Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Corrective Retrieval-Augmented Generation Adaptation in Closed Domains for Development of Intelligent Human Resource Management Systems

Devi Willieam Anggara
Institut Teknologi Bandung
Marcellino Joestiawan
18222115@std.stei.itb.ac.id

Institut Teknologi Bandung
Mohd Shafry Mohd Rahim
UTM-IRDA Digital Media Centre, Media and Game Innovation of Excellence (MaGICX), Universiti Teknologi Malaysia
Ibrahim Salim Sulaiman Al Farsi
Sohar University
Fetty Fitriyanti Lubis
Institut Teknologi Bandung
Dion Tanjung
Institut Teknologi Bandung

Research Article

Keywords: Retrieval-Augmented Generation, Corrective RAG, Structured RAG, Human Resource Management System, Small and Medium Enterprises, Ablation Study

Posted Date: September 2nd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10574280/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

https://doi.org/10.1007/xxx

Corrective Retrieval-Augmented Generation Adaptation in Closed Domains for Development of Intelligent Human Resource Management Systems

Devi Willieam Anggara[1] · Marcellino Joestiawan[1*] · Mohd Shafry Mohd Rahim[2] · Ibrahim Salim Sulaiman Al Farsi[3] ·Fetty Fitriyanti Lubis[1] ·Dion Tanjung[1]

Received:
Published online :
© Springer Science+Business Media, LLC, part of Springer Nature 2026

Abstract
Small and Medium Enterprises (SMEs) face persistent difficulties in managing human resources effectively, in part because they lack HR decision-support systems that are grounded in a trusted knowledge source. Standard Retrieval-Augmented Generation (Naive RAG) does not verify whether the documents it retrieves are genuinely relevant to the user's query, and its shallow vector representations handle comparative and predictive queries poorly. Using the Design Science Research Methodology (DSRM), this paper builds a multi-agent Smart HR system for three HR domains (recruitment, performance management, and career development) that answers only from Armstrong's Handbook of Human Resource Management Practice (16th edition). It makes two technical contributions: (1) a closed-domain CRAG adaptation (C1) that, instead of falling back to web search, re-retrieves from the same Qdrant knowledge base whenever retrieval is judged INCORRECT; and (2) a flan-t5-large retrieval evaluator trained with a same-chapter hard-negative strategy (C2) so that it can determine whether a chunk truly answers the query rather than merely sharing its topic. A three-variant ablation study (Naive RAG, SRAG-only, CRAG-only) over 270 synthetic queries (90 per domain; 810 inferences in total), evaluated with the RAGAS and LLM-as-Judge frameworks, confirms H1: CRAG-only is significantly more faithful than Naive RAG (Wilcoxon signed-rank, $p < 0.001$, Cohen's $d = 0.457$, medium effect; Corrective Action Rate 0.278). H2 is rejected, because SRAG-only does not improve Faithfulness on its own ($p = 0.388$, $d = 0.071$). This gain comes with a trade-off: every variant that adds a mechanism scores lower LLM-as-Judge Accuracy than Naive RAG. The paper contributes an auditable closed-domain CRAG artifact, empirical evidence that C1 and SRAG contribute unequally to answer quality, and guidance for choosing an architecture based on domain characteristics.

Keywords Retrieval-Augmented Generation · Corrective RAG · Structured RAG · Human Resource Management System · Small and Medium Enterprises · Ablation Study

- Marcellino Joestiawan1
  18222115@std.stei.itb.ac.id

[1] School of Electrical Engineering and Informatics, Institut Teknologi Bandung, Bandung, Indonesia
[2] UTM-IRDA Digital Media Centre, Media and Game Innovation of Excellence (MaGICX), Institute of Human Centered Engineering (iHumEn), Universiti Teknologi Malaysia, Johor, Malaysia
[3] Faculty of Computing and Information Technology, Sohar University, Sohar, Oman

1 Introduction

Micro, Small, and Medium Enterprises (MSMEs) are a pillar of the Indonesian economy, contributing 61.07% of national Gross Domestic Product and absorbing up to 97% of the national workforce [1]. Despite this contribution, MSMEs face persistent structural challenges in human resource (HR) management, including limited resources, low employee motivation, and high turnover rooted in weak managerial capability [2]. Informal work patterns and the absence of formal systems have been repeatedly identified as a key barrier to innovation and competitiveness [3, 4], and HR practices in MSMEs are generally reported as non-standardized and unprofessional [5, 6].

Three HR functions are simultaneously the most critical and the most constrained in MSMEs: recruitment, performance management, and career development. Recruitment frequently fails to attract and retain suitable candidates because selection is conducted informally without standardized criteria [2, 5]. Performance management tends to be subjective and reactive, lacking measurable indicators and consistent feedback [6, 7]. Career development is the weakest of the three: MSMEs rarely define career paths or maintain a continuous competency monitoring system, which sustains high turnover [2, 7].

Conventional Human Resource Information Systems (HRIS) improve HR efficiency in large organizations [8, 9], but adoption by Indonesian MSMEs is hindered by low digital literacy, limited technical human resources, and a mismatch between system complexity and organizational capacity [10, 11]. A Smart HR approach that combines artificial intelligence with automation is a better fit for this context, offering adaptive decision support with comparatively simple infrastructure [12, 13].

Retrieval-Augmented Generation (RAG) is a foundational technology for knowledge-grounded Smart HR systems, combining the generative capability of a Large Language Model (LLM) with retrieval from an external document base [14], thereby reducing the hallucination risk inherent to LLM-only generation. In this work, the external knowledge base is Armstrong's Handbook of Human Resource Management Practice (16th edition) [15], a comprehensive and widely recognized HRM reference.

However, standard RAG (Naive RAG) has two critical structural limitations in the HR domain [16]. First, it has no mechanism to evaluate the quality of retrieved documents: low relevance or ambiguous context is fed directly into the LLM, which is especially risky for HR decisions such as performance appraisal or promotion recommendations that directly affect employees [17]. Second, its vector representation captures only surface-level text semantics without deeper knowledge structure, so comparative queries (e.g., comparing competency-based selection methods) or

predictive queries (e.g., promotion readiness criteria) are often retrieved inaccurately [16].

Two advanced RAG architectures address these limitations. Corrective RAG (CRAG) introduces a retrieval-evaluator layer that scores document relevance before generation: high-quality documents are used directly, low-quality documents trigger re-search, and ambiguous results trigger query refinement [18]. Structured RAG (SRAG) enriches vector representations with structured metadata injected at indexing time, improving retrieval precision for analytical and comparative queries [19]. The two approaches are complementary: SRAG improves retrieval quality from the indexing stage onward, while CRAG provides a correction layer for cases in which retrieval still fails at inference time.

This paper makes two technical contributions realized as artifacts of a Smart HR system. First, a closed-domain CRAG adaptation (C1) that replaces the web-search fallback on the INCORRECT path with re-retrieval against the same curated Qdrant knowledge base, so that every answer remains auditable and confined to standardized knowledge. Second, a flan-t5-large retrieval evaluator trained with a same-chapter hard-negative strategy (C2), which teaches the model to discriminate relevance at the query level rather than at the chapter-topic level, a prerequisite for a reliable CRAG verdict mechanism on a curated knowledge base such as Armstrong's Handbook. Both contributions are implemented across three HR domains, and a three-variant ablation study (Naive RAG, SRAG-only, and the closed-domain CRAG adaptation) is used to empirically quantify their effect using the RAGAS and LLM-as-Judge frameworks.

The system is developed following the Design Science Research Methodology (DSRM) (Peffers et al., 2007); Fig. 1 summarizes how its six activities are instantiated in this work.

