IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Ingénierie des Systèmes d'Information
Vol. 31, No. 10, October, 2026, pp. 3297-3306
Journal homepage: http://iieta.org/journals/isi

Architectural Governance of Agentic Retrieval-Augmented Generation for Evidence-Grounded Generative AI Decision Support

Khalid Hakimi*, Omar Abahmane, Najlae Yachou

Interdisciplinary Laboratory for Applied Research, Doctoral Studies Center, International University of Agadir (LIDRA), Agadir 80000, Morocco

Corresponding Author Email: hakimi.khalid@isiam.ma

Copyright: ©2026 The authors. This article is published by IIETA and is licensed under the CC BY 4.0 license (http://creativecommons.org/licenses/by/4.0/).

https://doi.org/10.18280/isi.311005

Received: 20 May 2026
Revised: 3 August 2026
Accepted: 11 August 2026
Available online: 31 October 2026

Keywords:
agentic retrieval-augmented generation, architectural governance, evidence verification, generative AI decision support, faithfulness, task–technology fit, design science research

ABSTRACT

Generative AI decision-support systems must balance evidence fidelity against model and operating costs. Yet governance principles are rarely translated into architectural controls that verify outputs before release. This study designs and evaluates a seven-tier agentic retrieval-augmented generation (RAG) architecture based on a Verify-Before-Emit principle, combining evidence verification, bounded repair, diagnostic refusal, and audit logging. Following a design science research (DSR) approach, we conducted a 2 × 2 factorial evaluation of model tier (Gemini 2.5 Flash or Gemini 2.5 Pro) and governance (governed or ungoverned) on 500 stratified questions drawn equally from NQ-Open and HotpotQA. Faithfulness was assessed with retrieval-augmented generation assessment (RAGAS) using GPT-4o as an external judge, and practical equivalence was examined using two one-sided tests (TOST). The governed Flash configuration achieved 90.6% faithfulness, compared with 89.2% for governed Pro, and met the study-defined ±4.0-percentage-point equivalence margin (90% confidence interval (CI): −1.17 to 3.97 percentage points). Its normalized Application Programming Interface (API) cost per 1,000 queries was 76.3% lower than that of the governed Pro configuration. It also exceeded the ungoverned Pro baseline by 6.4 percentage points (p = 0.002). Ablation analysis identified verification as the largest contributor: removing it reduced faithfulness by 9.0 percentage points. These findings provide preliminary evidence that architectural governance can partly offset differences between model tiers in evidence-grounded question answering. However, the results are limited to one model family and two public question answering benchmarks; full deployment costs and organizational decision outcomes remain to be tested.

1. INTRODUCTION

Organizations increasingly use generative artificial intelligence (AI) to support knowledge creation, retrieval, and decision-making [1, 2]. However, adoption in high-stakes settings is constrained by a critical trade-off: frontier models offer greater reliability at high operational cost [3], whereas high-efficiency models are less costly but exhibit greater hallucination risk [4]. AI errors can erode user trust and complicate subsequent trust calibration [5].

This challenge is further amplified by limited transparency. Limited transparency can make it difficult for users to understand and validate AI-supported decisions [6, 7]. Although governance frameworks for trustworthy AI have received increased attention [8, 9], limited work explains how those principles can be operationalized through concrete architectural mechanisms.

Model scaling is widely used as a route to improved language-model performance [10]. However, for retrieval-augmented tasks, reliable performance depends on the system's ability to adhere to external evidence rather than default to internal memory [11]. From an Information Systems (IS) perspective, technical capability alone is often insufficient to produce organizational value without appropriate sociotechnical integration, governance, and accountability [12-14].

This study addresses that gap through the design, implementation, and evaluation of a seven-tier agentic retrieval-augmented generation (RAG) architecture that operationalizes a Verify-Before-Emit principle. The architecture embeds autonomous verification, adaptive repair, and diagnostic refusal directly into the decision-support workflow. The study is guided by three research questions:

RQ1: Can architectural governance enable a high-efficiency model to achieve practical equivalence in faithfulness to a frontier model?

RQ2: Which architectural governance mechanisms contribute most to grounded reliability?

RQ3: What theoretical and managerial implications follow from using architectural governance rather than simply scaling the model?

The paper makes three contributions. The first is a design artifact: the seven-tier governed architecture. The second is empirical evidence that a governed, efficient model reaches a

practically equivalent level of faithfulness at 76.3% lower cost. The third refines task–technology fit (TTF), since the results are consistent with a governance-contingent capability-fit relationship. In addition, the study proposes the Parametric Rigidity Index (PRI) as an exploratory indicator to distinguish contradiction-based grounding failures from other types of errors.

2. LITERATURE REVIEW

2.1 Design principles for agentic verification

Consistent with established design science research (DSR) principles [15] and recent guidance for AI design research in IS [16], the proposed artifact is guided by three design principles. First, adaptive governance through recursive verification requires generated outputs to be evaluated against retrieved evidence before emission; repair is initiated when misalignment is detected. Second, parametric accountability through diagnostic refusal requires the system to refuse when evidence is insufficient rather than fill gaps with unsupported parametric content. Third, transparent auditability requires verification outcomes, repairs, and refusals to be logged for human inspection. These three principles are consistent with agentic self-reflection frameworks [17, 18]. The Verify-Before-Emit principle extends Constitutional AI approaches [19] from training-time alignment to inference-time governance.

2.2 Parametric rigidity as a preliminary diagnostic indicator

One major issue with RAG is that a model may possess strong parametric knowledge yet fail to comply with the external evidence provided. This study introduces the PRI, which serves as an exploratory diagnostic indicator consistent with evidence that large language models (LLMs) often privilege internal memory over retrieved context [11, 20, 21]. PRI is calculated as $C/N \times 100$, where C represents queries containing contradiction-based failure codes, and N represents total queries evaluated. A contradiction-based failure was coded only when the generated answer directly conflicted with retrieved evidence; missing evidence, incomplete answers, retrieval sparsity, and general reasoning failures were excluded.

2.3 Task–technology fit in AI-mediated decision support

TTF predicts that performance improves when technology capabilities align with task requirements [22]. Recent organizational GenAI research likewise shows that task–technology alignment shapes performance [23]. TTF has also been applied to AI-mediated decision-support contexts, including clinical systems [24] and diagnostic tools [25]. In evidence-intensive contexts, a primary condition for success is disciplined adherence to externally retrieved evidence. A model with greater internal knowledge may therefore not achieve better fit when parametric knowledge conflicts with retrieved context [11, 26, 27].

