GenAI-Powered College Advisor – University Support Navigator

1st Christian M Bridge
College of Information
University of North Texas
Denton, USA
Christian.Bridge@unt.edu

2nd Chandrasekhar Pedalapu
College of Information
University of North Texas
Denton, USA
ChandrasekharPedalapu@my.unt.edu

3rd Pranay Rishith Bondugula
College of Information)
University of North Texas
Denton, USA
PranayRishithBondugula@my.unt.edu

4th Ravi Varma Kumar Bevara
College of Information
University of North Texas
Denton, USA
ravivarmakumarbevara@my.unt.edu

5th Krishna Annavaram
College of Information
University of North Texas
Denton, USA
krishnaannavaram@my.unt.edu

6th Dr. Stephen Wheeler
College of Information
University of North Texas
Denton, USA
stephen.wheeler@unt.edu

***Abstract*—This study investigates the integration of Generative AI (GenAI) into academic advising systems, with the objective of automating university and department-specific questions, providing continuous, timely, and accurate responses to students. This technology is especially useful for universities during high-demand academic periods, addressing challenges such as limited staff availability that cause delayed responses to questions from prospective students, current students, and international students. Our systems are designed to assist students with departmental and institutional information for the College of Information Data Science Master's program and information from departments like Student Accounting, Financial Aid, Dean of Students, and more. Our methodology combines LlamaIndex SimpleDirectoryReader, SentenceSplitter chunking using HuggingFace BAAI/bge-m3, and vector storage via ChromaDB. These components support context-aware chunking information retrieval from institutional datasets and student query logs, which is then used to generate responses through Retrieval-Augmented Generation Large Language Models (RAG-LLMs) with Meta's Llama 3.3 (70B) and OpenAI GPT 4o (70B). To further enhance context relevance, a re-ranking mechanism using cosine similarities was applied after retrieval that shows significantly improved quality and reduced hallucinations.**

The system performance was evaluated by both human feedback and qualitative analysis via tools such as Evidently AI using 50 questions. Llama 3.3's Evidently AI metrics were an average BERTScore of 0.74, a mean semantic similarity of 0.84, a response correctness of 28%, a context quality of 88% valid, and a faithfulness of 68%. Llama 3.3's human evaluations were an average response fluency of 4.86, coherence of 4.80, relevance of 4.42, and factuality of 3.88.

GPT-4o outperformed Llama 3.3 with an average BERTScore 0.85. In terms of response time, GPT-4o averaged 8.85 seconds per query, while Llama took significantly longer, averaging 179.07 seconds per query.

Our results highlight some of the limitations of GenAI such as factual hallucinations and partial faithfulness, which can be overcome with enhancements to our models like better prompt engineering, improved context retrieval logic, more efficient data structuring and text preprocessing. By training on huge institutional data, with improvements such as re-ranking enhanced context selection and fine-tuning, the model can automate academic advising, however, in the current state, it would not effectively replace human advisors because information that may be hallucinated could give students misinformation, possibly impacting grades, graduation timing, and visa status. We suggest that by continuing to improve on the model architecture, the model could create better more accurate responses, and in it's current state is best used as an internal tool for staff to assist them in information retrieval, and possibly for use by students with the caveat that all information should be verified

by respective departments and advisors.

***Index Terms*—Academic Advising, college advisor, large language model, AI, artificial Intelligence, major selection, GPT, LLM, educational technologies, sustainable educational practices**

I. Introduction & Statement of Problem

The global educational market is rapidly expanding due to digital transformations, resulting in the elimination of traditional institutional boundaries and significant increases in international studies, virtual universities, and online learning platforms (Tahvildari, 2025; Ashmel, Tlemsani, & Matthews, 2021; Rodríguez-Abitia & Bribiesca-Correa, 2021). As institutions expand, it becomes challenging to handle the volume of questions and requests from students who struggle to navigate university policies and enrollment procedures during peak academic periods such as enrollment or graduation. In traditional question-answering practices, students depend on email, phone calls, scheduled advising appointments, and finding the information buried in a wealth of student resources, like documents or web pages, which can struggle to meet the scale and immediacy of student needs(Khan, Khan, & Ali, 2025). Departments are also likely to receive the same question from different students adding to the volume, which are answered manually by staff, resulting in departments being inundated with calls, e-mails, and chats, increasing expected wait times for a response. In addition, delays in responses, inconsistent guidance between departments, and limited advisor availability contribute together to decision-making delays and student frustration. These challenges are exacerbated in large universities where high student-to-staff ratios and diverse departments create barriers for students to receive timely, consistent, and accessible support in academic guidance and administrative assistance. Moreover, handling repetitive and routine inquiries for multiple students results in an overburden to support staff and leaves limited time capacity for complex and urgent cases (L. Nguyen & Quan, 2025). International students are further affected by these inefficiencies due to urgency and complexity of support - which impact the student's visa process, immigration, financial aid, scholarships, housing, and international pre-arrival procedures.(Dinh & Tran, 2023).

Initially, many institutions experimented with rule-based chatbots and static FAQ pages, due to the lack of adaptability, contextual understanding, and semantic depth; however, these models struggle to understand the context of a query and only respond to questions that exactly match their programmed logic and predefined rules (Khan et al., 2025). To overcome these limitations, Artificial Intelligence (AI) and Natural language Processing (NLP) emerged as a transformative technology in generative Artificial Intelligence (GenAI) particularly Large Language Models (LLMs) combined with Retrieval-Augmented Generation (RAG) methods has emerged as significant potential in delivering context-aware adaptive responses(Khan et al., 2025; Abdelhamid, Bangura, & Shah, 2025; Lekan & Pardos, 2025). By integrating LLMs, Gen AI and RAG-LLM methodologies with institutional data, students will be able to receive 24/7, scalable academic support, with quick, improved guidance and university-tailored responses.

II. Literature Review

The integration of the GenAI pipeline in academic advice has gained rapid attention in recent years due to its ability to understand context and produce human-like responses. Chatbots have evolved from simple rule-based to sophisticated AI models such as ChatGPT, Gemini, etc (Khan et al., 2025). In this section, we will conduct a review of prior literature from the field.

