*applied sciences*

MDPI

*Article*

# Human-Rated Quality and Measured Token Use in Organizational RAG: A Field Comparison of Three Configurations

**Mihael Škarabot [1], Matjaž Gams [2] and Robert Leskovar [1,*]**

[1] Department of Informatics, Faculty of Organizational Sciences, University of Maribor, Kidričeva cesta 55a, 4000 Kranj, Slovenia; mihael.skarabot@student.um.si
[2] Department of Intelligent Systems, Jožef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia; matjaz.gams@ijs.si
* Correspondence: robert.leskovar@um.si

### Featured Application

Organizations deploying retrieval-augmented assistants can use the documented field-evaluation approach and practical recommendations to examine actual tool use, in-context user ratings, and input/output token consumption before choosing a default configuration. The approach combines turn-level random assignment with concealed configuration names, analysis of repeated, partially rated responses, and an explicit token-accounting boundary.

### Abstract

Retrieval-augmented generation (RAG) is commonly assessed on benchmarks, but its behavior under employees' workplace questions remains less clear. We compared single-pass, agentic, and hybrid RAG in a Slovenian public-sector assistant to examine user-perceived quality, actual tool use, and token consumption. Configurations were randomized per turn, and their names were not shown to users. We analyzed 701 responses from 131 accounts; 402 answers were evaluated by 73 accounts using five 1–5 rating items and a separate binary usefulness judgment. Resource use was application-recorded input plus output tokens, excluding re-ranking. Hybrid had the highest descriptive mean quality (4.15), followed by Single-pass (3.93) and Agentic (3.79), but neither dimension tests nor adjusted composite contrasts established superiority. Hybrid also had the highest mean token total among rated responses (41,143 versus 32,977 and 28,432). Agentic initiated document retrieval on 62.9% of turns; Hybrid used additional retrieval in 3.5%. Thus, the comparison showed a design–behavior gap without establishing an escalation benefit. The study provides field-evaluation evidence and practical deployment recommendations; incomplete rating coverage, configuration-specific prompts, and mixed conversation histories limit generalization and causal interpretation.

**Keywords:** retrieval-augmented generation; hybrid RAG; orchestration; human evaluation; organizational conversational agent; large language model

Check for updates

Academic Editor: Arkadiusz Gola

Received: 14 October 2026
Revised: 13 November 2026
Accepted: 16 November 2026
Published: 19 November 2026

**Copyright:** © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

## 1. Introduction

Organizational conversational agents provide natural-language access to information and require evaluation throughout deployment [1,2]. Earlier agent-design work studied natural-language scripting and semantic sentence similarity [3,4]. Organizational adoption raises knowledge-management and implementation challenges [5,6]; research on responsi-

ble artificial intelligence (AI) in human resources management also identifies governance considerations for organizational use [7].

The integration of large language models (LLMs) into conversational AI has expanded the range of possible applications in organizations. Transformer architectures underpin modern language models [8], while chatbot research provides a broader account of conversational-system designs [9]. However, their deployment in organizational environments also exposes well-documented limitations: lack of access to proprietary or domain-specific knowledge, temporal inconsistency, and the tendency to generate plausible but factually incorrect outputs (hallucinations) [10–13]. These limitations motivate attention to the risks and resource demands of large language models [14].

Retrieval-augmented generation (RAG) combines generative models with retrieved external passages, using a retriever to select evidence for answer generation [15,16]. While single-pass RAG follows a retrieve-then-generate paradigm, agentic designs allow iterative reasoning, tool use, and multi-step retrieval [17,18]. The quality and resource implications of these additional operations depend on the implemented policy and task.

Motivated by bounded rationality [19], hybrid designs can combine initial retrieval with optional additional tool use. We call this orchestration configuration *Hybrid RAG*; it is distinct from *hybrid retrieval*, which combines dense semantic and sparse lexical search within the retrieval stage. Whether this flexibility improves quality or reduces resource use remains an empirical question.

Most evidence on these design choices comes from benchmark datasets, controlled conversations, or ratings by external evaluators [20–23]. Such evidence is essential for isolating mechanisms, but it leaves open what happens when the same configurations meet the questions, follow-ups, and expectations of employees during ordinary work. In particular, an orchestration design only *permits* additional retrieval; whether a language model actually uses that option, how often, and at what token cost can be observed only in deployment. An organization choosing among RAG configurations therefore needs field evidence on three questions: how users judge the answers, how the configurations actually behave, and what they consume.

We address this need with a field comparison of three orchestration-and-prompt configurations in an assistant used by employees of a Slovenian public institution. For each turn, the configuration was selected at random without its name being shown to users, and all configurations shared the same generative pre-trained transformer (GPT) model, document corpus, and retrieval stack. Archival linkage yields a primary employee-use sample of 701 responses across 408 threads, including 402 complete five-item evaluations and usefulness judgments. Participant accounts provide the clustering unit for exploratory inference.

This study addresses the following research questions:

1. How do observed response-level human ratings and binary judgments differ among the three configurations?
2. How often do the configurations actually invoke document retrieval during use, and does observed tool use match the behavior each configuration was designed to permit?
3. How do measured input and output tokens, excluding re-ranking, compare across configurations, and what descriptive quality–token trade-offs are observed?

The study makes four contributions:

- **Field evidence from real use.** It provides a configuration-concealed, turn-randomized comparison of single-pass, agentic, and hybrid RAG within one deployed stack, rated by the employees who asked the questions, rather than by benchmark labels or external judges.

- **A design–behavior gap.** Tool logs show that the configurations behaved differently from what their labels suggest: the agentic model made no new document-search call on 37% of its turns, and the hybrid configuration used its additional-retrieval option in only 3.5% of responses. Configuration names are therefore insufficient descriptions of deployed RAG behavior.
- **Resource evidence.** It pairs user-rated quality with measured input/output tokens and shows that input context, not generated text, accounts for more than 98% of measured tokens in every configuration.
- **A documented field-evaluation approach and practical guidance.** It describes the analysis of repeated, partially rated in situ feedback using participant-cluster inference, bootstrap intervals, missingness bounds, and sensitivity checks, and derives practical recommendations for organizational RAG deployments (Section 5.3).

The analysis does not identify an optimal escalation policy or isolate orchestration logic from prompt differences, conversation history, and incomplete rating coverage.

The remainder of this paper is organized as follows. Section 2 reviews related work. Section 3 describes the configurations, study sample, and evaluation approach. Section 4 reports the results. Section 5 discusses their interpretation and limitations, and Section 6 concludes the paper.

## 2. Related Work and Research Gap

### 2.1. Organizational RAG and Orchestration

Organizational conversational agents are increasingly used to provide natural-language access to internal knowledge resources, operational procedures, policies, and domain-specific documentation. In such settings, conversational AI systems are expected not only to generate fluent responses but also to provide answers that are accurate, grounded, traceable, and appropriate for organizational decision-support contexts [1,2,24]. These requirements distinguish organizational conversational agents from open-domain chatbots because the relevant knowledge is often proprietary, locally maintained, and embedded in specific work practices.

Large language models have expanded the capabilities of conversational agents by improving language understanding, response generation, and contextual interaction. However, standalone LLMs are limited in organizational environments because they may lack access to internal documents, may not reflect recent organizational changes, and may generate plausible but unsupported responses [10–13]. These limitations are particularly consequential when generated responses are used to support operational decisions or interpret internal procedures.

Retrieval-augmented generation addresses these limitations by combining generative models with external knowledge retrieval. Instead of relying solely on model parameters, RAG systems retrieve relevant passages from an external corpus and condition answer generation on the retrieved evidence [15,16]. This approach conditions answers on domain-specific evidence rather than relying exclusively on model parameters. Subsequent work includes retrieval-augmented few-shot learning with Atlas [25], broader retrieval-generation methods surveyed by Gao et al. and Huang and Huang [26,27], and precise retrieval through the SAGE framework, which combines semantic segmentation with adaptive context selection [28].

Applied implementations illustrate these architectural choices in specific domains. Papageorgiou et al. [29] present a modular e-government assistant built with Haystack, combining source-specific RAG pipelines, conversational memory, and agent-based tool selection. Their case studies emphasize reproducibility and the integration of multiple information sources. Burbano et al. [22] describe a grocery-store chatbot that combines a

locally hosted LLM, vector retrieval, and structured database access to answer operational queries. These studies provide concrete examples of how retrieval and generation can be integrated into domain-specific applications; the present study compares orchestration-and-prompt configurations within a shared model, corpus, and retrieval environment.

In practice, however, RAG is not a single fixed architecture. Production systems differ in how retrieval is invoked, how queries are reformulated, how retrieved evidence is assessed, and whether the system performs additional reasoning or retrieval before generating an answer. These choices are part of the orchestration layer. For organizational deployments, orchestration is especially important because response quality must be balanced against latency, computational cost, explainability, and behavioral predictability.

Table 1 summarizes how the present study is positioned relative to the main related research areas.

**Table 1.** Summary of related research areas and positioning of the present study.

| Research Area | Representative Prior Work | Positioning of the Present Study |
| --- | --- | --- |
| RAG foundations and retrieval-augmented generation | [15,16,25]; RAG surveys [26,27] | Examines RAG not as a single architecture but as a set of orchestration choices in a deployed organizational conversational agent. |
| Retrieval extensions and tool use | Reasoning and Acting (ReAct) [17]; Toolformer [18]; knowledge graph-extended retrieval-augmented generation (KG-RAG) [30]; query expansion and knowledge-graph prompting [31,32]; agentic knowledge curation versus full-context retrieval [23] | Compares agentic, single-pass, and hybrid configurations within a shared deployment and reports how often the agentic and hybrid tool options were actually used. |
| Human-centered evaluation of conversational AI and RAG | Healthcare evaluation framework [33]; dialogue evaluation [34]; user expectations and workplace appropriation [35,36]; attributed question answering (QA) [37]; evidence-span extraction [38] | Uses step-level human evaluation in an organizational deployment, covering accuracy, information coverage, clarity, source linkage, and expression appropriateness. |
| Applied RAG system architectures | Modular e-government assistant [29]; retail RAG with structured database access [22] | Compares three orchestration-and-prompt configurations within a shared system stack, using field-user ratings and measured input/output tokens. |
| Field evaluation and organizational deployment | Organizational conversational-agent work [1,2,24]; applied deployment studies [39,40] | Evaluates natural user queries and response-level feedback in multi-turn organizational conversations. |
| Quality–cost trade-offs and bounded autonomy | Long-context and RAG trade-off analysis [41]; RAG evaluation and evaluation-framework work [21,42]; bounded rationality [19] | Adds a descriptive comparison of measured input-plus-output tokens and human-rated quality; the benefit of conditional resource allocation is not identified. |

### 2.2. Human Evaluation of Conversational AI and RAG

Evaluation of RAG systems commonly includes retrieval-oriented and generation-oriented measures, such as relevance, faithfulness, groundedness, answer correctness, and citation quality [21,26,42]. These measures are useful for controlled benchmarking, but they only partially capture the performance of conversational agents deployed in organizations. In real organizational use, response quality is judged not only by factual correctness but also by whether the response is complete enough for the task, understandable to the user, appropriately phrased, and clearly connected to source material.

Human-centered evaluation is therefore central for applied conversational AI. Tam et al.'s healthcare-focused literature review organizes human evaluation along multiple dimensions, including information quality, reasoning, expression, and trust-related factors [33]. It motivates multidimensional assessment here, without validating this study's organizational feedback

instrument. Dialogue-system evaluation research similarly shows that subjective user judgments remain important because conversational quality is shaped by coherence, clarity, appropriateness, and perceived usefulness, not only by automatic scores [34]. Bohnet et al. formulate attributed question answering and evaluate whether generated answers are supported by attributed evidence [37]. In a distinct task, Dimitriadis and Tsoumakas use weakly supervised evidence-span extraction to improve yes/no question answering [38]; this is not a general validation of citation quality in conversational RAG.

