Bridging Borders with Bots: Conversational AI for UTK Admissions

Seungwoo An
University of Tennessee
Knoxville, TN, USA

Devanshi Patel
University of Tennessee
Knoxville, TN, USA

Jihun Kim
University of Tennessee
Knoxville, TN, USA

Sai Deepika Dasari
University of Tennessee
Knoxville, TN, USA

Abstract

This project presents the development of a conversational AI chatbot aimed at providing accurate, timely, and student-friendly information to University of Tennessee, Knoxville (UTK) applicants, specifically international applicants. Students with unique obstacles, such as different time zones, unfamiliar terminology, and language barriers, often have limited access to critical resources. To solve this, we developed a scalable web scraping and Q&A generation and fine-tuning process applied to a pre-trained LM-based solution.

Our process started with scraping the UTK admissions website for clean, structured content. We used these data to generate high-quality question-answer pairs through the Claude API, which allowed us to form a domain-specific dataset that aligns with the questions potential students might have. Using Unsloth in Google Colab, we fine-tuned Meta's llama-3-8b-bnb-4bit model and then exported the trained model in GGUF format for local deployment with Ollama. A React-based front-end was created to give users a responsive real-time user interface to query the chatbot.

The fine-tuning method improves answer accuracy, as the language model learns from UTK-related data and can thus produce context-aware and factually correct responses with less hallucination. The initial assessment shows high accuracy, clarity, and usability results. While certain challenges—such as the absence of pre-existing datasets and limited training infrastructure—constrained the project scope, our end-to-end pipeline demonstrates a real-world, extensible, and modular approach to building AI-driven support systems in higher education.

1 Introduction

The university application process is daunting for potential students. The process is typically characterized by a vast array of forms, deadlines, qualification requirements, and school-specific procedures that vary depending on the background, program, and residency status of the candidate. For international students, these are added by geographical, language, and cultural barriers. As universities attempt to increase access and global reach, the demand for support systems that are reactive, scalable, and stable has become increasingly important.

On the University of Tennessee, Knoxville (UTK) campus, existing admissions support is primarily offered in the form of web pages, email messaging, and in-person or online advising. They are useful but not necessarily effective and timely, especially for geographically distributed learners or those who find accessing institutional websites difficult. This gap presents an opportunity to leverage conversational AI to deliver an improved information bridge and an inclusive and responsive admissions support experience.

This project takes into account the development of an AI chatbot to provide timely, relevant, and precise responses to questions about UTK admissions, particularly international students. The chatbot shall be a support system for the student, who could require support while making one of the most significant life choices they ever make. It is a value for institutions such as UTK in streamlining messages, reducing routine workload, and demonstrating dedication towards embracing digital advancements and inclusiveness.

The project aligns with broader trends in higher education as well, since more universities are adopting AI-powered solutions for tasks such as academic advising, enrollment support, and student engagement. Many of these possibilities, however, are generic or proprietary and expensive. Our objective is to provide an open-source, institution-centric, tailored solution that utilizes contemporary language models and retrieval systems to produce answers drawn from real UTK content—so they're accurate and institutionally spoken.

This project stands out by emphasizing local deployment and privacy, using tools like Ollama to run the model on local infrastructure. This not only ensures data security but also reduces the consumption of commercial APIs and gives more control over performance, cost, and customization.

2 Related work

Lambebo and Chen (2024) [3] examine AI chatbots in higher education, highlighting their role in admissions, counseling, and campus services. The study found that chatbots improve accessibility, improve student engagement, and reduce administrative workload. Inspired by this study, we developed a chatbot tailored specifically for admissions at the University of Tennessee, Knoxville, with a particular focus on creating a more convenient and engaging experience for international students.

Parthasarathy, Zafar et al. (2024) [5] fine-tuned LLM to significantly improve the chatbot functionality, improving contextual understanding, personalization, and efficiency. We adjusted the parameters of our model to provide answers about UTK admissions based on the dataset, and we plan to conduct more studies to provide more accurate and personalized answers.

Vakayil, Juliet et al. (2024) [6] found that LLM-based chatbots with Retrieval-Augmented Generation (RAG) can provide human-like conversations and even outperform humans in empathy. The chatbots also generate useful advice without judgment or bias. Considering that UTK's website is frequently updated, we are also considering introducing web scraping, as mentioned in the study.