Earlier, chatbots experimented with rule-based models that are relatively simple and straightforward, with limitations in adaptability and contextual interpretation. Rule-based chatbots generate responses to queries by identifying matching patterns and struggle to provide institutional or personalized information without hard-coded logic (Khan et al., 2025). Dinh et al. identified that this approach opens up to critical inaccuracies when processing sentences and emphasized the need for context-aware solutions (Dinh & Tran, 2023). In contrast to rule-based models, LLM-driven systems demonstrate their competencies in delivering context-aware and semantically rich human advice responses. This can be achieved after extensive training of large-scale datasets using high-capacity neural networks or a transformer-based architecture. This evaluation makes more modern chatbots not rely on pattern matching rules. In Dinh and Tran's *EduChat: An AI-Based Chatbot for University-Related Information Using a Hybrid Approach*, they mention that despite advances in the fields of AI and NLP, chatbot capabilities are restricted to trained data – The model has to train on new data every time the policies are changed or new information is added (Dinh & Tran, 2023).

In *NEU-chatbot: Chatbot for admission to National Economics University*, NEU-chatbot was developed for prospective students and parents with admission inquiries related to National Economic University in Vietnam (T. T. Nguyen, Le, Hoang, & Nguyen, 2021). They built on the Rasa platform and achieved 97.1% accuracy on the test set which is practically applied to use in real-time admissions. They stated that the model was trained on university data and policies, which are to be trained every year and every time whenever the policies and data changed or updated.

Despite their advanced capabilities, LLMs are often face concerns over factual accuracy and can easily be prone to hallucinations – responses that inaccurate or misleading responses which are critical in academic or administrative settings. To mitigate these issues, Retrieval-Augmented Generation (RAG) frameworks come into picture and emerged as a potential solution combining generative models with retrieval-based mechanisms to consult data sources (L. Nguyen & Quan, 2025). In supporting that, Khan et al. integrated the vector embeddings with a document retrieval mechanism led to significant improvement in response correctness generated by LLMs (Khan et al., 2025).

Lekan and Pardos explored the application of ChatGPT in providing academic major recommendations to students. By their research, they found that the decision making is enhanced by using ChatGPT as a collaborative tool when integrated into a human-in-the-loop process (Lekan & Pardos, 2023). This finding suggests the promising role of GenAI in augmenting, rather than replacing academic advisor. From their research, Akiba and Fraboni argued that AI-powered advisors play a crucial role in reducing the unnecessary burden on staff and ensure students receive timely assistance (Akiba & Fraboni, 2023). Similarly, Al-Ghonmein et al. discussed the importance of integrating the human emotional intelligence and sentiment analysis with AI capabilities in a hybrid advising model,(Al-Ghonmein, Al-Moghrabi, & Alrawashdeh, 2023).

By their work, Huang et al. noted a substantial gain in student engagement and administrative efficiency by their GPT-based system which was tested on 100 academic scenarios. However, the author also highlighted the transparency especially in high decision making (Huang et al., 2021). Likewise, Huang et al. warned over the concerns like racial bias in AI recommendations and suggested set of safeguards such as transparency, fairness-aware algorithms and inclusivity of training data (Huang et al., 2021).

Samad demonstrated that combining the LlamaIndex's sentence-aware chunking with cosine similarity reranking leads to better context selection before LLM invocation, results in minimizing hallucinations (Samad, 2024). To enable modular modifications of institutional policies without compromising model logic, Abdelhamid et al. described the advantages of isolating front-end user interface from vector stores such as ChromaDB (Abdelhamid et al., 2025). To enable quick retraining or expansion into multilingual and domain-specific settings, this has been used in a number of institutions.

Lekan and Pardos introduced a hybrid RAG model named URAG for university admissions, to highlight the role of chunking and reranking in boosting accuracy and lower false information. Their approach significantly helps in this project development in fields of chunking (data preprocessing), and reranking (post-retrieval verification) (Lekan & Pardos, 2023). There are two mains approaches to increase the knowledge base of LLM. They are RAG and fine tuning. In this project we have done both; RAG integration and limited with partial fine-tuning of the LLM. The above literature collectively aids in implementing the rule-based system design to intelligent, context-aware, retrieval-enhanced GenAI agents.

III. Objectives of the Study

The primary objective of this research is to develop a RAG-LLM academic advisor capable of answering common questions about the College of Information Data Science program, departmental information, course requirements, and more. In addition, these studies investigate the technical and functional aspects of integrating LLMs into university support services by training on institutional data. The system design integrates sentence-aware chunking vector embedding via HuggingFace's BAAI/bge-m3, ChromaDB vector embedding storage, and reranking mechanisms to retrieve context-rich answers. The models are powered by high-performance LLMs such as Meta's LLaMA 3.3 and OpenAI's GPT-4o and are deployed via a local Streamlit user interface (UI) for accessibility. We followed both Human-in-the-loop evaluation for fluency, coherence, relevance and factuality, as well as qualitative tools like Evidently AI to determine context quality, faithfulness, BERTScore, and semantic similarity.

