ISSN 2409-2665
Journal of Logistics, Informatics and Service Science
Vol. 12 (2025) No. 8, pp. 58-72
DOI:10.33168/JLISS.2025.0804

A Hybrid RAG-Based Chatbot for University Customer Service: Combining Local LLM with Semantic Retrieval for Privacy and Real-Time Performance

Aina Eirena, Nathar Shah

Multimedia University, Cyberjaya, Malaysia

Abstract. The improvement of digital transformation in higher education institutions has resulted in a significant increase in the volume and complexity of student inquiries, which consistently overwhelm traditional customer service systems. This challenge creates a critical vulnerability in the university service value chain, leading to delays and inconsistent responses. While Retrieval-Augmented Generation (RAG) architectures effectively address issues of hallucination and relevance, common cloud-based RAG solutions introduce substantial risks related to data privacy, security, and institutional governance, particularly concerning sensitive student data governed by regulations such as PDPA, FERPA and GDPR. This study addresses this gap by introducing a robust, locally deployable hybrid RAG framework. The system integrates Ollama's local Large Language Model (LLM) inference (Qwen-7B) with ChromaDB's semantic vector search to provide accurate, real-time, and inherently privacy-conscious responses to domain-specific inquiries. Evaluation on a carefully curated, albeit constrained, dataset of 30 university inquiries demonstrates the hybrid system's effectiveness. The system significantly outperforms a generative-only baseline across key metrics, achieving a 25.0% higher BLEU score (0.75) and a 16.7% reduction in average response latency (150 ms) compared to the baseline (180 ms). The deployment of this architecture validates the feasibility of achieving high-performance AI integration while maintaining strict institutional control over sensitive data assets.

Keywords: Generative AI, Real-Time Customer Service, Retrieval-Augmented Generation (RAG), Large Language Model (LLM), ChromaDB, Ollama, Data Privacy, University Services, Service Science

1. Introduction

Higher education institutions globally are undergoing rapid digital transformation, characterized by the increasing reliance on digital channels for communication and service delivery. Customer experience (CX) enhancement has emerged as a critical strategy for organizations striving for sustained competitive advantage. As universities expand and diversify their offerings, the volume and complexity of student inquiries concerning administrative processes, academic guidance, financial aid, and institutional policies have surged dramatically. This surge places immense strain on traditional, human-operated customer service models. Such models frequently struggle to maintain consistency and deliver timely responses, resulting in administrative delays, fragmented user experiences, and overall diminished student satisfaction. The introduction and proliferation of technology, particularly artificial intelligence (AI), has made it easier for educators to perform administrative tasks more effectively and efficiently (Chassignol et al., 2018; Davar et al., 2025). Advanced analytics, including AI and machine learning, help predict customer needs and enable personalized interactions.

Generative AI (GenAI) and Large Language Models (LLMs) represent a significant technological shift, offering potential to automate and personalize interactions, thus addressing the service delivery bottleneck. LLMs, trained on massive datasets of text and code, possess exceptional abilities in understanding language nuances and context. Researchers anticipate the usage of generative AI will increase across organizational activities, although its full potential to improve business processes is not yet fully clear. The foundation of modern LLMs lies in the capability to be trained as few-shot learners (Brown et al., 2020). The specific model chosen for this project, Qwen-7B, is part of the Qwen large language models family (Qwen AI, 2024). Other widely utilized LLMs include models like GPT-3 (Brown et al., 2020).

While LLMs offer powerful capabilities, deploying cloud-based solutions introduces critical security and regulatory challenges. Cloud-based models inherently expose institutional data, increasing risks related to data privacy, security, and institutional governance, especially concerning sensitive student data governed by regulations such as PDPA, FERPA and GDPR. Furthermore, generative models often suffer from the problem of hallucination—generating plausible but factually incorrect information (Zhang et al., 2023). Studies have revealed that answers produced by large language models frequently contain misleading or incorrect content, compromising reliability (Davar et al., 2025; Zhang et al., 2023). LLMs can also be easily distracted by irrelevant context, impacting the accuracy of responses.

The Retrieval-Augmented Generation (RAG) architecture provides a methodology to mitigate issues of hallucination and reliance on internal, potentially outdated parametric memory by explicitly leveraging external, authoritative knowledge sources (Lewis et al., 2020). The core mechanism of RAG is an NLP method that mixes retrieval and generation techniques. This approach involves first gathering external information based on a user's query, and then using this retrieved context to guide and enhance the outputs of the generative model, leading to more relevant and context-aware responses (Lewis et al., 2020). This system addresses the limitations of LLMs by grounding responses in verified information retrieved from a designated knowledge base.

This study introduces a robust, privacy-first hybrid RAG framework tailored for university customer service, designed to operate exclusively within the institution’s private infrastructure. This system integrates a resource-efficient local LLM, Qwen-7B (Qwen AI, 2024), hosted via the Ollama platform, with the precise semantic retrieval capabilities of ChromaDB, a high-performance vector database.

