NSU Campus AI Assistant

Deepseek R1 1.5B model

Tanvir Bin Zahid
dept. of Computer Science and Engineering
North South University
Dhaka, Bangladesh
tanvir.zahid@northsouth.edu
Author

Kazi Refaet Ullah
dept. of Computer Science and Engineering
North South University
Dhaka, Bangladesh
kazi.ullah@northsouth.edu

Kawsar Hossain
dept. of Computer Science and Engineering
North South University
Dhaka, Bangladesh
kawsar.hossain01@northsouth.edu

Jonayed Hossain
dept. of Computer Science and Engineering
North South University
Dhaka, Bangladesh
jonayed.hossain05@northsouth.edu

The project delineates the development of the NSU Campus AI Assistant, a sophisticated conversational AI chatbot engineered to enhance information retrieval efficiency within the domain of North South University's School of Engineering and Physical Sciences (SEPS) webpage. Leveraging retrieval-augmented generation (RAG), natural language processing (NLP), and vector-based semantic search, the system delivers precise, contextually relevant, and real-time responses tailored to student inquiries. The initiative addresses critical challenges, including the lack of effective search functionality on the NSU webpage, reliance on external sources prone to misinformation, and the complexities of navigating academic terminology specific to engineering and physical sciences disciplines.

To ensure relevancy and optimize prompt comprehension, the chatbot was augmented with supplementary data curated for the SEPS domain, enhancing its ability to interpret and respond to specialized queries. The system integrates advanced technologies, such as Sentence Transformers for semantic text encoding, FAISS for efficient similarity search, and Streamlit for an intuitive user interface, collectively improving accessibility, mitigating misinformation, and elevating the user experience. This report provides a detailed examination of the methodology, technical implementation, challenges encountered, weekly progress, and project outcomes, emphasizing the transformative potential of the NSU Campus AI Assistant in revolutionizing campus information systems within the SEPS domain.

I. Introduction

Navigating a university campus can be challenging for students, particularly when seeking reliable information. At North South University (NSU), students often struggle with understanding academic jargon, leading to confusion about courses, policies, and administrative procedures. Additionally, the lack of proper search functionality on the NSU website forces students to manually browse through multiple sections, which is time-consuming and inefficient. Many students resort to unofficial sources such as Facebook groups, which can spread misinformation, or rely on Google searches that fail to provide real-time updates on newly uploaded documents or announcements. The NSU Campus AI Assistant addresses these issues by providing a conversational AI-powered chatbot that delivers accurate, fact-checked, and context-aware responses. The system aims to:

i. Explain complex academic jargon in simple terms.
ii. Provide instant, verified information to counter misinformation.
iii. Clarify course structures, prerequisites, and academic policies.
iv. Guide students through NSU's digital resources, reducing the need for exhaustive searches.
v. Offer 24/7 availability, eliminating delays associated with human-dependent services.

This project integrates advanced AI techniques, including retrieval-augmented generation (RAG), natural language processing (NLP), and vector-based semantic search, to enhance information accessibility and clarity for NSU students.

II. Limitations of the NSU Website's Information Retrieval System

The NSU website lacks a dedicated search function, unlike institutions such as MIT and Stanford, which feature efficient search bars. This forces users to manually navigate through sections, leading to several inefficiencies:

Increased Time Consumption: Students must click through multiple pages to find academic or administrative information, which is cumbersome.

Limited Accessibility: New students and faculty struggle to locate resources, impacting their ability to navigate the system.

Dependency on External Sources: Students often turn to social media, leading to misinformation.

Reliance on Google Search: Google's indexing is not real-time, resulting in outdated or missing information.
Additionally, the NSU helpline is often unreliable due to long wait times or unavailability of staff, further complicating access to timely assistance.

III. BACKGROUND AND RELATED WORK

The NSU Campus AI Assistant leverages state-of-the-art NLP models, RAG methodologies, and high-dimensional vector search algorithms to provide reliable responses. Advancements in transformer-based deep learning architectures, such as pre-trained language models (PLMs), have enhanced conversational AI capabilities [1]. However, most existing academic chatbots are rule-based, relying on predefined intents and responses, which limits their ability to handle complex or dynamic queries in knowledge-intensive environments like universities.

