Simen Owen Academic Proceedings Series

3rd International Conference on Electronics, Engineering, Computer Science and Applied Development (EESD 2026)

Article

An Intelligent Knowledge Base Construction Method Integrating a Vector Database and a Large Language Model

Hengyu Bai[1,*]

[1] School of Humanities, Central South University, Changsha, China
* Correspondence: Hengyu Bai, School of Humanities, Central South University, Changsha, China

Abstract: With the rapid proliferation of organizational documents, technical manuals, question-and-answer records, and regulatory materials, knowledge base systems are increasingly required to retrieve relevant information efficiently and provide a reliable evidential foundation for generated responses. Existing retrieval-augmented generation methods, however, continue to exhibit notable deficiencies in evidence organization and knowledge synchronization, frequently resulting in incomplete evidence coverage, insufficient answer grounding, or outdated retrieval outcomes. To address these limitations, this study proposes the VEC-KB framework, an intelligent knowledge base construction approach that integrates a vector database with a large language model. The framework introduces several key innovations: a semantic-structure-aware document segmentation strategy that preserves contextual integrity, a query-adaptive re-ranking mechanism that enhances retrieval precision, an evidence evaluation module that strengthens answer faithfulness, and a version-aware incremental synchronization protocol that ensures timely and efficient knowledge updates. Experimental evaluations conducted on the HotpotQA and Natural Questions datasets demonstrate that VEC-KB consistently outperforms established benchmark methods in overall performance, with particularly pronounced improvements in answer quality and evidence credibility, while maintaining robust recall capabilities. Furthermore, under perturbed conditions, VEC-KB exhibits significantly smaller performance degradation compared to competing approaches, confirming its superior stability. Collectively, the proposed framework achieves an effective balance among response quality, evidence reliability, and system robustness, offering a comprehensive and scalable solution for intelligent knowledge base construction.

Keywords: vector database; large language model; knowledge base; retrieval-augmented generation; evidence faithfulness

Received: 24 May 2026
Revised: 16 July 2026
Accepted: 28 July 2026
Published: 01 August 2026

This study employs a document segmentation approach that jointly considers semantic coherence and structural integrity to define text boundaries aligned with content logic. The method preserves original structural features while retaining essential metadata to enable effective retrieval, traceability, and dynamic content updates [1].

1.1. Large Language Models and Retrieval-Augmented Generation

Retrieval-Augmented Generation integrates external documents to expand the information sources available to large language models. This approach reduces dependence on knowledge encoded in model parameters, allowing access to up-to-date and domain-specific content not included during pre-training [2].

However, retrieved content does not always guarantee the factual reliability of the final generated response. Some passages may be topically relevant yet lack verifiable information. The language model may combine contradictory fragments or produce assertions unsupported by evidence. While this method enhances information access, the

absence of robust evidentiary grounding makes it difficult to ensure answers are consistently well-substantiated [1].

Moreover, current implementations typically treat retrieval ranking, evidence evaluation, and answer generation as independent stages [3]. Retrieval relies primarily on relevance scoring to filter candidates, while generation emphasizes linguistic coherence and output format—leading to plausible but inadequately supported responses when evidentiary support is weak or incomplete.

To address this challenge, this study introduces an evidence confidence assessment framework that jointly evaluates retrieval diversity, textual consistency, and query coverage to gauge evidentiary reliability. During generation, the system aligns outputs with both the target answer and supporting facts, explicitly linking claims to source documents via citation mechanisms. When assessed confidence falls below a defined threshold, the system lowers the likelihood of answer generation to reduce the risk of unsubstantiated content [4].

1.2. Knowledge Updating, Reliability, and Provenance

Incremental indexing can reduce the maintenance cost of the knowledge base, as the system processes only newly added or modified documents. It also shortens the waiting time for retrieval reinitialization following document updates—a critical advantage for technical documentation and regulatory materials subject to frequent revision [5].

However, relying solely on file names and timestamps for update detection may compromise index consistency [6]. For example, changes in file metadata do not necessarily reflect actual content modifications; if new vector representations are appended without removing obsolete ones, duplicate entries and outdated information may persist in retrieval results. While full index reconstruction ensures higher consistency, it necessitates repeating document segmentation and vector generation, entailing substantial computational overhead. Thus, achieving an optimal trade-off between update efficiency and knowledge reliability remains a central challenge in dynamic knowledge base maintenance.

