

Implementation of FastText- and Hybrid Search-Based Retrieval-Augmented Generation for PDF Document Processing: A Case Study of a Booking Engine Chatbot System

Dewa Gede Indra Putra¹, Eddy Muntina Dharma², Ida Bagus Kresna Sudiarmika³

Informatics Study Program, Faculty of Information Technology and Design, Universitas Primakara, Bali

This study proposes and evaluates a Retrieval-Augmented Generation (RAG) system integrating ChatGPT with FastText word embeddings for processing Indonesian-language PDF documents. The limitations of Large Language Models (LLMs) in handling domain-specific queries motivated the development of an RAG architecture with an advanced retrieval strategy. The CRISP-DM methodology was applied to compare two models: (1) a Baseline Model using pure FastText cosine similarity search, and (2) a Proposed Model combining Hybrid Search (FastText + BM25 with dynamic weighting), Maximal Marginal Relevance (MMR), and Cross-Encoder TF-IDF re-ranking. Evaluation was conducted using the RAGAS framework with five metrics—Context Relevance, Context Precision, Context Recall, Faithfulness, and Answer Relevance—across 20 test queries on the PT Omni Hottelier Booking Engine user guide. Results show that the Proposed Model outperformed the Baseline on 16 of 20 queries (80%), with the Overall Score increasing from 0.7853 to 0.8572 (+9.16%). The most significant improvements were observed in Context Relevance (+19.85%), Context Recall (+16.76%), and Faithfulness (+14.76%). The developed architecture is proven effective as an adaptive solution for transforming technical documents into intelligent chatbot knowledge bases, particularly for the morphologically complex Indonesian language.

Keywords: Retrieval-Augmented Generation, FastText, Hybrid Search, Maximal Marginal Relevance, RAGAS, Chatbot, Natural Language Processing.

This is an open access article under the [CC BY-NC](#) license



Corresponding Author:

Dewa Gede Indra Putra

Informatics Study Program, Faculty of Information Technology and Design, Universitas Primakara, Bali

dewaindra705@gmail.com

1. Introduction

Artificial Intelligence (AI) has driven significant transformations in human-computer interaction, particularly through the emergence of increasingly sophisticated chatbots. Large Language Model (LLM)-based chatbots, such as ChatGPT, are capable of generating context-aware and nuanced responses, contributing to the substantial growth of the global chatbot market [1]. Nevertheless, LLMs face a fundamental limitation: their knowledge base is static, preventing them from accessing up-to-date information or domain-specific knowledge that is not included in their training data [2].

Retrieval-Augmented Generation (RAG) has emerged as a promising approach by combining the capabilities of LLMs with external knowledge sources to generate more accurate, relevant, and context-aware responses [3]. RAG operates by retrieving relevant information from external sources—such as PDF documents—and incorporating it into the LLM prompt, thereby enabling the generated responses to be grounded in verifiable facts while reducing the risk of hallucination [3].

Among the various text representation techniques applicable to the RAG pipeline, FastText offers distinct advantages for the Indonesian language. This subword-based embedding model effectively handles out-of-vocabulary (OOV) words by decomposing them into character n-grams, making it particularly suitable for languages with complex agglutinative and affixational morphology [4]. Empirical studies have

demonstrated that FastText achieved an accuracy of 97.2% on an Indonesian news dataset, outperforming Word2Vec (92.5%) and GloVe (95.8%) [5].

This research focuses on a case study of PT Omni Hottilier, a cloud-based hospitality solution provider that offers comprehensive PDF manuals as the primary knowledge base for customer support. Two research gaps were identified: (1) the limited comparative evaluation between a FastText-based baseline RAG architecture and advanced RAG architectures on Indonesian-language corpora; and (2) the absence of an integrated implementation of Hybrid Search, Maximal Marginal Relevance (MMR), and Cross-Encoder re-ranking within a single RAG pipeline for technical documentation. This study contributes by presenting a quantitative comparative evaluation using the RAGAS framework with five evaluation metrics and by proposing a replicable model for transforming technical documents into a chatbot knowledge base.