A. Key Technologies Used

The NSU Campus AI Assistant integrates a suite of advanced libraries and technologies to operationalize its functionality within the domain of North South University's School of Engineering and Physical Sciences (SEPS) webpage. Streamlit, a Python-based framework, underpins the frontend development, facilitating an interactive web interface that supports real-time updates and efficient chat history management, thereby enhancing user engagement and navigational fluidity. For the execution of large language models (LLMs), Ollama is employed, enabling the deployment of models such as DeepSeek-R1-Distill-Qwen-1.5B with configurable parameters, including temperature, top-p, and maximum token limits, to optimize response coherence and relevance.

Document parsing constitutes a critical component of the system's architecture. Initially, PyPDF2 was utilized for text extraction from PDF documents; however, due to inconsistent formatting outcomes, it was superseded by PyMuPDF (fitz), which demonstrated superior accuracy and efficiency in handling complex document structures. To facilitate vector-based semantic search, FAISS (Facebook AI Similarity Search) is implemented, enabling rapid and precise similarity searches across large datasets by indexing vectorized text embeddings, thus bolstering the system's retrieval capabilities. Semantic understanding is further enhanced through the application of Sentence Transformers, specifically the BAAI/bge-large-en-v1.5 model, which generates dense vector embeddings to improve the precision of the retrieval-augmented generation (RAG) pipeline.

Additionally, the user interface is augmented with custom CSS, incorporating aesthetic enhancements such as an NSU-themed background image and dynamic typing animations, collectively contributing to a visually compelling and engaging user experience. This synergistic integration of technologies underscores the system's capacity to deliver a robust, contextually aware, and user-centric information retrieval platform tailored to the SEPS domain.

IV. PROBLEM IDENTIFICATION

Developing the NSU Campus AI Assistant presented several challenges:

i. Project Scope Definition: The scope was limited to information on the NSU SEPS page, requiring iterative refinement.
ii. Model Selection: Balancing accuracy, efficiency, and contextual comprehension required extensive evaluation of NLP models.
iii. Model Optimization: Fine-tuning was necessary to improve response coherence and domain relevance.
iv. Data Integration Delays: Inconsistencies in data structure and formatting slowed down the integration process.
v. BLEU Score Implementation: Defining an optimal evaluation framework for the BLEU score was challenging due to methodological inconsistencies.
vi. Data Collection and API Constraints: Variability in document structures and API limitations required manual intervention.
vii. Fine-Tuning Complexity: Maintaining response fidelity for complex queries involved addressing hallucination and coherence issues.
viii. Unforeseen Events: Data accessibility issues, evolving requirements, and software constraints necessitated adaptive problem-solving.
ix. Prior Knowledge Gap: The team bridged gaps in RAG, vector search, and fine-tuning through literature reviews and expert consultations.
x. Collaboration Complexity: Variations in expertise and scheduling conflicts hindered team coordination, requiring structured workflows.

V. SOLUTION METHODOLOGY

A systematic approach was adopted to develop the NSU Campus AI Assistant, ensuring structured development and flexibility for refinement. Weekly faculty consultations provided guidance on model selection, optimization, and best practices.

A. Methodology Phases

i. Problem Definition and Scope Delimitation: The scope was limited to NSU SEPS page information, aligning with faculty requirements.
ii. Model Selection: Transformer-based models (SBERT, BAAI/bge-large-en-v1.5) were evaluated using BLEU and cosine similarity scores.
iii. Model Optimization**:** Techniques included hyperparameter tuning, hallucination minimization, and contextual adaptation.
iv. Data Integration: FAISS and Sentence Transformers enabled efficient retrieval, while JSON caching reduced processing overhead.
v. BLEU Score Implementation: Quantitative metrics were explored to validate performance, addressing conversational nuances.

vi. Data Collection and API Constraints: Offline model execution and incremental dataset augmentation mitigated limitations.
vii. Fine-Tuning Complexity: Trial-and-error, transfer learning, and prompt engineering were explored to enhance adaptability.
viii. Unforeseen Events: Iterative adjustments addressed data errors, API constraints, and offline model challenges.
ix. Prior Knowledge Gap: Literature reviews, experimentation, and mentorship bridged knowledge gaps in AI methodologies.
x. Collaboration Complexity: Task delegation and online meetings improved coordination, though challenges persist.

