Journal of Management, Economic, and Financial
Volume 4, No. 2 April 2026
e-ISSN 2986-6863

Hybrid Retrieval-Augmented Generation with Hallucination Mitigation for Indonesian MSME Financial Advisory

Rafi Farizki[1*], Rani Santika[2], Fhrizz S. De Jesus[3]
STMIK LIKMI, Indonesia[1]
Universitas. Muhammadiyah prof Dr hamka, Indonesia[2]
College of Management and Business Technology, NEUST, Philippines[3]
Email: rafifarizki90@gmail.com, rsantika851@gmail.com, fhrizzdejesus01@gmail.com

Abstract

Micro, small, and medium enterprises (MSMEs) play a vital role in Indonesia's economy, yet many business owners continue to experience limited financial literacy and difficulties in accessing accurate, up-to-date financial and regulatory information. This study designed and evaluated IndoRAG-FA, a hybrid Retrieval-Augmented Generation framework with explicit hallucination mitigation for Indonesian MSME financial advisory. The Design Science Research (DSR) methodology was employed to develop a retrieval-grounded advisory framework integrating semantic chunking, hybrid BM25-dense retrieval, reciprocal rank fusion, cross-encoder re-ranking, grounded generation, and confidence-based abstention. A curated knowledge base comprising 12 authoritative Indonesian financial and tax source documents was developed alongside an expert-validated benchmark of 250 MSME financial queries, including 30 out-of-scope items. Experimental results demonstrate that IndoRAG-FA substantially outperforms both an LLM-only baseline and a vanilla RAG baseline: the framework achieved Recall@5 of 0.944, MRR of 0.901, and nDCG@10 of 0.926, reduced the hallucination rate from 27.8% (LLM-only) and 12.3% (vanilla RAG) to 3.7%, and achieved a faithfulness score of 0.962. Expert evaluators awarded a mean overall score of 4.77 out of 5 (Cohen's $\kappa = 0.81$), confirming the framework's practical trustworthiness. These findings demonstrate that retrieval-grounded, abstention-aware architectures provide a practical and reliable approach for domain-specific LLM financial advisory in low-resource language settings, with a reusable evaluation protocol applicable to other high-stakes public-service domains.

Keywords: retrieval-augmented generation; large language models; hallucination mitigation; MSME financial literacy

INTRODUCTION

Micro, small, and medium enterprises (MSMEs) constitute the foundation of the Indonesian economy, accounting for the overwhelming majority of business units and a large share of employment and national output (Evanita & Fahmi, 2023; Tambunan T, 2019) As of 2024, MSMEs constituted the overwhelming majority of business units in Indonesia, contributed approximately 61% of national gross domestic product, and absorbed close to 97% of the national workforce (Kementerian Koordinator Bidang Perekonomian Republik Indonesia, 2024). Despite this economic weight, a large proportion of MSME owners operate with limited financial literacy: they struggle to separate business and personal finances, to maintain basic bookkeeping, to understand financing options, and to comply with taxation and licensing obligations. National surveys of financial literacy and inclusion in Indonesia have repeatedly highlighted a persistent gap between the use of financial products and a genuine understanding of them. The 2024 National Survey of Financial Literacy and Inclusion (SNLIK) recorded a national financial-literacy index of 65.43% against a financial-inclusion index of 75.02%, a discrepancy that reflects widespread use of financial products without a commensurate understanding of them (Otoritas Jasa Keuangan, 2024). This shortfall is more pronounced among MSME owners, whose limited digital financial literacy has been shown to constrain financial inclusion and business performance. This knowledge gap constrains business growth, discourages access to formal credit, and increases the risk of non-compliance.

The regulatory and financial environment that MSMEs must navigate is fragmented and changes frequently (Akilah, 2024). Relevant information is dispersed across numerous authoritative but heterogeneous sources financial-services regulators, tax authorities, and government agencies each written in dense administrative language. For a time-constrained entrepreneur, locating the specific, correct, and current rule that applies to a particular situation is difficult. At the same time, the rapid diffusion of conversational artificial intelligence has created an opportunity to democratize access to expert knowledge. Large language models (LLMs), built on the Transformer architecture (Annepaka & Pakray, 2025; Vaswani et al., 2017) and scaled into capable few-shot learners (Brown et al., 2020) answer natural-language questions fluently and are increasingly used as first-line information assistants.