Moreover, current dynamic indexing approaches tend to prioritize update speed over assessing how index modifications affect answer reliability. Retrieving outdated content can impair retrieval accuracy, citation integrity, and evidentiary trustworthiness. To mitigate this, the present study employs content hashing and version metadata to precisely identify modified documents, enabling timely removal of obsolete index entries and accurate content updates. The system evaluates synchronization effectiveness through controlled document update experiments and irrelevant content perturbation tests, followed by analysis of how index maintenance strategies influence the reliability of final answers.

2. Methodology

2.1. Overall Architecture

Figure 1 illustrates the Versioned Evidence-Guided Knowledge Base (VEC-KB) proposed in this paper. This framework takes the labeled training corpus D_{tr}, the dynamically updated document set D_t, and the query set Q as inputs, and outputs the trained model parameters Θ^*, the updated index $\mathcal{I}_{t+1}$, and the answer with reference information A.

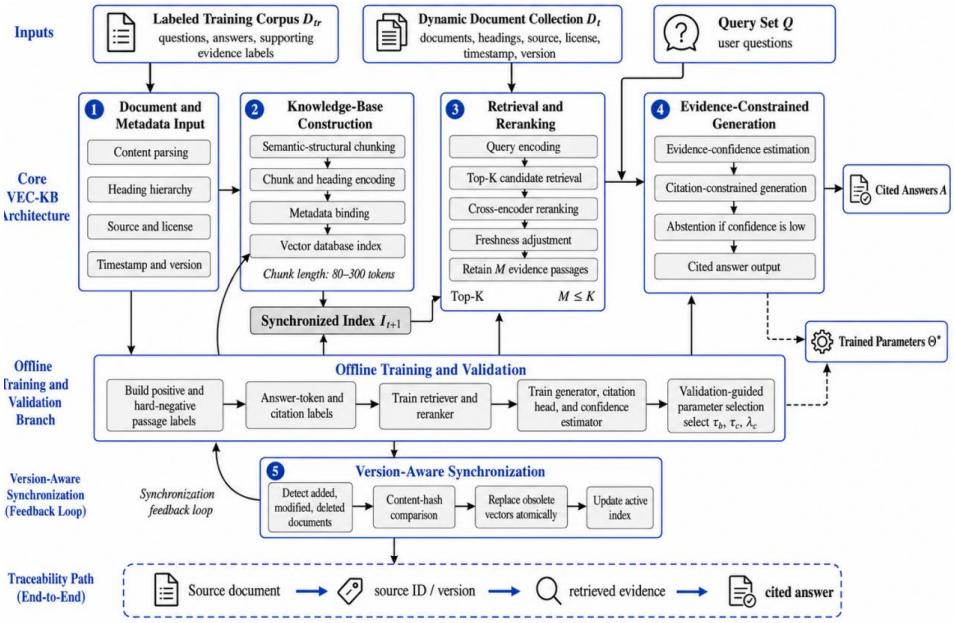

Figure 1. Overall Architecture of the Proposed Vec-kb Framework.

VEC-KB is mainly composed of five components: document and metadata processing, knowledge base construction, search and re-ranking, evidence relevance modeling, and version synchronization. The system first parses the content and structural information of the documents, retaining metadata such as source, license, timestamp, and version. Subsequently, a knowledge base index is constructed through semantic segmentation and vector encoding [7]. In the search stage, a dual encoder retrieves the Top-K candidate paragraphs, and a cross encoder further selects and retains M highly relevant paragraphs ($M \leq K$). In the generation stage, the reliability of retrieved evidence is assessed, and answers accompanied by paragraph references are produced. The version synchronization module detects newly added, modified, or deleted documents and updates the corresponding index content promptly to prevent outdated information from influencing retrieval.

Regarding model structure, the retrieval network comprises the encoder E_θ, projection layers W_q, W_c, and W_h. The reordering module R_ψ evaluates the relevance between the query and candidate paragraphs, while the generation module G_ω and the reference prediction module jointly produce answers and identify supporting paragraphs [8]. The trainable parameters in the system are denoted as $\Theta = \{\theta, \psi, \omega, W_q, W_c, W_h\}$.