In an applied retail setting, Burbano et al. [22] combine assessment using G-Eval, a language-model-based text-evaluation framework, with manual ratings of faithfulness, relevance, context precision, and context recall. Their evaluation comprises 52 controlled conversations with 312 query–response pairs, and their limitations identify the absence of real-customer validation. This distinction between evaluator assessment and feedback from users during operational use motivates the present study's response-level field evaluation, which also addresses clarity and expression appropriateness.

Martín et al. [23] compare agentic knowledge curation with full-context retrieval over the technical documentation of a Spanish legal semantic-search system, using blinded scoring by two independent external reviewers. Overall accuracy did not differ significantly between the two paradigms, yet the paradigms failed in different ways: full-context retrieval produced false negatives associated with the lost-in-the-middle effect [43], whereas agentic curation lost textual fidelity on queries requiring exact reproduction of formulas or sequences. Their study shows that a comparison of retrieval paradigms can be informative even without a significant difference in aggregate quality, because the mechanisms by which configurations succeed or fail differ. The present study takes a complementary perspective: instead of independent reviewers scoring controlled queries, the employees who posed naturally occurring questions rate the answers, and tool logs show how each configuration behaved.

Organizational use adds further evaluation challenges. Users interact with systems across multi-turn threads, ask follow-up questions, and adapt their usage patterns over time. Workplace research identifies heterogeneous patterns of chatbot appropriation based on users' modes of interaction and understanding of the technology [36]. Prior work on imperfect intelligent systems also shows that user acceptance depends on how system limitations are experienced and communicated [35]. Applied healthcare and cognitive-assessment studies provide domain-specific examples of usage analysis and user acceptability assessment [39,40].

These findings suggest that organizational RAG systems should be evaluated under realistic use conditions and with metrics that reflect user-perceived quality. Step-level evaluation is especially relevant because different turns within the same conversation may vary substantially in complexity, context dependence, and retrieval requirements. Treating an entire conversation as a single evaluation unit can obscure these differences and make it difficult to determine how orchestration strategies perform at specific interaction points.

### 2.3. Quality–Cost Trade-Offs and Bounded Autonomy

A further challenge in deployed RAG systems is the trade-off between response quality and operational cost. More retrieval, longer prompts, additional reasoning steps, and repeated re-ranking may improve some responses, but they also increase token consumption, latency, and infrastructure load. RAG surveys increasingly emphasize that practical deployments must account for efficiency, robustness, maintainability, and cost in addition to answer quality [26,27,42]. This is particularly important in organizational settings, where conversational agents may serve many routine queries and must remain economically sustainable.

The quality–cost trade-off is closely related to the degree of autonomy granted to the orchestration layer. Fully agentic systems can pursue additional evidence and reasoning paths, but the benefits of doing so are query-dependent. For simple or well-supported questions, additional reasoning may introduce unnecessary verbosity or instability. For ambiguous or under-supported questions, however, additional retrieval may improve completeness and grounding. This motivates orchestration strategies that adapt reasoning effort to task demands.

Recent work comparing RAG, fine-tuning, and hybrid optimization similarly emphasizes that RAG-based approaches must be evaluated as accuracy–efficiency trade-offs rather than as purely accuracy-maximizing designs [44].

The idea of adapting effort to constraints is consistent with bounded rationality, where decision-making is understood as resource-constrained rather than exhaustive [19]. In the context of RAG, bounded autonomy means that the system should not always maximize retrieval depth or agentic reasoning. Instead, it should apply additional reasoning only when the expected improvement in answer quality justifies the added cost and complexity. Hybrid RAG is motivated by this principle, but an implementation must be evaluated using logged decisions and counterfactual comparisons before resource savings or effective escalation can be established.

### *2.4. Research Gap and Positioning of This Study*

The literature establishes RAG as a practical approach for grounding LLM-based conversational agents and shows that agentic retrieval can improve flexibility in complex information-seeking tasks. The e-government case studies of Papageorgiou et al. [29], the controlled retail evaluation of Burbano et al. [22], and the reviewer-scored comparison of Martín et al. [23] demonstrate domain-specific integration and paradigm comparison. However, none compares single-pass, agentic, and hybrid configurations within a shared system stack using ratings from users in operational use, logged tool behavior, and token consumption. This comparison defines the specific gap addressed here.

Adaptive retrieval already has close precedents. Adaptive-RAG selects among no-retrieval, single-step, and iterative strategies using question complexity [20]. The Multi-arm Bandit-based framework for Adaptive Retrieval-Augmented Generation (MBA-RAG) learns strategy selection while considering quality and efficiency [45]. RAGate is particularly relevant to conversational use: it uses conversational context and human judgments to decide whether a turn requires external retrieval [46]. Stop-RAG studies value-based stopping in iterative retrieval [47], while the Recoverability-Aware Selective Escalation Router (RASER) preprint studies recoverability-aware escalation after one-shot retrieval [48]. These methods address decisions that the present comparison does not isolate or optimize.

These methods evaluate routing or stopping in their specified task settings, including benchmark QA and conversational retrieval decisions. Benchmark answer scores and retrieval labels characterize those settings; they do not directly estimate the frequency of model-directed tool activation in a different deployment. The implementation studied here delegates optional retrieval decisions to the language model through prompted tool use, without a separately trained router. This design assumes that the model exercises the option when it is needed. Whether this assumption holds in organizational use, and whether users perceive any resulting difference in quality, has received little empirical attention.

Workplace prompts can contain institution-specific shorthand, incomplete questions, and follow-ups whose meaning depends on earlier turns. Previously retrieved passages may remain in the conversation history, so a new question need not require a new search. Tool activation also depends on the deployed prompt, the evidence already supplied, and whether retrieval is enforced or left to the model. Consequently, activation rates need not

transfer from a benchmark's question and context distribution to ordinary workplace traffic, even when benchmark answer accuracy is high. This motivates measuring tool use directly in the field rather than treating benchmark accuracy as a prediction of retrieval frequency.

The contribution here is therefore empirical evidence on what happens when three established configurations are used by employees. It combines response-level perceptions, participant-cluster uncertainty, observed retrieval activity, and measured input/output tokens within a defined accounting boundary. It does not claim a new adaptive-retrieval principle, a trained router, or an isolated causal benefit of a sufficiency check. The shared stack reduces some implementation differences, while distinct prompts and mixed conversation histories remain part of the comparison.

## 3. Materials and Methods

The study procedure comprised six steps.

1. The Slovenian assistant was deployed for voluntary employee use after the study information and consent form were provided.
2. One configuration was selected uniformly at random for each user turn, without its name being shown to the employee.
3. The selected configuration ran on the shared stack, retaining conversation history and recording answer and tool activity.
4. Five rating items and a binary usefulness judgment were collected before another turn in the same thread; users could leave or start another thread without rating.
5. The 720 exported responses were linked to archived accounts and runs. Excluding the researcher's 19 responses left 701 primary responses, including 402 complete evaluations.
6. Quality, retrieval and token use were summarized; uncertainty was assessed by account clustering and whole-account bootstrap, with the stated sensitivity analyses.

The following subsections specify the deployment, configurations, analytic sample, instrument, token boundary and statistical methods.

### *3.1. Deployment Context and Study Population*

The assistant supported employees of a Slovenian public institution in consulting organizational documents. The application and system prompts used Slovenian. Employees supplied naturally occurring questions, including follow-up questions within persistent conversations, rather than answering an assigned benchmark. The matched primary-sample responses were recorded for 32 days.

Participants received an information notice and consent form before use. Participation was voluntary, but submitting feedback was required to continue within the same conversational thread. The same five-item feedback interface and binary usefulness buttons were used across configurations. The analysis uses account identifiers as the participant proxy; it does not independently verify that every account corresponds to one distinct eligible employee.

### *3.2. Configurations and Recorded Implementation Evidence*

Single-pass RAG performs initial retrieval before generation. Agentic RAG exposes retrieval tools to the language model without obligatory initial retrieval. Hybrid RAG combines initial retrieval with access to additional retrieval tools. A time tool is available independently of document retrieval. The configurations have distinct system prompts; the Agentic prompt instructs the model to use document tools, while the Hybrid prompt makes this instruction conditional on information not already supplied by the vector store. Both

permit the model to finish without another retrieval. No separate numeric sufficiency score or learned stopping controller is evidenced by the implementation.

Each configuration combines orchestration with instructions matching its available context and tools. The prompts share role, language, source-use and citation requirements, but differ in tool-use instructions. The comparison concerns these deployable packages rather than orchestration logic alone. Appendix D supplies English translations of the system-prompt templates.

*Hybrid RAG* is an orchestration configuration: it combines an initial retrieve-then-generate path with optional model-directed retrieval during answer generation. It therefore specifies *when and whether further retrieval occurs*. *Hybrid retrieval* instead specifies *how documents are retrieved within a search call*, by combining dense semantic retrieval with sparse lexical retrieval, for example with Best Matching 25 (BM25). These are separate design choices: hybrid retrieval can serve Single-pass, Agentic, or Hybrid RAG, and is not a feature exclusive to the Hybrid RAG configuration. The three configurations share the retrieval stack; this study compares orchestration and prompts, not dense-only, sparse-only, and hybrid retrieval strategies. The recorded dense/sparse settings and their provenance are reported separately in Table 2.

The deployed system used GPT-4o, text-embedding-3-large, and Qdrant [49–51]. Exact historical provider model snapshots and a frozen index manifest were not recovered. Table 2 distinguishes persisted thread metadata from source-code defaults and limits. Recorded configuration values are snapshots at thread persistence, rather than a per-call settings history; they cannot prove that every request used unchanged values.

**Table 2.** Implementation applicability and provenance. S, A and H denote Single-pass, Agentic and Hybrid. Retrieval settings apply whenever that configuration retrieves. Thread metadata are persistence snapshots, not per-call runtime logs; source-code defaults and limits do not establish per-call settings.

| Component | Applies to | Recorded Value or Behavior | Provenance |
|---|---|---|---|
| Generation setting | S/A/H | Temperature 0.15 in 407 of 408 primary threads; one thread has no settings metadata. | Persisted thread metadata |
| Retrieval settings | S/A/H | Final top-*k* 12, score threshold 0.65, sparse flag true in the same 407 thread snapshots. | Persisted thread metadata |
| Tool-step display | S/A/H | Display setting true in the same 407 thread snapshots; tool steps were shown to employees by default (Section 3.2). | Persisted thread metadata; source-code configuration |
| Retrieval path | S/H initial; A/H optional | S/H invoke retrieval before generation; A/H can call the same document-search/re-ranking implementation through a model-selected tool. | Source code; recorded tool steps |
| Candidate search | S/A/H | Default: up to 100 candidate passages before re-ranking. Dense embeddings and a Qdrant/BM25 sparse encoder are configured. | Source-code default |
| Re-ranking | S/A/H | Chat-model re-ranking: temperature 0, output cap 500, up to 100 snippets of 700 characters each; retain the requested final *k*. | Source-code default |
| Generation limits | S/A/H | At most five tool-loop iterations and 10,000 output tokens per answer-model call; neither bounds combined input-plus-output totals. | Source-code limit |
| Conversation history | S/A/H | Trim preceding history to 10 messages and a 100,000-token content threshold before adding the current question and retrieval context. | Source-code limit |

The application implements turn-level selection by a uniform random choice over three profiles. The selected configuration name was concealed from employees. However, the application was configured to display tool steps by default. When a document search or time lookup ran, the interface showed a collapsible step labeled with the tool name above the answer; expanding it showed the tool input and returned output. Single-pass

and Hybrid responses therefore always showed a document-search step, whereas Agentic responses showed one only when the model called the tool. The presence or absence of a step and differences in the final answer could provide indirect cues to the configuration, and employees' ability to identify configurations was not tested. We knew each response's configuration when analyzing the records. The configuration could change between turns within the same conversation.

Earlier responses remain in the history when the current configuration changes. The comparison consequently concerns current configurations in mixed histories, not the long-run effect of using one configuration for every turn. Appendix D reproduces English translations of the three system-prompt templates and the tool definitions exposed to the model; Appendix E describes record linkage, the researcher-account exclusion, and the computational checks.