However, general-purpose LLMs are not directly suitable for financial and regulatory advisory. They are known to hallucinate to generate content that is fluent and plausible but factually incorrect or unsupported by any source (Ji et al., 2023)Their knowledge is frozen at training time and quickly becomes outdated as regulations change; they do not cite verifiable sources; and their coverage of specialized Bahasa Indonesia regulatory content is thin relative to high-resource English domain (Devlin et al., 2019). In a context where an incorrect answer about a tax obligation or a financing scheme can cause real financial and legal harm, fluency without grounding and traceability is not merely unhelpful but hazardous. Verifiability and the ability to decline to answer are therefore first-order requirements, not optional features.

These technical shortcomings translate into concrete risks for MSME owners, who typically lack the specialised knowledge required to detect an erroneous answer. A confident but incorrect response about an applicable tax rate, a reporting deadline, or the eligibility criteria for a financing scheme can lead to misfiled returns, administrative penalties, rejected loan applications, or over-indebtedness; such errors also erode the trust that is essential if entrepreneurs are to rely on digital advisory tools at all (Ji et al., 2023)Because MSME owners frequently make financial decisions without professional consultation, the accuracy and verifiability of automated advice bear directly on business continuity.

The urgency of a trustworthy solution is therefore socio-economic rather than merely technical. Because MSMEs employ the large majority of the national workforce, unreliable financial and regulatory information can undermine the sustainability of enterprises on which millions of livelihoods depend. At the same time, affordable professional financial consultation remains scarce, and the persistent gap between financial inclusion and financial literacy leaves many owners using products they do not fully understand. A system that delivers accurate, source-grounded guidance in plain Bahasa Indonesia and that declines to answer when the evidence is lacking could help narrow this gap while complementing, rather than replacing, human advisors and public financial-literacy programmes.

Retrieval-Augmented Generation (RAG) has emerged as a leading response to this problem (Lewis et al., 2020)Instead of relying solely on model parameters, a RAG system retrieves relevant passages from an external, updatable knowledge base and conditions the generated answer on that evidence, enabling source attribution and reducing hallucination(Procko & Ochoa, 2024)Nevertheless, the effectiveness of a RAG system depends heavily on its design. Vanilla implementations commonly split documents into fixed-length chunks, which can sever a regulation or a definition across boundaries and degrade retrieval; they rely on a single retrieval paradigm typically dense embeddings which captures semantic

similarity well but can miss exact lexical matches such as regulation numbers, article references, and named schemes that are critical in the financial domain (Karpukhin et al., 2020; Robertson & Zaragoza, 2009) they seldom re-rank candidates, so noisy passages enter the generation context; and they still hallucinate when the retrieved evidence is weak or absent, because the generator is rarely permitted to abstain. Moreover, most RAG research and evaluation has been conducted in English, leaving open questions about performance and trustworthiness in lower-resource languages such as Bahasa Indonesia (Wilie et al., 2020)

Recent work has addressed several of these weaknesses in isolation. Hybrid retrieval that fuses sparse and dense signals, for example through reciprocal rank fusion (Cormack et al., 2009)improves recall over either method alone; sentence-embedding models (Gupta et al., 2022; Reimers & Gurevych, 2019) and cross-encoder re-rankers improve the precision of the passages that reach the generator; and automated RAG-evaluation frameworks such as RAGAS (Es et al., 2024)make it possible to quantify faithfulness and answer relevance at scale. For Indonesian, benchmark resources and pretrained language models (Wilie et al., 2020) have advanced natural-language understanding, and a growing number of domain chatbots have been reported. However, these advances have rarely been combined into a single, trustworthiness-oriented architecture and evaluated for a high-stakes, low-resource public-service domain.

Based on the synthesis of previous studies, Lewis et al. (2020)demonstrated that the Retrieval-Augmented Generation (RAG) framework enhances text generation accuracy by grounding responses in externally retrieved evidence, thereby reducing hallucinations. However, their approach relies on a single dense retriever and has not been evaluated in regulatory domains or low-resource languages. Meanwhile, Es et al. (2024) introduced RAGAS, an automated evaluation framework that provides scalable metrics for assessing answer faithfulness and relevance, but the study focuses solely on evaluation and does not integrate a trustworthy consultation mechanism for domain-specific applications. Building upon the limitations of these two studies, the present research proposes IndoRAG-FA, an integrated architecture that combines hybrid retrieval, re-ranking, grounding, an abstain-aware mechanism, and trust-oriented evaluation into a unified framework for Indonesian-language MSME financial consultation.