IV. Methodology

This study follows a modular AI pipeline, pictured in figure 1, powered by RAG and LLMs to build and evaluate a generative academic advising system. The methodology consists of six stages: Data acquisition and collection, data pre-processing, text chunking and embedding, vector database storage, query handling with re-ranking, and system evaluation.

label=0

1) Data Collection: We gathered institution-specific data from publicly available domains such as course catalog, university policies, degree requirements, FAQs, admissions information, transportation, and other department-specific pages. The data includes both structured and unstructured documents like .pdf, .json, .jsonl and .txt formats.
2) Data Preprocessing: The collected data is preprocessed by using standard natural language cleaning techniques such as removing some special characters from the text, cleaning up multiple blank spaces, deleting multiple period marks, and lowercasing all text. By incorporating Knowledge augmentation and sentence simplification techniques enhances the quality and usability of institutional text data and also optimized for downstream embedding and retrieval (Alva-Manchego, Scarton, & Specia, 2020).
3) Sentence-Aware Chunking and Embedding: The large text from each data source was segmented into document chunks using sentence-aware chunking. The SentenceSplitter from LlamaIndex is used in this system design with a maximum chunk size of 512 tokens. However, our collected data can be longer containing more than 5000 tokens. So, in order to maintain semantic similarity, we used an overlapping chunk technique of 50 tokens. These chunks are then embedded by using the HuggingFace model BAAI/bge-m3 from SentenceTransformers, which shows superior semantic retrieval performance in academic contexts (Reimers & Gurevych, 2019a; Karpukhin et al., 2020), resulting in high-dimensional vector representations.
4) Vector Database Storage: The high-dimensional vector embedding was stored in a scalable, open-source vector database named ChromaDB, which is optimized for fast similarity search. ChromaDB provides continuous storage and metadata tagging capabilities which enhances the traceability of retrieved response such as source document, and preprocessing details (*Chroma Documentation: Getting Started*, 2024).
5) Query Handling with Reranking: When user request a response or submit a query, it is first embedded using the same BAAI/bge-m3 model and compared with stored vectors in ChromaDB by using cosine similarities. In our system, we design in a way to retrieve the top k (k=5) similar chunks. We observed several sematic dissimilarity and hallucinations

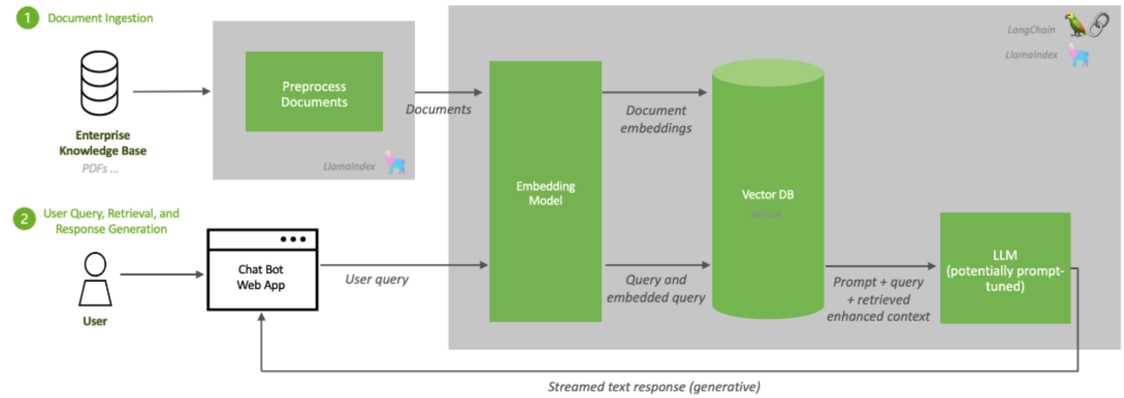

Fig. 1. Nividia Developer's RAG-LLM Architecture (Wolff, 2023)

in the generated responses/results. To address this, we introduced reranking step inspired by (L. Nguyen & Quan, 2025; Khan et al., 2025). This reranking module reorders the candidate chunks to improve the contextual relevance and helps in reduces hallucinations and enhance the faithfulness by categorizing the most relevant context fed into the LLM.

6) System Evaluation: System performance was evaluated by using the hybrid approach i.e., both human feedback and automated metrics. Human Feedback: In one of the methods, we collected the most common queries and responses from university human advisors, compared with generated responses. We rated accordingly based on relevance, correctness, completeness, and fluency. In another method, we asked groups of students to interact with the AI and rate the responses based on their satisfaction which provides us with a holistic evaluation. Automated Metrics: We used evaluation metrics like BERTScore to evaluate semantic similarity, F1 Score, EM (Exact Match) for Q&A performance, Evidently AI to analyze the correctness, context quality and faithfulness.

V. Data Collection

In order to develop a most reliable and context-aware system like academic advisor, a diverse amount of data has to gather from multiple sources while adhering to privacy and ethical standards. Generally, LLMs are inherently data-intensive technologies that demand substantial data and significant computational resources to train and maintain better accuracy and quick responsiveness. After considering these constraints, our data collection process intentionally focused on limited, domain-specific targeting institutions, most relevant content often accessed by students, such as information from the Data Science department, Transportation services and other relevant data. Even with a focused approach, it is challenging to collect information because universities often have a wealth of websites, FAQs, and resources to use to find information, however these information sources are subject to change per semester and are often unstructured. Additionally, there is no centralized FAQs for UNT, and its structure is completely different across departments.

label=0.

1) Institutional Data Sources: Primary source of our data obtained from official UNT websites, departments subdomains including:
   - Academic catalogs and program pages: (e.g., B.S., M.S., and Ph.D. in Data Science)
   - University-wide policies and student handbooks

- Frequently Asked Questions (FAQs): Sections across departments
- International student services: Including visa, health insurance, and orientation support
- Transportation Services and parking guidelines

We intentionally collected data from both academic and non-academic support domains to make the model capable of assisting with advising on administrative processes and campus services.

2) Data Formats and Acquisition Methods: Data were gathered by using web scraping tools and Python-based custom scripts and saved in a variety of formats such as:
   - Structured data in .json, .jsonl and .csv from dynamic web pages and open datasets.
   - Unstructured data in .pdf and .txt formats such as handbooks, policy guides, newsletters and Graduate Catalog PDF.

   Collected documents are pre-screened to ensure whether they are publicly accessible and free from confidential or sensitive content.
3) Query Logs and Common Inquiry Themes: A synthetic dataset of 50 frequently asked queries was framed for testing and simulating the real-world advising scenarios. They are commonly based on personal experience, common questions raised as a student during semester, and general FAQs related to course registration process, deadlines, Financial Aid, Student Accounting parking facilities and many.