This research treats fit as a property of the model-architecture combination rather than of model capability alone. This yields a testable implication for evidence-intensive settings: governance may condition how model capability relates to fit when parametric knowledge competes with retrieved evidence. The present study evaluates this possibility within the tested Gemini family. Accordingly, any observed crossover is interpreted as preliminary evidence of a governance-contingent capability-fit pattern rather than as proof of a general non-monotonic relationship [28].

The central proposition is therefore that, in grounding-intensive tasks, architectural governance may condition the relationship between model capability and TTF when parametric knowledge competes with retrieved evidence. The direction and magnitude of this relationship are treated as empirical questions rather than assumed a priori.

2.4 Governance, structural assurance, and trust

Recent organizational AI governance research defines governance through rules, practices, and processes that align AI use with organizational strategies, values, legal requirements, and accountability expectations [29]. This is particularly relevant to trustworthy AI because system trustworthiness depends not only on algorithmic improvement but also on systematic transparency and accountability [8, 14]. The EU AI Act [30] likewise includes traceability and human-oversight requirements for high-risk AI systems [31], reinforcing the institutional need for the architectural governance mechanisms discussed in this paper. Recent work also emphasizes enterprise-level AI governance and responsible architecture as complementary mechanisms for operationalizing such controls [32-34].

The transition from static pipelines to agentic architectures creates opportunities for embedded control. Agentic systems can observe, reason, and revise intermediate outputs, initiating corrective actions before a final answer is emitted [17, 35]. Corrective RAG illustrates this principle by assessing retrieved evidence before generation and triggering corrective actions when evidence quality is insufficient [36]. Specialized agent orchestration can improve both performance and interpretability [37-39].

Structural assurance refers to institutional and technological measures that increase the likelihood of successful outcomes [40]. This study uses the concept to describe verification, refusal, and repair as mechanisms that create conditions for justified trust. Trust is required for sustained reliance on AI, but AI errors and automation bias can erode or distort reliance [5, 41]. Recent AI trust research highlights reliability, safety, transparency, and institutional or technical trustworthiness cues [42, 43]. Explanation and decision-control mechanisms can improve understanding and, under appropriate conditions, support reliance and decision performance [44, 45]. Recent experiments further show that transparent decision-process information can increase trust in AI assistance [46]. This study addresses the architectural supply of trustworthiness rather than the demand side of perceived trust.

2.5 Research gap

The gap lies not in the general discussion of trustworthy AI, but in how governance principles become operational at the architectural level in generative decision-support settings under realistic economic constraints. Although RAG is widely recognized as a paradigm for reducing hallucinations [47, 48], the literature has focused primarily on pipeline enhancement rather than embedded governance. This study therefore views trustworthy generative AI as an architectural governance outcome. Two main arguments are advanced: architectural controls can enhance TTF by disciplining generation, and they

can create structural assurance through auditability and resistance to unsupported inference.

3. METHODS

3.1 Organizational context and design science research approach

This study was situated in organizational decision-support environments in which knowledge workers interpret and validate information from heterogeneous document collections under uncertainty. It focused on high-stakes, asynchronous contexts in which structural assurance takes priority over minimizing latency. DSR provided the methodological foundation [15, 49]. Following Hevner's [50] three-cycle view, the relevance cycle examined the reliability-cost trade-off, the rigor cycle drew on TTF, governance, and agentic reasoning, and the design cycle involved iterative refinement of the artifact.

3.2 The seven-tier agentic retrieval-augmented generation architecture

In traditional linear RAG systems [47], generation is the terminal pipeline stage. The main design principle is Verify-Before-Emit: every response passes through a structured assurance cycle (provisional generation, evidence-based verification, targeted repair, verified emission or diagnostic refusal) before release. Reliability is transformed from a post hoc concern to an ex-ante design constraint. Table 1 summarizes the seven tiers of the agentic RAG architecture. Figure 1 summarizes the interfaces among retrieval, reranking, generation, runtime verification, and targeted repair.

Table 1. Design science research (DSR) artifact specification with theoretical mapping

Tier	Component	Function	Theoretical Construct
1	Query receiver	Input normalization and trace initialization	User engagement
2	Query analyzer	Decompose multi-hop queries	Task decomposition
3	Hybrid retrieval	Best Matching 25 (BM25) + dense retrieval with reciprocal rank fusion (RRF)	Information quality
4	Reranking	Cross-encoder scoring, noise reduction	Transparency
5	Grounding validator	Evidence verification checks	Structural assurance
6	Generative node	Constrained synthesis on evidence	Parametric discipline
7	Output and audit	Delivery, refusal logging, traces	Accountability

