KatzBot - Intelligent ChatBot using Large Language Models

Haider Ali Deepa Paikar Kiran Vutukuri
Yeshiva University
Hali4@mail.yu.edu dpaikar@mail.yu.edu kvutukur@mail.yu.edu

Abstract

In the digital landscape of higher education, accessing relevant and timely information from university websites can be a laborious process. Yeshiva University, with its wealth of data spanning departments, academic programs, faculty details, and more, encounters challenges in delivering a seamless user experience. In response, this project introduces KatzBot, a sophisticated virtual assistant powered by Large Language Models (LLMs), designed to streamline information retrieval and enhance user engagement.

KatzBot is specifically tailored for Yeshiva's Katz School of Science and Health, serving as a dynamic and comprehensive information hub. Its primary objective is to offer real-time, personalized information to prospective students, current students, alumni, and anyone seeking insights into the university. The motivation behind KatzBot lies in alleviating the time-consuming nature of traditional information retrieval methods, especially for those interested in academic programs, admission processes, campus facilities, financial aid options, faculty details, and campus life.

1. Introduction

The rise of artificial intelligence (AI) and natural language processing (NLP) technologies has led to a paradigm shift in human-computer interactions. One area where this shift is particularly pronounced is the development and deployment of chatbots. Chatbots, powered by sophisticated algorithms and machine learning, have become integral components of various online platforms, ranging from customer service applications to personal virtual assistants. This research project delves into the design, implementation, and evaluation of an advanced chatbot system aimed at enhancing user experience and functionality. As the demand for intelligent conversational agents continues to grow, understanding the intricacies of chatbot development and user engagement becomes imperative.

This study seeks to contribute to the evolving landscape of conversational AI by presenting a comprehensive analysis of the architecture, capabilities, and user satisfaction of the developed chatbot. The subsequent sections will outline the methodology employed, present empirical results, and discuss the broader implications of the findings for the future of interactive AI systems. Moreover, the project aims to offer practical insights into optimizing chatbot performance, addressing challenges, and fostering user trust. By exploring the intersection of technology and human interaction, this research endeavors to shape the trajectory of AI-driven conversational interfaces, ensuring their seamless integration into our digital lives.

The development of KatzBot follows a robust Data Science pipeline. Initially, we engaged in extensive data collection, extracting pertinent information from diverse web pages both within and outside Yeshiva University's website. This dataset formed the foundation for KatzBot's knowledge base. The dataset consists of 5000 sentence completion pairs and 5000 question-answer pairs. To refine the model's understanding and response capabilities, we implemented fine-tuning exercises using Large Language Models. Largely, language models (LLMs) usher in a novel paradigm for constructing chatbots, adept at comprehending and responding to natural language prompts with a nuanced understanding, enabling more sophisticated and context-aware conversations [10]. This transformative capability of LLMs enriches the user experience, fostering a more intuitive and dynamic interaction in the realm of chatbot development.

2. Methodology

Our question answering chatbot development required data collection, processing, and model building. The main challenge was to collect the dataset and process it. The other challenge was the computational resources required for training the models for long time.

The first phase of fine-tuning focused on sentence completion, enhancing the model's proficiency in generating contextually relevant information, ad allowing model to understand the semantics. In the second phase, we fine-tuned the same model on a question-answer dataset, ensuring KatzBot's efficacy in responding to a diverse array of user questions. This approach ensures that KatzBot not only understands the context of the information but can also provide precise and accurate responses.

As KatzBot takes will be place on the university's website, the implications for user experience will be profound. Prospective students can receive personalized information about academic programs and admission processes, aiding in their decision-making journey. Current students can access details about campus facilities and campus life, contributing to a more enriched university experience. Alumni seeking information about ongoing developments can find relevant updates effortlessly.

2.1. Data collection and processing

In the intricate process of developing a chatbot grounded in the wealth of information from Yeshiva websites, the journey commenced with a meticulous definition of the project's scope. This initial step involved delineating the specific topics, subjects, and types of information that the chatbot would adeptly handle. The aim was to create a conversational agent well-versed in a diverse array of Yeshiva-related domains, including but not limited to course details, teacher profiles, school locations, and international rules.

To gather the requisite data for the fine-tuning of Language Models (LLMs), a robust and extensive dataset was imperative. Recognizing the significance of both quality and variety in data points, the approach to data acquisition commenced with automated scraping attempts. Python, with its versatile libraries, notably BeautifulSoup and requests, was employed for this task. However, the initial attempts encountered substantial challenges in the data engineering phase. The data extracted through automated means exhibited a notable presence of gibberish and inconsistencies, complicating the subsequent stages of development.