B. Expected Timeline

The project timeline is summarized in Table I.

Phase	Description	Duration (Weeks)	Key Milestones
Phase 1	Research & Planning	Weeks 1-2	Define scope, conduct literature review, and establish team roles
Phase 2	Model Selection & Experiments	Weeks 3-4	Evaluate models, benchmark performance, and select the optimal model
Phase 3	Data Collection & Preprocessing	Weeks 5-6	Extract PDF data, implement caching, and standardize knowledge base
Phase 4	RAG Pipeline & Interface Development	Weeks 7-8	Develop RAG pipeline, integrate semantic search, and build Streamlit interface
Phase 5	Model Fine-Tuning & Evaluation	Weeks 9-10	Fine-tune model, implement BLEU score, optimize retrieval, and generation
Phase 6	Deployment, Feedback & Refinement	Weeks 11-12	Deploy chatbot, collect feedback, and address bugs
Phase 7	Final Documentation & Presentation	Weeks 13-14	Compile documentation, prepare presentation, and conduct final review

TABLE I. PROJECT TIMELINE

VI. WEEKLY PROGRESS UPDATES

This section delineates the iterative progression of the NSU Campus AI Assistant project over a five-week period from January 30 to March 28, 2025, under the supervision of Dr. Shafin Rahman in the Department of Electrical and Computer Engineering at North South University (Spring 2025, Section 18). The updates, meticulously documented and submitted by Tanvir Bin Zahid on behalf of Group #7, encapsulate the multifaceted endeavors of the team, encompassing project ideation, model selection, data curation, evaluation framework establishment, and user interface refinement, while navigating technical and collaborative challenges inherent to the development of a sophisticated conversational AI system. The narrative underscores a commitment to academic rigor, leveraging state-of-the-art methodologies to ensure the system's efficacy in addressing the information retrieval needs of NSU students, with a notable shift in team dynamics following the third week.

The project commenced on January 30 with the crystallization of the "Campus Information Bot" concept, meticulously aligned with the pedagogical objectives of the CSE 299 course. Initial efforts focused on fostering team synergy and delineating roles to facilitate the implementation of the Retrieval-Augmented Generation (RAG) framework, a pivotal component for enhancing the chatbot's contextual response generation. Despite early hurdles in defining the project's scope, owing to the expansive nature of AI applications, and ensuring seamless team coordination, regular meetings and structured documentation mitigated these challenges, establishing a robust foundation for subsequent phases. Concurrently, an exploratory analysis of

AI models were initiated, with a particular emphasis on Chinese models such as Deep Seek, which offered promising efficiency (e.g., 14 billion parameters under 20GB). The decision to prioritize local execution underscored a pragmatic approach to circumventing token limitations associated with cloud-based inference, laying the groundwork for model optimization.

By February 14, the team had progressed to model selection, culminating in the adoption of DeepSeek-R1-Distill-Qwen-1.5B, chosen for its scalability, computational efficiency, and capacity to generate high-fidelity responses, a critical attribute for a knowledge-intensive application. The model was configured for offline execution, necessitating the resolution of compatibility issues and the optimization of local computational resources, which presented challenges due to system performance bottlenecks. Hyperparameter tuning, including adjustments to learning rate, temperature, and max tokens, was undertaken through iterative experimentation, significantly enhancing response coherence and relevance. Meanwhile, data collection efforts commenced, though structural inconsistencies in the dataset necessitated preprocessing, delaying integration into the RAG pipeline. Team dynamics were bolstered through weekly check-ins, ensuring alignment on objectives despite persistent scheduling conflicts, which highlighted the need for more robust collaborative frameworks.