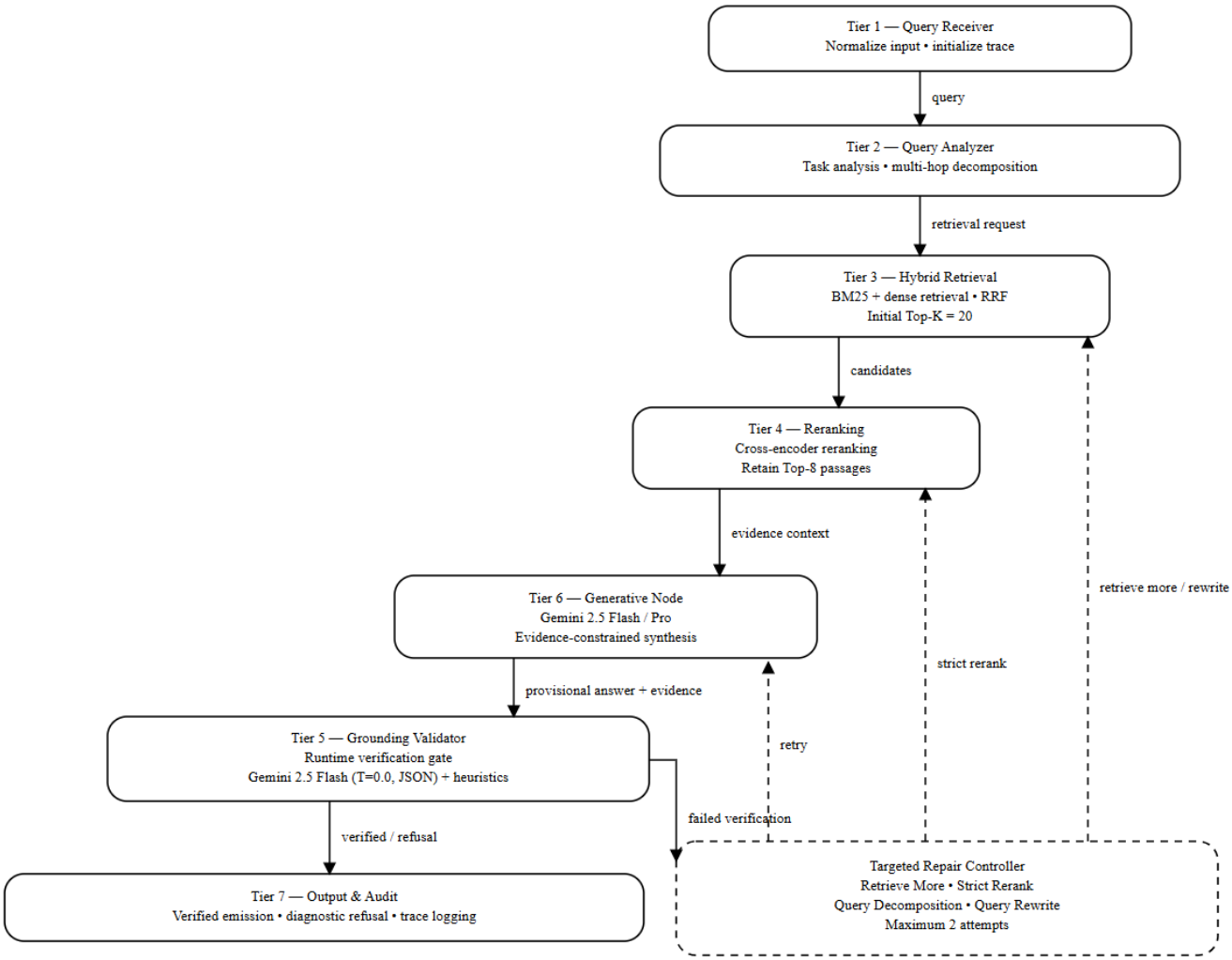

Figure 1. Formal component-interface view of the seven-tier agentic retrieval-augmented generation (RAG) architecture and bounded repair loop

When verification failed, the system initiated an autonomous repair phase extending iterative self-refinement principles [51]. Four primary repair strategies were triggered deterministically: Retrieve More (missing evidence), Strict Rerank (noisy evidence), Query Decomposition (incomplete multi-hop), and Query Rewriting (retrieval mismatch). Each query was limited to two generation attempts, with at most one bounded repair between attempts to preserve computational discipline. If reliable emission could not be achieved, the system produced an evidence-explicit, action-oriented diagnostic refusal. Figure 1 provides a formal component-interface specification: solid arrows indicate the primary processing path, whereas dashed arrows indicate conditional repair paths. The tier labels denote architectural responsibilities rather than a strict numerical execution sequence. At inference time, retrieval and reranking operated on a bounded candidate pool and the repair loop was capped at two attempts. Scalability beyond the tested corpus was not empirically evaluated; indexing, retrieval, context-management, and orchestration overhead are expected to increase with corpus size.

3.3 Benchmark and experimental design

A 500-query stratified benchmark was employed for evaluation: 250 NQ-Open single-hop queries [52] and 250 HotpotQA multi-hop queries [53]. Four primary configurations formed a 2 × 2 structure (Model Tier × Governance): Governed Efficient (Gemini 2.5 Flash), Governed Frontier (Gemini 2.5 Pro), Baseline Efficient (ungoverned Flash), and Baseline Frontier (ungoverned Pro). Targeted ablation variants (No Verification, No Repair, No Refusal, Vanilla) were assessed separately.

3.4 Evaluation metrics and statistical analysis

The principal metric was faithfulness, measured with retrieval-augmented generation assessment (RAGAS) [54], consistent with fine-grained factuality evaluation [55] and RAG assessment frameworks such as ARES [56]. Faithfulness was assessed with GPT-4o as an external judge and binarized at 0.55 as an operational rule for the statistical analysis rather than as a universal theoretical cutoff. Appendix Table A1 reports sensitivity at 0.50, 0.55, and 0.60. GPT-4o belongs to a different model family from the Gemini generators, reducing direct same-model circularity but not eliminating LLM-as-judge bias. McNemar's exact test assessed differences in paired binary outcomes between the governed configurations, and two one-sided tests (TOST) assessed practical equivalence using the study-specific ±4.0 percentage points (pp) operational tolerance. Ordinary estimates are reported with 95% confidence intervals (CI), whereas TOST uses a 90% CI, consistent with TOST at $\alpha = 0.05$. A two-proportion z-test compared the governed efficient configuration with the ungoverned frontier configuration. A supplemental aggregate binomial logistic model examined Model Tier, Governance, and their interaction across the four primary configuration totals. Additional metrics included PRI, cost per 1,000 queries, latency (P50, P95), repair success rate, and actionable refusal rate. A human validation audit was conducted on 150 cases by three PhD-level researchers in IS and natural language processing, blind to experimental conditions. The retained audit file records consolidated verdicts rather than individual ratings. Inter-rater reliability therefore cannot be recomputed, and the agreement reported below is between the automated classification and the consolidated human verdict. The sampling rule behind the 150-case set is not documented in the retained materials. Core system parameters are reported in Table B1.

4. RESULTS

4.1 Faithfulness and practical equivalence

Table 2 reports comparative performance across the four primary configurations on the shared 500-query benchmark.