Undeterred by the challenges posed by automated scraping, a strategic shift was made toward manual data collection. This involved hands-on visits to Yeshiva websites, ensuring a firsthand gathering of accurate and relevant information. The manual extraction process was thorough and comprehensive, covering a spectrum of topics that aligned with the predefined scope of the chatbot. Each piece of information, ranging from course details to teacher profiles, was meticulously organized and recorded.

The resulting dataset, born from this manual curation, took the form of structured text files. Each file was thoughtfully organized to align with specific categories or types of information. This meticulous structuring facilitated not only clarity but also the seamless integration of the dataset into the subsequent stages of development. The systematic arrangement ensured that the chatbot could draw upon a well-organized and categorized repository of information, enhancing its efficacy and responsiveness.

This manual curation process not only addressed the challenges encountered during automated scraping but also introduced a layer of human intuition and oversight. It allowed for the discernment of nuances, contextual understanding, and the extraction of information that might have eluded automated processes. The result was a dataset that not only met the quantitative requirements for fine-tuning but also boasted a qualitative richness born from the nuanced insights of human curation.

The shift to manual curation, while introducing an additional layer of effort, proved to be a strategic decision. It not only mitigated the challenges associated with automated scraping but also contributed to the development of a robust and reliable dataset. The organized structure of the dataset laid a solid foundation for subsequent stages of fine-tuning LLMs and the overall development of a chatbot poised to excel in delivering accurate and contextually relevant information within the realm of Yeshiva-related inquiries.

As the project advances, the dataset serves as the cornerstone upon which the chatbot's conversational abilities are honed and refined. The integration of human-curated data ensures that the chatbot not only responds accurately but also engages in meaningful and contextually aware conversations, embodying the essence of a knowledgeable virtual assistant within the domain of Yeshiva-related information.

After completing the data collection phase, the focus shifted to data pre-processing. A parsing mechanism was developed to extract information from the text files, utilizing techniques such as regular expressions or specific parsing methods tailored to the data's structure. For sentence completion data, each sentence underwent standardization by removing unnecessary symbols, punctuation, and special characters. Tokenization was then applied to break sentences into words or phrases. Similarly, question-answer pairs were carefully structured, with the removal of irrelevant information and noise to ensure each pair was well-formatted and consistent.

To enhance the quality of the dataset, duplicates, and noise were identified and removed, promoting diversity. Tokenization and lemmatization further refined the dataset by breaking down words and reducing them to their base or root form. Optional data augmentation techniques were considered to artificially boost dataset diversity. Quality control was applied through the manual review of a sample of the dataset, addressing any identified errors or inconsistencies.

In the data storage phase, the pre-processed data was organized for efficient retrieval and management. This involved storing the data in a structured database or using common formats such as CSV or JSON for compatibility with machine learning frameworks. These steps resulted in a clean, organized, and well-structured dataset, ready for training the chatbot. It's important to note that adjustments may have been made based on specific data characteristics and chatbot requirements, and adherence to data privacy and copyright considerations was crucial throughout the process.

Data Type	*Description*	*Quantity Generated*
Sentence Completion Pairs	Pairs created for model training	5,600
Question-Answer Pairs	QA pairs for detailed understanding	5,100
Test QA Pairs (Separate)	Pairs for model testing consistency	1000

Figure 1. Collected dataset

For fine-tuning purposes, two types of datasets were created. The first type consisted of pairs of sentence completions, and the second type comprised question-answer pairs. This approach resulted in the generation of a comprehensive dataset, including 5,600 sentence completion pairs and 5,100 question-answer pairs, providing diverse input for the subsequent stages of model training and refinement. For testing, we've created separate 1000 Questing Answers pairs and used them for each model for the comparison purpose.

2.2. Model Building

On this foundation of the Transformer architecture, specifically as articulated in the seminal "Attention Is All You Need" paper by Vaswani et al. [9], our models are constructed. This architecture is pivotal in granting our models the capability to process entire sentences at once, a departure from the traditional word-by-word processing employed by LSTMs and RNNs.

In the pursuit of fine-tuning and customization, we have identified and leveraged key parameters integral to our approach. The QLoRA parameters, for instance, encompass a LoRA attention dimension of 32, an alpha parameter for LoRA scaling set to 16, and a dropout probability for LoRA layers set at 0.1. Furthermore, our model adopts a 4-bit precision approach for base model loading, specifying a compute type of "float16" and a quantization type of "nf4" for bits and bytes. This deliberate choice serves to effectively reduce the overall model size.

The training process is orchestrated with meticulous attention to various parameters. Gradient checkpointing is employed to manage memory efficiently, and the maximum gradient norm is set at 0.3 to regulate gradient magnitudes. A cosine learning rate schedule is adopted, optimizing the learning rate over the course of training. Our optimization strategy involves the use of the paged-adamw-32bit optimizer.