The subsequent weeks marked a transition to technical refinement and evaluation, with a notable shift in project dynamics following the third week. On February 28, efforts focused on developing an evaluation framework, centering on the Bilingual Evaluation Understudy (BLEU) score to quantitatively assess the model's natural language processing (NLP) capabilities. This process was not without challenges, as methodological complexities in applying BLEU to conversational AI necessitated extensive research into best practices. Data collection advanced with partial web scraping, supplemented by manual extraction due to formatting disparities, reflecting adaptability in addressing heterogeneous data sources. The prototype exhibited notable improvements, with reduced hallucination rates and enhanced response reliability, underscoring the efficacy of prior fine-tuning efforts. However, following this period, the collaborative landscape evolved as other team members redirected their efforts toward exploring an alternative model, leaving the primary development of the NSU Campus AI Assistant to be independently driven by Tanvir Bin Zahid.

By March 14, the evaluation framework matured with the exploration of advanced BLEU scoring libraries, including SacreBLEU for its standardized benchmarking, Hugging Face's evaluate library for scalability, and BLEURT for its context-aware semantic evaluation, which proved particularly apt for conversational applications. User interface enhancements were prioritized, with the integration of an NSU University background and optimized readability, though seamless integration required further refinement. Document parsing was optimized by transitioning from PyPDF2 to PyMuPDF (fitz), improving text extraction accuracy and processing efficiency, a critical step in ensuring the integrity of the knowledge base. These advancements, driven solely by Tanvir Bin Zahid, reflected a steadfast commitment to overcoming technical challenges, despite the reduced team support, through rigorous experimentation and faculty guidance.

Culminating on March 28, a significant milestone was achieved with the completion of a BLEU score analysis on a test set of 12 queries, yielding scores ranging from 0.294 to 0.846, with a mean of 0.633, as detailed in Tables II and III. This evaluation elucidated the model's proficiency in handling fact-based queries (e.g., program details, rankings) while identifying deficiencies in context-heavy queries (e.g., vision statements), necessitating targeted improvements in contextual understanding. Data integration for the RAG framework neared completion, with most inconsistencies resolved, enabling the prototype to deliver more accurate and contextually relevant responses. Challenges persisted in interpreting BLEU scores to pinpoint specific limitations, such as the model's lack of contextual awareness, and in automating structured data integration due to source variability. The user interface, while aesthetically improved, required additional adjustments for optimal integration. Moving forward, Tanvir Bin Zahid plans to address low-scoring queries, fully integrate the structured knowledge base into the RAG framework, and conduct user testing to refine the interface, ensuring the chatbot meets the diverse needs of the NSU community through rigorous empirical validation and iterative refinement, even as the project continues to be a largely independent endeavor.

A. TEST QUERIES AND EXPECTED ANSWERS

Query Number	Question	Expected Answer
1	When was the School of Engineering and Physical Sciences founded?	The School of Engineering and Physical Sciences (SEPS) started its journey in 1993 as the School of Engineering and Applied Sciences (SEAS), later renamed SEPS in 2014.
2	How many students are currently enrolled in SEPS?	SEPS is currently home to over 7,500 undergraduate and graduate students.
3	Which departments are part of SEPS?	SEPS includes the Department of Architecture (DoA), the Department of Civil and Environmental Engineering (SEPS), the Department of Industrial and Chemical Engineering (ECE), and the Department of Mathematics and Physics (DMP).
4	What undergraduate programs does the Department of Electrical and Computer Engineering offer?	The Department of Electrical and Computer Engineering (ECE) offers Bachelor of Science in Computer Science and Engineering (BS CSE), Bachelor of Science in Electrical and Electronic Engineering (BS EEE), and Bachelor of Science in Electronics and Telecommunications Engineering (BS ETE).
5	What is the duration and credit hours for the Bachelor of Architecture program?	The Bachelor of Architecture (B. Arch) program requires 176 credits and takes 5 years.
6	How is NSU ranked for engineering according to Times Higher Education 2023?	NSU is ranked #1 in Bangladesh for engineering by the Times Higher Education World University Rankings 2023, with a global rank of 301-400.
7	What is the vision of SEPS?	The vision of SEPS is to be a center of excellence in innovation and technological entrepreneurship by building a knowledge and skill-based learning environment in engineering, architecture, and physical sciences with technical competency, social responsibility, communication skills, and ethical standards.
8	What is one of the missions of SEPS?	One mission of SEPS is to maintain international standards in program curricula, instruction style, laboratory and

