

Design and Evaluation of a Route-Aware, Evidence-Grounded RAG Chatbot for Accessing Village Monographs and Census Data within Data Desa Presisi

Ikhlasul Amal¹, Irman Hermadi², Yani Nurhadryani²

¹Computer Science, Postgraduate Program, IPB University, Bogor, Indonesia.

²Computer Science, School of Data Science, Mathematics, and Informatics, IPB University, Bogor, Indonesia.

Received:

May 23, 2026

Revised:

July 25, 2026

Accepted:

September 2, 2026

Published:

October 24, 2026

Corresponding Author:

Author Name*:

Ikhlasul Amal

Email*:

ikhlaslamal@apps.ipb.ac.id

DOI:

10.63158/journalisi.v8i5.1856

© 2026 Journal of Information Systems and Informatics. This open access article is distributed under a (CC-BY License)



Abstract. Data Desa Presisi integrates narrative village monographs and structured census records, creating challenges for contextual, numerical, and cross-source information access. This study designed and evaluated a locally deployed, route-aware, evidence-grounded retrieval-augmented generation (RAG) chatbot that combines deterministic DuckDB queries, document retrieval, provenance, and privacy-aware routing. Evaluation used 100 ground-truth questions from Setu and Gunung Putri. Eighty-four database, monograph, and mixed-evidence questions were assessed using RAGAS and ROUGE-L, whereas 16 conversational and policy questions were evaluated through deterministic checks. Elasticsearch at a temperature of 0.1 produced the highest composite score (0.592018), although post hoc Wilcoxon tests found no statistically significant superiority over alternative configurations. The selected configuration achieved a 98.0% functional pass rate and a 97.92% numeric-match rate. Public-mode usability evaluation involving 21 village officials and staff yielded 96.03% effectiveness. Scalability testing covered pipeline construction and retrieval across 30 census inputs and 28 monographs, but did not assess concurrent users or full end-to-end production workloads. These findings provide preliminary evidence that route-aware RAG can improve transparent, evidence-grounded access to heterogeneous village data while preserving deterministic handling of numerical queries and privacy-sensitive information.

Keywords: Data Desa Presisi, evidence-grounded RAG, privacy-aware routing, source provenance, structured-data QA

1. INTRODUCTION

Data Desa Presisi (DDP) integrates participatory census, spatial mapping, digital technology, and community-based qualitative inquiry to produce village information at individual, household, family, building, and territorial levels [1]. Its by-name, by-address, and by-coordinate characteristics support precise village planning but also create privacy obligations because the precision of data collection does not imply unrestricted public disclosure.

DDP knowledge is represented in two complementary forms. Village monographs contain historical narratives, institutional descriptions, maps, diagrams, tables, participatory findings, and interpretations of local conditions. Census records provide structured demographic, household, educational, occupational, welfare, infrastructure, and building attributes. Manual access requires users to search long PDF documents for contextual evidence and inspect structured data separately for numerical facts. Citizens generally require public monograph information and non-identifying aggregates, whereas authorized institutional users may require more detailed information under appropriate access controls. The primary user group evaluated in this study comprised village officials and staff using public mode; broader citizen access and authorized institutional use are intended deployment contexts rather than independently evaluated user populations.

Indonesian village-information initiatives have mainly focused on administrative services and digital public-service delivery. The Cihuni system supports web-based village administration [2], while Krandegan has implemented integrated smart-village services and an artificial-intelligence chatbot for continuous information access [3], [4]. The Popalia system applies generative artificial intelligence to administrative automation and interactive public services [5]. These systems demonstrate the value of conversational technology for villages, but they do not report an evidence-grounded architecture that combines narrative monographs, deterministic census querying, document retrieval, provenance, and privacy-aware routing.

Retrieval-Augmented Generation retrieves external information before response generation [6]. Dense retrieval captures semantic similarity [7], multilingual embeddings such as BGE-M3 support heterogeneous text [8], and FAISS enables vector similarity

search [9]. Lexical retrieval remains useful for explicit village names, years, administrative labels, table headings, and indicator terms. Reciprocal Rank Fusion combines independent rankings without requiring directly comparable retrieval scores [10].

Table 1. Comparison with Indonesian village-information and conversational systems

System	Main purpose	Evaluation	Distinctiveness
Cihuni [2]	Administrative services	User Acceptance Testing	No evidence-grounded monograph-census QA
Krandegan [3], [4]	Digital services and chatbot access	Qualitative and operational reporting	No reported route-aware RAG evaluation
Popalia [5]	AI-assisted administration	Functional and usability testing	Focuses on workflow and document automation
Proposed system	Monograph and census access	Layered technical and usability evaluation	Answer quality evaluated in two villages

Document retrieval alone is insufficient for DDP. Exact census questions require deterministic village filters, indicators, categories, household roles, aggregation operators, and units. Structured-data question answering similarly benefits from constraints that preserve valid database operations [11]. Conversely, database querying alone cannot explain historical, geographical, institutional, or participatory information contained in monographs.

A question such as “How many houses have shingle roofs?” requires a structured aggregate. A request to identify the buildings requires a more detailed lookup and an access-policy decision. Likewise, “How many household heads in Setu hold a doctoral degree?” can be answered as an aggregate, whereas a follow-up request for names changes the request into restricted individual-record access.

In this study, route-aware, evidence-grounded RAG refers to system-level orchestration rather than dense-lexical fusion alone. Structured questions use deterministic DuckDB queries, narrative questions use document retrieval, and mixed questions combine verified structured facts with monograph passages. FAISS, Elasticsearch, and Hybrid RRF

are implemented and evaluated as alternative document-retrieval strategies. Privacy-sensitive, cross-village, out-of-scope, and insufficient-evidence requests are redirected before ordinary answer generation.

Local deployment was selected to retain institutional control over village sources. However, local execution alone does not eliminate privacy leakage, malicious retrieved content, or provenance exposure [12], [13], [14]. The architecture therefore combines local models with village-scope controls, access-aware routing, refusal behavior, evidence validation, and human-readable source cards. This study addresses three research questions:

- 1) RQ1: How can narrative village monographs and structured census records be integrated within a locally controlled, route-aware, and evidence-grounded RAG architecture?
- 2) RQ2: How do FAISS, Elasticsearch, and Hybrid RRF perform at different generation temperatures for DDP questions?
- 3) RQ3: To what extent do deterministic routing and structured querying preserve numerical correctness, source provenance, functional behavior, privacy-policy compliance, and public-mode usability?

