Athens Journal of Technology & Engineering 2025, 12: 1-18
https://doi.org/10.30958/ajte.X-Y-Z

GPT-only vs. GPT with RAG: A Study on Accuracy in Handling University-Specific Queries

*By Meltem Cakar**

While Large Language Models (LLMs) have demonstrated impressive capabilities in general natural language processing, their accuracy often diminishes in domain-specific contexts where precise, factual responses are crucial. This study addresses this limitation within the higher education sector by comparing two approaches to handling university-specific queries. We evaluate a Generative Pre-trained Transformers (GPT)-only model that relies on prompt engineering against a Retrieval-Augmented Generation (RAG) model that incorporates external university documents, specifically program flyers and a module handbook, integrated using Langchain. We benchmark both systems using 90 academic queries categorized by the question difficulty and assess their performance through automatic metrics and blind expert ratings. Our results demonstrate that RAG significantly outperforms the GPT-only approach, particularly for complex questions concerning curriculum and program structure. This research offers valuable insights for higher education institutions seeking to implement reliable and effective AI-powered solutions for student support and information provision.

Introduction

The field of natural language processing has seen significant advancements with the development of large neural network models trained on vast datasets, commonly referred to as Large Language Models (LLMs) (Wu et al. 2023). These models are designed to generate human-like text and have demonstrated a capacity for various language understanding and generation tasks, including question answering and summarization (Naveed et al. 2025). GPT models, such as those developed by OpenAI, represent a prominent family of LLMs known for their ability to generate coherent and contextually relevant text based on extensive pre-training (Kalyan 2025).

Despite their capabilities, LLMs have inherent limitations. A notable concern is 'hallucination', where the models generate content that is factually incorrect, irrelevant or inconsistent with the input data (Bang et al. 2023, Ji et al. 2023). This issue poses risks in environments like academia and education, where factual accuracy and reliability are important. Even if the information is plausible, it can still be misleading and lead to incorrect conclusions or decisions, particularly when influencing academic advising or information retrieval.

In response to the issue of LLM hallucinations (Waldo & Boussar 2024) and to improve factual accuracy, RAG has emerged as a promising framework (Liang et al. 2025). RAG combines the generative capabilities of LLMs with a retrieval

*Academic Assistant, DHBW Heidenheim, Duale Hochschule Baden-Württemberg Heidenheim - Cooperative State University, Germany.

mechanism that retrieves relevant information from verifiable external data sources prior to generating a response. This enables models to base their outputs on actual, human-authored data, thereby reducing the occurrence of inaccuracies in the generated content (Gao et al. 2024). In educational contexts, structured content like university program flyers and module handbooks represent a rich source of context-specific information that LLMs often ignore. Integrating such retrieval mechanisms enables LLMs to adapt more effectively to specific organizational contexts while mitigating the risk of factual errors.

The need for accurate and context-specific information is particularly pronounced in higher education institutions, where students and staff frequently seek detailed answers on curricula, program structures, and administrative procedures. Current LLM-based chatbots, when deployed without a robust retrieval mechanism, may struggle to provide the necessary precision academic advising and support systems. This highlights a critical gap in the reliable application of LLMs for domain-specific information retrieval within universities.

This paper investigates the application of RAG in the context of academic information retrieval within a university setting. Specifically, we benchmark a RAG-based system against a GPT-only model, evaluating their respective performances in answering program-related questions derived from authentic university documents. Our methodology involves a structured approach outlined in three key steps.

1. System preparation and data ingestion: Initially, a Retrieval-Augmented Generation (RAG) system was implemented using the Langchain[1] framework. This involved ingesting documents specific to DHBW Heidenheim, specifically flyers for the Business Informatics and Computer Science programs, as well as the official Business Informatics module handbook. These documents form the knowledge base for the RAG system.
2. Structured benchmarking with automatic metrics: A structured benchmark was then conducted, comparing the RAG system (developed with Langchain) with the GPT-only model. For this purpose, a curated dataset of 90 university-specific questions was generated from the mentioned DHBW documents. These questions were categorized into three difficulty levels: easy, medium and difficult (detailed breakdown in Document Sources and Question Development). Reference answers (gold standards) were established and validated for accuracy by two independent domain experts for all questions. The responses generated by both the RAG and GPT-only systems were then evaluated quantitatively against these gold standards using the following automatic metrics: F1-score, BLEU and METEOR. These metrics were selected due to their widespread acceptance and complementary strengths in evaluating the quality of text generation.
3. Complementary Human Expert Evaluation: In addition to the automatic metrics, a qualitative assessment was performed by three human experts. These experts independently evaluated the answers from both systems against the described gold standards, unaware of which system generated

[1]LangChain Developers: Building Chatbots. LangChain v0.x Documentation. Accessed on October 28, 2025 from https://python.langchain.com/v0.2/docs/tutorials/chatbot/

which response (a detailed description is provided in Human Evaluation and Qualitative insights). This step provided crucial qualitative insights into aspects such as factual accuracy and completeness.

This comprehensive Langchain-RAG framework and evaluation process established a structured benchmark, combining automatic metrics with invaluable human expert assessment. This multidimensional approach enabled a thorough analysis of answer quality from both systems, considering various perspectives.