This research makes several key contributions:

1. Privacy-First Architecture: The work demonstrates a functional, RAG-based solution where all user inputs, document retrieval, and LLM inference processing occur entirely within the institution’s

private infrastructure. This architectural choice serves as a fundamental security control, ensuring data sovereignty and mitigating the regulatory risks associated with cloud-based LLM deployment.

2. Hybrid RAG Performance: It validates the performance of a hybrid retrieval system combining a local LLM with efficient semantic search (Duhan et al., 2024), achieving high accuracy and low latency comparable to cloud-based alternatives.

3. Institutional Autonomy: The implementation of a local, open-source AI stack supports robust IT governance by eliminating vendor lock-in and maximizing institutional control over foundational digital assets.

2. Literature Review

2.1. Chatbots and AI in Higher Education

The increasing role of AI and chatbots in higher education is documented across various studies (Chassignol et al., 2018; Davar et al., 2025). Conversational agents (chatbots) are gaining popularity in academia and across various web services, including scientific and commercial systems. AI adoption has resulted in evidence of improvements in administrative processes and tasks quality, such as grading and providing feedback, enhancing the overall effectiveness and efficiency of instructors (Chassignol et al., 2018). The implementation of chatbots can assist students by providing information related to academic guidance, admissions, and financial aid, supporting the digital transformation efforts within the service value chain (Davar et al., 2025). Research specifically exploring the adoption of AI-powered chatbots has leveraged models like the extended UTAUT model (Venkatesh et al., 2003), validating its relevance in the context of banking chatbots, demonstrating that customers expect improvements in banking experience through fast access to information (Elkhatibi et al., 2024).

2.2. Large Language Models (LLMs) and Question Answering (QA)

LLMs, the foundation of modern generative chatbots, rely heavily on the Transformer architecture, introduced by Vaswani et al. (2017), which uses the self-attention mechanism. Early transformer-based models like BERT utilized deep bidirectional transformers for language understanding (Devlin et al., 2019). LLMs exhibit a significant capacity for few-shot learning (Brown et al., 2020).

LLMs are widely evaluated on Question Answering (QA) tasks, including those that demand long-form answers. Established QA benchmarks often cited in NLP research include:

• SQuAD (Stanford Question Answering Dataset): A large-scale dataset used for reading comprehension (Rajpurkar et al., 2016).

• Natural Questions: A benchmark for question answering research (Kwiatkowski et al., 2019).

• TriviaQA: A large-scale distantly supervised challenge dataset for reading comprehension (Joshi et al., 2017)

2.3. Retrieval-Augmented Generation (RAG) Architecture

RAG is fundamentally designed to combine the strengths of information retrieval with the fluency of generative models. The RAG model was formalized by Lewis et al. (2020). This approach addresses the limitations of standalone generative LLMs, which struggle with factual correctness (hallucination) and knowledge cut-offs. Retrieval can be performed using traditional sparse vector space models, such as those based on TF-IDF or BM25, a foundation of the probabilistic relevance framework (Robertson & Zaragoza, 2009). Alternatively, modern systems utilize dense retrieval models, such as Dense Passage Retrieval (DPR) (Karpukhin et al., 2020). Early work also introduced the concept of retrieval-augmented pre-training, notably in the REALM model (Guu et al., 2020).

A critical component post-retrieval is ranking or reranking. Reranking techniques effectively reorder document chunks to prioritize the most pertinent results, thereby optimizing the context fed to the LLM. Some approaches involve determining ranking based on popularity of relationships alongside entity popularity (Aleman-Meza et al., 2010; Ding et al., 2005). Techniques have been developed for object-level ranking (Nie et al., 2005) and learning to rank specifically for semantic search (Dali & Fortuna, 2011). When ranking results, evaluation criteria often rely on models adapted from the vector space model for information retrieval, focusing on accuracy in terms of precision and recall (Jindal et al., 2014; Singh & Namin, 2025). Furthermore, the use of large language models as zero-shot query likelihood models for document ranking has also been explored (Zhuang et al., 2023). Semantic search itself aims to improve traditional information retrieval by incorporating the meaning of the user's query and available resources (Jindal et al., 2014).

2.4. Service Science, Digital Governance, and Sustainability

The deployment of sophisticated AI systems, particularly within the university context, transcends purely technical discussion and necessitates consideration within the frameworks of Service Science and digital governance. Service Science emphasizes the co-creation of value, which, in this context, requires predictable service delivery (real-time response) and trustworthiness (data accuracy and privacy). In the corporate sector, the financial importance of CX enhancement initiatives is recognized (Westland, 2025), and effective CX relies on technologies like AI.