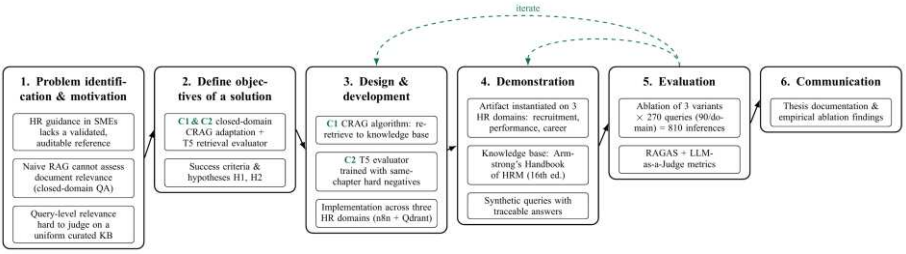


Fig. 1 Instantiation of the six Design Science Research Methodology (DSRM) activities (Peffers et al., 2007) in this work. Teal tags mark the two technical contributions (C1: closed-domain CRAG adaptation; C2: T5 hard-negative evaluator); dashed arrows denote DSRM iteration from evaluation back to design and demonstration.

The remainder of this paper is organized as follows. Section 2 reviews related work on RAG, CRAG, SRAG, and agentic RAG architectures. Section 3 presents the problem analysis and the design of the closed-domain CRAG adaptation and the SRAG comparison layer, including the ablation design and hypotheses. Section 4 describes the implementation. Section 5 reports the experimental results. Section 6 discusses the findings and limitations, and Section 7 concludes with contributions and future work.

2 Related work

2.1 HRM Challenges in SMEs and the Case for Smart HR

Human Resource Management (HRM) is a strategic, integrated approach to managing, developing, and supporting the well-being of an organization's workforce, spanning resourcing, learning and development, performance management, reward management, and employee relations [15]. This paper evaluates three of these functions (recruitment, performance management, and career development), which prior studies consistently identify as the most informally managed functions in MSMEs [2, 5, 6]. Conventional HRIS automate transactional record-keeping and payroll-adjacent workflows [8, 9, 20], including Electronic HRM self-service portals [21], but remain rule-based and non-conversational: they store transactional data rather than conceptual HRM knowledge and cannot answer open-ended questions such as "how should a competency framework be structured for a technology SME?" This gap between available HRIS data and the knowledge managers need to make quality HR decisions motivates an AI/RAG-based approach. Smart HRM, the application of AI to increase the effectiveness and accuracy of HR processes [12, 13], offers a path to democratize access to HR knowledge without requiring a large in-house HR department.

2.2 Retrieval-Augmented Generation

Large Language Models (LLMs) are transformer-based models trained on internet-scale corpora, capable of natural-language understanding and generation, multi-step reasoning, and zero- or few-shot task completion via in-context learning [16]. Their central weakness is hallucination: linguistically confident but factually ungrounded output [18], a serious risk in knowledge-intensive domains such as HRM, where a hallucinated KPI procedure or promotion recommendation carries real consequences. Metropolitansky and Larson [17] further distinguish closed-domain hallucination, where a model still produces unsupported claims despite being grounded in a supplied source document, and show that pipelines with multiple generative steps are more susceptible to this failure mode.

Retrieval-Augmented Generation (RAG) mitigates hallucination by grounding generation in an external knowledge base [14], combining parametric memory (the LLM's weights) with non-parametric memory (a retrievable corpus). Formally, Lewis et al. express the probability of generating response y given query x under the RAG-Sequence model as

$$p_{\text{RAG-Sequence}}(y|x) \approx \sum_{z \in \text{top-}k(p(\cdot|x))} p_\eta(z|x)\, p_\theta(y|x,z)$$
$$= \sum_{z \in \text{top-}k(p(\cdot|x))} p_\eta(z|x) \prod_{i}^{N} p_\theta(y_i|x,z,y_{1:i-1}) \quad (1)$$

where $p\eta(z|x)$ is the retriever (parameters η) that returns a distribution over the k documents z most relevant to query x, and $p\theta(y|x,z)$ is the generator (parameters θ) that

produces the response from the query and the retrieved context. The right-hand side factorizes the response probability into a product of per-token probabilities conditioned on x, the document z, and the preceding tokens, with N the response length; the RAG-Sequence model uses the same latent document for the entire response, unlike RAG-Token, which may use a different document per token. Gao et al. [16] trace RAG's evolution from Naive RAG (index, retrieve, and generate with no quality control), through Advanced RAG (pre-/post-retrieval techniques such as query rewriting and re-ranking), to Modular RAG, which allows retrieval, augmentation, and generation components to be independently configured. CRAG and SRAG, discussed next, are Modular RAG variants that specialize in retrieval quality.

2.3 Corrective RAG (CRAG)

CRAG [18] adds a retrieval evaluator that scores each retrieved document and triggers one of three actions: **CORRECT** (at least one document scores above the upper threshold; the system applies a decompose-filter-recompose knowledge-refinement procedure), **INCORRECT** (all documents score below the lower threshold; the original design discards retrieval and falls back to open web search), and **AMBIGUOUS** (score falls between the thresholds; refinement and web search are combined). Formally,

$$k = \begin{cases} \text{Refine}(x, \mathcal{D}) & \text{if CORRECT} \\ \text{WebSearch}(x) & \text{if INCORRECT} \\ \text{Refine}(x, \mathcal{D}) \oplus \text{WebSearch}(x) & \text{if AMBIGUOUS} \end{cases} \tag{2}$$

where x is the query, D the retrieved documents, and k the final context passed to the generator. CRAG improves consistently over standard RAG on open-domain QA benchmarks, particularly when retrieval quality cannot be guaranteed [18], a condition common to knowledge-intensive domains such as HRM.

2.4 Structured RAG (SRAG)

SRAG [19] improves chunk vector representations by injecting LLM-generated structured metadata (topics, chunk type, sentiment, knowledge-graph triples, and semantic tags) prior to embedding, so that both chunks and queries are represented in an enriched space rather than relying on surface-level similarity alone. The enriched representation is

$$e_i = t_i \oplus \text{Topics}\ (t_i) \oplus \text{Type}\ (t_i) \oplus \text{KGTriples}\ (t_i) \oplus \text{Tags}\ (t_i), \tag{3}$$

where t_i is the original chunk text. On a financial question-answering benchmark, SRAG improved LLM-as-Judge scores by 30% overall ($p = 2 \times 10{-}13$), with the largest gains on comparative (+38.2 points), analytical (+28.7 points), and predictive (+31.15 points) queries, the same query types that dominate the recruitment and career-development domains evaluated in this paper. The original ablation also reports an

asymmetric risk profile (mean gain +51.76 points vs. mean loss −1.95 points when SRAG underperforms), motivating its use as a low-risk retrieval-quality lever.

2.5 Closed-Domain Question Answering and Agentic RAG

Question-answering systems are commonly divided into open-domain QA, which answers factual questions from a large, unrestricted corpus such as Wikipedia [22], and closed-domain QA, which operates over a defined, curated knowledge base [23]. The Smart HR system in this paper is a closed-domain QA system: it answers exclusively from Armstrong's Handbook (16th edition), which makes the original CRAG's open-web fallback architecturally unsuitable, motivating the closed-domain adaptation presented in Section 3.

Agentic RAG architectures integrate autonomous agent decision-making into the RAG pipeline, allowing a system to evaluate retrieval results and adapt its strategy rather than following a fixed retrieve-read-generate path [24]. This paper adopts a two-tier multi-agent pattern: an orchestrator agent classifies and routes a query to a domain-specialist agent (recruitment, performance, or career development), and each domain agent invokes the closed-domain CRAG adaptation mechanism as a retrieval tool. Within the CRAG sub-workflow, the evaluate-correct-observe cycle follows the ReAct pattern of interleaving reasoning traces with tool actions [25].