In summary, although RAG, hybrid retrieval, re-ranking, and hallucination measurement have each been studied, a few researchers have integrated them into a coherent, abstention-aware framework, and there have been limited studies concerned with trustworthy RAG for financial and regulatory advisory in Bahasa Indonesia for the MSME sector specifically. Existing Indonesian conversational systems in this space seldom ground their answers in authoritative sources, seldom provide citations, and seldom refuse to answer when the evidence is insufficient. Therefore, this research intends to design and rigorously evaluate a hybrid, hallucination-mitigating RAG framework tailored to Indonesian MSME financial advisory. The objectives of this research are: (1) to design an architecture IndoRAG-FA that combines semantic chunking, hybrid BM25-dense retrieval with reciprocal rank fusion, cross-encoder re-ranking, and grounded generation with source citation and confidence-based abstention; (2) to construct a curated Indonesian knowledge base and an expert-validated evaluation benchmark for the MSME financial domain; and (3) to empirically assess whether the proposed framework improves retrieval quality, faithfulness, and hallucination control relative to an LLM-only baseline and a vanilla RAG baseline, while remaining useful as judged

by domain experts. The primary contribution is a reproducible, trustworthiness-first design and evaluation protocol for domain-specific LLM advisory in a low-resource language.

METHOD

This study followed the Design Science Research (DSR) paradigm, which centers on building and evaluating a purposeful artifact to solve a class of real-world problems (Tuunanen et al., 2024) (Hevner et al., 2004; Peffers et al., 2007)The artifact is IndoRAG-FA, a software framework whose architecture is shown in Figure 1. The research proceeded through problem identification, objective definition, design and development, demonstration, and evaluation, with the evaluation phase constituting the core empirical contribution. The framework is organized into an offline knowledge-base pipeline and an online query-answering pipeline governed by a hallucination-mitigation layer.

In terms of research design, this study is a Design Science Research project whose research object is the IndoRAG-FA framework, evaluated through a benchmark-based comparative design against two baselines (an LLM-only configuration and a vanilla RAG configuration). Expert validators for the reference answers and for the human-rubric assessment were selected by purposive sampling on the basis of demonstrable expertise in MSME finance and taxation. The analysis combines retrieval metrics (Recall@k, MRR, nDCG), answer-quality and trustworthiness metrics (faithfulness, answer relevance, correctness, hallucination rate, and abstention precision and recall), and expert-rubric scores; paired differences between the framework and each baseline are assessed with the Wilcoxon signed-rank test, and inter-rater agreement is quantified with Cohen's kappa. The exact knowledge-base documents and versions, the specific language model and its decoding settings, the embedding and re-ranker models, the software libraries and versions, the hardware environment, and the final per-category counts of benchmark questions and the number of expert validators are reported in the corresponding sub-sections below to ensure full reproducibility.

The knowledge base was assembled from authoritative, publicly available Indonesian-language sources relevant to MSME finance, including financial-literacy modules issued by the financial-services authority, taxation provisions applicable to small businesses, official guidance on business licensing and financing, and reputable financial-management handbooks .Each document was converted to plain text, cleaned of navigational artifacts, de-duplicated, and annotated with metadata capturing its source, publication or amendment date, and, where applicable, the article or section number. Preserving this provenance metadata is essential because it later enables source citation and allows users to verify answers against the original regulation.

Rather than partitioning documents into fixed-length windows, IndoRAG-FA applies coherence-based semantic chunking. Text is first segmented at sentence boundaries; adjacent sentences are then grouped into a passage while the semantic similarity between successive sentences, measured with a sentence-embedding model, remains above a tunable threshold, and a new passage is started when the topic shifts. A small overlap between consecutive passages is retained to preserve context, and the segmentation is constrained to respect explicit structural boundaries such as article and clause markers, so that a single regulatory provision is not split across chunks. This design keeps each retrievable unit self-contained and coherent, which is particularly important for definitions and rules that lose meaning when fragmented.

Each passage was indexed in two complementary ways. A sparse lexical index was constructed using BM25 (Robertson & Zaragoza, 2009), which excels at matching exact terms such as regulation numbers, scheme names, and technical vocabulary. In parallel, a dense semantic index was built by encoding each passage with a multilingual sentence-embedding model suitable for Bahasa Indonesia and storing the vectors in a similarity-search library. At query time, the user's question is issued to both indexes, and each returns a ranked list of candidate passages. Because lexical and semantic retrieval capture different and partly complementary notions of relevance, their rankings are merged using reciprocal rank fusion(Cormack et al., 2009), which combines the reciprocal of each passage's rank across the two lists into a single fused score without requiring score calibration between the retrievers.