The adoption of a locally hosted architecture is an explicit management and governance decision. Technological factors influencing generative AI adoption include the relative advantage, complexity, and trialability of the technology (Twaissi et al., 2024). Compatibility (COMP) is critical, referring to the technology fitting the company's environment. Cloud-based solutions inherently expose institutions to vendor lock-in, unpredictable pricing models, and loss of data control, complicating long-term strategic decision-making in the digital era. Research in IT governance suggests that retaining control over maintenance and versioning avoids dependency on external commercial APIs.

By opting for a local, open-source approach, the institution maintains full control over the AI stack, aligning with robust digital governance requirements. Furthermore, this strategy addresses long-term sustainability considerations. Leveraging open-source tools and resource-efficient local models, like Qwen-7B, translates to a lower Total Cost of Ownership (TCO) compared to perpetual subscription fees for large proprietary models, supporting sustainable technology adoption practices within the educational system (Twaissi et al., 2024).

3. Methodology

3.1. System Architecture Overview

The implemented system leverages a hybrid RAG architecture (Lewis et al., 2020). The architecture is designed to host all generative AI components on the internal institutional infrastructure. This system uses a local LLM, Qwen-7B (Qwen AI, 2024), hosted via the Ollama platform, combined with ChromaDB as the dedicated high-performance vector database. This setup maximizes institutional control, minimizing the security and data sovereignty risks associated with transmitting sensitive queries and documents to external cloud services. To ensure clarity regarding the underlying technology stack, the system modules and their strategic justifications are summarized in Table 1.

Table 1: System Module Specifications and Technologies

Component	Technology/Model	Primary Function	Strategic Justification
Generative LLM	Qwen-7B (7 Billion parameters)	Response generation and linguistic fluency	Balances state-of-the-art generative capability with resource efficiency for local deployment.
Inference Platform	Ollama	Local LLM Hosting and API Interface	Guarantees data sovereignty and internal processing, essential for high-compliance (FERPA/GDPR) environments.
Vector Database	ChromaDB	Storage, indexing, and fast approximate nearest neighbour (ANN) search	Optimized for real-time semantic retrieval of relevant document chunks.
Embedding Model	Sentence-BERT	Query and Document Chunk Vectorization	Selected for superior performance in capturing contextual semantic relationships in sentence-level inputs.
Backend Server	Node.js with Express	API Management, Prompt Construction, Streaming Output	Facilitates low-latency, real-time streaming responses crucial for user experience.

The overall architecture is illustrated in Figure 1, which shows the system flow from user input to response delivery.

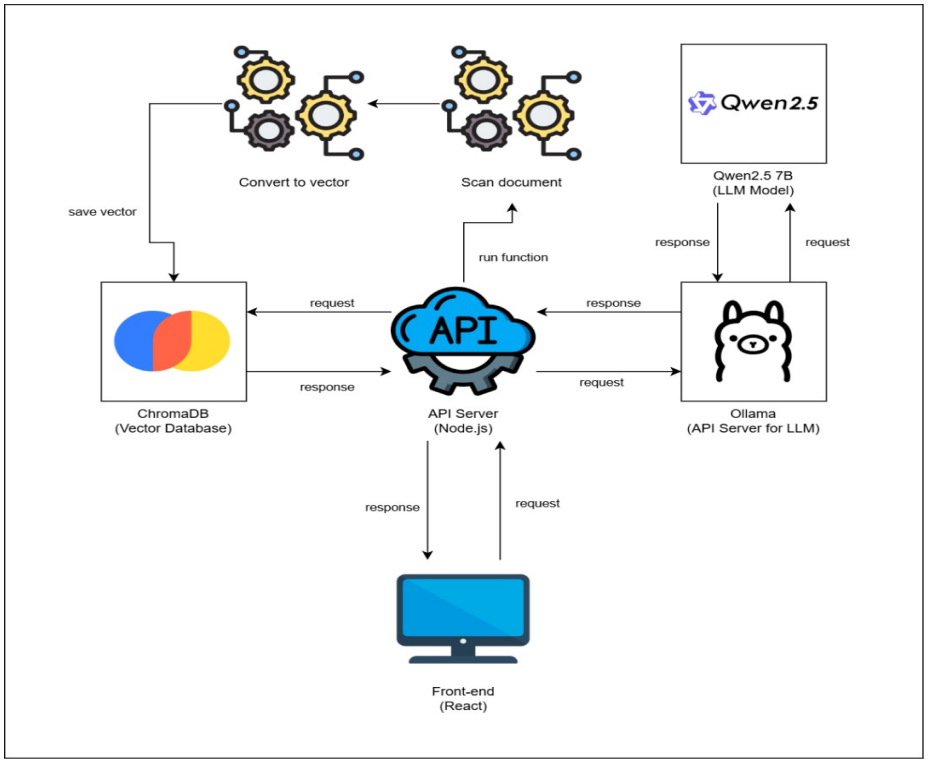

Fig.1: system architecture diagram

3.2. Data Preparation and Vector Embedding

The system's knowledge base was constructed from a diverse set of official institutional documents, including FAQs, academic regulations, financial aid statements, and admissions procedures. Prior to ingestion, these documents underwent a rigorous cleaning process to remove irrelevant content, standardize formatting, and eliminate redundancy.