### 3.3. Analytic Sample

The primary sample contains 701 responses from 131 accounts in 408 threads. There are 402 complete five-item ratings and binary usefulness judgments from 73 accounts in 176 threads: 122 Agentic, 144 Single-pass, and 136 Hybrid ratings. The remaining 299 responses have no ratings; the feedback requirement applied to continuing a thread, rather than to exiting or starting another one. Among rated accounts, the median contribution is 4 ratings and the maximum is 24. Forty-eight of the one hundred thirty-one accounts contributed responses under all three configurations. Among the 73 accounts providing complete evaluations, 24 rated one configuration, 16 rated two, and 33 rated all three. Agentic, Single-pass and Hybrid received ratings from 48, 57 and 50 distinct accounts, respectively; these sets overlap. The rated sample is therefore not a balanced three-condition within-person design. All primary responses enter activity and measured-token summaries; only complete ratings enter quality and usefulness comparisons. No missing ratings are imputed.

### 3.4. Human Evaluation Instrument

The author-designed interface collected five items on a 1–5 response scale. Multidimensional evaluation is motivated by dialogue-system and human-evaluation literature [33,34]; these references do not validate the specific instrument. Table 3 reproduces the Slovenian wording and its translation. The interface gave a common anchor sequence: *1–napačno, 2–zadostno, 3–dobro, 4–v redu, 5–popolno* (1–incorrect, 2–sufficient, 3–good, 4–okay, 5–perfect). These generic anchors do not define separate objective criteria for every dimension.

**Table 3.** Response-level feedback instrument. English wording is a translation, not an independently administered questionnaire.

| Analysis Label | Slovenian Interface | English Translation |
| --- | --- | --- |
| Accuracy | Dejanska točnost odgovora (1 = napačno, 5 = popolno) | Factual accuracy of the answer (1 = incorrect, 5 = perfect) |
| Coverage | Pokritost informacije | Information coverage |
| Clarity | Jasnost in razumljivost | Clarity and understandability |
| Source linkage | Povezava z virom | Connection with the source |
| Expression | Ustreznost izražanja | Appropriateness of expression |
| Binary judgment | Uporabno/Ni uporabno | Useful/Not useful |

Accuracy is the user's perceived factual accuracy, not independent verification of correctness. Source linkage likewise measures a perception of connection to evidence, rather

than an audited citation-support score. The ratings describe how users perceived answers in their own tasks; they do not estimate independently verified factual-error rates, citation entailment or successful task completion. Favorable scores therefore support conclusions about perceived quality, not audited correctness. For each rated response $i$, the descriptive composite is

$$Q_i = \frac{1}{5}\sum_{j=1}^{5} r_{ij}. \quad (1)$$

Equal weights define a descriptive index; they do not establish a validated latent scale. Ordinal ratings are coded numerically when estimating means, with full distributions reported in the Appendix A.2. The binary outcome is perceived usefulness, drawn directly from the recorded button response rather than a threshold on $Q_i$. The interface required all five items for a complete submission, and users had to submit feedback before continuing within the same conversational thread. However, users could exit the application, terminate the thread, or start a new thread without rating the preceding answer. The resulting incomplete coverage and its possible selection mechanisms are discussed in Section 5.5. No inter-rater reliability estimate is available because these are users' assessments of their own responses, not a crossed panel of independent judges. Feedback was used for offline evaluation rather than online model adaptation.

### 3.5. Measured Input and Output Tokens

Resource use was assessed using input and output tokens measured by the application. For response $i$, the total is

$$T_i = I_i + O_i, \quad (2)$$

where $I_i$ and $O_i$ are the recorded input and output token counts, respectively. Table 4 reports the two components and their sum for all responses and for the rated subset.

**Table 4.** Measured input and output tokens per response, excluding re-ranking. Means are rounded independently; total equals input plus output before rounding. See Section 3.5.

| Configuration | Subset | *n* | Mean Input | Mean Output | Mean Total |
|---|---|---|---|---|---|
| Agentic RAG | All | 210 | 26,423 | 467 | 26,890 |
| Agentic RAG | Rated | 122 | 27,971 | 462 | 28,432 |
| Single-pass RAG | All | 237 | 28,909 | 438 | 29,347 |
| Single-pass RAG | Rated | 144 | 32,525 | 453 | 32,977 |
| Hybrid RAG | All | 254 | 36,940 | 477 | 37,417 |
| Hybrid RAG | Rated | 136 | 40,660 | 482 | 41,143 |

The 10,000-token limit in Table 2 applies to output from an individual model call, not to the combined input-plus-output total. For example, Single-pass averages 28,909 input tokens and 438 output tokens per response, giving a combined mean of 29,347. Thus totals above 10,000 are compatible with the output limit and the Single-pass generation path in Figure 1.

Figure 2 summarizes sample reconciliation and evaluation coverage.

Re-ranking tokens are excluded. The same re-ranking procedure was used across configurations; the comparison focuses on the application's measured input/output tokens. Because retrieval frequency can differ across configurations, a shared re-ranking procedure does not necessarily imply equal omitted overhead per response. The reported totals therefore describe the included token consumption, rather than complete pipeline cost or monetary expenditure.

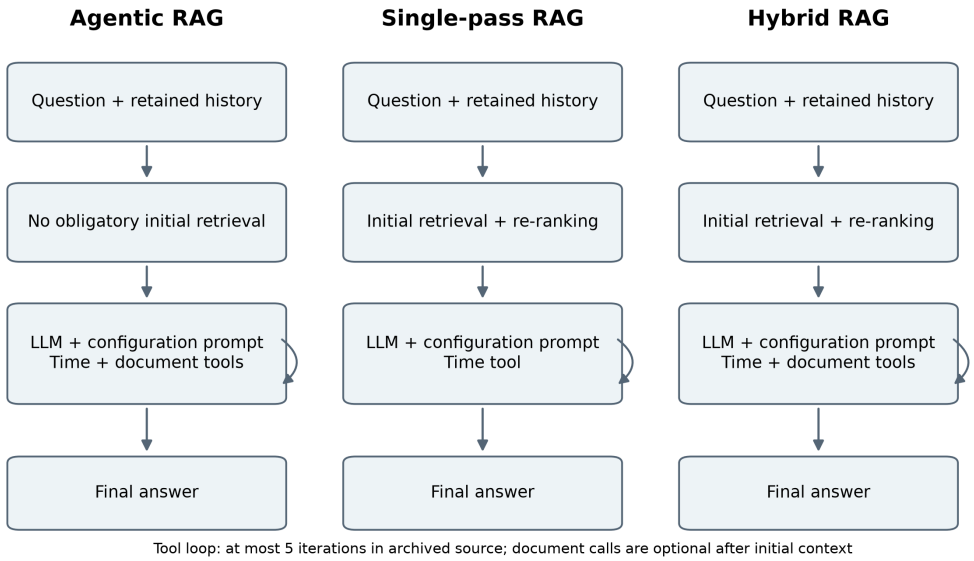


**Figure 1.** Implemented orchestration paths. Hybrid RAG denotes initial retrieval plus optional further tool use; hybrid retrieval denotes dense/sparse search within a retrieval call and can underlie any of the three paths (see Section 3.2 for terminology). Prompt wording, tool availability, and persistent history form part of each configuration; optional retrieval does not imply that a tool was called.

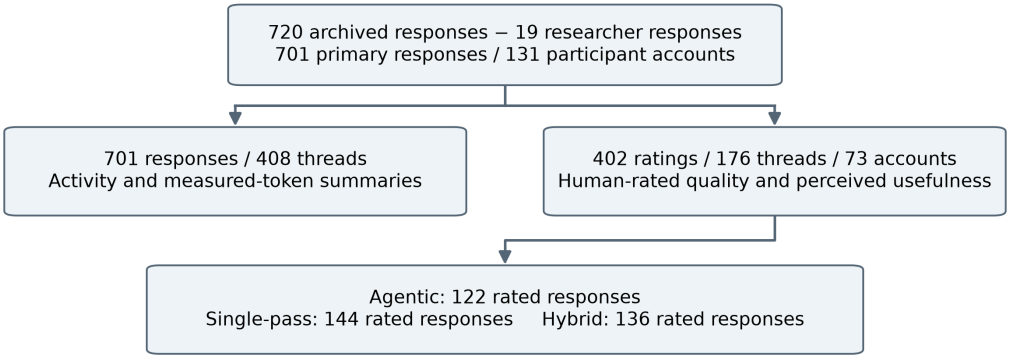


**Figure 2.** Sample reconciliation and evaluation coverage. The primary analysis excludes the researcher account; the full-export sensitivity analysis retains all 720 records.

We summarize response totals by median, first and third quartiles (Q1 and Q3), and interquartile range (IQR = Q3 − Q1), using linear interpolation at position $(n-1)p$ for quantile $p$. For account-level summaries, tokens are summed within each account and configuration. Concentration shares divide the top accounts' cumulative tokens by that configuration's total within the stated subset; the top 5% and 10% use the ceiling of the corresponding fraction of active accounts, with actual selected counts reported. First turns are identified from all archived user-message runs in each thread, including runs outside the analytic sample; other analyzed turns are follow-ups. All-primary and rated-only results are reported separately.

For account-influence sensitivity, each account in the stated subset is omitted in turn from every configuration, and the response-weighted Hybrid minus Single-pass mean token difference is recomputed. We report the minimum and maximum difference across these omissions. These are descriptive sensitivity ranges, not confidence intervals or estimates of a causal configuration effect.

### *3.6. Statistical Analysis*

The primary estimand is the response-weighted mean rating among retained responses with submitted ratings in each configuration. It is not the equally weighted mean participant experience or the mean quality of all generated responses. Configuration means and sample standard deviations (SDs) are computed from individual response composites. Repeated records are clustered by pseudonymized participant account, which also accounts for dependence among that account's threads [52]. Independence between accounts remains an assumption. The implications of shared accounts or multiple accounts per person are discussed in Section 5.5.

For each dimension, ordinary least squares fits three arm indicators without an intercept. The covariance uses the finite-sample-corrected cluster-robust estimator (CR1), applying the factor $G/(G-1)\,(N-1)/(N-3)$ to the covariance based on cluster-summed residual scores, where $G = 73$ and $N = 402$. Equality of the three means is assessed by a two-contrast Wald statistic using an approximate $F(2, 72)$ reference. Holm adjustment controls the five-dimension family at $\alpha = 0.05$ [53]. Numeric coding of the ordinal items targets mean scores, rather than an ordinal latent-variable model.

Composite contrasts use 10,000 whole-account bootstrap resamples. Each draw samples 73 accounts with replacement and retains all their rated responses and configuration labels. Percentile intervals are reported at 95% and at 98.33%, the latter using Bonferroni adjustment for three pairwise contrasts. Figure intervals for individual dimension means are pointwise 95% intervals. Binary usefulness and rating submission are separate exploratory secondary analyses with the same CR1 mean-comparison approach, using 73 and 131 accounts, respectively. There is no claim of a single familywise error guarantee over every secondary analysis.

Sensitivity analyses repeat the dimension tests with thread clustering and repeat participant-cluster analyses on all 720 exported records, including the researcher account. Missingness sensitivity bounds all-response quality by allowing each missing composite to lie anywhere in $[1, 5]$; a separate tipping-point calculation varies the unrated Hybrid mean relative to its rated mean. Neither procedure recovers missing outcomes. All analyses are exploratory. Non-significance does not establish equivalence.

## 4. Results

### *4.1. Response Quality and Rating Coverage*

Table 5 summarizes the 402 complete evaluations. Hybrid has the highest descriptive composite mean (4.15), followed by Single-pass (3.93) and Agentic (3.79). Positive usefulness judgments are 84.56%, 82.64%, and 80.33%, respectively.

**Table 5.** Composite response quality and perceived usefulness ($N = 402$). SD is computed across response-level five-item means.

| Configuration | *n* | Mean | SD | Positive/Rated | Positive (%) |
|---|---|---|---|---|---|
| Agentic RAG | 122 | 3.79 | 1.23 | 98/122 | 80.33 |
| Single-pass RAG | 144 | 3.93 | 1.23 | 119/144 | 82.64 |
| Hybrid RAG | 136 | 4.15 | 1.12 | 115/136 | 84.56 |

Complete ratings cover 402/701 responses (57.35%). Submission rates range from 53.54% to 60.76% across configurations (Table 6). The participant-cluster submission test gives $F(2, 130) = 1.1958$, $p = 0.30578$. The corresponding usefulness test gives $F(2, 72) = 0.4030$, $p = 0.66978$.