Governance increased faithfulness for both models: +9.0 pp for the efficient model (81.6% to 90.6%) and +5.0 pp for the frontier model (84.2% to 89.2%) relative to their respective ungoverned baselines. The governed efficient configuration exceeded the ungoverned frontier baseline by 6.4 pp ($z = 3.049$, $p = 0.002$; 95% CI [2.2, 10.6] pp), corresponding to Cohen's $h = 0.19$, or approximately 32 additional faithful responses per 500 queries. This result provides preliminary evidence consistent with the proposed governance-contingent crossover. McNemar's test between governed configurations was non-significant ($p = 0.443$). The TOST 90% CI [−1.17, +3.97] pp remained within the study-specific ±4.0 pp operational bounds (Cohen's $d = 0.04$); the equivalence conclusion is therefore margin-dependent and would not hold under a narrower ±3.0 pp bound. In the aggregate factorial check, governance had a strong effect in the Flash reference condition (OR = 2.173, 95% CI [1.492, 3.165], $p < 0.001$). Model Tier did not reach significance (OR = 1.202, 95% CI [0.864, 1.672], $p = 0.275$), and neither did the Model × Governance interaction (OR = 0.713, 95% CI [0.421, 1.209], $p = 0.210$). Human validation showed 98.7% agreement between the automated RAGAS/GPT-4o classification and consolidated human judgment (Cohen's $\kappa = 0.940$; 95% CI [0.839, 1.000]), with no false negatives and two false positives (1.3%).

Table 2. Comparative performance across evaluated configurations

System	Passes/500	Faithfulness (%)	95% Confidence Interval (CI)	Difference (Percentage Points (pp))	Cost/1,000	P50	P95
Agentic Flash (ours)	453	90.6	[88.0, 93.2]	–	$3.13	9.0 s	22.0 s
Agentic Pro	446	89.2	[86.5, 91.9]	–1.4	$13.25	14.0 s	35.0 s
Vanilla Pro	421	84.2	[81.0, 87.4]	–6.4	$8.75	8.2 s	15.3 s
Vanilla Flash	408	81.6	[78.2, 85.0]	–9.0	$2.14	4.5 s	8.0 s

4.2 Ablation analysis

Table 3 presents the findings of the ablation study. Removal of verification was the most detrimental to faithfulness (−9.0 pp) and returned performance to the Vanilla Flash level (81.6%). Removal of repair was more detrimental (−3.6 pp)

than removal of refusal (−2.4 pp). Verification is therefore the dominant assurance mechanism, while repair and refusal provide smaller incremental gains. Figure 2 visualizes this pattern.

Across the benchmark, repair was triggered in 40 cases (8.0% of queries). The overall repair success rate was 45.0%, and the repaired subset showed an average gain of +5.8 pp. Table 4 reports strategy-level effectiveness.

The dominant operational trigger was retrieval sparsity. Query decomposition produced the highest local gain (+9.0 pp) among the triggered repairs. Repair frequency by query type was not subjected to a pre-specified subgroup test and is therefore not interpreted inferentially.

Table 3. Ablation results

Configuration	Faithfulness (%)	Δ (Percentage Points (pp))	Cost/1,000	Mechanism
Full agentic	90.6	–	$3.13	Governed architecture
No refusal	88.2	−2.4	$3.13	Accountability
No repair	87.0	−3.6	$2.85	Adaptive governance
No verification	81.6	−9.0	$2.14	Structural assurance
Vanilla	81.6	−9.0	$2.14	No governance

Table 4. Repair strategy effectiveness

Strategy	Triggers	Success Rate	Average Δ Faithfulness	Share (%)
Retrieve More	20	50.0%	+6.5 pp	50.0
Strict Rerank	12	41.7%	+4.1 pp	30.0
Query Decomposition	4	50.0%	+9.0 pp	10.0
Query Rewrite	4	25.0%	+3.5 pp	10.0
Total	**40**	**45.0%**	**+5.8 pp**	**100.0**

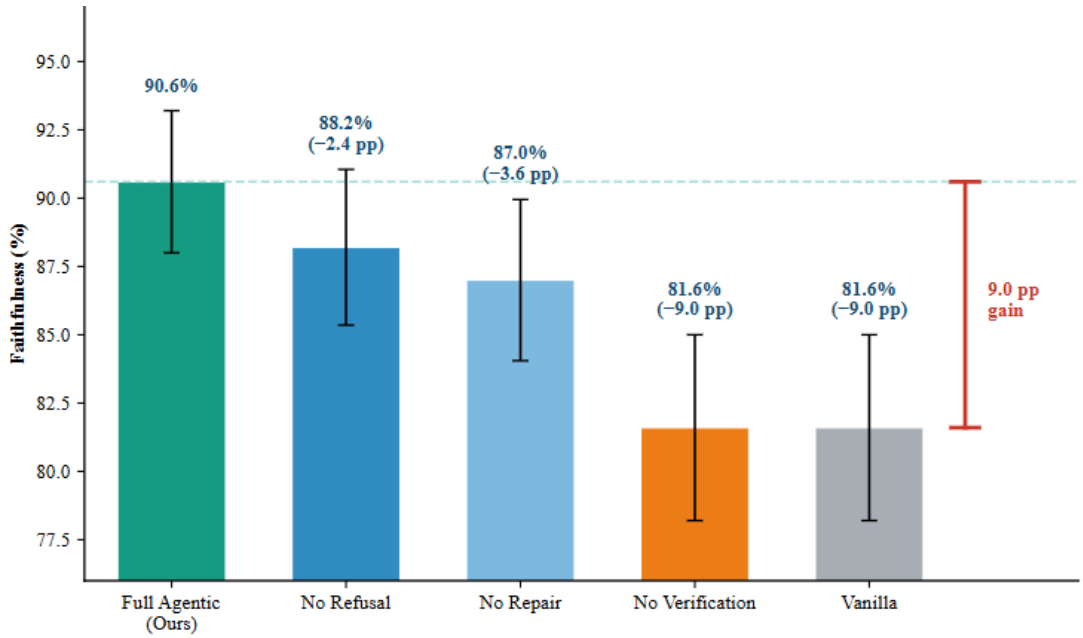

Figure 2. Faithfulness across full architecture and ablation conditions
Note: Verification removal yields the largest performance degradation. Error bars indicate 95% confidence intervals.

4.3 Failure analysis, refusal, and efficiency

There were 10/500 contradiction-based failures in the frontier model and 5/500 in the efficient model (PRI = 2.0% vs. 1.0%). The difference was not statistically significant (two-sided Fisher exact $p = 0.298$), so PRI is interpreted as an exploratory diagnostic indicator rather than a validated inferential metric. The efficient model experienced more reasoning-gap failures than the frontier model (10 vs. 2), while retrieval sparsity was similar (25 vs. 20). These directional patterns require replication at larger scale.