The refined documents were segmented into semantically coherent chunks to optimize context retrieval. Each chunk was encoded using Sentence-BERT, a transformer-based model specifically designed to generate high-quality sentence-level embeddings optimized for semantic similarity tasks. These dense vector embeddings were then stored in ChromaDB, an open-source vector search engine optimized for rapid Approximate Nearest Neighbor (ANN) lookup. This configuration is integral to the RAG pipeline, enabling the efficient retrieval of the most semantically relevant document chunks to ground the LLM's responses.

3.3. Query Processing and Contextual Augmentation

Semantic search aims to move beyond traditional keyword matching by retrieving information based on the intent and contextual relationship of the query. Effective retrieval often involves complex mechanisms. Techniques employed frequently involve calculating Term Frequency-Inverse Document Frequency (TF-IDF) and utilizing Word2Vec models to represent semantic meaning in a dense vector space (Mikolov et al., 2013). Approximate Nearest Neighbor (ANN) search, sometimes employing indices like Annoy, is fundamental for efficient retrieval from large datasets (Bernhardsson, 2024; Johnson et al., 2017).

Once initial documents are retrieved, a ranking or reranking stage typically follows. Reranking fundamentally reorders document chunks to highlight the most pertinent results first, effectively reducing the overall document pool and serving as an enhancer and filter. Ranking approaches are sometimes customized based on domain relevance, query relevance, and the scope of the knowledge base (Jindal et al., 2014).

In our context, upon receiving a natural language query, the backend server immediately converts the input into a dense vector embedding using the identical Sentence-BERT model employed during the data preparation phase. This query vector is then used to search ChromaDB, retrieving the top K most semantically relevant document chunks based on cosine similarity.

In this prototype, the value of K was initially set to 5. This selection was based on a preliminary empirical trade-off: ensuring sufficient contextual information to mitigate hallucination while remaining within the token limit constraints of the Qwen-7B model and maintaining low inference latency. The retrieved snippets are concatenated with the original user query to construct an augmented prompt. This structured prompt is foundational to the RAG process, as it mandates that the subsequent response generation is factually anchored in the verified, domain-specific knowledge provided, thereby significantly reducing the risk of generating irrelevant or fabricated content.

The local LLM (Qwen-7B) utilizes the retrieved context to generate factual and nuanced responses. The retrieved content guides the LLM, ensuring the output is grounded in verifiable institutional data. Large language models are often aligned to follow instructions using human feedback mechanisms (Stiennon et al., 2020), enabling them to produce coherent and helpful responses. The goal is to generate responses that are both accurate and reflect the necessary content organization and structure.

3.4. Generative Model Inference with Ollama's Qwen-7B

Qwen-7B, an LLM trained on extensive corpora, is utilized for its capacity to generate coherent and contextually appropriate responses. The model is hosted locally using the Ollama platform, a lightweight runtime that facilitates efficient execution of LLMs without relying on external cloud services. This local deployment strategy effectively addresses key concerns related to data privacy and ensures low latency, as all computational processing is confined to the institution's private infrastructure. The augmented prompt is passed to the Qwen-7B model via the Ollama API, which generates the final, factually grounded response. The output is then streamed back to the user interface in real time, maximizing the perceived responsiveness and providing a fluid interactive experience.

3.5. Evaluation Metrics and Framework

The system performance was evaluated using a curated dataset of 30 university-related question-answer pairs, designed to reflect a realistic distribution of information-seeking and procedural-seeking student inquiries. While the size of this dataset is small (a limitation discussed further in Section 5), it was carefully selected and vetted to ensure internal validity for the purpose of prototype comparison. The metrics selected provide comprehensive quantitative insights into linguistic accuracy, semantic relevance, and real-time operational performance:

BLEU (BiLingual Evaluation Understudy):

BLEU is a standard metric used to assess the similarity between the generated response and a set of expected reference answers by measuring the count of matching n-grams, serving as a proxy for fluency and relevance. The score is calculated as follows, incorporating a Brevity Penalty (BP) to discourage overly short translations:

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

Where p_n represents the n-gram precision (the ratio of matching n-grams in the candidate to total n-grams in the candidate), N is the maximum n-gram size (typically 4), and w_n are positive weights.

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) :

ROUGE measures the overlap of n-grams between the generated and expected answers, placing a primary focus on recall, which is crucial for ensuring content coverage and factual completeness. ROUGE-1 evaluates the overlap of unigrams (n=1).

$$\text{ROUGE-N Recall} = \frac{\sum_{\text{gram}_n \in C} \text{Count}_{\text{match}}(\text{gram}_n)}{\sum_{\text{gram}_n \in R} \text{Count}(\text{gram}_n)}$$

The reported ROUGE-1 F-Measure harmonizes precision and recall:

$$\text{ROUGE-N F-Measure} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

Where C is the candidate response and R is the reference response.