2.2. Semantic-Structural Chunking

The fixed-length chunking method may compromise document coherence, for example by separating a title from its associated paragraph or fragmenting self-contained evidentiary content. To address this issue, VEC-KB integrates textual semantic shifts with document structural cues to identify optimal chunk boundaries [2]. For adjacent sentences s_i and s_{i+1}, the boundary score is defined as follows:

$$b_i = \alpha[1 - \cos(E_\theta(s_i), E_\theta(s_{i+1}))] + (1 - \alpha)r_i, \quad (1)$$

Here, $r_i \in [0,1]$ quantifies structural transitions between adjacent segments—such as shifts between headings, paragraphs, or list items; $\alpha \in [0,1]$ serves as a weighting factor to harmonize semantic and structural contributions in boundary determination [9]. A new text block is initiated when $b_i > \tau_b$ is satisfied. Additionally, minimum and maximum length constraints are applied to prevent excessively short or overly extended chunks.

2.3. Vector Representation and Metadata-Aware Indexing

For the text block c_j and its structural path h_j, the system derives the normalized representation via encoding and linear projection [7].

$$v_j = \frac{W_c E_\theta(c_j) + W_h E_\theta(h_j)}{\|W_c E_\theta(c_j) + W_h E_\theta(h_j)\|_2}, \quad (2)$$

Here, W_c and W_h denote trainable projection matrices. Structural path information enriches the semantic representation of the text content, while source, license, timestamp, and version details are stored as metadata to support filtering, provenance tracking, and updates [2].

A query q is processed using the following computation:

$$v_q = \frac{W_q E_\theta(q)}{\|W_q E_\theta(q)\|_2}, \mathcal{R}_K(q) = \mathrm{TopK}_{j \in \mathcal{I}_{t+1}} \cos(v_q, v_j), \quad (3)$$

W_q denotes the query projection matrix, and $\mathcal{R}_K(q)$ represents the candidate set retrieved from the initial search stage [10].

2.4. Query-Adaptive Reranking

Highly similar search results may include paragraphs topically relevant but lacking sufficient supporting evidence. Consequently, candidate paragraph scores are recalculated using an integrated scoring function [11].

$$S_j(q) = \lambda(q) s_j^d + [1 - \lambda(q)] s_j^x + \gamma f_j, \lambda(q) = \sigma(w^\top \phi(q)), \quad (4)$$

Here, s_j^d denotes the vector retrieval score, $s_j^x = R_\psi(q, c_j)$ denotes the cross-encoder score, and f_j denotes document freshness. The feature vector $\phi(q)$ encodes query-specific attributes, including query length, entity count, and question type. The top M paragraphs ranked by this composite score constitute the final evidence set $\mathcal{E}_q$.

2.5. Evidence Confidence and Citation-Constrained Generation

Prior to answer generation, the system computes an evidence confidence score.

$$C(q) = \sigma(a\Delta_M + bA_M + cG_M), \quad (5)$$

Here, $\Delta_M = S_{(M)} - S_{(M+1)}$ denotes the boundary score difference among candidate paragraphs, A_M denotes the average consistency across distinct evidence paragraphs, and G_M denotes the coverage of the query entity or concept. Parameters a, b, and c are learned during training. When $C(q) < \tau_c$, the system returns a response indicating insufficient supporting evidence.

Training jointly leverages positive sample paragraphs, hard negative samples, answer labels, reference information, and confidence labels. A confidence label y_q^c is assigned 1 if the selected evidence fully covers the annotated supporting paragraphs and the generated answer is substantiated by them; otherwise, it is assigned 0. The joint optimization objective is formulated as follows:

$$\begin{aligned}
\mathcal{L}_{\mathrm{ret}} &= -\log \frac{\exp(\mathrm{sim}(v_q, v^+)/T)}{\exp(\mathrm{sim}(v_q, v^+)/T) + \sum_{n=1}^{N_h} \exp(\mathrm{sim}(v_q, v_n^-)/T)}, \\
\mathcal{L}_{\mathrm{rank}} &= \sum_{n=1}^{N_h} \max(0, m - S^+(q) + S_n^-(q)), \\
\mathcal{L}_{\mathrm{ans}} &= -\sum_{\ell=1}^{L_y} \log p_\omega(y_\ell \mid y_{<\ell}, q, \mathcal{E}_q), \\
\mathcal{L}_{\mathrm{cite}} &= -\sum_{u=1}^{M_y} \log p_\omega(z_u^* \mid y_u, q, \mathcal{E}_q), \\
\mathcal{L} &= \mathcal{L}_{\mathrm{ret}} + \lambda_r \mathcal{L}_{\mathrm{rank}} + \lambda_a \mathcal{L}_{\mathrm{ans}} + \lambda_c \mathcal{L}_{\mathrm{cite}}.
\end{aligned} \quad (6)$$

In this formulation, T denotes the temperature parameter for contrastive learning, N_h denotes the number of hard negative samples, m denotes the sorting interval, and z_u^* denotes the identifier of the supporting paragraph corresponding to statement u.