The contributions are an architecture that integrates deterministic and retrieval-based evidence paths, a layered evaluation separating textual and deterministic measures, an empirical comparison of nine retrieval-temperature configurations, and an application of locally deployed RAG to heterogeneous DDP knowledge. Answer-quality evaluation was limited to Setu and Gunung Putri, and the artefact is positioned as a research prototype rather than a fully validated public service.

2. METHODS

2.1. Research Design, Data Sources and Study Scope

The study followed Design Science Research Methodology, comprising problem identification, definition of objectives, artefact design, demonstration, evaluation, and communication [15]. The artefact was required to accept natural-language questions, distinguish structured and narrative information needs, maintain village scope, preserve

authoritative values and units, present provenance, apply access restrictions, and operate through locally controlled components.

The production case studies used the 126-page Setu monograph, the 142-page Gunung Putri monograph, and structured census records materialized in DuckDB. Public access covered monograph information and non-identifying census aggregates. Restricted names, identification numbers, household identifiers, complete addresses, and individual coordinates were excluded from the public retrieval corpus. End-to-end answer-quality evaluation was limited to Setu and Gunung Putri. Additional village artefacts were used only for pipeline scalability validation and were not treated as evidence of answer-quality generalization.

2.2. Ground-Truth Construction

The evaluation set was developed through LLM-assisted candidate generation followed by source-based author verification. ChatGPT received the monograph scope, available census indicators, supported evidence routes, and intended conversational and policy behaviors; its outputs were treated as candidate questions and draft answers rather than authoritative annotations. The first author removed duplicates, assigned identifiers and categories, specified the active village, expected answer, source type, expected behavior, required and prohibited elements, privacy expectation, and metric applicability, and verified the final 100-question set against the Setu and Gunung Putri monographs and DuckDB outputs. The questions were purposively constructed to exercise supported evidence routes and policy behaviors rather than sampled from production logs. The final set was reviewed by the first author and was not independently reviewed by village-data experts or double-annotated. Consequently, inter-annotator agreement was not calculated.

Table 2. Distribution and expected behavior of the evaluation questions

Group	n	Illustrative purpose	Expected behavior
DB	46	Structured census value	DuckDB query and deterministic value
MONO	24	Narrative monograph information	Document retrieval
HYB	14	Structured fact and monograph context	Mixed-evidence assembly

Group	n	Illustrative purpose	Expected behavior
CONV	3	Capability or clarification	Controlled conversational response
CV	4	Village-scope conflict	Scope clarification
OOS	1	Unsupported topic	Safe out-of-scope response
PRIV	8	Restricted personal information	Refusal or authorization guidance
Total	100		

Note: RAGAS and ROUGE-L were applied to 84 DB, MONO, and HYB questions. The remaining 16 questions were evaluated through deterministic behavioral checks.

Illustrative question patterns included a village-population aggregate for DB, the historical origin of the name Setu for MONO, a housing-and-electricity question requiring structured and narrative evidence for HYB, a system-capability request for CONV, a question referring to another village while the current village remained active for CV, an unrelated topic for OOS, and a request for resident names or identifiers for PRIV.

2.3. Knowledge Preparation

The monograph pipeline performed page extraction, layout normalization, hierarchy reconstruction, identification of tables and visual-object contexts, quality auditing, and construction of retrievable semantic units. It used structure-aware boundaries rather than a fixed sliding-window chunk size. Sections, table facts, indicator descriptions, captions, and adjacent explanatory text formed semantic units; no fixed token overlap was applied. Each unit retained its village, document, topic path, page range, source type, and text. The final monograph knowledge base contained 355 semantic units: 169 from Setu and 186 from Gunung Putri. Tables were represented as textual facts when their rows could be extracted reliably, while captions were associated with adjacent explanatory text. The system did not claim independent interpretation of unavailable image content.

Census preparation comprised schema profiling, column normalization, type conversion, village alignment, privacy classification, aggregate construction, and DuckDB materialization [16]. Numerical questions were resolved into village, indicator, category, unit, analytical level, role, and aggregation operation. Verified counts and units were computed deterministically; the language model verbalized them only when narrative

composition was required. The canonical production corpus contained 451 monograph-type records and 8,821 processed public census facts, producing 9,272 canonical records. The same identifiers were used in FAISS and Elasticsearch without missing or duplicate identifiers.

2.4. Route-Aware Architecture

The router combined deterministic guards and local semantic classification. Deterministic checks first evaluated active-village scope, privacy-sensitive content, out-of-scope requests, and resolvable structured indicators. Questions not conclusively resolved by these guards were processed by a locally deployed Llama 3.1 8B semantic intent classifier, after which the predicted intent was checked against indicator, evidence-source, and access-policy rules. Structured questions were directed to DuckDB, narrative questions to document retrieval, and mixed questions to both evidence paths. Conversational, ambiguous, village-mismatch, out-of-scope, insufficient-evidence, and privacy-sensitive requests used dedicated controlled responses. The document route implemented FAISS dense retrieval [9], Elasticsearch BM25 retrieval [17], and Hybrid RRF [10]. Production used Elasticsearch as its default document retriever based on the observed E1–E9 results; it did not run all three retrieval strategies for every request.

Verified structured values, retrieved passages, units, active-village scope, source type, document, page, and indicator metadata were assembled into an evidence package. Route-specific answer instructions required the generator to rely on this package, preserve verified values and units, avoid unsupported inference, and present human-readable provenance. When evidence was insufficient, village scope required clarification, or a request involved restricted information, the system returned a limitation, clarification, refusal, or authorization-guidance response instead of ordinary answer generation. Post-generation validation checked numerical and categorical claims, preserved authoritative structured facts, removed internal control text, normalized provenance, and constructed source cards [14]. Figure 1 summarizes this architecture.

Figure 1 separates question-level evidence routing from the choice among document-retrieval strategies. The application-level threat model considered prompt injection, accidental disclosure, unauthorized access, operational-log exposure, and source-card leakage [12], [13]. Implemented safeguards included route-first policy checks, exclusion of

personally identifiable records from the public corpus, village-scope enforcement, refusal or authorization guidance, evidence-package constraints, and source-card sanitization. Privacy-policy evaluation used eight predefined scenarios. Adversarial prompt-injection, data-exfiltration, authentication- or authorization-bypass, formal log-security auditing, and penetration testing were not performed.