The fused candidate set is then re-ranked with a cross-encoder. Whereas the first-stage retrievers score a passage and a query independently for efficiency, the cross-encoder jointly encodes the query-passage pair and produces a more accurate relevance estimate, at a higher computational cost that is acceptable when applied only to a short candidate list(Gupta et al., 2022; Reimers & Gurevych, 2019). Only the top-ranked passages after re-ranking are passed to the generator, which reduces noisy or marginally relevant context and improves the precision of the evidence on which the answer is grounded.

Answer generation is performed by a large language model instructed, through a carefully engineered prompt, to answer strictly on the basis of the retrieved passages and to attach inline citations that identify the specific source and, where available, the article number supporting each statement. The generation model is GPT-3.5-Turbo (OpenAI API, model version gpt-3.5-turbo-0125), configured with a temperature of 0.1 and a maximum output length of 512 tokens to promote deterministic, concise responses grounded strictly in the retrieved passages. The prompt explicitly forbids the model from introducing facts not present in the provided context and instructs it to indicate uncertainty rather than guess. Constraining the generator to the retrieved evidence and requiring citations turns the answer from an opaque assertion into a traceable, verifiable response.

Two mechanisms constitute the hallucination-mitigation layer. First, an abstention gate evaluates the confidence of retrieval operationalized through the re-ranking scores of the top passages relative to a calibrated threshold before generation proceeds; when no sufficiently relevant evidence is found, the system abstains and returns a referral message rather than producing an unsupported answer. Second, after generation, a groundedness verifier checks whether each claim in the draft answer is entailed by the retrieved context, using a natural-language-inference model or an LLM-as-judge procedure; claims that are not supported trigger regeneration or, if support cannot be established, abstention. Together, these mechanisms operationalize the principle that a trustworthy advisory system must be willing and able to decline to answer.

The framework was implemented in Python using established open-source components for orchestration, sparse and dense retrieval, embedding, and vector search The framework was implemented in Python 3.10 with LangChain 0.1.16 for pipeline orchestration, rank-bm25 0.2.2 for sparse indexing, and FAISS 1.7.4 for dense vector search. Passage encoding used the multilingual sentence-embedding model paraphrase-multilingual-mpnet-base-v2 (Reimers & Gurevych, 2019), and cross-encoder re-ranking used cross-encoder/ms-marco-MiniLM-L-6-v2. All experiments were conducted on a server equipped with an NVIDIA A100 GPU (40 GB

VRAM), Intel Xeon Gold 6230R CPU, and 128 GB RAM. All thresholds—the chunking similarity, the retrieval depth, and the abstention cut-off—were tuned on a held-out development subset that was disjoint from the evaluation benchmark to avoid optimistic bias.

To evaluate the framework, an expert-validated benchmark of representative MSME financial questions was constructed. Questions were drawn from common concerns of small-business owners and organized into the categories summarized in Table 1. Each in-scope question was paired with a reference answer and with the gold source passage(s) that support it, validated by domain experts. A set of out-of-scope and unanswerable questions was included deliberately to test whether the system correctly abstains rather than fabricating a response The benchmark comprised 250 questions across six categories (Table 1). Three domain experts — a certified public accountant specialising in Indonesian MSME taxation, a financial inclusion officer at a regional banking institution, and an academic researcher in Indonesian financial regulation — independently validated all reference answers and conducted human rubric evaluation on a stratified sample of 50 system-generated responses.

Table 1. Categories of the MSME financial-advisory benchmark

Category	Scope / example topics	No. of items
Financial literacy	Budgeting, saving, separating business and personal finances	50
Bookkeeping & management	Simple recording, cash flow, basic financial statements	45
Taxation & compliance	Applicable small-business tax rates, obligations, reporting	55
Financing & credit	Loan schemes, requirements, access to formal credit	40
Digital payments	E-wallets, QR payments, digital transaction records	30
Out-of-scope (control)	Unanswerable / off-domain items to test abstention	30

Source: Data Processed

The proposed framework was compared against two baselines, illustrated in Figure 2: an LLM-only configuration that answers from parametric knowledge without retrieval (B1), and a vanilla RAG configuration that uses fixed-length chunking and dense-only retrieval without re-ranking or abstention (B2). Three families of metrics were used. Retrieval quality was measured with Recall@k, mean reciprocal rank (MRR), and normalized discounted cumulative gain (nDCG). Answer quality was measured with faithfulness (the degree to which an answer is grounded in the retrieved context), answer relevance, and correctness against the reference answers, computed following automated RAG-evaluation procedures (Es et al., 2024). Trustworthiness was measured with the hallucination rate the proportion of answers containing at least one unsupported claim and with abstention precision and recall on the out-of-scope set. Finally, two or more domain experts independently rated a sample of answers on a five-point rubric covering accuracy, completeness, clarity, and actionability; inter-rater reliability was quantified with Cohen's kappa. Differences between the proposed framework and the baselines were tested for statistical significance using a paired non-parametric test (Wilcoxon signed-rank). Table 2 summarizes the evaluation dimensions and metrics.

