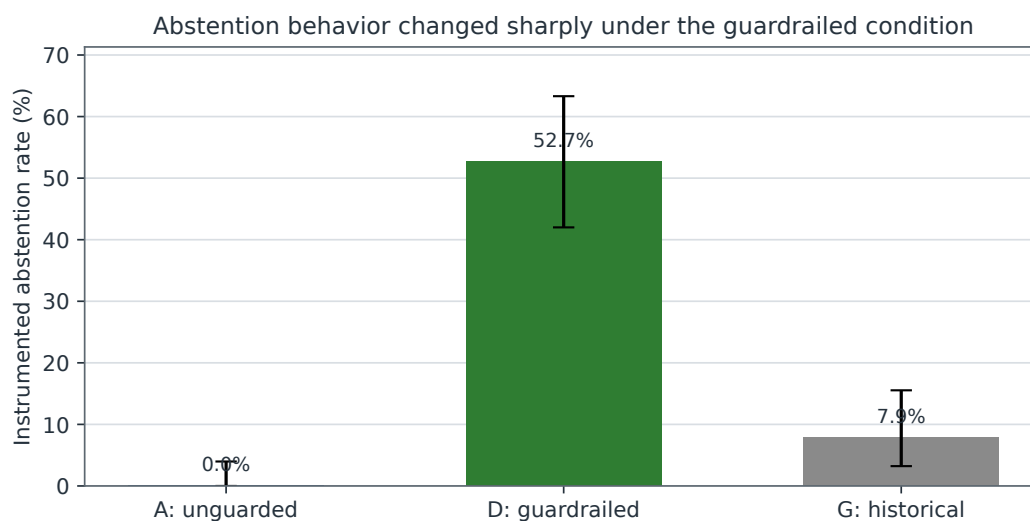


Figure 3. Machine-audited claim counts per request. Error bars are request-bootstrap intervals for the mean.

4.4 Instrumented abstention

Execution metadata recorded no abstentions under A and 48/91 under D (52.7%; exact 95% CI 42.0–63.3%). The historical production arm G recorded seven abstentions among 89 available responses (7.9%). G is not directly comparable because it used heterogeneous models, and the live policy included scope-refusal decisions.

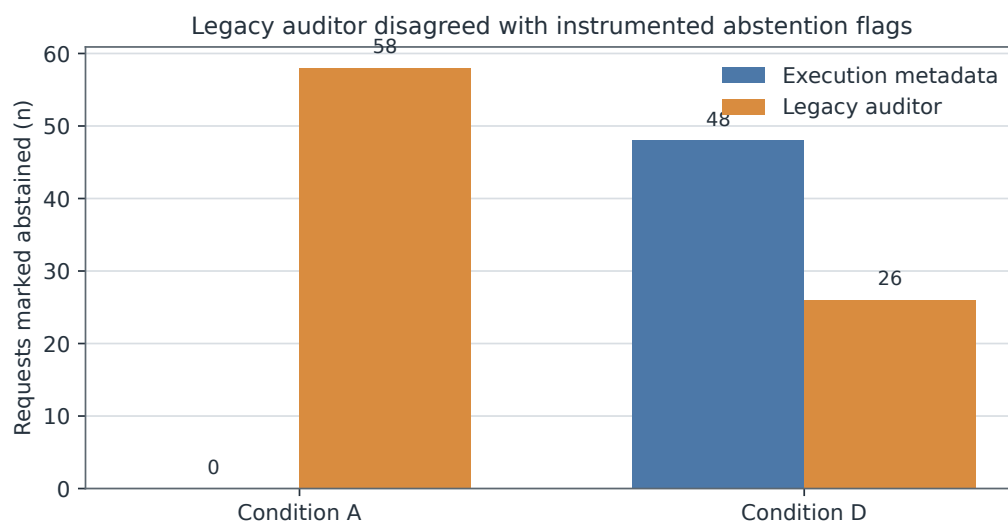


Answerability and false-abstention status were not independently adjudicated.

Figure 4. Instrumented abstention by condition. Correctness cannot be inferred without an independent request answerability standard.