Of the 51 diagnostic refusals, 42 (82.3%) reflected evidence gaps, 5 (9.8%) evidence conflicts, and 4 (7.9%) query ambiguity. Fifty refusals (98.0%) provided actionable next-step guidance, indicating that refusal functioned as a governed evidentiary response rather than as arbitrary abstention.

The governed efficient system achieved 90.6% faithfulness at $3.13 per 1,000 queries, compared with $13.25 for the governed frontier system, representing a 76.3% cost reduction. The ungoverned frontier baseline cost $8.75 per 1,000 queries and achieved 84.2% faithfulness. Cost per 1,000 queries is normalized token-based Application Programming Interface (API) costs under the pricing schedule used for the experiment; they exclude fixed infrastructure, maintenance, monitoring, orchestration, and broader governance overhead. Figure 3 depicts the cost-faithfulness relationship, and Figure 4 visualizes the observed crossover pattern. The governed efficient configuration had a P50 latency of 9.0 s and a P95 latency of 22.0 s. No direct comparison with human verification time was conducted.

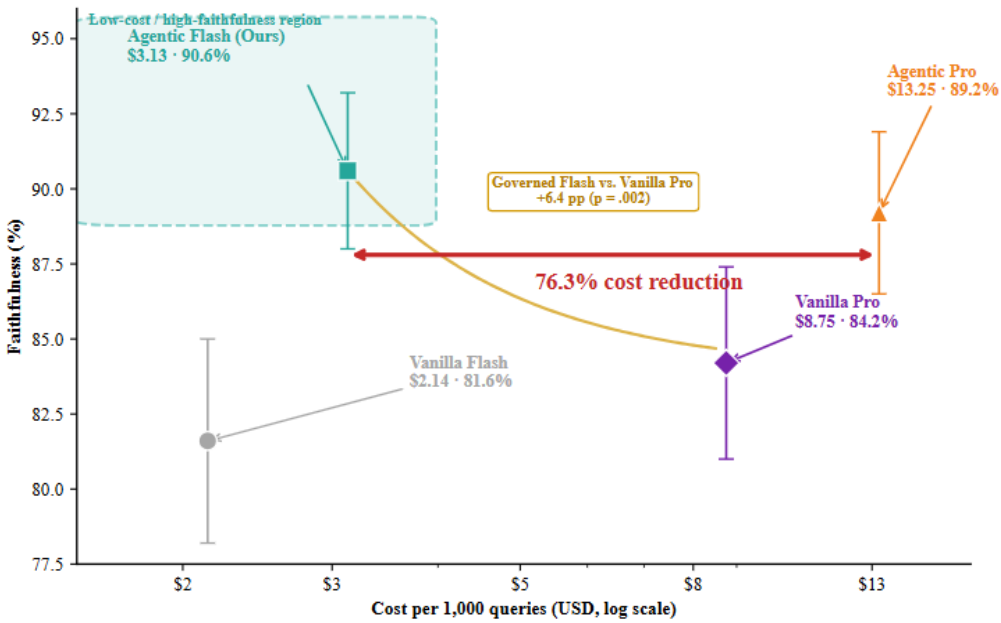

Figure 3. Cost-faithfulness relationship across configurations

Note: The governed efficient configuration occupies a low-cost, high-faithfulness region while maintaining frontier-comparable faithfulness. Vertical error bars indicate 95% confidence intervals for faithfulness.

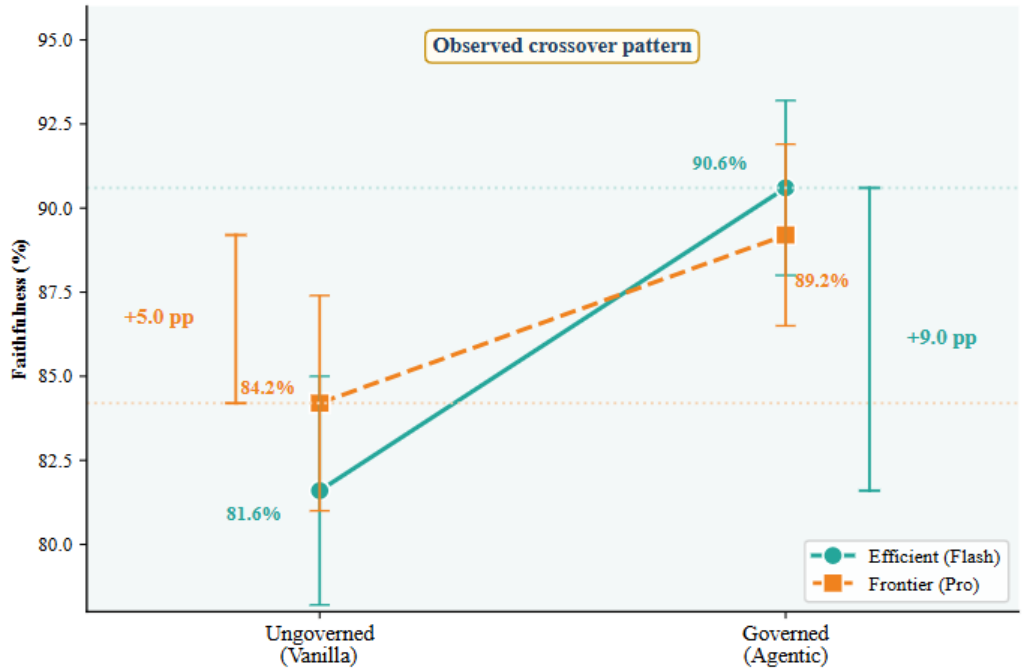

Figure 4. Observed capability-fit crossover

Note: Under ungoverned conditions, the frontier model leads (+2.6 pp); under governed conditions, the efficient model leads (+1.4 pp). The crossing lines illustrate the observed pattern within the tested Gemini family rather than establishing a general non-monotonic relationship. Point error bars indicate 95% confidence intervals; colored side brackets indicate the observed governance gains.

5. DISCUSSION

5.1 Key findings and theoretical implications

The core empirical pattern indicates that the reliability-cost trade-off is not determined by model scale alone. Together, the practical equivalence between governed configurations and the observed crossover provides preliminary evidence consistent with a governance-contingent capability-fit pattern within the tested setting. The ablation analysis identifies verification as the principal mechanism, with repair and refusal providing smaller incremental contributions. The models also exhibited different failure profiles: contradiction-based failures were more frequent in the frontier model, whereas reasoning-gap failures were more frequent in the efficient model. These findings suggest that architectural suitability cannot be inferred from general model capability alone. Two mechanisms are suggested: passive parametric interference, in which internal knowledge creates inconsistency, and active parametric rigidity, in which the model ignores retrieved evidence [26, 27].

