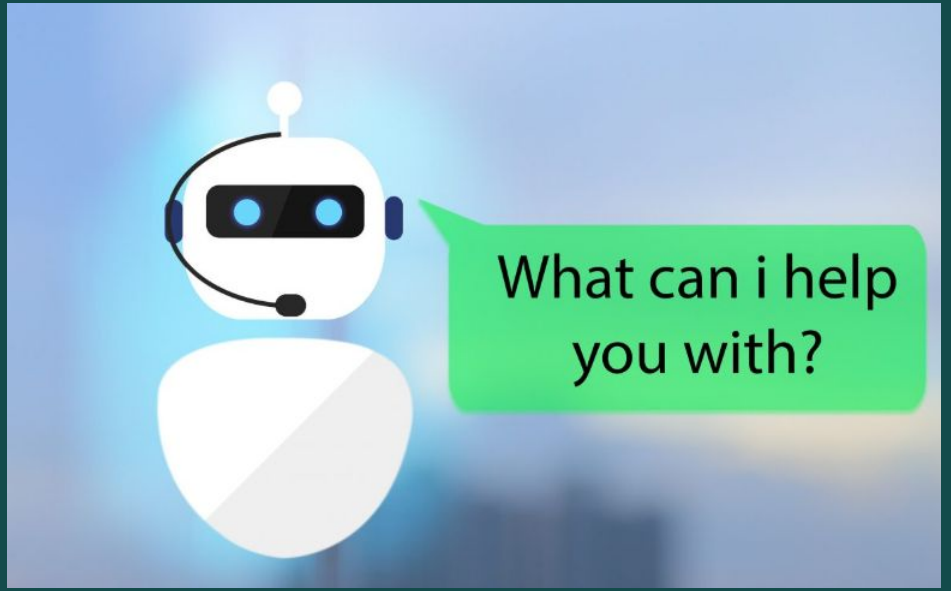
Katz Bot: Intelligent Chatbot

An LLM based Approach



Participants:

- Deepa Paikar
- Kiran Vutukuri
- Haider Ali

Data Collection

- Data Collection:
  - Manual exploration of Yeshiva websites.
  - Collection of course details, teacher profiles, locations, and international rules.
  - Organization of data into text files, categorized systematically.

- Data Pre-Processing:
  - Parsing mechanism development using regular expressions and tailored methods.
  - Standardization by removing symbols, punctuation, and special characters.
  - Tokenization of sentences and structuring of question-answer pairs.

- Quality Enhancement:
  - Removal of duplicates and noise to ensure diversity.
  - Consideration of data augmentation techniques.
  - Manual quality control through dataset sampling

- Data Storage and Organization
  - Use of structured databases, CSV, and JSON formats for ML framework compatibility.

Data Collection

- Dataset for Fine-Tuning
  - Creation of sentence completion and question-answer pair datasets.
  - Generation of 5,600 sentence completion pairs and 5,100 question-answer pairs.
  - Separate test dataset with 600 question-answer pairs for model consistency.

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

Main parameters and Model training techniques

- QLORA parameters:
  - R = 16: Rank for QLORA config – It defines the dimensions of the Low rank matrix.
  - Noise Efficient Fine Tuning (NEFT) = 5%: It adds noise to the input embeddings which has shown increase in the instruction fine tuning of the LLMs.
  - Double Quantization: This allows us to apply to apply double quantization to reduce the size 2 times.
  - 4bit quantization: Reduced model size to 4bit float precision
  - This has been used for Mistral Instruct 7B, Llama 7B, Microsoft Phi 1.5B, and Mistral Lite 7B.

Llama 2

- Model details:
  - Developed by Meta under the meta-llama repository.
  - Variants include 7B, 13B, and 70B models.
  - Auto-regressive, optimized transformer architecture.
  - Utilizes SFT and RLHF for alignment with human values of helpfulness and safety.