An emphasis on efficiency permeates the training approach. Sequences are intelligently grouped into batches of similar lengths, a strategy that minimizes padding, thereby conserving memory and expediting the training process. Moreover, our model architecture is designed to accommodate Supervised Finetuning Trainer (SFT), signifying its flexibility for further refinement in a supervised learning context, tailored to specific tasks or domains.

In summary, our model development approach is guided by the transformative principles of the Transformer architecture, augmented by carefully chosen parameters for fine-tuning, precision optimizations, and a strategic training configuration. The amalgamation of these elements results in a model that is not only accurate and efficient but also adaptable to diverse applications[7].

2.2.1 Microsoft Phi 1.5

Microsoft phi 1.5 B LLM model [3], is a Transformer with a formidable 1.3 billion parameters, which stands as a powerful tool intended for research purposes. This language model, crafted with the same data sources as its predecessor phi-1 and enriched with synthetic NLP texts, showcases near-state-of-the-art performance across common sense, language understanding, and logical reasoning benchmarks, all while maintaining a model size below 10 billion parameters. Notably, phi-1.5 prioritizes safety by excluding generic web-crawl data sources like common-crawl from training, minimizing exposure to potentially harmful online content. However, users are advised to exercise caution, given potential limitations such as generating inaccurate code snippets, limited scope for uncommon packages in code generation, and a preference for standard English. As the model is not trained on the web data it doesn't contain toxicity.

I've fine-tuned this model by applying hyperparameter tuning, but due to its size and architecture, it wasn't able to perform better.

2.2.2 Mistral 7B Instruct

Mistral AI introduces Mistral 7B Instruct LLM [2], a groundbreaking language model comprising 7.3 billion parameters—a remarkable achievement in packing immense computational power into a compact design. Renowned as the most formidable language model within its size category, Mistral 7B not only outperforms Llama 2 13B across all benchmarks but also competes favorably with Llama 1 34B on diverse language tasks. Noteworthy features include Grouped-query attention (GQA) and Sliding Window Attention (SWA), enhancing inference speed and efficiently managing extended sequences with reduced computational overhead. Released under the Apache 2.0 license, Mistral 7B provides users with unrestricted access, facilitating local deployment or integration with cloud platforms such as

AWS, GCP, or Azure. It uses the below format for the instruction. Below is an example prompt required for training a fine-tuning Mistral model.

[INST] Answer the following question truthfully: What is the official seal of Yeshiva University? [/INST] Torah U Madda

Beyond benchmark comparisons, Mistral 7B demonstrates innovation in attention mechanisms, employing SWA for a 2x speed improvement in certain scenarios and introducing local attention for optimized cache utilization during inference. Additionally, Mistral 7B showcases its adaptability through fine-tuning public instruction datasets, resulting in Mistral 7B Instruct—a model that surpasses other 7B models on MT-Bench.

Additionally, with fine-tuning, we applied noisy embeddings improve Instruction finetuning with 5 percent noise which helped improving the overall accuracy in Question Answering[5].

Inference based on different temperature We tested the Mistral Instruct 7B with different temperatures and as expected the model performance changes according to the temperature.

Here are some examples:

Question:
Yeshiva University's MS AI program is:

Answer with 0.9 temperature:
Yeshiva University's MS in AI is an interdisciplinary program designed for visionaries like you. It offers a blend of traditional Jewish studies and cutting-edge AI technologies, creating a unique environment for a global impact. The program prepares students to become pioneers in AI integration, driving innovation across various sectors. With a focus on ethics and responsibility, students learn to use AI for social impact.

Answer with 0.5 temperature:
Yeshiva University's MS AI program is an interdisciplinary program designed for students to master artificial intelligence and its applications in various fields. Students gain skills in AI technologies, including machine learning, computer vision, natural language processing, robotics, and more. The program is STEM-OPT eligible, and international students may be eligible for up to 36 months of Optional Practical Training (OPT). The program is also aligned with industry certifications, such as the Certified Artificial Intelligence Professional (CAIP). The MS AI program offers full-time, part-time, or evening options, and international students may be eligible for up to 36 months of Curricular Practical Training (CPT). The program is also STEM-OPT eligible. Students may enter the program in January, May, or September. Application deadlines are rolling, and there are multiple entry points throughout the year. Tuition for the program is 25,000 USD.

Answer with 0.1 temperature:
Yeshiva University's MS AI program is a 30-credit Master of Science in Artificial Intelligence, offered on campus in New York City, with full-time or part-time options and evening courses to accommodate working students. The program focuses on designing and building cutting-edge AI technologies for various applications, including finance, biotech, health care, cybersecurity, and ad-tech. Students gain hands-on experience with tools like Python, R, SQL/NoSQL, TensorFlow, Keras, and pyTorch. The program also covers AI product design, business strategy, and ethical considerations. It is STEM-OPT eligible, and international students may be eligible for up to 36 months of Optional Practical Training (OPT).