This work contributes to the IS literature on three fronts.

First, it extends TTF to the setting of generative AI by showing an observed crossover within the tested model family: the ungoverned frontier model (84.2%) was outperformed by the governed efficient model (90.6%; $p = 0.002$). This result is treated as preliminary evidence of a governance-contingent capability-fit pattern rather than as proof of a universal non-monotonic relationship. Second, this paper builds on governance and structural assurance by showing how trustworthy AI principles can be instantiated architecturally through verification-before-emission, bounded repair, and diagnostic refusal. This interpretation is consistent with recent enterprise AI governance and responsible-architecture work [32-34]. Third, the proposed PRI provides an exploratory diagnostic distinction between contradiction-based errors and other failure sources; the current directional findings (PRI: 1.0 versus 2.0) require replication at larger scale. Table 5 shows the mapping between empirical findings and corresponding TTF propositions.

Table 5. Empirical findings mapped to TTF propositions

TTF Proposition	Empirical Evidence
TTF improves individual performance	Governance raised faithfulness by +9.0 pp (efficient) and +5.0 pp (frontier)
Fit is task-contingent	Repair triggered in 8.0%; no pre-specified query-type subgroup test
Governance can complement model capability	Governed Flash and Governed Pro were practically equivalent under the study-specific ±4.0 pp criterion; Governed Flash had lower normalized API cost
Capability-fit may be governance-contingent	Observed crossover: Governed Flash (90.6%) > Vanilla Pro (84.2%), $p = 0.002$; interpretation limited to tested setting

Note: TTF = Task–Technology Fit, API = Application Programming Interface, pp = percentage points.

5.2 Managerial implications

Organizations should not assume that model scaling alone is sufficient for reliable decision support. In the tested configurations, architectural governance materially affected faithfulness and should therefore be considered alongside model selection. The added runtime associated with the governed workflow may be acceptable in asynchronous decision-support settings, although deployment-specific latency requirements should be evaluated. Diagnostic refusals identifying evidence gaps create a source of feedback for improving an organization's documentation and retrieval quality.

5.3 Limitations and future research

Although this study does not measure user-level trust or reliance, the contribution centers on structural assurance mechanisms. The main comparison was limited to one model family. Future research should replicate the architecture across other model families. Public question answering datasets were used instead of document-heavy enterprise corpora, limiting ecological validity for organizational deployment; field validation with proprietary enterprise collections is therefore needed. The cost analysis covers normalized API usage costs rather than full infrastructure, maintenance, monitoring, or governance overhead. GPT-4o evaluation may retain LLM-as-judge bias despite being external to the Gemini generation family, although the human audit provides an independent check. Repair frequency by query type was not formally tested. Appendix A demonstrates that the principal result is stable across threshold values (0.50, 0.55, 0.60).

6. CONCLUSIONS

The study shows that architectural governance can enable a high-efficiency generative AI model to achieve grounded reliability comparable to a governed frontier model while reducing normalized API cost. The seven-tier agentic RAG artifact operationalizes Verify-Before-Emit through verification, bounded repair, and diagnostic refusal. The observed crossover suggests, within the tested Gemini family, that reliability in grounding-intensive tasks depends on the model-architecture combination rather than model capability alone. For organizations, frontier-model adoption is therefore not the only route to reliable decision support; architectural governance provides an alternative design strategy. The findings remain limited to one model family and public question answering benchmarks. Future research should validate the architecture with enterprise corpora, other model families, and direct measures of user trust.

REFERENCES

[1] Alavi, M., Leidner, D.E., Mousavi, R. (2024). A knowledge management perspective of generative artificial intelligence. Journal of the Association for Information Systems, 25(1): 1-12. https://doi.org/10.17705/1jais.00859

[2] Leoni, L., Gueli, G., Ardolino, M., et al. (2024). AI-empowered KM processes for decision-making: Empirical evidence from worldwide organisations. Journal of Knowledge Management, 28(11): 320-347. https://doi.org/10.1108/JKM-03-2024-0262

[3] Chen, L., Zaharia, M., Zou, J. (2024). FrugalGPT: How to use large language models while reducing cost and improving performance. Transactions on Machine Learning Research. https://openreview.net/forum?id=cSimKw5p6R.

[4] Ji, Z., Lee, N., Frieske, R., et al. (2023). Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12): 1-38. https://doi.org/10.1145/3571730

[5] Pareek, S., Velloso, E., Gonçalves, J. (2024). Trust development and repair in AI-assisted decision-making during complementary expertise. Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency, Rio de Janeiro, Brazil, pp. 546-561. https://doi.org/10.1145/3630106.3658924

[6] Senoner, J., Schallmoser, S., Kratzwald, B., et al. (2024). Explainable AI improves task performance in human-AI collaboration. Scientific Reports, 14(1): 31150. https://doi.org/10.1038/s41598-024-82501-9

[7] Huynh, M.T. (2024). Using generative AI as

decision-support tools: Unraveling users' trust and AI appreciation. Journal of Decision Systems, 1-32. https://doi.org/10.1080/12460125.2024.2428166

[8] Díaz-Rodríguez, N., Del Ser, J., Coeckelbergh, M., et al. (2023). Connecting the dots in trustworthy artificial intelligence: From AI principles, ethics, and key requirements to responsible AI systems and regulation. Information Fusion, 99: 101896. https://doi.org/10.1016/j.inffus.2023.101896

[9] Papagiannidis, E., Mikalef, P., Conboy, K. (2025). Responsible artificial intelligence governance: A review and research framework. The Journal of Strategic Information Systems, 34(2): 101885. https://doi.org/10.1016/j.jsis.2024.101885

[10] Kaplan, J., McCandlish, S., Henighan, T., et al. (2020). Scaling laws for neural language models. arXiv preprint. https://doi.org/10.48550/arXiv.2001.08361

[11] Mallen, A., Asai, A., Zhong, V., Das, R., Khashabi, D., Hajishirzi, H. (2023). When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 9802-9822. https://doi.org/10.18653/v1/2023.acl-long.546