Additional studies further contextualize the relevance of chatbots in higher education. Liu (2025) [4] discusses how generative AI can support international students by mitigating language and cultural barriers, reinforcing the importance of domain-aware and context-sensitive chatbot responses. Wang et al. (2023) [7] similarly examine how AI chatbots can deliver personalized academic support to international learners.

Karimi et al. (2024) [2] present a university help desk chatbot that handled over 20,000 international applicant queries, demonstrating efficiency gains from targeted deployment. Likewise, Day and Shaw (2021) [1] propose a GPT-2-based admissions chatbot evaluated using BLEU scores. These studies underscore the growing adoption of AI support systems in admissions and informed the technical and structural choices of our project.

3 Methods

This section outlines the complete methodological workflow we followed to design, implement, and deploy a domain-specific language model capable of answering admission-related queries for the University of Tennessee, Knoxville (UTK). The process was divided into four key phases:

- Dataset collection,
- AI-powered question-answer generation,
- Fine-tuning a pre-trained language model, and
- Integration with a front-end interface for user interaction.

3.1 Dataset Collection and Preparation

3.1.1 Focused Web Scraping

We used a focused web scraping approach to extract content directly from UTK's official websites, such as `https://international.utk.edu` and other departmental admissions pages.

To automate this, we developed a Python-based crawler consisting of three core functions:

- `get_main_nav_links(homepage_url)`: Parses the homepage to find internal navigation links.
- `scrape_page(url)`: Fetches HTML content and cleans it by removing scripts, ads, footers, and converting tables into plain text.
- `run_focused_crawler(start_url, max_pages)`: Controls the crawl and collects content from up to `max_pages`, storing structured data in CSV format (URL, Title, Content).

This approach ensured that the data collected was both relevant and reliable.

3.2 AI-Powered Question-Answer Generation

3.2.1 Initial Attempts and Challenges

We initially attempted to use Hugging Face's text-to-text generation models for Q&A creation. However, this method proved inefficient due to slow generation times and low-quality outputs.

3.2.2 Using Claude API

To improve output quality, we employed the Claude API for better context-aware question-answer generation.

Steps:

1. **Content Preprocessing**: Cleaned web pages and converted tables into readable text.
2. **Prompt Engineering**: Designed prompts to guide Claude in generating relevant FAQs based on webpage content.
3. **Structured Output**: Stored Claude's outputs in a CSV file with columns for Instruction (Question), Input (Context), and Output (Answer).
4. **Dataset Expansion**: Repeated the process with additional UTK webpages to ensure broader topic coverage.

3.3 Fine-Tuning the Language Model

3.3.1 Training Process

In the training process, we began by converting the CSV file into a format compatible with Hugging Face datasets. The CSV file was structured with three key components: instruction, input, and output. After preparing the file, we created a Hugging Face dataset that integrated with the Hugging Face API, allowing for efficient data handling throughout the training pipeline.

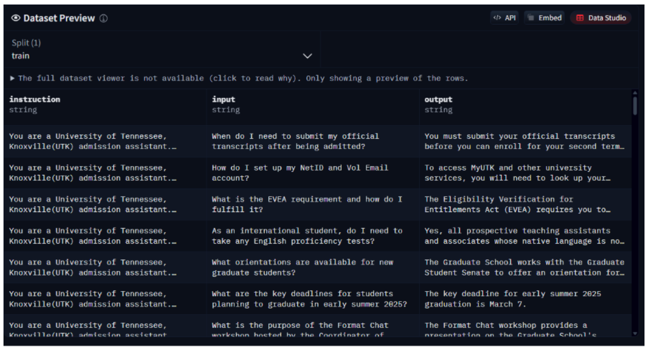

Figure 1: Hugging Face Dataset

After setting up the dataset, we proceeded to fine-tune the LLaMA model using the Unsloth framework, which is optimized for faster and more memory-efficient training of large language models. Fine-tuning involved adjusting several hyperparameters, such as learning rate, batch size, number of epochs, and sequence length. These settings were carefully selected to align the model's learning process with the size and complexity of our dataset. Hyperparameter tuning played a critical role in preventing overfitting while ensuring that the model learned meaningful patterns from the data. Unsloth's efficient training process made it possible to iterate quickly and monitor improvements during training.

