4th International Conference on Media, Economy, Communication and Intelligence Management (MECI 2026)

Article

Generative AI-Empowered ERP Systems: Semantic Retrieval and Business Decision Support for Unstructured Data

Xiaoyu Chen[1,*]

[1] Beijing Normal-Hong Kong Baptist University, Zhuhai, China
* Correspondence: Xiaoyu Chen, Beijing Normal-Hong Kong Baptist University, Zhuhai, China

Abstract: Enterprise Resource Planning (ERP) systems are effective in processing structured transactional data but remain limited in exploiting heterogeneous unstructured information such as financial reports, procurement descriptions, supplier notes, and internal textual records. Existing retrieval-augmented generation approaches improve information access, yet semantic retrieval, evidence utilization, and confidence control are often weakly coupled, which may reduce traceability in business decision support. This study proposes an ERP-oriented Semantic Retrieval and Decision Support (ERP-SRDS) framework integrating metadata-aware semantic encoding, hybrid lexical–semantic retrieval, cross-encoder evidence reranking, evidence-grounded generation, and confidence-based output control. Experiments on FiQA-2018 and FinQA show that ERP-SRDS achieves an nDCG@10 of 0.421 ± 0.008, a decision accuracy of 67.8 ± 1.3%, and an Evidence Coverage@5 of 92.4 ± 0.9%. Although BGE-M3 attains a slightly higher FiQA nDCG@10 of 0.429 ± 0.007, ERP-SRDS provides stronger downstream reasoning and evidence coverage. Under 30% irrelevant-context perturbation, its decision accuracy remains 63.9 ± 1.6%, with a 3.9-percentage-point decline compared with 7.0 points for RankRAG. These results indicate that combining retrieval relevance, evidence grounding, and confidence control can improve the traceability and robustness of generative AI-assisted ERP decision support without replacing managerial judgment.

Keywords: Generative AI; Enterprise Resource Planning; Semantic Retrieval; Retrieval-Augmented Generation; Business Decision Support

1. Introduction

1.1. Research Background and Problem Definition

Enterprise Resource Planning (ERP) systems integrate core organizational functions such as finance, procurement, inventory, sales, and production. Conventional ERP architectures are effective in processing structured information, including transaction records, inventory quantities, accounting entries, and standardized master data. These data can be stored in predefined fields and relational tables, making them suitable for deterministic querying, aggregation, and reporting [1].

However, many business decisions also depend on unstructured or semi-structured information that is not easily represented through fixed database fields. Typical examples include procurement descriptions, supplier notes, financial reports, customer-service records, internal textual reports, and policy or contract documents. Such information often contains inconsistent terminology, contextual dependencies, business-specific abbreviations, and relationships distributed across multiple documents. As a result, field-based queries and keyword search may fail to identify relevant evidence when users express the same business concept using different language or when a decision requires information from several sources [2].

Received: 05 July 2026
Revised: 15 August 2026
Accepted: 27 August 2026
Published: 30 August 2026

Generative artificial intelligence (AI), combined with semantic retrieval, provides a possible means of connecting these textual resources with ERP-oriented analytical processes [3]. Its practical usefulness, however, depends on whether retrieved information is relevant, generated outputs are supported by identifiable evidence, and unreliable results can be controlled. This study therefore addresses the following question: How can generative AI convert heterogeneous unstructured business information into retrievable, evidence-grounded, and decision-relevant information without weakening traceability?

1.2. Research Gap

Three limitations remain in existing approaches. First, conventional ERP retrieval primarily relies on structured attributes and lexical matching. Although effective for exact product codes, dates, amounts, and standardized terms, it provides limited semantic recall when equivalent concepts are expressed through synonyms, abbreviations, or context-specific language.

Second, retrieval-augmented generation (RAG) typically connects retrieval and generation as separate processing stages. Retrieving a relevant passage does not necessarily ensure that the generation model uses the correct evidence when producing a numerical answer or business interpretation. Consequently, decision errors may still occur even when document retrieval is successful [4].

Third, existing generative decision-support mechanisms often lack explicit controls for evidence sufficiency and output confidence. When retrieved information is incomplete, contradictory, or affected by irrelevant context, the model may still generate a definitive response. Such behavior reduces traceability and reliability in ERP-oriented applications [5].