Related Work

The field of natural language processing (NLP) has achieved significant progress, largely due to the development of large language models (LLMs). Trained on extensive text corpora, these models have demonstrated their ability to perform various NLP tasks, such as text generation, summarization and answering questions (Brown et al. 2020, Wu et al. 2023). Early iterations, such as GPT-3, demonstrated the advantages of few-shot learning, allowing models to perform new tasks with few examples or even through prompt engineering alone (Brown et al. 2020, Beltagy 2022). Despite their versatility, LLMs have limitations. The most notable of these is the phenomenon of hallucination, whereby models generate content that is factually incorrect, irrelevant or fabricated (Ji et al. 2023, Bang et al. 2023). This limitation poses significant challenges in fields requiring high levels of factual accuracy and reliability, such as education, healthcare and legal advice. In academic settings, for example, the generation of misleading or unverified information can undermine trust and hinder effective knowledge transfer.

To address these challenges, the RAG approach has emerged as a robust solution. RAG systems combine the generative capabilities of LLMs with a dynamic information retrieval component, enabling models to base their responses on external, verifiable data sources (Lewis et al. 2020, Gao et al. 2024). This approach reduces hallucinations and enhances factual consistency by anchoring outputs in real, human-authored content (Barnett et al. 2024). Research specifically on RAG in education highlights its potential to address these accuracy concerns. For instance, Calfoforo et al. (2024) investigated the integration of RAG and the Langchain framework to develop a question-answering system using the Llama-2 model. Their study aimed to improve information retrieval accuracy and relevance for policy-related questions based on a faculty handbook and FAQs in PDF format, demonstrating enhanced information accessibility and support efficiency within academic institutions. Similarly, Khan et al. (2025) developed an educational virtual assistant that leverages RAG with Llama-2 and Mistral models to provide university-related information. They evaluated the assistant's responses using BLEU scores and emphasized its potential for automating administrative support. Other studies have examined the use of RAG in generating precise answers from course materials. Furthermore, studies like Soygazi and Oguz (2023) have analyzed the performance of LLMs and Langchain-based models in mathematics education, highlighting the

challenges LLMs face in deterministic fields and the role of frameworks like Langchain in integrating specialized knowledge or tools to improve accuracy.

Building upon existing literature, including recent work by Calfoforo et al. (2024) and Khan et al. (2025) on RAG in educational contexts, this study addresses a critical research gap by providing structured benchmark. Previous studies have often focused on single RAG system implementations or general LLM comparisons. Our study compares a GPT-only model with a RAG system to identify their respective strengths and weaknesses when processing university-specific queries. Furthermore, we use a variety of university documents to assess performance and information complexity. We also use different quantitative metrics alongside a detailed, blind human expert evaluation.

Materials and Methods

Model Foundation

This study evaluates two LLM configurations: a GPT-only baseline model and RAG system. Both configurations are based on OpenAI's GPT-3.5 Turbo model. GPT-3.5 Turbo was chosen due to its optimal combination of linguistic capabilities in question-answering tasks. This model provides a robust foundation for both experimental setups, offering a cost-effective yet high-performing solution that is well-suited to research projects.

GPT-only Model Configuration

The GPT-only baseline model uses the GPT-3.5-turbo model provided by OpenAI directly. In this configuration, responses are generated based solely on the model's pre-trained internal knowledge. Prompt engineering was applied to guide the model's output, using a system prompt to direct its behavior and response style. The system prompt used was:

Figure 1. *System Prompt GPT-only*

> *You are an assistant for question-answering tasks. Use the following pieces of retrieved context to answer the question. If you don't know the answer, say that you don't know. Use three sentences maximum and keep the answer concise.*

RAG System Configuration

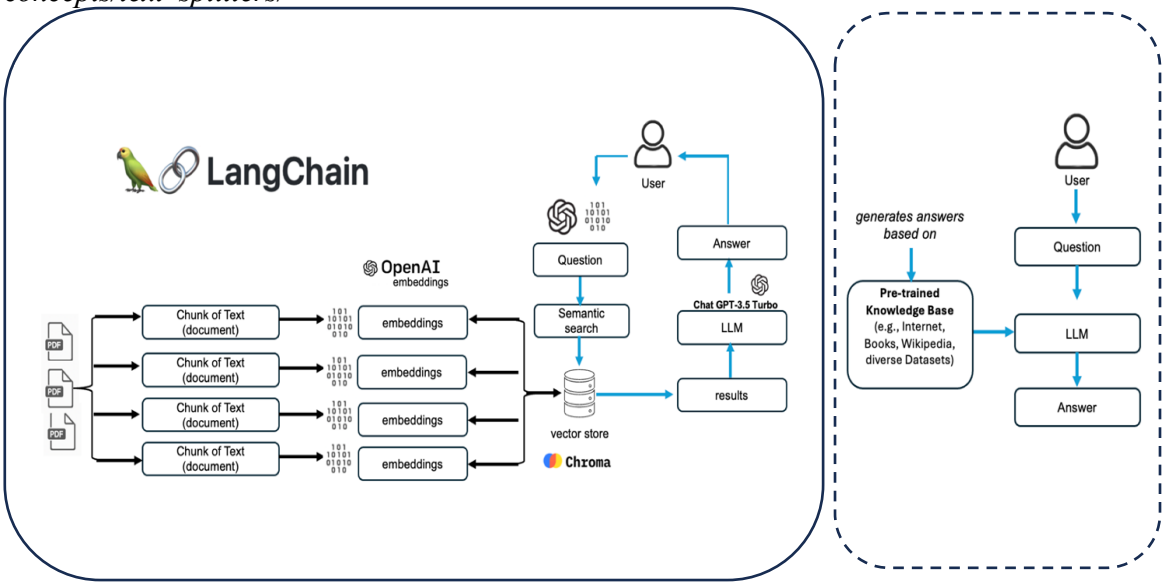
The RAG system also employs OpenAI's GPT-3.5-turbo but augments its capabilities with an information retrieval component. This system was implemented using the Langchain framework. The core components of the RAG (Langchain[2]) are detailed below.