Figure 2: Hyperparameter

Once training was complete, the final model was exported in Grok GGML Unified Format(GGUF), a binary format designed for efficient inference and fast loading in local environments. GGUF provides a compact and optimized representation of the model weights, which makes it well-suited for deployment in resource-constrained scenarios or when rapid response time is required. This format is widely supported by lightweight model runners and inference engines, allowing for easy distribution and local execution without the overhead of large frameworks.

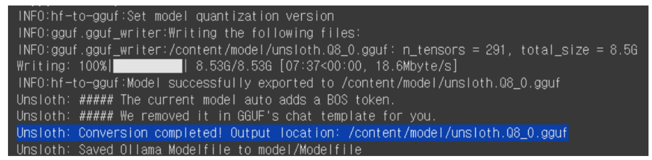

Figure 3: Model File Generation

3.4 User Interface and Integration

3.4.1 Ollama Integration

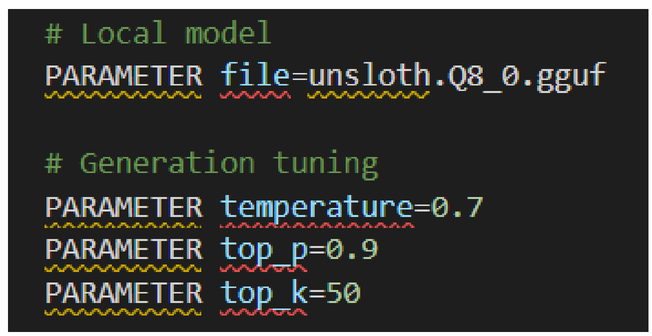

Figure 4: Parameter Fine-tuning

To provide a smooth and interactive user experience, we integrated the fine-tuned GGUF model with Ollama. Ollama is designed to simplify the deployment of GGUF models by managing model loading, memory usage, and inference speed efficiently. This setup makes it possible to use our model in production or development environments without the need for GPU-backed cloud infrastructure. Moreover, by adjusting the temperature, top_p, and top_k parameters as shown on 4, we can fine-tune the creativity, diversity, and relevance of its responses. Additionally, Ollama supports secure and customization configurations, giving developers control over how the model is served and how user interactions are handled.

3.4.2 React Front-End

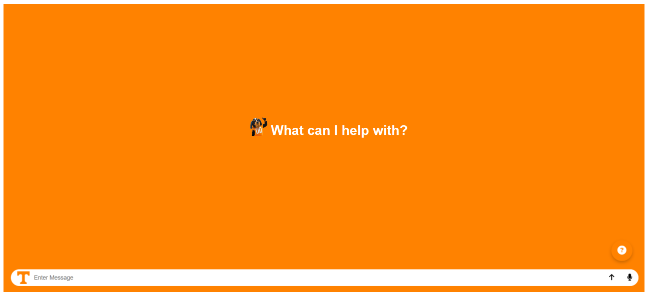

Figure 5: Chatbot Frontend

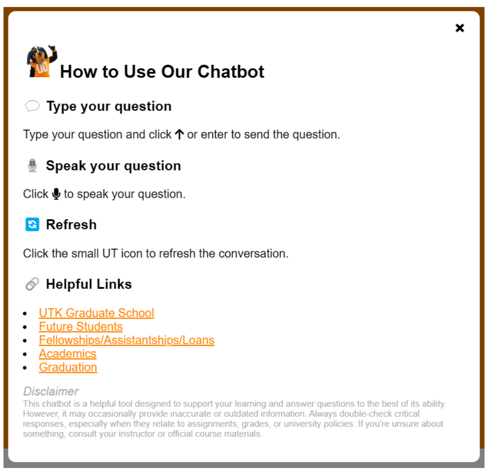

Figure 6: Chatbot User Guide

The front-end was developed using React, offering a simple and responsive chatbot interface that includes:

- Chat-style input/output interaction
- Dynamic loading of responses
- Optional source/context display
- Accessibility and mobile responsiveness
- Speech recognition for accessibility
- Chatbot User Guide, including how to use the chatbot and helpful links