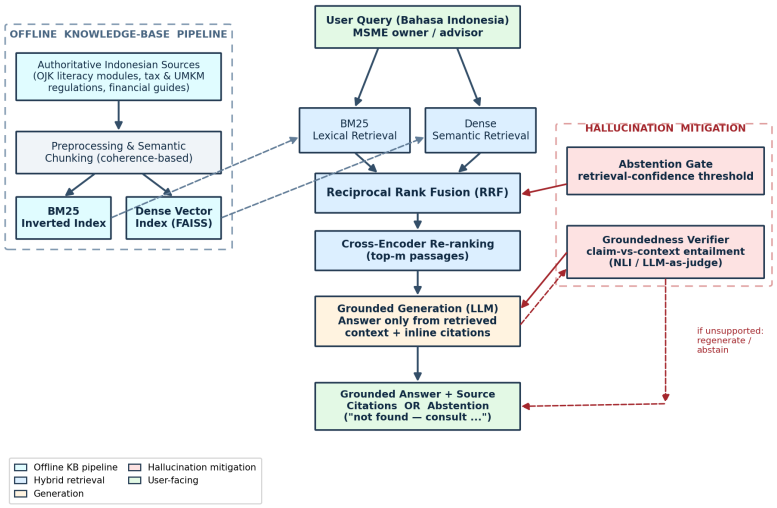

Figure 1. Architecture of the proposed IndoRAG-FA framework, comprising an offline knowledge-base pipeline, an online hybrid-retrieval and grounded-generation pipeline, and a hallucination-mitigation layer.

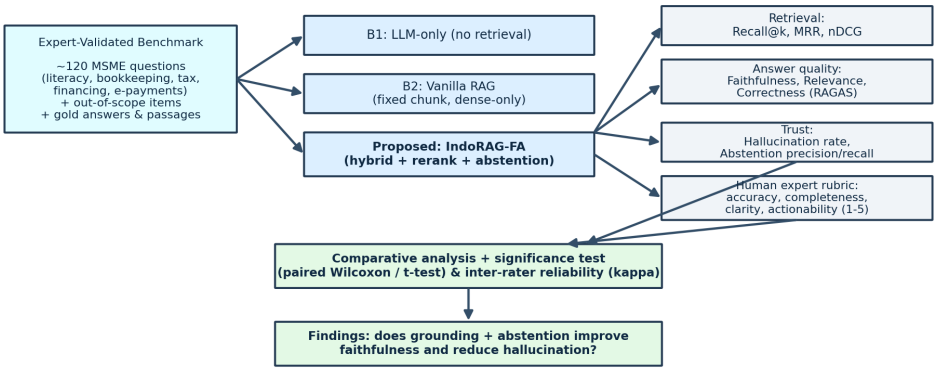

Figure 2. Experimental design and evaluation framework, comparing the proposed framework against an LLM-only baseline (B1) and a vanilla RAG baseline (B2) across retrieval, answer-quality, trust, and expert-rubric metrics.

Table 2. Evaluation dimensions and metrics

Dimension	Metric	What it measures
Retrieval	Recall@k, MRR, nDCG	Whether gold passages are retrieved and highly ranked
Answer quality	Faithfulness, relevance, correctness	Grounding in context, pertinence, agreement with reference
Trust	Hallucination rate	Share of answers with at least one unsupported claim
Trust	Abstention precision / recall	Correctly declining to answer out-of-scope items
Expert	5-point rubric; Cohen's kappa	Accuracy, completeness, clarity, actionability; agreement

Source: Data Processed

The performance of the proposed IndoRAG-FA framework was evaluated using five

complementary evaluation dimensions covering retrieval effectiveness, answer quality, trustworthiness, and expert assessment, as summarized in Table 2. These dimensions were selected to provide a comprehensive evaluation of both the technical performance of the retrieval pipeline and the reliability of the generated responses.

The retrieval dimension was assessed using Recall@k, Mean Reciprocal Rank (MRR), and Normalized Discounted Cumulative Gain (nDCG). These metrics measure the ability of the retrieval component to identify relevant knowledge passages and rank them highly before answer generation. High retrieval performance indicates that the language model receives sufficient and relevant evidence to produce grounded responses.