1.3. Research Objectives and Contributions

This study develops an ERP-oriented Semantic Retrieval and Decision Support (ERP-SRDS) framework with three experimentally verifiable contributions. (1) ERP-aware hybrid semantic retrieval. A metadata-aware retrieval mechanism combines lexical matching with dense semantic representations. The lexical component retains sensitivity to exact business terms and numerical identifiers, while semantic encoding improves retrieval across inconsistent terminology. Its effect is evaluated through the overall comparison, retrieval ablation, and convergence analysis. (2) Evidence-grounded generative decision support. A reranking and evidence-grounded generation mechanism connects retrieved passages with textual or numerical outputs. Candidate evidence is reordered according to query relevance before generation, reducing the gap between retrieval quality and decision correctness. This mechanism is evaluated through decision accuracy, Evidence Coverage@5, and module ablation. (3) Confidence-based decision control. A confidence controller evaluates retrieval quality and evidence sufficiency before releasing an output. When evidence is inadequate, the framework performs additional retrieval or suppresses low-confidence results. Its contribution is examined through controller ablation, noise-based stability testing, and cross-dataset robustness analysis.

1.4. Technical Route and Research Significance

The framework first preprocesses and synchronizes heterogeneous business texts and associated metadata. It then constructs semantic representations, combines lexical and dense retrieval, reranks candidate evidence, generates evidence-grounded outputs, and evaluates confidence before presenting the final result. Methodologically, the study integrates semantic retrieval, evidence grounding, and retrieval-generation coupling within a unified ERP-oriented framework. Practically, it focuses on traceability, robustness, auditable evidence, and controlled decision support. The system is intended to assist managerial analysis by improving access to dispersed business information rather than replacing managerial judgment.

2. Related Works

2.1. Generative AI in ERP and Enterprise Information Systems

Generative AI has expanded the interaction model of enterprise information systems by allowing users to access organizational information through natural-language queries rather than predefined menus, report templates, or database commands. Large language models can summarize heterogeneous documents, reformulate business information, and answer questions using textual context, thereby reducing the technical barrier between enterprise data and non-technical users [6]. Their ability to process distributed organizational knowledge also creates opportunities for supporting business-process analysis and decision-oriented information access.

However, enterprise information environments differ substantially from general-purpose question answering. ERP information is organized around functional modules, transaction periods, document origins, organizational responsibilities, and access-control boundaries. A semantically plausible response may therefore remain operationally inappropriate if it ignores the source, time, or business context of the underlying information. General-purpose generative models do not explicitly represent these ERP-specific relationships, which limits their direct use as reliable business information interfaces. The relevant problem is consequently not only language generation, but also the alignment of generated content with controlled enterprise evidence.

2.2. Semantic Retrieval for Unstructured Business Data

Information retrieval for unstructured business documents can broadly be divided into lexical, dense, and hybrid approaches. Keyword-based retrieval provides strong matching for exact expressions and remains useful for product identifiers, monetary values, dates, and specialized terminology. Its main limitation is lexical dependence: semantically equivalent expressions may receive low similarity when they share few surface terms. Dense retrieval addresses this limitation by representing queries and documents in continuous semantic spaces, allowing conceptually related texts to be matched even when their wording differs [7].

Dense representations, however, can weaken sensitivity to exact terms that carry substantial business meaning. Empirical retrieval research also indicates that hybrid strategies are not uniformly superior under every domain and configuration, suggesting that sparse and dense signals should be combined according to the characteristics of the information being retrieved rather than treated as interchangeable components [8]. For ERP-oriented documents, this complementarity motivates the lexical–semantic fusion mechanism adopted in this study.

2.3. Retrieval-Augmented Generation for Decision Support

Retrieval-Augmented Generation (RAG) extends language models with external evidence and can reduce dependence on information stored only in model parameters. Its practical advantage is that retrieved documents can provide domain-specific and potentially verifiable context for generation. Recent RAG studies have accordingly emphasized retrieval quality, reranking, grounding, and interpretability as important elements of domain-specific systems [9].

Nevertheless, retrieval relevance should not be interpreted as equivalent to decision correctness. In financial, procurement, or operational analysis, a passage may be topically relevant while lacking the figures, conditions, or relationships required for a specific judgment. Moreover, multiple retrieved passages may contain partially consistent or conflicting evidence. Recent multi-evidence RAG architectures therefore incorporate cross-encoder reranking and evidence refinement before answer generation [10]. This suggests that decision-oriented RAG requires explicit candidate ranking and evidence verification before generation rather than direct generation from an unfiltered retrieval set.

2.4. Reliability, Traceability, and Remaining Research Gap

Reliability remains a central limitation because RAG does not eliminate unsupported generation. Hallucinations can originate from retrieval errors, irrelevant context, conflicting evidence, or failures to correctly use retrieved information [11]. Existing approaches consequently leave four gaps for ERP-oriented decision support: limited integration of business metadata into semantic retrieval, weak constraints between retrieved evidence and final outputs, insufficient control of low-confidence decisions, and limited robustness evaluation under noisy or distribution-shifted documents. The ERP-SRDS framework addresses these gaps by integrating metadata-aware hybrid retrieval, evidence reranking, grounded generation, and confidence-based output control within a unified decision-support mechanism.