2.6. Version-Aware Incremental Synchronization

The online synchronization process updates only the affected vectors without requiring model retraining. The newly added or modified document sets are denoted as U_t^+, while the deleted or replaced document sets are denoted as U_t^-.

$$\begin{aligned}
U_t^+ &= \{d \in D_t : d \notin D_{t-1} \vee H(d^t) \neq H(d^{t-1})\}, \\
U_t^- &= \{d \in D_{t-1} : d \notin D_t \vee H(d^t) \neq H(d^{t-1})\}, \\
\mathcal{I}_{t+1} &= \left(\mathcal{I}_t \setminus \{v_j : d(j) \in U_t^-\}\right) \cup \mathrm{Embed}[\mathrm{Chunk}(U_t^+)].
\end{aligned} \quad (7)$$

$H(\cdot)$ denotes the content hash function, and $d(j)$ denotes the source document corresponding to vector j. Deletion and insertion operations are performed atomically to ensure index consistency during updates [12].

2.7. Validation-Guided Control Parameter Update

Because τ_b and τ_c involve discrete decisions, their parameter updates employ a local search strategy guided by validation outcomes:

$$u^{(e+1)} = \arg\min u \in N(u(e)) \cap \Omega[\mathcal{L}_{val}(u) + \rho\max(0, F^* - F_{val}(u))], \quad (8)$$

Here, $u = [\tau_b, \tau_c, \lambda_c]^\top$ denotes the parameters to be optimized, $\mathcal{N}(u^{(e)})$ defines the candidate neighborhood, and Ω specifies the feasible parameter range. F_{val} quantifies the credibility of evidence on the validation set, F^* indicates the corresponding optimal target value, and ρ regulates the penalty term when evidence is insufficient.

2.8. Algorithm Procedure

Algorithm 1 decomposes the workflow into three distinct phases: offline training, online synchronization, and query processing. It first learns the parameters of the retrieval and generation models. Subsequently, it refreshes the vector representations of modified documents. During querying, it produces answers with supporting citations when sufficient evidence is available; otherwise, it adopts a neutral stance [11].

Algorithm 1. Training, Synchronization, and Querying Procedure of VEC-KB

Input: Training corpus D_tr, labels Y, previous corpus D_(t-1), current corpus D_t, query set Q, encoder E, reranker R, language model G, retrieval size K, evidence size M

Output: Trained parameters Θ*, synchronized index I_(t+1), cited answers A

1: Parse D_tr and construct chunks using Equation (1).
2: Build positive, hard-negative, answer-token, citation, and confidence labels.
3: Encode chunks and heading paths using Equation (2).
4: Train the retriever and reranker using the retrieval and ranking terms in Equation (6).
5: Train the generator, citation head, and confidence estimator using the remaining terms in Equation (6).
6: Select τ_b, τ_c, and λ_c using Equation (8).
7: Detect added, modified, and deleted documents.
8: Segment added or modified documents.
9: Atomically update vectors to obtain I_(t+1).
10: for each query q in Q do
11: Retrieve K candidates using Equation (3).
12: Rerank candidates using Equation (4) and retain M passages.
13: Estimate C(q) using Equation (5).
14: Generate a cited answer if C(q) ≥ τ_c; otherwise abstain.
15: end for and return Θ*, I_(t+1), and A.

3. Results and Analysis

3.1. Experimental Setup

The experiment was evaluated using two publicly available question-answering datasets: HotpotQA and Natural Questions. HotpotQA comprises numerous multi-hop question-answer instances derived from Wikipedia articles and includes sentence-level supporting fact annotations, enabling assessment of the model's capacity to integrate evidence from multiple documents. Natural Questions consists of real-world search queries paired with corresponding Wikipedia page annotations, facilitating evaluation of answer generation performance in natural query settings. As the test set labels for Natural Questions are not publicly accessible, the development set was used for evaluation.