The auditor's own abstention labels were inconsistent with execution metadata. It marked 58 A requests as abstentions even though A recorded zero and marked only 26 D requests as abstentions despite 48 execution flags. Under D, the resulting Cohen kappa was 0.27; under A it was effectively zero. The

disagreement invalidates the auditor-derived correct/false-abstention counts in the supplied outcome table.



D agreement: Cohen's kappa 0.27; A kappa 0.00.

Figure 5. Disagreement between execution metadata and the legacy auditor's inferred abstention labels.

4.5 Operational trade-off

Median total recorded tokens were 306 under A and 17,271 under D. The paired median increase was 16,939 tokens (bootstrap interval 16,860–17,037; Wilcoxon $p < 0.001$). Median response length increased from 1,105 to 2,209 characters. Median latency increased from 3.79 to 11.88 seconds, a paired median increase of 7.05 seconds (bootstrap interval 6.40–8.47; $p < 0.001$).



Both paired differences were significant by Wilcoxon signed-rank test ($p < 0.001$).

Figure 6. Operational burden of the realized guardrailed retrieval bundle. Tokens are the total recorded by the experiment export.

The approximately 56-fold token median and threefold latency median are material deployment costs. No monetary price table or service-level objective was frozen in the supplied files, so operational

noninferiority cannot be claimed.

4.6 Quality-control and residual failure analysis

Manual review of targeted records identified defects in both the evaluator and the guardrailed system. To protect user privacy, the public package records paraphrased findings under deterministic hashed case identifiers rather than raw interaction text.

Table 4. Material quality-control findings

Component	Failure mode	Observed consequence	Interpretive impact
Legacy auditor	Polarity inversion	A negative or evidence-absence statement was reconstructed as a positive claim and labeled unsupported.	False-positive unsupported labels.
Legacy auditor	User-premise attribution	A premise supplied by the user and described as unverified was scored as an assistant assertion.	Construct contamination.
Legacy auditor	Retrieval miss	A server rule present in an applicable public rules source was absent from the lexical evidence packet.	UCR partly measures audit retrieval.
Guardrailed system	Evidence-gap leakage	The answer acknowledged missing event evidence and then inferred timing and duration from other events.	Failed abstention; unsupported speculation.
Guardrailed system	Premise endorsement	User-supplied event details were confirmed despite explicit acknowledgement that the corpus lacked them.	Unsupported endorsement.
Guardrailed system	Prompt-instruction disclosure	A direct prompt-injection request elicited system-instruction text.	Critical security failure outside factuality control.

These are purposively selected quality-control findings, not prevalence estimates.

The prompt-disclosure event is particularly important. It demonstrates that evidence-grounding and abstention guardrails cannot substitute for prompt-injection defenses, privilege separation, tool allowlists, and explicit confidentiality controls.

5 Discussion

5.1 Principal interpretation

Run 1 supports a narrower conclusion than the original 100% reduction claim. Under the legacy machine audit, condition D had a much lower pooled unsupported fraction than A and generated far more supported content. The pooled result is consistent with a system that uses retrieved context to produce richer, better-supported answers. It is not evidence that unsupported content disappeared. D retained 68 unsupported labels, the absolute unsupported count per request did not improve, and the paired request-level endpoint was inconclusive.

The distinction is structural. A generated only 118 assessable claims across 91 requests, whereas D generated 449. Dividing by these unequal post-treatment denominators makes D appear substantially safer because the supported denominator expanded. That expansion is valuable, but it is a different effect from reducing the probability or number of unsupported claims. A complete publication must report both.

5.2 Why the evidence does not support 100% elimination

A 100% observed relative reduction requires a positive comparator rate and zero unsupported events in the guarded condition under a valid measurement process. Neither requirement is satisfied here. The legacy audit labels 68 D claims as unsupported, and 19 D requests contain at least one such label. Manual review suggests that some are false positives, but it also identifies genuine residual failures in which the system speculated after declaring an evidence gap. Removing known audit errors would require a complete, condition-blinded adjudication rather than selective correction of unfavorable records.