3. Methodology

3.1. Overall Architecture of Erp-srds

The proposed ERP-oriented Semantic Retrieval and Decision Support (ERP-SRDS) framework is designed to transform heterogeneous business documents into traceable decision-support outputs while retaining explicit links between retrieved evidence and generated conclusions. As illustrated in Figure 1, the architecture contains seven functional components: unstructured business data processing, preprocessing and metadata alignment, ERP-aware semantic encoding, hybrid lexical–semantic retrieval, cross-encoder evidence reranking, evidence-grounded generation, and confidence-based decision control. The final output contains both the generated decision result and its supporting evidence trace.

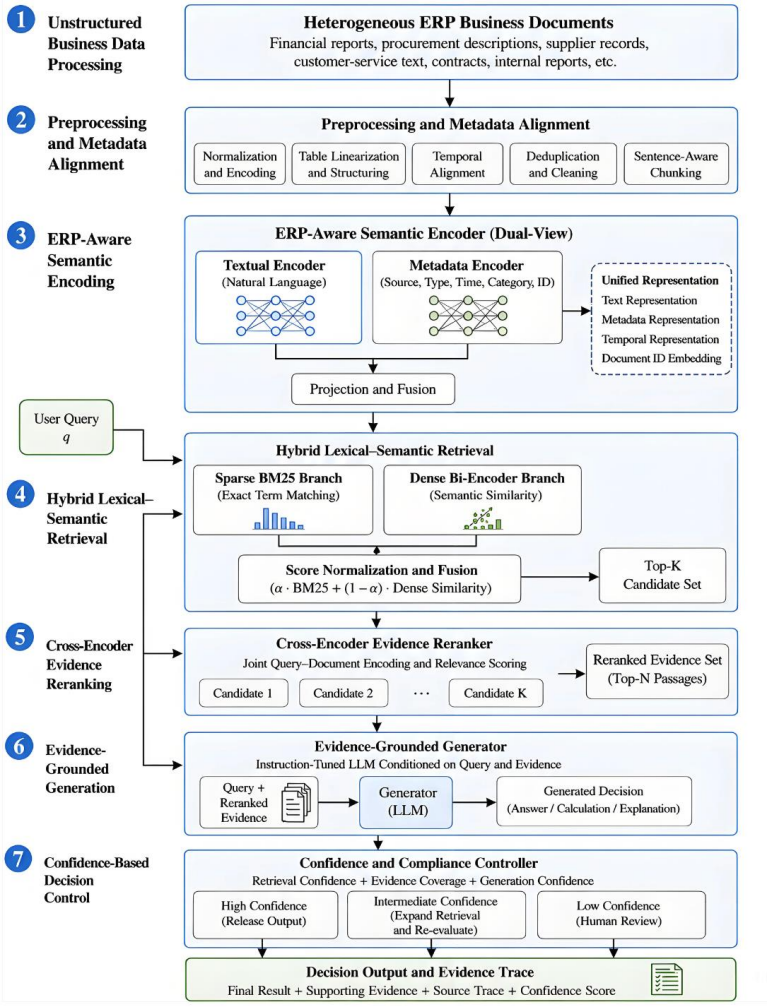

Figure 1. Overall Architecture of the Erp-srds Framework

The retrieval layer contains two complementary branches. A sparse BM25 branch preserves exact matches involving company names, account items, dates, monetary values, and business identifiers, whereas a dense bi-encoder maps queries and business documents into a shared semantic space. Candidate documents obtained from the two branches are fused and passed to a cross-encoder reranker. Unlike the bi-encoder, which encodes queries and documents independently for efficient large-scale retrieval, the cross-encoder processes each query–document pair jointly and estimates fine-grained evidence relevance. The highest-ranked passages are supplied to an instruction-tuned generator. A lightweight confidence-scoring head subsequently evaluates retrieval quality, evidence sufficiency, and generation certainty before the decision result is released.

3.2. Business Data Representation and Synchronization

ERP-oriented textual information may originate from financial reports, procurement descriptions, supplier records, customer-service text, contracts, or internal reports. Before encoding, Unicode characters and whitespace are normalized, tabular information is converted into structured text, timestamps are mapped to a consistent temporal representation, duplicates are removed, and long documents are divided using sentence-aware chunking.

For document i , the unified representation is defined as

$$\mathbf{z}_i = \mathrm{Concat}(\mathbf{x}_i, \mathbf{m}_i, \mathbf{t}_i, \mathbf{d}_i). \quad (1)$$

Here, $\mathbf{x}_i$ denotes the textual content representation, $\mathbf{m}_i$ represents source and business-category metadata, $\mathbf{t}_i$ denotes normalized temporal information, and $\mathbf{d}_i$ represents the document identifier embedding. $\mathrm{Concat}(\cdot)$ denotes feature concatenation. This representation preserves semantic content while retaining contextual attributes needed to distinguish documents containing similar language but different business meanings.