[12] Storey, V.C., Yue, W.T., Zhao, J.L., et al. (2025). Generative artificial intelligence: Evolving technology, growing societal impact, and opportunities for information systems research. Information Systems Frontiers, 27: 2081-2102. https://doi.org/10.1007/s10796-025-10581-7

[13] Al-kfairy, M. (2025). Strategic integration of generative AI in organizational settings: Applications, challenges, and adoption requirements. IEEE Engineering Management Review, 53(6): 80-97. https://doi.org/10.1109/EMR.2025.3534034

[14] Lu, Q., Zhu, L., Xu, X., et al. (2024). Responsible AI pattern catalogue: A collection of best practices for AI governance and engineering. ACM Computing Surveys, 56(7): 1-35. https://doi.org/10.1145/3626234

[15] Hevner, A.R., March, S.T., Park, J., et al. (2004). Design science in information systems research. MIS Quarterly, 28(1): 75-105. https://doi.org/10.2307/25148625

[16] Abbasi, A., Parsons, J., Pant, G., et al. (2024). Pathways for design research on artificial intelligence. Information Systems Research, 35(2): 441-459. https://doi.org/10.1287/isre.2024.editorial.v35.n2

[17] Yao, S., Zhao, J., Yu, D., et al. (2023). ReAct: Synergizing reasoning and acting in language models. In International Conference on Learning Representations (ICLR 2023). https://react-lm.github.io/.

[18] Shinn, N., Cassano, F., Gopinath, A., Narasimhan, K., Yao, S. (2023). Reflexion: Language agents with verbal reinforcement learning. Advances in neural information processing systems, 36: 8634-8652. https://doi.org/10.52202/075280-0377

[19] Bai, Y., Kadavath, S., Kundu, S., et al. (2022). Constitutional AI: Harmlessness from AI feedback. arXiv preprint arXiv:2212.08073. https://doi.org/10.48550/arXiv.2212.08073

[20] Xie, J., Zhang, K., Chen, J., Lou, R., Su, Y. (2024). Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In International Conference on Learning Representations, 2024: 35623-35646. https://openreview.net/forum?id=auKAUJZMO6.

[21] Bi, B., Liu, S., Wang, Y., et al. (2026). Parameters vs. context: Fine-grained control of knowledge reliance in language models. In International Conference on Learning Representations, 2026: 106786-106808. https://openreview.net/forum?id=TJ3DqFiGau.

[22] Goodhue, D.L., Thompson, R.L. (1995). Task-technology fit and individual performance. MIS Quarterly, 19(2): 213-236. https://doi.org/10.2307/249689

[23] Przegalinska, A., Triantoro, T., Kovbasiuk, A., et al. (2025). Collaborative AI in the workplace: Enhancing organizational performance through resource-based and task-technology fit perspectives. International Journal of Information Management, 81: 102853. https://doi.org/10.1016/j.ijinfomgt.2024.102853

[24] Parsons, C.S., Zuiderwijk, A., Orchard, N.A.O., et al. (2025). Task-technology fit of artificial intelligence-based clinical decision support systems: A review of qualitative studies. BMC Medical Informatics and Decision Making, 25(1): 397. https://doi.org/10.1186/s12911-025-03237-8

[25] Zheng, R., Jiang, X., Shen, L., et al. (2025). Investigating clinicians' intentions and influencing factors for using an intelligence-enabled diagnostic clinical decision support system in health care systems: Cross-sectional survey. Journal of Medical Internet Research, 27: e62732. https://doi.org/10.2196/62732

[26] Zhang, Q., Xiang, Z., Xiao, Y., et al. (2025). FaithfulRAG: Fact-level conflict modeling for context-faithful retrieval-augmented generation. Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 21863-21882. https://doi.org/10.18653/v1/2025.acl-long.1062

[27] Asai, A., Wu, Z., Wang, Y., Sil, A., Hajishirzi, H. (2024). Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In International conference on learning representations, 2024: 9112-9141. https://openreview.net/forum?id=hSyW5go0v8.

[28] Sambamurthy, V., Zmud, R.W. (1999). Arrangements for information technology governance: A theory of multiple contingencies. MIS Quarterly, 23(2): 261-290. https://doi.org/10.2307/249754

[29] Birkstedt, T., Minkkinen, M., Tandon, A., et al. (2023). AI governance: Themes, knowledge gaps and future agendas. Internet Research, 33(7): 133-167. https://doi.org/10.1108/INTR-01-2022-0042

[30] European Parliament. (2024). Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union, L 2024/1689. http://data.europa.eu/eli/reg/2024/1689/oj.

[31] Ayankoya, F.Y., Kuyoro, S.O., Adekola, O.D., et al. (2025). Scalable human oversight for aligned large language models: A hybrid framework for intent fidelity. Ingénierie des Systèmes d'Information, 30(8): 2011-2020. https://doi.org/10.18280/isi.300807

[32] Luna-Reyes, L.F., Harrison, T.M. (2024). An enterprise view for artificial intelligence capability and governance: A system dynamics approach. Digital Government: Research and Practice, 5(1): 1-23. https://doi.org/10.1145/3609228

