Issues in Information Systems
Volume 27, Issue 2, pp. 784-799, 2026

DOI: https://doi.org/10.48009/2_iis_2026_151

AskWhiz: Development and comparative evaluation of retrieval-augmented generation configurations for an institutional knowledge chatbot

Mel John S. Macabenta, *Mapúa Malayan Colleges Mindanao, mjmacabenta@mcm.edu.ph*
Jhouvann S. Morden, *Mapúa Malayan Colleges Mindanao, jmorden@mcm.edu.ph*
Marga G. Pilapil, *Mapúa Malayan Colleges Mindanao, mgpilapil@mcm.edu.ph*
Rogelio Jr. O. Badiang, DIT, *Mapúa Malayan Colleges Mindanao, rjobadiang@mcm.edu.ph*

Abstract

Institutions are frequently asked by students regarding admission, policy, and procedure information. Static websites and FAQ pages do not provide immediate answers to these questions. The purpose of this paper is to present the results of developing and testing "AskWhiz", a Retrieval-Augmented Generation (RAG) knowledge-based chatbot for Mapúa Malayan Colleges Mindanao (MMCM) located in Davao City, Philippines. In addition to developing and testing the chatbot, nine different RAG configurations were tested through use of RAGAS, an open-source automated evaluation system for retrieval-augmented generation systems. These nine configurations combined three different types of retrieval strategies (Basic RAG, Prompt-Engineered RAG, and Hybrid RAG) with three large language models (GPT-4o-mini, Gemini 2.5 Flash, and Claude Haiku 4.5). The performance of each RAG configuration was compared using RAGAS across five different metrics. Based on a One-Way Analysis of Variance (ANOVA) analysis, it was determined that there were statistically significant differences between configurations [$F(8, 437) = 3.874$, $p = .0002$]. A Tukey Honest Significant Difference test indicated that the poorest performing configuration (C2: Basic RAG/Gemini 2.5 Flash) had statistically significant differences from each of the three Hybrid RAG configurations. The best performing configuration was C9 (Hybrid RAG/Claude Haiku 4.5) with an average RAGAS score of .7713; thus, it was used as the operational configuration for AskWhiz. A usability study involving thirty-one academic users resulted in a Mean Rating of Strongly Agree ($M = 4.42/5.00$, Cronbach Alpha = .87) for all five dimensions evaluated. A comparative semantic similarity analysis of AskWhiz versus three general purpose chatbots (GPT-5.5, Gemini 2.5 Flash, and Claude Sonnet 4.6) confirmed that AskWhiz provided better responses than each of the other three chatbots using a Kruskal-Wallis H-test ($H = 66.93$, $p < .0001$), which indicates that the development of domain-specific retrieval based on the parameters of knowledge can be beneficial in terms of providing better quality responses than those generated solely based upon parametric knowledge.

Keywords: Retrieval-Augmented Generation, Large Language Models, Institutional Knowledge Chatbot, Hybrid Retrieval, RAGAS, Chatbot Usability Evaluation

Introduction

Educational institutions receive hundreds of repetitive questions each day regarding admissions, policies and procedures. A steady administrative burden on staff results from issues such as requirements for enrollment, allowable absence rules and other recurring problems (full fabric, 2023). The nature of static

websites and FAQ pages is that they are non-interactive and can not tailor answers to specific questions asked by users (Fakrudeen, 2025). This problem is exacerbated by the fact that 75 percent of student chatbot inquiries take place outside regular business hours, and 39 percent of students report difficulty finding institutional information even after actively searching for it (Peyton et al., 2025; Schnieders & Bobek, 2024).

In addition to advancing natural language processing through large language models (LLMs), LLMs suffer in that their knowledge base is established during the model's initial training and therefore cannot be updated. Additionally, LLMs can generate information which may appear correct; however, such information could be false, or hallucinated (Gao et al. 2024). The problem still stands in modern generations of models. In the latest surveys, there is evidence of the presence of problems with hallucination classification and identification (Huang et al., 2024), as well as problems of temporal knowledge lag where the training limit of the model makes it miss out on any changes to the institution since then (Pęzik et al., 2025). These facts indicate that the problem of institutional grounding outlined by Lewis et al. (2020) has not been solved yet by modern model generations. To overcome these limitations of LLMs, Retrieval-Augmented Generation (RAG) retrieves relevant information prior to generating an answer from an external source(s) which has been previously validated (Lewis et al. 2020).

Most of the previous research has tested multiple systems but almost all of it has evaluated a single system for the testing of RAG-based chatbot educational architectures (Khasanova & Suh, 2025; Antico et al., 2024). Most of the research that have been done recently to test retrieval strategies and LLMs have been part of an evaluation that were conducted on a single institutional knowledge base and/or via statistical inference. A second, related gap is that domain-specific RAG systems are rarely benchmarked against the general-purpose chatbots (ChatGPT, Gemini, Claude) that students already use informally, even though these chatbots rely on parametric knowledge with no access to an institution's current, verified documents. Thirdly, although there are many automated measures of performance like RAGAS, they have rarely been paired with an end user usability assessment. As Sáiz-Manzanares et al. (2023), noted, higher technical performance does not necessarily equate to higher customer satisfaction

We address these gaps by developing and evaluating AskWhiz, a RAG-based chatbot for Mapúa Malayan Colleges Mindanao (MMCM), a private academic institution in Davao City, Philippines. Nine configurations combining three retrieval strategies (Basic RAG, Prompt-Engineered RAG, Hybrid RAG) with three LLMs (GPT-4o-mini, Gemini 2.5 Flash, Claude Haiku 4.5) were built over MMCM's handbook and admission documents. The goal of the research project was to achieve six specific objectives. First, we used RAGAS metrics to evaluate each of the nine designs. Second, we conducted statistical analyses to determine if there were any statistically significant differences in performance among the nine designs. Third, we determined which design was superior. Fourth, we integrated the superior design into the chatbot platform called AskWhiz. Fifth, after developing AskWhiz, we evaluated the usability and effectiveness as perceived by the users with whom it would interact (academic end-users). We did this through an evaluation questionnaire for AI chatbots. Sixth, we compared AskWhiz's ability to answer questions from our institutional queries to that of three other general-purpose AI chatbots (GPT-5.5, Gemini 2.5 Flash, Claude Sonnet 4.6) based on semantic similarity to those queries using RAGAS metrics.