3.3. Metadata-Aware Semantic Encoding

The semantic encoder adopts a dual-view structure. The textual branch encodes natural-language content, whereas the metadata branch encodes document type, source, time, and business category. The two views are projected to the same latent dimension and combined before retrieval.

The fused document representation is calculated as

$$\mathbf{h}_i = \gamma, \mathbf{W}_x f_x(\mathbf{x}_i) + (1-\gamma), \mathbf{W}_m f_m(\mathbf{m}_i, \mathbf{t}_i, \mathbf{d}_i), \quad (2)$$

where $f_x(\cdot)$ is the text encoder, $f_m(\cdot)$ is the metadata encoder, $\mathbf{W}_x$ and $\mathbf{W}_m$ are trainable projection matrices, and $\gamma \in [0,1]$ controls the relative contribution of semantic text and business context. Relevance judgments from FiQA provide positive and negative query–document supervision, while FinQA supporting facts provide evidence-level relevance labels. This design allows the encoder to distinguish documents that are linguistically similar but belong to different business contexts.

3.4. Hybrid Lexical–Semantic Retrieval

A single retrieval paradigm is insufficient for heterogeneous enterprise data. Exact values, dates, account names, and identifiers benefit from lexical matching, whereas descriptive business queries often require semantic similarity. ERP-SRDS therefore combines sparse and dense retrieval scores.

For query q and document d_i , the fused retrieval score is

$$s_i^{ret} = \alpha, \tilde{s}_{BM25}(q, d_i) + (1-\alpha)\cos\left(\mathbf{h}_q, \mathbf{h}_i\right), \quad (3)$$

where $\tilde{s}_{BM25}$ is the normalized BM25 score, $\mathbf{h}_q$ and $\mathbf{h}_i$ are dense query and document representations, $\cos(\cdot)$ denotes cosine similarity, and $\alpha \in [0,1]$ is the lexical–semantic fusion coefficient. The highest-scoring K documents form the initial candidate set. The fusion mechanism therefore retains precise term matching without sacrificing retrieval of semantically related passages expressed using different terminology.

3.5. Cross-Encoder Evidence Reranking

The initial retrieval stage prioritizes recall and may return passages that are generally relevant but insufficient for a specific business decision. A cross-encoder is therefore

applied only to the top-(K) candidates, reducing computational cost while enabling more detailed interaction between query and document tokens.

The reranking score is defined as

$$s_i^{rank} = \sigma(\mathbf{w}_r^{\top} f_c([q; d_i]) + b_r), \quad (4)$$

where $[q; d_i]$ represents the joint query–document input, $f_c(\cdot)$ is the cross-encoder, $\mathbf{w}_r$ and b_r are trainable parameters, and $\sigma(\cdot)$ is the sigmoid function. The reranker does not attempt to retrieve additional documents. Instead, it identifies which candidates contain evidence most directly relevant to the required calculation, explanation, or managerial judgment.

3.6. Evidence-Grounded Decision Generation

After reranking, the highest-scoring passages form the supporting evidence set $E = e_1, \ldots, e_J$. The generator receives both the user query and selected evidence so that the output is conditioned on retrieved business information rather than generated solely from model parameters.

The probability of decision output y is expressed as

$$P(y \mid q, E) = \prod_{t=1}^{T} P\left(y_t \mid y_{<t}, q, E; \theta_g\right), \quad (5)$$

where T is the output length, y_t is the token generated at step t, $y_{<t}$ represents preceding output tokens, and θ_g denotes generator parameters.

To quantify whether the generated result is sufficiently supported by the retrieved evidence, ERP-SRDS calculates an evidence-grounding score:

$$g(y, E) = \frac{1}{J} \sum_{j=1}^{J} \max\left(0, \cos\left(\mathbf{u}y, \mathbf{u}e_j\right)\right), \quad (6)$$

where $\mathbf{u}y$ is the semantic representation of the generated result and $\mathbf{u}e_j$ is the representation of evidence passage e_j. Higher values indicate stronger semantic support between the decision output and retrieved evidence. For FinQA, gold answers and annotated reasoning programs additionally supervise numerical reasoning, allowing calculated results to be checked against supporting facts and intermediate operations.

3.7. Confidence-Based Decision Control

Generation quality alone is not sufficient for ERP decision support because an answer may appear linguistically coherent despite weak evidence. The confidence controller therefore combines retrieval confidence, evidence coverage, and generation confidence. It also adapts the release threshold according to validation behavior rather than using a permanently fixed value.