3 Problem Analysis and Architecture Design

Business-process mapping of MSME recruitment, performance management, and career-development workflows shows a common pattern: linear, manual hand-offs between the business-unit head, HR staff, employees, and (for recruitment) candidates, with no automation and no shared knowledge reference. Critical decisions such as competency criteria, selection methods, interview design, KPI definitions, appraisal criteria, and career-path recommendations depend entirely on individual, undocumented HR knowledge rather than a validated HRM reference, consistent with prior findings that MSME HR practice is informal and ad hoc [2, 5–7, 26]. A gap analysis using Critical-to-Quality (CTQ) criteria confirms one gap per domain, translated into functional and non-functional requirements emphasizing (i) answers grounded in a validated HR knowledge source, (ii) contextual multi-type query handling (comparative, procedural, predictive), and (iii) retrieval precision across query types.

Solution selection proceeded through three successive Weighted Scoring Model (WSM) decisions, each narrowing the design space. First, at the technical paradigm level, fine-tuning an LLM on HR data, a retrieval-free ("Naive") LLM, and Retrieval-Augmented Generation were compared: RAG was selected because it requires no retraining when HR knowledge is updated (unlike fine-tuning [14]) and grounds every answer in a verifiable document (unlike a retrieval free LLM, which is prone to hallucination on a specialized domain such as HRM [18]), directly satisfying the non-functional requirements for knowledge-grounded accuracy and source traceability derived from the CTQ gap analysis [16].

Second, within RAG, three retrieval architectures were compared using a WSM with weights closed-domain feasibility 40%, source auditability 30%, retrieval quality 20%, and implementation complexity 10% (closed-domain feasibility received the highest weight because it is a hard constraint: a system that depends on web search cannot

satisfy source auditability at all). Naive RAG scored 2.60 (no retrieval-quality control), off-the-shelf CRAG (with its open-web fallback) scored 1.90, disqualified primarily on closed-domain feasibility and auditability; and a closed-domain CRAG adaptation scored 4.50, the highest, by retaining CRAG's corrective mechanism while confining every fallback source to the curated Qdrant knowledge base.

Third, within indexing, plain indexing (text and minimal structural metadata only) was compared against SRAG-enhanced indexing (LLM-generated chunk-type and semantic-tag metadata) using a WSM weighted toward semantic representation richness (40%) and contextual completeness (30%): plain indexing scored 2.50 and SRAG-enhanced indexing scored 4.50, since chunk-type labels let the system associate a query's information need (e.g., a procedural "how do I..." query) with the matching chunk type, a signal plain indexing lacks entirely. SRAG-enhanced indexing was therefore adopted, operating complementarily to the CRAG adaptation at a different pipeline layer (indexing versus inference); its actual marginal contribution to answer quality is tested empirically via the ablation study detailed below.

3.1 System Overview

The proposed system operates as a closed-domain QA system grounded exclusively in Armstrong's Handbook (16th edition). It has two layers operating at different stages of the query lifecycle: an SRAG layer active at indexing time, and a CRAG layer active at inference time (Fig. 2). The system follows a two-tier multi-agent pattern: a domain orchestrator receives the user query and manages per-domain conversational context and invokes a CRAG sub-workflow as a tool to perform retrieval, quality evaluation, and correction. Separating the orchestrator from the CRAG sub-workflow allows either to be updated independently.

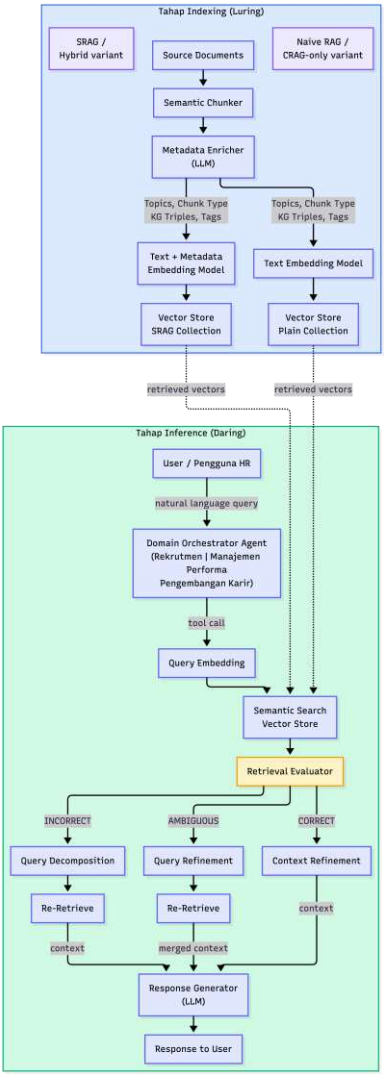


Fig.2 System overview of the Smart HR system: SRAG operates at the indexing layer and CRAG at the inference layer.

3.1.1 SRAG Layer Design

The SRAG layer enriches each chunk with LLM-generated metadata before embedding (Fig. 3), so that semantic search at inference time is more precise. Two adaptations were made relative to the original design [19], which targeted financial question-answering over corporate earnings reports. First, the sentiments metadata field was dropped: HR queries in this system are overwhelmingly factual or procedural, and the original paper explicitly notes that "sentiment is intrinsically less relevant for factual queries," a finding corroborated by its own ablation (removing sentiment slightly

improved scores by +0.63, p = 0.37, not significant). Second, the chunk type taxonomy was re-mapped from finance-oriented categories (quantitative, comparative, analytical, information lookup, informational, predictive) to a content-type taxonomy suited to an academic HRM handbook: conceptual (replacing factual, since HRM knowledge is largely conceptual rather than numeric), procedural (retained), comparative (retained), and two additions, evaluative (for performance-appraisal criteria and best practices) and example (for case studies).

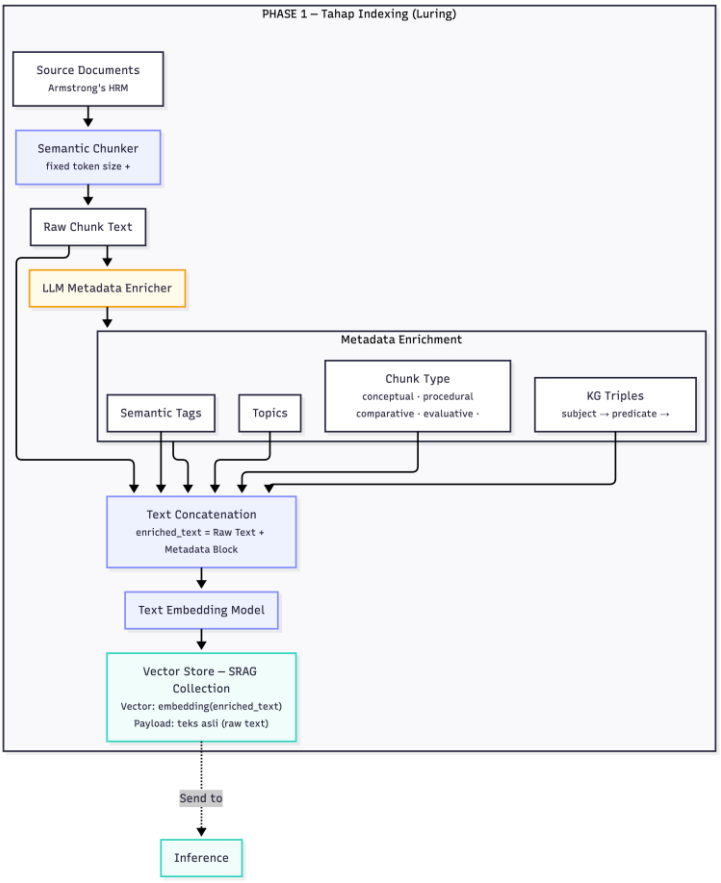


Fig.3 SRAG metadata-enrichment mechanism used at indexing time.