**Table 6.** Complete-rating submission in the study sample.

| Configuration | Responses | Complete Ratings | Rated (%) |
|---|---|---|---|
| Agentic RAG | 210 | 122 | 58.10 |
| Single-pass RAG | 237 | 144 | 60.76 |
| Hybrid RAG | 254 | 136 | 53.54 |

### *4.2. Dimension and Composite Comparisons*

Hybrid has the highest descriptive mean on all five dimensions (Table 7). Figure 3 shows participant-bootstrap uncertainty; Appendix A supplies complete distributions. Section 4.5 and Appendix B report thread-cluster and full-export sensitivity checks and missing-outcome analyses alongside these primary comparisons.

**Table 7.** Dimension-level results (mean ± sample SD); Agentic $n = 122$, Single-pass $n = 144$, Hybrid $n = 136$.

| Dimension | Agentic RAG | Single-Pass RAG | Hybrid RAG |
|---|---|---|---|
| Accuracy | 3.63 ± 1.43 | 3.78 ± 1.44 | 3.93 ± 1.39 |
| Information coverage | 3.43 ± 1.49 | 3.69 ± 1.45 | 3.93 ± 1.40 |
| Clarity | 4.02 ± 1.28 | 4.10 ± 1.26 | 4.35 ± 1.04 |
| Source linkage | 3.69 ± 1.52 | 3.83 ± 1.46 | 4.06 ± 1.38 |
| Expression appropriateness | 4.15 ± 1.24 | 4.26 ± 1.19 | 4.46 ± 1.02 |

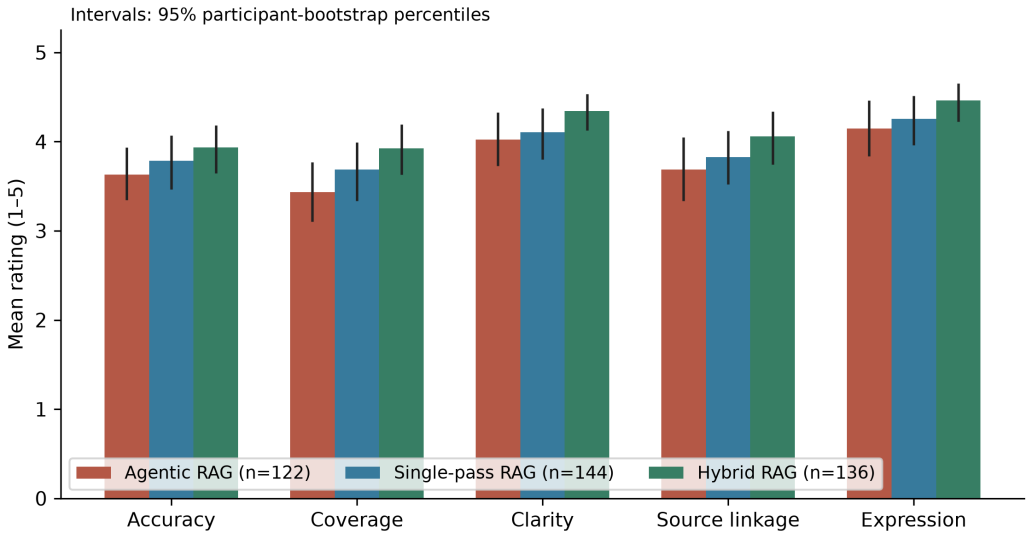


**Figure 3.** Mean ratings for Agentic ($n = 122$), Single-pass ($n = 144$), and Hybrid ($n = 136$). Lines are pointwise 95% whole-participant bootstrap intervals, accounting for repeated responses across each account's threads.

No dimension comparison rejects equality of configuration means after Holm adjustment. Information coverage has the smallest unadjusted probability ($p = 0.01862$), but its adjusted probability is $p_{\text{Holm}} = 0.09310$ (Table 8). All three multiplicity-adjusted composite intervals include zero (Table 9). The pointwise 95% Hybrid–Agentic interval excludes zero, whereas its three-contrast adjusted interval does not.

**Table 8.** Exploratory dimension tests using participant-cluster covariance, $F(2, 72)$, and Holm adjustment across five outcomes.

| Dimension | $F$ | $p$ | $p_{\text{Holm}}$ |
|---|---|---|---|
| Accuracy | 1.5018 | 0.22962 | 0.49804 |
| Information coverage | 4.2123 | 0.01862 | 0.09310 |
| Clarity | 2.0810 | 0.13225 | 0.49804 |
| Source linkage | 2.1448 | 0.12451 | 0.49804 |
| Expression appropriateness | 2.0447 | 0.13686 | 0.49804 |

**Table 9.** Composite mean differences with whole-participant bootstrap confidence intervals (CIs). The 98.33% intervals adjust for three contrasts by Bonferroni.

| Contrast | Difference | 95% CI | 98.33% CI |
|---|---|---|---|
| Hybrid-Single-pass | 0.214 | [−0.033, 0.463] | [−0.086, 0.520] |
| Hybrid-Agentic | 0.360 | [0.043, 0.656] | [−0.030, 0.720] |
| Single-pass-Agentic | 0.147 | [−0.128, 0.414] | [−0.193, 0.480] |

### *4.3. Observed Retrieval Behavior: Design Versus Use*

All Single-pass and Hybrid responses have an initial document-tool step (Table 10). Agentic records document retrieval for 132/210 responses (62.86%) and no document-tool call for 78 (37.14%). Additional retrieval occurs in 9/254 Hybrid responses (3.54%); 4 of those responses are rated.

**Table 10.** Retrieval activity in the primary sample. Counts refer to recorded document-tool steps, not successful retrieval or causal escalation benefit.

| Configuration | Responses | Any Retrieval | Retrieval Steps |
|---|---|---|---|
| Agentic RAG | 210 | 132 | 132 |
| Single-pass RAG | 237 | 237 | 237 |
| Hybrid RAG | 254 | 254 | 263 |

Table 11 separates current-turn document searches by original thread position. Agentic made a new search on 98/112 first turns (87.5%) and 34/98 follow-ups (34.7%). Of its 78 responses without a new document-search call, 64 were follow-ups.

**Table 11.** Agentic document-search activity by original thread turn among all 210 primary Agentic responses. A new search denotes a recorded document-tool call on the current turn, not verified successful retrieval. First-turn classification uses the same archived chronology as the token summaries.

| Turn | Responses | New Search | No New Search | New Search (%) |
|---|---|---|---|---|
| First | 112 | 98 | 14 | 87.5 |
| Follow-up | 98 | 34 | 64 | 34.7 |

Table 12 describes rated quality by observed retrieval behavior. Agentic responses without a document-tool call have a higher descriptive mean (3.97) than those with retrieval (3.65). The four rated Hybrid responses with additional retrieval have a mean composite of 1.85, and one of them was judged useful, compared with 4.22 for the 132 rated Hybrid responses without additional retrieval. No rated Hybrid response with additional retrieval received a composite above 3.0. These are descriptive, model-selected and non-randomized subgroups; the comparison does not estimate the effect of retrieving. Selection and causal interpretation are addressed in Section 5.1.

**Table 12.** Rated quality and measured tokens by observed retrieval behavior (descriptive; subgroups are model-selected and not randomized). Quality is the mean composite and tokens the mean input-plus-output total among rated responses, excluding re-ranking.

| Configuration | Observed Behavior | Responses | Rated | Quality | Tokens |
|---|---|---|---|---|---|
| Agentic RAG | No new document search | 78 | 52 | 3.97 | 29,452 |
| Agentic RAG | New document search | 132 | 70 | 3.65 | 27,675 |
| Hybrid RAG | Initial retrieval only | 245 | 132 | 4.22 | 41,025 |
| Hybrid RAG | Additional retrieval | 9 | 4 | 1.85 | 45,024 |

### *4.4. Measured Tokens*

Hybrid has the highest mean measured input-plus-output token total both in the primary sample and among rated responses (Table 4). Among rated responses, the means are 41,143 for Hybrid, 32,977 for Single-pass, and 28,432 for Agentic. Among all primary responses, they are 37,417, 29,347, and 26,890, respectively. Figure 4 pairs rated-response quality and measured input-plus-output tokens from the same subset; re-ranking is excluded.

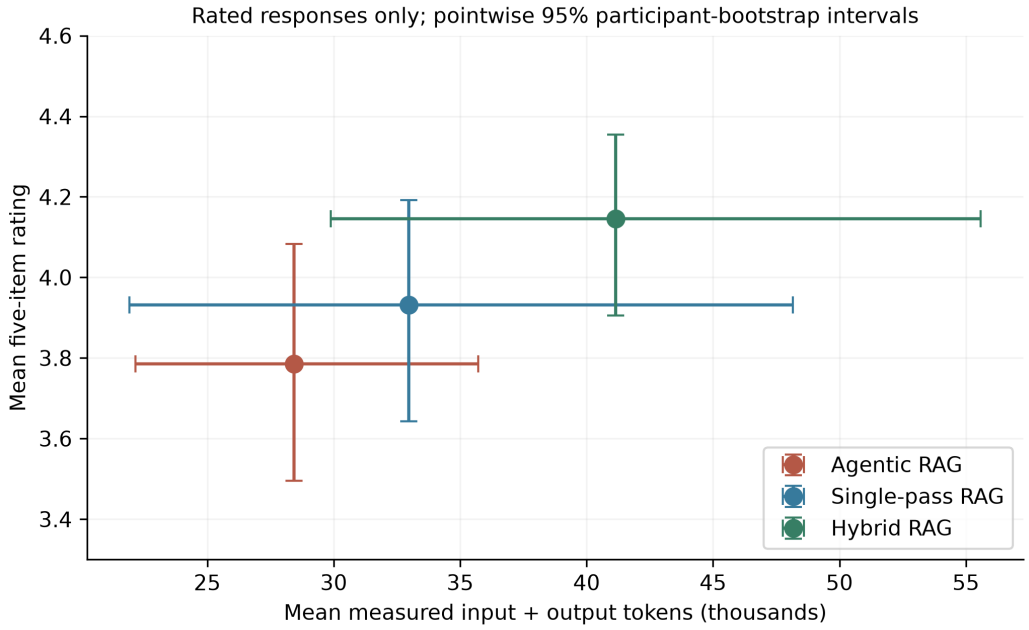


**Figure 4.** Configuration means for the same 402 rated responses. Horizontal and vertical bars are separate pointwise 95% participant-bootstrap intervals for measured input-plus-output tokens and composite quality. They quantify sampling uncertainty within the stated accounting boundary; they are not a simultaneous confidence region or a validated routing policy. Re-ranking tokens are excluded.

Input tokens account for more than 98% of the measured totals in every configuration; mean output is below 500 tokens per response (Table 4).

Table 13 reports the response-level distributions. Among rated responses, medians are 15,990.5 tokens for Agentic, 19,942.5 for Single-pass, and 27,168.5 for Hybrid; their IQRs are 22,992.25, 25,263 and 44,636.25 tokens, respectively. Appendix C reports cumulative account consumption and concentration (Table A8) and original-turn summaries (Table A9).

**Table 13.** Response-level input-plus-output token distributions, excluding re-ranking. All comprises 701 primary responses; Rated comprises the 402 complete evaluations. Quartiles use linear interpolation; IQR is Q3 minus Q1. Interpolated fractions are retained.

| Configuration | Subset | *n* | Median | Q1 | Q3 | IQR |
|---|---|---|---|---|---|---|
| Agentic | All | 210 | 14,189 | 10,332 | 27,759 | 17,427 |
| Single-pass | All | 237 | 16,959 | 10,755 | 34,958 | 24,203 |
| Hybrid | All | 254 | 18,519.5 | 11,787.5 | 41,800.25 | 30,012.75 |
| Agentic | Rated | 122 | 15,990.5 | 10,698.75 | 33,691 | 22,992.25 |
| Single-pass | Rated | 144 | 19,942.5 | 12,729.5 | 37,992.5 | 25,263 |
| Hybrid | Rated | 136 | 27,168.5 | 12,375.5 | 57,011.75 | 44,636.25 |

The Hybrid minus Single-pass mean token difference remains positive in every single-account omission (Table 14). Across all 131 omissions in the primary sample it ranges from 6137 to 8699 tokens; across the 73 rated-account omissions it ranges from 7161 to 9107 tokens.