3.5 Workflow

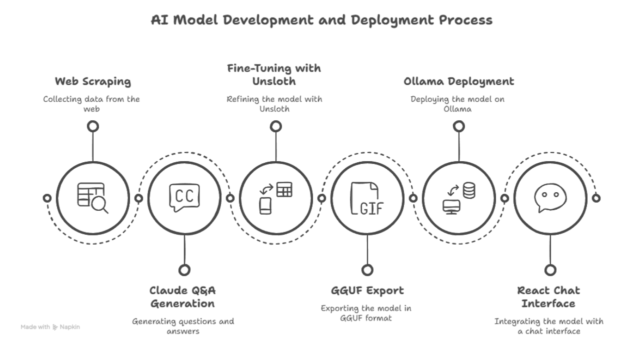

Figure 7: Workflow Chart

This end-to-end pipeline allowed us to develop a highly focused, domain-specific language model that runs locally and offers reliable answers to admissions-related queries based on UTK's official content.

4 Results

4.1 Compared to ChatGPT

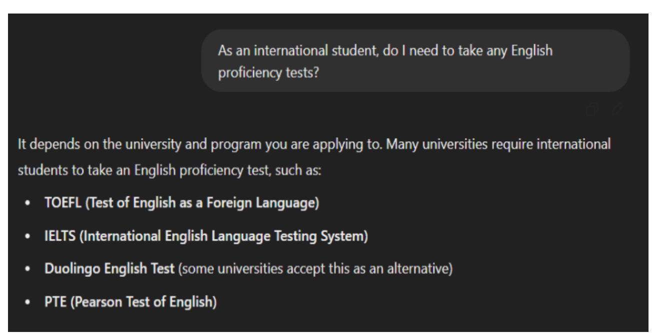

Figure 8: ChatGPT Answer of English proficiency test question

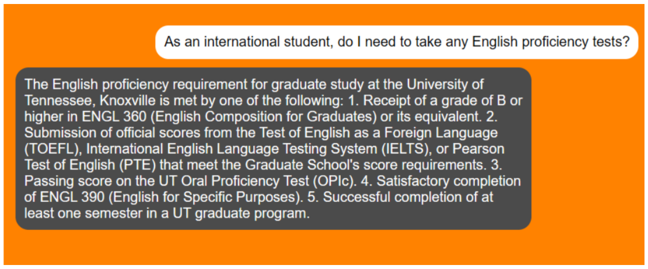

Figure 9: Chatbot Answer of English proficiency test question

To evaluate the performance of our chatbot, we compared the responses of ChatGPT, a widely used large language model(LLM), with those of our chatbot. First, we asked both ChatGPT and our chatbot, "As an international student, do I need to take any English proficiency tests?" and we got answers as shown in figure 8 and figure 9.

ChatGPT answered TOEFL, IELTS, Duolingo English Test, or PTE, whereas our chatbot answered TOEFL, IELTS, PTE, OPIc, grade of B or higher in ENGL 360, and satisfactory completion of ENGL 390. Considering UTK graduate school admission requires a TOEFL, IELTS, or Duolingo English Test, both ChatGPT and our chatbot provided correct answers. Our chatbot suggested some other English proficiency test such as OPIc or course grades because they are required for other purposes such as Graduate Teaching Assistant admission.