The system stores the enriched text's embedding vector but returns the clean chunk text as generator context ("embed enriched, retrieve clean"), so the generator is never distracted by metadata markup:

Topics: *topic1*; *topic2*; *topic3*
Type: *chunk_type*
Concepts: *subj1-pred1-obj1* | *subj2-pred2-obj2*

Tags: *tag1, tag2, tag3, tag4, tag5*

[original chunk text]

3.1.2 Closed-Domain CRAG Adaptation

The CRAG layer verifies retrieval quality at inference time (Fig. 4). The original CRAG design assumes an open-domain setting with access to web search as a fallback source [18, 27]; on this system's closed-domain knowledge base [23], an open-web fallback is unacceptable for two reasons: web content is off-domain and not standardized to Armstrong's HRM terminology, and it cannot be guaranteed to originate from the validated reference, violating the source-auditability requirement. The corrective action is therefore adapted: on INCORRECT, the query is reformatted with a fixed HRM-context prefix (e.g., "human resource management framework principles: {query}") and re-retrieved against the same Qdrant collection (rather than searched on the web); on AMBIGUOUS, the query is reformatted with a different fixed prefix ("HRM best practices: {query}"), re-retrieved, and the results are merged with the original retrieval and de-duplicated. The CORRECT path is unchanged from the original design. This adaptation (contribution C1) is the paper's primary technical contribution: it preserves CRAG's corrective mechanism while confining every answer's provenance to the curated knowledge base.

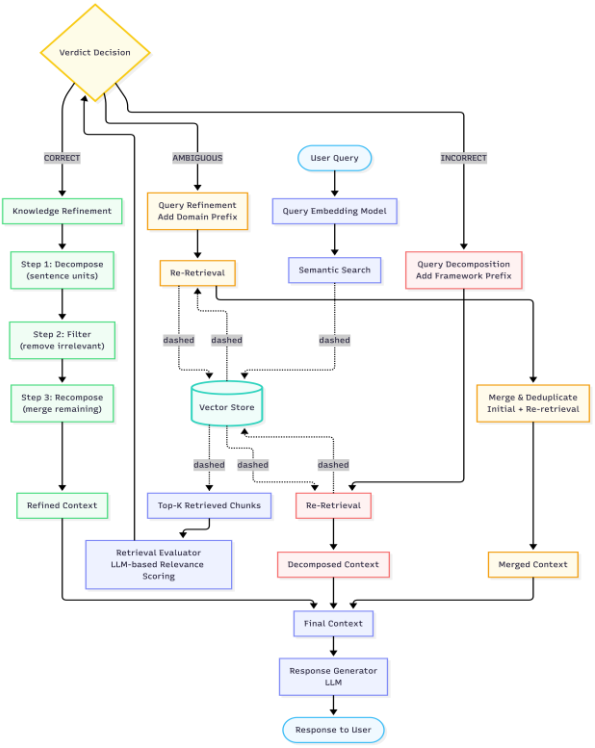


Fig.4 Closed-domain CRAG architecture: the INCORRECT and AMBIGUOUS paths re-retrieve against the same Qdrant collection instead of querying the open web..

3.1.3 T5 Retrieval Evaluator with Same-Chapter Hard Negatives

The closed-domain CRAG adaptation requires a retrieval evaluator that distinguishes chunks that actually answer a query from chunks that merely share its topic. On a curated knowledge base such as Armstrong's Handbook, nearly every retrieved chunk is topically relevant because the source corpus is already domain-filtered; an evaluator trained only on topic similarity would classify almost everything as CORRECT, and the corrective mechanism would never trigger. This is addressed with model choice and training-data design (contribution C2).

The evaluator is a fine-tuned google/flan-t5-large model. It reads a "query [SEP] chunk" pair and produces a single relevance score (a regression head trained with MSE loss): +1.0 when the chunk is relevant to the query and −1.0 when it is not, following the original CRAG evaluator design [18]. A local T5 model was chosen over LLM-API-based judging for latency (milliseconds vs. hundreds of milliseconds per call), determinism (LLM judgments are probabilistic and can vary between calls on identical inputs), and cost (810 ablation inferences × per-pair evaluation would be API-cost-prohibitive). The training set (2,271 labelled examples; 33.1% positive, 66.9% negative) combines six sources (Table 1), the key one being 600 same-chapter hard negatives: chunks drawn from the same chapter as a reference positive but that do not answer the query. Without this signal, a model tends to learn chapter-level topical similarity rather than query-level relevance. The 26.4% share of hard negatives is large enough to force query-level discrimination without dominating the negative-class distribution needed to also recognize domain- and cross-domain-level negatives.

Table 1. Comparison of the features extraction & features matching result

Source	Label	N	Description
Reference chunks	Positive	375	Ground-truth chunks from Armstrong's Handbook
Retrieved overlap ≥60%	Positive	376	Verified-relevant retrieval results
Same-chapter hard negatives	Negative	600	Same chapter, does not answer the query
Same-domain, other chapter	Negative	375	Medium-difficulty negative
Cross-domain	Negative	375	Easy negative
Retrieved overlap <30%	Negative	170	Verified-irrelevant retrieval results

The two decision thresholds (θupper and θlower) are set automatically from the distribution of the model's logit scores on the validation set, rather than set manually. A chunk is judged CORRECT when its logit is at least θupper, AMBIGUOUS when the logit falls between θlower and θupper, and INCORRECT when the logit is below θlower.

3.1.4 Knowledge Base and Ablation Design

Armstrong's Handbook (16th edition; 83 chapters, 1,671 pages) is split into chunks by semantic chunking (about 500 tokens each, with a 50-token overlap). The chunks are spread across four domain collections: recruitment, performance management, career development, and a shared HRM-foundations collection. Each collection is stored twice, once as a plain variant and once as an SRAG-enriched variant, giving eight

Qdrant collections in total (1,032 chunks). The 500-token size is a deliberate balance: it is large enough to keep multi-paragraph procedures intact, yet small enough for precise retrieval, and it matches the typical subsection length of 300 to 700 words in the handbook.

Three domains were chosen because they exhibit distinct epistemological profiles that stress different aspects of retrieval: recruitment is dominated by analytical/comparative queries requiring multi-aspect matching; performance management by evaluative/procedural queries where an incorrect procedural answer has direct employee impact; and career development by informational/predictive queries that often require cross-chapter context.

To isolate the individual contribution of CRAG and SRAG against the Naive RAG baseline, a three-variant ablation is used (Table 2): Naive RAG (baseline), SRAG-only (isolating metadata enrichment), and the closed-domain CRAG adaptation (C1+C2, the paper's primary contribution).

Table 2. Three-variant ablation design.

Variant	Collection	T5 evaluator	Role
Naive RAG	plain	No	Lower baseline
SRAG-only	srag	No	Isolates metadata-enrichment contribution
CRAG (adapted)	plain	Yes (C2)	Primary C1+C2 contribution

Two hypotheses are tested, using Faithfulness as the primary metric because it most directly reflects whether the retrieval/correction mechanism grounds the response in retrieved context:

- H1. The closed-domain CRAG adaptation achieves significantly higher Faithfulness than Naive RAG, because relevance evaluation and corrective action filter out irrelevant chunks before generation.
- H2. SRAG-only achieves significantly higher Faithfulness than Naive RAG, because structured metadata enrichment improves query-chunk vector alignment.

The evaluation dataset comprises 270 synthetic queries (90 per domain), stratified into a dominant query type (54 queries) and a secondary type (36 queries) per domain, each domain's dominant type matching its epistemological profile (Table 3). Each query is paired with an LLM-generated reference answer produced exclusively from the top-five retrieved chunks of Armstrong's Handbook, with an explicit instruction confining the answer to that context. Reference-answer quality was verified on a stratified random 10% sample (27 queries, proportional per domain and query type) against the handbook text through an AI-assisted check whose findings were all manually adjudicated: 21 answers were fully accurate, 5 contained minor inaccuracies (composite quotations or explicitly flagged extrapolation), and 1 exceeded the source scope, with none contradicting the handbook. Exhaustive manual verification of all 270 reference answers was not performed; the implications of this limitation are discussed in Section 6.