[2]LangChain Developers: Building Chatbots. LangChain v0.x Documentation. Accessed on October 28, 2025 from https://python.langchain.com/v0.2/docs/tutorials/chatbot/

- Document ingestion and pre-processing (OCR): University-specific documents, including program flyers for Business Informatics and Computer Science and the official Business Informatics module handbook, were used as the knowledge base. As these flyers contained images and text, an Optical Character Recognition (OCR) process was applied to them first using the *Tesseract OCR* engine to accurately extract text from image based content within the PDFs. This ensured that all textual information, regardless of its original format, was available for processing. The extracted text was then prepared for the next step.
- Chunking: The pre-processed text from all documents was divided into smaller, semantically meaningful units, or 'chunks', to optimize retrieval accuracy. For this purpose, the *RecursiveCharacterTextSplitter* from Langchain was implemented with the following parameters: chunk_size=500 characters and chunk_overlap= 200 characters. This overlap helps maintain contextual continuity across consecutive chunks and avoids critical information being split.
- Embedding Model: To convert the text chunks into numerical vector representations (embeddings), an embedding model is essential. These embeddings allow for efficient semantic search. We utilized OpenAI's *text-embedding-ada-00* model for generating these embeddings.
- Vector Storage: The generated embeddings of the document chunks were then stored in a vector database. For vector storage and efficient similarity search, ChromaDB was chosen due to its ease of integration within the Langchain framework This vector store facilitates rapid retrieval of relevant document segments based on the semantic similarity to a user's query.
- Retrieval Method: When a user query is submitted, its embedding is generated and used to perform a similarity search within the vector store. The retrieval method identifies the most relevant chunks based on vector similarity. These retrieved chunks are then passed as context to the LLM.
- System Prompt for RAG: The same system prompt as in the GPT-only configuration was used to ensure consistent conversation style and response constraints (Figure 1).

Figure 2 provides a visual overview of both the GPT-only and the RAG system architectures, illustrating the distinct workflow of each approach. The diagram highlights how the RAG system augments the LLM's capabilities by integrating external document retrieval and a dedicated vector store, contrasting with the GPT-only model's reliance only on its internal pre-trained knowledge.

Figure 2. *Comparative Architecture of GPT-only and RAG System based on Langchain, own illustration based on Langchain: https://python.langchain.com/docs/concepts/text_splitters/*



Document Sources and Question Development

To simulate realistic student query scenarios and ensure the practical relevance of our evaluation, we compiled a comprehensive dataset of questions based on authentic university documents from DHBW Heidenheim. The primary sources of these documents were two program flyers (Business Informatics and Computer Science) and the official Business Informatics module handbook. These documents are typical sources of information for students seeking details on study options.

A total of 90 questions were manually developed from this content. To ensure a structured assessment of the models' capabilities across different cognitive tasks, these questions were systematically classified into three difficulty levels: Easy, Medium and Difficult.

- Easy level (remembering): Questions at this level target factual information that is stated directly within the source documents. They align with the 'Remembering' cognitive level of Bloom's Taxonomy (Bloom et al. 1956). Examples include questions about the length of studies, the number of credits, and the basic structure of dual study courses.
- Medium level (understanding): Questions categorized as 'medium' require a broader understanding and interpretation of the content of the document. These correspond to the 'Understanding' cognitive level of Bloom's Taxonomy (Bloom et al. 1956). Examples of such questions include identifying overarching program goals or highlighting the differences between similar study program based on their key characteristics.
- Difficult level (challenging questions) These questions were designed to go beyond direct recall or simple interpretation. Drawing upon information within the documents, they required comparison or a deeper understanding of

interrelationships. Examples include inquiring about learning outcomes across modules, the role of specific modules in competence development and drawing detailed comparisons between different study options. They were designed to evaluate the models' capacity to address more complex, academic questions that transcend the explicit levels of 'Remembering' and 'Understanding' as defined by Bloom's Taxonomy.

This multi-level classification system allowed us to assess how accurately each system could retrieve explicit answers and reflecting the diverse informational needs of a university setting.

Results

Quantitative Results by Document Type and Question Difficulty

The quantitative evaluation was conducted using a dataset comprising 90 questions, which were structured to simulate the diverse queries that students might bring to a university environment. The dataset was derived from three distinct document types representative of DHBW Heidenheim.

- Business Informatics Flyer (n=30 questions),
- Computer Science/Software Engineering Flyer (n=30 questions),
- Business Informatics Module Handbook (n=30 questions).

Each set of 30 questions for each document type was divided equally into three difficulty levels, based on the cognitive demands posed to the models.

- Easy (10 questions per document type),
- Medium (n = 10 questions per document type),
- Difficult (n = 10 questions per document type).

A precise reference answer (gold standard) was established for every question, ensuring a reliable basis for automated evaluation. The responses generated by the GPT-only and RAG-enhanced models were then assessed quantitatively using the F1-score, BLEU and METEOR metrics. Table 1 shows the performance metrics achieved by both models for different document types and question difficulty levels.