Background and literature review

Retrieval Strategies in RAG Systems

The Basic RAG system performs a semantic similarity search using dense vector representations to retrieve document chunks based upon a user’s question or prompt (Lewis et al., 2020). This method is effective at answering straightforward user prompts but suffers from two key shortcomings. The first problem is that the outputted document pieces will likely not align well with what was asked. Secondly, Basic RAG does not allow for iterative reasoning when a user's query requires multiple steps. As opposed to modifying the

search component, Prompt Engineered RAG adds an additional layer to the prompting layer in order to make it better at producing high quality output. This new layer contains instructions to facilitate thinking in a "chain-of-thought," guidelines on how to handle uncertainty, and directives on the user's desired tone (Park et al., 2025). Hybrid RAG, on the other hand, combines two different types of retrieval models; one semantic and one lexical (BM25), using a weighted score fusion approach (Sawarkar et al., 2024). The reason for doing so is that Hybrid RAG can take advantage of each type of retrieval model depending upon whether the query being asked is looking for something with an exact match of keywords or if the query is asking for something that is related in some sort of semantic manner.

RAG-Based Institutional Chatbots

As Table 1 illustrates, there have been several studies examining the deployment of RAG systems as part of institutional chatbots. The studies discussed here demonstrate the promise of RAG deployments; however, they each have limitations regarding their architectural scope. For example, BarkPlug version 2 (Mississippi State University) demonstrated a mean RAGAS score of .96 when processing campus facility-related queries (Neupane et al., 2024); however, it evaluated only one possible retrieval methodology and utilized only one large language model (LLM). Additionally, the Unimib Assistant was assessed as being high usability; however, it also produced inaccurate responses during evaluation which emphasizes the need for precise development of RAG systems (Antico et al., 2024). Khasanova and Suh (2025) examined several possible RAG architectures and LLMs; however, their study occurred in a general domain admissions context and did not use statistical inference. Finally, none of the studies previously reviewed conducted statistical analyses to determine if the performance differences observed among the different configurations were statistically significant — the void this study seeks to address.

Table 1: Summary of Related Studies

Study Authors	Year	Focus	Relevance to this study
Lewis et al.	2020	Basic RAG framework combining parametric and non-parametric memory	Established semantic similarity search as the RAG baseline used across all three strategies
Neupane et al. (BarkPlug v.2)	2024	Single-configuration hybrid RAG chatbot for university queries	RAGAS = 0.96 on campus queries; one retrieval strategy and LLM only, no cross-config comparison
Antico et al. (Unimib Assistant)	2024	ChatGPT-based RAG chatbot for university students	High usability but occasional inaccuracies; single-system validation, no statistical testing
Khasanova & Suh	2025	Comparative evaluation of RAG pipelines for admissions chatbots	Multi-configuration comparison, but general-domain and without inferential statistics
Sawarkar et al.	2024	Blended RAG via dense-sparse fusion	Hybrid retrieval outperforms single-method approaches; no institutional grounding tested
Es et al. (RAGAS)	2024	Open-source automated RAG evaluation across five metrics	Adopted as the primary evaluation instrument in this study

General-Purpose Chatbots versus Domain-Specific RAG

The increasing informal use of general-purpose chatbots, such as ChatGPT, Gemini, and Claude for answering Institutional Questions by Students is interesting given that the systems do not have direct access to their Institution's current verified documentation at Inference Time (Gao et al., 2024; Pęzik et al., 2025). As a result, they may produce answers that sound credible but are incomplete, outdated, or incorrect. Recent flagship general-purpose releases — GPT-5.5 (OpenAI, 2026), Gemini 2.5 Flash (Google DeepMind, 2025), and Claude Sonnet 4.6 (Anthropic, 2026) — have substantially improved reasoning and knowledge-work performance since the early LLMs that originally motivated RAG research, raising the question of

whether the same institutional-grounding gap still persists in newer models. Previous studies comparing domain-specific RAGs to general purpose chatbots were mostly based on qualitative, demonstration-type evaluations (Khasanova & Suh, 2025; Neupane et al., 2024). No previous study has recently compared these two types of systems using a common, quantifiable evaluation method for institutional knowledge queries as described in Objective 6 of this research.

Evaluation Framework

RAGAS is a publicly available quantitative framework that will allow for the automatic evaluation of RAG based on five other measures, eliminating the necessity for manual annotation and providing objectivity and replicability (Es et al., 2024). In addition to RAGAS' use in evaluating semantic similarity as part of the four system comparative analysis, we also applied its Semantics Similarity measure - the Cosine Measure between the vectors representing the original answers and the responses produced by each of the four systems; because it does not require a retrieval step, it can be applied to both RAG-based systems and general purpose systems. Previous studies have employed RAGAS in conjunction with one-way ANOVA and Tukey's HSD (α=0.05) when comparing configurations (Abbas et al., 2024; Dietrich & Stubbert, 2025); therefore, the same statistical approach was utilized in our nine configuration comparison.

Statistical Methods for Multi-System Comparison

The nine RAG configurations were evaluated using one-way ANOVA and Tukey's HSD post-hoc testing due to the continuous nature of RAGAS scores and that our experimental design met the necessary conditions of normality and independence for a parametric test. Since the chatbots being compared had inherently different response behaviors (i.e., retrieval grounded vs. purely parametric) which could result in nonnormal or unequal variance similarity distributions, we replaced one-way ANOVA with the non-parametric Kruskal-Wallis H-test with Dunn's post-hoc test (Bonferonni corrected) for the four-chatbot comparative analysis. The Kruskal-Wallis test compares median ranks across the different groups without requiring normality and Dunn's test then identifies which pairs differed from one another while controlling family-wise error rates (Kruskal & Wallis, 1952; Dunn, 1964).