Table 3. Query type distribution in the evaluation dataset (270 queries).

Domain	Dominant type (n)	Secondary type (n)	Total

Recruitment	Analytical (54)	Comparative (36)	90
Performance Mgmt.	Evaluative (54)	Procedural (36)	90
Career Dev.	Informational (54)	Predictive (36)	90
Total			**270**

4 Experiment Results

The system is deployed on a self-hosted Virtual Private Server using Docker. The n8n low-code workflow engine orchestrates the domain orchestrators and the CRAG sub-workflow via visual tool-call composition. Qdrant is used as the vector database for its per-collection payload filtering, needed to keep the eight ablation collections isolated. The bge-m3 embedding model (via Ollama) produces 1,024-dimensional multilingual vectors, bridging the mismatch between the English-language knowledge base and the predominantly Indonesian-language user queries. DeepSeek Chat serves as the generator and metadata-enrichment agent. The fine-tuned flan-t5-large evaluator is served as a FastAPI HTTP endpoint (port 8001), invoked for each retrieved query-chunk pair during evaluation by the ablation script. Supabase (PostgreSQL) stores operational data (e.g., employee KPI records), and a Next.js frontend/backend provides the end-user conversational interface.

4.1 Indexing Pipeline and SRAG Enrichment

Armstrong's Handbook is extracted from a text-extractable PDF into 1,671 pages across 83 chapters, each chapter mapped to one of four domain collections, then chunked (500 tokens, 50 token overlap, *cl100k_base* tokenizer) into 1,032 chunks (recruitment 95, performance management 122, career development 104, foundations 711). For every chunk, a DeepSeek Chat agent extracts four metadata fields (topics, chunk type, knowledge-graph triples, semantic tags; the sentiments field from the original SRAG design is intentionally omitted, as discussed in Section 3.3). The resulting *enriched_text* is embedded with bge-m3 and stored in the *_srag collections, while the *_plain collections store embeddings of the unmodified chunk text. Both collection payloads retain the clean chunk text, so the generator always receives unenriched context regardless of which vector space was searched.

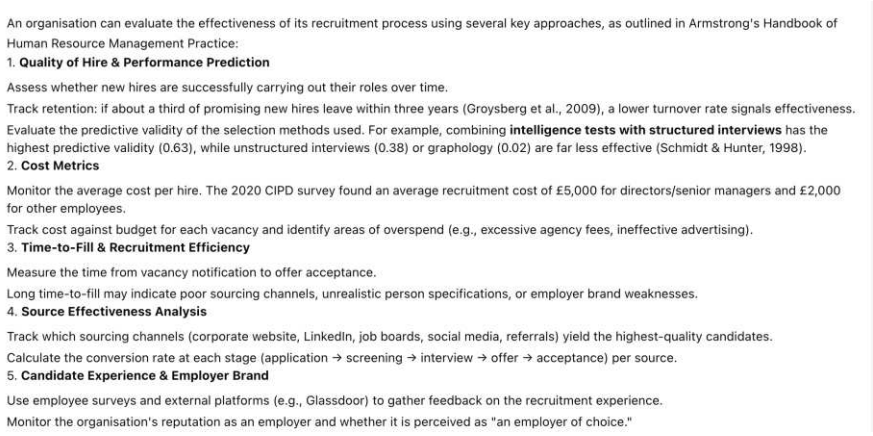
4.2 Closed-Domain CRAG Implementation

The fine-tuned flan-t5-large evaluator reached 91.8% validation accuracy, with automatically derived verdict thresholds of $\theta_{upper} = -0.55$ and $\theta_{lower} = -0.65$. At inference, the CRAG sub-workflow routes each query-chunk pair through one of three paths: CORRECT chunks proceed to Knowledge Refinement (decompose into sentence-level strips, capped at 15 per chunk, re-score each strip with the same T5 evaluator, and recompose the surviving strips); INCORRECT reformats the query with the fixed prefix "human resource management framework principles: {query}" and re-retrieves on the plain collection; AMBIGUOUS reformats the query with the fixed prefix "HRM best practices: {query}", re-retrieves, and merges the result with the original retrieval, de-duplicated. This is the operational realization of contribution C1:

every fallback path stays inside the closed Qdrant knowledge base instead of querying the open web.

4.3 Multi-Domain Orchestration and Operational Data Integration

Three separate n8n workflows implement one orchestrator per HR domain, each with a domain-specific system prompt and conversational context, all calling the same CRAG sub-workflow as a tool with a domain-appropriate Qdrant collection parameter. This realizes a hierarchical agentic RAG pattern: the orchestrator plays the role of the router/master agent, and the CRAG sub-workflow the specialized retrieval sub-agent [24]. For the performance-management domain, an additional Supabase tool retrieves company-specific employee KPI records (table employee_kpi), which the orchestrator combines with normative guidance retrieved from Armstrong's Handbook, so responses can be grounded in both real operational data and standardized HRM practice. Fig. 5 shows example interactions with the deployed system across the three domains, each answer grounded in Armstrong's Handbook.


An organisation can evaluate the effectiveness of its recruitment process using several key approaches, as outlined in Armstrong's Handbook of Human Resource Management Practice:

1. **Quality of Hire & Performance Prediction**

Assess whether new hires are successfully carrying out their roles over time.

Track retention: if about a third of promising new hires leave within three years (Groysberg et al., 2009), a lower turnover rate signals effectiveness.

Evaluate the predictive validity of the selection methods used. For example, combining **intelligence tests with structured interviews** has the highest predictive validity (0.63), while unstructured interviews (0.38) or graphology (0.02) are far less effective (Schmidt & Hunter, 1998).

2. **Cost Metrics**

Monitor the average cost per hire. The 2020 CIPD survey found an average recruitment cost of £5,000 for directors/senior managers and £2,000 for other employees.

Track cost against budget for each vacancy and identify areas of overspend (e.g., excessive agency fees, ineffective advertising).

3. **Time-to-Fill & Recruitment Efficiency**

Measure the time from vacancy notification to offer acceptance.

Long time-to-fill may indicate poor sourcing channels, unrealistic person specifications, or employer brand weaknesses.

4. **Source Effectiveness Analysis**

Track which sourcing channels (corporate website, LinkedIn, job boards, social media, referrals) yield the highest-quality candidates.

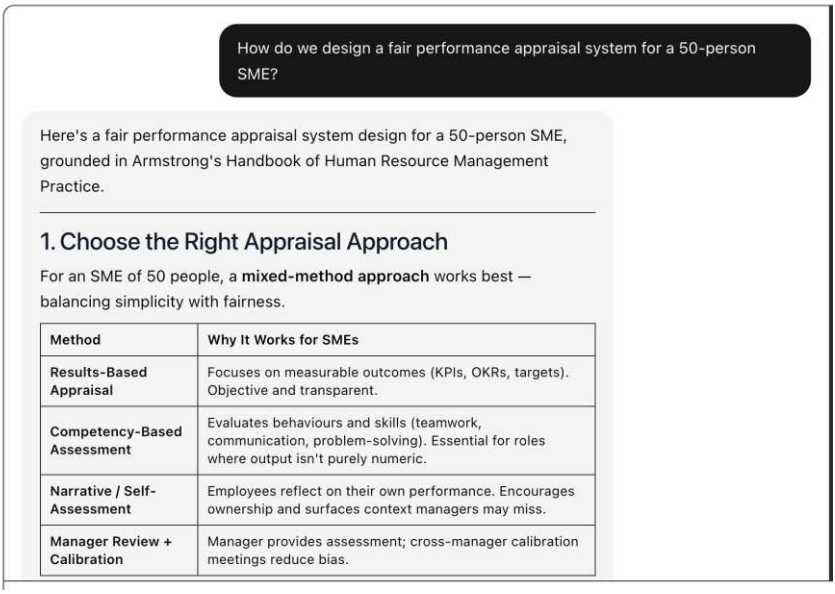
Calculate the conversion rate at each stage (application → screening → interview → offer → acceptance) per source.