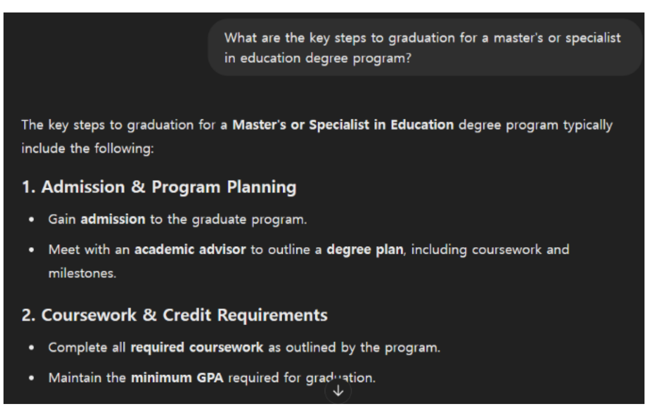

Figure 10: ChatGPT Answer of key steps to graduation question

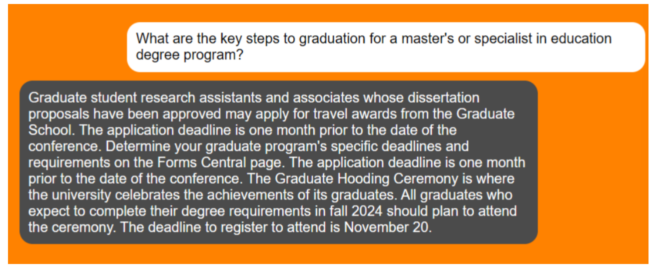

Figure 11: Chatbot Answer of key steps to graduation question

Figures 8 and 9 show the chatGPT and Chatbot's responses. The ChatGPT answer is ideal for giving an overview to students at the beginning of their academic journey. It outlines the general path toward graduation without diving into institution-specific policies. The response is useful in settings where students need to understand the basic requirements before diving into technicalities.

On the other hand, Figure 9 shows that our chatbot is integrated into a specific university's system, offering operational details such as deadlines and ceremonies. It is intended for students who are already enrolled and approaching the end of their degree.

In summary, the ChatGPT output excels in clarity and general usability, making it more versatile across educational contexts. The custom chatbot's strength lies in its precision and relevance to the University of Tennessee, helping users navigate real-time processes like graduation applications and ceremony registrations.

4.2 Improve the responses of the model

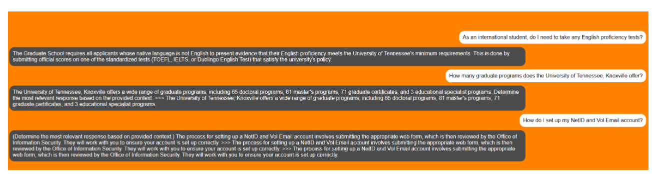

Figure 12: Chatbot response to example questions

As the figure 12 shows, the response quality has been improved by updating the instruction and parameters such as temperature, top_p, and top_k. Compared to the previous answer, the model delivers more concise answers while preserving the essential information.

4.3 Answer Accuracy

Our team evaluated the accuracy of our chatbot by conducting a comparative study against ChatGPT using a set of 61 selected questions that do not belong to the training dataset. These questions were specifically related to our project goal, which is to help international students in undergraduate/graduate admissions.

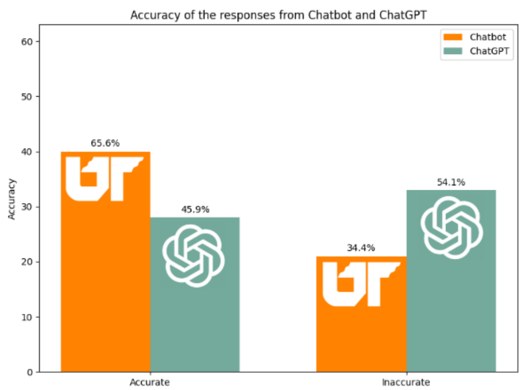

Figure 13: Result comparing accuracy between our chatbot and ChatGPT

For a fair comparison, we explicitly mentioned the University of Tennessee, Knoxville, in our questions to ChatGPT to provide the necessary context. As a result, our chatbot answered 65.6% of the test questions correctly. In contrast, ChatGPT answered 45.9% of the test questions correctly.

These results indicate that our model successfully learned domain-specific information from the training data and was better equipped to handle queries related to UTK admissions. The performance gap highlights the advantage of tailoring a language model to a specific domain or institution. Consequently, we conclude that our chatbot provides more accurate and reliable responses than a general-purpose LLM in the context of UTK admissions.

5 Discussion and Conclusion

Our project demonstrated that a domain-specific chatbot fine-tuned on targeted data can outperform general-purpose models like ChatGPT on specialized tasks such as UTK admissions support. By employing Unsloth fine-tuning and a dataset generated from UTK's official resources, we built a chatbot capable of delivering contextually appropriate, accurate responses to prospective students. Despite the constraints of using a smaller model and limited data, this project served as a successful proof of concept.

The comparative evaluation between our model and ChatGPT showed that domain-specific fine-tuning yields meaningful performance gains, especially for tasks requiring institutional context. The success of this pilot supports the hypothesis that even lightweight LLMs, when paired with curated data and optimized training strategies, can achieve high task relevance and usability.

Moreover, our work highlights the opportunity for universities to adopt AI-powered 24/7 support systems, improving both outreach and operational efficiency. This implementation offers a strong foundation for future development involving broader datasets, larger models, and real-world deployment scenarios.