Table 1. *Model Performance in structured Benchmarks by Document Type and Question Difficulty*

GPT- only vs. RAG mean scores

Document Type	Model	Ouestion Difficulty	F1-Score	Bleu-Score	Meteor-Score
Business Informatics Flyer	GPT-only	easy	0.72	0.56	0.64
	RAG	easy	0.68	0.42	0.61
	GPT-only	medium	0.74	0.56	0.65
	RAG	medium	0.63	0.39	0.55
	GPT-only	difficult	0.66	0.39	0.53
	RAG	difficult	0.60	0.34	0.49
Computer Science Flyer	GPT-only	easy	0.70	0.62	0.63
	RAG	easy	0.96	0.94	0.95
	GPT-only	medium	0.66	0.52	0.58
	RAG	medium	0.94	0.88	0.90
	GPT-only	difficult	0.59	0.52	0.55
	RAG	difficult	0.90	0.84	0.86
Module Handbook	GPT-only	easy	0.62	0.60	0.64
	RAG	easy	0.97	0.89	0.92
	GPT-only	medium	0.58	0.41	0.48
	RAG	medium	0.93	0.87	0.94
	GPT-only	difficult	0.52	0.43	0.49
	RAG	difficult	0.88	0.81	0.88

Interpretation of Automatic Metric Scores

In order to interpret the results presented in Table 1, it is important to understand what high or low scores for each metric indicate in the context of question answering.

F1-score: Ranges from 0 to 1, with 1 indicating a perfect match between the generated answer and the reference answer (Chauhan & Daniel 2023). A higher score suggests that the model's response is both complete and concise.

BLEU: Also ranges from 0 to 1, indicating the n-gram overlap with the reference answer. A higher BLEU score indicates greater similarity in phrasing and word order to the ground truth, and better fluency. BLEU is particularly good at capturing exact matches of phrases (Chauhan & Daniel 2023).

METEOR (Metric for Evaluation of Translation with Explicit Ordering): It typically ranges from 0 to 1 and measures semantic similarity by considering synonyms, as well as exact word matches. A higher METEOR score indicates that the generated answer expresses the same meaning as the reference, even if different words are used (Lavie & Denkowski 2009).

Analysis of Results by Document Type and Difficulty

The quantitative results reveal varying performance differences between the GPT-only and RAG systems, depending on the document type and question difficulty.

Performance on the Business Informatics Flyer

For questions based on the Business Informatics Flyer, the GPT-only model outperformed the RAG system consistently across all difficulty levels. As shown in Table 1, the GPT-only model achieved higher F1, BLEU and METEOR scores for easy (F1: 0.72 vs. 0.68), medium (F1: 0.74 vs. 0.63) and difficult (F1: 0.66 vs. 0.60) questions. This unexpected outcome suggests that, for this specific document which is likely designed to contain widely known information (e.g. standard dual study advantages and general program descriptions), GPT-only's extensive pre-training enabled it to provide semantically similar answers without the need for external retrieval. The information required for these questions may have been sufficiently covered by GPT-only's pre-existing knowledge base, rendering the retrieval step less beneficial and potentially introducing minor noise, given that some of the information in the flyer is common knowledge.

Performance on the Computer Science/Software Engineering Flyer

By contrast, when questions from the Computer Science/Software Engineering Flyer were evaluated, the RAG system demonstrated a clear and significant advantage. RAG showed competitive performance for easy questions, but its strength became more apparent as the questions became more difficult. RAG achieved an F1 score of 0.94 for medium questions, compared to 0.66 for GPT-only. For difficult questions, RAG scored 0.90 (F1), compared to 0.59 for GPT-only. Similar trends were observed for BLEU and METEOR scores. This suggests that the Computer Science flyer, despite also being a 'flyer', probably contained more specialized or less commonly covered information that GPT-only found difficult to recall precisely from its general training data. This meant that it significantly profited from RAG's targeted retrieval.

Performance on the Module Handbook

The most significant performance difference of the RAG system was observed with questions from the Business Informatics Module Handbook. Across all difficulty levels, RAG consistently outperformed GPT-only. RAG achieved an F1 score of 0.97 for easy questions (compared to GPT-only's 0.62), 0.93 for medium questions (compared to GPT-only's 0.58), and 0.88 for difficult questions (compared to GPT-only's 0.52). These results strongly confirm the hypothesis that retrieval augmentation is critical for highly specific, detailed and voluminous domain knowledge. The module handbook contains intricate and unique curriculum details, learning outcomes and module descriptions that are unlikely to be fully or accurately represented in the general training data of GPT-only, thereby highlighting the ability of RAG to ground responses in precise information.

Examples of response Generation

To illustrate the quantitative findings and provide qualitative context for the performance differences, specific examples of model responses at different difficulty levels are presented below. These examples demonstrate how the RAG mechanism enables more accurate and contextually relevant answers, especially when dealing with domain-specific or complex information.

Example 1: Easy question (Business Informatics flyer)
Prompt: What are the advantages of alternating between theoretical and practical phases?
GPT-only completion: 'Theory is applied directly in practice, which increases learning success.'
Quantitative analysis: While this answer is semantically correct, it is also generic. GPT-only's F1, BLEU and METEOR scores for the easy Business Informatics flyer questions are comparatively high, (F1: 0.72, BLEU: 0.56, METEOR: 0.64) but reflect this general nature. A reference answer from the flyer might use more specific phrasing that is not precisely captured by GPT-only.
RAG completion: 'The combination of theoretical and practical elements ensures the course is both varied and practical.'
Quantitative analysis: The RAG-based response utilizes adjectives ('varied', 'practical') that are aligned with the specific terminology found in DHBW's official materials. Although the F1, BLEU and METEOR scores are slightly lower than those for GPT-only in this particular category, they demonstrate RAG's ability to ground answers in document-specific phrasing.
Example 2: Difficult question (computer science/software engineering flyer).
Prompt: In which field of study does the development of secure software belong?
GPT-only completion: 'The development of secure software is usually part of the degree program.'
Quantitative analysis: This response is highly unspecific and lacks the precise institutional detail necessary. The lower F1, BLEU and METEOR scores for GPT-only on difficult questions from the Computer Science Flyer (e.g. F1: 0.59) clearly reflect this generalization and the inability to provide accurate and specific information from its internal knowledge base alone.
RAG Completion: 'The Software Engineering course provides instruction on the development of secure software.'
Quantitative analysis: In stark contrast, the RAG-based response precisely attributes the topic to the 'Software Engineering' course, directly referencing the structure of DHBW's curriculum. This accuracy is a key driver for RAG's substantially higher scores (e.g., F1: 0.90) in this difficult category, showcasing its ability to retrieve and integrate specific domain knowledge.
These examples reinforce the quantitative findings, showing that while GPT-only can provide generic answers, RAG is essential for providing specific, accurate and contextually relevant information derived directly from university documents — especially for complex, domain-specific queries.