The threshold is updated as

$$\tau_{t+1} = \mathrm{clip}(\tau_t + \eta_\tau(\rho_t - \rho^*), \tau_{\min}, \tau_{\max}), \quad (7)$$

where τ_t is the confidence threshold at update step t, η_τ is the update rate, ρ_t is the observed false-acceptance rate on validation samples, ρ^* is the target false-acceptance rate, and $\tau_{\min}$ and $\tau_{\max}$ restrict the threshold to a predefined range. When incorrectly accepted outputs increase, the threshold becomes more conservative.

Outputs with high confidence are released with their supporting evidence. Intermediate-confidence cases trigger an enlarged retrieval set and another evidence evaluation. Results remaining below the threshold are withheld and marked for human review. The controller therefore treats uncertainty as an explicit system state rather than forcing a prediction for every query.

3.8. Joint Optimization and Algorithm Procedure

The trainable components are optimized jointly using retrieval, reranking, generation, and evidence-grounding objectives:

$$\mathcal{L} = \lambda_r \mathcal{L}_{ret} + \lambda_c \mathcal{L}_{rank} + \lambda_g \mathcal{L}_{gen} + \lambda_e \mathcal{L}_{evid}, \quad (8)$$

where $\mathcal{L}_{ret}$ is the contrastive retrieval loss, $\mathcal{L}_{rank}$ measures candidate-ranking error, $\mathcal{L}_{gen}$ is the token-level generation loss, and $\mathcal{L}_{evid}$ penalizes insufficient correspondence between generated outputs and supporting evidence. The coefficients λ_r, λ_c, λ_g, and λ_e control the contribution of the four objectives. Validation performance is used to select these coefficients and prevent any single task from dominating joint optimization.

Algorithm 1. ERP-SRDS Semantic Retrieval and Decision Support Procedure

Input: query q, document collection D metadata M, top- K value K, confidence threshold τ

Output: decision result y, supporting evidence E, confidence score c

Normalize and segment documents in D, and align each segment with metadata M.

Construct unified text, temporal, category, and identifier representations.

Encode q and all document segments using the semantic bi-encoder.

Compute sparse BM25 scores for q against the document collection.

Compute dense semantic similarity between q and document embeddings.

Fuse sparse and dense scores using coefficient α.

Select the K highest-scoring candidate passages.

Jointly encode each candidate with q using the cross-encoder.

Rerank candidates and construct supporting evidence set E.

Generate decision result y conditioned on q and E.

Calculate evidence-grounding and retrieval-confidence values.

Combine retrieval, grounding, and generation signals into confidence c.

If c is intermediate, enlarge the retrieval set and repeat evidence evaluation.

If $c < \tau$, withhold the result and assign the case for human review.

Return y, E, and c when the confidence requirement is satisfied.

4. Results and Analysis

4.1. Experimental Setup

Experiments were conducted on FiQA-2018 and FinQA, representing complementary financial retrieval and reasoning tasks. The BEIR version of FiQA-2018 contains approximately 57.6K financial documents and 648 test queries with relevance judgments and was used to evaluate semantic retrieval and ranking. FinQA contains 8,281 expert-annotated question–answer pairs derived from 2,789 financial-report pages, together with supporting facts and executable reasoning programs. Its official split comprises 6,251 training, 883 validation, and 1,147 test examples, with no overlapping source reports across subsets. The official FinQA resources are publicly available under the MIT License.

Financial numbers, percentages, currency symbols, and document identifiers were preserved during preprocessing. Unicode and whitespace were normalized, tables were linearized into structured text, and documents were divided into sentence-aware chunks of 256 tokens with a 32-token overlap. Queries were limited to 128 tokens and generator inputs to 2,048 tokens. No test-set tuning was performed.

Experiments used Python 3.10, PyTorch 2.3.1, Transformers 4.45.2, and FAISS 1.8.0 on one NVIDIA RTX 4090 with 24 GB GPU memory and 32 GB system memory. The initial retrieval depth was $K = 10$, with five passages retained as candidate evidence. The lexical–semantic coefficient α, metadata coefficient γ, and initial confidence threshold τ were set to 0.35, 0.70, and 0.68, respectively. AdamW optimization used a learning rate of 2×10^{-5}, batch size 16, and a maximum of 15 epochs.

Four baselines were evaluated: BM25 + Generator [12], DPR + Generator [13], BGE-M3 + Generator [14], and RankRAG [15]. The first three baselines and ERP-SRDS used the same 8B instruction-tuned generator and retrieval depth, while RankRAG retained its native integrated ranking mechanism. Performance was evaluated using nDCG@10,

Decision Accuracy (DA), and Evidence Coverage@5 (EC@5). All results report mean ± standard deviation over five runs, with paired two-sided t-tests conducted at $\alpha = 0.05$.

4.2. Overall Performance Comparison

Table 1 compares retrieval quality and downstream decision performance across the five methods.