2.2.3 Amazon's MistralLite

MistralLite, a refined iteration of the Mistral-7B-v0.1 [2] language model, excels in processing extended contextual information with a capacity to handle up to 32K tokens. The model leverages an adapted Rotary Embedding and sliding window mechanism during fine-tuning, showcasing superior performance in long context retrieval and answering tasks, all while preserving the simplicity of the original model structure. MistralLite finds utility in various applications such as long context line and topic retrieval, summarization, and question-answering. Deployable on a single AWS g5.2x instance with Sagemaker Huggingface Text Generation Inference (TGI) endpoint, MistralLite caters to high-performance demands in resource-constrained environments. Comparatively, MistralLite outshines Mistral-7B-Instruct-v0.1 by addressing limitations in handling longer contexts, as evidenced in evaluations about topic retrieval, line retrieval, pass key retrieval, and question-answering. Notably, MistralLite achieves a test set accuracy of 64.4 percent and a hard subset accuracy of 56.2 percent, surpassing Mistral-7B-Instruct-v0.1 in handling long input texts[6].

I've fine-tuned this model parallelly, but there wasn't a major difference from the original Mistral 7B Instruct. Below is the graph from eval loss during the fine-tuning phase.

Not many experimentations were done with Mistral Lite model as there was no improvement in the Rouge Score. Hence, no good results received.

2.2.4 Llama 7B

Llama 2 [8], a creation of Meta under the meta-llama repository, represents a remarkable leap in the field of large language models (LLMs). Developed and publicly released by Meta, Llama 2 is a collection of pre-trained and fine-tuned

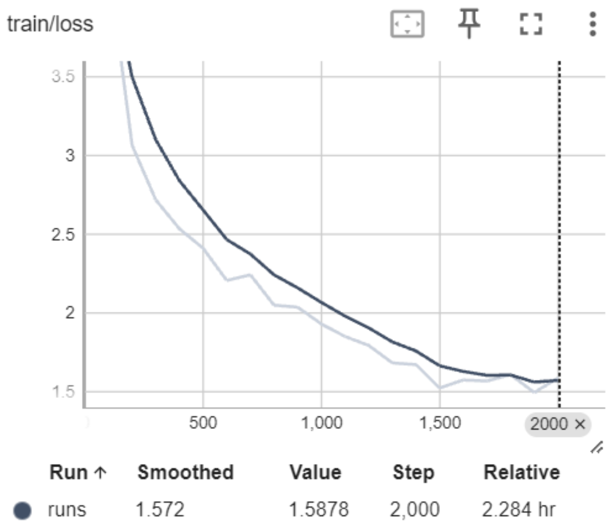


Figure 2. Training Loss during fine tuning

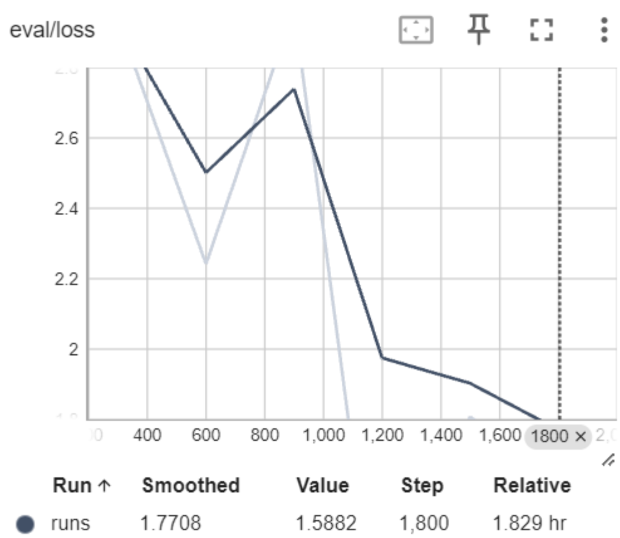


Figure 3. Evaluation Loss during fine tuning

generative text models that vary in scale, ranging from 7 billion to 70 billion parameters.Llama 2 offers a range of variations, including 7B, 13B, and 70B models. The input models are designed to handle input text only, while the output models specialize in generating text exclusively. The architecture of Llama 2 is auto-regressive, employing an optimized transformer architecture. The tuned versions leverage supervised fine-tuning (SFT) and reinforcement learning with human feedback (RLHF) to align with human preferences for helpfulness and safety.

For training data, the Llama 2 family utilizes a new mix of publicly available online data, with token counts referring to pretraining data only. The 70B models incorporate Grouped-Query Attention (GQA) for improved inference scalability. Llama 2 stands as a significant advancement in generative text models, offering a diverse range of models catering to various scales and applications.