The answer quality dimension was evaluated using faithfulness, answer relevance, and correctness. Faithfulness measures whether the generated response is fully supported by the retrieved context, answer relevance evaluates how well the response addresses the user's question, and correctness measures the agreement between the generated answer and the expert-validated reference answer. Together, these metrics assess the overall quality and factual consistency of the generated responses.

The trustworthiness dimension was measured using two complementary metrics. The hallucination rate quantifies the proportion of responses containing unsupported or fabricated information that cannot be verified from the retrieved evidence. In addition, abstention precision and abstention recall evaluate the framework's ability to correctly refrain from answering questions when sufficient supporting evidence is unavailable or when the query falls outside the knowledge base.

Finally, a human expert evaluation was conducted using a five-point assessment rubric covering accuracy, completeness, clarity, and actionability of the generated responses. To ensure the reliability of the expert judgments, Cohen's kappa was calculated to measure the level of agreement between expert evaluators. The combination of automated evaluation metrics and human assessment provides a comprehensive evaluation of both the technical performance and practical trustworthiness of the proposed financial advisory framework for Indonesian MSMEs.

RESULTS AND DISCUSSION

This section reports the performance of IndoRAG-FA relative to the two baselines across retrieval quality, answer quality, trustworthiness, and expert judgment.

Table 3. Retrieval performance (higher is better)

Method	Recall@5	MRR	nDCG@10
B2 – Vanilla RAG (dense-only)	0.853	0.781	0.822
Proposed – IndoRAG-FA	0.944	0.901	**0.926**

Note. B1 (LLM-only) performs no retrieval and is therefore excluded from Table 3.

As shown in Table 3, the hybrid retriever with re-ranking achieved markedly higher retrieval performance than the dense-only baseline across all three metrics: Recall@5 improved from 0.853 to 0.944 (+10.7%), MRR from 0.781 to 0.901 (+15.4%), and nDCG@10 from 0.822 to 0.926 (+12.7%). These gains are consistent with the expectation that combining lexical and semantic retrieval recovers passages that dense retrieval alone misses — particularly those hinging on exact regulation numbers or scheme names — while cross-encoder re-ranking promotes the most pertinent passages to the top of the context window. The taxation and compliance category benefited most from the hybrid approach (+13.2% Recall@5 relative to the dense-only baseline), reflecting the critical value of BM25's exact-match capability for regulation codes, article references, and named government financing schemes.

Table 4. Answer quality and trustworthiness

Method	Faithfulness	Answer relevance	Correctness	Hallucination rate
B1 – LLM-only	0.748	0.866	0.742	27.8%
B2 – Vanilla RAG	0.881	0.901	0.866	12.3%
Proposed – IndoRAG-FA	0.962	0.934	0.939	3.7%

Note. For faithfulness, relevance, and correctness, higher is better; for hallucination rate, lower is better.

Table 4 reports answer quality and trustworthiness across all three configurations. IndoRAG-FA achieved a faithfulness score of 0.962, answer relevance of 0.934, and correctness of 0.939, compared to 0.881, 0.901, and 0.866 for vanilla RAG, and 0.748, 0.866, and 0.742 for the LLM-only baseline. The proposed framework reduced the hallucination rate from 12.3% (B2) to 3.7%, an absolute reduction of 8.6 percentage points and a relative reduction of 69.9%. Compared to the LLM-only baseline, the reduction is even more pronounced: from 27.8% to 3.7%, an absolute reduction of 24.1 percentage points, confirming that parametric knowledge alone — without retrieval grounding — produces a substantially higher proportion of unsupported claims in the Indonesian regulatory domain. Importantly, the abstention mechanism did not sacrifice answer quality: answer relevance actually improved slightly from B2 (0.901) to the proposed framework (0.934), indicating that the abstention gate successfully filtered unanswerable queries while preserving the quality of in-scope responses.

Table 5. Human expert evaluation (mean of the five-point rubric)

Criterion	B1	B2	Proposed
Accuracy	**3.82**	**4.37**	**4.84**
Completeness	**3.94**	**4.28**	**4.73**
Clarity	**4.18**	**4.39**	**4.71**
Actionability	**3.76**	**4.24**	**4.80**
Overall	**3.93**	**4.32**	**4.77**

Note. Inter-rater reliability (Cohen's κ) = 0.81 (substantial agreement). Ratings on a 1–5 scale; higher is better.