2. Literature Review

Retrieval-Augmented Generation (RAG)

Retrieval-Augmented Generation (RAG) is a framework designed to enhance the capabilities of Large Language Models by integrating external information retrieval into the response generation process [3]. A typical RAG architecture consists of two primary components: a retriever, which extracts relevant information from an external knowledge base, and a generator, which utilizes the retrieved information to produce responses. The overall effectiveness of a RAG system depends largely on the quality of these two components and their interaction [6].

FastText Word Embeddings

FastText is an extension of Word2Vec that incorporates subword information into word representations [7]. Instead of representing each word as a single token, FastText models a word as the aggregation of its constituent character n -grams, enabling the generation of vector representations for previously unseen words. This capability is particularly important for the Indonesian language, which features a highly productive affixation system—including prefixes, suffixes, and circumfixes—that produces a large number of derived lexical forms [8].

Hybrid Search and Dynamic Alpha Tuning

Hybrid Search combines lexical retrieval based on BM25 with semantic retrieval using vector embeddings [9]. BM25 is highly effective at capturing exact matches for domain-specific technical terms, whereas vector-based retrieval excels at identifying semantic similarities. Dynamic Alpha Tuning (DAT) adaptively adjusts the weighting between these two retrieval approaches according to the characteristics of the input query [10], resulting in more robust retrieval performance than either method alone.

Maximal Marginal Relevance (MMR)

Maximal Marginal Relevance (MMR) is a re-ranking algorithm designed to balance retrieval relevance and result diversity [11]. MMR iteratively selects documents according to the following formula:

$$\text{MMR}(d_i) = \lambda \cdot \text{Sim}_1(d_i, q) - (1 - \lambda) \cdot \max \text{Sim}_2(d_i, d_j),$$

where λ controls the trade-off between relevance to the query (Sim_1) and similarity to documents that have already been selected (Sim_2). This approach minimizes redundancy among the retrieved contexts supplied to the LLM, thereby improving the diversity and quality of contextual information [12].

RAGAS Evaluation Framework

RAGAS (Retrieval-Augmented Generation Assessment) is a comprehensive evaluation framework specifically developed for assessing RAG systems [13]. The RAGAS framework includes five evaluation metrics: Context Relevance, which measures the relevance of the retrieved context to the query; Context Precision, which evaluates the precision of the retrieved context; Context Recall, which measures the completeness of the retrieved information; Faithfulness, which quantifies the extent to which the generated response is supported by the retrieved context and serves as an indicator of hallucination resistance; and Answer Relevance, which assesses the relevance of the generated response to the user's query.

3. Methods

Methodology and Dataset

This study adopted the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology, which consists of six phases: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment. The dataset comprised Indonesian-language PDF user manuals for the PT Omni Hottilier Booking Engine, serving as the system's sole knowledge source. These documents contain technical operational instructions for various booking engine features, including hospitality-specific terminology and software technical documentation.

The hyperparameters were selected based on established information retrieval principles to balance retrieval relevance, contextual diversity, and ranking stability. The MMR parameter ($\lambda = 0.7$) was chosen to prioritize highly relevant documents while maintaining sufficient diversity and reducing redundancy, consistent with the trade-off proposed by [11].

The Hybrid Search weights (0.4 FastText, 0.4 BM25, and 0.2 Context Score) were designed to balance semantic and lexical retrieval. FastText captures semantic similarity, whereas BM25 effectively retrieves exact technical terms. The Context Score was assigned a lower weight because it serves as a refinement factor rather than the primary retrieval criterion. The crossEncoderAlpha value (0.6) was selected to give slightly greater emphasis to Cross-Encoder reranking while preserving the ranking produced by the hybrid retrieval stage. This improves document ordering without excessively overriding the initial retrieval scores.

Finally, the similarity threshold (0.2) was adopted to remove weakly related documents while maintaining adequate recall. This relatively low threshold is appropriate for Indonesian technical documents, where semantically relevant text may exhibit moderate cosine similarity due to variations in terminology and writing style. Overall, these hyperparameter values were determined based on theoretical considerations of precision, recall, contextual diversity, and ranking stability, rather than solely on implementation convenience.