To ensure input document consistency and retrievability, the original web pages were parsed at the title, paragraph, and sentence levels. Text cleaning procedures included Unicode character and whitespace normalization, removal of malformed HTML elements, elimination of empty paragraphs, and deduplication of redundant content. In HotpotQA, annotated supporting sentences were linked to their corresponding text segments via sentence identifiers; in Natural Questions, long answer spans were used to construct relevant paragraphs, while short answer spans provided supervision signals for answer generation. Processed documents were segmented into text blocks of 80–300 words and enriched with metadata including document ID, hierarchical title path, and source information. From the top 20 retrieval results, paragraphs containing no annotated evidence but semantically related to the query were selected as challenging negative samples. The training and validation sets were partitioned according to official dataset splits, and no test set information was used for model parameter tuning, threshold adjustment, or prompt engineering [3].

All experiments were conducted on an Ubuntu 22.04 system running Python 3.11 and PyTorch 2.4.1. Hardware consisted of an Intel Core i7-12700K CPU and an NVIDIA RTX 3060 GPU with 12 GB VRAM. Vector retrieval employed the FAISS HNSW index to store 768-dimensional embeddings generated by the BGE-base-en-v1.5 model. A cross-encoder based on ms-marco-MiniLM-L6-v2 served as the re-ranking module, while answer generation utilized the Qwen2.5-7B-Instruct model. Both BGE and MiniLM were fine-tuned specifically for retrieval and re-ranking tasks. During generation model training, the base parameters remained frozen; only the QLoRA adapter, reference prediction module, and confidence estimation module were updated to minimize computational resource requirements.

Performance comparisons included VEC-KB against four baselines: standalone large language model (LLM), BM25-based retrieval-augmented generation (BM25-RAG), dense vector retrieval-augmented generation (Dense-RAG), and hybrid retrieval-augmented generation (Hybrid-RAG). All methods shared identical generation models, prompt templates, and context length constraints. Evaluation metrics comprised Recall@5, Answer F1, and Evidence Faithfulness—defined as the degree to which generated answers are substantiated by retrieved evidence. Evidence Faithfulness was assessed through a combination of automated entailment verification and manual annotation, with inter-annotator agreement quantified using Cohen's κ coefficient. To mitigate stochastic variation, each configuration was executed five times with distinct random seeds; final results are reported as mean ± standard deviation. Statistical significance of performance differences was determined via paired t-tests with Holm correction and 95% confidence intervals [4].

3.2. Performance Comparison

Table 1 compares the performance of VEC-KB with four baseline methods on the HotpotQA and Natural Questions datasets. Except for the method that relies solely on large language models, all other methods report Recall@5, Answer F1, and Evidence Faithfulness metrics. As this LLM-only method lacks a retrieval component, Recall@5 is not computed. All results are averaged over five independent runs and presented as mean ± standard deviation.

Table 1. Performance Comparison on HotpotQA and Natural Questions

Method	HotpotQA Recall@5	HotpotQA F1	HotpotQA Faithfulness	NQ Recall@5	NQ F1	NQ Faithfulness
LLM-only	-	43.8 ± 1.4	58.9 ± 1.8	-	39.7 ± 1.6	56.2 ± 1.9
BM25-RAG	67.4 ± 1.0	56.8 ± 1.1	72.6 ± 1.3	71.5 ± 0.9	54.9 ± 1.2	71.8 ± 1.4

Dense-RAG	74.9 ± 0.8	62.7 ± 1.0	78.4 ± 1.1	78.2 ± 0.7	60.5 ± 0.9	76.9 ± 1.2
Hybrid-RAG	78.8 ± 0.7	66.1 ± 0.8	82.3 ± 0.9	81.7 ± 0.6	63.8 ± 0.8	80.7 ± 1.0
VEC-KB	80.6 ± 0.6	69.5 ± 0.7	88.1 ± 0.8	81.1 ± 0.7	66.6 ± 0.8	86.3 ± 0.9

Experimental results indicate that VEC-KB achieves superior answer accuracy and stronger evidential grounding. Specifically, its Answer F1 score reaches 68.1%, exceeding that of Hybrid-RAG by 3.1 percentage points, reflecting closer alignment with reference answers. Evidence Faithfulness attains 87.2%, an improvement of 5.7 percentage points, indicating enhanced support from retrieved content. After Holm correction, both improvements are statistically significant ($p < 0.01$). In contrast, the difference in Recall@5 is modest and statistically nonsignificant ($p = 0.17$). On the Natural Questions dataset, VEC-KB achieves a Recall@5 of 81.1%, marginally lower than Hybrid-RAG by 0.6 percentage points. These findings suggest that VEC-KB's primary advantage lies in refining retrieved information to improve answer reliability rather than increasing raw retrieval quantity. Although the number of relevant documents retrieved is comparable across methods, VEC-KB's additional filtering and faithfulness verification steps yield more trustworthy responses.