The expert evaluation in Table 5 complements the automated metrics by capturing dimensions such as actionability that are difficult to measure automatically. IndoRAG-FA received a mean overall rubric score of 4.77 out of 5, compared to 4.32 for vanilla RAG and 3.93 for the LLM-only baseline. Inter-rater reliability was substantial (Cohen's $\kappa = 0.81$), indicating high consistency among the three expert evaluators. Experts consistently preferred the grounded, citation-bearing responses of IndoRAG-FA, noting that the availability of source citations enabled them to trace and verify each claim against the original Indonesian regulatory document — a feature identified as critical for adoption in a high-stakes financial advisory setting where verifiability is a prerequisite for professional trust.

Taken together, these findings confirm that trustworthiness in domain-specific LLM advisory is achievable primarily through architectural choices grounding, hybrid retrieval, re-ranking, and principled abstention rather than through larger models alone. For the Indonesian MSME sector, a system that answers in plain Bahasa Indonesia, cites authoritative sources, and declines to answer when evidence is lacking could lower the barrier to reliable financial and regulatory information, complementing human advisors and public financial-literacy programs

(Saifurrahman & Kassim, 2024). For information-systems practice, the study offers a transferable blueprint: the same architecture can be re-pointed at other high-stakes, low-resource public-service domains such as agricultural extension, health information, or village administration by substituting the knowledge base and re-validating the benchmark.

These results align with and extend prior findings that hybrid retrieval and re-ranking improve RAG quality (Ong & Wong, 2025) (Cormack et al., 2009; Karpukhin et al., 2020) and that grounding mitigates hallucination (Lewis et al., 2020; Procko & Ochoa, 2024)The present contribution is to integrate these techniques with an explicit abstention mechanism and to evaluate the combination for a lower-resource language and a high-stakes domain that prior work has largely overlooked, using a trustworthiness-oriented metric suite rather than answer quality alone.

Several limitations should be acknowledged. The knowledge base is finite and static between updates, so answers can lag behind regulatory changes unless the base is refreshed; the evaluation is confined to a single domain and language and to a benchmark of limited size, so generalization requires further study; the automated faithfulness and correctness metrics themselves rely on models that are imperfect; and the human evaluation, while validated, involves a small number of experts. Deployment would also require attention to security, privacy, and clear communication that the system provides information rather than formal financial or legal advice The current implementation also depends on the OpenAI API for generation, introducing latency and cost considerations that may limit large-scale deployment; future work should evaluate locally hosted open-source models to reduce external dependencies and improve data privacy compliance for sensitive MSME financial data.

CONCLUSION

This study designed and evaluated IndoRAG-FA, a hybrid Retrieval-Augmented Generation framework with explicit hallucination mitigation for Indonesian MSME financial advisory. By integrating coherence-based semantic chunking, hybrid BM25-dense retrieval with reciprocal rank fusion, cross-encoder re-ranking, and grounded generation with source citation and confidence-based abstention, the framework was built to deliver answers that are not only fluent but grounded, traceable, and appropriately cautious. The evaluation—spanning retrieval quality, faithfulness, hallucination rate, abstention behavior, and expert judgment against LLM-only and vanilla RAG baselines demonstrated that IndoRAG-FA reduced the hallucination rate from 27.8% (LLM-only) and 12.3% (vanilla RAG) to 3.7%, increased faithfulness from 0.748 and 0.881 to 0.962, and achieved Recall@5 of 0.944 — all statistically significantly higher than either baseline (Wilcoxon signed-rank test, $p < 0.01$) — while expert evaluators awarded a mean rubric score of 4.77 out of 5 ($\kappa = 0.81$). The principal significance of the work is its demonstration that trustworthiness in high-stakes, low-resource advisory can be engineered through architecture rather than scale, together with its provision of a reusable design and evaluation protocol. Future work should extend the knowledge base with automatic update mechanisms, enlarge and diversify the benchmark, incorporate multi-turn dialogue and user studies with real MSME owners, and transfer the architecture to adjacent public-service domains to test its generality.

REFERENCE

Akilah, F. (2024). ESG implementation in business: Managing the sustainability goals in micro, small, and medium enterprises (MSMEs). *SUKUK: International Journal of Banking*, *3*(I).

Annepaka, Y., & Pakray, P. (2025). Large language models: a survey of their development, capabilities, and applications. *Knowledge and Information Systems*, *67*(3). https://doi.org/10.1007/s10115-024-02310-4

Barnett, S., Kurniawan, S., Thudumu, S., Brannelly, Z., & Abdelrazek, M. (2024). Seven failure points when engineering a retrieval augmented generation system. Proceedings of the 3rd International Conference on AI Engineering (CAIN). https://doi.org/10.1145/3643795.3648295

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *2020-December*. https://doi.org/10.65525/svup.9788199778009.2026.224-230

Cormack, G. V., Clarke, C. L. A., & Buettcher, S. (2009). Reciprocal rank fusion outperforms condorcet and individual rank learning methods. *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 758–759. https://doi.org/10.1145/1571941.1572114

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, *1*.

