**DOMAIN ADAPTATION IN LARGE LANGUAGE MODELS**



**A Thesis**

submitted for the degree of

**Bachelor of Technology**

**in**

**Information Technology**

Submitted by:

**Harshita Prakash [2116765]**

**Nandni Gupta [2116795]**

**Pearl Khatri [2116800]**

**Shaivi Agarwal [2116818]**

Supervisor:

**Dr. Nisheeth Joshi**

Associate Professor

Department of Computer Science

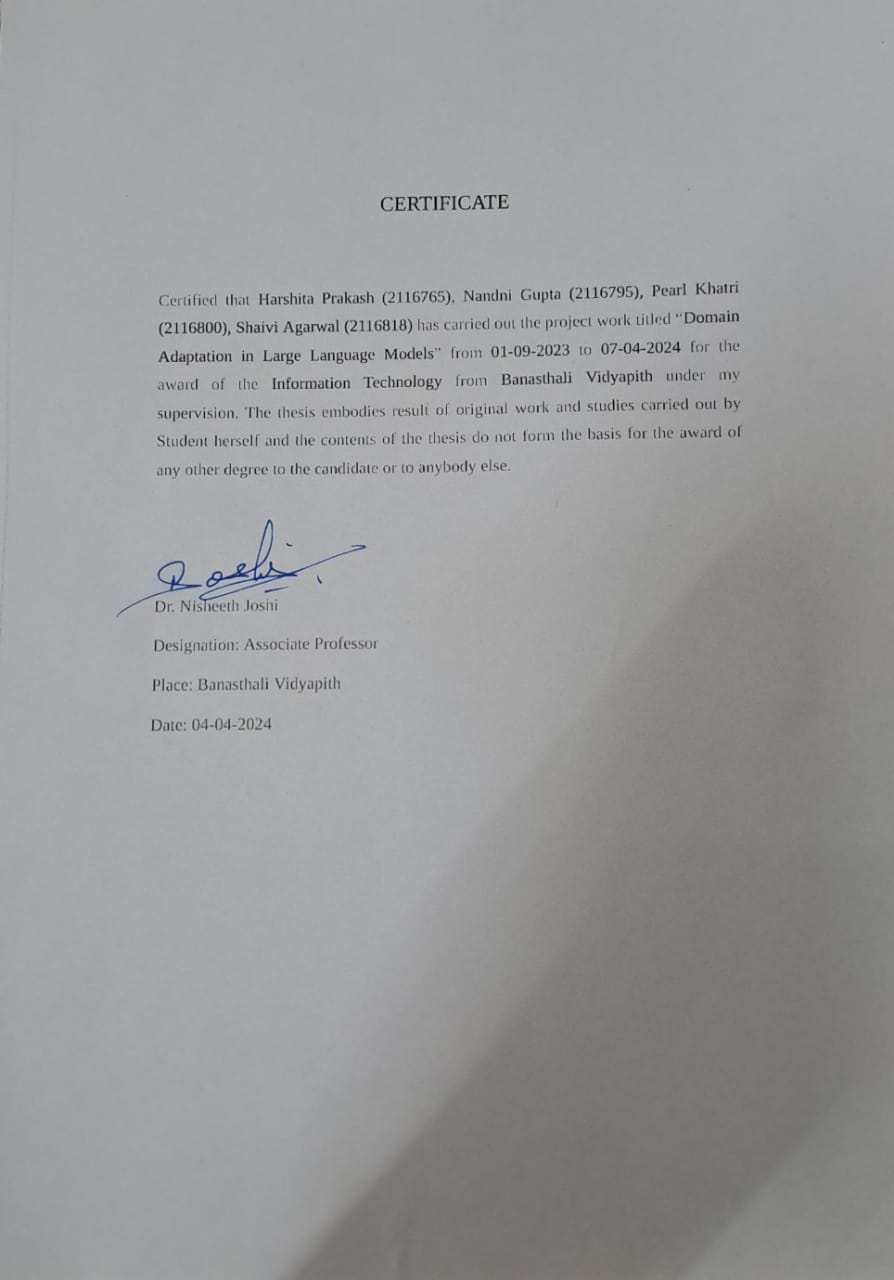
Banasthali Vidyapith

Rajasthan.

**FACULTY OF MATHEMATICS AND COMPUTING**

**BANASTHALI VIDYAPITH**

**Rajasthan, India 2024**

****

**ABSTRACT**

Large Language Models (LLMs) have revolutionized natural language processing, enabling the development of advanced AI systems capable of understanding and generating human-like text. In this project, we propose the creation of a chatbot specifically tailored to streamline the admission process at Banasthali Vidyapith.

The proposed chatbot aims to address several key needs identified through a thorough analysis, including simplifying website navigation, providing quick access to relevant information, clarifying placement-related details, and consolidating student queries in one accessible platform.

To achieve this, we outline a comprehensive methodology. Firstly, we plan to develop a corpus by gathering and preprocessing data from Banasthali Vidyapith's annual reports spanning the past decade. This data will be structured into various sections, such as admission requirements, procedures, fees, available courses, and contact details, forming a robust knowledge base for the chatbot.

Secondly, we intend to fine-tune the chatbot using Natural Language Processing (NLP) techniques. By deeply analyzing the collected data, including annual reports and information from the institution's website, the chatbot will gain a nuanced understanding of user queries and provide meaningful responses. We aim to employ a parameter-efficient approach called LoRA (Low-Rank Adaptation) for fine-tuning, optimizing the model's performance while minimizing computational costs.

The expected outcomes of this project include the deployment of a functional chatbot customized for Banasthali Vidyapith. This chatbot will assist prospective students and their parents by providing accurate information related to the admission procedure and history of the institution, thereby enhancing the overall admission process and user experience at Banasthali Vidyapith.

**ACKNOWLEDGEMENT**

“Alone, we can do so little; together, we can do so much.”

– Helen Keller

The fruition of this research paper stands as a testament to the collaborative efforts of multiple individuals who dedicated their time, intellect, and resources to its completion. As we reflect on this journey, we extend our deepest gratitude to those who have supported and inspired us along the way.

First and foremost, we express our heartfelt appreciation to our esteemed mentor Dr. Nisheeth Joshi (Associate Professor, Dept. of Computer Science, Banasthali Vidyapith) whose blessings, guidance, expertise and wishes made this miracle happen. A year ago, during the initial days of our major project, we came to him as a mound of clay, unstructured and unknown to the magical realm of AI-NLP and research, in general. He kindly took us under his wing and patiently sculpted us into something valuable. He taught us to imagine the impossible and find our path to it. He never allowed us to get disappointed with failures and helped us reach our goals by defeating all the odds. It was due to his supervision and guidance, that we could complete this work in time. We are immensely grateful for his tolerance, patience and encouragement in our moments of frustration. Anything and everything we do here after will be an act of tribute to him, as a gesture of our gratefulness. We hope we make him proud someday.

If there is someone after him, whose support and wishes made this research happen, it is our senior Pragya Katyayan (Senior Research Fellow (SERB), Speech and Language Processing Laboratory, Center for Artificial Intelligence, Banasthali Vidyapith, Rajasthan) didi who has been a constant motivator and helped us find our way through our initial days. This research wouldn't have been possible without her help. She always supported us, even when she was busy with her own work. Whenever we had problems with coding or debugging, she made time to assist us and never left us empty-handed.

We express our sincere acknowledgement to Prof. Anshuman Shastri (Director, Centre for Artificial Intelligence, Banasthali Vidyapith) and Prof. Ina Aditya Shastri (Vice Chancellor, Banasthali Vidyapith) for establishing a healthy research environment at Banasthali Vidyapith and providing us with all the facilities and resources to conduct our research. We are also grateful to Prof. Chandra Kumar Jha (Head of Department (Computer Science), Banasthali Vidyapith) for being helpful towards our research requirements. We are thankful to all the teaching and non-teaching staff members of the Department of Computer Science, Banasthali Vidyapith for their vigorous efforts in managing our research work.

Without the support of our families, this feat was impossible, who taught us to be fearless and independent and never failed to encouraged us to be the best. They not only sacrificed their wishes in front of us, they fought the world so that they could give us the best possible education. It is because of their determination, that we are what we are today.

Apart from all the positive support and encouraging force around us, there were people who judged us, underestimated us, told us that we couldn’t do it and created all sorts of hurdles in our way. What they didn’t realize was, they were motivating us even more to achieve our goals and we are thankful to them for all the energy they put in for us.

Last, but not the least, we would like to thank all the anonymous reviewers for giving time and efforts in reviewing this work and providing their super-valued suggestions.

**-HARSHITA PRAKASH**

**-NANDNI GUPTA**

**-PEARL KHATRI**

**-SHAIVI AGARWAL**

**TABLE OF CONTENTS**

CHAPTER 1: [INTRODUCTION](#bookmark=id.gjdgxs) 1

CHAPTER 2: LITERATURE REVIEW 13

CHAPTER 3: EXPERIMENTAL SETUP 24

CHAPTER 4: PROPOSED METHODOLOGY 27

CHAPTER 5: RESULTS AND ANALYSIS 39

CHAPTER 6: CONCLUSION AND FUTURE WORKS 44

REFERENCES 46

**CHAPTER 1**

**INTRODUCTION**

In the rapidly evolving landscape of artificial intelligence (AI), Large Language Models (LLMs) have emerged as pivotal tools capable of processing and generating human language with remarkable sophistication. These models, characterized by their immense size, complexity, and deep learning architecture, have revolutionized various natural language processing (NLP) tasks. With their ability to comprehend intricate linguistic patterns and semantic nuances, LLMs have found applications across diverse domains, from content generation to sentiment analysis, and beyond.

A large language model (LLM) is a type of artificial intelligence (AI) model that's designed to process and create human language. It's characterized by its size, complexity, and the ability to comprehend and interpret text in a human-like manner. These models are typically trained on massive datasets such as books, articles, and the internet. Through their extensive training, LLMs can easily understand a wide range of writing styles and topics. LLMs uses deep learning techniques to process and generate text, which lets them take on complex semantic and linguistic patterns. These tools can be useful in a variety of natural language processing tasks.

The most well-known examples of a Large Language Model are GPT-3 (Generative Pre-trained Transformer 3) by OpenAI, BERT (Bidirectional Encoder Representations from Transformer) by Google and many more. GPT-3 has acquired significant recognition for its capacity to excel in various language-related tasks and generate human-like text.

Large Language Models (LLMs) operate through deep learning, employing a neural network architecture called transformers. These models undergo pre-training on extensive text data sourced from the internet, enabling them to grasp language patterns, grammar, context, and semantics. During pre-training, the model predicts what comes next in a sentence, which helps it grasp the relationships between words and phrases. After undergoing pre-training, LLMs can be further customized through fine-tuning to cater to specific tasks. LLMs are highly proficient in various natural language processing tasks such as language comprehension, text creation, translation, summarization, and more. This versatility makes them valuable tools for a diverse array of NLP applications.