Usability Evaluation of Educational Chatbots

Automated methods like RAGAS assess quality in terms of retrieval and generation but do not reflect the experience of users interacting with a chatbot; past research has demonstrated that while better automated scores can be found when students are more satisfied with their interaction with a chatbot (Sáiz-Manzanares et al., 2023) there is no guarantee they will be. To address Objective 5 of this study, the Students' Perceived Satisfaction with the Use of Chatbot Survey (SPSUCS), which was initially designed for use in assessing student satisfaction related to using a veterinary consultation chatbot (Huang & Chueh, 2021), had been previously adapted and validated in a university setting for assessing student satisfaction with university chatbots. It had shown high levels of reliability (internal consistency) as an assessment tool (Sáiz-Manzanares et al., 2023).

Methodology

Research Design

We used a quantitative, experimental-developmental design with two independent factors — retrieval strategy and LLM selection — crossed into nine configurations. Development followed Agile Scrum across three phases and six sprints: Phase 1 (Experimental) comprised Sprint 1 Data Foundation, Sprint 2 Experimentation, Sprint 3 Evaluation, and Sprint 4 Statistical Analysis; Phase 2 (Developmental) comprised Sprint 5 System Integration; and Phase 3 (Evaluative) comprised Sprint 6 User and Comparative Evaluation, which addressed Objectives 5 and 6. An Input-Process-Output (IPO) model, shown in Figure

1, guided this transformation of institutional documents into a validated, deployed, and independently evaluated chatbot.

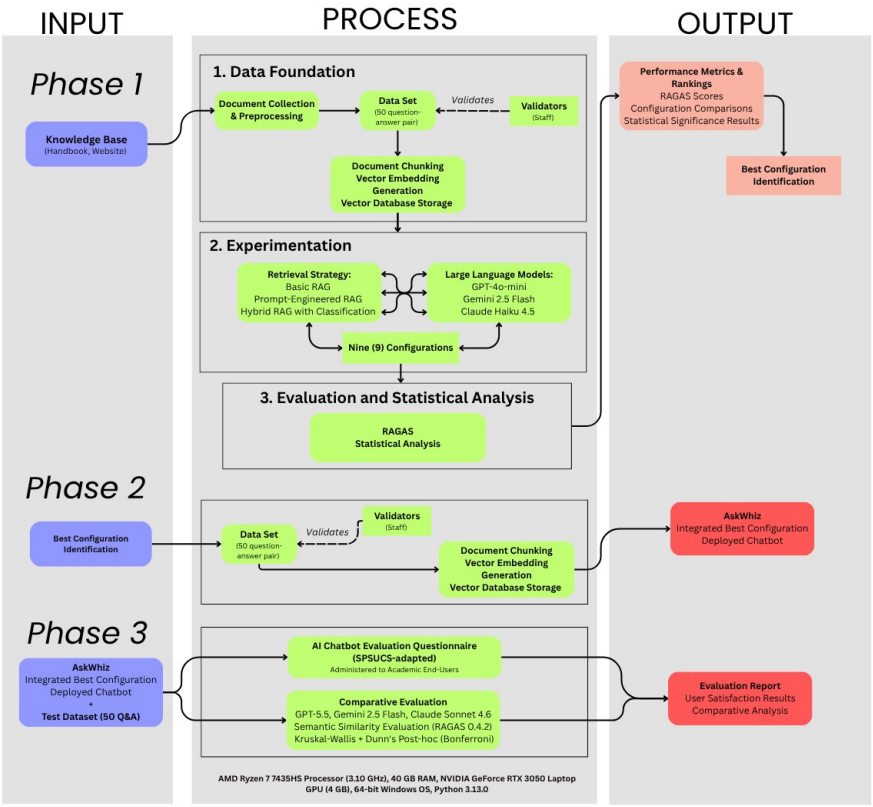


Figure 1. Conceptual Framework of the System Development.

RAG Configurations

Three retrieval strategies of increasing complexity were tested. Basic RAG uses dense-only FAISS retrieval to return the top 5 most similar chunks with a minimal prompt. The prompt-engineered RAG utilizes a similar retrieval process as prompt-free RAG and incorporates an additional chain of thought prompt, an output format directive, an uncertainty directive and an institutional tone directive. The hybrid RAG utilized a combination of the faiss dense search method (60%) and bm25 sparse search method (40%) using a weighted score fusion approach. Each methodology was used to pair each of the three large language models (LLMs) we tested—Gpt-4o-mini (openai), Gemini 2.5 flash (google), and claude haiku 4.5 (anthropic)—that were chosen due to being production-ready, cost effective, and have high performance on the relevant benchmarks. This yielded the nine combinations that are shown in Table 2.

Table 2. RAG Configuration Design

Config.	RAG strategy	LLM
C1	Basic RAG	GPT-4o-mini (OpenAI)
C2	Basic RAG	Gemini 2.5 Flash (Google)
C3	Basic RAG	Claude Haiku 4.5 (Anthropic)
C4	Prompt-Engineered RAG	GPT-4o-mini (OpenAI)
C5	Prompt-Engineered RAG	Gemini 2.5 Flash (Google)
C6	Prompt-Engineered RAG	Claude Haiku 4.5 (Anthropic)

Config.	RAG strategy	LLM
C7	Hybrid RAG	GPT-4o-mini (OpenAI)
C8	Hybrid RAG	Gemini 2.5 Flash (Google)
C9	Hybrid RAG	Claude Haiku 4.5 (Anthropic)

Knowledge Base and Dataset

The knowledge base was built by collecting MMCM's publicly available Student Handbook, Admission Policy, Enrollment Procedure, FAQs, as well as all institutional policies using Python libraries Beautiful Soup 4.12.0 and PyMuPDF 1.23.0. Documents were chunked with LangChain's RecursiveCharacterTextSplitter (512-token chunks, 50-token overlap) and embedded with OpenAI's text-embedding-3-small (1,536 dimensions) into a FAISS v1.13.2 IndexFlatIP index. A validated 50-question answer (question/answer) test item pool that covered four areas (factual, procedural, comparative, and policy) was developed using three methods; these were GPT-4 generated questions, existing FAQs, and hand-written questions. The correct answers for all questions were taken directly from source documentation and validated by an expert panel in the area of student services, at the Office of Student Services. Questions were categorized by type (factual, procedural, comparative, policy) to ensure topical coverage, but were not formally graded or stratified by difficulty; this constraint is discussed further in the Limitations section.