Table 1. Overall Performance Comparison on FiQA and FinQA (Mean ± SD, N = 5)

Method	FiQA nDCG@10 ↑	FinQA DA (%) ↑	EC@5 (%) ↑
BM25 + Generator	0.327 ± 0.011	55.6 ± 1.7	82.4 ± 1.6
DPR + Generator	0.371 ± 0.010	60.1 ± 1.6	86.7 ± 1.2
BGE-M3 + Generator	0.429 ± 0.007	64.5 ± 1.4	89.9 ± 1.0
RankRAG	0.407 ± 0.009	66.4 ± 1.5	91.3 ± 1.1
ERP-SRDS	0.421 ± 0.008	67.8 ± 1.3	92.4 ± 0.9

ERP-SRDS did not obtain the highest FiQA retrieval score. BGE-M3 achieved an nDCG@10 of 0.429, compared with 0.421 for ERP-SRDS, indicating that the additional metadata and control components do not automatically improve pure ranking performance. BGE-M3 is specifically designed to support multiple retrieval functions, including dense, sparse, and multi-vector retrieval, which provides a plausible explanation for its stronger ranking result.

In contrast, ERP-SRDS achieved the highest FinQA DA and EC@5. Compared with DPR, its gains in nDCG@10, DA, and EC@5 were significant at p=0.003, p=0.002, and p=0.001, respectively. The DA advantage over RankRAG was smaller and did not reach the p<0.01 level (p=0.028). This pattern suggests that the principal benefit of ERP-SRDS lies less in maximizing retrieval ranking alone than in preserving useful evidence during reranking and subsequent decision generation.

4.3. Ablation and Mechanism Validation

Table 2 evaluates the contribution of the four principal mechanisms.

Table 2. Ablation Results of Erp-srds (Mean ± SD, N = 5)

Configuration	nDCG@10 ↑	DA (%) ↑	EC@5 (%) ↑
Full ERP-SRDS	0.421 ± 0.008	67.8 ± 1.3	92.4 ± 0.9
w/o metadata-aware encoding	0.414 ± 0.010	65.9 ± 1.5	90.8 ± 1.1
w/o lexical fusion	0.399 ± 0.011	64.6 ± 1.8	88.9 ± 1.4
w/o evidence reranking	0.408 ± 0.009	63.8 ± 1.7	87.6 ± 1.5
w/o confidence controller	0.422 ± 0.009	66.2 ± 1.4	92.1 ± 1.0

Removing lexical fusion reduced DA by 3.2 percentage points (p=0.004), showing that exact lexical information remains useful for financial amounts, dates, and named entities. Removing evidence reranking produced the largest DA reduction, from 67.8% to 63.8% (p=0.002), while EC@5 fell by 4.8 points. By contrast, removing the confidence controller slightly increased nDCG@10 from 0.421 to 0.422, well within run-to-run variation. This is expected because confidence control operates after retrieval rather than directly optimizing document ranking. Its main effect was instead observed in decision reliability.

4.4. Convergence Analysis

Figure 2 presents the training objective and validation DA across 15 epochs, with shaded regions indicating ±SD over five runs.

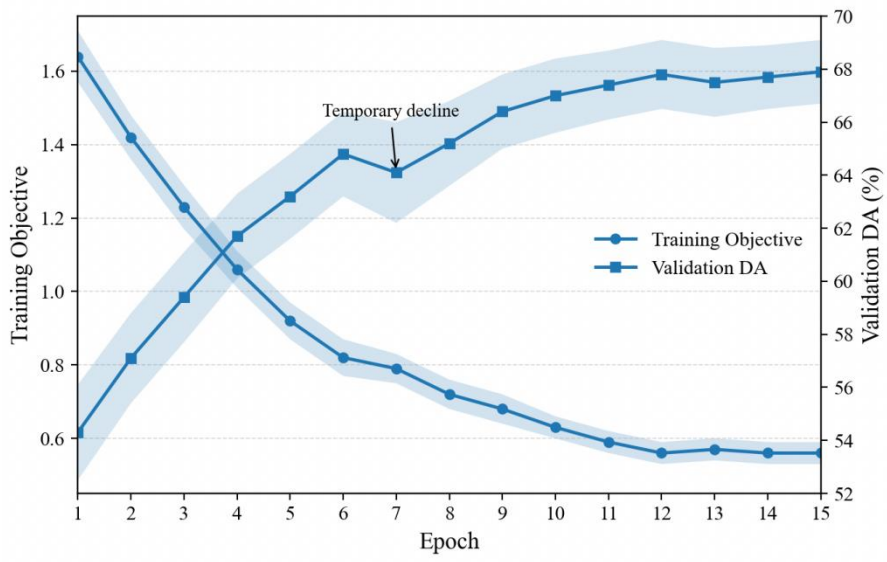

Figure 2. Training Convergence of Erp-srds over Five Runs (Mean ± SD)