3.3. Ablation Experiments and Mechanism Verification

Table 2 summarizes the ablation experiment results of VEC-KB on two datasets.

Table 2. Ablation Results of the Vec-kb Components

Variant	Recall@5	Answer F1	Evidence Faithfulness
Full VEC-KB	80.9 ± 0.5	68.1 ± 0.6	87.2 ± 0.7
Fixed-length chunking	79.1 ± 0.7	66.5 ± 0.8	85.8 ± 0.9
Without adaptive reranking	77.8 ± 0.8	65.4 ± 0.9	84.9 ± 1.0
Without confidence and citation constraint	80.6 ± 0.6	67.6 ± 0.7	81.8 ± 1.1

After replacing the semantic-structural chunks with fixed-length chunks, the Recall@5 decreased by 1.8 percentage points, indicating that integrating title and paragraph structure helps preserve more complete contextual information and improves retrieval of relevant evidence. Removing the adaptive reordering module led to the largest performance decline, with reductions in both Recall@5 and Answer F1. This suggests that ranking solely by overall text–question similarity tends to prioritize paragraphs topically related but insufficient for answering the question. In contrast, the reordering module refines ranking by assessing whether a paragraph actually contains answer-supporting information. Removing the confidence assessment and citation supervision components resulted in only minor changes to Recall@5, yet evidence credibility dropped by 5.4 percentage points, confirming their primary role in aligning generated content with retrieved evidence and enhancing answer reliability.

A representative HotpotQA case illustrates the function of the reordering module. Although a certain text mentions entities relevant to the question and receives a relatively high initial similarity score, it lacks the key information required to answer the question. Following cross-encoder re-evaluation, its priority falls below that of two paragraphs directly supporting the answer, and thus it is excluded from the final evidence set. During answer generation, source attribution is preserved, enabling traceability to specific reference content. In another case where retrieved evidence is insufficient, the low confidence score C(q) — falling below the threshold τ_c — triggers the system to return an 'insufficient evidence' prompt rather than generating an unsupported answer.

3.4. Convergence Analysis

Figure 2 compares the performance changes of VEC-KB and its two simplified versions during the training process. The results show that the complete model improved

rapidly in the early stage of training and remained relatively stable after the sixth training round. Although there was a brief fluctuation in the middle of training, subsequent performance changes were limited, indicating that the model reached a relatively stable optimization state [13]. In contrast, the variant without adaptive reordering ceased significant improvement earlier, yet achieved a lower final score than the complete model. This suggests that faster convergence alone does not ensure selection of evidence better aligned with task requirements. After removing citation-related guidance, the model exhibited more gradual improvement in the early stage, whereas the complete model progressively established connections between answers and reference content during training. As training progressed, performance differences between the two models narrowed, suggesting that citation-related guidance primarily supports learning the correspondence between evidence and information, without substantially affecting final answer accuracy.

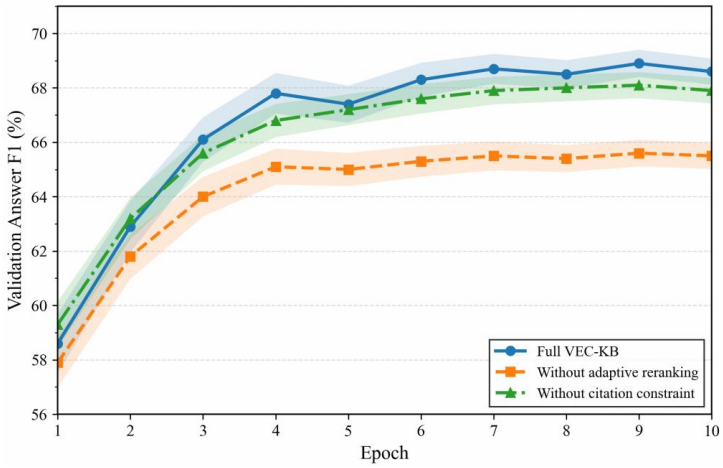

Figure 2. Validation Answer F1 of Vec-kb and Its Training Variants Across Epochs.