RAGAS Evaluation Metrics

Each of the nine configurations was assessed using version 0.4.2 of the RAGAS Framework. Each configuration received normalized scores on five evaluation metrics, defined in equations 1 through 5. In general, when comparing two or more configurations that have been evaluated against the same set of criteria, the configuration with the larger value will be considered superior.

Context Precision

Context Precision looks at the retriever's performance in ranking relevant chunks higher than the irrelevant ones. A score closer to 1.0 means that close to all chunks that were retrieved were directly relevant to the query.

$$Context\ \ Precision\ =\ \frac{\left(Number\ \ of\ \ relevant\ \ chunks\ \ retrieved\right)}{\left(Total\ \ chunks\ \ retrieved\right)} \quad (1)$$

Context Recall

Context Recall is a measurement of the extent to which chunks retrieved answer a query. A high level of recall means that the duration of the context retrieved was sufficient to give detail of the entire referent context piece; low recall indicates that there was some important detail in the reference context that was not retrieved.

$$Context\ \ Recall = \frac{\left(Claims\ \in\ reference\ \ supported\ \ by\ \ retrieved\ \ context\right)}{\left(Total\ \ claims\ \in\ reference\right)} \quad (2)$$

Context Entity Recall

Context Entity Recall is an indicator of how much of the ground truth answer (e.g., office names, deadlines, institutional requirements, etc.) has been explicitly named in the retrieved context. This is a measure of assertions with regards to specific entities that can be detected whereas Context Recall addresses claims in general.

$$Context\ \ Entity\ \ Recall\ =\ \frac{\left(Common\ \ entities\ \ between\ \ retrieved\ \ context\ \wedge\ \ reference\right)}{\left(Total\ \ entities\ \in\ \ reference\right)} \quad (3)$$

Faithfulness

Faithfulness is a measure of how factually coherent a generated answer is with the retrieved context. This is vital for institutional contexts: the lower the faithfulness score, the higher the risk of the system hallucinating answers in its responses, which impacts the credibility of answers.

$$Faithfulness = \frac{\left(Claims\ \in\ response\ \ supported\ \ by\ \ retrieved\ \ context\right)}{\left(Total\ \ claims\ \in\ response\right)} \quad (4)$$

Answer Relevancy

Answer Relevancy is the degree of relevance of the answer to the user's question. Comparative processing between the User Question (UQ) and the Questions obtained from the Answer (QA) is averaged to obtain it. High (relevancy) means that the outcome matches the question being asked, low (relevancy) means they may be factually correct, but in some way incomplete or unrelated to the question asked.

$$Answer\ \ Relevancy\ =\ Average\ \ similarity\ \ between\ \ UQ\ \wedge\ QA \quad (5)$$

Overall RAGAS Score

The Overall RAGAS Score is the arithmetic mean of the five individual metric scores, enabling holistic comparison of configurations while maintaining interpretability through its constituent metrics.

$$Overall\ \ RAGAS\ \ Score\ =\ \frac{\left(Precision\ +\ Recall\ +\ Entity\ \ Recall\ +\ Faithfulness\ +\ Relevancy\right)}{5}$$

(6)

In each of the configurations, the knowledge base, embedding model, retrieval count (k=5), generation parameters (temperature: 0.1 and max tokens: 1024), and vector index were all fixed.

Statistical Analysis

Based on prior evidence that retrieval strategy affects RAG performance (Sawarkar et al., 2024; Gao et al., 2024), we hypothesized that overall RAGAS scores would differ significantly across the nine configurations. Formally, H0 stated that no significant difference in mean overall RAGAS score exists among the nine configurations, while H1 stated that at least one configuration differs significantly from the others. We performed a one-way ANOVA ($\alpha = .05$) to test whether overall RAGAS scores differed across the nine configurations. Where significant, we applied Tukey's HSD to all 36 pairwise comparisons, controlling the family-wise error rate.

Usability Evaluation Instrument

We addressed Objective 5 with an AI Chatbot Evaluation Questionnaire adapted from the SPSUCS (Huang & Chueh, 2021; Sáiz-Manzanares et al., 2023): 10 items on a 5-point Likert scale (1 = Strongly Disagree, 5 = Strongly Agree) spanning five dimensions — Chatbot Performance Accuracy (CPA), Perceived Benefit (PB), Chatbot Helpfulness and Usefulness (CHU), Motivation (M), and Overall Satisfaction (OS). Item wording was minimally adapted from a course-specific to an institutional-knowledge context. Internal consistency was assessed with Cronbach's alpha.

Population, Sample, and Sampling Procedure
The usability-evaluation population comprised MMCM academic end-users — students, faculty, and administrative staff who regularly interact with the institution's information systems. Thirty-one (31) respondents were recruited through convenience sampling, appropriate given AskWhiz's single-institution, closed-domain deployment and consistent with recommended minimums for survey-based usability pilots (Moran & Budiu, 2025) and with comparably sized prior institutional-chatbot usability studies (Qiu et al., 2025; Vanichvasin, 2022).

Respondent Selection Criteria
Inclusion required current enrollment at MMCM (any year level or program) or employment as MMCM faculty or administrative staff who interact with institutional information systems; participation was voluntary, with informed consent obtained in accordance with the Data Privacy Act of 2012 (Republic Act No. 10173). Respondents who did not complete the survey in full were excluded from analysis to preserve consistency across all ten SPSUCS items.

Comparative Evaluation Methodology
We addressed Objective 6 by submitting the same 50-item institutional test set to three general-purpose chatbots — GPT-5.5 (OpenAI, 2026), Gemini 2.5 Flash (Google DeepMind, 2025), and Claude Sonnet 4.6 (Anthropic, 2026) — under default settings, with no system prompt and no retrieval augmentation, simulating typical unassisted student use. AskWhiz was queried under its deployed C9 configuration (temperature = 0.1, max tokens = 1,024); maximum token output was the only parameter held equal across all four systems. Each response was scored against the ground-truth answer using RAGAS's Semantic Similarity metric with OpenAI's text-embedding-3-small as the embedding backbone. Differences across the four systems were tested with a Kruskal-Wallis H-test, followed by Dunn's post-hoc test with Bonferroni correction, at $\alpha = .05$.