Human Evaluation and Qualitative insights

To complement the quantitative results presented in Quantitative Results by Document Type and Question Difficulty, a structured evaluation by human experts was conducted. This qualitative assessment aimed to evaluate the practical usefulness and factual accuracy of the systems' responses from the perspective of experienced academic advisors, identifying nuances that automated metrics might overlook.

Evaluation Setup

Three academic experts, all holding leadership positions as program directors in dual study programs (Business Informatics and Software Engineering) at DHBW, participated in a blind review process. Their deep familiarity with the university's curricula and official documents ensured a highly relevant assessment.

The evaluation focused exclusively on student query scenarios derived from the Business Informatics program flyer. This specific focus enabled an in-depth qualitative analysis, allowing direct comparisons to be made between system responses and the experts' domain knowledge. For each pre-determined query, the experts were presented with two anonymized answers: one from the RAG-based model and one from the GPT-only model. They were instructed to:

- compare each response against a human-annotated gold standard (previously defined based on official study documents).
- Assign a score from 0 to 2 for factual correctness and completeness (*0 = not fulfilled, 1 = partially fulfilled, 2 = fully fulfilled*).
- They were also asked to provide a concise written justification for each score, highlighting any missing information, inaccuracies, or overly generic language.

Findings and Expert Perspectives

The human expert evaluation of the questions on the Business Informatics program flyer revealed critical insights that complement and sometimes nuance the automatic metric results. Table 2 provides an overview of the average human ratings alongside the automated F1, BLEU and METEOR scores for this category of document.

Table 2. *Combined Automatic Metrics and Average Human Ratings for Business Informatics Flyer*

GPT- only vs. RAG mean scores

Document type	Model	Question Difficulty	human rating	F1-Score	Bleu-Score	Meteor-Score
Business Informatics Fyler	GPT-only	easy	0.45	0.72	0.56	0.64
	RAG	easy	0.71	0.68	0.42	0.61
	GPT-only	medium	0.55	0.74	0.56	0.65
	RAG	medium	0.57	0.63	0.39	0.55
	GPT-only	difficult	0.47	0.66	0.39	0.53
	RAG	difficult	0.48	0.60	0.34	0.49

Analysis of Combined Metrics for Business Informatics Flyer

The results of questions based on the Business Informatics Flyer indicate a dynamic performance landscape, with varying strengths for GPT-only and RAG depending on the difficulty of the question. This section critically examines the interplay between automatic metrics and human expert judgements.

Easy questions: For straightforward queries, the RAG system achieved a notably higher average human rating (0.71) than GPT-only (0.45). Despite GPT-only achieving superior scores in automatic metrics (F1: 0.72 vs RAG: 0.68; BLEU: 0.56 vs RAG: 0.42; METEOR: 0.64 vs RAG: 0.61), human experts perceived RAG's responses as more accurate and reliable, or more grounded in the content of the flyer. This discrepancy highlights that, for easy factual questions, human experts prioritize precise factual alignment over broader semantic coverage or stylistic fluency, which automatic metrics might favour.

Medium questions: For questions of medium difficulty, both models performed competitively in terms of human ratings. RAG achieved an average rating of 0.57, just above GPT-only's 0.55. However, GPT-only maintained a clear lead in all automatic metrics: F1 score (0.74 vs. 0.63 for RAG); BLEU score (0.56 vs. 0.39 for RAG); and METEOR score (0.65 vs. 0.55 for RAG). This suggests that, for moderately complex questions on this particular flyer, GPT-only's vast pre-trained knowledge enabled it to generate accurate answers that aligned well with general linguistic patterns.

Difficult questions: When it came to the most challenging questions related to the Business Informatics Flyer, the RAG system marginally outperformed GPT-only in terms of human ratings (0.48 versus 0.47). This is a notable finding, given that GPT-only had a consistent lead in terms of automatic metrics (F1: 0.66 versus RAG's 0.60; BLEU: 0.39 versus RAG's 0.34; METEOR: 0.53 versus RAG's 0.49). This suggests that, despite the concise nature of the flyer, RAG's ability to retrieve limited relevant context provided a slight edge in perceived accuracy and grounding by human experts for highly complex questions. Conversely, GPT-only's tendency to generalize or infer from its broader knowledge base, while achieving higher automatic metric scores for linguistic similarity, may have been perceived as less accurate human evaluators for these specific, demanding queries from such a concise document.

Figure 3 visualizes the individual expert scores for each question, providing a granular understanding of the expert agreement and discrepancies. This level of detail is essential for understanding the accuracy and consistency of the responses, showing where the models consistently met expectations and where they struggled in the eyes of the experts.