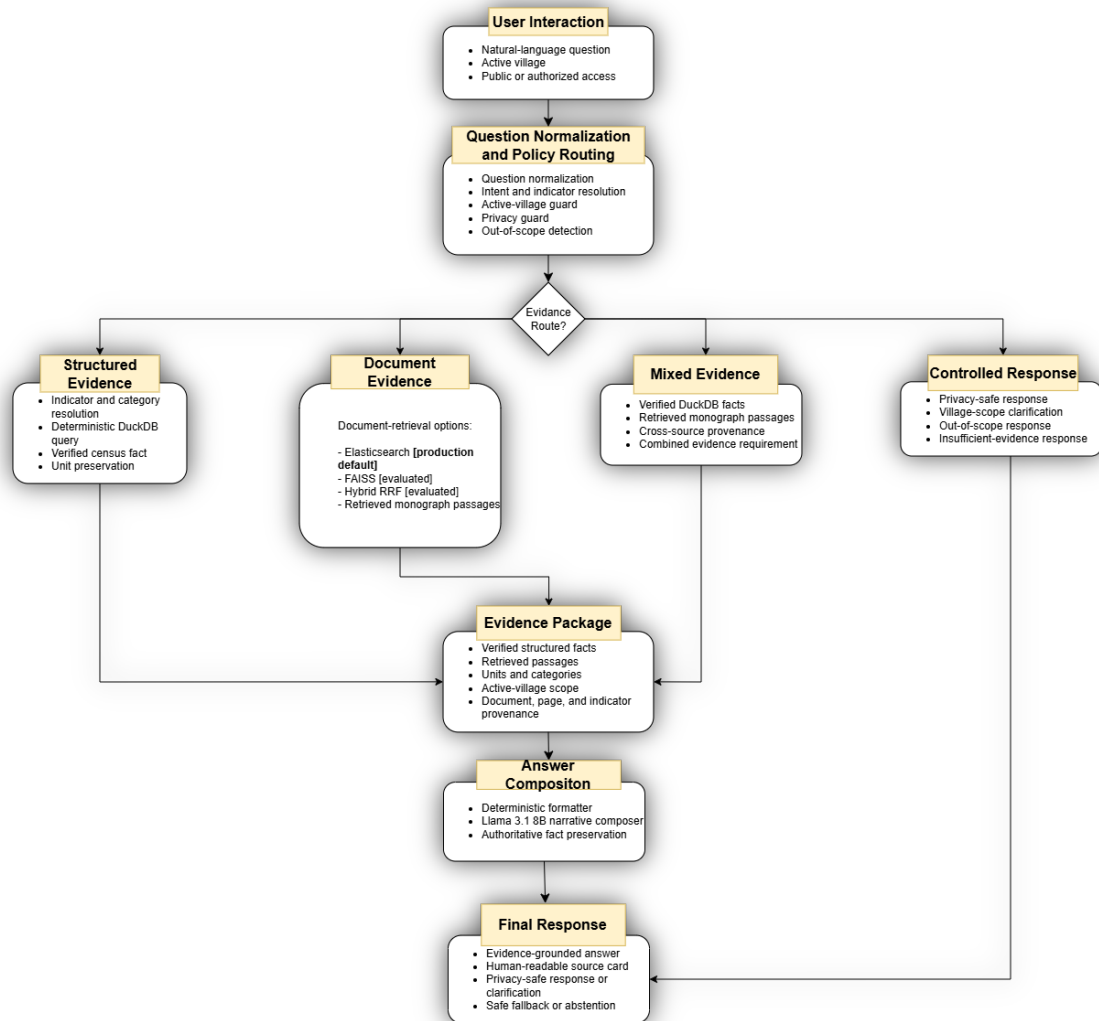


Figure 1. Route-aware, evidence-grounded architecture integrating deterministic census querying, document retrieval, policy controls, and source provenance.

2.5. Implementation and Experimental Design

The prototype ran on Ubuntu 22.04.4 LTS using an AMD Ryzen 7 5800X processor, 125 GiB RAM, and an NVIDIA RTX 3060 with 12 GiB VRAM. The stack included Python 3.10.12, FastAPI 0.115.0, Streamlit 1.37.1, DuckDB 1.0.0, FAISS 1.7.4, Elasticsearch 7.17.25, and RAGAS

0.4.3. Llama 3.1 8B served as the local runtime model, Qwen2.5 14B as the local RAGAS evaluator, and BAAI/bge-m3 as the 1,024-dimensional embedding model [8]. FAISS used an IndexFlat inner-product index, while Elasticsearch used the standard analyzer and BM25 over the same canonical records [17]. No additional embedding-normalization or custom field-boosting configuration was documented for the evaluated experiments; therefore, these settings were not assumed in the reported implementation.

The experiment used a 3×3 design combining FAISS, Elasticsearch, and Hybrid RRF with temperatures 0.1, 0.2, and 0.3. Each document retriever returned three candidates. Hybrid mode fused FAISS and Elasticsearch rankings with RRF using a rank-smoothing constant of 60. Generation used top-p 0.9, a 2,048-token context window, and route-dependent output limits of approximately 180–760 tokens. Each configuration generated one response set without a fixed random seed. The design therefore compared observed runs rather than repeated estimates of stochastic variability. The E1-E9 matrix constituted a controlled retrieval-temperature ablation within the route-aware architecture. It did not include separate LLM-only, no-routing, DuckDB-only, or generation-free architectural baselines.

Table 3. Compact implementation and experimental configuration

Component	Configuration
Runtime/evaluator	Llama 3.1 8B / Qwen2.5 14B, local Ollama
Embedding	BAAI/bge-m3, 1,024 dimensions
Dense retrieval	FAISS IndexFlat, inner product, top 3
Lexical retrieval	Elasticsearch 7.17.25, standard analyzer, BM25, top 3
Hybrid retrieval	FAISS + Elasticsearch, RRF k = 60
Corpus	9,272 canonical records
Temperatures	0.1, 0.2, 0.3
Generation	top-p 0.9; context 2,048; 180–760 output tokens
Generation seed	Not fixed
Production default	Elasticsearch, temperature 0.1

2.6. Evaluation Procedures

RAGAS measured faithfulness, answer relevancy, context precision, and context recall for the 84 DB, MONO, and HYB questions [18]. Non-applicable values from the remaining conversational and policy-oriented questions were not converted into zeros. The four RAGAS dimensions were combined into an unweighted descriptive composite for each experimental configuration i :

$$Composite_i = \frac{F_i + AR_i + CP_i + CR_i}{4} \quad (1)$$

where F_i , AR_i , CP_i , and CR_i denote the mean faithfulness, answer relevancy, context precision, and context recall of configuration i , respectively. The composite was used only as a descriptive summary of the four RAGAS dimensions and was not treated as a universally validated measure of overall answer accuracy. ROUGE-L F1 measured sequence-level similarity between the generated and reference answers using the longest common subsequence [19]. ROUGE-L was reported separately and was not included in the composite score.