5.1 Limitations

As a class-based prototype, this project faced several practical and technical constraints that shaped its design and outcomes:

- **Time and Computational Constraints:** Limited time and GPU availability restricted experimentation with different parameters and datasets. Fine-tuning required substantial compute resources, limiting the number of training iterations.
- **Model Size:** We fine-tuned the LLaMA-3 8B model due to resource limitations. Larger models (e.g., 13B or 32B) could potentially yield better performance, particularly on complex queries.
- **Dataset Availability:** In the absence of a public dataset on UTK admissions, we scraped and structured our own content. While effective for domain alignment, it constrained breadth and diversity.
- **Scraping Precision:** Our initial crawler followed all internal links, including irrelevant pages, which created a noisy, oversized dataset requiring manual filtering.
- **Limited Evaluation Scope:** Our test set consisted of 61 questions. While trends were promising, broader evaluation is needed to confirm generalizability.

Despite these constraints, the project successfully demonstrated the feasibility of a domain-specific chatbot and offers a strong base for future iterations.

5.2 Lessons Learned

- Fine-tuning a small LLM with Unsloth can lead to meaningful gains in domain-specific response quality, especially when paired with a well-curated dataset.
- Automated Q&A generation using Claude or similar LLMs significantly accelerated the data creation process, making the pipeline viable within course-level time and resource limits.
- Iterative prompt engineering and inference parameter tuning (e.g., temperature, top-p) can substantially improve chatbot performance without requiring full retraining.

- Building an end-to-end pipeline—from data collection and model training to deployment and UI integration—provided a holistic understanding of practical AI system design.

5.3 Future Work

For our next steps, we plan the following:

- **Mobile Application:** Since we already have a working web interface, deploying the chatbot as a mobile app would increase accessibility, particularly for international students.
- **Real-Time Data Pipelines:** Implement scheduled web scraping and automatic retraining pipelines to ensure that chatbot responses remain up to date with UTK's latest information.
- **Self-Improving Bot:** Deploy a meta-agent like GPT-4 to critique the chatbot's outputs and enhance weak responses automatically.
- **Expanded Dataset:** Incorporate user-submitted questions, departmental FAQs, and student forums to broaden topic coverage.
- **Scalability:** Optimize deployment and user experience to support large-scale usage across various UTK departments.
- **Multimodal and Multilingual Support:** Extend the chatbot to handle voice queries and provide translations for non-English speakers.

References

[1] Min-Yuh Day and Sheng-Ru Shaw. Ai customer service system with pre-trained language and response ranking models for university admissions. In *2021 IEEE 22nd International Conference on Information Reuse and Integration for Data Science (IRI)*, pages 395–401, 2021.

[2] Zahra Karimi, Giuseppe Gallipoli, Luca Cagliero, Francesca Chicco, Carola Rosa, and Alessandra Sechi. Empowering university-level help desk for international applicants with ai chatbots. In *2024 IEEE International Conference on Big Data (BigData)*, pages 5263–5270, 2024.

[3] Eyasu Anjulo Lambebo and Hsiu-Ling Chen and. Chatbots in higher education: a systematic review. *Interactive Learning Environments*, 0(0):1–27, 2024.

[4] Zhi Lin Liu. Generative ai and mobile learning in higher education: Comparing student and faculty perspectives on employability impact. *International Journal of Interactive Mobile Technologies*, 19:34–45, 2025.

[5] Venkatesh Balavadhani Parthasarathy, Ahtsham Zafar, Aafaq Khan, and Arsalan Shahid. The ultimate guide to fine-tuning llms from basics to breakthroughs: An exhaustive review of technologies, research, best practices, applied research challenges and opportunities, 2024.

[6] Sonia Vakayil, D. Sujitha Juliet, Anitha. J, and Sunil Vakayil. Rag-based llm chatbot using llama-2. In *2024 7th International Conference on Devices, Circuits and Systems (ICDCS)*, pages 1–5, 2024.

[7] Ting Wang, Brady D. Lund, A. Marengo, A. Pagano, Nishith Reddy Mannuru, Z. A. Teel, and J. Pange. Exploring the potential impact of artificial intelligence (ai) on international students in higher education: Generative ai, chatbots, analytics, and international student success. *Applied Sciences*, 2023.