Results and Findings

RAGAS Performance Across Configurations
Mean scores of the RAGAS metrics and their SDs are reported in Table 3. In almost all metrics, Hybrid RAG (C7-C9) outperformed both Basic and Prompt-engineered RAG. The most notable improvement was seen in Context Recall, where the score increased from 0.7120 - 0.7387 for C1-C6 to 0.9567– 0.9667 for C7-C9 (a 22%- 25% increase). This shows that the fusion of dense-sparse retrievals captured much of the reference context. Faithfulness was also best in Hybrid Context Precision was fairly stable across strategies (0.6942–0.7869), while Context Entity Recall was the weakest metric overall (0.2028–0.3607) for every configuration, indicating a chunking-level limitation — entities such as office names, deadlines, and thresholds are dispersed across sections of a document and are not reliably captured in a single retrieved chunk, regardless of retrieval strategy or LLM.

Table 3. Mean RAGAS metric scores and score variability by configuration

ID	Strategy	LLM	Ctx. Prec.	Ctx. Rec.	Ent. Rec.	Faith.	Ans. Rel.	Overall	SD
C1	Basic	GPT-4o-mini	0.7097	0.7270	0.2378	0.7889	0.6802	0.6287	0.3068
C2	Basic	Gemini 2.5 Flash	0.7081	0.7370	0.2028	0.7108	0.6105	0.5939	0.3060
C3	Basic	Claude Haiku 4.5	0.7097	0.7370	0.2383	0.8771	0.6707	0.6466	0.2761
C4	Prompt-Eng.	GPT-4o-mini	0.7124	0.7387	0.2486	0.7756	0.6842	0.6319	0.3132

ID	Strategy	LLM	Ctx. Prec.	Ctx. Rec.	Ent. Rec.	Faith.	Ans. Rel.	Overall	SD
C5	Prompt-Eng.	Gemini 2.5 Flash	0.7048	0.7120	0.2531	0.9419	0.6231	0.6470	0.2535
C6	Prompt-Eng.	Claude Haiku 4.5	0.6942	0.7279	0.2460	0.9277	0.6790	0.6550	0.2655
C7	Hybrid	GPT-4o-mini	0.7764	0.9567	0.3342	0.9334	0.8445	0.7690	0.1262
C8	Hybrid	Gemini 2.5 Flash	0.7654	0.9667	0.3607	0.9338	0.7832	0.7620	0.1304
C9	**Hybrid**	**Claude Haiku 4.5**	**0.7869**	**0.9645**	**0.3373**	**0.9758**	**0.7918**	**0.7713**	**0.1176**

Note. C6 and C9 reflect reduced valid samples (n = 49 and n = 47, respectively) due to the judge LLM's output token limit on four verbose test items; means use valid samples only. All other configurations were evaluated on the full 50-item set.

Analysis of Variance and Post-Hoc Testing

We performed a one-way ANOVA to test the null hypothesis that there are no differences in the RAGAS scores across the nine configurations. From this analysis, the F-ratio obtained was $F(8, 437) = 3.874$, $p = .0002$, thereby rejecting the null hypothesis and confirming that the design of the configuration matters. Using Tukey's HSD as a post hoc test revealed only three of the 36 pairs of comparison were statistically significant, where C2 (Basic RAG with Gemini 2.5 Flash) was compared with all Hybrids.No other comparisons — including all three within the Hybrid tier — reached significance ($p = 1.000$ for C7 vs. C8, C7 vs. C9, and C8 vs. C9), reflecting both the closeness of those configurations' scores and the limited statistical power of a 50-question dataset rather than true equivalence.

Table 4. Selected Tukey HSD Pairwise Comparison Results (α = .05)

Comparison	Mean difference	Adjusted p	Significant
C2 vs. C7	0.1752	0.0124	Yes
C2 vs. C8	0.1681	0.0201	Yes
C2 vs. C9	0.1774	0.0128	Yes

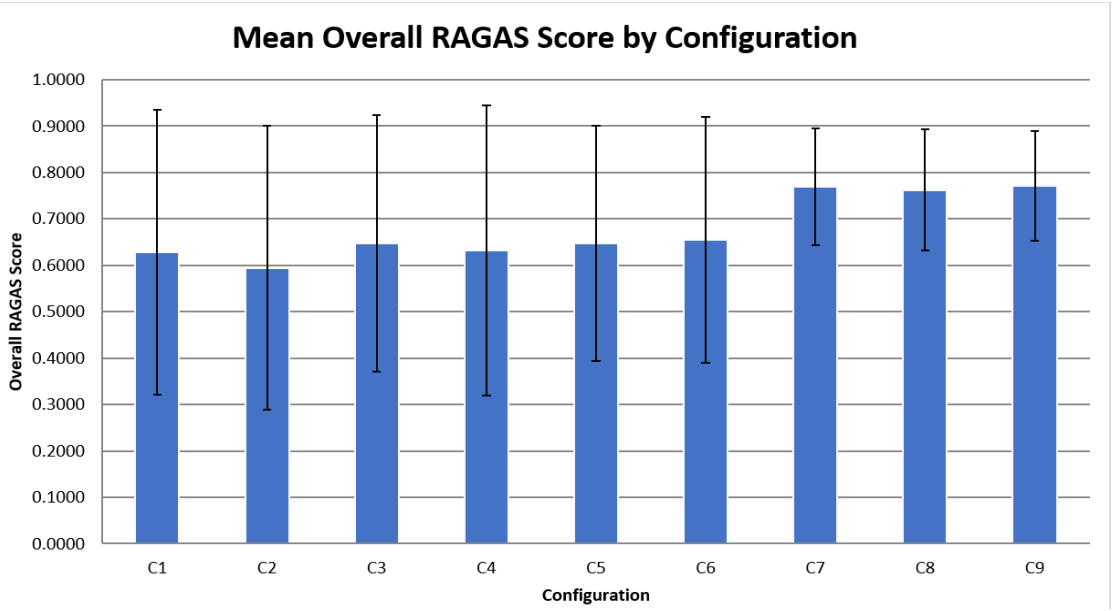


Figure 2. ***Mean Overall RAGAS Score by Configuration***
Note. Error bars represent standard deviation (SD). Asterisks (*) indicate configurations significantly different from C2 (Tukey's HSD, $p < .05$)