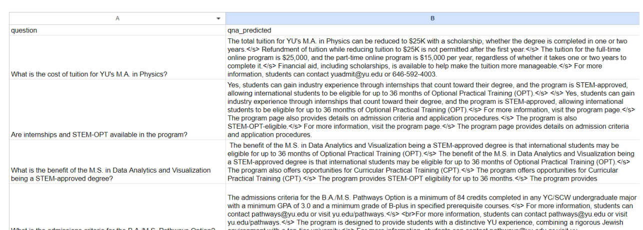
Fine tuning Llama2- We organized our dataset in the specified format on Hugging Face, tailoring it for the sentence completion model.

[INST] Complete the statement: The Innocence Project moved from the 11th floor of Brookdale Center to a new office space. [/INST] The move allowed the Innocence Project to hire more staff and significantly increase the number of cases it takes.

Initially, we fine-tuned the Llama 2 7B model using the complete set of sentence completion data. Subsequently, we saved the resulting model on Hugging Face, yielding a model file of approximately 9 GB. and creating question-answer pairs for subsequent fine-tuning on Llama.

[INST] What is the primary research focus of Dr. Assa Cohen? [/INST] Dr. Assa Cohen's research primarily focuses on financial stability and market microstructure.

Following this, we embarked on a second fine-tuning phase, this time utilizing the entire question-answer dataset. The computational demands were notable, with each epoch taking approximately 40 minutes. Given these constraints, we opted to train the model for a duration of 5 epochs. During this process, we observed discernible differences in model performance between training for 2 epochs and the extended 5-epoch training regimen. These findings underscore the impact of training duration on model behavior and effectiveness. The initial loss rate commences at 3.5 and gradually diminishes to nearly 0.2.



A	B
question	qna_predicted
What is the cost of tuition for YU's M.A. in Physics?	The total tuition for YU's M.A. in Physics can be reduced to $25K with a scholarship, whether the degree is completed in one or two years.</s> Refundment of tuition while reducing tuition to $25K is not permitted after the first year.</s> The tuition for the full-time online program is $25,000, and the part-time online program is $15,000 per year, regardless of whether it takes one or two years to complete it.</s> Financial aid, including scholarships, is available to help make the tuition more manageable.</s> For more information, students can contact yuadmit@yu.edu or 646-592-4003.
Are internships and STEM-OPT available in the program?	Yes, students can gain industry experience through internships that count toward their degree, and the program is STEM-approved, allowing international students to be eligible for up to 36 months of Optional Practical Training (OPT).</s> </s> Yes, students can gain industry experience through internships that count toward their degree, and the program is STEM-approved, allowing international students to be eligible for up to 36 months of Optional Practical Training (OPT).</s> For more information, visit the program page.</s> The program page also provides details on admission criteria and application procedures.</s> The program is also STEM-OPT-eligible.</s> For more information, visit the program page.</s> The program page provides details on admission criteria and application procedures.
What is the benefit of the M.S. in Data Analytics and Visualization being a STEM-approved degree?	The benefit of the M.S. in Data Analytics and Visualization being a STEM-approved degree is that international students may be eligible for up to 36 months of Optional Practical Training (OPT).</s> The benefit of the M.S. in Data Analytics and Visualization being a STEM-approved degree is that international students may be eligible for up to 36 months of Optional Practical Training (OPT).</s> The program also offers opportunities for Curricular Practical Training (CPT).</s> The program offers opportunities for Curricular Practical Training (CPT).</s> The program provides STEM-OPT eligibility for up to 36 months.</s> The program provides
What is the admissions criteria for the B.A./M.S. Pathways Option?	The admissions criteria for the B.A./M.S. Pathways Option is a minimum of 84 credits completed in any YC/SCW undergraduate major with a minimum GPA of 3.0 and a minimum grade of B-plus in specified prerequisite courses.</s> For more information, students can contact pathways@yu.edu or visit yu.edu/pathways.</s> For more information, students can contact pathways@yu.edu or visit yu.edu/pathways.</s> The program is designed to provide students with a distinctive YU experience, combining a rigorous Jewish environment with a top-tier university.</s> For more information, students can contact pathways@yu.edu or visit yu.

Figure 4. Results after finetuning

2.2.5 GPT-2

Generative Pretrained Transformer 2(GPT-2) is an advanced language model by OpenAI, succeeding the original GPT. With a range of models varying in parameters from millions to billions, it demonstrates exceptional natural language understanding and generation capabilities[1]. The transformer-based architecture allows it to excel in diverse language tasks, making it a versatile tool for NLP applications[4].

Standing out for its scalability, this model offers variations in size to suit various computational needs. This flexibility makes it ideal for tasks like fine-tuning and creative text generation, catering to a broad user base. Researchers and developers appreciate its adaptability, enabling them to choose a model that aligns precisely with their specific requirements.