5. **Candidate Experience & Employer Brand**

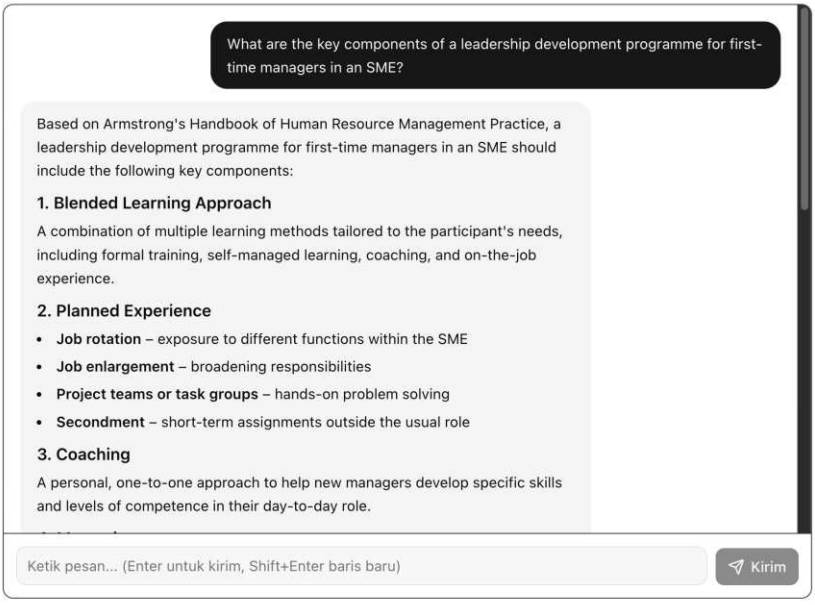
Use employee surveys and external platforms (e.g., Glassdoor) to gather feedback on the recruitment experience.

Monitor the organisation's reputation as an employer and whether it is perceived as "an employer of choice."

(a)



(b)



(c)

Fig. 5 Example interactions with the deployed Smart HR system across the three HR domains, each answer grounded in Armstrong's Handbook (16th edition): (a) recruitment, (b) performance management, and (c) career development.

4.4 Ablation Environment

The three ablation variants are implemented as selectable pipeline configurations rather than duplicated workflows: Naive RAG embeds the raw query and searches the plain collection with no evaluator; SRAG-only embeds an enrich_query() representation (matching the chunk's Topics/Type/Concepts/Tags format) and searches the srag collection with no evaluator; and the CRAG variant searches plain, scores results with the T5 evaluator, and applies the corrective action matching the verdict. All three variants are run against the same 270-query dataset, producing 810 total inferences for the analysis in Section 5. The T5 evaluator is used in full for every CRAG inference in this ablation via the FastAPI endpoint. In the production n8n deployment on a four-vCPU CPU-only VPS, however, flan-t5-large inference takes about 63 s per chunk, so the production evaluator is substituted with a zero-temperature LLM prompt that yields the identical CORRECT/AMBIGUOUS/INCORRECT verdict schema; the reference T5-evaluator configuration is included in the repository as a workflow ready to activate on adequate hardware. This latency constraint is consistent with the latency trade-off discussed in Section 6.

4.5 Evaluation Procedure

RAGAS [28] Faithfulness, Answer Relevancy, Context Precision, and Context Recall, and LLM as-Judge [29] Accuracy, Completeness, and Hallucination (0 to 10 scale; lower Hallucination is better) are reported, together with per-query Latency and the Corrective Action Rate (CAR), the share of CRAG queries that trigger an AMBIGUOUS or INCORRECT verdict (CAR is 0 by construction for Naive RAG and SRAG-only, which have no evaluator). Hypotheses H1 and H2 are tested with a two-sided paired Wilcoxon signed-rank test ($\alpha = 0.05$) on Faithfulness, at $n = 90$ per domain and $n = 270$ pooled across domains, with Cohen's d effect sizes (small ≥ 0.2, medium ≥ 0.5, large ≥ 0.8). A Kruskal-Wallis test supplements the paired comparisons as a simultaneous three-variant check.

4.6 Quantitative Results

Table 4 reports results per variant and domain, and Table 5 the three-domain averages.

Table 4. Evaluation results per variant and domain (n=90 per cell; R = Recruitment, P = Performance, C = Career; Hallucination on a 0 to 10 scale, lower is better).

Variant	Dom.	Faith.	Ans.Rel.	Acc.	Comp.	Hall.	CtxPrec.	CtxRec.
Naive	R	0.618	0.965	8.80	7.60	1.30	0.367	0.585
Naive	P	0.642	0.961	8.56	7.17	1.53	0.429	0.558
Naive	C	0.585	0.955	8.73	7.23	1.64	0.558	0.469
SRAG	R	0.697	0.969	8.20	6.97	1.84	0.327	0.526
SRAG	P	0.636	0.964	7.86	6.50	2.08	0.330	0.429
SRAG	C	0.581	0.967	8.12	6.76	2.04	0.524	0.415
CRAG	R	0.793	0.970	8.13	6.71	1.86	0.420	0.585
CRAG	P	0.765	0.971	8.07	6.84	1.50	0.446	0.544

CRAG	C	0.714	0.956	8.30	6.82	1.80	0.558	0.452

Table 5. Mean evaluation results per variant (n=270, three domains pooled).

Variant CAR	Faith.	Ans.Rel.	Acc.	Comp.	Hall.	CtxPrec.	CtxRec.	Lat.(s)
Naive 0.00	0.615	0.960	8.70	7.33	1.49	0.451	0.537	21.9
SRAG 0.00	0.638	0.967	8.06	6.74	1.99	0.394	0.457	18.7
CRAG 0.28	**0.757**	0.966	8.17	6.77	1.72	0.475	0.527	42.1

Table 6 summarizes the Wilcoxon signed-rank test for H1 and H2. H1 is confirmed: the closed-domain CRAG adaptation achieves significantly higher Faithfulness than Naive RAG ($p < 0.001$, $d = 0.457$, medium effect), and this holds in every domain individually (Table 7: Recruitment $d = 0.561$, Performance $d = 0.397$, Career $d = 0.413$). H2 is not confirmed: SRAG-only does not significantly outperform Naive RAG on Faithfulness ($p = 0.388$, $d = 0.071$); in Performance and Career, SRAG-only's mean Faithfulness is in fact marginally lower than Naive RAG (0.636 vs. 0.642 and 0.581 vs. 0.585).

Table 6. Mean evaluation results per variant (n=270, three domains pooled).

Hyp.	Comparison	Metric	*p*	Cohen's Status
H1	CRAG vs. Naive	Faithfulness	< 0.001	+0.457 **Confirmed**
H2	SRAG-only vs. Naive	Faithfulness	0.388	+0.071 Not confirmed

Table 7. H1 results per domain: CRAG vs. Naive RAG (Faithfulness)

Domain	Mean CRAG	Mean Naive	*p*	Cohen's *d*
Recruitment	0.793	0.618	< 0.001	+0.561
Performance	0.765	0.642	0.004	+0.397
Career Dev.	0.714	0.585	0.003	+0.413

Exploratory per-metric, per-domain analysis (beyond the two formal hypotheses) additionally shows that CRAG significantly suppresses Hallucination relative to SRAG in the Performance domain (1.50 vs. 2.08, $p = 0.035$, $d = -0.331$), consistent with the role of Knowledge Refinement in filtering context, that SRAG-only approaches significance against Naive RAG in Recruitment (0.697 vs. 0.618, $p = 0.063$), and that both augmented variants (CRAG and SRAG) show significantly lower LLM-as-Judge Accuracy than Naive RAG across all domains ($p \leq 0.020$, d between −0.28 and −0.75), discussed further below.