For *easy* questions, RAG responses tend to be highly consistent across experts, often receiving a score of 2, which indicates strong agreement on full fulfilment. In contrast, GPT-only responses to *easy* questions demonstrate greater variability and lower scores, particularly from Experts 1 and 3. This confirms the lower average human rating for GPT-only responses in this category.

While the average human ratings for *medium* questions are very close, Figure 3 reveals mixed expert opinions for both models. This indicates that neither model consistently achieved full agreement on high scores.

For *difficult* questions, the average human rating shows that RAG has a minimal lead over GPT-only. This is supported by instances where RAG's responses achieved a higher agreement on "partially fulfilled" or "fully fulfilled" scores, compared to GPT-only which received more "not fulfilled" scores (0) from certain experts. This suggests that, despite the challenges of concise summaries, RAG occasionally provided the critical information that GPT-only missed entirely for difficult queries.

Figure 3. *Expert Evaluation Scores by Question Difficulty for Business Informatics Flyer (0 = not fulfilled, 1 = partially fulfilled, 2 = fully fulfilled)*

Difficulty level	Question (Q)	Reference (R)	Answer Model	Expert 1	Expert 2	Expert 3	Answer Model	Expert 1	Expert 2	Expert 3
easy	Q1	R1	**RAG Answer**	**2**	**2**	**2**	**GPT-only Answer**	**2**	**2**	**2**
	Q2	R2	**RAG Answer**	**2**	**2**	**2**	**GPT-only Answer**	**2**	**2**	**2**
	Q3	R3	**RAG Answer**	**2**	**2**	**2**	**GPT-only Answer**	**2**	**2**	**2**
	Q4	R4	**RAG Answer**	**2**	**2**	**2**	**GPT-only Answer**	**2**	**2**	**2**
	Q5	R5	**RAG Answer**	**2**	**2**	**2**	**GPT-only Answer**	1	1	1
	Q6	R6	**RAG Answer**	1	1	1	**GPT-only Answer**	1	0	0
	Q7	R7	**RAG Answer**	1	1	1	**GPT-only Answer**	0	0	0
	Q8	R8	**RAG Answer**	1	1	1	**GPT-only Answer**	**2**	**2**	0
	Q9	R9	**RAG Answer**	1	1	**2**	**GPT-only Answer**	**2**	**2**	0
	Q10	R10	**RAG Answer**	0	0	**2**	**GPT-only Answer**	0	0	0

Difficulty level	Question (Q)	Reference (R)	Answer Model	Expert 1	Expert 2	Expert 3	Answer Model	Expert 1	Expert 2	Expert 3
medium	Q1	R1	**RAG Answer**	**2**	**2**	**2**	**GPT-only Answer**	**2**	**2**	1
	Q2	R2	**RAG Answer**	1	1	1	**GPT-only Answer**	**2**	**2**	**2**
	Q3	R3	**RAG Answer**	1	1	1	**GPT-only Answer**	**0**	**0**	**0**
	Q4	R4	**RAG Answer**	**2**	**2**	**2**	**GPT-only Answer**	1	1	1
	Q5	R5	**RAG Answer**	**2**	**2**	**2**	**GPT-only Answer**	1	1	1
	Q6	R6	**RAG Answer**	**2**	**2**	**2**	**GPT-only Answer**	**2**	**2**	**2**
	Q7	R7	**RAG Answer**	**2**	**2**	**2**	**GPT-only Answer**	1	**2**	1
	Q8	R8	**RAG Answer**	**2**	**2**	**2**	**GPT-only Answer**	**2**	**2**	**2**
	Q9	R9	**RAG Answer**	**2**	**2**	**2**	**GPT-only Answer**	**2**	**2**	**2**
	Q10	R10	**RAG Answer**	**2**	**2**	**2**	**GPT-only Answer**	**2**	**2**	**2**

Difficulty level	Question (Q)	Reference (R)	Answer Model	Expert 1	Expert 2	Expert 3	Answer Model	Expert 1	Expert 2	Expert 3
difficult	Q1	R1	**RAG Answer**	1	1	1	**GPT-only Answer**	**2**	**2**	**2**
	Q2	R2	**RAG Answer**	1	1	1	**GPT-only Answer**	**2**	**2**	2
	Q3	R3	**RAG Answer**	**2**	**2**	**2**	**GPT-only Answer**	0	0	0
	Q4	R4	**RAG Answer**	**2**	**2**	**2**	**GPT-only Answer**	**2**	**2**	0
	Q5	R5	**RAG Answer**	0	0	0	**GPT-only Answer**	0	0	0
	Q6	R6	**RAG Answer**	**2**	**2**	**2**	**GPT-only Answer**	**0**	**0**	0
	Q7	R7	**RAG Answer**	1	1	1	**GPT-only Answer**	1	0	0
	Q8	R8	**RAG Answer**	0	0	0	**GPT-only Answer**	**2**	**2**	0
	Q9	R9	**RAG Answer**	0	0	0	**GPT-only Answer**	**2**	**2**	0
	Q10	R10	**RAG Answer**	2	2	2	**GPT-only Answer**	1	1	1

The qualitative justifications provided by the experts further elaborate on these observations:

Expert 1 (Head of Degree Program Business Informatics): 'RAG provided answers that mirrored the structure of the flyer, particularly with regard to questions about learning outcomes and course flow. GPT-Base provided plausible, but occasionally inaccurate, information.'

Interpretation: This expert's comment reinforces RAG's strength in mirroring document structure. The human ratings suggest this structural alignment is highly valued for easy questions.

Expert 2 (Head of the Informatics Degree Program): 'For complex or curriculum-related queries in particular, RAG was clearly more reliable. GPT tended to generalize.'