Deterministic evaluation measured functional pass, source pass, must-contain pass, must-not-contain pass, numeric match, unit pass, village-context behavior, out-of-scope behavior, and privacy-policy compliance. Numeric and unit matching were applicable to 48 questions: all 46 DB questions and two HYB questions, SETU-HYB-001 and GP-HYB-001, that contained directly verifiable single structured values. The original must-not-contain check used case-insensitive literal substring matching against the generated answer. Privacy-sensitive questions expected refusal, limitation, or authorization guidance rather than disclosure. Configuration-level deterministic acceptance results are reported for E4 because it was selected as the operational configuration, whereas the comparative E1-E9 analysis focused on RAGAS and ROUGE-L over the 84 applicable questions.

The runtime's fine-grained routes did not map one-to-one to the seven evaluation groups. Several internal routes could validly satisfy the same high-level evidence requirement. A conventional single-label confusion matrix would therefore classify some acceptable routes as errors. Routing was evaluated through the observed category-route

relationship together with functional outcomes, source compliance, village-scope behavior, and privacy-policy outcomes.

Passage-level retrieval metrics such as Recall@k, MRR, and nDCG were not calculated for the E1-E9 answer-quality experiment because ranked passage identifiers and independent passage-relevance judgments were not retained consistently across configurations. These metrics were therefore not retrospectively reconstructed for the E1-E9 comparison.

Nonparametric bootstrap confidence intervals used 10,000 question-level resamples for the composite and ROUGE-L results. Wilson score intervals were calculated for binary proportions. Paired Wilcoxon signed-rank tests compared E4 with each of the other eight configurations, followed by Holm correction for multiple comparisons. These tests were supplementary post hoc analyses and were not used to claim that E4 was statistically superior.

2.7. Usability and Pipeline Validation

Public-mode usability evaluation involved 21 village officials and staff, comprising eight participants from Setu and 13 from Gunung Putri. Each participant completed six tasks covering active-village selection, monograph access, geographical information, public census aggregates, village potentials or development problems, and source inspection. Testing was conducted exclusively in public mode, and participants were not instructed to retrieve restricted personal records. The source-inspection task assessed whether participants could locate the displayed provenance information; it did not independently assess source-authority comprehension, evidential usefulness, or trust calibration.

Effectiveness used binary task completion x_{ij} , where 1 denoted successful completion and 0 denoted non-completion. Efficiency used an assistance score s_{ij} , where 3 denoted independent completion, 2 minor assistance, 1 substantial assistance, and 0 non-completion. Satisfaction used 15 five-point Likert items q_{il} . The measures were calculated as shown in Equation (2) to (5).

$$Effectiveness (\%) = \left(\frac{\sum_{i=1}^N \sum_{j=1}^T x_{ij}}{N \times T} \right) \times 100 \quad (2)$$

$$Efficiency (\%) = \left(\frac{\sum_{i=1}^N \sum_{j=1}^T s_{ij}}{3 \times N \times T} \right) \times 100 \quad (3)$$

$$Satisfaction (\%) = \left(\frac{\sum_{i=1}^N \sum_{l=1}^Q q_{il}}{5 \times N \times Q} \right) \times 100 \quad (4)$$

$$Overall Usability (\%) = \frac{Effectiveness + Efficiency + Satisfaction}{3} \quad (5)$$

where $N = 21$, $T = 6$, and $Q = 15$. The resulting overall value was treated as a study-specific descriptive index rather than a standardized System Usability Scale score. The evaluation dimensions followed ISO 9241-11 [20], and the task-observation and questionnaire procedures were adapted from Nurhadryani et al. [21].

Pipeline scalability validation used five staged scenarios—S5, S10, S20, S28, and S30—to examine preprocessing completion, canonical-corpus construction, FAISS and Elasticsearch index parity, retrieval accuracy, and retrieval-only latency as the number of village artefacts increased. S30 contained census inputs for 30 villages but monographs for 28 villages. The validation did not measure concurrent users, language-model generation latency, end-to-end interaction latency, or answer quality for the additional villages.

3. RESULTS AND DISCUSSION

3.1. Developed Artefact and Routing Outcomes

The study produced an operational route-aware RAG prototype for Setu and Gunung Putri. Its runtime integrated a web interface, FastAPI, deterministic DuckDB querying, FAISS and Elasticsearch retrieval, local generation, evidence validation, and source cards. The public corpus contained 9,272 canonical records and excluded personally identifiable census records. Figure 2 presents the public-mode desktop and mobile interface. It displays the active village, question, answer, and human-readable source provenance. Authorized access was represented as a separate policy context but was not evaluated in the public-mode usability study.

All nine experimental configurations produced responses for the same 100 question identifiers. The evaluation groups represented high-level evidence or policy

requirements, whereas runtime routes represented finer-grained internal processing decisions. Their relationship was therefore not strictly one-to-one. Table 4 presents the observed category–route relationship under E4.

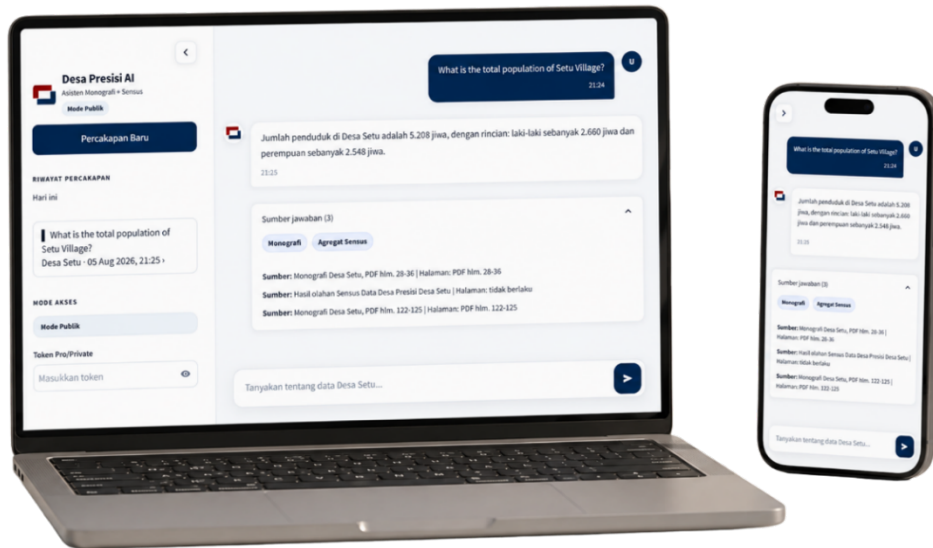


Figure 2. Final public-mode desktop and mobile interface showing an evidence-grounded answer and human-readable source provenance.