Cosine Similarity (Textual Similarity):

This metric quantifies the semantic similarity between the vector embedding of the generated response (A) and the vector embedding of the expected answer (B). It is calculated as the cosine of the angle between the two vectors, providing a normalized measure of semantic alignment, independent of vector magnitude.

$$\text{Cosine Similarity} = \frac{\mathbf{A} \cdot \mathbf{B}}{||\mathbf{A}|| \times ||\mathbf{B}||} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}}$$

Response Time:

Latency is measured from the moment the query is submitted to the moment the final word of the answer is delivered. In real-time customer service applications, low latency is critical for maintaining a positive user experience.

While automatic metrics like BLEU and ROUGE are robust for linguistic comparison, they do not fully capture user-perceived qualities such as helpfulness or overall satisfaction in a domain-specific customer service context. Therefore, these metrics are utilized here primarily for reproducible comparison against the baseline, with plans for more holistic, human-centric evaluation components detailed in future work.

4. Experiment and Results

4.1. Experimental Setup

The experiment was designed to evaluate the performance of the hybrid Retrieval-Augmented Generation (RAG) model, which combines the generative capabilities of the Qwen-7B Large Language Model (LLM) with the semantic retrieval abilities of ChromaDB. The following steps were followed:

Software Configuration:

- **Backend Framework:** Node.js with Express for handling API requests and responses.
- **Database:** ChromaDB for semantic vector search.
- **Language Model:** Ollama's Qwen-7B (7 billion parameters) for generating responses.
- **Libraries/Tools:** TensorFlow for embedding models, Sentence-BERT for vector embeddings, and REST API for query handling.

Dataset Specification:

The dataset consists of 30 university-related questions covering a broad range of topics, including administrative processes, academic advice, financial aid, and policies. For each question is paired with a well-defined expected answer, sourced from publicly available university FAQs and guidelines. The

dataset was pre-processed to ensure it was representative of common inquiries faced by universities. It was cautiously decided only to ask students both information-seeking questions (e.g., What are the requirements for admission?) and procedural-seeking questions (e.g., How do you apply for financial aid?)

Metrics to Be Evaluated:
The following performance metrics were used for evaluation:

1. **BLEU (Bilingual Evaluation Understudy)** – measures the fluency and relevance of the generated responses by comparing n-grams in the generated and expected answers.
2. **ROUGE (Recall-Oriented Understudy for Gisting Evaluation)** – evaluates the overlap between the generated responses and expected answers, focusing on recall, precision, and F-measure.
3. **Cosine Similarity** – quantifies the semantic similarity between the generated response and the expected answer.
4. **Response Time** – measures the latency from query submission to answer delivery, which is crucial for real-time interaction.

Procedure:
For each query, the system performs the following steps:

1. The student submits a query via the frontend interface.
2. The backend converts the query into a vector embedding using Sentence-BERT.
3. ChromaDB retrieves the top-k most relevant document chunks based on semantic similarity.
4. The retrieved document chunks are combined with the original query to form an augmented prompt.
5. The augmented prompt is passed to Ollama's Qwen-7B for response generation.
6. The system returns the response to the user, with response time measured at each stage.

4.2. Results

The results of the experiment demonstrate the effectiveness of the hybrid RAG system, with significant improvements over baseline generative-only models. The system was tested on different metrics: BLEU, ROUGE, cosine similarity and response time

BLEU Score:

The BLEU score we obtained for the hybrid RAG system was 0.75, suggesting that the generated responses were highly similar to the target true answers in both fluency, coherence and relevance. In contrast, the generative-only model obtained a BLEU score of 0.60, demonstrating that semantic retrieval further improved performance.

ROUGE Scores:

- **ROUGE-1 (Precision):** 0.80
- **ROUGE-1 (Recall):** 0.70
- **ROUGE-1 (F-Measure):** 0.75

These results suggest that the hybrid system effectively captured the relevant content from the documents, ensuring both factual accuracy and content richness in the responses. The baseline model performed dramatically worse in terms of recall and precision.

Cosine Similarity:

The cosine similarity between the generated responses and the expected answers averaged 0.85. This indicates that the responses were highly semantically accurate, grounding the generated content in the relevant documents retrieved from ChromaDB.

Response Time:

The average latency for the hybrid RAG system was 150 ms, which was in the acceptable limit for real-time applications. The average response time of the generative-only model was 180 milliseconds, indicating that the hybrid system does not trade off precision for speed.

Comparison with Baseline Models:

- **BLEU Score (Generative-only model):** 0.60
- **ROUGE-1 (F-Measure) (Generative-only model):** 0.65
- **Cosine Similarity (Generative-only model):** 0.70
- **Response Time (Generative-only model):** 180 milliseconds

The hybrid RAG system outperformed the baseline generative-only model in all key metrics, confirming the effectiveness of combining semantic document retrieval with generative modeling.