Interpretation: This observation from Expert 2 aligns with the general benefits of RAG for specific, complex content, as confirmed by the Module Handbook results in Quantitative Results by Document Type and Question Difficulty. Even for the concise Business Informatics flyer, RAG's lead in human ratings for difficult questions suggests that its clarity for specific points are perceived as more valuable than GPT's generalizations.

Expert 3 (Head of Degree Program Business Informatics): 'The answers from RAG felt more grounded — they referenced specific content from the documents we use for guidance. GPT alone guessed too much.'

Interpretation: The emphasis on 'grounded' answers is crucial. For easy questions, this grounding gives RAG a clear human rating advantage. For medium and difficult questions, both models struggle to achieve consistently high human scores on this concise flyer. However, RAG's marginal lead for difficult questions suggests that its 'grounded' approach, even when yielding limited information, is slightly preferred to GPT's potentially less verifiable inferences or predictions.

These qualitative insights, when combined with the detailed visual representation of the individual expert scores, show that the performance differences between RAG and GPT-only depend highly on the nature of the source document and the specific demands of the query. This emphasizes the importance of analyzing specific use cases and document characteristics when designing LLM-based advisory systems.

Discussion & Educational Implications

This section provides a holistic interpretation of the quantitative and qualitative results presented in the Results Section, discussing their wider implications for the use of Large Language Models (LLMs) in higher education. The aim is to summarize the findings, explain the reasons behind the observed results, and derive practical recommendations for academic information systems.

Results and Interpretation of Key Findings

Performance on the Business Informatics Flyer: For questions derived from the Business Informatics Flyer, the GPT-only model demonstrated competitive performance and in some areas (automatic metrics and medium human ratings), superior performance compared to RAG. This suggests that, for documents containing broadly accessible information or relating to general concepts (e.g. the advantages of dual studies), GPT-only's extensive pre-trained knowledge base is often sufficient. The information required for these queries may be adequately represented in its training data, enabling it to generate accurate responses without external retrieval. The slight divergence, whereby RAG achieved higher human ratings for easy questions

despite lower automatic scores, suggests that human evaluators prioritize direct factual alignment, even if the fluency of the response is not optimized.

Performance on Computer Science/Software Engineering Flyer: In contrast, the evaluation of questions from the Computer Science/Software Engineering Flyer showed a clear and significant superiority of the RAG system across all metrics and difficulty levels. This difference is particularly pronounced for medium and difficult questions. This suggests that the content of this flyer likely contained more specialized or less commonly encountered information that was not as robustly represented in GPT-only's general training data. In this case, RAG's capacity to accurately retrieve specific details from the target document became critical, enabling it to provide accurate and contextually relevant answers where GPT-only struggled to recall correctly. Performance on the Business Informatics Module Handbook: The RAG system demonstrated the most significant performance improvements with questions derived from the Business Informatics Module Handbook. RAG performed better than GPT-only across all difficulty levels and metrics. This finding strongly supports the idea that retrieval augmentation is essential for highly specific, detailed and extensive domain knowledge.

Added Value and Educational Implications

The results of this study show that adding RAG to academic information systems is valuable, especially in higher education. Addressing LLM limitations and enhancing trust: A key challenge with GPT-only models is that they are liable to hallucinations and rely on static training data, which makes it difficult to control the recency and source of information. In dynamic educational environments where curricula and policies frequently change, this can result in the generation of outdated or inaccurate advice. RAG addresses this directly by enabling the integration of current institutional documents, such as up-to-date module handbooks or recently revised degree plans.

This on-demand update capability gives university administrators greater transparency and control over content, ensuring that AI-powered services provide reliable information. From a university's perspective, the ability to base responses on verified documents fundamentally enhances the quality of answers, fostering greater trust among students and staff. The study shows that a one-size approach to LLM deployment is not ideal. Instead, a differentiated deployment strategy is efficient. GPT-only models can be suitable for simple queries (e.g. contact details or general study facts) or questions related to widely available information (as demonstrated by the Business Informatics flyer). Their extensive pre-trained knowledge enables them to provide plausible and accurate responses. However, for complex, context-dependent queries that require high levels of precision and specific institutional knowledge (e.g. detailed curriculum planning, module competencies or unique program comparisons), RAG is essential. Its superior performance on the Computer Science flyer and the Module Handbook demonstrates its critical role in providing accurate responses in such scenarios. Qualitative expert feedback consistently reinforces this, highlighting RAG's ability to mirror the structure of the flyer and provide accurate answers.

Limitations and Future Work

Although this study provides valuable insights, it is subject to certain limitations. The evaluation was conducted using documents from a single university and focused on specific program types. Future work could involve scaling the system to support additional degree program by integrating a wider range of institutional content, including dynamic online sources (e.g. official web pages and news feeds), to ensure real-time content alignment. Additionally, it would be valuable to explore the impact of different chunking strategies or embedding models for diverse document types. Investigating user satisfaction and the long-term impact of RAG-based chatbots on administrative efficiency through different studies would also provide further practical insights.

Conclusion

This study systematically evaluated the accuracy of RAG against a GPT-only in handling university-specific queries, demonstrating the critical role of retrieval augmentation in educational question-and-answer scenarios. Through a structured benchmark involving 90 categorized questions derived from various university documents, our findings reveal the nuances of performance. While the GPT-only model proved sufficient for surface-level inquiries and questions related to broadly accessible information, RAG consistently demonstrated superior performance when deeper comprehension was required.