Structured database questions were predominantly handled through structured-indicator queries, with smaller numbers processed through exact structured queries or general evidence composition. Monograph questions were distributed across narrative, table-summary, concept-explanation, and general document-evidence routes. All mixed-evidence questions used cross-source evidence composition. Village-context and out-of-scope questions were assigned dedicated controlled routes. Privacy-sensitive questions were associated with structured-lookup intent, but access-policy checks overrode ordinary lookup completion and returned controlled responses without disclosing restricted records.

Table 4. Observed relationship between evaluation groups and runtime routes under E4

Evaluation group	n	Observed runtime routes and outcomes
DB	46	structured_indicator_query (38); exact_structured_primary (6); open_rag (2)
MONO	24	narrative_monograph_primary (9); open_rag (7); pdf_table_summary (5); ddp_concept_explanation (3)

Evaluation group	n	Observed runtime routes and outcomes
HYB	14	open_rag (14)
CONV	3	small_talk_identity (2); clarification_needed (1)
CV	4	village_scope_mismatch (4)
OOS	1	out_of_scope_safe_response (1)
PRIV	8	Structured-lookup intent with controlled privacy response (8)

Note: Evaluation groups represent expected evidence or policy behavior, whereas runtime routes represent internal processing decisions. Multiple fine-grained routes may validly satisfy one high-level evaluation requirement; therefore, this table is not interpreted as a conventional one-label confusion matrix.

3.2. Retrieval-Temperature Performance

Table 5 reports the E1–E9 results. E4, Elasticsearch at temperature 0.1, obtained the highest observed composite score of 0.592018. E2 produced the highest faithfulness, E3 the highest answer relevancy, and E8 the highest ROUGE-L. The metric rankings therefore did not identify one configuration as dominant across all dimensions.

Table 5. RAGAS, ROUGE-L, and composite results

Config.	Retrieval	Temp.	Faith.	Ans. rel.	Ctx. prec.	Ctx. rec.	Composite	ROUGE-L
E1	FAISS	0.1	0.359413	0.759016	0.640450	0.567460	0.581585	0.466270
E2	FAISS	0.2	0.367257	0.763418	0.640450	0.567460	0.584646	0.469139
E3	FAISS	0.3	0.340195	0.764466	0.640450	0.567460	0.578143	0.472154
E4	Elasticsearch	0.1	0.365721	0.761206	0.655827	0.585317	0.592018	0.478107
E5	Elasticsearch	0.2	0.358877	0.763371	0.640450	0.567460	0.582540	0.468040
E6	Elasticsearch	0.3	0.357465	0.748037	0.640450	0.567460	0.578353	0.459249
E7	Hybrid RRF	0.1	0.327817	0.753799	0.640450	0.567460	0.572382	0.480773
E8	Hybrid RRF	0.2	0.362026	0.734385	0.640450	0.567460	0.576081	0.483727
E9	Hybrid RRF	0.3	0.349938	0.729566	0.640450	0.567460	0.571854	0.461011

Note: Metrics were calculated from 84 applicable questions. Composite is the mean of four RAGAS dimensions; ROUGE-L was reported separately.

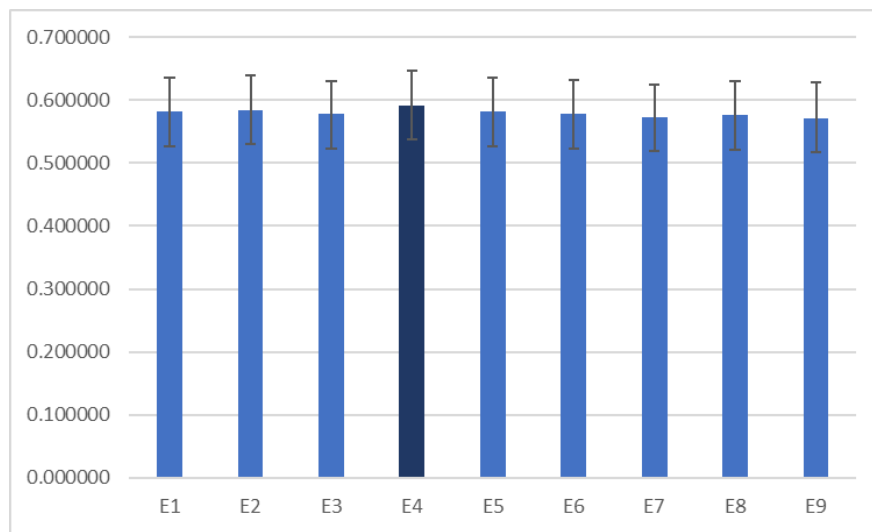


Figure 3. Observed composite scores and 95% bootstrap confidence intervals across nine retrieval-temperature configurations.

Question-level context auditing showed identical stored contexts for E4 and E7 across all 100 questions. Within each retrieval mode, stored contexts were identical for 98–100% of the questions across generation temperatures, while most cross-mode comparisons differed for only two questions. The extensive context overlap coincided with the repeated aggregate context-precision and context-recall values. The few recorded context differences were retained as observed and were not interpreted as evidence that generation temperature systematically changed retrieval. Composite scores ranged from 0.571854 to 0.592018. E4 obtained the highest observed score, with a bootstrap 95% confidence interval of 0.537090–0.646709; its ROUGE-L interval was 0.417287–0.539220. The confidence intervals overlapped substantially, and paired Wilcoxon tests found no significant difference between E4 and the other configurations. Raw p-values ranged from 0.174876 to 0.908923, while all Holm-adjusted p-values were 1.000. E4 was therefore selected descriptively rather than as a statistically proven optimum. Table 6 reports the complete post hoc comparisons.

Table 6. Paired Wilcoxon and Holm-adjusted comparisons between E4 and the other configurations

Comparison	Wilcoxon statistic	Raw p-value	Holm-adjusted p-value	Significant
E4 vs E1	402.0	0.535946	1.000	No

Comparison	Wilcoxon statistic	Raw p-value	Holm-adjusted p-value	Significant
E4 vs E2	449.0	0.591373	1.000	No
E4 vs E3	569.0	0.845487	1.000	No
E4 vs E5	601.0	0.908923	1.000	No
E4 vs E6	360.5	0.505826	1.000	No
E4 vs E7	369.0	0.209174	1.000	No
E4 vs E8	419.0	0.514355	1.000	No
E4 vs E9	343.0	0.174876	1.000	No

Note: Tests were conducted over the 84 question-level composite scores. Statistical significance was evaluated after Holm correction at $\alpha=0.05$.