VI. Exploratory Data Analysis (EDA) and Hypothesis For the Study

We had many different types of input documents and document length, including some documents that have extremely large amounts of tokens, which completely skewed our initial token box plots seen in figure 2. The average document token length was 1,414 with a standard deviation of 13,566, and after filtering out the documents with extremely large token lengths (List of Fall and List of Summer Classes), we can see a much closer distribution in figure 3.

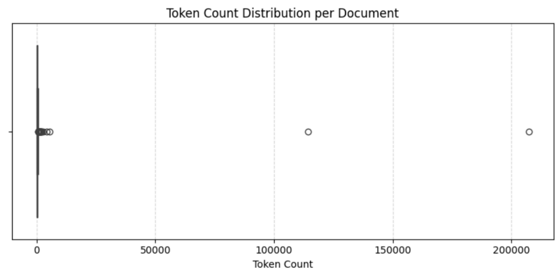

Fig. 2. Non-Filtered Box Plot of Document Text Tokens

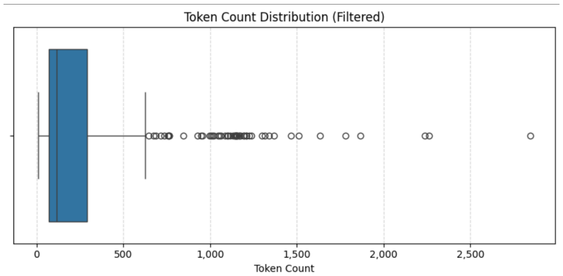

Fig. 3. Filtered Box Plot of Document Text Tokens

We can see in figure 4 after we create the document chunks of size 512 with a chunk overlap of 50, the breakdown of number of chunks per document type is heavily skewed to our excel spreadsheets which contained the largest amount of tokens. While this could seriously impact the results by overwhelming our vector database with similar text from the list of classes, we believe it is important to keep the data for class schedules, meeting times, professors, and building locations for all College of Information classes in the Summer and Fall semesters.

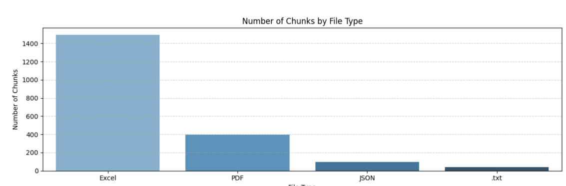

Fig. 4. Breakdown of Number of Chunks per Document Type

We can also see which words occurred the most in our chunks in figure 5. Many of these top words are repeated in the excel document of class schedule per class such as ”class”, ”description”, ”session”, ”start date”, etc. however we do not want to clean these words as they can have meaning in our dataset. Even ”words” like the character ”n” are used to indicate in the course schedule dataset that those specified classes are not in attendance on that day. In future iterations, it would beneficial to further specify the character ”n” means ”no”, or restructure the encoding for days the course is in attendance.

Our null hypothesis is that the information that our RAG-LLM generates will not be significantly different than information given by an academic advisor, or information that can be found from our data sources. The alternative hypothesis is that there will be a significant difference between the text generated by our models compared to the answers given by our academic advisor, and found in the university data sources.

VII. Data Analytics

This section focused on evaluating the effectiveness and reliability of the proposed GenAI-powered academic advising system across two dimensions: (i) the retrieval and response quality of large language models (LLMs), and (ii) the impact of reranking mechanisms in improving context relevance and reducing factual hallucinations. Both quantitative metrics and qualitative metrics (human feedback) have taken into consideration to assess the performance of the system with comparisons made between two LLMs: Meta’s LLaMA 3.3 (70B) and OpenAI’s GPT-4o.

label=0.

1) Evaluation framework: To ensure comprehensive analysis, the system was evaluated using a dual-mode retrieval comparison:
   - 1.1 Baseline vector search, and
   - 1.2 Rerank-Enhanced Retrieval – a post-retrieval filtering using cosine similarity.

   These dual mode concept provides a significant comparative analysis of post-retrieval filtering on generative accuracy and semantic relevance where each mode was evaluated on following metrics and tools.
   - BERTScore (Precision, Recall, F1) used to captures sematic similarity between the LLM generated response and human practices, particularly used to evaluate the effectiveness of the LLMs (Zhang, Kishore, Wu, Weinberger, & Artzi, 2019).
   - SBERT Cosine Similarity measures the embedding-level alignment by computing the sentence-level semantic closeness between context and output embeddings (Reimers & Gurevych, 2019b).
   - Evidently AI analyses the quality and reliability of the generated responses based on fluency, faithfulness, context validity, and risk of hallucination, results the deeper understanding, accuracy and trustworthiness of the retrieved content.
2) Human Evaluation: It is the complimentary evaluation criteria to automated metrics. A set of 50 student queries used with human responses have taken into considerations and rated each system generated responses on four qualitative dimensions:
   - Fluency - Clarity and naturalness of language
   - Coherence – Logical consistency and structure of the answer
   - Relevance - Alignment with the original user query
   - Factuality – Adherence to institutional facts and documentation

   Outputs from both GPT-4o and LLaMA 3.3 model are reviewed and rated separately on same queries.
3) Dataset Composition and Query Selection: Mixed data used in evaluation from the following areas/components such as, a set of 50 student queries from the common themes in academic advising interactions, publicly available Institutional documents, chunked content generated during responses, and finally RAG-based responses generated