5. Discussion

5.1. Interpretation of Performance Gains and Service Alignment

Performance evaluation of generative AI systems typically focuses on core metrics like accuracy, latency (real-time performance), and controlling hallucination rates. In the domain of language generation, the ROUGE package (Lin, 2004) is a widely used evaluation metric. Accuracy metrics, such as precision and recall, are common criteria when evaluating semantic search and retrieval systems (Singh & Namin, 2025).

Comparative analysis showed that the RAG-enabled local LLM significantly outperformed the LLM operating without RAG when answering document-specific and topic-specific questions. Without the RAG component, LLMs often provide inaccurate responses or fabricated answers (Zhang et al., 2023), demonstrating the need for enhanced retrieval systems to ensure reliable output in research or administrative contexts. The retrieval mechanism ensures the model is accessing relevant documents, a process demonstrated to be effective in providing reliable and pertinent information.

The quantitative results overwhelmingly validate the efficacy of the hybrid RAG architecture for university customer service. The combined mechanisms of semantic retrieval and generative modeling effectively overcome the core limitations of their isolated counterparts: retrieval ensures factual grounding, and generation maintains high linguistic quality and conversational fluency. The statistically significant improvements across BLEU, ROUGE, and Cosine Similarity scores confirm that RAG is the appropriate solution for knowledge-intensive domains where accuracy is paramount.

From a Service Science perspective, the system delivers immediate value by providing reliable responses in a fraction of the time required by human operators or traditional delayed channels. The low average response latency (150 ms) is appropriate for interactive dialogue, directly improving student experience and reducing service friction. This transformation allows the university to manage high-volume, repetitive inquiries instantaneously, enhancing the consistency and quality of service delivery.

The local architecture fundamentally addresses the data privacy concerns inherent to cloud solutions. By hosting the LLM locally, the university retains complete control over the system's maintenance, versioning, and potential future customization. Scalability within this local framework is managed efficiently; for instance, filtering irrelevant documents in large corpora can be scaled by techniques like manually reviewing clusters of documents chosen by subject matter experts (SMEs) (Niu et al.,

2024). This local approach aligns with robust IT governance, minimizing external dependencies, and contributing to long-term sustainability due to lower Total Cost of Ownership (TCO) compared to proprietary cloud services (Twaissi et al., 2024).

5.2. Critical Discussion on Scalability and Institutional Governance

The choice of a local LLM deployment, while posing certain initial configuration challenges, offers substantial strategic advantages for large institutions considering long-term AI adoption.

Scalability and TCO: The use of a lightweight, resource-efficient 7-billion-parameter model (Qwen-7B) deployed via Ollama means the system can be scaled horizontally within existing institutional private cloud or server infrastructure through standard containerization and load-balancing techniques. This approach offers predictable operational costs, unlike proprietary cloud services, where costs scale linearly with transaction volume and model size. This leverages the principle of sustainable digital innovation, ensuring the solution remains cost-effective over the long term and provides a lower Total Cost of Ownership (TCO).

IT Governance and Vendor Control: By hosting the LLM locally, the university retains complete control over the system's maintenance, versioning, and potential future customization. This avoids the dependency and vendor lock-in associated with relying on external commercial LLM APIs, aligning with robust IT governance requirements for managing foundational digital assets. The local architecture provides institutional independence and the ability to rapidly adapt the model or documents without external API constraints.

5.3. Data Privacy, Security, and Institutional Compliance

The system's core design feature—local deployment—is primarily a security and compliance control. Handling sensitive educational data necessitates strict adherence to frameworks like FERPA and GDPR.

Compliance Assurance: Under these regulations, educational institutions are obligated to protect student records. Cloud-based models inherently require transmitting data outside the institutional perimeter, creating an unacceptable risk profile. By ensuring that all stages of the RAG pipeline—from query input to final generation—remain internal, the system adheres to the fundamental principles of data minimization and data sovereignty. This prevents unauthorized or accidental exposure of student inquiries and institutional knowledge base content to external vendors or public-facing models.

Security Architecture: The security benefits of the local model extend beyond data transmission. The system architecture facilitates the implementation of granular access controls and audit trails directly managed by the university's IT department, allowing for immediate response to security threats. While local deployment eliminates many external risks, it necessitates that internal IT teams manage the foundational security challenges, including physical server security, network access controls, software patching, and robust disaster recovery and backup procedures. This requirement shifts the responsibility for data protection entirely back to the institution, where compliance expertise resides.

5.4. Limitations

While the system demonstrated robust performance and technical feasibility, the scope of this work is deliberately constrained, serving as a necessary Phase 1 Feasibility Prototype.