[33] Lu, Q., Zhu, L., Xu, X., et al. (2024). Toward responsible

AI in the era of generative AI: A reference architecture for designing foundation model-based systems. IEEE Software, 41(6): 91-100. https://doi.org/10.1109/MS.2024.3406333
[34] Batool, A., Zowghi, D., Bano, M. (2025). AI governance: A systematic literature review. AI and Ethics, 5: 3265-3279. https://doi.org/10.1007/s43681-024-00653-w
[35] Wang, L., Ma, C., Feng, X., et al. (2024). A survey on large language model based autonomous agents. Frontiers of Computer Science, 18: 186345. https://doi.org/10.1007/s11704-024-40231-1
[36] Yan, S.Q., Gu, J.C., Zhu, Y., et al. (2024). Corrective retrieval augmented generation. arXiv preprint arXiv:2401.15884. https://doi.org/10.48550/arXiv.2401.15884
[37] Wu, Q., Bansal, G., Zhang, J., et al. (2023). AutoGen: Enabling next-gen LLM applications via multi-agent conversation. arXiv preprint arXiv:2308.08155. https://doi.org/10.48550/arXiv.2308.08155
[38] Xi, Z., Chen, W., Guo, X., et al. (2025). The rise and potential of large language model based agents: A survey. Science China Information Sciences, 68(2): 121101. https://doi.org/10.1007/s11432-024-4222-0
[39] Joshi, D., Gupta, N., Patwardhan, N., et al. (2026). MediMate: An agent-based AI framework for healthcare e-commerce with multi-agent retrieval-augmented generation and technical validation. Ingénierie des Systèmes d'Information, 31(3): 731-737. https://doi.org/10.18280/isi.310307
[40] McKnight, D.H., Choudhury, V., Kacmar, C. (2002). Developing and validating trust measures for e-commerce: An integrative typology. Information Systems Research, 13(3): 334-359. https://doi.org/10.1287/isre.13.3.334.81
[41] Romeo, G., Conti, D. (2026). Exploring automation bias in human-AI collaboration: A review and implications for explainable AI. AI & Society, 41: 259-278. https://doi.org/10.1007/s00146-025-02422-7
[42] Afroogh, S., Akbari, A., Malone, E., et al. (2024). Trust in AI: Progress, challenges, and future directions. Humanities and Social Sciences Communications, 11: 1568. https://doi.org/10.1057/s41599-024-04044-8
[43] Agbabiaka, O., Ojo, A., Connolly, N. (2025). Requirements for trustworthy AI-enabled automated decision-making in the public sector: A systematic review. Technological Forecasting and Social Change, 215: 124076. https://doi.org/10.1016/j.techfore.2025.124076
[44] Westphal, M., Vössing, M., Satzger, G., et al. (2023). Decision control and explanations in human-AI collaboration: Improving user perceptions and compliance. Computers in Human Behavior, 144: 107714. https://doi.org/10.1016/j.chb.2023.107714
[45] Wang, P., Ding, H. (2024). The rationality of explanation or human capacity? Understanding the impact of explainable artificial intelligence on human-AI trust and decision performance. Information Processing & Management, 61(4): 103732. https://doi.org/10.1016/j.ipm.2024.103732
[46] Zhao, Y., Xiao, C. (2025). Overcome medical algorithm aversion: Conditional joint effect of direct and indirect information. International Journal of Human-Computer Interaction, 41(7): 3889-3900. https://doi.org/10.1080/10447318.2024.2344146
[47] Lewis, P., Perez, E., Piktus, A., et al. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems, 33: 9459-9474. https://doi.org/10.48550/arXiv.2005.11401
[48] Wu, S., Xiong, Y., Cui, Y., et al. (2026). Retrieval-augmented generation for natural language processing: A survey. Artificial Intelligence Review, 59: 192. https://doi.org/10.1007/s10462-026-11605-7
[49] Peffers, K., Tuunanen, T., Rothenberger, M.A., et al. (2007). A design science research methodology for information systems research. Journal of Management Information Systems, 24(3): 45-77. https://doi.org/10.2753/MIS0742-1222240302
[50] Hevner, A.R. (2007). A three-cycle view of design science research. Scandinavian Journal of Information Systems, 19(2): 87-92. https://aisel.aisnet.org/sjis/vol19/iss2/4.
[51] Madaan, A., Tandon, N., Gupta, P., et al. (2023). Self-refine: Iterative refinement with self-feedback. Advances in Neural Information Processing Systems, 36: 46534-46594. https://doi.org/10.52202/075280-2019
[52] Kwiatkowski, T., Palomaki, J., Redfield, O., et al. (2019). Natural Questions: A benchmark for question answering research. Transactions of the Association for Computational Linguistics, 7: 452-466. https://doi.org/10.1162/tacl_a_00276
[53] Yang, Z., Qi, P., Zhang, S., et al. (2018). HotpotQA: A dataset for diverse, explainable multi-hop question answering. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP), Brussels, Belgium, pp. 2369-2380. https://doi.org/10.18653/v1/D18-1259
[54] Es, S., James, J., Espinosa Anke, L., et al. (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, St. Julian's, Malta, pp. 150-158. https://doi.org/10.18653/v1/2024.eacl-demo.16
[55] Min, S., Krishna, K., Lyu, X., et al. (2023). FActScore: Fine-grained atomic evaluation of factual precision in long-form text generation. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP), Singapore, pp. 12076-12100. https://doi.org/10.18653/v1/2023.emnlp-main.741
[56] Saad-Falcon, J., Khattab, O., Potts, C., et al. (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, Mexico City, Mexico, pp. 338-354. https://doi.org/10.18653/v1/2024.naacl-long.20

APPENDIX

A. Threshold sensitivity analysis

To examine whether the principal result depends on a single faithfulness cutoff, the paired binary analysis was recalculated at nearby thresholds. Table A1 shows that the substantive conclusion remains stable: the difference remains small, McNemar's test remains non-significant, and the TOST 90% CI remains within the ±4.0 pp equivalence bounds.

Table A1. Threshold sensitivity analysis

Threshold	Flash Passes	Pro Passes	Difference (pp)	McNemar *p*	TOST 90% CI
0.50 (Loose)	92.6%	91.0%	+1.6	0.322	Equivalent [−0.73%, +3.93%]
0.55 (Chosen)	90.6%	89.2%	+1.4	0.443	Equivalent [−1.17%, +3.97%]
0.60 (Strict)	85.6%	84.6%	+1.0	0.583	Equivalent [−1.40%, +3.40%]

Note: CI = confidence interval, pp = percentage points.

The practical-equivalence conclusion held across all three cutoffs, indicating that the finding is not an artifact of a single threshold choice.

B. Core system reproducibility details

Table B1 reports the core parameters needed to interpret and reproduce the reported experiment.

Table B1. Core system hyperparameters used in the reported experiment

Parameter	Reported Value
Retrieval strategy	Hybrid BM25 + dense retrieval with RRF
BM25 parameters	k1 = 1.5; b = 0.75
Fusion weights	BM25 = 0.6; Dense = 0.4
RRF parameter	k = 60
Initial retrieval	Top-K = 20 candidates
Reranker	cross-encoder/ms-marco-MiniLM-L-6-v2; retain Top-8
Runtime verification judge	Gemini 2.5 Flash; T = 0.0; structured JSON + heuristics
Maximum agentic attempts	2
Retrieve-more repair	Top-K increased from 20 to 30
Generation models	Gemini 2.5 Flash/Gemini 2.5 Pro
Generation temperature	Gemini API default; not explicitly overridden
Post-hoc evaluator	RAGAS with GPT-4o
Faithfulness binarization	RAGAS score ≥ 0.55 = Pass