Employing auto-regressive principles, this model predicts the next word based on context. Trained through unsupervised learning on extensive text corpora, it excels in learning grammar, context, and semantics. Its creative text generation capabilities have been demonstrated across tasks such as translation, summarizing, and storytelling, showcasing its impact on advancing natural language processing applications.

3. Related Work

Our project is an extension of the Summer 2023 initiative conducted by students who systematically gathered crucial data from Yeshiva University, forming a text-based dataset and finalizing 133 text files including the information about programs, and professors. The initial phase involved employing the MITIE transformer model used by RASA, chosen for its compact design. However, limitations emerged as the model exhibited constraints in comprehensively capturing question-related information, leading to subpar performance, especially in generating precise answers. Additionally, the model was only able to generate answers for questions provided during the training phase. This extension using LLMs is used to overcome these challenges and elevate the project's effectiveness.

The integration of Large Language Models (LLMs) in the development of chatbots has garnered significant attention in recent research endeavors. Notably, LLMs such as GPT-3 and BERT have been employed to enhance the natural language understanding and generation capabilities of chatbot systems. These models, pre-trained on vast amounts of diverse text data, offer a rich understanding of contextual nuances, enabling chatbots to engage in more contextually relevant and coherent conversations. Studies have explored the adaptability of LLMs in various chatbot applications, ranging from customer support to virtual assistants. Additionally, research efforts have delved into fine-tuning strategies, aiming to tailor pre-trained LLMs to specific chatbot tasks, further optimizing their performance. The incorporation of LLMs in chatbot development represents a promising avenue for advancing conversational AI, providing a foundation for more sophisticated and context-aware interactions between users and automated systems.

4. Methods

In our project, we employ a double fine-tuning methodology on large language models (LLMs) to enhance their adaptability for specific tasks, focusing on sentence completion and question-answering (QA). Initially, a dedicated dataset for sentence completion is curated, emphasizing text completion portions within input sentences. The chosen LLM is then fine-tuned on this dataset to grasp the intricacies of completing sentences. Subsequently, the model's performance is evaluated on sentence completion tasks. Following this, a distinct dataset tailored for QA, consisting of question-answer pairs, is prepared. The pre-trained model from the sentence completion phase serves as the starting point for a second fine-tuning iteration, now emphasizing QA requirements. The model is trained to comprehend questions and generate corresponding answers. Evaluation metrics specific to QA are applied to assess the model's performance in this task. Comparative analysis is conducted between the model's proficiency after the first fine-tuning (sentence completion) and the second fine-tuning (QA). The study concludes by selecting the final model based on its adeptness in both sentence completion and QA, providing insights into its iterative refinement across diverse language understanding tasks.

The evaluation of model-generated outputs was conducted using ROUGE (Recall-Oriented Understudy for Gisting Evaluation), a set of metrics commonly employed for the automatic assessment of text summarization and natural language generation tasks. ROUGE measures the overlap between the model-generated outputs and reference summaries regarding n-gram matches, recall, and precision. Specifically, we utilized ROUGE L, ROUGE 1, and ROUGE 2 scores to assess the quality of our models' outputs. ROUGE L focuses on the longest common subsequence, capturing the overall content similarity between the model-generated and reference sequences. Higher ROUGE scores, particularly in ROUGE L, ROUGE 1, and ROUGE 2, indicate better alignment and similarity between generated and ground truth answers. A ROUGE L score close to 1 signifies a high degree of content overlap, while elevated ROUGE 1 and ROUGE 2 scores suggest accurate reproduction of unigrams and bigrams. The formulas to calculate Rouge Scores are defined below:

4.1. Rouge-L

The Rouge-L score is defined as the longest common subsequence divided by the total number of words in the reference summary.

$$\text{Rouge\_L} = \frac{\text{Longest Common Subsequence}}{\text{Total Words in Reference Summary}} \tag{1}$$

4.2. Rouge-1

The Rouge-1 score is calculated as twice the product of recall and precision divided by their sum.

$$\text{Rouge\_1} = \frac{2 \cdot \text{Recall} \cdot \text{Precision}}{\text{Recall} + \text{Precision}} \tag{2}$$

4.3. Rouge-2

The Rouge-2 score is computed as twice the product of bigram recall and precision divided by their sum.

$$\text{Rouge\_2} = \frac{2 \cdot \text{Bigram Recall} \cdot \text{Bigram Precision}}{\text{Bigram Recall} + \text{Bigram Precision}} \quad (3)$$

5. Results

In first iteration we curated a comprehensive test dataset comprising a blend of question-answer pairs and sentence completion pairs to assess the quality of generated texts. We evaluate the models after sentence completion fine tuning and QA pairs fine tuning through inference. The models subjected to testing are listed below, with the ROUGE values for comparative analysis.