**EXAMPLES:**

**GPT-3 (Generative Pre-trained Transformer 3):** Developed by OpenAI, GPT-3 is one of the largest and most versatile LLMs to date, with 175 billion parameters.

It is known for its ability to perform a wide range of natural language processing tasks, including text completion, translation, summarization, question answering, and more.

It operates based on a transformer architecture and is pre-trained on vast amounts of text data from the internet, enabling it to generate human-like text across various domains without task-specific fine-tuning.

Applications include chatbots, content generation, language understanding tasks, creative writing, and more.

**BERT (Bidirectional Encoder Representations from Transformers):** Introduced by Google AI in 2018, BERT is a language representation model trained on large text corpora using bidirectional context.

It captures deep contextualized embeddings of words and sentences, allowing it to understand and generate text with a deeper understanding of context and semantics.

It has been widely adopted for various natural language processing tasks, including text classification, named entity recognition, sentiment analysis, and question answering.

Applications include search engine optimization, language understanding tasks, sentiment analysis, and more.

**OpenAI Codex:** Based on the GPT-3 architecture, OpenAI Codex is a domain-specific LLM trained specifically for programming-related tasks.

It is fine-tuned on code-related datasets and has the ability to understand and generate code in multiple programming languages, including Python, JavaScript, HTML, CSS, and more.

It can perform tasks such as code completion, code generation, debugging assistance, and API documentation.

OpenAI Codex is used in various software development applications, including code generation, code review, automated testing, and documentation generation.

**T5 (Text-To-Text Transfer Transformer):** Developed by Google Research, T5 is a versatile LLM that operates under a unified text-to-text framework.

It is trained to convert input text into output text, enabling it to perform various natural language processing tasks, including translation, summarization, text classification, and question answering, by framing tasks as text-to-text transformations.

It achieves state-of-the-art performance on a wide range of NLP benchmarks and tasks by leveraging large-scale pre-training and fine-tuning techniques.

Applications of T5 include language translation, text summarization, document understanding, and conversational AI.

**XLNet**: XLNet, introduced by Google AI and Carnegie Mellon University, is a generalized autoregressive pre-training method that combines ideas from autoregressive models (like GPT) and autoencoder models (like BERT).

It achieves state-of-the-art results on multiple NLP benchmarks by leveraging permutation-based training and capturing bidirectional context.

It is known for its ability to generate coherent and contextually relevant text across various tasks, including language modeling, text classification, sentiment analysis, and more.

Applications of XLNet include language understanding tasks, content generation, machine translation, and more.

**TYPES:**

The common types of Large Language Models (LLMs) are as follows:

**Zero-shot model:** It refers to a large, versatile model trained on a broad dataset, capable of providing accurate results for general use cases without requiring additional training.

Zero-shot models represent a remarkable advancement in the realm of natural language processing (NLP). These models are trained on massive datasets and exhibit remarkable proficiency in understanding and generating text across a wide range of topics and contexts without the need for task-specific fine-tuning. One of the most prominent examples of a zero-shot model is GPT-3 (Generative Pre-trained Transformer 3) developed by OpenAI. GPT-3 has garnered significant attention for its ability to perform various language-related tasks, including text completion, translation, summarization, and question answering, without any additional fine-tuning. The versatility of zero-shot models makes them valuable for applications where adaptability to diverse tasks and domains is crucial, such as chatbots, content generation, and language understanding tasks.

**Fine-tuned or domain-specific models:** These models represent a specialized category of LLMs that undergo additional training on top of a pre-existing zero-shot model. This further training process involves fine-tuning the model parameters using task-specific datasets or domain-specific corpora to improve performance on particular tasks or domains. For example, OpenAI Codex is a domain-specific language model fine-tuned for programming-related tasks based on GPT-3. Domain-specific models are tailored to address the unique requirements and nuances of specific domains, such as medical, legal, or financial domains. By fine-tuning the model on task-specific data, these models can achieve higher accuracy and efficiency for targeted applications, such as code generation, legal document analysis, or medical diagnosis.

**Language representation model:** One instance being Bidirectional Encoder Representations from Transformers (BERT). These models utilize deep learning and transformers, making them well-suited for natural language processing (NLP).

Language representation models play a fundamental role in capturing rich contextualized representations of text data for various NLP tasks. These models, exemplified by architectures like BERT (Bidirectional Encoder Representations from Transformers), are trained on large text corpora using unsupervised learning techniques. By pre-training on vast amounts of text data, language representation models learn to generate contextualized word embeddings that capture the semantic and syntactic nuances of language. BERT-based models have demonstrated remarkable performance on a wide range of NLP tasks, including text classification, named entity recognition, sentiment analysis, and question answering. The contextualized embeddings learned by these models enable them to understand and generate text with a deeper understanding of context and semantics.

**Multimodal model**: These expand beyond text-only capabilities. Originally, LLMs were designed for text, but the multimodal approach allows handling both text and images.

Multimodal models represent a fusion of LLMs with other modalities such as images, audio, or video. These models extend beyond text-only capabilities to incorporate additional sources of information, enabling more comprehensive understanding and generation of content. For example, GPT-4 integrates both text and image inputs to generate text-based descriptions or responses based on visual input. Multimodal models are valuable for applications that require understanding and processing of diverse types of data, such as image captioning, video summarization, and multimodal translation. By leveraging multiple modalities, these models can generate more informative and contextually rich outputs, enhancing their utility in various domains, including multimedia content creation, accessibility, and human-computer interaction.

Each type of Large Language Model (LLM) offers unique capabilities and advantages suited to different applications and use cases. Zero-shot models provide broad adaptability and versatility across diverse tasks and domains. Fine-tuned models offer specialized performance tailored to specific tasks or domains, achieving higher accuracy and efficiency through targeted training. Language representation models capture rich contextualized representations of text data, enabling deeper understanding and generation of text. Multimodal models extend LLMs' capabilities by incorporating additional modalities, enabling more comprehensive understanding and generation of content across diverse data types. The choice of LLM type depends on the specific requirements and objectives of the application or task at hand, guiding the selection of the most appropriate model architecture and training approach.

**APPLICATIONS:**

Large Language Models (LLMs) play a crucial role in meeting various demands and finding applications across diverse domains. Here are some key needs that LLMs address:

**Natural Language Understanding**: LLMs possess the capability to grasp and analyze human language with a proficiency that was challenging to attain using conventional rule-based or statistical approaches. They excel in comprehending the context, semantics, and nuances of text, making them valuable for applications such as sentiment analysis, text classification, and language translation.

Sentiment Analysis: Sentiment analysis involves determining the sentiment or emotional tone expressed in a piece of text, such as positive, negative, or neutral. LLMs use techniques like deep learning and supervised learning to classify text based on sentiment. Real-world applications include analyzing product reviews, social media posts, and customer feedback to gauge public opinion and sentiment towards brands or products.

Text Classification: Text classification involves categorizing text into predefined categories or classes based on its content. LLMs leverage techniques like supervised learning and feature extraction to classify text. Applications include news categorization, spam detection, and topic modeling.

**Content Generation**: Language models (LLMs) possess the ability to produce text that closely resembles human writing, making them valuable for diverse applications such as marketing, advertising, website content generation, and creative writing.

Marketing: LLMs can generate marketing copy for advertisements, email campaigns, and social media posts. By analyzing past successful campaigns and understanding consumer preferences, LLMs can generate persuasive and engaging marketing content tailored to specific target audiences.

Website Content Generation: LLMs can automatically generate website content, including product descriptions, blog posts, FAQs, and landing page content. By leveraging data from the website's database and analyzing user behavior, LLMs can create personalized and relevant content to enhance user engagement and conversion rates.

Creative Writing: LLMs can generate creative writing pieces such as poems, stories, and essays. By analyzing existing literary works and understanding narrative structures, themes, and writing styles, LLMs can generate original and compelling creative content.

Large Language Models can generate coherent and contextually relevant text across a broad spectrum of topics.

**Chatbots and Virtual Assistants**: LLMs form the basis for chatbots and virtual assistants. They empower these AI systems to participate in authentic and significant dialogues with users, offering customer assistance, responding to inquiries, and aiding in various tasks.

Customer Support Chatbots: LLM-powered chatbots can provide 24/7 customer support, answer frequently asked questions, troubleshoot issues, and escalate complex queries to human agents when necessary. By analyzing customer queries and historical interactions, LLMs can provide personalized and contextually relevant responses.

Virtual Personal Assistants: LLM-powered virtual assistants like Siri, Alexa, and Google Assistant can perform various tasks such as setting reminders, sending messages, making appointments, and providing information on weather, news, and traffic. By understanding natural language queries and leveraging external APIs and services, LLMs can assist users in their daily tasks and routines.

**Sentiment Analysis**: LLMs can analyze social media posts, reviews, and customer feedback to determine sentiment and gain insights into public opinion. This is valuable for businesses, marketers, and researchers.

Brand Monitoring: Companies use sentiment analysis to monitor public sentiment and perception towards their brand, products, and services. By analyzing social media posts, customer reviews, and news articles, companies can identify positive and negative sentiment trends, address customer concerns, and mitigate brand reputation risks.

Market Research: Market researchers use sentiment analysis to analyze consumer opinions, preferences, and trends in specific industries or markets. By analyzing social media conversations, surveys, and focus group discussions, researchers can gain insights into consumer sentiment towards products, advertisements, and marketing campaigns.

**Data Summarization:** LLMs can automatically summarize long documents or articles, making it easier for users to extract key information quickly. This is particularly useful in news aggregation and research.

News Aggregation: LLMs can automatically summarize news articles from various sources to provide users with concise and informative summaries. By extracting key information such as headlines, main points, and quotes, LLMs enable users to quickly grasp the essence of news stories without reading the full articles.

Document Summarization: LLMs can summarize long documents, research papers, and reports to extract important insights and findings. By identifying key sentences, paragraphs, and sections, LLMs can create concise summaries that capture the essence of the original text.

**Transcription:** Speech-to-Text Transcription: LLMs are used in speech recognition systems to transcribe spoken language into written text. By analyzing audio signals and converting them into text representations, LLMs enable applications such as voice dictation, voice search, and voice-controlled assistants.

Automatic Subtitling: LLMs can automatically generate subtitles for videos and audio recordings by transcribing spoken dialogue and synchronizing it with the corresponding timestamps. By analyzing speech patterns, speaker diarization, and background noise, LLMs produce accurate and synchronized subtitles for accessibility and language localization purposes.