Identification of the Best-Performing Configuration

Since Tukey's HSD test failed to statistically differentiate between C7, C8, and C9, selection of the configuration was based on convergent evidence from all five metrics, along with the Overall RAGAS score and variance (Table 5). Configuration C9 (Hybrid RAGAS using Claude Haiku 4.5) had the highest Overall RAGAS score (0.7713), highest Faithfulness (0.9758), highest Context Precision (0.7869), and lowest variance (SD = 0.1176), and scored in the top two in each of these categories – no other configuration was able to do as well. Most weight was given to Faithfulness, since in the institutional context an unsubstantiated claim has repercussions for the student who uses it.

Table 5. Performance Profile of C9 — Best-Performing Configuration

Metric	C9 score	All-configuration rank
Context Precision	0.7869	1st (highest)
Context Recall	0.9645	2nd (C8: 0.9667)
Context Entity Recall	0.3373	3rd (C8: 0.3607; C7: 0.3342)
Faithfulness	0.9758	1st (highest)
Answer Relevancy	0.7918	2nd (C7: 0.8445)
Overall RAGAS Score	0.7713	1st (highest)
Score variability (SD)	0.1176	Lowest (most consistent)

Askwhiz System Implementation

The AskWhiz model was launched as a web-based chatbot using the C9 pipeline framework, with a modular three-layered structure, which included a retrieval layer made up of a 60% dense FAISS and a 40% sparse BM25 model, a generation layer based on Claude Haiku 4.5 and using Chain-of-Thought prompt restricted to institutional knowledge, and finally a FastAPI web layer setting a rate limit of ten requests per minute per IP, storing credentials in the environment variable, and automatically scaling at Render.

The C9 pipeline underwent a functional testing process by posing 25 different queries from the validated 50-query test set. The queries posed to the C9 pipeline covered all of the question types mentioned above, namely factual, procedural, comparative, and policy. Additionally, the 25 queries posed to the C9 pipeline covered the institutional service areas of admission, enrollment, tuition & payment, academic policy, and student services. These queries were sent automatically to the deployed pipeline API, and correctness was assessed through programmatic means, where a result was considered to be correct if it contained an HTTP status code of 200, which indicated success, along with any of the expected keywords or phrases based on the source document for that particular query (for example, the keywords/phrase expected for "How much is tuition?" were "34000," "45000," or "PHP"). However, two queries were not counted in the results due to being blocked by the rate limiter of the API (status code 429), despite retrying the queries after exponential backoff. This resulted in 23 queries to test the correctness of, of which 22 were correct (95.7%).

Usability Evaluation Of Askwhiz (Objective 5)

A total of thirty-one respondents from MMCM answered the modified version of the SPSUCS survey. There was high internal consistency among items ($\alpha = .87$). All dimensions were classified as Strongly Agree on the 5-point scale, with an average rating of 4.42 (Table 6). In terms of mean scores for each dimension, Chatbot Helpfulness and Usefulness (CHU) received the highest rating of M = 4.58 since the responses indicated a high level of agreement about the efficacy of AskWhiz in decreasing the waiting time for answering institutional questions, while the lowest rating of M = 4.32 was given to the Overall Satisfaction (OS) dimension. However, OS also fell into the Strongly Agree category, suggesting that there may be unfulfilled expectations despite overall satisfaction with AskWhiz.

Table 6. Usability Evaluation Results By Dimension (n = 31)

Dimension	Items	Mean	SD	Interpretation
Chatbot Performance Accuracy (CPA)	3	4.48	0.52	Strongly Agree
Perceived Benefit (PB)	4	4.36	0.53	Strongly Agree
Chatbot Helpfulness & Usefulness (CHU)	1	4.58	0.62	Strongly Agree
Motivation (M)	1	4.35	0.75	Strongly Agree
Overall Satisfaction (OS)	1	4.32	0.70	Strongly Agree
Overall (10 items)	**10**	**4.42**	**0.70**	**Strongly Agree**

Comparative Evaluation Against General-Purpose Chatbots (Objective 6)

The Semantic Similarity scores for AskWhiz and the three general-purpose bots on the same set of 50 questions are shown in Table 7 below. The highest average semantic similarity rating for the bot compared to the actual answer is demonstrated by AskWhiz (M = 0.7772, SD = 0.0789), and it also has the lowest standard deviation among the four bots, indicating consistency in the document-based answers. The best result among general-purpose bots was gained by Claude Sonnet 4.6 (M = 0.6052), followed by Gemini 2.5 Flash (M = 0.5575) and GPT-5.5 (M = 0.4619).

Table 7. Semantic Similarity Scores by Chatbot (n = 50)

Chatbot	Mean	SD	Min	Max
AskWhiz (C9)	**0.7772**	**0.0789**	**0.4627**	**0.8979**
GPT-5.5	0.4619	0.3184	0.0160	0.8710
Gemini 2.5 Flash	0.5575	0.1382	0.1357	0.8206
Claude Sonnet 4.6	0.6052	0.1266	0.3177	0.8124

We conducted a Kruskal-Wallis H-test to assess the statistical significance of the four systems; $H = 66.93$, $p < .0001$. The post hoc comparison using Dunn's procedure with Bonferroni correction showed that the system AskWhiz differed significantly from all three general-purpose chatbots ($p < .0001$ for all comparisons), whereas there were no significant differences between the three general-purpose chatbots ($p = 1.0000$ for all comparisons). This implies that the systems GPT-5.5, Gemini 2.5 Flash, and Claude Sonnet 4.6 did not differ statistically in their responses to institutional questions and performed worse than the domain-specific AskWhiz model.