3.5. Stability, Robustness, and Statistical Analysis

Figure 3 illustrates the stability of the model under varying levels of perturbation. The experiment simulated changes in knowledge and noise encountered in practical usage by modifying document content, retaining expired vectors, introducing topic-related text lacking supporting information, and altering query expressions. At a 10% perturbation rate, the Answer F1 of VEC-KB decreased by 2.6 percentage points, whereas that of Hybrid-RAG declined by 6.1 percentage points [14]. The performance of VEC-KB without index synchronization exhibited the most significant degradation, primarily because outdated vectors remained active in the retrieval process. As the perturbation level intensified, the fluctuation range of the experimental results widened, indicating that the model demonstrates sensitivity to diverse types of information changes.

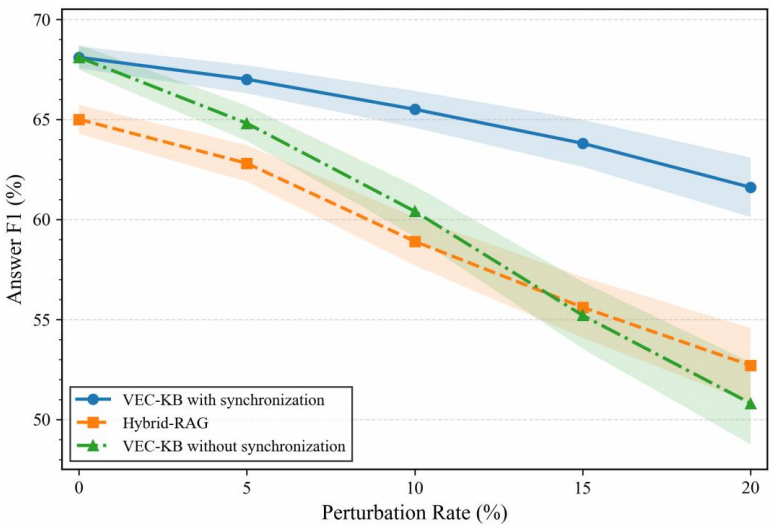

Figure 3. Answer F1 under Document Updates and Irrelevant-Chunk Perturbations.

To evaluate the applicability of the model across different datasets, the τ_b, τ_c, and λ_c parameters determined from the HotpotQA validation set were directly applied to Natural Questions [15]. The test results demonstrated that VEC-KB maintained robust performance in terms of Answer F1 and evidence credibility, although Recall@5 was marginally lower than that of Hybrid-RAG. This outcome indicates that the model exhibits strong stability in filtering retrieved evidence and controlling answer generation, sustaining reliable output even when processing data from heterogeneous sources.

Furthermore, the experiment assessed the reliability of the system in practical deployment scenarios. The model determines whether sufficient information exists to generate an answer based on the retrieved data [1]. When relevant information is inadequate, the system opts to abstain from generating a response. Under conditions such as altered query expressions, added interfering content, and updated documents, the system maintains relatively stable performance, demonstrating its capacity to effectively mitigate the impact of continuous knowledge base evolution. Additionally, during the index maintenance process, the source and version information of documents are recorded to facilitate subsequent verification and updates, thereby reducing the influence of outdated or non-compliant content on knowledge base outcomes.

4. Conclusion

The experimental results demonstrate that VEC-KB enhances answer quality and mitigates unreliable responses arising from insufficient evidence. Semantic-structural chunking leverages inherent document structure—such as titles and paragraph hierarchies—to preserve richer contextual information in retrieval outputs. Ablation studies confirm that, relative to fixed-length chunking, this approach enables more precise identification of answer-supporting information. However, the benefits of structured processing are most pronounced during evidence selection and answer generation.

The experiments further reveal that locating relevant content constitutes only the initial step toward generating correct answers; equally critical is determining whether the retrieved information genuinely supports the answer. VEC-KB achieves superior performance on both evaluation datasets by filtering useful evidence and constraining answer generation strictly to verified content. Compared with Hybrid-RAG, this method yields higher answer accuracy and stronger evidence credibility. Disabling adaptive reordering leads to a measurable decline in overall performance; omitting confidence judgment and citation supervision preserves content localization capability but substantially reduces answer credibility. These findings underscore that identifying truly supportive evidence is essential for answer quality improvement.