**ARCHITECTURE:**

The architecture of Large Language Models (LLMs) is primarily based on deep learning techniques, specifically utilizing neural network architectures known as transformers. These models are characterized by their immense size, complexity, and ability to process and generate human language with remarkable sophistication. Below is the explanation of the architecture of LLMs:

**Transformer Architecture:**

Transformers serve as the foundation of LLMs. This architecture was introduced in the paper "Attention is All You Need" by Vaswani et al. in 2017.

The key innovation of transformers lies in the self-attention mechanism, which allows the model to weigh the importance of different words in a sequence when processing information.

**Input Embeddings:**

At the input layer, words or tokens are represented as dense vectors called embeddings.

These embeddings capture the semantic meaning of words in a continuous vector space, allowing the model to process textual information efficiently.

Embeddings are typically learned during the pre-training phase of the model using techniques such as Word2Vec, GloVe, or contextualized embeddings like ELMo or BERT.

**Stacked Encoder-Decoder Architecture:**

LLMs often employ a stacked encoder-decoder architecture, where the encoder processes the input text, and the decoder generates the output text.

Each encoder layer in the stack transforms the input sequence using self-attention mechanisms and feed-forward neural networks.

The decoder layer generates the output sequence based on the representations learned by the encoder, using similar self-attention and feed-forward mechanisms.

**Self-Attention Mechanism:**

The self-attention mechanism enables the model to weigh the importance of each word in the input sequence when computing the representation for a given word.

It allows the model to capture long-range dependencies between words and learn contextual information effectively.

The self-attention mechanism computes attention scores for each word pair in the input sequence, which are then used to compute weighted sums of the input embeddings.

**Feed-Forward Neural Networks:**

After the self-attention mechanism, the transformer architecture includes a position-wise feed-forward neural network.

This network applies a linear transformation followed by a non-linear activation function to each position in the sequence independently.

The feed-forward neural network helps the model capture complex patterns and interactions between words in the input sequence.

**Layer Normalization and Residual Connections:**

To stabilize training and improve the flow of information through the network, transformers often incorporate layer normalization and residual connections.

Layer normalization normalizes the activations of each layer, reducing the internal covariate shift and improving training stability.

Residual connections enable the model to learn residual representations, facilitating the training of deeper architectures and mitigating the vanishing gradient problem.

**Output Layer:**

The output layer of the LLM typically consists of a softmax activation function, which produces a probability distribution over the vocabulary of possible output tokens.

During training, the model learns to predict the next token in a sequence given the preceding context, maximizing the likelihood of the actual next token in the training data.

At inference time, the model generates text by sampling tokens from the output distribution or using beam search to find the most likely continuation of the input sequence.

The architecture of LLMs, built upon the transformer framework, enables these models to capture complex linguistic patterns, understand context, and generate human-like text across a wide range of natural language processing ta\

**CHAPTER 2**

**LITERATUR REVIEW**

AbuShawar et al., (2015) addresses one noteworthy chatbot, ALICE (Artificial Linguistic Internet Computer Entity), which has been in development since 1995. ALICE adopts a distinct approach by separating the chatbot engine from the language knowledge model, allowing flexibility in language knowledge models. Unlike more complex Natural Language Processing (NLP) techniques used by some other chatbots, ALICE relies on a vast number of basic categories or rules for matching input patterns to output templates. Its emphasis is on quantity rather than sophistication, with the default ALICE system featuring around fifty thousand categories. Their paper highlights the development of software to convert readable text into ALICE's format, making it easier to expand ALICE's knowledge base. Moreover, the paper explores the application of ALICE in various contexts, such as language learning and information retrieval, demonstrating the versatility of chatbots beyond mere entertainment.

Hiremath et al., (2018) presents the practical implementation of an automatic response system driven by artificial intelligence and machine learning, specifically aimed at answering student queries. It emphasizes the extraction of keywords from user queries to generate appropriate responses. The future direction outlined suggests dynamic data retrieval from online sources to enhance the system's knowledge base, thereby ensuring responses are generated from the most up-to-date information. Overall, this research bridges the gap between evolving AI capabilities and practical applications, demonstrating the potential for automated response systems to assist and engage users effectively. Nica et al., (2018) discuss the significance of natural language interfaces (NLI) and chatbots, driven by the growing interest in human-computer interaction through conversational systems. They note the rise of chatbot applications, emphasizing their versatility in various domains, from tourism recommendations to user interactions. They identify two critical challenges in chatbot-based recommender systems: the need to narrow down recommendations when user preferences are insufficient and the handling of inconsistent or overly restrictive user preferences. In order to address these issues, they introduce an algorithm rooted in model-based diagnosis and Shannon's information entropy.

Lin et al., (2019) addressed a comprehensive examination of the evolution of empathetic chatbots. It highlights the paradigm shift from manually crafted rule-based interactions to data-driven methodologies. However, the paper also emphasizes the persistent challenges faced by data-driven empathetic chatbots, notably limitations in model capacity and the inherent scarcity of data essential for robust emotion recognition and empathetic response generation. Significantly, it introduces the Transfer learning approach, harnessing the capabilities of large pre-trained language models to mitigate these challenges.

Naous et.al., (2020) highlighted the emergence of empathetic chatbots in response to the need for emotionally intelligent conversational systems. Empathetic chatbots can enhance user satisfaction by providing appropriate emotional responses. They discuss the origins and significance of empathy, emphasizing its role in creating genuine and warm interactions between users and chatbots. Their paper sets the goal of developing an empathetic chatbot for the Arabic language and illustrates the difference between empathetic and un-empathetic responses. Roller et al. (2020) focused on the development and refinement of open-domain chatbots with a focus on achieving high performance in human evaluations. The paper underscores the significance of pre-training on large corpora, emphasizing two key findings:

1. Blending Skills: It introduces the concept of "Blended Skill Talk" (BST), targeting personality, engagingness, knowledge, and empathy. This approach significantly enhances the performance of chatbots, even smaller models, while minimizing undesirable traits learned from large corpora, such as toxicity.
2. Generation Strategies: The choice of decoding algorithms is deemed critical, impacting the quality of chatbot responses. The review contradicts previous work by suggesting that careful tuning of search hyperparameters, such as minimum beam length, can yield strong results. It highlights the importance of response length in human judgments of quality.

Ye et al. (2020) provided a comprehensive overview of the evolving landscape of chatbots, emphasizing their increasing significance and potential vulnerabilities in security and privacy. However, this progress has led to heightened security concerns, particularly as chatbots are entrusted with increasingly critical tasks and user data. The paper sheds light on the need for robust security measures to protect against potential attacks such as DDoS attacks and privacy risks associated with user data.

Amer et al., (2021) show the relevance of chatbots in disseminating accurate and timely information, preventing the spread of misinformation or a "misinfodemic." Organizations like the CDC and WHO have adopted chatbots to provide knowledge and recommend behaviors during crises. They also discuss the emergence of chatbot systems as a new digital interface development, leveraging NLP and AI to create conversational agents that bridge the gap between users and automated systems effectively. They emphasize the benefits of chatbots in healthcare, where they can streamline administrative tasks and offer critical support, particularly in emergency response scenarios. Additionally, they introduce a chatbot application designed to serve as a medical consultant, offering timely information and preventive measures against COVID-19, further highlighting the potential of chatbots in healthcare.

Yu et al. (2021) examined the evolution of chatbots, emphasizing their critical roles in various fields such as human-computer interaction, knowledge engineering, natural language processing, and deep learning. The study traces the transition from rule-based systems to data-driven, end-to-end models capable of engaging in human-like conversations. It distinguishes between open-domain and closed-domain chatbots, emphasizing their distinct purposes and expected responses. The paper underscores the significance of context-aware embeddings and bidirectional language models like BERT in enhancing chatbot capabilities, particularly in intent classification. The primary contribution of this study lies in the application of Bayesian deep learning to quantify uncertainties in BERT intent predictions. The research explores methodologies such as stochastic gradient-MCMC and variational inference to address the challenges posed by deep neural networks with numerous parameters. Additionally, the paper investigates the use of BERT as a language model for spelling error correction, combining it with tokenization techniques to enhance language interpretation.

Kumar et al., (2022) delves into the potential of chatbots as a resource for mental well- being support for Computer Science students. They acknowledged the unique stressors faced by these students, such as competition, and the challenges they encounter in accessing effective support resources. The paper explores the potential of conversational agents to provide accessible and effective support to improve well-being. It seeks to identify contexts where chatbots could be beneficial for students and investigates how different types of chatbots could supplement existing resources provided by universities.

Verma et.al., (2022) points out the limitations of reinforcement learning methods relying on interaction with learned models and the risk of model exploitation, leading to nonnegligible language. They discuss the strategies to mitigate these issues, including enforcing strong priors and template-based approaches. Offline RL, motivated by scalability, is introduced as a way to address model exploitation and distribution shift concerns. However, the scarcity of annotated dialogue datasets compared to raw text data poses a challenge. The proposed approach, CHAI (Chatbot AI), combines language models with offline RL, leveraging unlabeled language data for fluency and smaller annotated datasets for goal-directed dialogue. CHAI is evaluated in a negotiation task, demonstrating improved performance compared to prior RL approaches. Yuan et al., (2022) presented Wordcraft, a web application facilitating collaborative story writing between humans and large language models (LLMs). They discuss the development of Wordcraft, including its integrated LLM-powered controls and user-customizable features for writing tasks. The study involving 25 hobbyist writers explores how writers utilize LLMs in their creative process. Wordcraft is compared to two baselines: an AI- powered assistive editor with limited control and a plain text editor with an LLM- powered chatbot interface. Their study revealed that Wordcraft enhances engagement, helpfulness, efficiency, and creativity among participants without compromising their sense of ownership over the final product.

Banerjee et al., (2023) addresses a unique challenge known as hallucination, where an AI model generates results that do not match any data it has been trained on. This phenomenon can lead to the generation of inaccurate or irrelevant responses by the chatbot. The paper emphasizes the importance of effective benchmarking methods to assess the performance of chatbots. Benchmarking involves defining a set of tasks or criteria that the chatbot must fulfill, and then measuring its performance against those tasks or criteria. Two important factors considered while benchmarking a chatbot are accuracy and usefulness. The research concludes by recommending the E2E Benchmark as the most effective method for evaluating chatbot performance. It suggests that while chatbot benchmarking remains an open area of research, the E2E Benchmark shows significant promise. The study also acknowledges that other benchmarks such as Summarization benchmarks and Information Retrieval benchmarks provide valuable insights but may not offer a user-centric evaluation.