The training objective declined from 1.64 ± 0.07 at epoch 1 to 0.56 ± 0.03 at epoch 12. Validation DA increased from 54.3 ± 1.8% to 64.8 ± 1.6% by epoch 6, but decreased to 64.1 ± 1.9% at epoch 7 before recovering to 66.4 ± 1.4% at epoch 9. It reached 67.8 ± 1.3% at epoch 12 and remained between 67.5% and 67.9% during epochs 13–15. The temporary deterioration at epoch 7 indicates that optimization was not monotonic. The subsequent stabilization is consistent with progressive adjustment between retrieval, reranking, generation, and evidence-grounding objectives rather than continued improvement in every component.

4.5. Robustness, Generalization, and Reliability Analysis

Robustness was evaluated by adding irrelevant passages to the retrieved context. Figure 3 reports FinQA DA under four perturbation levels.

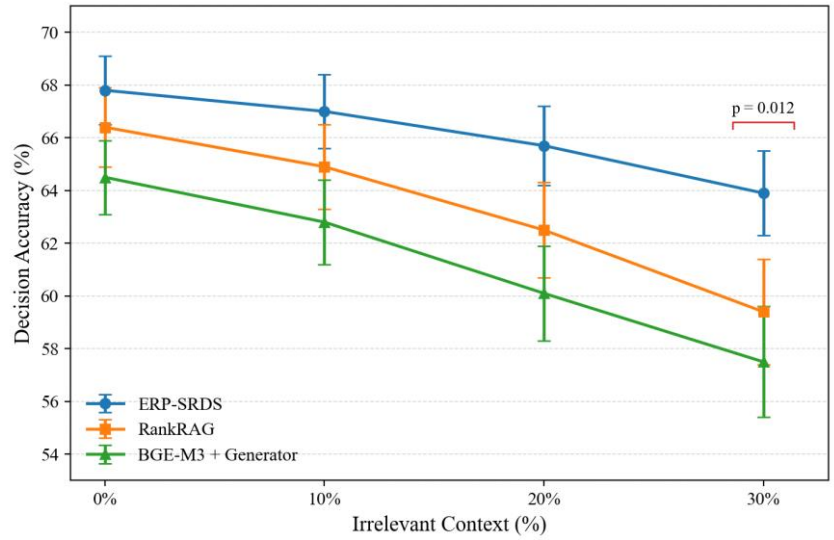

Figure 3. Decision Accuracy under Irrelevant-Context Perturbation (Mean ± SD, N = 5)

Performance declined for every method as irrelevant evidence increased. ERP-SRDS therefore cannot be regarded as insensitive to retrieval noise. Nevertheless, its reduction from 0% to 30% noise was 3.9 percentage points, compared with 7.0 points for RankRAG. At the 30% perturbation level, the difference between ERP-SRDS and RankRAG was significant at p=0.012.

Cross-dataset transfer produced a further performance decrease. When retrieval-related parameters trained on FiQA were applied to FinQA without additional retrieval fine-tuning, ERP-SRDS achieved 58.6 ± 1.9% DA, compared with 56.2 ± 2.0% for RankRAG and 54.9 ± 2.1% for BGE-M3 + Generator. The drop from the standard FinQA setting confirms that distribution shift remains a material limitation rather than being eliminated by semantic retrieval.

Interpretability was examined using the evidence trace for a FinQA case involving the change in vested shares between 2005 and 2006. The original FinQA example requires evidence from both textual fair-value information and tabulated weighted-average values and uses an executable multi-step calculation. ERP-SRDS retained the required evidence in the final evidence set before generating the calculation, illustrating how reranking can expose the evidence used for a numerical result. More generally, the framework provides source identifiers, ranked evidence, confidence-controlled output, and a human-review condition for low-confidence cases. These mechanisms improve auditability, but they should be interpreted as decision-support controls rather than guarantees of regulatory compliance.

5. Conclusion

5.1. Main Findings

The results show that the three mechanisms introduced in ERP-SRDS contribute to different aspects of ERP-oriented decision support. First, ERP-aware lexical–semantic retrieval improved retrieval effectiveness over conventional sparse and dense baselines. ERP-SRDS achieved an nDCG@10 of 0.421 ± 0.008 on FiQA, exceeding BM25 and DPR, although BGE-M3 obtained a slightly higher score of 0.429 ± 0.007. This result indicates that exact lexical information and semantic representations are complementary for business documents, while also showing that the proposed framework does not necessarily dominate specialized retrieval models in pure ranking performance.