Discussion of findings

The main source of performance differences was the type of retrieval strategy, not the LLM used. The dense-sparse fusion of Hybrid RAG achieved major performance gains in terms of Context Recall and Faithfulness through the capability of finding relevant context at both the lexical and semantic level, while Prompt-Engineered RAG barely exceeded the capabilities of Basic RAG, since sophisticated prompting is unable to replace the lack of relevant context (Park et al., 2025). Among Hybrid retrieval strategies, none of the LLMs made a difference ($p = 1.000$ for all pairwise comparisons), thus implying that once the retrieval is performed well enough, the performance of any of the LLMs is roughly equal.

The weakest metric of all configurations was Context Entity Recall (0.2028 – 0.3607); however, we consider it to be an issue related to chunking rather than retrieval, since the entities of interest are institutional entities such as names of offices, deadlines, and thresholds, scattered around tables, appendices, and policies in the document, and hence are usually included in separate chunks. This problem also became evident in the usability test results, as the most negative survey response was related to answering the most relevant questions.

Usability results align with the findings from the RAGAS metrics. Students strongly agreed with all usability items about the AskWhiz powered by C9 model ($M = 4.42/5.00$), indicating that the strong technical metrics have been reflected positively in the user experience. The comparative analysis also provides us with another piece of evidence: GPT-5.5, Gemini 2.5 Flash and Claude Sonnet 4.6 performed statistically identically, significantly lower than AskWhiz, when being used without any institutional grounding. Therefore, the institutional grounding gap, identified by Lewis et al. (2020) in previous RAG studies, persists in the current generation of general-purpose LLMs.

Implications for Research and Practice

For educational institutions

The hybrid RAG chatbot, built on the institution's official documentation, provides accurate, compliance-based answers without interruption. The model exhibits a high faithfulness score of 0.9758 for Criterion C9, indicating that almost all answers are supported by official documentation. In addition, the large gain in Objective 6 indicates that institutions have evidence that the domain-specific RAG chatbot is better than advising students to use commercial AI chatbots.

For researchers

It provides an example of a replicable, statistical methodology — the 3×3 factorial design used alongside the RAGAS evaluation method and inferential statistics, along with the usability test and a comparative objective against general-purpose chatbots — that could be adapted to other institutional settings (e.g., healthcare, government) where answering policy-related questions correctly matters.

For system designers
Since the retrieval strategy had a greater impact on the model's performance than the chosen LLM, optimizing the retrieval pipeline (e.g., hybrid fusion, future entity-aware chunking) would be the first step before choosing a better LLM. Since there were no significant differences among within-tier LLMs when retrieval was high-quality, designers could choose an LLM based on cost and latency.

Limitations

Limitations of the current study include: the study was conducted at only one institution, where the document structure, language, or policy complexity may vary. Secondly, the 50-item evaluation dataset used in this study although validated by experts is small in size and did not consider stratifying the questions using an empirical measure of question difficulty, which may contribute to some extent to the lower statistical power found; in this case, Tukey's HSD was able to show significance only between C2 and Hybrid tier, and Hybrid RAG configuration is to be considered as equal to each other. Thirdly, in all nine configurations, the same embedding model, chunking, and the number of retrievals were used, and therefore, any pattern seen, including Context Entity Recall, can be attributed partly to the use of the same infrastructure rather than the retrieval strategy itself. Fourthly, the evaluation of the generated models used only automatic RAGAS metric evaluation for the technical aspect, and no human evaluation for errors was made. In the same way, the function test after deployment of C9 pipe used automated evaluation through matching presence of keywords to the expected HTTP status code and reference, hence this can only be regarded as an operational test and not a correctness one. Fifthly, the usability evaluation was conducted with only 31 students using convenience sampling in a single guided session; no faculty or administrative staff subgroups were included. Lastly, the comparative evaluation was conducted on only three general-purpose chatbots using a single metric (Semantic Similarity) with non-retrieval default settings.

Conclusions

Nine RAG configurations were created and assessed for their suitability for use in an institutional knowledge chatbot application, before being extended in the evaluation process to include an assessment of user-perceived usability and a controlled comparison against general-purpose AI chatbots. From the results obtained, six conclusions can be drawn. First, the Hybrid RAGs performed better than both Basic RAG and Prompt-Engineered RAG on nearly every RAGAS criterion, with the retrieval strategy, not the particular choice of LLM, driving the better performance. Second, a one-way ANOVA found significant differences among the nine models ($F(8, 437) = 3.874$, $p = .0002$), with C2 differing significantly from all Hybrid RAG models according to Tukey's HSD test. Third, C9 (Hybrid RAG with Claude Haiku 4.5) turned out to be the best-performing configuration, with the highest overall RAGAS score (0.7713), the highest Faithfulness score (0.9758), and the lowest variability (SD = 0.1176). Fourth, C9 was successfully incorporated into AskWhiz, a working, operational institutional chatbot, where 22 of 23 non-rate-limited test queries worked as intended. Fifth, 31 academic end-users assessed AskWhiz as Strongly Agree ($M = 4.42/5.00$, $\alpha = .87$) on all five criteria of usability. Finally, a Kruskal-Wallis test showed that AskWhiz had statistically superior Semantic Similarity to institutional ground truth compared to GPT-5.5, Gemini 2.5 Flash, and Claude Sonnet 4.6 ($H = 66.93$, $p < .0001$); at the same time, the three general-purpose chatbots did not differ in their performance on this criterion. This is the first controlled, metric-based proof that a domain-specific RAG chatbot outperforms general-purpose chatbots at answering institutional knowledge questions. Further research should be conducted to optimize chunking in an entity-aware manner to reduce the Context Entity Recall score, incorporate qualitative evaluation, evaluate the chatbot in multi-turn

dialogue and longitudinal use, and compare it against more general-purpose chatbots using additional metrics.