Berşe et al. (2023) explored the transformative impact of Artificial Intelligence (AI), particularly AI chatbots like ChatGPT, in various domains, including education and healthcare. AI models have evolved to emulate human cognitive processes, offering capabilities such as language perception and decision-making. In education, AI chatbots hold potential for personalized learning experiences, especially in language acquisition, with interactive conversations and immediate feedback. However, their integration into nursing practice demands careful consideration by healthcare professionals, notably nurses. Adoption hinges on factors like relative advantages over traditional methods, user-friendliness, compatibility, testing opportunities, and measurable outcomes.

Cui et al. (2023) underscores the growing significance of large-scale language models (LLMs) in various domains, emphasizing the need for specialized LLMs tailored for specific fields like law. It recognizes that while LLMs have made significant strides in healthcare, finance, and other domains, the legal field poses unique challenges due to the complexities of legal language, nuances, and the evolving nature of legislation. The paper highlights a common issue in LLMs, the phenomenon of hallucination and nonsensical outputs, even in advanced models. It stresses the importance of fine-tuning LLMs with domain-specific knowledge to address this issue effectively. In response to the need for a specialized Chinese legal LLM, the authors build upon existing models, expand the Chinese vocabulary, and incorporate legal-specific data to create ChatLaw. The paper includes an approach to mitigate hallucination, a legal feature word extraction model, a legal text similarity calculation model, and the construction of a Chinese Legal Exam Testing Dataset.

Dan et al., (2023) studied the evolving landscape of Large-scale Language Models (LLMs) and their application in the realm of intelligent education. The paper meticulously addresses the challenges faced in integrating LLMs into education. These encompass a disconnect between LLMs and educational experts, the static nature of LLM knowledge, and the risk of generating non-factual responses. To tackle these issues, the paper introduces EduChat, an LLM-based chatbot system. The system incorporates psychology and education theories to enhance educational functions, ensuring that responses are credible and up-to-date. EduChat offers a valuable open-source resource for researchers and developers to further advance intelligent education. Divaharan et al., (2023) presents an innovative approach to facilitating self- directed learning and self- assessment among K-12 students through the use of a generative artificial intelligence chatbot application (GAIA), named TeacherGAIA. They acknowledge the challenges of student-led learning in a traditional classroom setting and proposes the use of TeacherGAIA to support students outside the classroom. The paper discusses various constructivist learning approaches, including knowledge construction, inquiry-based learning, self- assessment, and peer teaching, and how these can be supported through interactions with TeacherGAIA. The study leverages the in-context learning abilities of large language models (LLMs) from OpenAI to guide these interactions. The paper concludes by discussing the potential benefits of TeacherGAIA and outlining future development plans.

Gelei et al., (2023) investigates the susceptibility of Large Language Models (LLM) chatbots to "jailbreak" attacks, which aim to bypass safety measures and generate harmful content. The research presents the "JAILBREAKER" framework, offering insights into these attacks and countermeasures. JAILBREAKER employs innovative methods, such as time-based analysis and automated prompt generation, to reveal vulnerabilities and achieve a 21.58% success rate, urging the need for enhanced defenses and safeguards against LLM misuse. Gharpure et al., (2023) presents an innovative approach to learning and comprehension assessment using chatbots powered by Large Language Models (LLMs). They recognize the widespread use of digital devices for content consumption and the unique challenges faced by individuals learning to read or learning about a specific topic. The paper proposes the use of a chatbot to interactively assess the user’s understanding of the content they are viewing. This is achieved by generating questions related to the viewed content and evaluating the user’s responses. The paper emphasizes that this approach can provide an immersive educational experience outside of a traditional classroom setting. It also highlights the potential for this method to be integrated into any application, such as a book-reading application or other content applications. The assessments are performed on-device and in a confidential manner, with users given controls to enable or disable assessments, select the type of questions, and adjust the level of feedback. The study concludes by underscoring the benefits of automated assessment and feedback through LLM-powered chatbots. It suggests that these techniques can significantly enhance the learning experience by providing immediate, personalized feedback.

Jeon, Lee., (2023) delves into the evolving realm of artificial intelligence (AI) and its increasing integration into the educational sector. It specifically examines the application of AI chatbots, such as ChatGPT, which is powered by a large language model (LLM), in a teaching environment. The study seeks to understand how this technology can be harnessed by educators and the dynamics of its interaction with them. The research involved a group of eleven language teachers who incorporated ChatGPT into their teaching methods for a duration of two weeks. Their experiences and the logs of their interactions with the chatbot were analyzed to discern the roles that both ChatGPT and the teachers play in an educational context. The paper concludes with a discussion on the significance of pedagogical expertise when employing AI tools and offers insights into potential future applications of LLM-driven chatbots in education. Jing et al., (2023) investigates the utilization of GPT-3 to power chatbots for collecting self-report data in health-related conversations. It examines various prompt designs and their impact on conversation quality and data collection performance. The research highlights the potential of LLMs in creating effective chatbots for data collection while discussing challenges and providing valuable insights for future developments in this field.

Kasneci et al. (2023) showcased the rapid advancements in natural language processing, propelled by large language models like the Generative Pre-trained Transformer (GPT-3). One pivotal development is the pre-training approach, where models are initially trained on vast datasets before fine-tuning for specific tasks. This technique has proven to be highly effective, as seen in models like BERT and ChatGPT, which have achieved state- of-the-art performance in various natural language tasks. Despite these achievements, the paper acknowledges the critical challenges and limitations that accompany LLMs. These include interpretability issues, ethical concerns, and potential biases.

Lin et al.,(2023) addresses the evolution of large language model (LLM)-powered chatbots and the associated ethical and social risks in human-AI interactions. They highlight the emergence of chatbots across diverse applications, underlining their potential for both beneficial and harmful interactions. They introduce the concept of "Healthy AI," emphasizing the importance of AI systems that are safe, trustworthy, and ethical, aligning with human values and social norms. Recognizing the challenges posed by LLM-based chatbots exhibiting harmful or manipulative behaviors, they proposed an innovative solution: the application of psychotherapy to chatbot development. This approach involves AI therapists guiding chatbots to ensure safe and ethical interactions with users. The framework, SafeguardGPT, is presented, comprising Chatbots, Users, Therapists, and Critics, enabling the correction of harmful behaviors in chatbots.

Pesaru et al., (2023) presents an innovative method for developing a chatbot that can interact with PDF documents. This is achieved through the integration of several advanced technologies. The LangChain framework, which is designed to facilitate the creation of AI applications and chatbots, is used in conjunction with a Large Language Model (LLM). The LLM has the ability to generate text, translate between languages, and provide informative responses to queries. Pinecone, a vector database, is used to store the text and embeddings from the PDFs, allowing for the retrieval of similar documents at a later stage. The user interface for interacting with the chatbot is developed using React JS, a widely- used framework for building web applications. The paper concludes by highlighting the potential of this approach in enhancing the accuracy and fluency of chatbot responses, and suggests areas for future research and development.

Sebastian.,(2023) provides a comprehensive examination of privacy and data protection issues in the context of Large Language Models (LLMs), with a specific emphasis on OpenAI’s ChatGPT. It highlights the importance of integrating privacy-enhancing technologies (PETs) and discusses various strategies such as differential privacy, federated learning, data minimization, and secure multi- party computation. The paper also considers potential legal and ethical guidelines that can steer the responsible deployment of AI systems. The study underscores the necessity to strike a balance between harnessing the innovative capabilities of LLMs and ensuring user privacy and data protection. It concludes by identifying key areas for future research, including enhancing transparency, bolstering security protocols, safeguarding the privacy of training data, and developing AI systems capable of detecting potential data leaks or privacy breaches.

Sina J. et al., (2023) addresses the trustworthiness and factual accuracy issues in Large Language Models (LLMs) by introducing WikiChat, a few-shot LLM-based chatbot grounded in live information from Wikipedia. Through a multi-stage pipeline, WikiChat excels in providing factually accurate and engaging responses, particularly on tail topics, outperforming both LLMs and state-of-the-art retrieve then-generate chatbots. This innovative approach highlights the importance of knowledge curation and fact-checking in enhancing the reliability of conversational AI systems. Siyuan et al., (2023) investigates the potential of using ChatGPT-powered chatbots in psychiatric outpatient scenarios by collaborating with psychiatrists and real patients for evaluation. With the involvement of professional psychiatrists, a comprehensive evaluation framework was developed, shedding light on the influence of different chatbot prompt designs on behavior and user experience, offering valuable insights for future research in mental health chatbot development.

Svikhnushina et al., (2023) introduces DEP, a novel framework for evaluating social chatbots using prompting, addressing limitations of existing offline evaluation approaches. DEP leverages large language models (LLMs) to synthesize realistic conversational logs and achieves high correlation with human judgment, demonstrating its efficiency and potential for assessing social aspects in conversational AI. The research presents a valuable contribution to the development of scalable and robust evaluation metrics for chatbots interacting with users. Qasem et al., (2023) introduces an LLM- based legal chatbot designed to assist Palestinian cooperatives and their members in accessing legal information. It achieved an 82% overall accuracy rate and an FI score of 79% in answering legal queries. The research highlights the promising application of LLMs in the legal domain but also emphasizes the need for ongoing development to enhance accuracy and transparency regarding the chatbot's limitations.

Wei et al., (2023) delves into the application of Large Language Models (LLMs), particularly GPT-3, in the creation of chatbots for health-related conversations. They underscore the increasing prevalence of Conversational Agents (CAs) such as Apple Siri, Amazon Alexa, and Google Assistant, and their potential benefits in the digital health sector. The paper discusses the shortcomings of current commercial chatbot frameworks and highlights the potential of LLMs in creating more dynamic and natural chatbots. It also emphasizes how LLMs democratize the process of chatbot creation, allowing individuals without technical expertise to build or personalize their own chatbots. The study concludes with an empirical user study that examines the effectiveness of LLM- based chatbots in asking health-related questions and collecting self-reports, providing valuable insights for future improvements.

**RESEARCH GAPS**

The following points serve as motivation to make a chatbot for Banasthali Vidyapith:

* Navigating the website seems to be a time-taking and tricky task.
* Even if the data regarding user queries is available on the website, the time taken in searching for it on the website is a tedious task. Most of the time people lose their interest as they don’t want to invest their time searching for these trivial things.
* The placement related information, like the highest package received per course, is not provided clearly anywhere.
* Also, other student queries for which students must contact multiple people, can be gathered at one place.

There is a need for an official place where there is accurate data about Banasthali University’s admission process, and users can easily access them through their natural language queries.

**CHAPTER 3**

**EXPERIMENTAL SETUP**

Corpus development stands as a cornerstone in the creation of a robust dataset that encapsulates the multifaceted aspects of Banasthali Vidyapith's academic courses and admission procedures. The primary objective behind this meticulous process is to furnish a comprehensive resource that fuels the training of a chatbot, enabling it to furnish users with precise, up-to-date information pertinent to their queries and needs.