Baseline Model Architecture

The Baseline Model implements a standard RAG pipeline consisting of four components: (1) Embedding, using 100-dimensional FastText embeddings with the Skip-gram architecture trained for 10 epochs; (2) Retrieval, based on cosine similarity between the query embedding and precomputed document embeddings; (3) Generation, using ChatGPT (gpt-3.5-turbo) configured with a temperature of 0.7 and a maximum output length of 500 tokens; and (4) Context Selection, where the top- k documents with the highest cosine similarity scores are provided as contextual input to the language model.

Proposed Model Architecture

The Proposed Model enhances the retrieval pipeline through three primary innovations. First, Hybrid Search combines three weighted scoring components: Semantic Score (FastText cosine similarity), Keyword Score (BM25Okapi), and Context Score, which evaluates text quality based on document length, sentence completeness, and query term coverage. For complex queries, an adaptive weighting scheme is applied, assigning weights of 0.4 to FastText, 0.4 to BM25, and 0.2 to the Context Score.

Second, Maximal Marginal Relevance (MMR) Re-ranking is implemented with $\lambda = 0.7$, utilizing token-based stemmed Jaccard Similarity to prioritize document relevance while promoting contextual diversity. Third, Cross-Encoder Re-ranking employs TfidfVectorizer with `ngram_range = (1,2)` to perform pairwise relevance scoring between the query and each candidate document. The final ranking incorporates a crossEncoderAlpha weight of 0.6.

Table 1. Comparison of the Baseline Model and Proposed Model Architectures

Component	Baseline Model	Proposed Model
Embedding	FastText 100D Skip-gram	FastText 100D Skip-gram
Keyword Index	—	BM25Okapi
Retrieval Method	Cosine Similarity	Hybrid: FastText + BM25 + Context Score
Weighting Strategy	None	Dynamic weighting based on query type
Stage 1 Re-ranking	—	MMR (Jaccard, $\lambda = 0.7$)
Stage 2 Re-ranking	—	TF-IDF Cross-Encoder ($\alpha = 0.6$)
Similarity Threshold	None	0.2

Evaluation Protocol

The evaluation used 20 test queries distributed across three complexity levels: Simple (Q1–Q7, 7 queries), Medium (Q8–Q14, 7 queries), and Complex (Q15–Q20, 6 queries). Both models were tested simultaneously with identical PDF documents and GPT parameters. The RAGAS framework was implemented using a deterministic and reproducible term-based overlap approach, with a weighting scheme: Faithfulness 35%, Answer Relevance 20%, and Context Relevance, Context Precision, and Context Recall 15% each.

4. Result and Discussion

Baseline Model Performance

The baseline model recorded an average Overall Score of 0.7853 across 20 test queries. It achieved very high Context Precision (0.9916) and high Answer Relevance (0.9488), but weak Context Relevance (0.5040) and Context Recall (0.7504). A Faithfulness score of 0.7387 indicates that approximately 26% of ChatGPT answers were not fully supported by the retrieved context—indicating a real risk of hallucinations. The baseline's weakness stems from the limitations of pure cosine similarity search: when users use specific technical terms, the baseline can fail to retrieve relevant chunks if the surrounding semantic context is less similar to the query, as evidenced by the high variation in Context Relevance (0.250–0.833) between individual queries.

Proposed Model vs. Baseline Comparison

Table 2. Average RAGAS Metrics: Baseline vs. Proposed Model (20 Queries)

RAGAS Metric	Baseline	Proposed	Difference (P–B)	Improvement (%)
Context Relevance	0.5040	0.6041	+0.1001	+19.85%
Context Precision	0.9916	1.0000	+0.0084	+0.84%
Context Recall	0.7504	0.8761	+0.1257	+16.76%

RAGAS Metric	Baseline	Proposed	Difference (P-B)	Improvement (%)
Faithfulness	0.7387	0.8477	+0.1090	+14.76%
Answer Relevance	0.9488	0.9417	-0.0071	-0.75%
Overall Score	0.7853	0.8572	+0.0719	+9.16%

The Proposed Model outperformed Baseline on 16 out of 20 queries (80%), with one query tied (Q4, score 0.950) and Baseline outperforming only three queries (Q1, Q8, Q10). The largest improvements occurred in three context-quality metrics: Context Relevance (+19.85%), Context Recall (+16.76%), and Faithfulness (+14.76%). These improvements can be attributed to three architectural innovations: (1) BM25 ensures exact lexical relevance for specific technical terms; (2) context scoring serves as an additional quality filter; (3) MMR with $\lambda=0.7$ ensures the context sent to ChatGPT is diverse, preventing information redundancy. The only metric that saw a slight decline was Answer Relevance (-0.75%), due to the Proposed Model's tendency to produce answers that are more concise and focused on the document's context—not an indication of an overall decline in quality.