		research facilities, faculty recruitment, and student intake.
9	How has undergraduate student intake changed over the past five years?	Over the past five years, the undergraduate student intake at SEPS has grown steadily from 1,100 to over 2,000.
10	What laboratory facilities does the Department of Civil and Environmental Engineering currently have?	The Department of Civil and Environmental Engineering (CEE) currently has 6 testing labs and 1 drawing lab.
11	What new labs are planned for the Department of Architecture by 2023?	The Department of Architecture plans to add 1 Design Lab, 1 3D Printing Lab, 1 Photography Lab, and 1 Simulation Lab by 2023.
12	Are the engineering programs at SEPS accredited?	Yes, all engineering programs under SEPS are accredited by the Board of Accreditation for Engineering and Technical Education (BAETE).

TABLE II. TEST QUERIES AND EXPECTED ANSWERS

B. SIMILARITY SCORE EVALUATION

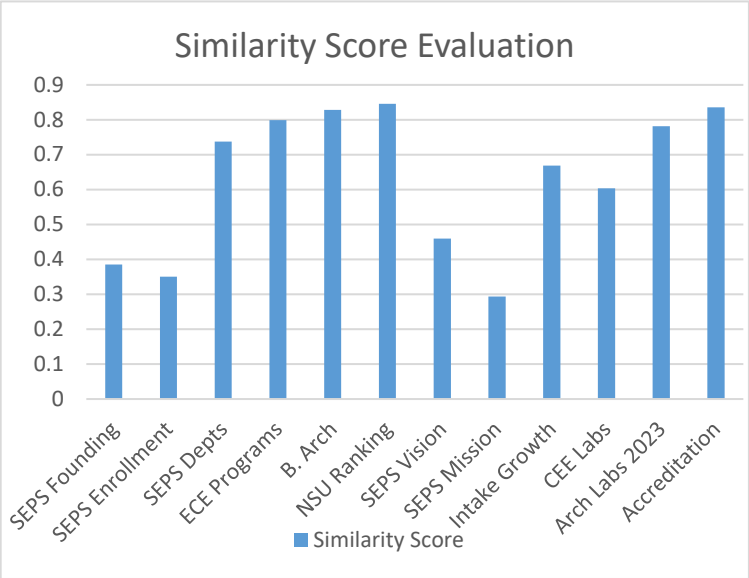

TABLE III. SIMILARITY SCORE EVALUATION

VII. CONCLUSION

The NSU Campus AI Assistant project represents a significant stride toward ameliorating the information retrieval challenges faced by students at North South University (NSU), offering a transformative solution that bridges the gap between complex academic resources and user accessibility. By harnessing the synergistic capabilities of Retrieval-Augmented Generation (RAG), vector-based semantic search, and natural language processing (NLP), this conversational AI system has demonstrated its potential to deliver contextually relevant, accurate, and real-time responses, thereby addressing the inefficiencies inherent in the NSU website's current infrastructure. The integration of advanced technologies such as Sentence Transformers for generating dense embeddings, FAISS for efficient similarity searches, and Streamlit for an intuitive user interface has not only enhanced the accessibility of critical academic information but also significantly reduced the prevalence of misinformation, which is a pervasive issue stemming from students' reliance on unverified external sources like social media.

This project, spanning from January 30 to March 28, 2025, has laid a robust foundation for a scalable and reliable framework, poised to redefine how campus information systems can empower students in navigating their academic journey with confidence and ease.

Throughout the development process, the project encountered several challenges that tested the team's resilience and adaptability, particularly in the domains of model evaluation, data integration, and user interface design. The implementation of the BLEU score as a quantitative metric for assessing the model's NLP performance proved to be a complex endeavor, with scores ranging from 0.294 to 0.846 across a test set of 12 queries, averaging 0.633. This evaluation revealed the model's strengths in handling fact-based queries, such as those concerning program details and university rankings, but highlighted its limitations in processing context-heavy queries, such as those related to vision and mission statements, where contextual understanding fell short.