5 Discussion and future work

5.1 Indexing Pipeline and SRAG Enrichment

H1's medium effect size ($d = 0.457$, $p < 0.001$) and its consistency across all three domains ($d = 0.561$, 0.397, 0.413) indicate that the Faithfulness gain reflects a stable contribution of the corrective mechanism rather than an artifact of a single domain. The mean CAR of 0.28 (Table 5) is itself evidence that the T5 evaluator functions as designed: on a curated knowledge base such as Armstrong's Handbook, an evaluator that only learns chapter-level topic similarity would classify nearly everything as CORRECT and CAR would be close to zero. A CAR of 0.28 shows the same-chapter hard-negative strategy (Section 3.5) successfully taught query-level discrimination, allowing the corrective mechanism to actually trigger.

RAGAS Context Precision and Context Recall (Table 5) reinforce this interpretation. CRAG and Naive RAG show comparable Context Precision and Context Recall (CRAG: 0.475/0.527; Naive: 0.451/0.537), so the substantial Faithfulness gap (0.757 vs. 0.615) cannot be attributed to better initial retrieval quality. This indicates that CRAG's Knowledge Refinement mechanism, rather than superior initial retrieval, is the primary source of the Faithfulness improvement.

5.2 Metadata Enrichment Alone Is Insufficient

H2's rejection ($p = 0.388$, $d = 0.071$) is not merely a null statistical result; it exposes a structural limitation of SRAG without a relevance-evaluation layer. Structured metadata enrichment pulls in chunks that are topically related to a query but does not guarantee that the retrieved chunk specifically answers it. Without an evaluator to make that distinction, the generator receives more varied but not necessarily more accurate context, consistent with SRAG-only's higher Hallucination score (1.99) relative to Naive RAG (1.49), suggesting that more varied context gives the generator more room to fill informational gaps with claims not fully grounded in the retrieved content [19]. In a closed-domain system with an already-curated knowledge base, this suggests that investing in metadata enrichment alone does not measurably improve answer quality unless paired with a relevance-evaluation layer.

5.3 Faithfulness vs. Accuracy Trade-off

Both CRAG and SRAG show significantly lower LLM-as-Judge Accuracy than Naive RAG ($p \leq 0.020$, d between -0.28 and -0.75), so CRAG's Faithfulness gain comes with an Accuracy cost. This follows from the differing definitions of the two metrics: Faithfulness measures the share of claims traceable to retrieved context, while LLM-as-Judge Accuracy scores factual correctness holistically, including knowledge not necessarily present in the retrieved context. CRAG's Knowledge Refinement trims context to what is judged relevant, producing answers more faithful to the source but potentially less complete from a holistic-judge perspective; Naive RAG's wider, unfiltered context gives the generator more background information, raising perceived Accuracy at the cost of traceability. For an SME HR system, Faithfulness is arguably the more critical property: an unverifiable answer carries a larger compliance risk than an answer limited to what is available in Armstrong's Handbook.

5.4 Limitations

SRAG-only approaches significance only in Recruitment ($p = 0.063$), suggesting its benefit may be more measurable for domains with highly variable comparative terminology; this warrants a larger sample or a query design more targeted at the query types metadata enrichment is expected to help most. The evaluation has two further limitations. First, all evaluation queries are synthetic, LLM-generated queries, so their distribution may not fully reflect the variation of real queries from MSME HR staff. Second, validation from an HR practitioner's perspective (whether the differences measured automatically are also meaningful for real decision-making in an MSME context) was not conducted within the scope of this paper, due to time and evaluator access constraints; only one LLM provider (DeepSeek Chat) is evaluated as well.

5.5 Conclusion and Future Work

This paper proposed and evaluated a closed-domain adaptation of Corrective RAG combined with a Structured RAG metadata-enrichment layer for a multi-agent Smart HR system spanning recruitment, performance management, and career development, grounded exclusively in Armstrong's Handbook of HRM Practice (16th edition) [15]. Two technical contributions were realized: (C1) a closed-domain CRAG adaptation that replaces the original web-search fallback with re-retrieval against the same curated Qdrant knowledge base, and (C2) a flan-t5-large retrieval evaluator trained with a same-chapter hard-negative strategy so that it discriminates query-level, not merely chapter-level, relevance. The corrective mechanism was shown to be operationally active, with a mean Corrective Action Rate of 0.28.

A three-variant ablation study ($n = 90$ per domain, 810 inferences) confirmed one of two hypotheses at $\alpha = 0.05$: the closed-domain CRAG adaptation achieves significantly higher Faithfulness than Naive RAG across all three domains (H1, $p < 0.001$, $d = 0.457$), while SRAG-only metadata enrichment alone does not (H2, $p = 0.388$, $d = 0.071$). Both augmented variants show a consistent trade-off of lower LLM-as-Judge Accuracy relative to Naive RAG. Together, these results indicate that C1 (the closed-domain CRAG adaptation) is the primary contributor to improved Faithfulness, C2 (the T5 hard-negative evaluator) is the enabling component that lets C1 operate reliably, and metadata enrichment without a relevance-evaluation layer is not, by itself, sufficient to improve closed-domain answer quality. CRAG's effect is also domain-dependent, strongest in Recruitment ($d = 0.561$), whose comparative queries most require precise source grounding.

Future work should (i) enlarge the sample size and extend to additional HR domains to strengthen generalizability, particularly for the smaller observed effects; (ii) explore mechanisms, such as generation-prompt adjustments or blending original and corrected context, to mitigate the Faithfulness/Accuracy trade-off; (iii) explore per-domain adaptive CRAG configuration, given that the H1 effect size varies across domains; and (iv) add a fourth ablation variant combining SRAG and the CRAG adaptation. This paper deliberately keeps SRAG and CRAG isolated so their individual contributions can be attributed cleanly; H2's rejection shows SRAG has no significant effect on its own, but whether SRAG's richer indexing-time representation strengthens the T5 evaluator's discrimination at inference time, that is, whether the two mechanisms interact rather than merely coexist, remains untested.

6 Declaration

Funding

This research was funded by Hibah P2MI ITB (Project ID: STEI.PPMI-1-47-2026).

Competing Interest

The authors have no relevant financial or non-financial interests to disclose. All authors certify that they have no affiliations with or involvement in any organization or entity with any financial or non-financial interest in the subject matter or materials discussed in this manuscript.

Ethics approval

Not applicable. This study did not involve human participants, their identifiable data, or biological material, nor did it involve animals.

Consent to participate

Not applicable. This study did not involve human participants; therefore, informed consent to participate was not required.

Consent to publish

Not applicable. This manuscript does not contain any individual person's identifiable data, images, or personal details; therefore, consent to publish was not required.

Availability of data and material

The knowledge source used in this study, Armstrong's Handbook of Human Resource Management Practice (16th edition), is a commercially published, copyrighted work and cannot be redistributed by the authors; it is available from the publisher as cited in the References. The 270 synthetic evaluation queries, the per-query evaluation results, and the custom code developed for the system, including the fine-tuned flan-t5-large retrieval evaluator, are available from the corresponding author on reasonable request.