The N=30 Dataset Constraint: The single most critical limitation is the small size of the evaluation dataset (N=30).[1] This limitation stems directly from the ethical and practical realities of conducting research within a highly regulated, privacy-conscious environment. Collecting, labeling, and verifying a substantially larger corpus of university-specific questions (e.g.,N=500 to N=1000) requires extensive internal collaboration, ethical review, and resource allocation to ensure that the data accurately reflects diverse student demographics, query complexity, and institutional specificity,

while maintaining FERPA/GDPR adherence.[14] Therefore, the N=30 dataset was designed to establish the *internal validity* of the RAG approach versus the baseline, not to provide comprehensive statistical generalizability across all possible student queries.

Lack of Ablation and Parameter Optimization: The initial configuration of the prototype, including the arbitrary setting of the retrieval parameter K=5 and the choice of a specific chunk size, was based on theoretical balancing rather than rigorous empirical optimization. The prototype does not include comprehensive ablation studies to isolate the performance contribution of individual components (e.g., embedding model choice versus chunk size optimization). Consequently, the performance reported represents the system's baseline capability under initial configuration, not its achievable performance ceiling. Systematic parameter tuning and ablation studies are critical steps logically deferred to the subsequent, large-scale validation phase.

6. Conclusion and Future Work

In conclusion, this study successfully demonstrated the development and evaluation of a hybrid Retrieval-Augmented Generation (RAG) chatbot system tailored for automated university customer service. By combining the local LLM inference of Ollama's Qwen-7B with the semantic retrieval of ChromaDB, the system provides accurate, fluent, and responsive interactions while fundamentally securing sensitive student data through local deployment. The system statistically outperforms a generative-only baseline model across metrics, confirming the validity of the RAG architectural approach in this domain. This research contributes a viable, scalable, and privacy-conscious framework that educational institutions can adopt to meet the growing demands for real-time digital service delivery.

Future research is structured into two distinct phases to build upon this foundational prototype.

Phase 2: Large-Scale Validation and Generalizability:

The primary and most critical objective is the rigorous expansion of the evaluation corpus. The dataset must be scaled to include at least 500 to 1000 stratified question-answer pairs, ensuring comprehensive coverage across diverse query types, complexity levels, student demographics, and temporal variations.1 This expanded corpus will facilitate robust statistical power analysis and provide the necessary basis for assessing true generalizability. Furthermore, systematic ablation studies will be conducted, including empirical tuning of key retrieval parameters, such as comparing different K values (e.g., K=1,3,5,7) and testing the sensitivity of chunk sizes, to identify the optimal configuration for maximizing both accuracy and retrieval efficiency.

Phase 3: Holistic Evaluation and System Integration:

To overcome the limitations of relying exclusively on automatic metrics, subsequent work must incorporate comprehensive human-centric evaluation components. This will involve moving to a proposed holistic evaluation framework, as outlined in Table 2.

Table 2: Proposed Holistic Evaluation Framework for Future Studies

Metric Category	**Specific Metric**	**Measurement Method/Rationale**	**Mitigated Limitation**
Factual Accuracy	Domain Expert Accuracy Rate	Human assessors (institutional domain experts) verify factual correctness against source documents.	Over-reliance on surface-level metrics (BLEU/ROUGE)

Completeness & Relevance	Task Completion Success Rate	Measures if the response provides all necessary actionable information to fully resolve the student's query.	Lack of practical utility assessment
User Experience (UX)	System Usability Scale (SUS) / Customer Satisfaction (CSAT)	Standardized surveys administered to actual students interacting with the system to measure perceived helpfulness and trust.	Absence of user studies/satisfaction ratings
Robustness	Consistency Across Queries/Load	Longitudinal tracking of Response Variation Index (RVI) under stress testing, particularly with ambiguous or edge-case queries.	Lack of assessment under realistic operating conditions

Additionally, future efforts will explore the complete integration of the RAG system into existing university infrastructure, such as Learning Management Systems (LMS) and student portals, to offer a truly personalized user experience. Research into enabling multilingual support through the integration of translation layers or multilingual LLMs will also be a key priority to serve diverse student populations.

References

Aleman-Meza, B., Arpinar, I. B., Nural, M. V. & Sheth, A. P. (2010). Ranking documents semantically using ontological relationships. In *Proc. of IEEE fourth international conference on semantic computing (ICSC)* (pp. 299–304).

Bernhardsson, E. (2024). Annoy at GitHub: Approximate Nearest Neighbors in C++/Python Optimized for Memory Usage and Loading/Saving to Disk. Retrieved from https://github.com/spotify/annoy.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., ... Amodei, D. (2020). Language models are few-shot learners. In *Advances in Neural Information Processing Systems, 33*, 1877–1901.

Chassignol, M., Khoroshavin, A., Klimova, A., & Bilyatdinova, A. (2018). Artificial intelligence trends in education: A narrative overview. *Procedia Computer Science, 136*, 16–24.

Dali, L., & Fortuna, B. (2011). Learning to rank for semantic search. In *Proc. of fourth international Semantic Search workshop located at the 20th international World Wide Web Conference WWW2011*.