**Table 14.** Account-influence sensitivity of the Hybrid minus Single-pass mean input-plus-output token difference, excluding re-ranking. Each account is omitted in turn from every configuration in the stated subset; remaining responses retain response weighting. Ranges summarize those omissions, not confidence intervals or causal effects. Token differences are rounded to whole tokens.

| Subset | Responses | Accounts | Original Gap | Omission Gap Range |
|---|---|---|---|---|
| All | 701 | 131 | 8070 | 6137–8699 |
| Rated | 402 | 73 | 8165 | 7161–9107 |

### *4.5. Sample, Dependence, and Missingness Sensitivity*

Neither thread-cluster analysis of the primary sample nor participant-cluster analysis retaining all 720 exported records yields a significant dimension comparison after its five-outcome Holm adjustment (Appendix B). The full-export composite means remain 4.14 for Hybrid, 3.97 for Single-pass, and 3.84 for Agentic. The descriptive ordering is retained when the researcher account is included.

Allowing the 299 missing primary-sample composites to take any value from 1 to 5 produces all-response mean bounds of [2.618, 4.294] for Agentic, [2.781, 4.351] for Single-pass, and [2.684, 4.543] for Hybrid. These are identification bounds rather than confidence intervals. If unrated Single-pass responses have their rated arm's mean, an unrated Hybrid mean approximately 0.460 points below its rated arm's mean would eliminate the full-sample Hybrid–Single-pass difference. This is an assumed tipping point, not an estimate of missing quality.

## 5. Discussion

### *5.1. What the Field Comparison Establishes*

The study provides a traceable comparison of implemented configurations in organizational use, and three findings stand out.

First, user-rated quality shows a consistent descriptive ordering: Hybrid leads Single-pass, which leads Agentic, on the composite and on every one of the five dimensions. The ordering is unchanged when the researcher account is retained, but no dimension comparison passes Holm adjustment under either participant or thread clustering, and all multiplicity-adjusted composite intervals include zero. The broad intervals permit potentially useful differences as well as small differences in the opposite direction, so failure to reject is not evidence of equivalence. Similar outcomes have been reported in controlled settings: Martín et al. [23] found no significant difference in overall accuracy between agentic curation and full-context retrieval while observing distinct failure patterns. The

five correlated ratings are not five independent replications of the ordering. Similar rating-submission rates do not establish representative feedback, and non-significant usefulness tests do not establish equal performance. The observed quality ordering concerns rated responses rather than an identified ordering for all generated answers. This study did not establish reliable differences in user-perceived quality. Observed tool behavior provides complementary evidence about how the configurations operated.

Second, recorded tool activity reveals a gap between design and behavior that configuration names conceal. The agentic model made no new document-search call on more than a third of its turns despite instructions to always use tools and to search whenever institutional knowledge was needed, and the hybrid model almost never used its option to retrieve further. This is the study's clearest result because it does not depend on rating coverage: it concerns all 701 responses. The designed decision mechanism, namely the language model's own judgment of whether further evidence is needed, therefore often produced no new search in Agentic and rarely added retrieval in Hybrid; the records do not show whether these decisions were appropriate. Consequently, the mean-rating differences cannot be interpreted as a measured benefit of repeated retrieval or an effective sufficiency gate; Hybrid behaved mostly as obligatory initial retrieval under a prompt that also permitted tools. The two configurations with obligatory initial retrieval were not rated lower than the configuration that left retrieval to the model, and the few escalations occurred in poorly rated answers (Table 12); this pattern is compatible with obligatory initial retrieval as a default for document-grounded organizational questions, without establishing its superiority. Several mechanisms could account for low tool activation. The Agentic prompt asks the model to search the knowledge base *whenever* it needs knowledge about the institution, so the instruction itself leaves the model to judge whether such knowledge is needed. The application programming interface (API) also permits automatic tool choice; the instruction is therefore not an execution constraint. The Hybrid prompt makes further search conditional on evidence not already supplied, so initial retrieval may already satisfy the model's apparent information need. Conversation history can retain earlier passages, particularly for short follow-up questions. Institution-specific wording, question difficulty and corpus coverage could also affect the decision. These are plausible explanations rather than demonstrated causes: the records show whether tools were called, not why the model made each choice, and no prompt or context ablation was performed.

The much lower new-search rate on Agentic follow-ups than on first turns (Table 11) is compatible with reuse of evidence in conversation history. However, the records do not establish whether sufficient evidence was available or used for each answer. Absence of a new search therefore does not establish absence of source context or inappropriate retrieval behavior.

Additional-retrieval subgroups are selected by the model after configuration assignment; they were not separately randomized. Their low ratings could reflect difficult or poorly supported questions that prompted escalation. With only four rated Hybrid escalations, the data cannot establish that extra retrieval reduces quality, or that stopping would have improved those same answers. Estimating an escalation effect would require matched initial contexts and a controlled stop-versus-continue comparison.

Third, measured token consumption is dominated by input context. Answer length varies little across configurations, so measured token consumption is governed by how much history, instruction, and retrieved text is placed in the model's context, rather than by the orchestration label. Hybrid's higher mean token total coexists with rare additional retrieval. These observations do not isolate the contributions of prompt wording, question mix and inherited context. The slightly higher token mean of Agentic answers without a document-tool call likewise shows why a retrieval label alone is insufficient to explain consumption.

### 5.2. Implications for Evaluation and Deployment

The current observations do not justify an automatic deployment preference for Hybrid. Its higher rated-response mean accompanies a higher measured token mean, while quality uncertainty and missing feedback remain substantial. A deployment decision additionally requires verified transaction costs, failure severity, and application-specific quality requirements. Such requirements were not elicited as utility weights in this study. The reported token totals exclude re-ranking and other pipeline expenditure, so they cannot be interpreted as full operational or monetary costs (Section 3.5).

The additional summaries show substantial account concentration and much larger token totals on follow-up turns. No single account accounts for most of any configuration's consumption: among rated responses, the largest account contributes 10.25%, 24.10% and 21.69% of Agentic, Single-pass and Hybrid tokens, and the top three accounts contribute 25.82%, 42.66% and 36.71%. In the rated subset, concentration is therefore highest in Single-pass, not Hybrid, and the six Single-pass accounts in the top 10% contribute 53.86% of its tokens. Across all primary responses, concentration is similar and slightly highest for Agentic (top three accounts 29.42%, 29.13% and 26.33%; Table A8). On first turns, rated medians are almost identical (11,109, 11,259.5 and 11,492 tokens), while the descriptive Hybrid–Single-pass mean gap is approximately 2325 tokens. On follow-ups, both the mean gap (approximately 9592 tokens) and the median gap (12,465 tokens) are larger, and follow-ups form a slightly larger share of rated Hybrid responses (61% versus 57% for Single-pass). The observed Hybrid–Single-pass gap is therefore larger among follow-up responses. These descriptive summaries do not isolate configuration effects from conversation length, question mix or account composition, and the causes of the observed gap remain unresolved.

The account-influence check further shows that no single account explains the positive Hybrid–Single-pass mean token difference: it persists when each account is removed from all configurations (Table 14). This does not rule out collective influence from several accounts or separate configuration effects from differences in conversation history and question mix.

### 5.3. Practical Recommendations for Organizational RAG Deployments

The findings translate into six recommendations for practitioners who develop or operate organizational RAG assistants. They follow from the observed behavior and measurement experience of this deployment and should be validated in other settings.

1. **Log and audit tool use, not only configuration labels.** An agentic or hybrid label describes what a model is *permitted* to do. In this deployment, the agentic model made no new document-search call on 37% of turns, and the hybrid escalation path was used in 3.5% of responses. Recording every tool call per response is the only way to know which behavior users actually rated and paid for.
2. **Do not rely on prompt instructions alone to enforce retrieval.** An explicit instruction to always search the knowledge base did not ensure retrieval. Where answers must be grounded in organizational documents, obligatory initial retrieval in the orchestration code is a more dependable guarantee than a prompt instruction.
3. **Verify that an escalation path is exercised before attributing benefits to it.** The rarity of additional retrieval limits reliable estimation of its contribution to quality differences. Before an adaptive design is credited with quality gains or cost savings, its trigger rate and the outcomes of triggered cases should be reported.
4. **Control context size, the dominant driver of measured token use.** Because input tokens made up more than 98% of measured consumption, history trimming, prompt length, and the number of retrieved passages are the main levers for reducing measured token use, whereas limits on answer length affect only a small share. The monetary savings and quality effects of such limits were not evaluated.

5. **Randomize and conceal configuration assignment, and plan for clustering.** Turn-level random assignment with configuration names concealed from users is feasible in a production assistant; interface elements that differ between configurations, such as tool-step displays, should also be hidden. Assignment per thread or per user would avoid mixed conversation histories, and analyses should treat users, not responses, as independent units.
6. **Design feedback collection for coverage.** Requiring a rating only to continue a thread left 43% of responses unrated. Lightweight ratings that can also be given when a thread ends, together with reporting of submission rates per configuration, reduce the risk that quality estimates reflect only the conversations users chose to continue.

These recommendations also define what a reusable evaluation report should contain: tool-use rates per configuration, rating coverage, participant-level uncertainty, and a token-accounting boundary. They prevent architecture labels or a selected low-token subgroup from substituting for evidence about deployed behavior.

### *5.4. Conditional Retrieval as a Design Hypothesis*

Bounded rationality motivates additional computation when its expected benefit justifies its cost [19]. Adaptive-RAG, MBA-RAG, RAGate, and Stop-RAG operationalize related choices in their respective settings [20,45–47]. Here this remains a design hypothesis. A causal evaluation would compare stopping and continuing from matched initial contexts, with the same prompts and complete per-call accounting. The present configurations do not supply that controlled contrast.

### *5.5. Limitations*

First, ratings cover only 57.35% of primary responses. Although scoring was required to continue within a thread, users could leave the application or end or switch threads before submitting feedback. A dissatisfied user might abandon a thread after a poor answer, whereas a satisfied user might stop after an excellent answer completed the task. Both low and high ratings could therefore be selectively missing; the direction of bias is unknown. Neither similar submission rates nor cluster-robust inference establishes that the rated responses are representative of all answers.

Second, account-level clustering assumes independence between accounts; it does not establish it. If one person used multiple accounts, dependence could cross the recorded clusters, reducing the effective number of independent participants and potentially making uncertainty intervals too narrow. A shared account would combine judgments from different people within one cluster. Neither pattern was verified, and thread-cluster sensitivity does not validate account identity. Response weighting gives frequent raters greater influence; the estimand remains the response-weighted mean among observed ratings.

Third, the outcomes are user perceptions rather than audited factual accuracy or source support. They cannot establish factual-error rates or verified task success, and the five-item index is not an independently validated scale.

Fourth, each configuration combines orchestration with its own system prompt, and turn-level assignment produces conversations whose earlier turns may come from other configurations. Differences between configurations therefore reflect deployable packages in mixed histories; they cannot be attributed to orchestration logic alone or to the long-run use of one configuration. Because additional Hybrid retrieval was rare, the study also has little information on the value of escalation itself. Because tool steps were displayed by default, the presence or absence of a document-search step could have offered a partial cue to the configuration (Section 3.2). Agentic responses without a document-tool step

were not rated lower descriptively than those with one (Table 12), but this does not establish that masking was effective.

Fifth, application-measured token totals exclude re-ranking (Section 3.5). The comparison therefore concerns the included input/output tokens, rather than complete pipeline cost. Shared re-ranking procedures do not guarantee equal omitted overhead when retrieval frequency differs.