3.3. Deterministic, Category-Level, and Error Results

E4 achieved 98% functional pass, 99% source pass, 97.92% numeric match, 97.92% unit pass, and policy-compliant responses for all eight privacy questions. The must-contain pass rate was 62%, meaning that 38 responses missed at least one required element.

Table 7. Deterministic and policy results under E4

Metric	Passed/applicable	Result (%)	Wilson 95% CI (%)
Functional pass	98/100	98.00	93.00–99.45
Source pass	99/100	99.00	94.55–99.82
Must-contain pass	62/100	62.00	52.21–70.90
Must-not-contain pass	94/100	94.00	87.52–97.22
Numeric match	47/48	97.92	89.10–99.63
Unit pass	47/48	97.92	89.10–99.63
Privacy-policy pass	8/8	100.00	67.56–100.00

Note: The must-not-contain result reflects the original case-insensitive literal substring rule and may therefore include lexical false positives. The reported 94/100 result is retained unchanged for traceability. Privacy-policy pass covers eight predefined scenarios and does not represent security certification.

Table 8 shows that DB questions had the highest answer relevancy and ROUGE-L but the lowest faithfulness. Thirty-four of the 46 DB questions received zero textual faithfulness.

Across the 48 numerically applicable questions, comprising 46 DB questions and two HYB questions, numeric and unit matching were both 47/48.

Table 8. E4 results by RAGAS-applicable group

Group	n	Faithfulness	Answer relevancy	Context precision	Context recall	ROUGE-L
DB	46	0.202899	0.849210	0.637854	0.586957	0.586437
HYB	14	0.511593	0.607238	0.826389	0.369048	0.180236
MONO	24	0.592706	0.682349	0.590782	0.708333	0.444233

Table 9. Representative E4 success and failure cases

Case and question	Retrieved evidence	Generated and expected answer	Evaluation outcome
GP-DB-001 — “Berapa jumlah penduduk Desa Gunung Putri?”	Structured census evidence: “Jumlah penduduk: 14.764 jiwa.”	Generated: “Jumlah penduduk Desa Gunung Putri adalah 14.764 jiwa.” Expected: the same value and unit.	Functional, source, numeric, and unit checks passed; ROUGE-L = 1.0, while textual faithfulness = 0.
SETU-MONO-008 — “Jelaskan asal-usul nama Desa Setu.”	A general village and census overview was selected instead of the required historical monograph passage.	Generated excerpt: “Desa Setu dapat dijelaskan dari beberapa sisi...” followed by general census information. Expected: the name Setu derives from a Sundanese term for a lake or reservoir and refers to a former lake near Bukit Angsana.	Must-contain failure; the required historical evidence was not selected for the benchmark wording.
SETU-DB-021 — “Berapa keluarga yang mengikuti program KB di Desa Setu?”	Generic household and census evidence rather than the required family-planning indicator.	Generated excerpt: “Desa Setu memiliki 1.594 keluarga...” followed by unrelated population statistics. Expected: 500 KK participating in family planning.	Numeric and unit checks failed; the required structured indicator was not resolved correctly.

Case and question	Retrieved evidence	Generated and expected answer	Evaluation outcome
SETU-HYB-006 – "Bagaimana gambaran kondisi papan dan listrik rumah tangga di Desa Setu?"	Retrieved contexts included relevant housing and electricity evidence.	Generated: a generic household-condition response that omitted the required values. Expected: 979 KK living in owner-occupied homes and 275 KK using PLN 900 W electricity.	Mixed-evidence composition failure; relevant evidence was not preserved in the final response.

3.4. Usability and Pipeline-Scalability Results

Participants completed 121 of 126 tasks. Overall effectiveness was 96.03%, efficiency was 76.98%, satisfaction was 80.51%, and the descriptive usability index was 84.51%. T05, involving open-ended village-potential questions, produced the lowest effectiveness and efficiency. All participants completed the source-inspection task, demonstrating that they could locate the displayed provenance information. The study did not independently evaluate whether participants correctly judged source authority, evidential usefulness, or trustworthiness.

Table 10. Public-mode usability results

Evaluation scope	Participants	Effectiveness (%)	Efficiency (%)	Satisfaction (%)	Overall usability (%)
Setu	8	100.00	84.03	85.17	89.73
Gunung Putri	13	93.59	72.65	77.64	81.29
Overall	21	96.03	76.98	80.51	84.51

Note: The overall value is study-specific and is not a standardized SUS score. Evaluation covered public mode only.

Pipeline construction completed all five scenarios. Indexed items increased from 1,663 at S5 to 5,842 at S28 and remained unchanged at S30 because only 28 monographs were available. FAISS and Elasticsearch retained identical item counts.

Table 11. Pipeline scalability-validation results

Stage	Census villages	Monographs	Monograph chunks	Census Facts	Index items	Status
S5	5	5	863	800	1,663	Pass
S10	10	10	1,913	800	2,713	Pass
S20	20	20	3,769	800	4,569	Pass
S28	28	28	5,042	800	5,842	Pass
S30	30	28	5,042	800	5,842	Pass

Note: S28 reports the completed resumed run. FAISS and Elasticsearch had equal item counts. The validation measured pipeline construction and retrieval, not concurrent users or multi-village end-to-end answer quality.

All census and monograph benchmark groups achieved Hit@5 of 1.0. FAISS and Hybrid RRF achieved Top-1 accuracy of 1.0 across the benchmark groups, whereas Elasticsearch monograph Top-1 accuracy ranged from 0.80 to 0.95. These values belong to the separate scalability benchmark, not the E1-E9 answer-quality experiment.

Elasticsearch recorded the lowest retrieval-only latency across the scalability scenarios. Its mean latency ranged from approximately 5.0 to 5.7 ms for census retrieval and from 5.4 to 6.5 ms for monograph retrieval. FAISS mean latency ranged from approximately 76.6 to 78.6 ms for census retrieval and from 67.5 to 77.2 ms for monograph retrieval. Hybrid RRF required approximately 81.5–84.2 ms for census retrieval and 73.5–82.7 ms for monograph retrieval. At S30, p95 latency was 5.82–6.50 ms for Elasticsearch, 80.79–82.72 ms for FAISS, and 86.61–88.30 ms for Hybrid RRF. These measurements covered retrieval execution only and excluded routing, DuckDB processing, language-model generation, post-generation validation, network transfer, and interface rendering.