- Fine-Tuning Llama 2 on sentence completion
  - Dataset (5,600 rows) formatted and hosted on Hugging Face for the sentence completion model.
  - Trained on A100 GPU
  - Fine-tuned 7B model with complete sentence completion data, resulting in a 9 GB model file.

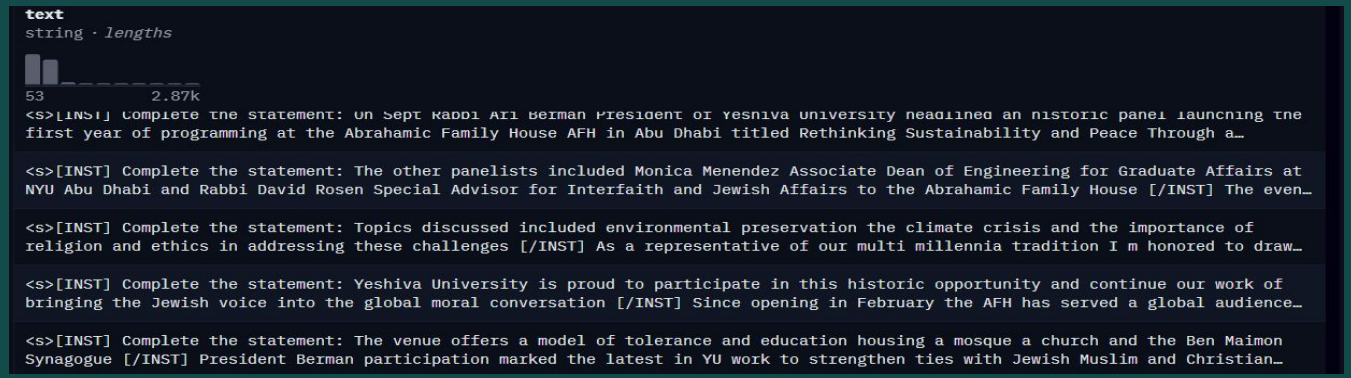
No of epochs	***Time taken***	***Loss start***	***End loss***
500	3 hours	3.5	1.2

Llama 2 Double Fine Tuning

- Double Fine-Tuning Llama 2 &B on QA pairs:
  - Second phase of fine-tuning using the entire question-answer dataset (5100 rows)
  - Each epoch took approximately 40 minutes, with a total of 10 epochs for training.
  - Marked performance improvements observed from 2 epochs.
  - Loss rate decreased from 3.5 to 0.2, indicating significant learning.
  - Fine-tuned 7B model with complete QA data, resulting in a 13 GB model files.