Second, evidence reranking and grounded generation produced clearer advantages in downstream reasoning. ERP-SRDS achieved a FinQA decision accuracy of 67.8 ± 1.3% and Evidence Coverage@5 of 92.4 ± 0.9%. Removing evidence reranking reduced decision accuracy to 63.8 ± 1.7%, representing the largest decline in the ablation study. The results suggest that retrieving additional relevant text alone is insufficient for reliable decision support; candidate evidence must also be filtered and prioritized according to its contribution to the requested calculation or interpretation.

Third, confidence-based control primarily improved robustness rather than retrieval accuracy. Under 30% irrelevant-context injection, ERP-SRDS decision accuracy decreased to 63.9 ± 1.6%, confirming that the framework remains affected by noisy evidence. However, its 3.9-percentage-point decline was smaller than the 7.0-point reduction observed for RankRAG. Confidence control is therefore most useful for identifying evidence-deficient cases and limiting unsupported outputs rather than directly maximizing nominal accuracy. The practical value of generative AI in ERP systems lies not in replacing managerial judgment, but in improving access to dispersed business evidence and providing more traceable support for routine analytical decisions.

5.2. Limitations

Several limitations remain. FiQA and FinQA represent financial retrieval and reasoning tasks rather than complete ERP workflows, and financial reports differ from operational procurement, inventory, production, and supply-chain records. The metadata structure used in ERP-SRDS is an ERP-oriented representation rather than a schema extracted from a deployed ERP database. Generation errors may also persist when retrieval or reranking provides incomplete evidence. In addition, the appropriate confidence threshold depends on business context and risk tolerance. The reported experiments therefore do not establish suitability for autonomous high-risk business decisions.

5.3. Future Research

Future research should evaluate the framework on enterprise-specific ERP datasets covering additional functional modules and document types. Further work could incorporate multimodal invoices and scanned documents, access-control-aware retrieval, and temporally updated ERP knowledge. Human-in-the-loop evaluation should also examine whether evidence traces and confidence indicators improve managerial verification and decision quality in realistic organizational settings.

References

1. M. A. Mannan et al., "Transforming ERP systems with collaborative AI: Paving the path to strategic growth and sustainability," *Array*, p. 100517, 2025.
2. L. Silva and L. Barbosa, "Improving dense retrieval models with LLM augmented data for dataset search," *Knowledge-Based Systems*, vol. 294, p. 111740, 2024.
3. S. Pan et al., "Unifying large language models and knowledge graphs: A roadmap," *IEEE Transactions on Knowledge and Data Engineering*, vol. 36, no. 7, pp. 3580–3599, 2024.
4. S. Wang, H. Yang, and G. Bai, "Construction of intelligent decision support systems through integration of retrieval-augmented generation and knowledge graphs," *Scientific Reports*, vol. 15, no. 1, p. 35462, 2025.
5. J. Qu et al., "PNCD: Mitigating LLM hallucinations in noisy environments–A medical case study," *Information Fusion*, vol. 123, p. 103328, 2025.
6. S. Feuerriegel et al., "Generative AI," *Business & Information Systems Engineering*, vol. 66, no. 1, pp. 111–126, 2024.
7. T. Breuer et al., "Large language models for information retrieval: Challenges and chances," *Datenbank-Spektrum*, vol. 25, no. 2, pp. 71–81, 2025.
8. S. Frihat and N. Fuhr, "Integration of biomedical concepts for enhanced medical literature retrieval," *International Journal of Data Science and Analytics*, vol. 20, no. 5, pp. 4409–4422, 2025.
9. M. Hindi et al., "Enhancing the precision and interpretability of retrieval-augmented generation (RAG) in legal technology: A survey," *IEEE Access*, vol. 13, pp. 46171–46189, 2025.
10. S. Xu et al., "MEGA-RAG: A retrieval-augmented generation framework with multi-evidence guided answer refinement for mitigating hallucinations of LLMs in public health," *Frontiers in Public Health*, vol. 13, p. 1635381, 2025.
11. L. Huang et al., "A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions," *ACM Transactions on Information Systems*, vol. 43, no. 2, pp. 1–55, 2025.
12. S. Robertson and H. Zaragoza, "The probabilistic relevance framework: BM25 and beyond," *Foundations and Trends® in Information Retrieval*, vol. 4, no. 1–2, pp. 1–174, 2009.
13. V. Karpukhin et al., "Dense passage retrieval for open-domain question answering," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Nov. 2020, pp. 6769–6781.
14. J. Chen et al., "M3-Embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation," in *Findings of the Association for Computational Linguistics: ACL 2024*, 2024, pp. 2318–2335.
15. Y. Yu et al., "RankRAG: Unifying context ranking with retrieval-augmented generation in LLMs," in *Advances in Neural Information Processing Systems*, vol. 37, 2024, pp. 121156–121184.