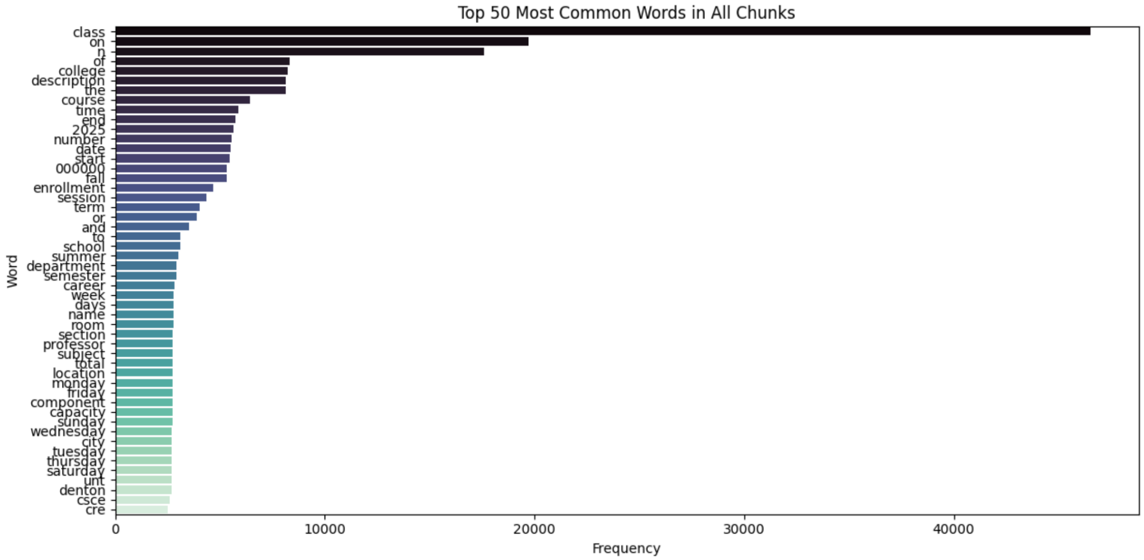

Fig. 5. Top 50 Words in the Document Chunks

by GPT-4o and LLaMA 3.3 under standard prompting conditions.

4) Pre-Evaluation Validations: Prior to the full-scale analysis, a series of sanity checks were implemented to reduce bias, uphold data integrity, and validate the reliability of the evaluation results. • Context Coverage Verification ensures that each test query successfully retrieved semantically relevant chunks from the vector database.

The results that we received from the model appeared to perform well with the data presented previously. We can see immediately that there is a difference in the response time between the locally run Llama 3.3 model, and the API run ChatGPT-4o model; With responses taking upwards of 22.47 seconds ChatGPT is able to return fast responses, whereas the Llama model can take upwards of 670.42 seconds to produce one response. In our academic advising appointment, which lasts 30 minutes, the academic advisor was able to answer 14 questions or about one per 128 seconds. We can see when comparing these timings that ChatGPT is able to answer these questions significantly faster, where as the Llama model would struggle to keep the pace.

Reviews of the model responses also indicate that chunk retrieval may be more effective in the ChatGPT model which appears to pull more diverse chunks, compared to Llama which appears to frequently pull chunks from the list of courses excel spreadsheet which, as indicated above, has many more representation based on chunks and word frequencies.

The following are chunks retrieved by both models for the following question.

Question 6: Can I take online courses in summer as an international student?

Llama Chunks Retrieved (truncated for brevity): 1: summer 2025 semester: summer 2025 college or school: libr college description: college of information department: lib and info science class subject: info class number: 5900 class: info 5900 course: info 5900 class section: 410 career: graduate course class description: special problems enrollment capacity: 30 total enrolled: 0 component: cre room number: internet days of the week: class start time: class end time: class on monday: n class on tuesday: n class on wednesday: n class on thursday: n class on friday: n class on saturday: n class on sunday: n class start date: 2025-06-23 00:00:00 class end date: 2025-07-25 00:00:00 professor name: kimjeonghyun summer session: 5w2 class location description: internet course - texas city: denton....

ChatGPT Chunks Retrieved (truncated for brevity): if not enrolled in classes. Student must maintain full-time status before the vacation and must intend to register for the following semester to work on campus full-time during summer break when not enrolled. You may not not work full-time during summer break if enrolled in classes. You do not need authorization to work on-campus. No special permission from International Student and Scholar Services is required to work on-campus. No special permission from immigration is required to work on-campus. You...

We can see from the above results that Llama's chunk retrieved does not appear to answer the question well, were as ChatGPT's retrieved chunks provide more meaningful information. While it may be important for future models to remove the list of classes to improve general question answering, we may also lose the ability of giving students class information, that we believe would be important for actual implementation.

VIII. Results

Our models were evaluated using multiple metrics and methods, human evaluation was conducted for fluency, coherence, relevancy, and factuality. As human evaluation results can often be difficult to obtain due to factors like time, cost, and requiring setup to distribute results or access to the LLM models, we also leveraged Evidently AI, a LLM evaluation tool that used ChatGPT to evaluate context quality, correctness, BERTScore, and semantic similarity.

Factuality is the evaluation of the model's response being true in relation to the source document or the ground truth answer (Do et al., 2024). This evaluation is used to ensure that the model is not hallucinating answers which when fluent and coherent can appear to be true, especially for users who are not an expert in the subject matter, like students looking for answers from a university chatbot. Fluency is the judgment by an individual if the "response [is] written correctly" and judges the quality of each sentence, whereas relevance is the judgment if "the response is relevant given the context", and selection of important content from the source (Mendonça, Lavie, and Trancoso (2024), p.1); Coherence is "the collective quality of all sentences" (Shen, Cheng, Nguyen, You, and Bing (2023), p.5). Each metric was ranked on a scale from 1 to 5, 1 being poor performance and 5 being excellent performance. Using these evaluations, we can determine if the responses generated by our models have high-quality sentences, are relevant to the context, and are factually correct. Due to time constraints, only one person reviewed evaluated each model; this limitation will be discussed further in the Discussion section.

Each model was given the same 50 questions for both human and Evidently AI evaluation. The questions were created by asking advisors within the UNT College of Information the most commonly asked questions they receive, as well as questions created based on the data sources collected; the following is a sample of the questions:

1) What are my course options?
2) Can I take all online courses as an international student?
3) What types of assistanship are there?
4) I am an international student with an F-1 visa, what are my US tax responsibilities?

1) Performance of Retrieval Strategies:

   The effectiveness of the generated response is evaluated by comparing on a baseline vector-only search and a reranking-enhanced pipeline. As discussed earlier, the reranking strategy reorder the top k (k=5) retrieved chunks based on cosine similarity between the query and each context embedding, followed by refining semantic alignment before generating the final response.