No of epochs	***Time taken***	***Loss start***	***End loss***
10	7 hours	1.2	0.2

Data Format for Llama



Sentence pairs

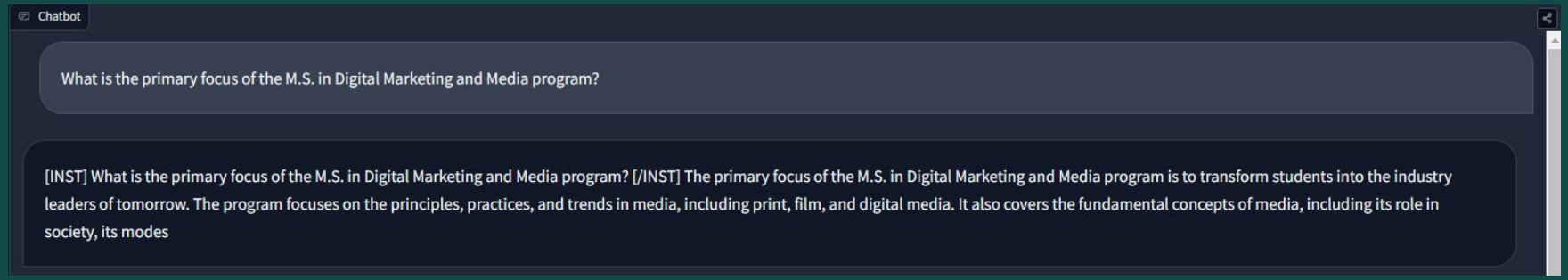
QA pairs

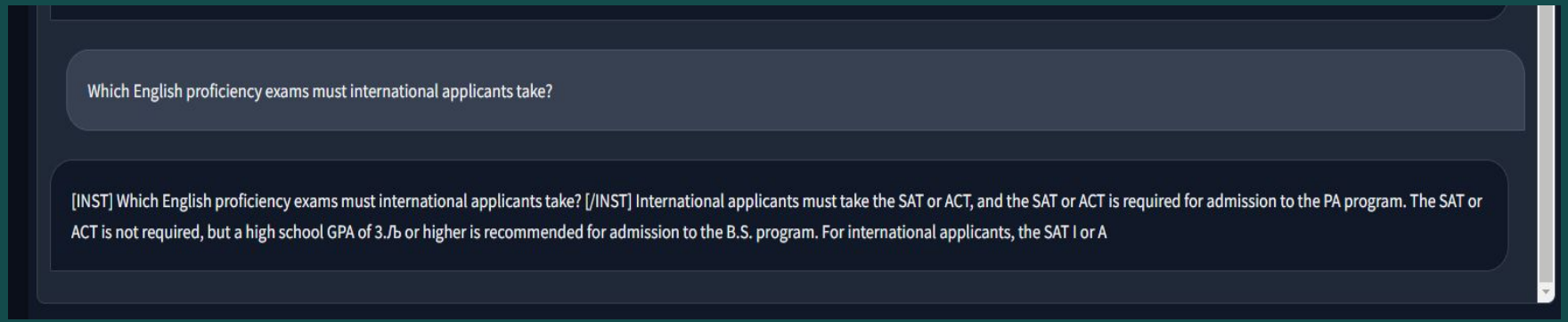
Llama 2 Results

- Metrics

Metric	Precision	Recall	F1-Score
rouge-1	0.25397	0.665229	0.342769
rouge-2	0.152191	0.502428	0.212769
rouge-l	0.247213	0.650588	0.334161

Llama 2 Predictions





Microsoft Phi 1.5B

- Model details:
  - It is successor model of Microsoft Phi 1 and predecessor of Phi 2 model.
  - The model is only trained on **synthetic** books data which keeps it away from internet toxicity.
  - The model contains total 1.3 B parameters and total size of 2.84 GB.
  - This model isn't **instruction trained or RLHF.**
- Training:
  - Input format: <s> [INST] question/instruction [/INST] answer/ground truth [INST]
  - Training hours: 5 hours approx.
  - Epochs: 40
  - Trainable parameters: 14,286,848

Question Answer

Metric	Precision	Recall	F1-Score
rouge-1	0.163489	0.117869	0.130689
rouge-2	0.0277622	0.0189221	0.0216161
rouge-l	0.141821	0.100103	0.111831

Sentence Completion

Metric	Precision	Recall	F1-Score
rouge-1	0.442034	0.395166	0.403886
rouge-2	0.301005	0.240321	0.256521
rouge-l	0.416284	0.371399	0.379973

Amazon Mistral 7B

- Model details:
  - MistralLite is a fine-tuned version of the Mistral-7B-v0.1 language model, specifically designed to excel in handling long contexts, supporting up to 16K tokens.
  - Utilizes an adapted Rotary Embedding and sliding window during fine-tuning, resulting in improved performance on tasks requiring extensive context comprehension.
  - MistralLite introduces adaptations, including an increased RotaryEmbedding adaptation (rope_theta = 1000000) and a larger sliding window size (16384), enabling it to efficiently process longer contexts.
  - In our dataset, our context length is 1050 tokens which is enough for Mistral Instruct
- Training:
  - Input format: <s> [INST] question/instruction [/INST] answer/ground truth [INST]
  - Epochs: 2 epochs
  - Training hours: 5 hours approx.
  - Trainable parameters: 23,646,256