Finally, the single organization, Slovenian language, and specific historical implementation limit generalization. These limitations constrain claims about policies and mechanisms even if larger samples were to narrow the statistical intervals.

### *5.6. Future Work*

Prospective studies should specify the target population in advance and account for repeated observations from participants.

Independent annotation of answers can assess factual or source-support errors if conducted with an explicit sampling protocol, but would be a new evaluation rather than a historical user measure. The design–behavior gap motivates a specific follow-up: comparing model-decided retrieval with retrieval decided by an explicit controller, such as a trained sufficiency classifier [20,46], under the same prompt, so that trigger rates and their consequences can be measured directly. Controlled comparisons should evaluate additional retrieval from matched contexts and measure monetary cost and latency before testing a routing policy across organizations.

## 6. Conclusions

This study examined what happens when three established RAG configurations (single-pass, agentic, and hybrid) are used by employees for real organizational questions, with turn-level assignment and concealed configuration names on a shared stack. Across 701 responses and 402 complete user evaluations from 73 participant accounts, Hybrid has the highest descriptive composite mean (4.15), followed by Single-pass (3.93) and Agentic (3.79), and leads on all five dimensions, but it also has the highest mean measured input-plus-output token total. No dimension passes Holm adjustment and all multiplicity-adjusted composite intervals include zero; the results establish neither superiority nor equivalence.

The clearest finding concerns behavior rather than ratings. The agentic model made no new document-search call on 37% of turns, and the hybrid configuration used additional retrieval in only 9 of 254 responses. The adaptive mechanisms that distinguish these configurations by design therefore often produced no new search in Agentic and rarely added retrieval in Hybrid, and Hybrid's descriptive lead cannot be credited to escalation. Measured token use was driven by input context, not by answer length.

For practitioners, the study shows that configuration labels are insufficient evidence of deployed behavior. Tool use should be logged and audited, retrieval that must occur should be enforced in orchestration code rather than requested in prompts, and context size should be managed as the main target for reducing measured input-token consumption. For researchers, it documents a field-evaluation approach for repeated, partially rated user feedback and derives practical recommendations. Establishing when further retrieval is worthwhile requires controlled comparisons from matched contexts and complete cost accounting beyond this field sample.

**Author Contributions:** Conceptualization, M.Š.; methodology, M.Š. and R.L.; software, M.Š.; formal analysis, M.Š.; investigation, M.Š.; writing—original draft preparation, M.Š.; writing—review and editing, M.Š., M.G. and R.L.; supervision, M.G. and R.L. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was supported by ongoing national and European Union-funded projects at the E9 Department of the Jožef Stefan Institute and University of Maribor, Faculty of Organizational Sciences. The authors also acknowledge funding from the Slovenian Research and Innovation Agency (ARIS), basic core funding P2-0209 and P5-0018. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

**Institutional Review Board Statement:** The research protocol was reviewed and approved by the Ethics Committee for Research in the Field of Organizational Sciences at the University of Maribor, Faculty of Organizational Sciences (decision no. 514-18/2025/1/250-DJ, dated 18 December 2025). The committee confirmed that the research poses no ethical concerns and granted formal approval for its execution. All participants were informed about the purpose and nature of the study, and informed consent was obtained prior to participation. Participation was voluntary.

**Informed Consent Statement:** Informed consent was obtained from all participants before their first use of the system. All participants were informed about the purpose and nature of the study and received an information notice and consent form. Participation was voluntary. Consent for publication is not applicable because the manuscript does not contain data that identify any individual person.

**Data Availability Statement:** The source records contain organizational questions and pseudonymized account-linked interactions and are not publicly available. Access requests may be addressed to the corresponding author, subject to institutional permission and applicable privacy and consent restrictions.

**Acknowledgments:** GPT-6 (OpenAI) and Claude Opus 5.5 (Anthropic) (used in October and November 2026) assisted with manuscript revision and data analysis. Numerical summaries and cluster tests were checked using a separate implementation. The authors are responsible for the accuracy, interpretation, declarations, and final content of the manuscript.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## Appendix A. Usage Statistics and Detailed Rating Distributions

### *Appendix A.1. Usage Statistics*

The primary sample contains 701 responses across 408 threads. The mean number of responses per thread is 1.72 (sample SD 1.80), the 90th percentile is 3, and the maximum is 18. Complete evaluations occur in 176 threads and come from 73 accounts.

### *Appendix A.2. Full Rating Distributions*

Tables A1–A5 describe the same 122 Agentic, 144 Single-pass, and 136 Hybrid rated responses. Percentages can differ from 100% by rounding.

**Table A1.** Accuracy rating distribution (percentages within each configuration).

| Configuration | Mean | SD | 1 | 2 | 3 | 4 | 5 | Median |
|---|---|---|---|---|---|---|---|---|
| Agentic RAG | 3.63 | 1.43 | 13.9% | 9.0% | 15.6% | 23.0% | 38.5% | 4 |
| Single-pass RAG | 3.78 | 1.44 | 13.9% | 6.9% | 11.8% | 21.5% | 45.8% | 4 |
| Hybrid RAG | 3.93 | 1.39 | 11.8% | 5.1% | 14.0% | 16.2% | 52.9% | 5 |

**Table A2.** Information coverage rating distribution (percentages within each configuration).

| Configuration | Mean | SD | 1 | 2 | 3 | 4 | 5 | Median |
|---|---|---|---|---|---|---|---|---|
| Agentic RAG | 3.43 | 1.49 | 16.4% | 13.1% | 16.4% | 18.9% | 35.2% | 4 |
| Single-pass RAG | 3.69 | 1.45 | 14.6% | 7.6% | 14.6% | 20.8% | 42.4% | 4 |
| Hybrid RAG | 3.93 | 1.40 | 11.0% | 8.1% | 10.3% | 18.4% | 52.2% | 5 |

**Table A3.** Clarity rating distribution (percentages within each configuration).

| Configuration | Mean | SD | 1 | 2 | 3 | 4 | 5 | Median |
|---|---|---|---|---|---|---|---|---|
| Agentic RAG | 4.02 | 1.28 | 7.4% | 7.4% | 13.1% | 19.7% | 52.5% | 5 |
| Single-pass RAG | 4.10 | 1.26 | 7.6% | 5.6% | 11.8% | 18.8% | 56.2% | 5 |
| Hybrid RAG | 4.35 | 1.04 | 3.7% | 2.9% | 11.0% | 19.9% | 62.5% | 5 |

**Table A4.** Source linkage rating distribution (percentages within each configuration).

| Configuration | Mean | SD | 1 | 2 | 3 | 4 | 5 | Median |
|---|---|---|---|---|---|---|---|---|
| Agentic RAG | 3.69 | 1.52 | 16.4% | 8.2% | 11.5% | 18.0% | 45.9% | 4 |
| Single-pass RAG | 3.83 | 1.46 | 12.5% | 10.4% | 9.7% | 16.7% | 50.7% | 5 |
| Hybrid RAG | 4.06 | 1.38 | 10.3% | 6.6% | 10.3% | 12.5% | 60.3% | 5 |

**Table A5.** Expression appropriateness rating distribution (percentages within each configuration).

| Configuration | Mean | SD | 1 | 2 | 3 | 4 | 5 | Median |
|---|---|---|---|---|---|---|---|---|
| Agentic RAG | 4.15 | 1.24 | 6.6% | 6.6% | 10.7% | 18.0% | 58.2% | 5 |
| Single-pass RAG | 4.26 | 1.19 | 6.2% | 4.9% | 8.3% | 18.1% | 62.5% | 5 |
| Hybrid RAG | 4.46 | 1.02 | 3.7% | 2.9% | 8.1% | 14.0% | 71.3% | 5 |

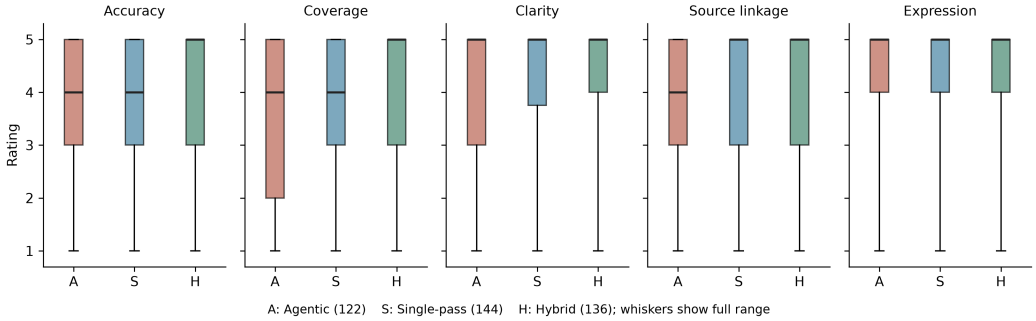


**Figure A1.** Complete-rating distributions. Boxes show medians and quartiles; whiskers show the full range. A, S, and H denote Agentic, Single-pass, and Hybrid.

## Appendix B. Sensitivity Analyses

The full-export sensitivity includes the researcher account and all 416 ratings from 74 accounts. The primary thread sensitivity uses 176 rated threads, whereas the primary account analysis uses 73 rated accounts. Each column of Table A6 applies Holm adjustment to its own five-dimension family. They are alternative analytic specifications, not additional independent replications.

**Table A6.** Sensitivity of dimension comparisons to clustering and sample choice. Entries are Holm-adjusted probabilities for the five-outcome family.

| Dimension | Primary: User | Primary: Thread | Full Export: User |
|---|---|---|---|
| Accuracy | 0.49804 | 0.45850 | 1.00000 |
| Information coverage | 0.09310 | 0.14821 | 0.39972 |
| Clarity | 0.49804 | 0.45850 | 1.00000 |
| Source linkage | 0.49804 | 0.45850 | 1.00000 |
| Expression appropriateness | 0.49804 | 0.45850 | 1.00000 |

**Table A7.** Worst-case bounds on all-response composite means, allowing each unrated score to lie between 1 and 5. These are identification bounds, not confidence intervals.

| Configuration | Lower Bound | Upper Bound |
|---|---|---|
| Agentic RAG | 2.618 | 4.294 |
| Single-pass RAG | 2.781 | 4.351 |
| Hybrid RAG | 2.684 | 4.543 |

## Appendix C. Token Distributions and Exploratory Groups

### *Appendix C.1. Account Concentration and Original Turn Position*

Table A8 summarizes tokens accumulated by active accounts within each configuration. Accounts can occur in several configurations; each concentration denominator is the token total for the stated configuration and subset. In the rated subset, the top three accounts contribute 25.82% of Agentic tokens, 42.66% of Single-pass tokens and 36.71% of Hybrid tokens. The top 10% groups contain five, six and five accounts, respectively, and contribute 38.37%, 53.86% and 45.57%. The largest single account contributes 10.25%, 24.10% and 21.69%, respectively. Account-level summaries are aggregate distributions, not disclosures of individual accounts.

**Table A8.** Cumulative input-plus-output tokens per account within each configuration and subset. Top-group entries are percentages of that configuration's subset token total; parentheses give selected account counts. The top 5% and 10% use the ceiling of the corresponding fraction of active accounts. Re-ranking is excluded.

| Configuration | Subset | Accounts | Median | Q1 | Q3 | Top 5% | Top 10% | Top 3 (%) |
|---|---|---|---|---|---|---|---|---|
| Agentic | All | 83 | 26,581 | 11,905.5 | 73,319 | 38.60 (5) | 51.64 (9) | 29.42 |
| Single-pass | All | 84 | 39,848 | 14,222 | 95,711.75 | 35.74 (5) | 47.18 (9) | 29.13 |
| Hybrid | All | 92 | 42,098.5 | 17,198.75 | 115,263.5 | 34.75 (5) | 49.62 (10) | 26.33 |
| Agentic | Rated | 48 | 41,529.5 | 16,046.75 | 87,844.5 | 25.82 (3) | 38.37 (5) | 25.82 |
| Single-pass | Rated | 57 | 31,355 | 15,139 | 94,082 | 42.66 (3) | 53.86 (6) | 42.66 |
| Hybrid | Rated | 50 | 57,149 | 24,918.25 | 121,943.75 | 36.71 (3) | 45.57 (5) | 36.71 |

Across all 701 primary responses, 385 are first turns and 316 are follow-ups. In 23 of the 408 sampled threads, the first archived user run is outside the analyzed response sample, so the earliest selected response is correctly classified as a follow-up. Table A9 uses the full archived user-run chronology rather than rating order. Among rated responses, mean first-turn totals are 11,354, 13,388 and 15,713 for Agentic, Single-pass and Hybrid; corresponding follow-up means are 41,994, 47,788 and 57,380.