Acknowledgements

The authors thank the faculty of the College of Computer and Information Science, particularly Mr. Christopher Josh L. Dellosa, Mr. Clyde Chester R. Balaman, and Ms. Genevieve A. Pilongo, for their constructive feedback, and the panel members — Dr. Cherry B. Lisondra, Mr. Martzel Baste, and Dr. Patrick D. Cerna — for their thorough evaluation and recommendations. Sincere thanks are due to the Information and Technology Office and the Office of Student Services of Mapúa Malayan Colleges Mindanao for their support in facilitating and validating this study, and to the 31 student participants in the usability evaluation.
The authors acknowledge the use of Grammarly and QuillBot for grammar checking and language improvement, and GitHub Copilot and/or Claude for coding assistance during system development. GPT-4 was additionally used to generate a portion of the evaluation dataset, as described in the Methodology. All research design, analysis, interpretation, and conclusions are solely the work of the authors.

References

Abbas, A., Rehman, M. S., & Rehman, S. S. (2024). Comparing the performance of popular large language models on the national board of medical examiners sample questions. Cureus. https://doi.org/10.7759/cureus.55991

Anthropic. (2025). Introducing Claude Haiku 4.5. https://www.anthropic.com/news/claude-haiku-4-5

Anthropic. (2026). Introducing Claude Sonnet 4.6. https://www.anthropic.com/news/claude-sonnet-4-6

Antico, C., Giordano, S., Koyuturk, C., & Ognibene, D. (2024). UNIMIB assistant: Designing a student-friendly RAG-based chatbot for all their needs. arXiv preprint arXiv:2411.19554.

Dietrich, N., & Stubbert, B. (2025). Evaluating adherence to Canadian radiology guidelines for incidental hepatobiliary findings using RAG-enabled LLMs. Canadian Association of Radiologists Journal. https://doi.org/10.1177/08465371251323124

Dunn, O. J. (1964). Multiple comparisons using rank sums. Technometrics, 6(3), 241–252. https://doi.org/10.2307/1266041

Es, S., James, J., Anke, L. E., & Schockaert, S. (2024). RAGAS: Automated evaluation of retrieval augmented generation. Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics, 150–163.

Fakrudeen, M. (2025). Evaluation of the accessibility and usability of university websites: A comparative study of the Gulf region. Universal Access in the Information Society, 24, 1883–1898. https://doi.org/10.1007/s10209-024-01160-9

Full Fabric. (2023). How universities are using chatbots to improve the student admissions process. https://www.fullfabric.com/articles/how-universities-are-using-chatbots-to-improve-the-student-admissions-process

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., & Wang, H. (2024). Retrieval-augmented generation for large language models: A survey. arXiv preprint arXiv:2312.10997.

Google DeepMind. (2025). Gemini 2.5 Flash & 2.5 Flash Image model card. https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-2-5-Flash-Model-Card.pdf

Huang, D.-H., & Chueh, H.-E. (2021). Chatbot usage intention analysis: Veterinary consultation. Journal of Innovation & Knowledge, 6(3), 135–144. https://doi.org/10.1016/j.jik.2020.09.002

Huang, L., et al. (2024). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. ACM Transactions on Information Systems, 43(2). https://doi.org/10.1145/3703155

Khasanova, M. Z., & Suh, Y. (2025). Design and performance evaluation of LLM-based RAG pipelines for chatbot services in international student admissions. Electronics, 14(15), 3095. https://doi.org/10.3390/electronics14153095

Kruskal, W. H., & Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. Journal of the American Statistical Association, 47(260), 583–621. https://doi.org/10.2307/2280779

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

Moran, K., & Budiu, R. (2025). How many participants for quantitative usability studies: A summary of sample-size recommendations. Nielsen Norman Group. https://www.nngroup.com/articles/summary-quant-sample-sizes/

Neupane, S., Jarold, A., Dhungana, S., McCann, J., Gautam, M., Wang, L., Gao, X., Vatsa, N., Bhusal, S., & Rawls, S. (2024). From questions to insightful answers: Building an informed chatbot for university resources. arXiv preprint arXiv:2405.08120.

OpenAI. (2026). Introducing GPT-5.5. https://openai.com/index/introducing-gpt-5-5/

Park, J., Atarashi, K., Takeuchi, K., & Kashima, H. (2025). Emulating retrieval augmented generation via prompt engineering for enhanced long context comprehension in LLMs. arXiv preprint arXiv:2502.12462.

Peyton, K., Unnikrishnan, S., & Mulligan, B. (2025). A review of university chatbots for student support: FAQs and beyond. Discover Education, 4(1). https://doi.org/10.1007/s44217-025-00397-7

Pęzik, P., et al. (2025). LLMLagBench: Identifying temporal training boundaries in large language models. arXiv. https://arxiv.org/abs/2511.12116

Qiu, W., et al. (2025). A systematic approach to evaluate the use of chatbots in educational contexts: Learning gains, engagements and perceptions. Computers, 14(7), 270. https://doi.org/10.3390/computers14070270

Sáiz-Manzanares, M. C., Marticorena-Sánchez, R., Martín-Antón, L. J., González Díez, I., & Almeida, L. (2023). Perceived satisfaction of university students with the use of chatbots as a tool for self-regulated learning. Heliyon, 9(1). https://doi.org/10.1016/j.heliyon.2023.e12843

Sawarkar, K., Mangal, A., & Solanki, S. R. (2024). Blended RAG: Improving RAG accuracy with semantic search and hybrid query-based retrievers. Proceedings of the International Conference on Multimedia Information Processing and Retrieval (MIPR), 155–161. https://doi.org/10.1109/MIPR62202.2024.00031

Schnieders, J., & Bobek, B. (2024). High school students gather college information from diverse sources. ACT Research. https://www.act.org/content/act/en/research/pdfs/R2413

Soliman, H., Kotte, H., Kravčík, M., Pengel, N., & Duong-Trung, N. (2025). Retrieval-augmented chatbots for scalable educational support in higher education. Proceedings of the 2nd Workshop on Generative AI for Learning Analytics (GenAI-LA).

Vanichvasin, P. (2022). Impact of chatbots on student learning and satisfaction in the entrepreneurship education programme in higher education context. International Education Studies, 15(6), 15. https://doi.org/10.5539/ies.v15n6p15