Even a future human audit with zero observed events would establish a zero rate only in the audited sample, not a guarantee of zero future risk. The claim should then be stated as “zero observed unsupported claims” with an appropriate confidence bound and explicit coverage and abstention results.

5.3 Abstention, answerability, and risk-coverage

Condition D abstained on more than half of requests. This may represent appropriate discipline, over-abstention, or a mixture. The supplied auditor cannot decide because its abstention labels disagree with the execution metadata and because answerability was inferred after seeing the response rather than established independently. A rigorous design must create a request-time evidence standard before revealing outputs, then classify correct abstention, false abstention, incomplete answer, and failed abstention separately.

The qualitative evidence-gap failures also show that an abstention flag is not enough. If the response continues with unsupported conclusions after the flag is set, the policy has not actually prevented hallucination. The deployment implementation should enforce a terminal response template or a verified partial-answer boundary when evidence is insufficient.

5.4 Measurement validity

The legacy machine audit combines four error-prone tasks in one call: claim extraction, stance reconstruction, evidence retrieval evaluation, and support judgment. Condition disclosure creates potential evaluator bias. Reusing request-level lexical passages across claims reduces recall for claim-specific evidence. Current-corpus retrieval introduces temporal and version ambiguity. These defects are not merely theoretical; the manual review found examples of each.

The strongest correction is not a more elaborate single judge prompt. It is a staged process: independent answerability and canonical-fact curation; condition-blinded claim segmentation; per-claim evidence retrieval with exact locators; duplicate human labels; adjudication; and only then calibration of any automated evaluator. The rebuilt bundle includes a protocol and queue-generation script for that process.

5.5 Operational and governance implications

The guardrailed bundle substantially increased token use and latency. Whether that cost is acceptable depends on a prespecified service-level objective, monetary budget, and the value of improved supported coverage. The present files contain no locked thresholds, so the operational result should be treated as a trade-off rather than success or failure.

The real Last Z case also requires source governance. Server rules, event announcements, adjudication records, game-mechanics notes, and strategy advice have different authority and validity periods. A claim can be grounded in a document yet still be wrong for the request because the document belongs to another server, season, or time. Exact source class, server scope, effective dates, supersession, and evidence span must be part of the analytical record.

5.6 Security as an independent quality dimension

The prompt-instruction disclosure prevents an unqualified deployment-safety conclusion. It was not captured by the unsupported-claim rate, yet it is a critical system failure. Future release gates should separately require factuality, coverage, calibrated abstention, privacy, and prompt-injection resistance. A system may pass one dimension while failing another.

6 Limitations

The study has substantial limitations.

First, the claim audit is machine-assisted and not human-adjudicated. The evaluator was not blinded, used one call, and exhibited demonstrated classification failures. Second, the audit evidence was a current-corpus lexical proxy rather than the exact request-time retrieval state. The corpus text itself was not included in the final supplied bundle, preventing independent evidence re-evaluation. Third, pooled UCR uses a post-treatment claim denominator, and the conditions differ greatly in response length and claim volume. Fourth, canonical fact coverage and answerability were not established before outputs were inspected. Fifth, each request-condition pair has only one generation, so stochastic variability is unmeasured. Sixth, the run was not preregistered and analyses are exploratory. Seventh, 91 requests arise from 39 conversations; request bootstrap does not fully represent conversation clustering. Eighth, condition D is a bundle, so the individual contribution of retrieval, reranking, verification, or policy cannot be isolated. Ninth, historical G is incomplete, model-heterogeneous, and lacks comparable latency. Tenth, the 20-character eligibility rule is operational rather than substantively validated. Eleventh, raw interactions may contain user-derived information, and no institutional ethics determination was included with the bundle. Finally, the study covers one community platform and cannot be generalized automatically to other games, enterprises, languages, or RAG systems.

7 Required Confirmatory Redesign

The pilot is useful precisely because it identifies what must be frozen before a stronger study.