3.5. Discussion

3.5.1. Retrieval, Routing, and Structured Evidence

E4 produced the highest observed composite score, but the difference was small and non-significant after Holm correction. Elasticsearch at temperature 0.1 was therefore retained as the production document-retrieval default on an operational basis, reflecting

the highest observed E1–E9 composite rather than statistically proven or universal superiority.

DDP questions frequently contain explicit administrative terms, indicator names, years, village names, and headings that may favor BM25 matching, while dense retrieval remained competitive. Hybrid RRF did not improve the composite in the evaluated run. The stored E4 and E7 contexts were identical for all 100 questions, suggesting that fusion often provided limited additional evidence within the final context. Because RRF combines ranked evidence from multiple retrieval systems, its benefit depends on the component rankings contributing useful complementary items [10]. This observation does not establish a causal explanation because independent passage-relevance judgments and RRF parameter ablations were not performed. The fixed top-three retrieval depth and untuned RRF constant of $k = 60$ may also have influenced the observed result.

Importantly, Hybrid RRF is one document-retrieval strategy, whereas the route-aware RAG contribution of this study concerns system-level orchestration among deterministic structured, document, mixed-evidence, and controlled-response paths. Therefore, the absence of a Hybrid RRF advantage does not by itself imply that heterogeneous evidence orchestration was ineffective.

RAGAS faithfulness was evaluated against the supplied textual contexts rather than the entire knowledge base. For structured routes, these contexts were often concise field-value evidence; numerically correct answers could therefore receive low textual faithfulness despite passing value, unit, and source checks. Thirty-four of the 46 DB questions received zero faithfulness, while numeric and unit matching across the 48 applicable questions were both 47/48. Because claim-level evaluator reasoning was not retained, no single cause is asserted. Textual grounding and deterministic structured-data checks should therefore be interpreted as complementary measures [18], [22], [23].

3.5.2. Provenance, Privacy, and User Implications

The 99% source-pass rate shows that expected evidence types were usually presented, but the 62% must-contain rate shows that source presence did not guarantee complete answer composition. RAG systems can still omit or contradict retrieved evidence [24], while citation visibility alone does not establish claim-level support [14], [25].

All eight privacy-sensitive scenarios produced controlled responses. However, the evaluation covered application-policy behavior rather than formal security resistance. Prompt injection remained a potential risk because malicious instructions could attempt to override evidence or access rules. Accidental disclosure could arise from incorrect privacy classification, while unauthorized access could result from weaknesses in authentication or role enforcement. Operational logs could expose sensitive questions if they were not protected, and source cards could leak restricted provenance if private records entered the public evidence layer.

The implemented controls included route-first policy checks, exclusion of personally identifiable records from the public corpus, village-scope enforcement, refusal or authorization guidance, evidence-package constraints, and source-card sanitization. Nevertheless, prompt-injection attacks, database exfiltration, authorization-bypass testing, formal log-security auditing, and penetration testing were outside the evaluated scenarios. The prototype is therefore described as privacy-aware at the application-policy level, not as formally privacy-secure [12], [13].

Usability results indicate that village officials and staff could complete most public-mode tasks, although assistance was more frequently required for open-ended questions. All participants completed the source-inspection task, but this establishes source-card discoverability rather than independent validation of source authority, evidential usefulness, or trust calibration. The usability findings therefore apply to the evaluated institutional participants using public mode and should not be generalized to all citizens or authorized private-mode users.

3.5.3. Scalability, Governance, and Limitations

Pipeline validation showed that preprocessing and index construction remained operational as source artefacts increased. It did not assess concurrent requests, GPU contention, model queueing, service saturation, or end-to-end response time. The 5,842-item S30 corpus is also not directly comparable with the 9,272-record production corpus because the scalability scenarios retained only 800 retrieval-oriented census facts.

Operational deployment requires governance for source updates, version periods, aggregate approval, index rebuilding, authorization, backups, and security monitoring.

Census records may be updated more frequently than monographs, and systematic conflict-resolution evaluation remains future work.

The study has several limitations. Answer-quality evaluation covered only Setu and Gunung Putri. The 100-question set was purposively constructed with LLM assistance and verified by the first author without independent expert review or double annotation. Each E1-E9 configuration was generated once without a fixed seed. Ranked passage identifiers and independent relevance judgments were not retained consistently, preventing retrospective Recall@k, MRR, and nDCG. The E1-E9 matrix evaluates retrieval and temperature within the route-aware architecture but does not isolate orchestration against LLM-only, no-routing, DuckDB-only, or generation-free baselines. Usability involved 21 village officials and staff in public mode only, and privacy testing comprised eight predefined policy scenarios rather than adversarial security testing.

Future evaluation should retain ranked passage identifiers and independent relevance judgments from the outset, define controlled architectural baselines prospectively, and use independent expert or multi-annotator ground-truth review. Further work should examine source-card comprehension, broader villages and user groups, adversarial prompt injection and authorization controls, monograph-census conflicts, repeated generation runs, and end-to-end load performance.

4. CONCLUSION

This study developed and evaluated a locally deployed, route-aware, evidence-grounded RAG chatbot integrating narrative village monographs with deterministic census querying for Data Desa Presisi. Answer-quality evaluation was limited to Setu and Gunung Putri and used 100 ground-truth questions. Elasticsearch at temperature 0.1 achieved the highest observed descriptive composite score of 0.592018, although post hoc Wilcoxon-Holm analyses did not establish statistical superiority. The selected configuration achieved 98% functional pass, 99% source pass, and 47/48 numeric and unit matches, while all eight predefined privacy-policy scenarios produced compliant responses. Usability testing was limited to public mode and involved 21 village officials and staff, reaching 96.03% effectiveness and an 84.51% study-specific usability index. Pipeline validation covered 30 census inputs and 28 monographs and supports preprocessing and

retrieval-pipeline extensibility rather than concurrent-user capacity or multi-village answer-quality generalization. The practical contribution is a research prototype providing a provenance-aware interface for public monograph information and non-identifying census aggregates while separating restricted data. Broader deployment requires independent expert validation, stronger architectural comparisons, additional village and user evaluation, adversarial security testing, and end-to-end load assessment.

ACKNOWLEDGMENT

The authors acknowledge IPB University, the Data Desa Presisi research ecosystem, and the participating village stakeholders for supporting this study.