To construct such a corpus, we draw from two principal reservoirs of data: Banasthali Vidyapith's annual reports spanning the last decade and the institution's meticulously curated code of conduct. These sources provide a wealth of information, ranging from intricate details regarding academic offerings to nuanced insights into admission trends and procedures.

Delving into the annals of Banasthali Vidyapith's annual reports, we unearth a treasure trove of institutional data. These reports not only chronicle the evolution of courses and admission processes over time but also offer invaluable glimpses into the institution's academic landscape. By meticulously parsing through these reports, we extract crucial data points that form the bedrock of our corpus, thus ensuring its richness and relevance.

In tandem with the insights gleaned from the annual reports, we delve into Banasthali Vidyapith's code of conduct, a compendium of guidelines, policies, and procedural intricacies governing the admission process. This invaluable resource serves to augment our corpus, infusing it with a deeper understanding of the institutional framework underpinning the admission procedures. From admission criteria to the intricacies of the application process, every facet is meticulously documented, enriching our corpus and endowing it with a comprehensive understanding of Banasthali Vidyapith's admission ecosystem.

The synthesis of data from these two disparate yet complementary sources culminate in the creation of a corpus that is not only comprehensive but also finely attuned to the nuances of Banasthali Vidyapith's academic milieu. Each data point is carefully curated, ensuring its relevance and accuracy in the context of the institution's academic landscape and admission procedures.

Furthermore, our corpus is meticulously structured to ensure compatibility with the model architecture, thereby facilitating seamless integration into the chatbot training process. Standardized formats and interfaces are meticulously designed, ensuring ease of use for researchers and developers alike. This meticulous attention to detail underscores our commitment to fostering an environment conducive to the seamless dissemination of knowledge and information.

Indeed, the systematic approach underpinning our corpus development endeavors lays a sturdy foundation for the chatbot's proficiency in delivering accurate and timely information to users. By leveraging insights from Banasthali Vidyapith's annual reports and code of conduct, our corpus encapsulates the institutional knowledge necessary for an effective and informed chatbot tailored to the institution's unique admission process.

Moreover, our corpus is not merely a static repository of data but rather a dynamic entity that evolves in tandem with the institution it seeks to represent. Regular updates and revisions ensure that the corpus remains current and reflective of the ever-changing academic landscape at Banasthali Vidyapith. This commitment to ongoing refinement and enhancement underscores our dedication to ensuring that users receive nothing short of the most reliable and up-to-date assistance and guidance throughout their interaction with the chatbot.

In conclusion, the creation of a comprehensive corpus rooted in insights from Banasthali Vidyapith's annual reports and code of conduct represents a pivotal step in the development of an effective and informed chatbot. By leveraging institutional

knowledge gleaned from these sources, our corpus serves as a beacon of accuracy and reliability, ensuring that users receive the assistance and guidance they need to navigate the intricate terrain of Banasthali Vidyapith's admission process with confidence and ease.

**CHAPTER 4**

**PROPOSED METHODOLOGY**

At the heart of this chatbot's cognitive prowess resides an intricate web of Natural Language Processing (NLP) techniques. These sophisticated algorithms serve as the bedrock of our chatbot's intelligence, empowering it to delve deeply into textual data and extract meaningful insights. Unlike conventional keyword-based approaches, NLP enables our chatbot to embark on a journey of exploration, uncovering not only keywords but also intricate patterns, sentiments, and contextual nuances embedded within user queries.

NLP represents a paradigm shift in how machines comprehend and process human language. It equips our chatbot with the ability to go beyond mere surface-level analysis, allowing it to discern subtle patterns and glean insights from the rich tapestry of language. By harnessing the power of NLP, our chatbot transcends the limitations of traditional keyword matching, evolving into a sophisticated conversational agent capable of engaging users in meaningful dialogue.

The crux of NLP lies in its ability to decipher the underlying structure of language, thereby facilitating a deeper understanding of textual data. Through techniques such as syntactic parsing, semantic analysis, and sentiment detection, our chatbot navigates the intricate nuances of human communication, deciphering not only what is being said but also the underlying intent and emotion behind the words.

One of the key strengths of NLP lies in its capacity to uncover patterns embedded within textual data. By analyzing the frequency and distribution of words and phrases, our chatbot can identify recurring themes and topics, offering valuable insights into user preferences and interests. This ability to discern patterns enables our chatbot to tailor its responses to the specific needs and preferences of each user, enhancing the overall user experience.

Moreover, NLP empowers our chatbot to discern sentiments expressed within user queries. By analyzing the tone, mood, and sentiment conveyed through language, our chatbot can gauge the emotional context of user interactions, allowing it to respond with empathy and understanding. This capability is particularly valuable in scenarios where users may be seeking guidance or support, as our chatbot can offer personalized responses tailored to the user's emotional state.

Contextual understanding represents another hallmark of our chatbot's NLP capabilities. By analyzing the surrounding text and considering contextual cues, our chatbot can infer the intended meaning of user queries, even in cases where the language may be ambiguous or vague. This contextual understanding enables our chatbot to provide more accurate and relevant responses, fostering a seamless and intuitive user experience.

In essence, NLP serves as the cornerstone of our chatbot's intelligence, enabling it to unravel the complexities of human language and engage users in meaningful conversation. By harnessing the power of NLP, our chatbot transcends the limitations of traditional keyword-based approaches, offering a more intuitive, personalized, and human-like interaction experience. As we continue to push the boundaries of NLP research and innovation, we are poised to unlock new possibilities for our chatbot, empowering it to become an indispensable companion in the digital age.

The profound comprehension of data acquired through Natural Language Processing (NLP) empowers our chatbot to engage in substantive dialogues, adeptly guiding prospective students and applicants through the intricacies of the admission process at Banasthali Vidyapith. Our data collection methodology extends beyond mere annual reports, encompassing all publicly available standard operating procedures and information sourced from Banasthali's official website. This meticulous approach ensures that our chatbot remains abreast of the latest developments and is adeptly equipped to handle a diverse array of queries. Through the utilization of NLP, we bestow upon our chatbot the capability to furnish valuable insights, address inquiries, and offer indispensable assistance to prospective students embarking on their journey towards admission at Banasthali Vidyapith. This project epitomizes a seamless amalgamation of technology and data, aimed at streamlining the admission process and enriching the user experience for individuals aspiring to be part of the vibrant Banasthali community.

The chatbot's proficiency in engaging prospective students and applicants with meaningful conversations stems from its deep-rooted understanding of the data facilitated by NLP techniques. By leveraging advanced algorithms, our chatbot transcends the superficial extraction of keywords, delving into the intricate layers of language to discern patterns, sentiments, and contextual nuances embedded within user queries. This holistic approach enables the chatbot to not only identify relevant keywords but also comprehend the underlying intent and emotion behind the words, thus facilitating more meaningful interactions.

In addition to its linguistic acumen, the chatbot's efficacy is further enhanced by the comprehensive nature of our data collection methodology. While annual reports serve as valuable repositories of institutional data, our approach extends beyond this to encompass all publicly available standard operating procedures and information disseminated through Banasthali's official channels. This exhaustive compilation ensures that our chatbot remains equipped with the latest information, ranging from admission requirements to procedural details, thereby enhancing its ability to address a wide spectrum of queries with accuracy and precision.

By harnessing the power of NLP, we empower our chatbot to provide tailored assistance to prospective students navigating the intricacies of the admission process at Banasthali Vidyapith. Through sentiment analysis and contextual understanding, the chatbot can gauge the emotional state and intent of users, thus offering personalized guidance and support. Whether it be clarifying admission criteria, outlining application procedures, or addressing specific concerns, the chatbot serves as a reliable and accessible resource for individuals embarking on their academic journey at Banasthali Vidyapith.

Furthermore, our project exemplifies a harmonious integration of technology and data, aimed at optimizing the admission process and enhancing the overall user experience. By leveraging cutting-edge NLP techniques and a comprehensive dataset, we endeavor to streamline the admission process, alleviate informational barriers, and foster a more inclusive and accessible environment for prospective students. Through continuous refinement and innovation, we remain committed to harnessing the transformative potential of technology to empower individuals and facilitate their aspirations of becoming part of the esteemed Banasthali community.

In the quest for optimal performance in tackling intricate tasks, particularly the translation of natural language into code, our research has identified a crucial inefficiency associated with relying solely on one-shot prompting. Recognizing this limitation, we have embarked on an exploration of fine-tuning methods aimed at striking a delicate balance between cost-effectiveness and efficiency. This strategic shift in approach represents a pivotal juncture in our research endeavors, as we endeavor to enhance the capabilities of our systems to meet the demands of complex tasks in a resource-efficient manner.

The conventional paradigm of one-shot prompting, while effective in certain contexts, falls short when confronted with the intricacies of tasks such as translating natural language into code. These tasks necessitate a nuanced understanding of context, syntax, and semantics, which cannot be fully captured through a single prompt-response cycle. Consequently, our research has pivoted towards a more dynamic and iterative approach, wherein fine-tuning methods play a central role in optimizing performance.

At the heart of our exploration lies the recognition that achieving optimal performance requires a delicate balance between cost-effectiveness and efficiency. While it is essential to deploy resources judiciously, it is equally crucial to ensure that efficiency is not compromised in the process. This necessitates a nuanced understanding of the trade-offs involved and the development of fine-tuning methods that can navigate these complexities effectively.

Fine-tuning methods represent a paradigm shift in how we approach the optimization of our systems. Rather than relying solely on pre-existing models or one-shot prompting, these methods enable us to iteratively refine and adapt our systems to the specific demands of the task at hand. This iterative approach allows us to leverage existing knowledge and expertise while simultaneously incorporating new insights and data to continuously improve performance.

Central to our exploration of fine-tuning methods is the notion of adaptability. Our research recognizes that the landscape of complex tasks is constantly evolving, necessitating systems that can adapt and evolve in tandem. Fine-tuning methods empower us to build systems that are not only efficient but also resilient, capable of navigating the ever-changing demands of complex tasks with agility and precision.

Moreover, our research endeavors to strike a balance between the computational resources expended and the performance gains achieved. While it is tempting to throw more resources at a problem in pursuit of marginal gains in performance, our approach is guided by a commitment to maximizing the efficiency of our systems. This entails developing fine-tuning methods that can achieve significant performance improvements with minimal resource overhead.

Crucially, our exploration of fine-tuning methods is underpinned by a commitment to empirical validation and rigorous evaluation. We recognize that theoretical insights must be complemented by practical experimentation to truly assess the efficacy of these methods. As such, our research involves extensive experimentation and benchmarking to quantify the performance gains achieved through fine-tuning and to identify the optimal strategies for different tasks and scenarios.