Our results confirm that integrating a document retrieval mechanism directly improves the accuracy of responses. This provides clear evidence that augmenting LLMs with retrieval capabilities is a significant technical advancement for educational institutions. By basing responses on verified official documents, RAG systems effectively reduce the risk of misinformation and promote transparent communication.

This paper provides a replicable blueprint for creating domain-specific chatbots for academic contexts. It illustrates how openly available documents can be systematically integrated into AI systems to create context-aware educational assistants. Future work will expand the system's knowledge base to include a wider range of institutional content, such as dynamic online sources. It will also explore integrating multimodal inputs to enhance the chatbot's capabilities as a comprehensive academic assistant.

Acknowledgements

The author would like to express their sincere gratitude to Prof. Dr. Anke Hutzschenreuter (DHBW Heidenheim) for her invaluable guidance and thoughtful suggestions during the development of this manuscript. Her support and critical feedback were instrumental in refining the work.

References

Asai A, Min S, Zhong Z & Chen D (n.d.) *Tutorial proposal: Retrieval-based language models and applications.*

Barnett S, Kurniawan S, Thudumu S, Brannelly Z & Abdelrazek M (2024) Seven failure points when engineering a retrieval augmented generation system. *Proceedings of the IEEE/ACM 3rd International Conference on AI Engineering – Software Engineering for AI*, 14 April 2024, ACM, pp. 194–199. doi:10.1145/3644815.3644945.

Bloom BS, Engelhart MD, Furst EJ, Hill WH & Krathwohl DR (1956) *Taxonomy of educational objectives: The classification of educational goals. Handbook I: The cognitive domain.* Philadelphia.

Brown TB, Mann B, Ryder N & et al. (2020) Language models are few-shot learners. *arXiv preprint* arXiv:2005.14165. doi:10.48550/arXiv.2005.14165.

Calfoforo JC & Raga RC (2024) Unleashing AI in education: A pre-trained LLMs for accurate and efficient question-answering systems. *2024 21st International Conference on Information Technology Based Higher Education and Training (ITHET)*, 6 November 2024, IEEE, pp. 1–6. doi:10.1109/ITHET61869.2024.10837606.

Chauhan S & Daniel P (2023) A comprehensive survey on various fully automatic machine translation evaluation metrics. *Neural Processing Letters* 55(9): 12663–126717. doi:10.1007/s11063-022-10835-4.

Fan W, Ding Y, Ning L & et al. (2024) A survey on RAG meeting LLMs: Towards retrieval-augmented large language models. *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 25 August 2024, ACM, pp. 6491–6501. doi:10.1145/3637528.3671470.

Gao Y, Xiong Y, Gao X & et al. (2024) Retrieval-augmented generation for large language models: A survey. *arXiv preprint* arXiv:2312.10997. doi:10.48550/arXiv.2312.10997.

Huang Y & Huang J (2024) A survey on retrieval-augmented text generation for large language models. *arXiv preprint* arXiv:2404.10981. doi:10.48550/arXiv.2404.10981.

Kalyan KS (2024) A survey of GPT-3 family large language models including ChatGPT and GPT-4. *Natural Language Processing Journal* 6: 100048. doi:10.1016/j.nlp.2023.100048.

Khan UH, Khan MH & Ali R (2025) Large language model based educational virtual assistant using RAG framework. *Procedia Computer Science* 252: 905–911. doi:10.1016/j.procs.2025.01.051.

Lavie A & Denkowski MJ (2009) The METEOR metric for automatic evaluation of machine translation. *Machine Translation* 23(2–3): 105–115. doi:10.1007/s10590-009-9059-4.

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

Li H, Su Y, Cai D, Wang Y & Liu L (2022) A survey on retrieval-augmented text generation. *arXiv preprint* arXiv:2202.01110. doi:10.48550/arXiv.2202.01110.

Liang H, Zhou Y & Gurbani VK (2024) Efficient and verifiable responses using retrieval augmented generation (RAG). *Proceedings of the 4th International Conference on AI-ML Systems*, 8 October 2024, ACM, pp. 1–6. doi:10.1145/3703412.3703431.

Naveed H, Khan AU, Qiu S & et al. (2025) A comprehensive overview of large language models. *ACM Transactions on Intelligent Systems and Technology* (online first). doi:10.1145/3744746.

Quidwai MA & Lagana A (2024) A RAG chatbot for precision medicine of multiple myeloma. *Genetic and Genomic Medicine*, preprint, 18 March 2024. doi:10.1101/2024.03.14.24304293.

Soygazi F & Oguz D (2023) An analysis of large language models and LangChain in mathematics education. *Proceedings of the 2023 7th International Conference on Advances in Artificial Intelligence*, 13 October 2023, ACM, pp. 92–97. doi:10.1145/3633598.3633614.

Vakayil S, Sujitha Juliet D, Anitha J & Vakayil S (2024) RAG-based LLM chatbot using Llama-2. *2024 7th International Conference on Devices, Circuits and Systems (ICDCS)*, 23 April 2024, IEEE, pp. 1–5. doi:10.1109/ICDCS59278.2024.10561020.

Waldo J & Boussard S (2025) GPTs and hallucination. *Communications of the ACM* 68(1): 40–45. doi:10.1145/3703757.

Zhao P, Zhang H, Yu Q, Wang Z, Geng Y, Fu F, Yang L, Zhang W & Cui B (2024) Retrieval-augmented generation for AI-generated content: A survey. *arXiv preprint* arXiv:2402.19473.

Zhong N, Qian Z & Zhang X (2021) Deep neural network retrieval. *Proceedings of the 29th ACM International Conference on Multimedia*, 17 October 2021, ACM, pp. 3455–3463. doi:10.1145/3474085.3475505.