SC pairs

```
Sentence completion results
Test mean F measures
RougeL mean F measure:  0.07479042919391612
Rouge1 mean F measure:  0.09923955487718501
Rouge2 mean F measure:  0.0016528925619834708
************************************************
Train mean F measures
RougeL mean F measure:  0.07235893223976217
Rouge1 mean F measure:  0.09778782898344184
Rouge2 mean F measure:  0.0042109929078014184
```

QA pairs

```
Question Answers results
Test mean F measures
RougeL mean F measure:  0.23427533090582467
Rouge1 mean F measure:  0.2699527805426028
Rouge2 mean F measure:  0.1617031378864939
************************************************
Train mean F measures
RougeL mean F measure:  0.205389810231495
Rouge1 mean F measure:  0.24756714980473765
Rouge2 mean F measure:  0.14096743438208675
```

Mistral Instruct 7B

- Model Details:
  - Built on top of Llama 7B model.
  - Innovative Sliding Window Attention **(SWA)** mechanism enables faster inference and significant memory savings. It also allows model to consider past tokens which are 2 times window size distant from the current token which allows it for the extended attention.
  - Utilizes Grouped-query attention (GQA) for efficient inference.
  - Outperforms Llama2 13B on all benchmarks and Llama1 on many benchmarks.
- Training:
  - Input format: <s> [INST] question/instruction [/INST] answer/ground truth [INST]
  - Epochs: 7 epochs
  - Training hours: 24 hours approx.
  - Trainable parameters: 23,646,208

SC pairs

Metric	Precision	Recall	F1-Score
rouge-1	0.263258	0.0805735	0.117046
rouge-2	0.0445077	0.0112352	0.0169852
rouge-l	0.236762	0.0715334	0.10436

QA pairs

Metric	Precision	Recall	F1-Score
rouge-1	0.335577	0.783925	0.44855
rouge-2	0.242113	0.647393	0.331086
rouge-l	0.329796	0.770991	0.440821

Mistral Instruct sentence completion on test set

```
Sentence:
 <s>[INST] Complete the following text: The Katz School s entry-level Occupational Therapy Doctorate, the first full
y accredited entry-level OT doctoral program in the New York
----------------------------------------------------------------------------------------------------
Ground truth completion:
  metropolitan area, trains caring and competent scholar-practitioners to develop and deliver holistic services thro
ugh the integration of theory, research and knowledge-based practice and clinical skills. </s>
----------------------------------------------------------------------------------------------------
Generated answer:  The Katz School s entry-level Occupational Therapy Doctorate, the first fully accredited entry-le
vel OT doctoral program in the New York metropolitan area, offers a unique blend of traditional and emerging researc
h, creativity, and knowledge-based practice. It is designed to meet the needs of the diverse and dynamic occupationa
l therapy community. The program is rooted in a strong foundation of science and the arts, providing students with a
comprehensive understanding of occupational therapy. The
CPU times: user 12.1 s, sys: 278 ms, total: 12.4 s
Wall time: 12.4 s
{'rouge1': Score(precision=0.18421052631578946, recall=0.5, fmeasure=0.2692307692307692),
 'rouge2': Score(precision=0.05333333333333334, recall=0.14814814814814814, fmeasure=0.0784313725490196),
 'rougeL': Score(precision=0.11842105263157894, recall=0.32142857142857145, fmeasure=0.17307692307692307)}
```

Details about MS AI program from Mistral Instruct (Temp)