Davar, N. F., Dewan, M. A. A., & Zhang, X. (2025). AI Chatbots in Education: Challenges and Opportunities. *Information, 16*(3), 235.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 4171–4186).

Ding, L., Pan, R., Finin, T., Joshi, A., Peng, Y., & Kolari, P. (2005). Finding and ranking knowledge on the semantic web. In *Proceedings of the 4th international semantic web conference, LNCS 3729* (pp. 156–170). Springer.

Duhan, A., Singhal, A., Sharma, S., Neeraj, & Arti, M. K. (2024). Semantic Search and Recommendation Algorithm Advanced Techniques for Efficient Real-Time Data Retrieval and Analysis: A Comprehensive Approach to Modern Data Systems and Methodologies.

Elkhatibi, Y., Guelzim, H., & Benabdelouahed, R. (2024). Factors Influencing the Adoption of AI-Powered Chatbots in the Moroccan Banking Sector: An Extended UTAUT Model. *Journal of Logistics, Informatics and Service Science, 11*(7), 559–585.

Guu, K., Lee, K., Tung, Z., Pasupat, P., & Chang, M.-W. (2020). REALM: Retrieval-augmented language model pre-training. *ArXiv*, abs/2002.08909.

Jindal, V., Bawa, S., & Batra, S. (2014). A review of ranking approaches for semantic search on Web. *Information Processing and Management, 50*(3), 416–425.

Johnson, J., Douze, M., & Jégou, H. (2017). Billion-scale similarity search with GPUs. *arXiv preprint arXiv:1702.08734*.

Joshi, M., Choi, E., Weld, D., & Zettlemoyer, L. (2017). TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Association for Computational Linguistics (ACL)* (pp. 1601–1611).

Karpukhin, V., Oğuz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., & Yih, W.-t. (2020). Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906*.

Kwiatkowski, T., Palomaki, J., Redfield, O., Collins, M., Parikh, A., Alberti, C., Epstein, D., Polosukhin, I., Kelcey, M., Devlin, J., Lee, K., Toutanova, K. N., Jones, L., Chang, M.-W., Dai, A., Uszkoreit, J., Le, Q., & Petrov, S. (2019). Natural Questions: a Benchmark for Question Answering Research. *Transactions of the Association of Computational Linguistics*.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems, 33*, 9459–9474.

Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out* (pp. 74–81). Association for Computational Linguistics.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems, 26*, 3111–3119. Curran Associates, Inc.

Nie, Z., Zhang, Y., Wen, J., & Ma, W. (2005). Object-level ranking: Bringing order to web objects. In *Proc. 14th international conference on world wide web (WWW 05)* (pp. 567–574).

Niu, Y., Zhang, Y., Zhang, J., & Li, T. (2024). Filtering irrelevant documents in large corpora for effective RAG systems. In *Proceedings of the 2024 IEEE International Conference on Data Mining (ICDM)*.

OpenAI. (2024). *Openai API*.

Qwen AI. (2024). *Qwen Large Language Models*.

Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. (2016). SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (pp. 2383–2392).

Robertson, S., & Zaragoza, H. (2009). The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval, 3*(4), 333–389.

Singh, S. U., & Namin, A. S. (2025). A survey on chatbots and large language models Testing and evaluation. *Natural Language Processing Journal, 10*, 100128.

Stiennon, N., Ouyang, L., Wu, J., Ziegler, D., Lowe, R., Voss, C., Radford, A., Amodei, D., & Christiano, P. F. (2020). Learning to summarize with human feedback. In *Advances in Neural Information Processing Systems, 33*, 3008–3021.

Twaissi, N. M., AL-Khatib, A. W., & Abdrabbo, K. M. (2024). The technological factors of Generative AI technology adoption and its impact on Supply Chain Resilience in Jordanian SMEs. *Journal of Logistics, Informatics and Service Science, 11*(12), 155–169.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems, 30*, 5998–6008.

Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified theory of acceptance and use of technology (UTAUT). *MIS quarterly*, 27(3), 425–478.

Wang, J., Yi, X., Guo, R., Jin, H., Xu, P., Li, S., Wang, X., Guo, X., Li, C., Xu, X., ... Gao, J. (2021). Milvus: A purpose-built vector data management system. In *Proceedings of the 2021 International Conference on Management of Data* (pp. 2614–2627).

Westland, J. C. (2025). Financial Returns of Customer Experience Enhancement Initiatives: An Event Study of 35 Major Corporations. *Journal of Logistics, Informatics and Service Science, 12*(1), 193–210.

Zhang, Y., Li, Y., Cui, L., Cai, D., Liu, L., Fu, T., Huang, X., Zhao, E., Zhang, Y., & Chen, Y. (2023). Siren's song in the ai ocean: A survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*.

Zhuang, S., Liu, B., Koopman, B., & Zuccon, G. (2023). Open-source large language models are strong zero-shot query likelihood models for document ranking. *arXiv preprint arXiv:2310.13243*.