**Table A9.** Input-plus-output tokens by original thread turn, excluding re-ranking. First denotes the first archived user-message run, not the first rated or exported response. All other analyzed responses are Follow-up. Means are rounded; interpolated quartile fractions are retained.

| Configuration | Subset | Turn | *n* | Mean | Median | Q1 | Q3 |
|---|---|---|---|---|---|---|---|
| Agentic | All | First | 112 | 11,236 | 11,103.5 | 9336.5 | 14,176 |
| Agentic | All | Follow-up | 98 | 44,780 | 30,530.5 | 15,195 | 54,323.5 |
| Single-pass | All | First | 132 | 14,865 | 11,507.5 | 9001.5 | 15,062.25 |
| Single-pass | All | Follow-up | 105 | 47,553 | 35,625 | 23,757 | 57,229 |
| Hybrid | All | First | 141 | 16,353 | 12,260 | 9846 | 16,147 |
| Hybrid | All | Follow-up | 113 | 63,700 | 45,114 | 27,690 | 74,386 |
| Agentic | Rated | First | 54 | 11,354 | 11,109 | 9363.75 | 14,340.75 |
| Agentic | Rated | Follow-up | 68 | 41,994 | 31,780 | 15,953.75 | 55,611.5 |
| Single-pass | Rated | First | 62 | 13,388 | 11,259.5 | 8889 | 14,976.75 |
| Single-pass | Rated | Follow-up | 82 | 47,788 | 33,162 | 22,467.75 | 56,236.5 |
| Hybrid | Rated | First | 53 | 15,713 | 11,492 | 9616 | 14,925 |
| Hybrid | Rated | Follow-up | 83 | 57,380 | 45,627 | 27,168.5 | 73,221.5 |

*Appendix C.2. Exploratory Configuration–Token Groups*

The 402 rated responses form 46 observed configuration–token groups using intervals of 10,000 measured tokens of $T_i$ (Section 3.5). A group is non-dominated when no other observed group has at least as high a mean quality and at most as large a mean measured input-plus-output token total, with a strict improvement in one criterion. Five groups satisfy this descriptive definition. The lowest-token Hybrid group contains 16 responses, has mean quality 4.400, and averages 8230 tokens. Higher-token frontier points contain five, one, and two responses. These small, selected groups cannot establish an optimal budget or the return to giving a question more tokens. Their non-dominance concerns measured input/output tokens only; re-ranking and other pipeline costs are outside the comparison.

**Table A10.** Observed non-dominated measured-token groups among the 402 rated responses. Labels indicate the lower bound of a 10,000-token interval. Re-ranking and other pipeline costs are excluded.

| Configuration | Bucket Lower Bound | *n* | Mean Tokens | Mean Quality |
|---|---|---|---|---|
| Agentic RAG | 0 | 24 | 6390.96 | 3.7167 |
| Hybrid RAG | 0 | 16 | 8230.06 | 4.4000 |
| Single-pass RAG | 60,000 | 5 | 65,239.40 | 4.6800 |
| Agentic RAG | 80,000 | 1 | 82,058.00 | 4.8000 |
| Hybrid RAG | 90,000 | 2 | 93,944.50 | 5.0000 |

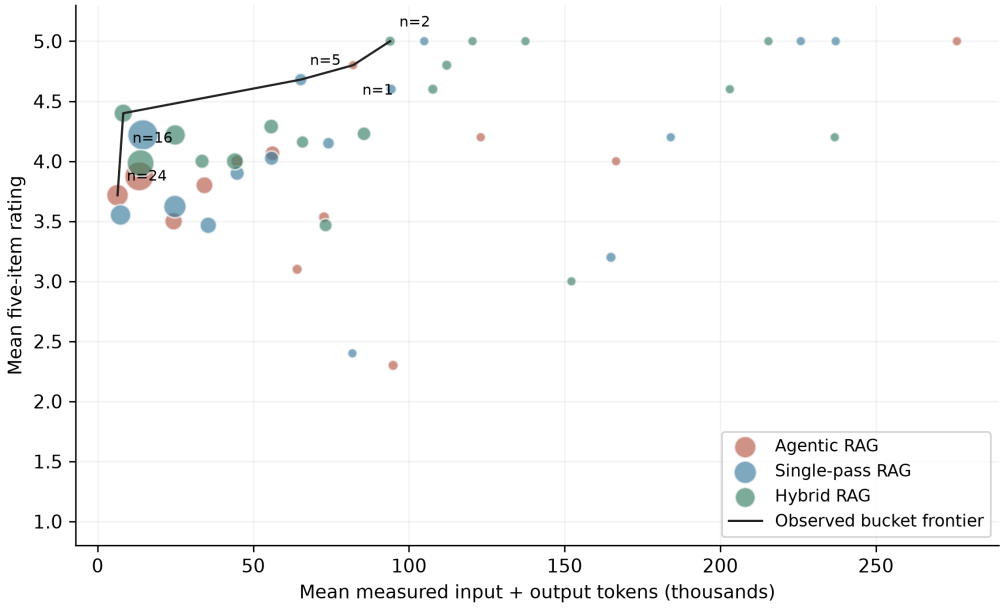


**Figure A2.** Exploratory means of measured-token groups in the primary rated sample. Bubble area increases with group size; the line connects observed non-dominated means. It is not a causal response curve or a deployable policy.

## Appendix D. System Prompts and Tool Definitions

The deployed prompts were written in Slovenian; the text below is an English translation of the archived templates, not the literal text sent to the model. The name of the institution is replaced by *[institution]*. Placeholders in braces were filled with the current date and day of the week at request time. Retrieved passages were supplied to the model as additional system-role messages.

### *Appendix D.1. Instructions Shared by All Configurations*

> Communicate only in Slovenian. You are a friendly assistant at [institution]. You are for internal use by [institution] employees only. You were created by [institution] development. If you make queries through functions, expect answers in Slovenian. Also expect all messages with the system role in Slovenian; these will usually be document queries or explanations to help your understanding. Do not answer a question unless you obtain a source for the answer. Ask for clarification if the question or the user's intent is ambiguous. [Configuration-specific instructions; Table A11.] The current date and time is {current_datetime}. The day of the week is {day_of_week}. Always list all relevant sources from which you obtained information at the end of the answer. Be sure to complete the source list. If you obtain information from sources, do not cite them immediately, but at the end of your answer in this form: "Sources: [file name 1](link to file 1), [file name 2](link to file 2)".

### *Appendix D.2. Configuration-Specific Instructions*

**Table A11.** Configuration-specific system-prompt instructions (English translation). A dash indicates that the instruction is absent from that configuration's prompt.

| Instruction | Single-Pass RAG | Agentic RAG | Hybrid RAG |
|---|---|---|---|
| Amendments | Documents in the knowledge base also have amendments. If you suspect that deeper understanding of the original document is necessary, ask the user for further clarification of the purpose and context of the question. | Documents in the knowledge base also have amendments. If you suspect that deeper understanding of the original document and subsequent amendments is necessary, search the knowledge base again for everything that may relate to amendments and original documents/laws and use it in the answer. | Identical to Agentic RAG. |
| Tool availability | — | Tools for document search and knowledge queries are available to help you answer questions. | Identical to Agentic RAG. |
| Tool use | — | Always use tools and functions. In particular, always search the knowledge base whenever you need knowledge about anything related to [institution]. You may use several tools or functions together in one answer. | Use the tools and functions available to you. In particular, always search the knowledge base whenever you need knowledge about anything related to [institution], *if it has not already been provided in the vector-database results*. You may use several tools or functions together in one answer. |
| Tool disclosure | — | Do not mention that you are using tools or functions; use them and answer the question. | Identical to Agentic RAG. |

The Agentic and Hybrid prompts therefore differ only in the tool-use instruction, which the Hybrid prompt makes conditional on information not already supplied by initial retrieval (italicized). The Single-pass prompt contains no tool-use instruction and redirects amendment-related uncertainty to the user rather than to further retrieval.

*Appendix D.3. Tool Definitions*

All three configurations expose a time function, `get_current_system_time`, described as "Returns the current system time", with a single required `timezone` parameter fixed to Central European Time. The Agentic and Hybrid configurations additionally expose one document-search function with a single required string parameter, `search_string`, which the model is told to formulate as "a rich retrieval query including synonyms, related legal terminology, temporal or contextual qualifiers to maximize recall before reranking". The function description states that it retrieves up to 100 potentially relevant passages using dense and optional sparse search, re-ranks them with the chat model for semantic relevance, and returns only the most relevant passages. The description mentions the top 10 passages, whereas the recorded thread settings specify a final $k$ of 12 (Table 2); this discrepancy is reported as recovered rather than corrected. Single-pass RAG performs the same retrieval and re-ranking procedure once in the orchestration code before generation, without exposing it as a tool. An instruction to use a tool is not evidence that it was invoked; observed invocations are reported in Table 10.

## Appendix E. Record Linkage and Exclusion Criteria

*Appendix E.1. Record Linkage*

All 720 exported feedback records were linked to archived answers using the thread identifier and question, the parent run, the recorded configuration, the answer length in characters, and the five-item and binary feedback vector. Two otherwise identical signatures were resolved by chronological ordering within the thread, and the ordering was checked across all matched records. Every archived answer identifier occurs at most once in the resulting mapping. The application's recorded input and output token counts sum to the recorded total in all 720 records.

*Appendix E.2. Researcher-Account Exclusion*

The researcher/developer account contributed 19 records, 14 of them rated. It was excluded from the primary analysis because the target population is employee use independent of the investigators, not because of any observed rating; the exclusion was introduced during archival analysis and was not preregistered. The exclusion is applied to linked account identifiers rather than to user-step names, which prevents answer records from remaining in the sample after their parent user step has been filtered. The full-export sensitivity analysis retains all 720 records (Appendix B).