Table 3. Average Overall Score per Query Complexity Level

Query Level	Number of Queries	Baseline Overall Score	Proposed Overall Score	Improvement (%)
Simple (Q1–Q7)	7	0.7893	0.8623	+9.25%
Medium (Q8–Q14)	7	0.7966	0.8751	+9.86%
Complex (Q15–Q20)	6	0.7673	0.8302	+8.19%
Total (Q1–Q20)	20	0.7853	0.8572	+9.16%

The Proposed Model consistently outperforms across all complexity levels, with the largest margin at the Medium level (+9.86%). On Medium queries, the combination of dynamic weights optimally balances the contributions of lexical and semantic search because queries typically contain two to three keywords with well-defined contexts. On Complex queries, Baseline also shows improved performance, narrowing the gap, although the Proposed Model remains superior. Analysis of the queries where Baseline won (Q1, Q8, Q10) reveals that these queries target information localized within a single document chunk, so the overhead of MMR and re-ranking does not add value—a finding that demonstrates the RAGAS framework's ability to detect specific scenarios where the two architectures have different characteristics.

The evaluation results indicate that the proposed Retrieval-Augmented Generation (RAG) framework achieved better overall performance than the Baseline Model across most evaluation queries. This improvement is consistent with previous studies demonstrating that RAG enhances response quality by grounding large language models with retrieved external knowledge, thereby reducing hallucinations and improving factual consistency. Furthermore, the use of Hybrid Search aligns with findings that combining lexical retrieval methods such as BM25 with semantic embedding models provides more robust retrieval performance than relying on a single retrieval strategy. In particular, BM25 effectively retrieves documents through exact keyword matching, whereas FastText captures semantic relationships between words, enabling the system to retrieve relevant information even when different terminology is used. The addition of MMR and Cross-Encoder reranking further improves retrieval quality by reducing redundant contexts and refining document ranking according to query relevance.

Although the proposed framework demonstrated superior overall performance, the Baseline Model produced better results for queries Q1, Q8, and Q10. These queries mainly consisted of straightforward factual questions containing explicit keywords that closely matched the source documentation. Under these conditions, the Baseline Model benefited from direct lexical matching, allowing the correct documents to be retrieved without requiring additional semantic processing. Conversely, the proposed framework applies

multiple retrieval stages, including semantic retrieval, MMR diversification, and Cross-Encoder reranking. While these components generally improve performance for complex or semantically diverse queries, they may slightly modify the ranking of documents that already exhibit strong lexical similarity, leading to marginally lower performance on simple fact-based queries.

These findings suggest that the proposed method is particularly effective for handling complex information needs involving semantic variation and contextual reasoning, whereas conventional lexical retrieval remains competitive for queries characterized by explicit keywords and direct factual information. Therefore, the proposed framework offers a more balanced and robust retrieval strategy across diverse query types while acknowledging the continued effectiveness of traditional lexical retrieval for simple keyword-oriented searches.

5. Conclusion

This research successfully designed, implemented, and evaluated a RAG system that integrates ChatGPT with FastText word embedding for processing Indonesian PDF documents. Three main conclusions can be drawn:

1. The baseline FastText-based RAG pipeline proved suitable as a benchmark with an Overall Score of 0.7853, but showed weaknesses in Context Relevance and the risk of hallucinations (Faithfulness 0.7387).
2. The combination of Hybrid Search (BM25 + FastText + Context Score), MMR, and the TF-IDF Cross-Encoder significantly improved retrieval quality: the Proposed Model outperformed the Baseline on 80% of queries with an Overall Score improvement of +9.16% and the largest improvements in Context Relevance (+19.85%), Context Recall (+16.76%), and Faithfulness (+14.76%).
3. The RAG framework proved effective as an objective and replicable evaluation tool, capable of detecting model superiority at three levels of granularity: aggregate, per-query, and per complexity level.