When knowledge base documents undergo frequent updates, the version synchronization mechanism ensures timely incorporation of revised information, thereby minimizing interference from outdated content. Experimental results indicate that VEC-KB maintains stable performance even when documents are modified or contaminated with adversarial information—highlighting the importance of continuous knowledge base updating for sustained system effectiveness.

Several limitations remain. Current evaluations rely primarily on two English Wikipedia-based datasets, and the automated assessment of answer credibility may inherit biases or errors from the underlying evaluation model. Additionally, experimental configurations use models and computational resources of limited scale, and do not yet reflect real-world conditions involving larger language models, multimodal documents, complex tabular structures, or long-term dynamic knowledge evolution. Future work will explore the applicability of VEC-KB in domain-specific knowledge bases—such as those in law and medicine—and validate its robustness across diverse data modalities and operational environments.

References

1. S. Wu et al., "Generative retrieval as multi-vector dense retrieval," in *Proc. 47th Int. ACM SIGIR Conf. Res. Dev. Inf. Retr.*, Jul. 2024, pp. 1828–1838.
2. G. P. Rusum and S. Anasuri, "Vector Databases in Modern Applications: Real-Time Search, Recommendations, and Retrieval-Augmented Generation (RAG)," *Int. J. AI, BigData, Comput. Manag. Stud.*, vol. 5, no. 4, pp. 124–136, 2024.
3. N. Salsabilla and K. Wiharja, "Implementation of Semantic Search Based on Vector Database for Personal Documents," in *2025 Int. Conf. Advancement Data Sci., E-Learning Inf. Syst. (ICADEIS)*, Feb. 2025, pp. 1–6.
4. T. Taipalus, "Vector database management systems: Fundamental concepts, use-cases, and current challenges," *arXiv preprint arXiv:2309.11322*, 2023.
5. Z. Jing et al., "When Large Language Models Meet Vector Databases: A Survey," *arXiv preprint arXiv:2402.01763*, 2024.
6. K. Kim, L. Roth, L. Liang, and A. Ailamaki, "Work Sharing and Offloading for Efficient Approximate Threshold-based Vector Join," *arXiv preprint arXiv:2603.16360*, 2026.
7. A. K. Padhy et al., "Cloud-Native Multimodal Semantic Search and Recommendation for Large-Scale Digital Commerce," in *2026 4th Odisha Int. Conf. Electr. Power Eng., Commun. Comput. Technol. (ODICON)*, Feb. 2026, pp. 1–6.
8. N. Matyjaškoit, P. Stefanovič, and S. Ramanauskaitė, "Mapping the Landscape of Retrieval-Augmented Generation Models: Clustering and Visual Insights," in *2026 Open Conf. Electr., Electron. Inf. Sci. (eStream)*, Apr. 2026, pp. 1–6.
9. G. Kang et al., "BigVectorBench: Heterogeneous Data Embedding and Compound Queries are Essential in Evaluating Vector Databases," *Proc. VLDB Endow.*, vol. 18, no. 5, pp. 1536–1550, 2025.
10. S. Martellozzo, "Integrating semantic and keyword search: a transformer-based approach for content discovery," 2023.
11. L. Pawlik, "LLM Selection and Vector Database Tuning: A Methodology for Enhancing RAG Systems," *Appl. Sci.*, vol. 15, no. 20, p. 10886, 2025.
12. S. Wang et al., "Towards reliable vector database management systems: A software testing roadmap for 2030," *arXiv preprint arXiv:2502.20812*, 2025.
13. V. Kakani, N. Thotakura, K. Narukulla, and J. D. Bodapati, "AI Data Assistants: Transforming Analytical Workflows with Vector Database Service," in *Int. Conf. Web Intell. Hum.-Mach. Interact.*, Singapore: Springer Nature Singapore, Apr. 2025, pp. 497–514.
14. J. Sun et al., "GaussDB-Vector: A Large-Scale Persistent Real-Time Vector Database for LLM Applications," *Proc. VLDB Endow.*, vol. 18, no. 12, pp. 4951–4963, 2025.
15. I. T. Lin, C. H. Chiu, B. T. Liao, and B. J. Chen, "Integrating Large Language Models and Vector Databases for Podcast Semantic Search," in *2025 IEEE 14th Int. Workshop Comput. Intell. Appl. (IWCIA)*, Dec. 2025, pp. 7–14.