## References

1. Lewandowski, T.; Kučević, E.; Leible, S.; Poser, M.; Böhmann, T. Enhancing conversational agents for successful operation: A multi-perspective evaluation approach for continuous improvement. *Electron. Mark.* **2023**, *33*, 39. [CrossRef]
2. Klesel, M.; Wittmann, H.F. Retrieval-Augmented Generation (RAG). *Bus. Inf. Syst. Eng.* **2025**, *67*, 551–561. [CrossRef]
3. O'Shea, K.; Crockett, K.; Bandar, Z.; O'Shea, J. Natural language scripting within conversational agent design. *Appl. Intell.* **2014**, *40*, 189–197. [CrossRef]
4. O'Shea, K.; Crockett, K.; Bandar, Z.; O'Shea, J. An Approach to Conversational Agent Design Using Semantic Sentence Similarity. *Appl. Intell.* **2012**, *37*, 558–568. [CrossRef]
5. Rezaei, M. Artificial intelligence in knowledge management: Identifying and addressing the key implementation challenges. *Technol. Forecast. Soc. Change* **2025**, *217*, 124183. [CrossRef]
6. Jarrahi, M.H.; Askay, D.; Eshraghi, A.; Smith, P. Artificial intelligence and knowledge management: A partnership between human and AI. *Bus. Horiz.* **2023**, *66*, 87–99. [CrossRef]
7. Bujold, A.; Roberge-Maltais, I.; Parent-Rocheleau, X.; Boasen, J.; Sénécal, S.; Léger, P.M. Responsible artificial intelligence in human resources management: A review of the empirical literature. *AI Ethics* **2024**, *4*, 1185–1200. [CrossRef]
8. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. *arXiv* **2017**, arXiv:1706.03762. [CrossRef]

9. Adamopoulou, E.; Moussiades, L. An Overview of Chatbot Technology. In *Artificial Intelligence Applications and Innovations*; IFIP Advances in Information and Communication Technology; Maglogiannis, I., Iliadis, L., Pimenidis, E., Eds.; Springer International Publishing: Cham, Switzerland, 2020; Volume 584, pp. 373–383. [CrossRef]
10. Dwivedi, Y.K.; Kshetri, N.; Hughes, L.; Slade, E.L.; Jeyaraj, A.; Kar, A.K.; Baabdullah, A.M.; Koohang, A.; Raghavan, V.; Ahuja, M.; et al. Opinion Paper: 'So what if ChatGPT wrote it?' Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *Int. J. Inf. Manag.* **2023**, *71*, 102642. [CrossRef]
11. Dhingra, B.; Cole, J.R.; Eisenschlos, J.M.; Gillick, D.; Eisenstein, J.; Cohen, W.W. Time-Aware Language Models as Temporal Knowledge Bases. *Trans. Assoc. Comput. Linguist.* **2022**, *10*, 257–273. [CrossRef]
12. Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Bang, Y.J.; Madotto, A.; Fung, P. Survey of Hallucination in Natural Language Generation. *ACM Comput. Surv.* **2023**, *55*, 248. [CrossRef]
13. Huang, L.; Yu, W.; Ma, W.; Zhong, W.; Feng, Z.; Wang, H.; Chen, Q.; Peng, W.; Feng, X.; Qin, B.; et al. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Trans. Inf. Syst.* **2025**, *43*, 42. [CrossRef]
14. Bender, E.M.; Gebru, T.; McMillan-Major, A.; Shmitchell, S. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, Virtual Event, Canada*; Association for Computing Machinery: New York, NY, USA, 2021; pp. 610–623. [CrossRef]
15. Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.t.; Rocktäschel, T.; et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *arXiv* **2020**. [CrossRef]
16. Karpukhin, V.; Oğuz, B.; Min, S.; Lewis, P.; Wu, L.; Edunov, S.; Chen, D.; Yih, W.T. Dense Passage Retrieval for Open-Domain Question Answering. *arXiv* **2020**. [CrossRef]
17. Yao, S.; Zhao, J.; Yu, D.; Du, N.; Shafran, I.; Narasimhan, K.; Cao, Y. ReAct: Synergizing Reasoning and Acting in Language Models. *arXiv* **2022**. [CrossRef]
18. Schick, T.; Dwivedi-Yu, J.; Dessì, R.; Raileanu, R.; Lomeli, M.; Hambro, E.; Zettlemoyer, L.; Cancedda, N.; Scialom, T. Toolformer: Language Models Can Teach Themselves to Use Tools. In *Proceedings of the Advances in Neural Information Processing Systems*; Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2023; Volume 36, pp. 68539–68551. [CrossRef]
19. Simon, H.A. Rational decision making in business organizations. *Am. Econ. Rev.* **1979**, *69*, 493–513.
20. Jeong, S.; Baek, J.; Cho, S.; Hwang, S.J.; Park, J. Adaptive-RAG: Learning to Adapt Retrieval-Augmented Large Language Models through Question Complexity. In *Proceedings of the NAACL-HLT (Volume 1: Long Papers)*; Association for Computational Linguistics: Kerrville, TX, USA, 2024; pp. 7036–7050. [CrossRef]
21. Es, S.; James, J.; Espinosa-Anke, L.; Schockaert, S. Ragas: Automated Evaluation of Retrieval Augmented Generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*; Association for Computational Linguistics: Kerrville, TX, USA, 2024; pp. 150–158. [CrossRef]
22. Burbano, J.; Landeta-López, P.; Guevara-Vega, C.; Quiña-Mera, A. Design and Evaluation of a Retrieval-Augmented Generation LLM Chatbot with Structured Database Access. *Appl. Sci.* **2026**, *16*, 3147. [CrossRef]
23. Martín, C.A.; Torres, J.M.; Aguilar, R.M.; Alayón, S.; Bacallado, M.A. Agentic Knowledge Curation Versus Full-Context Retrieval: An Empirical Study of Retrieval Failure Topology in Long-Context LLM Systems. *Appl. Sci.* **2026**, *16*, 6793. [CrossRef]
24. Škarabot, M.; Leskovar, R. Conceptual Framework for Utilizing Chatbots as Domain Experts in Organizations. In *Proceedings of the Conference on Organizational Science Development: Green and Digital Transition—Challenge or Opportunity*; Univerzitetna založba Univerze v Mariboru: Maribor, Slovenia, 2024; pp. 935–945. [CrossRef]
25. Izacard, G.; Lewis, P.; Lomeli, M.; Hosseini, L.; Petroni, F.; Schick, T.; Dwivedi-Yu, J.; Joulin, A.; Riedel, S.; Grave, E. Atlas: Few-shot Learning with Retrieval Augmented Language Models. *arXiv* **2022**. [CrossRef]
26. Gao, Y.; Xiong, Y.; Gao, X.; Jia, K.; Pan, J.; Bi, Y.; Dai, Y.; Sun, J.; Wang, M.; Wang, H. Retrieval-Augmented Generation for Large Language Models: A Survey. *arXiv* **2024**. [CrossRef]
27. Huang, Y.; Huang, J. A Survey on Retrieval-Augmented Text Generation for Large Language Models. *arXiv* **2024**, arXiv:2404.10981. [CrossRef]
28. Zhang, J.; Li, G.; Su, J. SAGE: A Framework of Precise Retrieval for RAG. In Proceedings of the 2025 IEEE 41st International Conference on Data Engineering (ICDE), Hong Kong, China, 19–23 May 2025; pp. 1388–1401. [CrossRef]
29. Papageorgiou, G.; Sarlis, V.; Maragoudakis, M.; Tjortjis, C. Enhancing E-Government Services through State-of-the-Art, Modular, and Reproducible Architecture over Large Language Models. *Appl. Sci.* **2024**, *14*, 8259. [CrossRef]
30. Linders, J.; Tomczak, J.M. Knowledge graph-extended retrieval augmented generation for question answering. *Appl. Intell.* **2025**, *55*, 1102. [CrossRef]
31. Esposito, M.; Damiano, E.; Minutolo, A.; De Pietro, G.; Fujita, H. Hybrid Query Expansion Using Lexical Resources and Word Embeddings for Sentence Retrieval in Question Answering. *Inf. Sci.* **2020**, *514*, 88–105. [CrossRef]
32. Li, X.; Peng, S.; Yada, S.; Wakamiya, S.; Aramaki, E. GenKP: Generative Knowledge Prompts for Enhancing Large Language Models. *Appl. Intell.* **2025**, *55*, 464. [CrossRef]

33. Tam, T.Y.C.; Sivarajkumar, S.; Kapoor, S.; Stolyar, A.V.; Polanska, K.; McCarthy, K.R.; Osterhoudt, H.; Wu, X.; Visweswaran, S.; Fu, S.; et al. A Framework for Human Evaluation of Large Language Models in Healthcare Derived from Literature Review. *npj Digit. Med.* **2024**, *7*, 258. [CrossRef]
34. Deriu, J.; Rodrigo, A.; Otegi, A.; Echegoyen, G.; Rosset, S.; Agirre, E.; Cieliebak, M. Survey on Evaluation Methods for Dialogue Systems. *Artif. Intell. Rev.* **2021**, *54*, 755–810. [CrossRef]
35. Kocielnik, R.; Amershi, S.; Bennett, P.N. Will You Accept an Imperfect AI?: Exploring Designs for Adjusting End-user Expectations of AI Systems. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, Glasgow, UK, 4–9 May 2019; pp. 1–14. [CrossRef]
36. Gkinko, L.; Elbanna, A. The appropriation of conversational AI in the workplace: A taxonomy of AI chatbot users. *Int. J. Inf. Manag.* **2023**, *69*, 102568. [CrossRef]
37. Bohnet, B.; Tran, V.Q.; Verga, P.; Aharoni, R.; Andor, D.; Baldini Soares, L.; Ciaramita, M.; Eisenstein, J.; Ganchev, K.; Herzig, J.; et al. Attributed Question Answering: Evaluation and Modeling for Attributed Large Language Models. *arXiv* **2022**. [CrossRef]
38. Dimitriadis, D.; Tsoumakas, G. Enhancing Yes/No Question Answering with Weak Supervision via Extractive Question Answering. *Appl. Intell.* **2023**, *53*, 27560–27570. [CrossRef]
39. Gams, M.; Horvat, T.; Kolar, Ž.; Kocuvan, P.; Mishev, K.; Misheva, M.S. Evaluating a Nationally Localized AI Chatbot for Personalized Primary Care Guidance: Insights from the HomeDoctor Deployment in Slovenia. *Healthcare* **2025**, *13*, 1843. [CrossRef]
40. Pacheco-Lorenzo, M.R.; Anido-Rifón, L.E.; Fernández-Iglesias, M.J.; Valladares-Rodríguez, S.M. Will senior adults accept being cognitively assessed by a conversational agent? A user-interaction pilot study. *Appl. Intell.* **2024**, *54*, 7897–7912. [CrossRef]
41. Li, X.; Cao, Y.; Ma, Y.; Sun, A. Long Context vs. RAG for LLMs: An Evaluation and Revisits. *arXiv* **2024**. [CrossRef]
42. Yu, H.; Gan, A.; Zhang, K.; Tong, S.; Liu, Q.; Liu, Z. Evaluation of Retrieval-Augmented Generation: A Survey. In *Proceedings of the Big Data*; Springer: Singapore, 2025; pp. 102–120. [CrossRef]
43. Liu, N.F.; Lin, K.; Hewitt, J.; Paranjape, A.; Bevilacqua, M.; Petroni, F.; Liang, P. Lost in the Middle: How Language Models Use Long Contexts. *Trans. Assoc. Comput. Linguist.* **2024**, *12*, 157–173. [CrossRef]
44. Budakoglu, G.; Emekci, H. Unveiling the Power of Large Language Models: A Comparative Study of Retrieval-Augmented Generation, Fine-Tuning, and Their Synergistic Fusion for Enhanced Performance. *IEEE Access* **2025**, *13*, 30936–30951. [CrossRef]
45. Tang, X.; Gao, Q.; Li, J.; Du, N.; Li, Q.; Xie, S. MBA-RAG: A Bandit Approach for Adaptive Retrieval-Augmented Generation through Question Complexity. In *Proceedings of the 31st International Conference on Computational Linguistics*; Association for Computational Linguistics: Kerrville, TX, USA, 2025; pp. 3248–3254.
46. Wang, X.; Sen, P.; Li, R.; Yilmaz, E. Adaptive Retrieval-Augmented Generation for Conversational Systems. In *Proceedings of the Findings of the Association for Computational Linguistics: NAACL 2025*; Association for Computational Linguistics: Kerrville, TX, USA, 2025; pp. 491–503. [CrossRef]
47. Park, J.; Cho, S.; Lee, J.Y. Stop-RAG: Value-Based Retrieval Control for Iterative RAG. *arXiv* **2025**. [CrossRef]
48. Li, Y.; Yan, Z.; Käfer, T. RASER: Recoverability-Aware Selective Escalation Router for Multi-Hop Question Answering. *arXiv* **2026**. [CrossRef]
49. OpenAI. GPT-4o: Multimodal Large Language Model. 2024. Available online: https://openai.com/index/hello-gpt-4o/ (accessed on 22 December 2025).
50. OpenAI. Text-Embedding-3-Large. 2024. Available online: https://platform.openai.com/docs/models/text-embedding-3-large (accessed on 22 December 2025).
51. Qdrant Team. Qdrant: Open-Source Vector Database for AI Applications. 2025. Available online: https://qdrant.tech (accessed on 22 December 2025).
52. Cameron, A.C.; Miller, D.L. A Practitioner's Guide to Cluster-Robust Inference. *J. Hum. Resour.* **2015**, *50*, 317–372. [CrossRef]
53. Holm, S. A Simple Sequentially Rejective Multiple Test Procedure. *Scand. J. Stat.* **1979**, *6*, 65–70.


**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.