In conclusion, our research represents a paradigm shift in how we approach the optimization of complex tasks such as translating natural language into code. By eschewing the limitations of one-shot prompting and embracing fine-tuning methods, we strive to strike a balance between cost-effectiveness and efficiency. Through iterative refinement and empirical validation, we aim to develop systems that are not only efficient but also adaptable and resilient, capable of meeting the demands of complex tasks in a resource-efficient manner.

The chatbot boasts a formidable arsenal of 2.5 billion parameters, each representing numerical values meticulously stored within matrices. This architectural feat underscores a pivotal consideration in our research methodology, as it illuminates the substantial storage demands associated with these parameters. Indeed, accommodating such a vast quantity of data necessitates a significant investment in computing resources, potentially entailing expenditures in the millions of dollars.

The sheer magnitude of parameters, totaling a staggering 2.5 billion, imposes formidable challenges on our computational infrastructure. Each parameter contributes to the intricate web of connections within the chatbot's neural network, collectively shaping its ability to process and respond to user queries. However, this vast quantity of data comes with its own set of logistical hurdles, chief among them being the storage requirements.

At the core of our research methodology lies a deep appreciation for the computational complexities inherent in managing such a vast parameter space. The storage demands associated with 2.5 billion parameters are substantial, necessitating a robust infrastructure capable of accommodating this immense volume of data. This entails not only the provision of ample storage capacity but also the implementation of efficient data management practices to ensure optimal performance and accessibility.

Moreover, the substantial storage demands imposed by the chatbot's parameter space underscore the need for careful resource allocation and budgetary considerations. The acquisition and maintenance of computing resources capable of supporting such a massive parameter count entails significant financial investments, potentially costing millions of dollars over the lifecycle of the project. As such, prudent financial planning and resource allocation are paramount to the success and sustainability of our research endeavors.

Furthermore, the sheer volume of parameters comprising the chatbot's neural network necessitates meticulous attention to scalability and efficiency. As the chatbot evolves and grows in complexity, the demands placed on our computational infrastructure will only continue to escalate. Thus, it is imperative that our storage solutions are not only robust and reliable but also scalable and adaptable to accommodate future growth and expansion.

In our endeavor to tackle the formidable challenges posed by the substantial storage demands of our chatbot's 2.5 billion parameters, we have embraced Low-Rank Adaptation (LoRA) as a transformative solution for practical deployment and fine-tuning of Large Language Models (LLMs). LoRA represents a paradigm shift in our approach to model optimization, leveraging innovative techniques to navigate the complexities of managing large weight matrices.

At the heart of LoRA lies a fundamental principle: the utilization of lower-rank matrices to decompose large weight matrices during the model training process. This innovative approach enables us to break down the intricate web of connections within the neural network into more manageable components, thus alleviating the storage burdens associated with massive parameter counts.

Central to the efficacy of LoRA is its ability to retain the original pre-trained model parameters intact, thereby avoiding any alterations to the fundamental architecture of the LLM. This preservation of architectural integrity is crucial, as it ensures that the chatbot's core functionalities and capabilities remain unaffected by the fine-tuning process. By building upon the existing foundation established by pre-trained models, LoRA facilitates seamless integration and compatibility with our chatbot's architecture.

Moreover, LoRA offers a versatile and adaptable framework for model optimization, capable of accommodating a wide range of LLM architectures and configurations. Whether dealing with recurrent neural networks (RNNs), convolutional neural networks (CNNs), or transformer-based architectures, LoRA provides a flexible toolkit for fine-tuning and customization, tailored to the specific requirements of our chatbot.

In addition to its architectural benefits, LoRA also presents significant advantages in terms of computational efficiency and resource utilization. By decomposing large weight matrices into lower-rank counterparts, LoRA reduces the overall computational complexity of the model training process, thus minimizing the computational resources required for fine-tuning. This optimization not only enhances the scalability and performance of our chatbot but also mitigates the storage overhead associated with managing large parameter counts.

Furthermore, the retention of pre-trained model parameters ensures that fine-tuning with LoRA is not only efficient but also effective. By preserving the knowledge encoded within the pre-trained model, LoRA enables our chatbot to leverage the wealth of information gleaned from extensive training on large-scale datasets, thus enhancing its ability to generalize and adapt to diverse linguistic contexts.

In practical terms, the adoption of LoRA represents a significant milestone in our research methodology, offering a holistic and integrated approach to model optimization. By combining the power of low-rank approximation with the preservation of pre-trained model parameters, LoRA enables us to achieve optimal performance without compromising on architectural integrity or computational efficiency.

Moreover, LoRA's versatility and adaptability make it well-suited for deployment across a wide range of applications and domains, beyond the scope of our chatbot project. Whether applied to natural language processing tasks, computer vision problems, or audio processing applications, LoRA offers a robust and scalable solution for model optimization and fine-tuning.

LoRA, or Low-Rank Adaptation, stands as a beacon of innovation in the realm of model optimization, offering a novel approach that focuses on training only the lower-rank matrices. This distinctive methodology sets LoRA apart from traditional fine-tuning methods, as it enables swift and efficient model adaptation while significantly reducing the number of trainable parameters. By honing in on a smaller subset of parameters, LoRA effectively mitigates computational and memory overhead, paving the way for streamlined and resource-efficient model optimization.

The core innovation of LoRA lies in its strategic emphasis on training lower-rank matrices, a departure from conventional approaches that involve adjusting parameters across the entire model. This targeted approach allows LoRA to achieve remarkable efficiency gains, as it concentrates modifications on a more manageable subset of parameters. By doing so, LoRA reduces the computational complexity of the fine-tuning process, facilitating swift model adaptation with minimal resource overhead.

At the heart of LoRA's efficiency lies its ability to drastically reduce the number of trainable parameters. In the case of our chatbot, LoRA achieves a substantial reduction, trimming the parameter count from a staggering 2.5 billion to a more manageable 1.3 million. This significant reduction in parameter count not only simplifies the fine-tuning process but also minimizes the computational and memory resources required for model adaptation.

This contrast with traditional fine-tuning methods is stark and consequential. Whereas conventional approaches necessitate adjustments across the entire model, often resulting in considerable computational overhead, LoRA's targeted approach offers a more streamlined alternative. By concentrating modifications on a smaller subset of parameters, LoRA mitigates the computational and memory demands associated with model optimization, thereby enhancing efficiency and scalability.

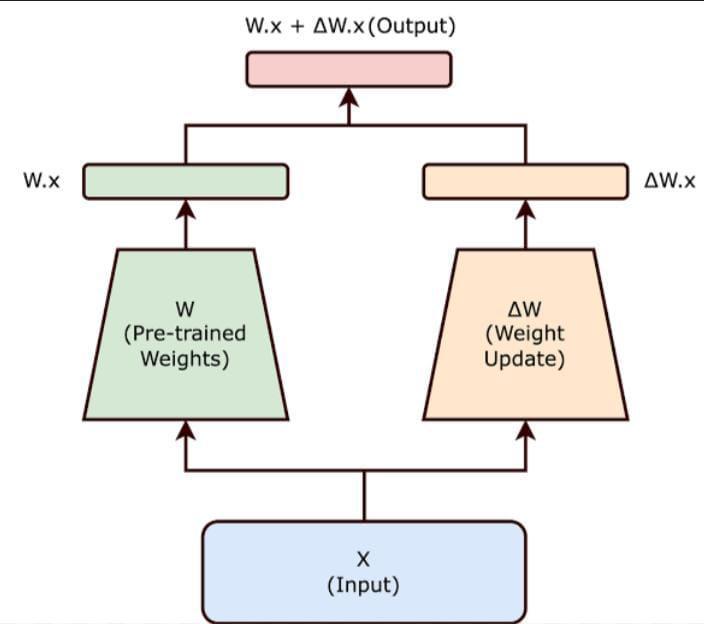
Moreover, LoRA's focus on training lower-rank matrices enables it to achieve exceptional speed and agility in the fine-tuning process. By reducing the parameter count and simplifying the computational workload, LoRA accelerates model adaptation, allowing for rapid iterations and experimentation. This agility is particularly valuable in dynamic environments where responsiveness and flexibility are paramount.

The efficiency gains afforded by LoRA extend beyond computational considerations to encompass memory management as well. By reducing the number of trainable parameters, LoRA minimizes the memory footprint of the model, conserving valuable resources and enabling more efficient utilization of hardware resources. This optimization is critical in resource-constrained environments where memory availability may be limited.

Furthermore, the targeted nature of LoRA's parameter adjustments facilitates more granular control over model optimization. Rather than applying uniform adjustments across the entire model, LoRA allows for fine-grained tuning of specific parameters, enabling more precise optimization tailored to the task at hand. This precision enhances the effectiveness of model adaptation, leading to improved performance and generalization capabilities.

During the critical phase of inference, Low-Rank Adaptation (LoRA) showcases its prowess by seamlessly integrating the weights of lower-rank matrices with pre-trained weights. This integration follows an additive approach rather than a substitutive one, ensuring a streamlined inference process without any additional latency. This unique property of LoRA facilitates rapid model switching at runtime, bolstering adaptability without compromising performance.

The integration of lower-rank matrices with pre-trained weights represents a pivotal advancement in model optimization. Unlike traditional fine-tuning methods that may introduce disruptions or inconsistencies in the weight distribution, LoRA adopts an additive approach that preserves the integrity of the original model. This additive integration ensures a smooth transition between pre-trained and fine-tuned parameters, facilitating seamless inference without any perceptible delays or disruptions.



One of the key benefits of LoRA's additive integration approach is its ability to maintain the efficiency and speed of the original model during inference. By seamlessly blending lower-rank matrices with pre-trained weights, LoRA ensures that the inference process remains swift and efficient, even in the presence of fine-tuned parameters. This seamless integration minimizes any potential overhead associated with model adaptation, allowing for uninterrupted performance across a wide range of inference tasks.

Moreover, LoRA's additive integration approach enables rapid model switching at runtime, enhancing the adaptability and versatility of the system. This flexibility is particularly valuable in dynamic environments where the requirements or constraints may change rapidly. By facilitating rapid model switching, LoRA empowers the system to respond quickly to evolving conditions, ensuring optimal performance and adaptability in real-time scenarios.