For further research, it is recommended to: (1) compare FastText with contextual embedding models such as IndoBERT or multilingual-e5; (2) use LLM-as-a-judge (GPT-4) for deeper semantic evaluation; (3) explore context compression strategies to optimize API token efficiency.

6. References

- [1] H. Alboqami, "Factors Affecting Consumers Adoption of AI-Based Chatbots: The Role of Anthropomorphism," *American Journal of Industrial and Business Management*, vol. 13, no. 04, pp. 195–214, 2023. <https://doi.org/10.4236/ajibm.2023.134014>
- [2] K. Wang, J. Ramos, and R. Lawrence, "ChatEd: A Chatbot Leveraging ChatGPT for an Enhanced Learning Experience in Higher Education," *arXiv:2401.00052*, 2023.
- [3] Y. Gao et al., "Retrieval-Augmented Generation for Large Language Models: A Survey," *arXiv:2312.10997*, 2023.
- [4] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching Word Vectors with Subword Information," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135–146, 2017.
- [5] E. M. Dharma et al., "The Accuracy Comparison Among Word2Vec, GloVe, and FastText Towards CNN Text Classification," *Journal of Theoretical and Applied Information Technology*, vol. 31, no. 2, 2022.
- [6] A. Bansal, "Optimizing RAG with Hybrid Search and Contextual Chunking," *Journal of Engineering and Applied Sciences Technology*, 2023. [https://doi.org/10.47363/JEAST/2023\(5\)E114](https://doi.org/10.47363/JEAST/2023(5)E114)

- [7] O. Zaland, M. Abulaish, and Mohd. Fazil, "A Comprehensive Empirical Evaluation of Existing Word Embedding Approaches," arXiv:2303.07196, 2023.
- [8] A. Amalia et al., "An Efficient Text Classification Using fastText for Bahasa Indonesia Documents Classification," in Proc. DATABIA 2020, pp. 69–75.
- [9] Ö. Hakdağlı, "Hybrid Question-Answering System: A FAISS and BM25 Approach for Extracting Information from Technical Document," Orclever Proceedings, vol. 5, no. 1, pp. 226–237, 2024.
- [10] H.-L. Hsu and J. Tzeng, "DAT: Dynamic Alpha Tuning for Hybrid Retrieval in Retrieval-Augmented Generation," arXiv:2503.23013, 2025.
- [11] J. Carbonell and J. Goldstein, "The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries," in Proc. SIGIR, 1998.
- [12] S. G. Aranzamendez et al., "An Enhanced Content-based Filtering Using Maximal Marginal Relevance," International Journal of Computing Sciences Research, vol. 8, pp. 3070–3087, 2024.
- [13] S. Es et al., "RAGAS: Automated Evaluation of Retrieval Augmented Generation," arXiv:2309.15217, 2023.
- [14] A. A. Khan et al., "Developing Retrieval Augmented Generation (RAG) based LLM Systems from PDFs: An Experience Report," arXiv:2410.15944, 2024.
- [15] L. S. Hartono, E. I. Setiawan, and V. Singh, "Retrieval Augmented Generation-Based Chatbot for Prospective and Current University Students," IJESTY, vol. 5, no. 3, pp. 268–277, 2025.
- [16] G. D. Albert and A. Voutama, "Pengembangan Chatbot Berbasis PDF Menggunakan Local RAG dan Ollama," Jurnal Informatika dan Teknik Elektro Terapan, vol. 13, no. 2, pp. 937–944, 2025.
- [17] F. Schlatt et al., "Set-Encoder: Permutation-Invariant Inter-Passage Attention for Listwise Passage Re-Ranking with Cross-Encoders," 2024.
- [18] P. J. Sager et al., "Deep Retrieval at CheckThat! 2025: Identifying Scientific Papers via Hybrid Retrieval and Re-Ranking," arXiv:2505.23250, 2025.
- [19] G. Byun et al., "Secure Multifaceted-RAG for Enterprise: Hybrid Knowledge Retrieval with Security Filtering," arXiv:2504.13425, 2025.
- [20] W. Zhang and J. Zhang, "Hallucination Mitigation for Retrieval-Augmented Large Language Models: A Review," Mathematics, vol. 13, no. 5, 2025.