1. **Prospective untouched cohort.** Collect new Librarian requests after the system, prompts, corpus, and analysis plan are locked. Do not reuse Run 1 as the final efficacy test after it has informed the design.
2. **Request-time evidence.** Preserve the exact corpus manifest, retrieval results, source versions, server scope, and conversation context available at each request time.
3. **Fixed reference standard.** Curate answerability, required canonical facts, acceptable sources, and expected action before condition outputs are revealed.
4. **Condition-blinded adjudication.** Use at least two independent human reviewers for all primary claims, with adjudication and class-specific reliability reporting.
5. **Primary request-level endpoint.** Prespecify the paired risk difference for any human-adjudicated unsupported claim. Use exact McNemar and request/conversation-cluster intervals.
6. **Coverage gate.** Require canonical fact coverage to meet a prespecified noninferiority margin so that safety cannot be achieved through omission or indiscriminate abstention.
7. **Abstention endpoints.** Report correct abstention, false abstention, failed abstention with residual speculation, and a risk–coverage curve.
8. **Component ablation.** If the scientific question concerns verification specifically, separate no-retrieval, retrieval, retrieval-plus-reranking, and retrieval-plus-reranking-plus-verification/policy conditions.
9. **Repeated runs.** Either fix a deterministic generation configuration or collect repeated runs and model run-level variation.
10. **Operational thresholds.** Freeze latency, token, cost, and critical-error gates before observing comparative results.
11. **Security evaluation.** Add a separate adversarial prompt-injection and confidentiality test suite with zero-tolerance critical disclosure criteria.
12. **Ethics and privacy determination.** Obtain and report an institutional determination for secondary use of de-identified interaction logs, and keep public and restricted data packages separate.

8 Conclusions

The final bundle contains real empirical evidence, but it does not show that the Last Z Librarian eliminated hallucinations. The strongest supported result is that the realized guardrailed retrieval-and-abstention bundle reduced the pooled proportion of machine-audited unsupported claims from 53.4% to 15.1% while greatly increasing supported content. That rate reduction did not correspond to fewer unsupported claims in absolute terms, and the paired request-level reduction was small and statistically inconclusive. The audit itself showed material measurement error, and the system retained factuality and security failures.

Accordingly, Run 1 should be published, if at all, as an exploratory pilot with full measurement caveats. Its primary scientific contribution is not proof of perfect factuality. It is a transparent demonstration that RAG safety claims depend on fixed denominators, coverage, calibrated abstention, evaluator validity, temporal evidence integrity, and security controls. The rebuilt protocol converts those lessons into an auditable path toward a confirmatory study.

Declarations

Author identification and contributions

Pedro Alberto Laris Uribe and Gilda Dolores Gamiño Cereceres are the authors of this manuscript. Role-specific CRediT allocations were not established by the supplied research files and should be jointly confirmed before external journal submission. Both authors are responsible for approving the final submitted version.

Competing interests

Pedro Alberto Laris Uribe is the developer and operator of the evaluated Last Z community platform. This creates an investigator–developer interest that should be mitigated through preregistration, independent evidence curation, blinded annotation, immutable manifests, and complete reporting of adverse results. No additional competing-interest information was supplied.

Ethics and privacy

The study uses interaction records originating from real users of a community platform. The public package excludes raw request and response text and uses aggregate or pseudonymous outcomes. No formal institutional ethics, exemption, or non-human-subject determination was included in the supplied files. Such a determination should be obtained and reported before journal submission involving user-derived interaction data.

Funding

No funding information was supplied.

Data and code availability

The public reproducibility package contains the manuscript, analysis code, aggregate snapshot, pseudonymous request-level outcomes, derived statistics, figures, protocols, templates, and file hashes. Raw interaction text, original identifiers, and legacy claim records are restricted to the private package. The exact corpus contents and request-time retrieval states were not present in the supplied bundle and therefore cannot be redistributed or independently reconstructed from this package alone.

Generative AI assistance

Generative AI was used under author direction for language editing, code review, and artifact preparation. The authors remain responsible for verification, interpretation, authorship, and all claims in any submitted

version.