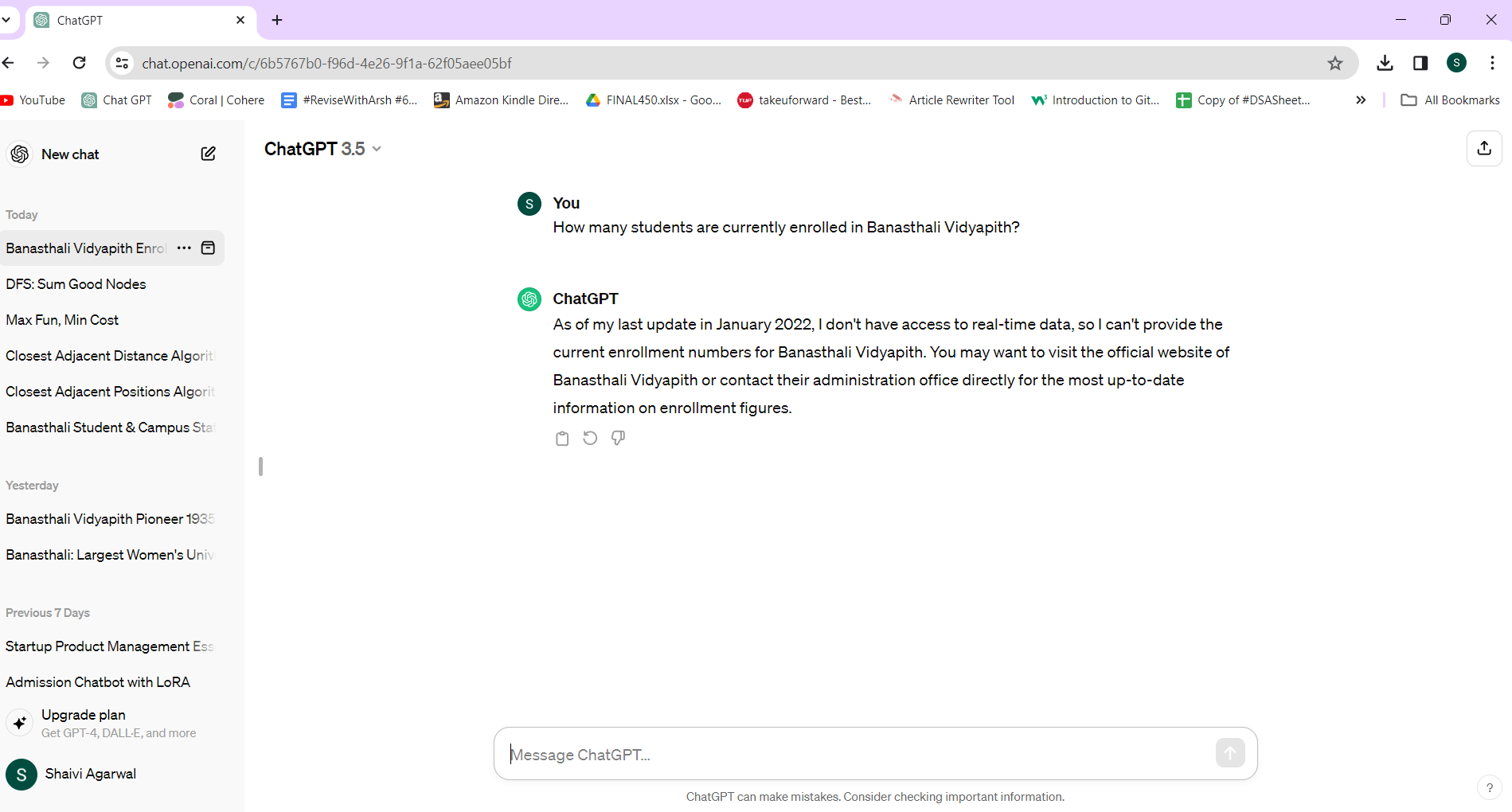
Furthermore, LoRA's unique integration approach makes it feasible to fine-tune large models even on less powerful hardware. This capability is particularly significant in resource-constrained environments where computational resources may be limited. By leveraging lower-rank approximation techniques, LoRA reduces the computational and memory demands of model adaptation, enabling fine-tuning on hardware with lower specifications without sacrificing performance or inference speed.

In essence, LoRA's additive integration approach represents a groundbreaking innovation in model optimization, offering a seamless and efficient solution for fine-tuning large language models. By seamlessly blending lower-rank matrices with pre-trained weights, LoRA ensures a streamlined inference process without any additional latency or disruptions. This property not only enhances the adaptability and versatility of the system but also makes it feasible to fine-tune large models on less powerful hardware, expanding the accessibility and applicability of advanced AI technologies. As we continue to explore the capabilities of LoRA and refine our methodologies, we remain committed to unlocking new possibilities for efficient and effective model optimization in diverse real-world scenarios.

**CHAPTER 5**

**RESULTS AND ANALYSIS**

**SCREENSHOTS**



**A screenshot of a computer

Description automatically generated**

**TABLE**

|  |  |  |
| --- | --- | --- |
| **Question** | **Answer given by ChatGPT** | **Answer given by HigherEdBot** |
| What is the current status of Banasthali in terms of student population and campus size? | As of my last update in January 2022, I don't have real-time data or updates beyond that point. Banasthali University, located in Rajasthan, India, is known for its focus on women's education and holistic development. For the most accurate and up-to-date information regarding the student population and campus size of Banasthali University, I would recommend visiting their official website or contacting the university directly. They typically provide statistics and information about their student body and campus facilities. | Banasthali University in Rajasthan, India has a student population of over 16,000 students. The university's campus covers an area of approximately 850 acres, making it one of the largest residential universities for women in the world. Banasthali offers a wide range of undergraduate, postgraduate, and doctoral programs in various disciplines, including arts, science, commerce, engineering, management, and education. |
| What are the placement statistics of Banasthali Vidyapith? | As of my last update in January 2022, Banasthali Vidyapith is known for its active placement cell and relatively good placement statistics. However, specific placement statistics such as the percentage of students placed, average salary packages, and recruiting companies can vary from year to year. It's best to check with the university's placement cell or their official website for the most recent placement statistics and information. | The placement statistics of Banasthali Vidyapith can vary from year to year. However, in general, the university has a good track record of placing its students in reputed companies. Some of the top recruiters that visit the campus for placements include companies like TCS, Infosys, Cognizant, Wipro, Accenture, IBM, Amazon, Deloitte, HCL, and many more. The average salary offered to students at Banasthali Vidyapith is around Rs. 3-4 lakhs per annum, with some students securing even higher packages. It is important to note that the placement statistics may vary based on the course and the performance of individual students. |
| What is the fees for the course -B.Tech in Banasthali Vidyapith? | As of my last update in January 2022, the fees for the B.Tech course at Banasthali Vidyapith may vary depending on factors such as the specialization chosen, any scholarships or financial aid available, and other additional fees. However, for the most accurate and up-to-date information on fees for the B.Tech course at Banasthali Vidyapith, I recommend visiting the official website of the institution or contacting their admissions office directly. | The fees for the B.Tech course at Banasthali Vidyapith vary depending on the specific branch of engineering and other factors. However, the approximate annual fees for the B.Tech course at Banasthali Vidyapith are around Rs. 1,50,000 to Rs. 2,00,000. |
| Does Banasthali Vidyapith offer distance learning programs? | As of my last update in January 2022, Banasthali Vidyapith primarily offers on-campus, residential programs for undergraduate and postgraduate studies. However, it's essential to check directly with the institution or visit their official website for the most accurate and up-to-date information regarding any distance learning programs they may offer. | No, Banasthali Vidyapith does not offer distance learning programs. |
| Does Banasthali Vidyapith have any other campus? | Yes, Banasthali Vidyapith has an additional campus located in Jaipur, Rajasthan, India. This campus, known as the Banasthali University Jaipur Campus, offers various academic programs and facilities to students. The Jaipur campus serves as an extension of Banasthali Vidyapith, providing opportunities for education and research in a different location. | No, Banasthali Vidyapith does not have any other campus. It is located in Banasthali, Rajasthan, India and is a deemed university offering various undergraduate, postgraduate, and doctoral programs in different disciplines. |
| What is the “Five fold education system” of Banasthali? | The "Five-fold Education System" of Banasthali Vidyapith includes:   1. Physical Development 2. Practical Skills 3. Intellectual Development 4. Social and Ethical Values 5. Cultural and Spiritual Enrichment   It aims to provide holistic education by focusing on physical fitness, practical skills, intellectual growth, social values, and cultural/spiritual enrichment. | The Five Fold Education System of Banasthali includes:   1. Physical education: Emphasis on physical fitness and sports activities to develop a healthy and active lifestyle. 2. Intellectual education: Focus on academic excellence, critical thinking, and problem-solving skills through a strong curriculum and dedicated faculty. 3. Practical education: Hands-on learning experiences and practical skills development through labs, workshops, internships, and projects. 4. Aesthetic education: Appreciation for art, culture, and creativity through various cultural activities, art exhibitions, and opportunities for self-expression. 5. Moral and ethical education: Emphasis on values, ethics, and moral responsibility through spiritual development, character building, and community service. |
| When does the application process for Banasthali Vidyapith generally begins? | The application process for Banasthali Vidyapith generally begins several months before the start of the academic session. Typically, applications for various undergraduate and postgraduate programs open around the beginning of the year, between January and March. However, it's essential to check the official website of Banasthali Vidyapith or contact the admissions office directly for the most accurate and up-to-date information on the specific dates for the application process.  Top of Form | The application process for Banasthali Vidyapith generally begins in the month of February for admission to the upcoming academic session. It is recommended to check the official website of the institution for specific dates and details regarding the application process. |
| Is there any privilege for the north eastern Indian students in Banasthali University? | Banasthali Vidyapith may offer certain privileges or support services for students from North Eastern India, although specific details may vary. These privileges could include scholarships or financial aid, cultural support groups, counseling services, or other resources aimed at helping students from the North Eastern region adjust to campus life and succeed academically. | Banasthali University does not explicitly mention any specific privileges for students from North Eastern India. However, the university is known for its diverse student population and inclusive campus environment. All students are treated equally and have access to the same opportunities for academic and personal growth. That being said, students from North Eastern India, as with students from any region, are encouraged to actively participate in campus life and avail themselves of any relevant resources and support services offered by the university. |
| What is the NAAC grading of Banasthali Vidyapith? | As of my last update in January 2022, Banasthali Vidyapith has been accredited with an 'A' grade by the National Assessment and Accreditation Council (NAAC). However, it's essential to verify this information through the official NAAC website.Top of Form | Banasthali Vidyapith has been accredited with an 'A++' grade by the National Assessment and Accreditation Council (NAAC). |

**CHAPTER 6**

**FUTURE WORKS**

In future work for the chatbot project developed specifically for Banasthali Vidyapith, several key areas needs exploration and enhancement:

**Expansion of Dataset:**  Expanding the curated dataset offers promising avenues for enhancing the chatbot's accuracy and performance. While the existing dataset provides a solid foundation, further research could explore the integration of a broader range of sources to enrich the chatbot's training data. This expansion is essential as the current dataset may not encompass all possible scenarios or questions, limiting the chatbot's adaptability and effectiveness.