   As shown in Table I the reranking-enhanced approach consistently outperformed the baseline across all evaluated metrics. Specifically: Average BERTScore F1 indicating better overall semantic similarity between the generated and reference responses. SBERT Cosine Similarity, which reflects embedding-level alignment between the retrieved context and generated output. BERT Precision demonstrated higher context relevance. BERT Recall reflecting better coverage of the expected information.

TABLE I
PERFORMANCE METRICS FOR VECTOR VS. RERANK PIPELINES

Metric	Vector Only	Rerank Enhanced
BERTScore F1 (avg)	0.8439	0.8673
SBERT Cosine Similarity	0.5785	0.7122
BERT Precision	0.8352	0.8680
BERT Recall	0.8540	0.8744

2) **Human Evaluation Results:**
The human evaluations for both models' responses can be seen below in table II. Both models performed well in fluency and coherence; however, ChatGPT model did a superb job at creating fluent, coherent, relevant, and factual answers to the questions.

TABLE II
AVERAGE HUMAN EVALUATION SCORES (50 QUERIES)

Criteria	GPT-4o	LLaMA 3.3
Fluency	4.81	4.86
Coherence	4.79	4.80
Relevance	4.85	4.42
Factuality	4.91	3.88

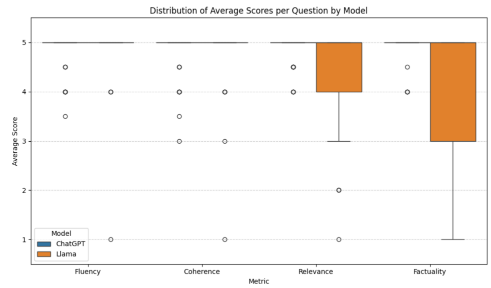

Fig. 6. Box plot for ChatGPT and Llama Model Human Evaluation Results.

We can see from the box plot in figure 6 that although the Llama model's overall fluency and coherence were consistently high, this could be coupled with responses that could be slightly relevant or lowly factual. We can also see that for the ChatGPT responses, that the ranges were much closer and response relevance and factuality were consistently high across all questions.

We can also review the results of individual questions to review the bottom 10 responses based on factuality, seen in 7, to further breakdown and analyze responses. It appears even when answers had lower relevance or factuality, the model still had higher fluency and coherence scores. This is another limitation we can discuss, as fluency and coherence may look acceptable to users, increasing perceived validity, while giving out misinformation.

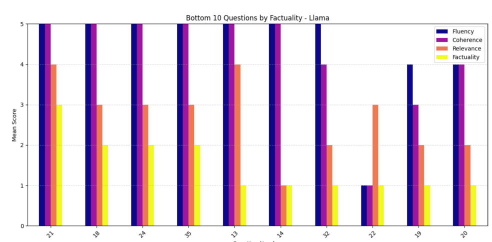

Fig. 7. Scores for Bottom 10 Factual Responses

Here we can examine question 14, to understand a response with high fluency and coherence but low relevance and factuality.

Question: Are certain courses offered in specific semesters?
Ground Truth: Yes, there is a rotation for the Data Science degree plan, some of the courses that are required for the degree but not in the college of information/DTSC there is less notice about courses such as business or college of engineering. https://datascience.unt.edu/students/data-science-course-rotation.html
Llama 3.3 Response (truncated for brevity): Based on the provided text, it appears that certain courses are offered in specific semesters. The text mentions various course sections with different semester offerings, such as: CSCE 5030 - Computer Systems (offered in Spring) CSCE 5040 - Operating Systems (offered in Fall) However, since this information is not present in the given text, we can look at the provided courses and their details...
ChatGPT's Response for comparison: Yes, certain courses may be offered in specific semesters. It is common for universities to schedule courses for particular terms/semesters each year. However, it's important to note that any course may be withdrawn from current offerings if the number of registrants is too small to justify conducting the course. For detailed information on when specific courses are offered, students should consult the university's course schedule or contact the relevant departmental advisors.

3) Hallucination and Faithfulness Analysis:
Using Evidently AI, the faithfulness and context validity scores were calculated as shown in below Table III and figures 8 and figure 9:

TABLE III
CONTEXT ADHERENCE EVALUATION USING EVIDENTLY AI

Model	Context Valid (%)	Faithful Answers (%)
GPT-4o	67	84
LLaMA 3.3	86	66

Compare to LLaMA, GPT-4o showed better faithfulness in retrieving context, even though LLaMA pulled more contextually valid chunks from the vector embedding database. This represnts, even GPT-4o may have pulled some unnecessary context, it is good at minimizing the critical factor named hallucinations.

Fig. 8. Evidently AI Evaluation of Llama 3.3's Context Quality & Faithfulness

Fig. 9. Evidently AI Evaluation of ChatGPT's Context Quality & Faithfulness

4) **Semantic Similarity:**
We can also see that the Llama responses, shown in figure 10, had a BERTScore mean of 0.87 with a standard deviation of 0.10, and an average semantic similarity score of 0.84 with a standard deviation of 0.09.

TABLE IV
SEMANTIC SIMILARITY METRICS

Model	BERTScore Mean	SBERT Similarity Mean
GPT-4o	0.87	0.82
LLaMA 3.3	0.74	0.84

Whereas the ChatGPT responses, shown in figure 11, had a BERTScore mean of 0.87 with a standard deviation of 0.08, and an average semantic similarity score of 0.82 with a standard deviation of 0.09. We interpret these values as the overall meaning of the responses closely resembled the ground truth (semantic similarity), but there may have been a token level difference between the words used in the response (BERTScore).