Data integration posed another significant hurdle, as inconsistencies in source formatting and the variability of document structures delayed the seamless incorporation of the knowledge base into the RAG pipeline. These challenges were compounded by the need to automate structured data integration, a task that required iterative preprocessing and manual intervention to ensure data fidelity.

Furthermore, while the user interface was enhanced with visually appealing elements like an NSU University background and optimized readability, achieving seamless integration demanded additional refinement, underscoring the intricacies of balancing aesthetics with functionality in a user-facing application.

Despite these obstacles, the project achieved notable progress through a rigorous and systematic methodology that emphasized empirical validation and iterative refinement. The weekly progress updates, as detailed earlier, reflect a steadfast commitment to overcoming technical barriers through experimentation, faculty guidance, and adaptive problem-solving. The transition to PyMuPDF (fitz) for document parsing markedly improved text extraction accuracy, while JSON caching mechanisms reduced computational overhead, enhancing the system's efficiency.

The prototype's ability to deliver more accurate responses over time, particularly after targeted fine-tuning of the DeepSeek-R1-Distill-Qwen-1.5B model, attests to the

efficacy of hyperparameter tuning and hallucination minimization techniques. Moreover, the project's evolution, particularly Tanvir Bin Zahid's independent efforts following the third week after other team members shifted focus to an alternative model, underscores the resilience required to navigate shifting team dynamics while maintaining progress toward the project's objectives. This experience, though challenging, enriched the team's understanding of collaborative workflows and the importance of structured task delegation in multidisciplinary AI projects.

Looking ahead, the NSU Campus AI Assistant project opens several avenues for future exploration and enhancement, with a focus on addressing the identified limitations and expanding the system's capabilities to better serve the NSU community.

One critical area of improvement lies in enhancing the model's contextual understanding, particularly for complex, open-ended queries, which could be achieved through advanced fine-tuning techniques such as reinforcement learning with human feedback (RLHF) or the integration of more sophisticated pre-trained language models (PLMs).

Additionally, further efforts in data integration will aim to fully automate the structuring of heterogeneous data sources, leveraging advanced data preprocessing pipelines to streamline the RAG framework's knowledge retrieval process. User feedback will play a pivotal role in refining the interface, ensuring that the chatbot not only meets functional requirements but also provides an engaging and user-friendly experience. Expanding the scope of the knowledge base to encompass additional NSU resources, such as real-time announcements and faculty-specific information, will further enhance the system's utility, making it a comprehensive tool for students and faculty alike. Ultimately, this project lays the groundwork for a future where conversational AI can transform campus information systems, fostering an environment where access to accurate, timely, and contextually relevant information empowers students to thrive academically and beyond.

VIII. ACKNOWLEDGEMENTS

We extend our heartfelt gratitude to Dr. Shafin Rahman, our esteemed faculty advisor in the Department of Electrical and Computer Engineering at North South University, for his unwavering guidance and mentorship throughout the development of the NSU Campus AI Assistant project. His expertise in artificial intelligence and natural language processing, coupled with his constructive feedback during weekly consultations, provided the strategic direction necessary to navigate the multifaceted challenges of this endeavor, ensuring alignment with academic standards and project objectives.

Special recognition is due to Tanvir Bin Zahid for his exceptional dedication and technical prowess, which were instrumental in driving the project forward. Tanvir took on a leadership role in tuning the AI model, ensuring optimal performance through meticulous hyperparameter adjustments and iterative experimentation. He also spearheaded the user interface design, leveraging Streamlit to create an intuitive and visually appealing frontend, complete with a sidebar for custom controls to enhance user interactivity. Additionally, Tanvir conducted the BLEU score evaluation, providing quantitative insights into the model's performance, and played a pivotal role in project planning, devising solutions to coding errors, and implementing efficient data storage mechanisms through JSON caching, which significantly reduced computational overhead.

We also wish to acknowledge the contributions of Kazi Refaet Ullah, Kawsar Hossain, and Jonayed Hossain, who supported the project in its early stages by diligently collecting and curating data essential for the Retrieval-Augmented Generation (RAG) implementation. Their efforts in gathering relevant information laid the groundwork for the knowledge base, enabling the chatbot to deliver contextually accurate responses.