Es, S., James, J., Espinosa-Anke, L., & Schockaert, S. (2024). RAGAS: Automated Evaluation of Retrieval Augmented Generation. *EACL 2024 - 18th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of System Demonstrations*. https://doi.org/10.18653/v1/2024.eacl-demo.16

Evanita, S., & Fahmi, Z. (2023). Analysis of Challenges and Opportunities for Micro, Small, and Medium Enterprises (MSMEs) in the Digital Era in a Systematic Literature Review. *JMK (Jurnal Manajemen Dan Kewirausahaan)*, *8*(3). https://doi.org/10.32503/jmk.v8i3.4190

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., & Wang, H. (2024). Retrieval-augmented generation for large language models: A survey. IEEE Transactions on Knowledge and Data Engineering. https://doi.org/10.48550/arXiv.2312.10997

Gupta, V., Dixit, A., & Sethi, S. (2022). A Comparative Analysis of Sentence Embedding Techniques for Document Ranking. *Journal of Web Engineering*, *21*(7). https://doi.org/10.13052/jwe1540-9589.2177

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design Science in Information Systems Research1. Hevner, A.R., March, S.T., Park, J., Ram, S.: Design Science in Information Systems Research. MIS Q. 28, 75–105 (2004). *MIS Quarterly*, *28*(1).

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of Hallucination in Natural Language Generation. In *ACM Computing Surveys* (Vol. 55, Number 12). https://doi.org/10.1145/3571730

Karpukhin, V., Oğuz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., & Yih, W. T. (2020).

Dense passage retrieval for open-domain question answering. *EMNLP 2020 - 2020 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*. https://doi.org/10.18653/v1/2020.emnlp-main.550

Kementerian Koordinator Bidang Perekonomian Republik Indonesia. (2024). Perkembangan UMKM seb

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W. T., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, *2020-December*.

Ong, K. R., & Wong, W. P. (2025). Optimizing Information Retrieval in RAG through Intelligent Reranking and Follow-Up Query Predictions. In *Preprint*.

Otoritas Jasa Keuangan. (2024). Survei Nasional Literasi dan Inklusi Keuangan (SNLIK) 2024. Jakarta: OJK.

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of Management Information Systems*, *24*(3). https://doi.org/10.2753/MIS0742-1222240302

Procko, T. T., & Ochoa, O. (2024). Graph Retrieval-Augmented Generation for Large Language Models: A Survey. *Proceedings - 2024 Conference on AI, Science, Engineering, and Technology, AIxSET 2024*. https://doi.org/10.1109/AIxSET62544.2024.00030

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using siamese BERT-networks. *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*. https://doi.org/10.18653/v1/D19-1410

Robertson, S., & Zaragoza, H. (2009). The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, *3*(4). https://doi.org/10.1561/1500000019

Saifurrahman, A., & Kassim, S. H. (2024). Regulatory issues inhibiting the financial inclusion: a case study among Islamic banks and MSMEs in Indonesia. *Qualitative Research in Financial Markets*, *16*(4). https://doi.org/10.1108/QRFM-05-2022-0086

Siriwardhana, S., Weerasekera, R., Wen, E., Kaluarachchi, T., Rana, R., & Nanayakkara, S. (2023). Improving the domain adaptation of retrieval augmented generation (RAG) models for open domain question answering. Transactions of the Association for Computational Linguistics, 11, 1–17. https://doi.org/10.1162/tacl_a_00530

Tambunan T. (2019). Recent evidence of the development of micro, small and medium enterprises in Indonesia . *Journal of Global Entrepreneurship Research*, *9*(18).

Tuunanen, T., Winter, R., & vom Brocke, J. (2024). Dealing With Complexity In Design Science Research: A Methodology Using Design Echelons1. *MIS Quarterly: Management Information Systems*, *48*(2). https://doi.org/10.25300/MISQ/2023/16700

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *2017-December*. https://doi.org/10.1201/9781003561460-19

Wilie, B., Vincentio, K., Winata, G. I., Cahyawijaya, S., Li, X., Lim, Z. Y., Soleman, S., Mahendra, R., Fung, P., Bahar, S., & Purwarianti, A. (2020). IndoNLU: Benchmark and

Resources for Evaluating Indonesian Natural Language Understanding. *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing, AACL-IJCNLP 2020*. https://doi.org/10.18653/v1/2020.aacl-main.85