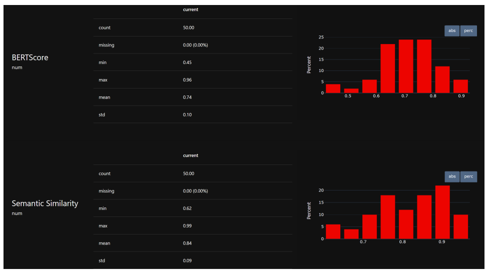

Fig. 10. Evidently AI Evaluation of Llama 3.3's BERTScore & Semantic Similarity

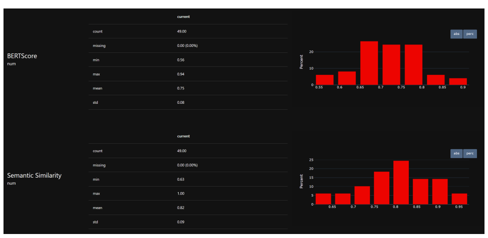

Fig. 11. Evidently AI Evaluation of ChatGPT's BERTScore & Semantic Similarity

5) Average Query Response Time Comparison:
The average response times to queries highlighted substantial efficiency differences. As shown in Table V Llama took significantly longer, averaging 179.07 seconds per query. We believe that this difference can be attributed to ChatGPT-4o being run via dedicated runtimes and an API, compared to the local installation of Llama 3.3. This difference showcases the benefits of running an API versus local installation, however there is an associated cost with using an API versus personal, dedicated hardware.
We believe at this time we would reject the null hypothesis that there is no difference between Llama and a human academic advisor; there are significant differences that are given between the academic advisor/ground truth that being incorrect on may cause severe issues for a student or lead to misinformation (website URLs, phone numbers, office locations, etc.). We believe that on the other hand with ChatGPT that there is no significant difference between the information it provides and the answers given by an academic advisor; however, it is important to note that information the models is trained on is constantly changing, and misinformation could be unnoticed until severe misinformation could be spread, so at this time it would be a tool that is best utilized internally by staff to assist in question answering, until it can be further refined for student facing operation.

TABLE V
AVERAGE QUERY RESPONSE TIME COMPARISON

Model	Avg. Response Time (seconds/query)
GPT-4o	8.85
LLaMA 3.3	179.07

IX. CONCLUSION & DISCUSSION

The GenAI-Powered College Advisor – University Support Navigator project focuses on upgrading the present, university-level support system to an AI-driven model. The primary objective of this study was to provide a scalable resource where students, faculty, or parents could resolve their queries 24/7, with minimal to no reliance on human assistance. This research mainly focuses on addressing common student-university issues such as delayed responses, streamlining common repeated questions to academic and administrative departments, emergency contacts, university policies, international student visa inquires, and more. Our project specifically caters to being able to answer the most, common queries received by the College of Information advising department from domestic and international students who seek precise guidance with quick, accurate support related to travel processes and more. Additionally, by automating the most common inquiries that may be sent by hundreds of students annually, the university support staff can dedicate more time to handling unique, emergency, or complex cases.

Our project showed that it is possible to receive fluent, cohesive, relevant, factual and semantically similar results from our Chat GPT-4o model with minimal response time. Our other model, Llama, displayed that it was also able to generate fluent, cohesive, and relevant responses, but that factuality could be variable, but semantically similar to the ground truth. We believe that both models have the ability to serve as an internal tool which could be validated and improved by domain-knowledgeable experts like academic advising staff, and with further improvements could be eventually deployed to be customer/student-facing.

Next, we discuss some of the limitations of our project. First, the cost of deploying a LLM model is either from needing the computational power to have dedicated server running the LLM, or needing to purchase API access which has a cost associated per token. While smaller models can also be deployed which can reduce API token cost, or computational resource needs, they may also drastically reduce the quality of response generated. Next, our evaluations were conducted with a very small sample for human evaluations, with only one person evaluating each model's response. With more time and funding, the models could be hosted and the streamlit UI distributed out for more user interaction beyond the created 50 questions, or a survey could be created to distribute to individuals to evaluate the questions, responses, ground truth, and retrieved context. Another issue is that information changes rapidly, especially at a University where courses being offered, webpages, FAQs, dates and deadlines change every semester and sometimes cannot be calculated years in advance. In order to maintain a RAG-LLM model a university would need a data pipeline or unification/standardization of resources to be able to automate the collection and preprocessing of data to be ingested by the model. In addition, the model could be further enhanced by being able to handle more types of data like video, audio, OCR, and better table parsing which could generate more coherent text from source materials, and add to information sources. Also, pre-processing is very difficult when you have information stored in characters like ”n” which could be removed with rudimentary text preprocessing like removal of stopwords or characters less than 3; to create a better model, we would need to spend more time analyzing the documents information parsed and possibly handle documents individually to ensure the best comprehensive text is extracted. Lastly, we could experiment with embedding models, chunking methods, and chunking sizes. If there are embedding models which deal with institutional data, we can create better vectors which are able to pull context fine-tuned for education and universities. With the chunking methods like Semantic Chunking, we may be able to create more coherent text chunks with smaller chunk sizes, which could make querying easier and more efficient.

As we have discussed above there are still improvements that can be iterated on our project, but that does not mean it performed poorly. RAG-LLMs are at the forefront of experimentation with LLMs and architectures continue to improve and evolve quickly. Most recently Ontology-Grounded RAG-LLMs (OG-RAGs) have been researched by Sharma et al. at Microsoft Research; the idea is that domain-expert created ontologies, structures of terminology that have relationships built within them, can enhance basic RAG-LLMs which only use a document set with no underlying structure (Sharma, Kumar, & Li, 2024). Sharma et al. indicate that using ontologies to bolster the document set create easier and faster context attribution, and better factual deductions as well, reducing hallucinations, and increasing transparency through better context attribution (Sharma et al., 2024).