Model	Task	Rouge L	Rouge 1	Rouge 2
Microsoft Phi 1.5B	QA	0.18	0.19	0.06
Microsoft Phi 1.5B	SC	0.09	0.10	0.01
Mistral 7B Instruct	QA	0.43	0.43	0.32
Mistral 7B Instruct	SC	0.24	0.27	0.18
AMZ MistralLite	SC	0.24	0.27	0.17
AMZ MistralLite	QA	0.14	0.16	0.10
Llama 7B	SC	0.66	0.66	0.58
Llama 7B	QA	0.71	0.71	0.62
GPT-2	SC	0.10	0.12	0.02
GPT-2	QA	0.85	0.76	0.85

Table 1. Comparison of models based on Rouge Scores for different tasks

Llama 7B: This model exhibits superior performance on the QA task, achieving high scores on ROUGE L (0.71), ROUGE 1 (0.71), and ROUGE 2 (0.62). The high ROUGE L and ROUGE 1 scores suggest that Llama 7B is particularly adept at generating contiguous text sequences that closely match the reference material in terms of both structure and content. In the SC task, Llama 7B still performs well, though not as strongly as in QA, suggesting that the model's architecture may be particularly tuned for the complexities of generating responsive texts based on queries.

GPT-2: In the QA task, GPT-2's performance is noteworthy with ROUGE L at 0.85, ROUGE 1 at 0.76, and ROUGE 2 at 0.85. These scores indicate that GPT-2 is highly proficient in generating text that aligns closely with the reference summaries, not just in terms of the longest common subsequence but also in capturing exact unigrams and bigrams. This aligns with GPT-2's known strength in producing coherent and contextually relevant text. However, the model's SC task performance is relatively lower, which may indicate its capabilities are more suited to direct question answering rather than filling in gaps within sentences.

Mistral 7B: For the QA task, Mistral 7B's scores are moderate (ROUGE L at 0.43, ROUGE 1 at 0.43, and ROUGE 2 at 0.32), indicating a fair performance in generating relevant answers. However, its performance in SC is notably lower (ROUGE L at 0.24, ROUGE 1 at 0.27, and ROUGE 2 at 0.18), which suggests that the model may struggle with tasks that require inserting contextually appropriate content within a pre-existing sentence structure.

Microsoft Phi 1.5B: This model displays weaker performance in both tasks, with particularly low scores in the SC task (ROUGE L at 0.09, ROUGE 1 at 0.10, and ROUGE 2 at 0.01).

AMZ MistralLite: The AMZ MistralLite shows low ROUGE scores in both tasks, with a slight edge in the QA over the SC task. These results suggest that while MistralLite has some capability in language tasks, it is outperformed by other models, especially in sentence completion, where generating contextually fitting and contiguous text seems to be more challenging.

In the second iteration, we assembled an extensive test dataset consisting of 1000 question-answer pairs to evaluate the proficiency of generated texts. Our assessment involved appraising the final models resulting from sentence completion and those subsequently refined with QA pairs during the inference stage. The listed models underwent testing, and we present their ROUGE values for comparative analysis.

Model	Rouge L	Rouge 1	Rouge 2
Microsoft Phi 1.5B	0.18	0.40	0.25
Mistral 7B Instruct	0.44	0.44	0.33
Llama 7B	0.33	0.34	0.21
GPT-2	0.65	0.66	0.54

Table 2. Comparison of double fine tuned models based on Rouge Scores

Microsoft Phi 1.5B: The scores indicate a moderate performance with a ROUGE L of 0.18, ROUGE 1 of 0.40, and ROUGE 2 of 0.25. The relatively higher ROUGE 1 score suggests that while Phi 1.5B can replicate individual words from the reference text with some success, it struggles more with capturing longer sequences of text.

Mistral 7B Instruct: This model shows a balanced performance with ROUGE L at 0.44, ROUGE 1 also at 0.44, and ROUGE 2 at 0.33. These scores suggest that Mistral 7B is capable of understanding and generating text that closely matches the references at the level of unigrams and longer subsequences, though it is less proficient at capturing bi-gram relationships.

Llama 7B: With ROUGE L at 0.33, ROUGE 1 at 0.34, and ROUGE 2 at 0.21, Llama 7B's performance is moderate. These scores suggest that while the model is capable of generating relevant content, there may be room for improvement in its ability to construct responses that closely follow the phrasing and sequence structure of the reference material, particularly in the construction of bi-grams.

GPT-2: GPT-2 displays strong performance with a ROUGE L score of 0.65, ROUGE 1 of 0.66, and ROUGE 2 of 0.54. These high scores across all metrics indicate that GPT-2 is particularly adept at generating coherent sequences that match the reference text at the level of both unigrams and bigrams. The high ROUGE 2 score, in particular, points to an ability to maintain the integrity of two-word phrases, which is important for preserving the contextual meaning of responses.