REFERENCES

- [1] S. Sjaif *et al*, "Data Desa Presisi: A new method of rural data collection," *MethodsX*, vol. 9, Art. no. 101868, 2022, doi: 10.1016/j.mex.2022.101868.
- [2] K. F. Siman and J. Wiratama, "A web-based village administrative information system for improving service quality toward the smart village concept," *G-Tech: J. Teknol. Terap*, vol. 7, no. 4, pp. 1519–1528, 2023, doi: 10.33379/gtech.v7i4.3196.
- [3] N. Saidah, L. Khasanah, Asriyatzahra, and S. Ridloah, "Analisis strategi kesuksesan Kampung Digital Krandegan dalam mendukung program smart village," *J. Reg. Rural Dev. Plan.*, vol. 6, no. 2, pp. 123–135, 2022, doi: 10.29244/jp2wd.2022.6.2.123-135.
- [4] Pemerintah Desa Krandegan, "Pemerintah Desa Krandegan maksimalkan pelayanan warga dengan chatbot artificial intelligence (AI)," 2025.
- [5] A. Arafat, R. Rasyid, S. Hestiana, R. Rendi, S. Bantun, and J. Y. Sari, "Designing an AI-based village information system using research and development approach for public governance modernization in Popalia Village," *J. Tek. Inform. (JUTIF)*, vol. 6, no. 5, pp. 5251–5269, 2025, doi: 10.52436/1.jutif.2025.6.5.5257.
- [6] P. Lewis *et al*, "Retrieval-augmented generation for knowledge-intensive NLP tasks," *Adv. Neural Inf. Process. Syst.*, vol. 33, pp. 9459–9474, 2020.
- [7] V. Karpukhin *et al*, "Dense passage retrieval for open-domain question answering," in *Proc. 2020 Conf. Empirical Methods Nat. Lang. Process. (EMNLP)*, 2020, pp. 6769–6781, doi: 10.18653/v1/2020.emnlp-main.550.

- [8] J. Chen, S. Xiao, P. Zhang, K. Luo, D. Lian, and Z. Liu, "M3-Embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation," in *Findings Assoc. Comput. Linguistics: ACL 2024*, 2024, pp. 2318–2335, doi: 10.18653/v1/2024.findings-acl.137.
- [9] J. Johnson, M. Douze, and H. Jégou, "Billion-scale similarity search with GPUs," *IEEE Trans. Big Data*, vol. 7, no. 3, pp. 535–547, 2021, doi: 10.1109/TBDATA.2019.2921572.
- [10] G. V. Cormack, C. L. A. Clarke, and S. Büttcher, "Reciprocal rank fusion outperforms Condorcet and individual rank learning methods," in *Proc. 32nd Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2009, pp. 758–759, doi: 10.1145/1571941.1572114.
- [11] T. Scholak, N. Schucher, and D. Bahdanau, "PICARD: Parsing incrementally for constrained auto-regressive decoding from language models," in *Proc. 2021 Conf. Empirical Methods Nat. Lang. Process. (EMNLP)*, 2021, pp. 9895–9901, doi: 10.18653/v1/2021.emnlp-main.779.
- [12] S. Zeng *et al.*, "The good and the bad: Exploring privacy issues in retrieval-augmented generation," in *Findings Assoc. Comput. Linguistics: ACL 2024*, 2024, pp. 4505–4524, doi: 10.18653/v1/2024.findings-acl.267.
- [13] Z. Tan *et al.*, "Glue pizza and eat rocks—Exploiting vulnerabilities in retrieval-augmented generative models," in *Proc. 2024 Conf. Empirical Methods Nat. Lang. Process. (EMNLP)*, 2024, pp. 1610–1626, doi: 10.18653/v1/2024.emnlp-main.96.
- [14] J. Qi, G. Sarti, R. Fernández, and A. Bisazza, "Model internals-based answer attribution for trustworthy retrieval-augmented generation," in *Proc. 2024 Conf. Empirical Methods Nat. Lang. Process. (EMNLP)*, 2024, pp. 6037–6053, doi: 10.18653/v1/2024.emnlp-main.347.
- [15] K. Peffers, T. Tuunanen, M. A. Rothenberger, and S. Chatterjee, "A design science research methodology for information systems research," *J. Manage. Inf. Syst.*, vol. 24, no. 3, pp. 45–77, 2007, doi: 10.2753/MIS0742-1222240302.
- [16] M. Raasveldt and H. Mühleisen, "DuckDB: An embeddable analytical database," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, 2019, pp. 1981–1984, doi: 10.1145/3299869.3320212.
- [17] S. Robertson and H. Zaragoza, "The probabilistic relevance framework: BM25 and beyond," *Found. Trends Inf. Retrieval*, vol. 3, no. 4, pp. 333–389, 2009, doi: 10.1561/15000000019.

- [18] S. Es, J. James, L. Espinosa-Anke, and S. Schockaert, "RAGAs: Automated evaluation of retrieval-augmented generation," in *Proc. 18th Conf. Eur. Chapter Assoc. Comput. Linguistics: Syst. Demonstrations*, 2024, pp. 150–158, doi: 10.18653/v1/2024.eacl-demo.16.
- [19] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*, 2004, pp. 74–81.
- [20] International Organization for Standardization, *Ergonomics of Human-System Interaction—Part 11: Usability: Definitions and Concepts*, ISO 9241-11:2018, 2018.
- [21] Y. Nurhadryani, S. K. Sianturi, I. Hermadi, and H. Khotimah, "Pengujian usability untuk meningkatkan antarmuka aplikasi mobile," *J. Ilmu Komput. Agri-Inform*, vol. 2, no. 2, pp. 83–93, 2013, doi: 10.29244/jika.2.2.83-93.
- [22] J. Saad-Falcon, O. Khattab, C. Potts, and M. Zaharia, "ARES: An automated evaluation framework for retrieval-augmented generation systems," in *Proc. 2024 Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol*, 2024, pp. 338–354, doi: 10.18653/v1/2024.naacl-long.20.
- [23] D. Ru *et al*, "RAGChecker: A fine-grained framework for diagnosing retrieval-augmented generation," *Adv. Neural Inf. Process. Syst.*, vol. 37, 2024, doi: 10.52202/079017-0692.
- [24] C. Niu *et al*, "RAGTruth: A hallucination corpus for developing trustworthy retrieval-augmented language models," in *Proc. 62nd Annu. Meeting Assoc. Comput. Linguistics (ACL)*, 2024, pp. 10862–10878, doi: 10.18653/v1/2024.acl-long.585.
- [25] T. Gao, H. Yen, J. Yu, and D. Chen, "Enabling large language models to generate text with citations," in *Proc. 2023 Conf. Empirical Methods Nat. Lang. Process. (EMNLP)*, 2023, pp. 6465–6488, doi: 10.18653/v1/2023.emnlp-main.398.