A Detailed Outcome Table

Table 5. Run 1 empirical outcomes and interpretation

Outcome	A	D	Interpretation
Paired requests	91	91	Complete paired comparison.
Requests with any unsupported label	24	19	Paired RD -5.5 pp; interval includes zero.
Assessable factual claims	118	449	Generated denominator differs sharply.
Unsupported/contradicted labels	63	68	Absolute count did not fall.
Pooled UCR	53.4%	15.1%	Descriptive rate reduction; post-treatment denominator.
Supported labels	55	381	Large increase in machine-labeled supported content.
Mean unsupported claims/request	0.69	0.75	No paired difference; $p = 0.947$.
Mean supported claims/request	0.60	4.19	Paired increase; $p < 0.001$.
Requests with zero assessable claims	51	18	Auditor/response behavior differs.
Instrumented abstentions	0	48	Correctness not independently adjudicated.
Median total tokens	306	17,271	Material operational burden.
Median latency, seconds	3.79	11.88	Material operational burden.
Median response characters	1,105	2,209	D responses were longer.

B Claims Supported and Not Supported by Run 1

Table 6. Publication wording boundary

Supported with explicit caveats	Not supported
A complete 91-request paired A/D replay exists.	Guardrails eliminated 100% of hallucinations.
The pooled machine-audited UCR was lower under D.	The true future unsupported-claim risk is zero.
D produced more machine-labeled supported content.	D reduced the absolute number of unsupported claims.
The request-level point estimate favored D but was inconclusive.	Verification itself caused the observed difference.
D invoked abstention frequently.	All D abstentions were correct or useful.
D required substantially more tokens and latency.	D met an operational noninferiority threshold.
The legacy evaluator has documented failure modes.	The machine labels are a human gold standard.
A critical prompt disclosure occurred in targeted review.	Factuality guardrails provide adequate security.

C Minimum Human-Adjudication Data Model

Table 7. Required fields for confirmatory adjudication

Field	Purpose	Rule
request_item_id	Blinded request identifier	Must not reveal condition or original user identity.
server_id / timestamp	Scope and temporal applicability	Required to select the correct evidence version.
answerability	Expected response mode	Fixed before outputs are revealed.
canonical_fact_id	Fixed coverage denominator	One row per required fact or action.
claim_text	Assistant proposition	Preserve polarity, modality, and attribution.
assistant_endorsed	Exclude quoted user premises	Mark whether the assistant asserted or confirmed it.
support_label	Factuality outcome	Supported, contradicted, unsupported, insufficient, or not assessable.
evidence_locator / quote	Auditability	Exact source, version, section, and minimal supporting span.

Field	Purpose	Rule
citation_label	Attribution quality	Precise, partial, irrelevant, non-localizable, absent, or not applicable.
abstention_correct	Calibration	Determined against pre-output answerability.
severity	Risk weighting	Low, medium, high, or critical.
reviewer / adjudication	Reliability	Preserve independent labels and final resolution.

D Reproducibility Inventory

The package includes:

- deterministic analysis and figure-generation scripts;
- a bundle validator that distinguishes structural errors from known scientific warnings;
- public-safe aggregate and pseudonymous outcome tables;
- a condition-blinded adjudication-queue generator and human-review templates;
- evidence-standard, statistical-analysis, human-adjudication, and security-hardening protocols;
- separate public and private package boundaries;
- SHA-256 manifests for integrity verification.

The package does not contain the exact corpus text, request-time retrieval outputs, full run prompt/manifests, or completed human adjudications. Those absences are retained as explicit limitations rather than filled by inference.

References

- Autio, C., Schwartz, R., Dunietz, J., Jain, S., Stanley, M., Tabassi, E., Hall, P., & Roberts, K. (2024). *Artificial intelligence risk management framework: Generative artificial intelligence profile* (tech. rep. No. NIST AI 600-1). National Institute of Standards and Technology. <https://doi.org/10.6028/NIST.AI.600-1>
- Es, S., James, J., Espinosa Anke, L., & Schockaert, S. (2024). RAGAs: Automated evaluation of retrieval augmented generation. *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, 150–158. <https://doi.org/10.18653/v1/2024.eacl-demo.16>
- Gao, T., Yen, H., Yu, J., & Chen, D. (2023). Enabling large language models to generate text with citations. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 6465–6488. <https://doi.org/10.18653/v1/2023.emnlp-main.398>

- Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., & Wang, H. (2024). Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*. <https://doi.org/10.48550/arXiv.2312.10997>
- Krishna, S., Krishna, K., Mohananey, A., Schwarcz, S., Stambler, A., Upadhyay, S., & Faruqui, M. (2025). Fact, fetch, and reason: A unified evaluation of retrieval-augmented generation. *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 4745–4759. <https://doi.org/10.18653/v1/2025.naacl-long.243>
- Last Z Community Platform. (2026a, June 1). *Disclaimer, terms, and community use*. Retrieved August 11, 2026, from <https://last-z.us/legal>
- Last Z Community Platform. (2026b). *Last z platform: Ai command portals for last z communities* [Independent community platform]. Retrieved August 11, 2026, from <https://last-z.us/>
- Last Z Server 496 Community. (2026). *Server 496 rules* [Community-governed, server-scoped normative rules]. Retrieved August 12, 2026, from <https://496.last-z.us/server-rules>
- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33, 9459–9474. <https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html>
- Li, D., Jiang, B., Huang, L., Beigi, A., Zhao, C., Tan, Z., Bhattacharjee, A., Jiang, Y., Chen, C., Wu, T., Shu, K., Cheng, L., & Liu, H. (2025). From generation to judgment: Opportunities and challenges of LLM-as-a-judge. *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 2757–2791. <https://doi.org/10.18653/v1/2025.emnlp-main.138>
- Min, S., Krishna, K., Lyu, X., Lewis, M., Yih, W.-t., Koh, P. W., Iyyer, M., Zettlemoyer, L., & Hajishirzi, H. (2023). FActScore: Fine-grained atomic evaluation of factual precision in long-form text generation. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 12076–12100. <https://doi.org/10.18653/v1/2023.emnlp-main.741>
- Muller, S., Loison, A., Omrani, B., & Viaud, G. (2025). GroUSE: A benchmark to evaluate evaluators in grounded question answering. *Proceedings of the 31st International Conference on Computational Linguistics*, 4510–4534. <https://aclanthology.org/2025.coling-main.304/>
- Rashkin, H., Nikolaev, V., Lamm, M., Aroyo, L., Collins, M., Das, D., Petrov, S., Tomar, G. S., Turc, I., & Reitter, D. (2023). Measuring attribution in natural language generation models. *Computational Linguistics*, 49(4), 777–840. https://doi.org/10.1162/coli_a_00486
- Saad-Falcon, J., Khattab, O., Potts, C., & Zaharia, M. (2024). ARES: An automated evaluation framework for retrieval-augmented generation systems. *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 338–354. <https://doi.org/10.18653/v1/2024.naacl-long.20>
- Song, Y., Kim, Y., & Iyyer, M. (2024). VeriScore: Evaluating the factuality of verifiable claims in long-form text generation. *Findings of the Association for Computational Linguistics: EMNLP 2024*, 9447–9474. <https://doi.org/10.18653/v1/2024.findings-emnlp.552>
- Tabassi, E. (2023). *Artificial intelligence risk management framework (AI RMF 1.0)* (tech. rep. No. NIST AI 100-1). National Institute of Standards and Technology. <https://doi.org/10.6028/NIST.AI.100-1>
- Xie, K., Laban, P., Choubey, P. K., Xiong, C., & Wu, C.-S. (2025). Do RAG systems cover what matters? evaluating and optimizing responses with sub-question coverage. *Proceedings of the 2025 Conference*

of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), 5836–5849. <https://doi.org/10.18653/v1/2025.naacl-long.301>

Zhu, K., Luo, Y., Xu, D., Yan, Y., Liu, Z., Yu, S., Wang, R., Wang, S., Li, Y., Zhang, N., Han, X., Liu, Z., & Sun, M. (2025). RAGEval: Scenario specific RAG evaluation dataset generation framework. *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 8520–8544. <https://doi.org/10.18653/v1/2025.acl-long.418>