Temp 0.1

```
 Yeshiva University's MS AI program is a 30-credit Master of Science in Artificial Intelligence, offered on campus i
n New York City, with full-time or part-time options and evening courses to accommodate working students. The progra
m focuses on designing and building cutting-edge AI technologies for various applications, including finance, biotec
h, 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 considerat
ions. It is STEM-OPT eligible, and international students may be eligible for up to 36 months of Optional Practical
Training (OPT).
```

Temp 0.5

```
 Yeshiva University's MS AI program is an interdisciplinary program designed for students to master artificial intel
ligence 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 internationa
l students may be eligible for up to 36 months of Optional Practical Training (OPT). The program is also aligned wit
h industry certifications, such as the Certified Artificial Intelligence Professional (CAIP). The MS AI program offe
rs full-time, part-time, or evening options, and international students may be eligible for up to 36 months of Curri
cular Practical Training (CPT). The program is also STEM-OPT eligible. Students may enter the program in January, Ma
y, or September. Application deadlines are rolling, and there are multiple entry points throughout the year. Tuition
for the program is $25,000
```

Temp 0.9

```
' 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. T
he 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.'
```

Mistral Instruct question answering on test set

What did Dr. Youshan Zhang do as a postdoctoral associate at Cornell University?

```
Ground truth answer:
  Youshan Zhang applied deep learning models in veterinary science to detect the diseases of animals and to analyze different animal behaviors. </s>
Generated answer:  Youshan Zhang applied deep learning models in veterinary science to detect the diseases of animals and to analyze different animal behaviors. He developed a
system that can recognize the sounds of different bird species and studied the effects of environmental factors on the behavior of birds. He also worked on a project to estima
te the age of wild animals using their vocalizations. As a postdoctoral associate, Zhang contributed significantly to the field of animal behavior and conservation.
CPU times: user 11.4 s, sys: 294 ms, total: 11.6 s
Wall time: 11.6 s

{'rouge1': Score(precision=0.2727272727272727, recall=0.9545454545454546, fmeasure=0.4242424242424242),
 'rouge2': Score(precision=0.2631578947368421, recall=0.9523809523809523, fmeasure=0.4123711340206186),
 'rougeL': Score(precision=0.2727272727272727, recall=0.9545454545454546, fmeasure=0.4242424242424242)}
```

What are some specific AI and machine learning models that students will learn in the program?

```
Ground truth answer:
  Students will learn supervised and unsupervised learning, deep learning and neural networks, and reinforcement learning. </s>
Generated answer:  Students will learn supervised and unsupervised learning models, including regression, classification, clustering, and dimensionality reduction. They will a
lso explore deep learning models such as neural networks, convolutional neural networks, and reinforcement learning. Additionally, students will gain hands-on experience with
tools like Python, R, SQL/NoSQL, TensorFlow, Keras, and pyTorch.
CPU times: user 10.3 s, sys: 260 ms, total: 10.6 s
Wall time: 10.6 s

{'rouge1': Score(precision=0.3, recall=0.9375, fmeasure=0.4545454545454453),
 'rouge2': Score(precision=0.22448979591836735, recall=0.7333333333333333, fmeasure=0.34375),
 'rougeL': Score(precision=0.28, recall=0.875, fmeasure=0.4242424242424425)}
```

Independent study in MS AI program (Prompt Engineering)

Temp 0.45

Can you explain the Independent study course in MS AI program?

```
Generated answer:  Yes, the Independent study course in MS AI program allows students to work on a project in collab
oration with a faculty member, developing skills in research, critical thinking, and professional presentation. The
course provides an opportunity for students to explore a specific area of interest within AI, tailoring their projec
t to gain hands-on experience and contribute to the field. The project can involve developing an AI application, con
ducting a feasibility study, or engaging in other activities related to AI. The course is taken over 12 weeks, and s
tudents receive guidance from a faculty member throughout the process.
CPU times: user 14.3 s, sys: 281 ms, total: 14.5 s
Wall time: 14.5 s
```