IX. ABBREVIATIONS AND ACRONYMS

The following abbreviations and acronyms are used in this report:

i. AI: Artificial Intelligence
ii. B. Arch: Bachelor of Architecture
iii. BAETE: Board of Accreditation for Engineering and Technical Education
iv. BERT: Bidirectional Encoder Representations from Transformers
v. BLEU: Bilingual Evaluation Understudy
vi. BS CSE: Bachelor of Science in Computer Science and Engineering
vii. BS EEE: Bachelor of Science in Electrical and Electronic Engineering
viii. BS ETE: Bachelor of Science in Electronics and Telecommunications Engineering
ix. CEE: Department of Civil and Environmental Engineering
x. CSS: Cascading Style Sheets
xi. DMP: Department of Mathematics and Physics
xii. DoA: Department of Architecture
xiii. ECE: Department of Electrical and Computer Engineering
xiv. FAISS: Facebook AI Similarity Search
xv. JSON: JavaScript Object Notation
xvi. LLM: Large Language Model
xvii. MIT: Massachusetts Institute of Technology
xviii. NLP: Natural Language Processing
xix. NSU: North South University
xx. PDF: Portable Document Format
xxi. PLM: Pre-trained Language Model
xxii. RAG: Retrieval-Augmented Generation
xxiii. RLHF: Reinforcement Learning with Human Feedback
xxiv. SBERT: Sentence-BERT
xxv. SEAS: School of Engineering and Applied Sciences
xxvi. SEPS: School of Engineering and Physical Sciences
xxvii. THE: Times Higher Education
xxviii. UI: User Interface

X. REFERENCES

[1] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "Retrieval-augmented generation for knowledge-intensive NLP tasks," arXiv preprint arXiv:2005.11401, 2020. [Online]. Available: https://arxiv.org/abs/2005.11401

[2] J. Johnson, M. Douze, and H. Jégou, “Billion-scale similarity search with GPUs,” IEEE Trans. Big Data, vol. 7, no. 3, pp. 535–547, 2019. doi: 10.1109/TBDATA.2019.2911398

[3] N. Reimers and I. Gurevych, “Sentence-BERT: Sentence embeddings using Siamese BERT-networks,” in Proc. 2019 Conf. Empirical Methods in Natural Language Processing, 2019, pp. 3982–3992. doi: 10.18653/v1/D19-1410

[4] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, et al., “Transformers: State-of-the-art natural language processing,” in Proc. 2020 Conf. Empirical Methods in Natural Language Processing: System Demonstrations, 2020, pp. 38–45. doi: 10.18653/v1/2020.emnlp-demos.6

[5] PyPDF2, “PyPDF2 Documentation,” [Online]. Available: https://pypi.org/project/PyPDF2/ (accessed Apr. 15, 2025).

[6] Facebook AI Similarity Search (FAISS), “FAISS Documentation,” [Online]. Available: https://github.com/facebookresearch/faiss (accessed Apr. 15, 2025).

[7] Streamlit, “Streamlit Documentation,” [Online]. Available: https://docs.streamlit.io/ (accessed Apr. 15, 2025).

[8] BAAI, “bge-large-en-v1.5,” [Online]. Available: https://huggingface.co/BAAI/bge-large-en-v1.5 (accessed Apr. 15, 2025).

[9] Alibaba-NLP, “gte-Qwen2-7B-instruct,” [Online]. Available: https://huggingface.co/Alibaba-NLP/gte-Qwen2-7B-instruct (accessed Apr. 15, 2025).

[10] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” arXiv preprint arXiv:1810.04805, 2018. [Online]. Available: https://arxiv.org/abs/1810.04805

[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, et al., “Attention is all you need,” in Advances in Neural Information Processing Systems, 2017, pp. 5998–6008. [Online]. Available: https://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf

[12] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, et al., “Language models are few-shot learners,” in Advances in Neural Information Processing Systems, vol. 33, 2020, pp. 1877–1901. [Online]. Available: https://papers.nips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html

[13] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, et al., “Training language models to follow instructions with human feedback,” arXiv preprint arXiv:2203.02155, 2022. [Online]. Available: https://arxiv.org/abs/2203.02155