6. Discussion

Model Name	Predicted Response
Llama 2	A CRN is a course that meets regularly, often in the evening, and is often offered in the summer……………………
GPT2	A CRN is a course registration number, which is unique to each course
Microsoft Phi	A CRN is a course registration number, which is unique to each course.
Mistral Instruct	A CRN, or Course Registration Number, is a unique identifier for each course offered at Yeshiva University…………….

Figure 5. Predictions after dpuble finetuning

The Rouge scores presented in the above Results tables provide valuable insights into the comparative performance of different models across sentence completion (SC) and question-answering (QA) datasets.

In the initial evaluation, Llama 7B appears to perform exceptionally well on the QA task, with high scores across all ROUGE metrics, indicating a strong ability to generate contiguous, relevant text. GPT-2 comparison shows a high ROUGE L and ROUGE 1 score, which implies its effectiveness in generating semantically and syntactically coherent responses.Mistral 7B in the instruction-based QA task exhibits relatively lower Rouge L, Rouge 1, and Rouge 2 scores compared to its performance in the instruction-based SC task. This discrepancy suggests potential challenges or limitations in capturing nuanced linguistic contexts when generating answers compared to completing sentences, highlighting the model's sensitivity to task intricacies.Microsoft Phi's performance in the question-answering (QA) task reveals lower Rouge scores compared to other models, suggesting potential challenges in capturing relevant information and linguistic nuances for this specific dataset.Similarly, Amazon's MistralLite demonstrates lower Rouge scores in both SC and QA tasks compared to Mistral 7B. While MistralLite exhibits a decline in performance, its scores still provide valuable information regarding its relative efficacy in these language generation tasks.

In summary, the ROUGE scores from the double fine-tuned models demonstrate that GPT-2 stands out for its ability to generate text that is closely aligned with reference summaries, which may make it particularly suitable for applications requiring high-quality natural language generation, such as creating summaries or generating answers to questions in natural language processing tasks.Under a temperature value of 0.8, the responses generated by the GPT model demonstrated unique and discernible characteristics when exposed to the test dataset. The other models, while showing varying levels of capability, generally have lower scores, indicating that they may have specific limitations in the text generation process that could be addressed through further training or refinement. These results can guide future improvements in model training and the selection of models for specific language generation.

The difference in performance between the initial evaluation and the double fine-tuned models suggests that iterative fine-tuning and targeted testing can significantly enhance model performance.

7. Conclusion

In conclusion, the evaluation demonstrates that while some models excel in language generation tasks, the effectiveness of fine-tuning is model-dependent, and certain models may require task-specific optimization. GPT-2's exceptional performance indicates a robust ability to generate text that aligns well with human-generated references, making it a strong option for high-stakes applications where the quality of text generation is paramount. However, the varying performance across models also highlights the complexity of natural language processing tasks and the need for careful consideration when selecting and tuning models for specific purposes. There is a lot room for improvement for further refining our models to enhance their performance by training them further, applying hyperparameters tuning, and adding system prompt such as Chain of Thought Prompting. Notably, we have observed that fine-tuning our models for sentence completion leads to an expansion of the model's knowledge base, accompanied by the observation that the loss function starts at a lower value after this fine-tuning process. Importantly, fine-tuning the model on question-answering (QA) tasks after sentence completion yields improved results compared to directly fine-tuning the models solely on QA pairs. This sequential fine-tuning approach proves advantageous, suggesting that the initial training on sentence completion contributes to a more effective foundation for subsequent QA tasks.

References

[1] Baglivo F. Arzilli G. Privitera G. P. Ferragina P. Tozzi A. E. Rizzo C. De Angelis, L. hatgpt and the rise of large language models: The new ai-driven infodemic threat in public health., 2023. 5

[2] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023. 3, 4

[3] Yuanzhi Li, Sébastien Bubeck, Ronen Eldan, Allie Del Giorno, Suriya Gunasekar, and Yin Tat Lee. Textbooks are all you need ii: phi-1.5 technical report, 2023. 3

[4] Di Li Li Zhou, Jianfeng Gao and Heung-Yeung Shum. The design and implementation of xiaoice an empathetic social chatbot, 2019. 5

[5] S.Karithk Sakthivel V, K. Srihari and C.D Anisha. Intelligent counselling bot using ranking algorithm in ai, 2018. 4

[6] I Shwartz, V. Dagan. Evaluating text representations on lexical composition, 2019. 4

[7] Y.Xiang D.Dai S.Zang, Y. Wu and YWu. Multi-task learning with multilingual transformers. 3

[8] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023. 4

[9] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017. 3

[10] Jing Wei, Sungdong Kim, Hyunhoon Jung, and Young-Ho Kim. Leveraging large language models to power chatbots for collecting user self-reported data. *arXiv preprint arXiv:2301.05843*, 2023. 1