Can you explain the Independent study course in MS AI program with it's requirements and prerequisites?

```
Generated answer:  The Independent study course in MS AI program requires 6 credits and is taken in either the fall
or spring semester. It has prerequisites of 2000 or 3000 level AI, Data Analytics, or Mathematics courses, and one 2
000 or 3000 level programming course. Students must obtain permission from the program director and the chair of the
department. The course involves working with a faculty member on a project, which may include research, design, impl
ementation, and/or experimentation. The project report is the final deliverable. The Independent study course is ava
ilable only for students enrolled in the full-time MS AI program.
CPU times: user 17.4 s, sys: 287 ms, total: 17.6 s
Wall time: 17.6 s
```

GPT 2

- The GPT-2 small language model , a lightweight variant within the GPT family, features 117 million parameters, catering to tasks that demand efficiency.
- Initially, GPT-2 was trained for 75 epochs on sentence completion pairs, establishing a foundational understanding of language contexts.
- Subsequently, fine-tuning continued with an additional 60 epochs on question-answer pairs.
- Training:
  - Training hours: 9 hours approx.
  - Epochs: 135

GPT 2 Metrics

- Sentence pairs

Metric	Precision	Recall	F1-Score
rouge-1	0.123113	0.130636	0.113528
rouge-2	0.0187608	0.0202557	0.0174924
rouge-l	0.109326	0.11581	0.100334

- QA pairs

Metric	Precision	Recall	F1-Score
rouge-1	0.681369	0.669716	0.667471
rouge-2	0.551696	0.549399	0.546123
rouge-l	0.672503	0.661576	0.659397

Generated test data responses

```
UserInput: At the Ferkauf Graduate School of Psychology, what is the primary area of research for Dr. Sarah Bloch-Elkouby?

Reference:  Dr. Sarah Bloch-Elkouby's primary area of research is enhancing suicide prevention and effective treatments.

Predict: Dr. Sarah Chutick's research primarily focuses on suicide prevention and effective treatments.
```

```
UserInput: What academic qualifications does Dr. Peter Franz hold?

Reference:  B.S. at Cornell University, M.A., and Ph.D. from Harvard University.

Predict: B.S., M.A., and Ph.D.
```

Metrics Comparison of tested LLMs

- GPT-2 stands out in both metrics, possibly due to its pre-training regimen.
- Microsoft Phi and Llama 7B show that larger models may not always lead to proportionally higher scores.

Metric	Microsoft Phi	GPT 2	Llama 7B	Mistral Instruct
rouge-1	0.403886	0.667471	0.342769	0.44855
rouge-2	0.256521	0.549339	0.212769	0.331086
rouge-l	0.379973	0.659397	0.334161	0.440821

Predictions Comparison of tested LLMs

- Question: what is a CRN?
- Ground truth: A CRN is a course registration number, which is unique to each course.

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…….

Discussion

- ***GPT-2 Performance***
  - Shows the highest Rouge-1, Rouge-2 and Rough-l scores, indicating strong unigram and bigram overlap.

- ***Microsoft Phi 1.5B***
  - The Rouge scores gets better per epoch, but the loss stabilizes at 0.5. After that model starts to overfitting.

- ***Llama 7B***
  - Has the lowest Rouge-1 and Rouge-2 scores, suggesting room for improvement in capturing linguistic details.Training it for more epochs might have some improvements or using 13B Llama might give better results but we need a strong GPU.

- ***Mistral Instruct 7B***
  - Displays lower Rouge scores compared to GPT-2 due to less epochs but is outperforming Llama 7B and Phi 1.5B. Can be expected to perform better with better instructions during fine tuning and inference stages.

Future Developments

- *Finetune the best working model once again and make it publicly available*
- *Optimizing hyperparameters for our models*
- *Comparison with document retrieval*
- *Try other newly available Question answer models for comparison*
- *Create an app interface/website interface for the chatbot and try to integrate with other platforms like Yeshiva website/telegram*

Thank you for your time