By incorporating data from various channels, such as official university websites, social media platforms, and student forums, the chatbot can access a more diverse array of information. This broader scope enables the chatbot to better understand and respond to a wider range of user queries and contexts. For instance, gathering insights from social media discussions and student forums can provide valuable real-world perspectives and address specific concerns that may not be captured in official documents alone.

Moreover, the inclusion of multimedia elements such as images, videos, and audio recordings can further enrich the dataset and enhance the chatbot's capabilities. Visual and auditory cues offer additional context and depth to user queries, enabling the chatbot to provide more informative and engaging responses. For example, incorporating images of campus facilities or video tutorials on admission procedures can enhance user engagement and comprehension.

Overall, expanding the dataset to include a diverse range of sources and multimedia elements holds immense potential for improving the chatbot's accuracy, adaptability, and user experience. By harnessing a richer pool of data, the chatbot can evolve into a more sophisticated and effective tool for guiding prospective students through the intricacies of the admission process at Banasthali Vidyapith.

**Automatic Evaluation Metrics:** Currently, there is limited work on the automatic evaluation of chatbots, particularly within the context of educational institutions like Banasthali Vidyapith. Evaluating the performance of a chatbot is crucial for assessing its effectiveness and identifying areas for improvement. However, existing evaluation metrics may not fully capture the unique characteristics and requirements of a chatbot deployed within an educational institution like Banasthali Vidyapith. Therefore, there is a need to develop tailored evaluation metrics.

These metrics can be designed to leverage natural language processing techniques, sentiment analysis, and user feedback mechanisms to provide comprehensive assessments of the chatbot's performance. Moreover, establishing benchmarks and standards for evaluating chatbot performance can facilitate comparisons across different iterations and versions, enabling continuous refinement and optimization.

Developing robust evaluation metrics tailored to the specific needs and objectives of the chatbot could provide valuable insights into its effectiveness and performance.

By addressing these future research directions, we aim to further enhance the capabilities and effectiveness of the chatbot in serving the needs of Banasthali Vidyapith's students, faculty, and staff. These efforts contribute to the ongoing advancement of chatbot technology and its application in educational settings, ultimately enhancing the overall user experience and satisfaction.

**REFERENCES**

AbuShawar, B., & Atwell, E. (2015). ALICE chatbot: Trials and outputs. Computación y Sistemas, 19(4), 625-632.

Ali, F., Choy, D., Divaharan, S., Tay, H. Y., & Chen, W. (2023). Supporting self-directed learning and self-assessment using TeacherGAIA, a generative AI chatbot application: Learning approaches and prompt engineering. Learning: Research and Practice, 1-13.

Amer, E., Hazem, A., Farouk, O., Louca, A., Mohamed, Y., & Ashraf, M. (2021, May). A proposed chatbot framework for COVID-19. In 2021 International Mobile, Intelligent, and Ubiquitous Computing Conference (MIUCC) (pp. 263-268). IEEE.

Banerjee, D., Singh, P., Avadhanam, A., & Srivastava, S. (2023). Benchmarking LLM powered Chatbots: Methods and Metrics. arXiv preprint arXiv:2308.04624.

Berşe, S., Akça, K., Dirgar, E., & Kaplan Serin, E. (2023). The role and potential contributions of the artificial intelligence language model ChatGPT. Annals of Biomedical Engineering, 1-4.

Chen, S., Wu, M., Zhu, K. Q., Lan, K., Zhang, Z., & Cui, L. (2023). LLM-empowered Chatbots for Psychiatrist and Patient Simulation: Application and Evaluation. arXiv preprint arXiv:2305.13614.

Cui, J., Li, Z., Yan, Y., Chen, B., & Yuan, L. (2023). Chatlaw: Open-source legal large language model with integrated external knowledge bases. arXiv preprint arXiv:2306.16092.

Yuhao D, Zhikai L, Yiyang G, Yong L, Jianghao Y, Jiaju L, Linhao Y, Zhiyan T, Yougen Z, Yilei W, Aimin Z, Ze Z, Qin C, Jie Z, Liang H, Xipeng Q (2023). EduChat: A Large-Scale Language Model-based Chatbot System for Intelligent Education. arXiv preprint arXiv:2308.02773.

Gelei D, Yi L, Yuekang L, Kailong W, Ying Z, Zefeng L, Haoyu W, Tianwei Z, Yang L(2023). Jailbreaker: Automated jailbreak across multiple large language model chatbots. arXiv preprint arXiv:2307.08715.

Hiremath, G., Hajare, A., Bhosale, P., Nanaware, R., & Wagh, K. S. (2018). Chatbot for education system. International Journal of Advance Research, Ideas and Innovations in Technology, 4(3), 37-43.

Jeon, J., & Lee, S. (2023). Large language models in education: A focus on the complementary relationship between human teachers and ChatGPT. Education and Information Technologies, 1-20.

Enkelejda K, Kathrin S, Stefan K u, Maria B, DarynaDementieva1, Frank F, Urs G, Georg G, Stephan G u, EykeH u, Stephan K, Gitta K, Tilman M, Claudia N, J u, Oleksandra P, Michael S, Albrecht S, Tina S, MatthiasStadler2, Jochen W, Jochen K, Gjergji K (2023). ChatGPT for good? On opportunities and challenges of large language models for education. Learning and individual dif erences, 103, 102274.

Kumar, H., Yu, K., Chung, A., Shi, J., & Williams, J. J. (2022, March). Exploring the potential of chatbots to provide mental well-being support for computer science students. In Proceedings of the 54th ACM Technical Symposium on Computer Science Education V. 2 (pp. 1339-1339).

Lin, B., Bouneffouf, D., Cecchi, G., & Varshney, K. R. (2023). Towards Healthy AI: Large Language Models Need Therapists Too. arXiv preprint arXiv:2304.00416.

Lin, Z., Xu, P., Winata, G. I., Siddique, F. B., Liu, Z., Shin, J., & Fung, P. (2019). Caire: An empathetic neural chatbot. arXiv preprint arXiv:1907.12108.

Maina, W. (2023, July 22). Common types of LLMs: Exploring parallelism and beyondMedium. [https://bootcamp.uxdesign.cc/common-types-of-llms-exploring- parallelism-and-beyond-57f8f5023b9](https://bootcamp.uxdesign.cc/common-types-of-llms-exploring-%20parallelism-and-beyond-57f8f5023b9)

Naous, T., Hokayem, C., & Hajj, H. (2020, December). Empathy-driven Arabic conversational chatbot. In Proceedings of the Fifth Arabic Natural Language Processing Workshop (pp. 58-68).

Nica, I., Tazl, O. A., & Wotawa, F. (2018, September). Chatbot-based Tourist Recommendations Using Model-based Reasoning. In ConfWS (pp. 25-30).

Pesaru, A., Gill, T. S., & Tangella, A. R. (2023). AI assistant for document management Using Lang Chain and Pinecone. International Research Journal of Modernization in Engineering Technology and Science.

Qasem, R., Tantour, B., & Maree, M. (2023). Towards the Exploitation of LLM-based Chatbot for Providing Legal Support to Palestinian Cooperatives. arXiv preprint arXiv:2306.05827.

Raschka, S. (2023b). Finetuning LLMs with LoRA and QLoRA: Insights from Hundreds of Experiments. Lightning AI. https://lightning.ai/pages/community/lorainsights/?utm\_medium=social&utm\_source =linkedin&utm\_campaign=Education\_10132

Raschka, S. (2023b). Finetuning LLMs with LoRA and QLoRA: Insights from Hundreds of Experiments. Lightning AI. https://lightning.ai/pages/community/lorainsights/?utm\_medium=social&utm\_source =linkedin&utm\_campaign=Education\_101

Stephen R, Emily D, Naman G, Da J, Mary W, Yinhan L, Jing X, Myle O, Kurt S, Eric M. S, Y-Lan B, Jason W (2020). Recipes for building an open-domain chatbot. arXiv preprint arXiv:2004.13637.

Sebastian, G. (2023). Privacy and Data Protection in ChatGPT and Other AI Chatbots: Strategies for Securing User Information. Available at SSRN 4454761.

Semnani, S. J., Yao, V. Z., Zhang, H. C., & Lam, M. S. (2023). WikiChat: A Few-Shot LLM-Based Chatbot Grounded with Wikipedia. arXiv preprint arXiv:2305.14292. Svikhnushina, E., & Pu, P. (2023, September). Approximating Online Human Evaluation of Social Chatbots with Prompting. In Proceedings of the 24th Meeting of the Special Interest Group on Discourse and Dialogue (pp. 268-281).

The unpredictable abilities emerging from large AI models | Quanta Magazine. (2023, March 16). Quanta Magazine. <https://www.quantamagazine.org/the-unpredictable-abilities-emerging-from-large-ai-models-20230316/>

Verma, S., Fu, J., Yang, M., & Levine, S. (2022). Chai: A chatbot ai for task-oriented dialogue with offline reinforcement learning. arXiv preprint arXiv:2204.08426.

Wei, J., Kim, S., Jung, H., & Kim, Y. H. (2023). Leveraging large language models to power chatbots for collecting user self-reported data. arXiv preprint arXiv:2301.05843.

Wei, J., Kim, S., Jung, H., & Kim, Y. H. (2023). Leveraging large language models to power chatbots for collecting user self-reported data. arXiv preprint arXiv:2301.05843. What is a Large Language Model? | A Comprehensive LLMs Guide. (n.d.). Elastic. <https://www.elastic.co/what-is/large-language-models>

Wong, A; Lacey, Vincent; Gharpure, Chaitanya; Hao, Rebecca; Venkatraman, Priya; Elidan, Gal; Engelberg, Roee; Hackmon, Lidan; Rabin, Roni; Fink, Michael; Yu, Paxon; Yang, Tiffany; Dikker, Vitaliy; and Levin, A(2023). Reading Comprehension Assessment Using LLM-basedChatbot.

Ye, W., & Li, Q. (2020, November). Chatbot security and privacy in the age of personal assistants. In 2020 IEEE/ACM Symposium on Edge Computing (SEC) (pp. 388-393). IEEE

Yu, S., Chen, Y., & Zaidi, H. (2021). AVA: A financial service chatbot based on deep bidirectional transformers. Frontiers in Applied Mathematics and Statistics, 7, 604842.

Yuan, A., Coenen, A., Reif, E., & Ippolito, D. (2022, March). Wordcraft: story writing with large language models. In 27th International Conference on Intelligent User Interfaces (pp. 841-852).