REFERENCES

Abdelhamid, S., Bangura, J., & Shah, S. (2025, 03). Advisely: Ai-powered academic advising using large language models (llms). In (p. 1-8). Retrieved from https://www.researchgate.net/publication/389989528_Advisely_AI-Powered

_Academic_Advising_Using_Large_Language_Models_LLMs
Akiba, D., & Fraboni, M. (2023, 08). Ai-supported academic advising: Exploring chatgpt's current state and future potential toward student empowerment. *Education Sciences*, *13*, 885. doi: 10.3390/educsci13090885
Al-Ghonmein, A., Al-Moghrabi, K., & Alrawashdeh, T. (2023, 06). Students' satisfaction with the service quality of academic advising systems. *Indonesian Journal of Electrical Engineering and Computer Science*, *vol. 30*, 1838-1845. doi: 10.11591/ijeecs.v30.i3.pp1838-1845
Alva-Manchego, F., Scarton, C., & Specia, L. (2020). Data-driven sentence simplification: Survey and benchmark. *Computational Linguistics*, *46*(1), 135–187. Retrieved from https://aclanthology.org/2020.cl-1.4/ doi: 10.1162/coli_a_00370
Ashmel, M., Tlemsani, I., & Matthews, R. (2021, 09). Higher education strategy in digital transformation. *Education and Information Technologies*, *27*, 3171-3195. doi: 10.1007/s10639-021-10739-1
Chroma documentation: Getting started. (2024). Chroma. Retrieved from https://docs.trychroma.com/getting-started (Retrieved on 2025-05-04)
Dinh, H., & Tran, T. K. (2023). Educhat: An ai-based chatbot for university-related information using a hybrid approach. *Applied Science*, *13*(22). Retrieved from https://doi.org/10.3390/app132212446
Do, H. J., Ostrand, R., Weisz, J. D., Dugan, C., Sattigeri, P., Wei, D., … Geyer, W. (2024). *Facilitating human-llm collaboration through factuality scores and source attributions.* Retrieved from https://arxiv.org/abs/2405.20434
Huang, X., Jeon, H., Liu, J., Yao, J., Wei, M., Han, W., … Han, J. (2021, 12). Correction: Huang et al. deep-learning based label-free classification of activated and inactivated neutrophils for rapid immune state monitoring. sensors 2021, 21, 512. *Sensors*, *21*, 8360. doi: 10.3390/s21248360
Karpukhin, V., Oguz, B., Min, S., Wu, L., Edunov, S., Chen, D., & Yih, W. (2020). Dense passage retrieval for open-domain question answering. *CoRR*, *abs/2004.04906*. Retrieved from https://arxiv.org/abs/2004.04906
Khan, U. H., Khan, M. H., & Ali, R. (2025). Large language model based educational virtual assistant using rag framework. *Procedia Computer Science*, *252*, 905-911. Retrieved from https://www.sciencedirect.com/science/article/pii/S1877050925000511 (4th International Conference on Evolutionary Computing and Mobile Sustainable Networks) doi: https://doi.org/10.1016/j.procs.2025.01.051
Lekan, K., & Pardos, Z. A. (2023). Ai-augmented advising: A comparative study of chatgpt-4 and advisor-based major recommendations. *NeurIPS'23 Workshop on Generative AI for Education (GAIED)*, 1-11. Retrieved from https://gaied.org/neurips2023/files/41/41_paper.pdf
Lekan, K., & Pardos, Z. A. (2025, Mar.). Ai-augmented advising: A comparative study of gpt-4 and advisor-based major recommendations. *Journal of Learning Analytics*, *12*(1), 110-128. Retrieved from https://learning-analytics.info/index.php/JLA/article/view/8593 doi: 10.18608/jla.2025.8593
Mendonça, J., Lavie, A., & Trancoso, I. (2024). *On the benchmarking of llms for open-domain dialogue evaluation.* Retrieved from https://arxiv.org/abs/2407.03841
Nguyen, L., & Quan, T. (2025). Urag: Implementing a unified hybrid rag for precise answers in university admission chatbots – a case study at hcmut. *International Journal of Human–Computer Interaction*, 1-16. Retrieved from https://arxiv.org/pdf/2501.16276

Nguyen, T. T., Le, A. D., Hoang, H. T., & Nguyen, T. (2021). Neu-chatbot: Chatbot for admission of national economics university. *Computers and Education: Artificial Intelligence*, *2*, 1-6. Retrieved from https://doi.org/10.1016/j.caeai.2021.100036

Reimers, N., & Gurevych, I. (2019a). Sentence-bert: Sentence embeddings using siamese bert-networks. *CoRR*, *abs/1908.10084*. Retrieved from http://arxiv.org/abs/1908.10084

Reimers, N., & Gurevych, I. (2019b). Sentence-bert: Sentence embeddings using siamese bert-networks. *CoRR*, *abs/1908.10084*. Retrieved from http://arxiv.org/abs/1908.10084

Rodríguez-Abitia, G., & Bribiesca-Correa, G. (2021, 02). Assessing digital transformation in universities. *Future Internet*, *13*, 52. doi: 10.3390/fi13020052

Samad, M. (2024, October 27). *Top 10 breakthrough papers from icml 2024: Advancements in ai, machine learning, and security.* Medium. Retrieved from https://medium.com/@mirzasamaddanat/top-10-breakthrough-papers-from-icml-2024-advancements-in-ai-machine-learning-and-security-3b8f175afc4a (Accessed: 2025-05-04)

Sharma, K., Kumar, P., & Li, Y. (2024). *Og-rag: Ontology-grounded retrieval-augmented generation for large language models.* Retrieved from https://arxiv.org/abs/2412.15235

Shen, C., Cheng, L., Nguyen, X.-P., You, Y., & Bing, L. (2023). *Large language models are not yet human-level evaluators for abstractive summarization.* Retrieved from https://arxiv.org/abs/2305.13091

Tahvildari, M. (2025, 04). Implementing robo-advisory systems in virtual universities for smart student counselling. In (p. 203-218). doi: 10.58503/icvl-v20y202517

Wolff, H. (2023). *Rag 101: Demystifying retrieval-augmented generation pipelines.* Retrieved from https://developer.nvidia.com/blog/rag-101-demystifying-retrieval-augmented-generation-pipelines/

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2019). Bertscore: Evaluating text generation with BERT. *CoRR*, *abs/1904.09675*. Retrieved from http://arxiv.org/abs/1904.09675