Authors' contributions

All authors contributed to the study conception and design. System design, implementation, data preparation, and experiments were carried out by Devi Willieam Anggara and Marcellino Joestiawan. Methodology guidance, supervision, and critical review of the manuscript were provided by Mohd Shafry Mohd Rahim, Ibrahim Salim Sulaiman Al Farsi, Fetty Fitriyanti Lubis, and Dion Tanjung. The first draft of the manuscript was written by Marcellino Joestiawan, and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Acknowledgment

This research is financially supported by Hibah P2MI ITB (Project ID: STEI.PPMI-1-47-2026). The authors gratefully acknowledge the research facilities and collaborative framework provided by Institut Teknologi Bandung, Universiti Teknologi Malaysia, and Sohar University. Furthermore, we are deeply thankful to our academic lecturers and supervisors for their continuous mentorship, rigorous evaluations, and insightful feedback throughout the drafting and revision of this manuscript

7 References

[1] Kementerian Koordinator Bidang Perekonomian RI. *UMKM Jadi Penopang Utama Perekonomian Indonesia*. 2023. URL: https://ekon.go.id/publikasi/detail/4749/umkmjadi-penopang-utama-perekonomian-indonesia.

[2] S. S. b. Omar, L. Arokiasamy, and M. Ismail. "The Background and Challenges Faced by the Small Medium Enterprises: A Human Resource Development Perspective". In: *International Journal of Business and Management* 4.10 (2009), pp. 95–107.

[3] T. Tranggono, E. E. Y. Amriel, E. Sholihatin, and N. Haryono. "Small and Medium Micro Enterprises Innovation Strategies in Efforts to Face the Industry Revolution 4.0". In: *Journal of Economics, Business, and Government Challenges* 4.1 (2021), pp. 24–31.

[4] H. Hernita, B. Surya, I. Perwira, H. Abubakar, and M. Idris. "Economic Business Sustainability and Strengthening Human Resource Capacity Based on Increasing the Productivity of Small and Medium Enterprises (SMEs) in Makassar City, Indonesia". In: *Sustainability* 13 (2021), p. 3177. DOI: 10.3390/su13063177.

[5] T. S. Anuradha and H. N. Ramesh. "HRM Practices in MSMEs: Remained Unprofessional". In: *International Journal for Research in Applied Science and Engineering Technology* 10.9 (2022). ISSN: 2321-9653.

[6] A. Mukhlis, A. Moeins, and W. Sunaryo. "Development Strategies for Micro, Small, and Medium Enterprises (MSME) by Improving the Quality of Human Resources". In: *International Journal of Economy, Education and Entrepreneurship* 2.2 (2022), pp. 525–556. DOI: 10.53067/ije3.v2i2.91.

[7] L. C. Nawangsari, Siswanti, and M. Soelton. "Human Resources Management Strategy for Business Sustainability in MSMEs". In: *International Conference on Community Development (ICCD)* (2023). DOI: 10.33068/iccd.v5i1.599.

[8] A. Abuhantash. "The Impact of Human Resource Information Systems on Organizational Performance: A Systematic Literature Review". In: *European Journal of Business and Management Research* 8.3 (2023). DOI: 10.24018/ejbmr.2023.8.3.1992.

[9] N. Panjaitan. “Implementing HRIS for Efficient Human Resource Management”. In: *International Journal of Science and Society* 5.2 (2023), pp. 128–141.

[10] L. Anatan and Nur. “Micro, Small, and Medium Enterprises’ Readiness for Digital Transformation in Indonesia”. In: *Economies* 11.6 (2023), p. 156. DOI: 10.3390/economies11060156.

[11] D. Schmid, J. Bueechl, R.-C. Härting, Y. Beldarrain, R. Schwörer, and J. Naeve. “Opportunities and Barriers of Agility and Digital Processes in HR in SMEs”. In: *Procedia Computer Science* 246 (2024), pp. 3605–3613. DOI: 10.1016/j.procs.2024.09.197.

[12] R. Mithun and S. Roopadarshini. “A Study on Impact of Smart HRM in Bengaluru Smart City”. In: *International Journal of Applied and Scientific Research* 1.4 (2023), pp. 417–426.

[13] R. Panchbhai, M. Chauhan, E. Pal, and S. Goel. “Exploring Smart HR 4.0 in Digital HR Transformation for Performance, Engagement, and Sustainable Organizational Goals”. In: *Journal of International Commercial Law and Technology* 6.1 (2025), pp. 1796–1802.

[14] P. Lewis et al. “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks”. In: *arXiv preprint arXiv:2005.11401* (2020).

[15] M. Armstrong and S. Taylor. *Armstrong’s Handbook of Human Resource Management Practice*. 16th ed. London: Kogan Page, 2023.

[16] Y. Gao et al. “Retrieval-Augmented Generation for Large Language Models: A Survey”. In: *arXiv preprint arXiv:2312.10997v5* (2024).

[17] D. Metropolitansky and J. Larson. “VeriTrail: Closed-Domain Hallucination Detection with Traceability”. In: *The Fourteenth International Conference on Learning Representations (ICLR 2026)*. OpenReview.net, 2026.

[18] S.-Q. Yan, J.-C. Gu, Y. Zhu, and Z.-H. Ling. “Corrective Retrieval Augmented Generation”. In: *arXiv preprint arXiv:2401.15884v3* (2024).

[19] P. Shah, S. Venkatesan, R. Anand, and N. Goel. “Structured RAG: Enhancing Retrieval-
Augmented Generation with Metadata-Enriched Representations”. In: *arXiv preprint arXiv:2603.26670v* (2026).

[20] T. G. Mauro and J. E. Borges-Andrade. “Human Resource System as Innovation for Organisations: A Literature Meta-analysis”. In: *Innovation & Management Review* 16.1 (2019), pp. 2–19. DOI: 10.1108/INMR-06-2018-0039.

[21] T. Bondarouk and H. Ruël. “Electronic Human Resource Management: Challenges in the Digital Era”. In: *International Journal of Human Resource Management* 20.3 (2009), pp. 505–514.

[22] D. Chen, A. Fisch, J. Weston, and A. Bordes. “Reading Wikipedia to Answer OpenDomain Questions”. In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by R. Barzilay and M.-Y. Kan. Vancouver, Canada: Association for Computational Linguistics, July 2017, pp. 1870–
1879. DOI: 10.18653/v1/P17-1171. URL: https://aclanthology.org/P171171/.

[23] D. Mollá and J. L. Vicedo. “Question Answering in Restricted Domains: An Overview”. In: *Computational Linguistics* 33.1 (2007), pp. 41–61. DOI: 10.1162/coli.2007.33.1.41. URL: https://aclanthology.org/J07-1004/.

[24] A. Singh, A. Ehtesham, S. Kumar, and T. T. Khoei. “Agentic Retrieval-Augmented Generation: A Survey on Agentic RAG”. In: *arXiv preprint arXiv:2501.09136* (2025).

[25] S. Yao et al. “ReAct: Synergizing Reasoning and Acting in Language Models”. In: *The Eleventh International Conference on Learning Representations (ICLR 2023)*. 2023.

[26] S. S. Idrus, S. M. Yusof, N. Abdullah, and W. N. W. Osman. “Career Development Strategies to Improve Human Resources Sustainability in Organisations”. In: *International Journal of Academic Research in Business and Social Sciences* 13.5 (2023). DOI: 10.6007/IJARBSS/v13-i5/17022.

[27] D. Chen, A. Fisch, J. Weston, and A. Bordes. “Reading Wikipedia to Answer OpenDomain Questions”. In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics* (2017), pp. 1870–1879. DOI: 10.18653/v1/P17-1171.

[28] S. Es, J. James, L. Espinosa-Anke, and S. Schockaert. “RAGAS: Automated Evaluation of Retrieval Augmented Generation”. In: *arXiv preprint arXiv:2309.15217* (2023).

[29] L. Zheng et al. “Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena”. In: *arXiv preprint arXiv:2306.05685v4* (2023).