Finetuning Llama on custom data for QA tasks

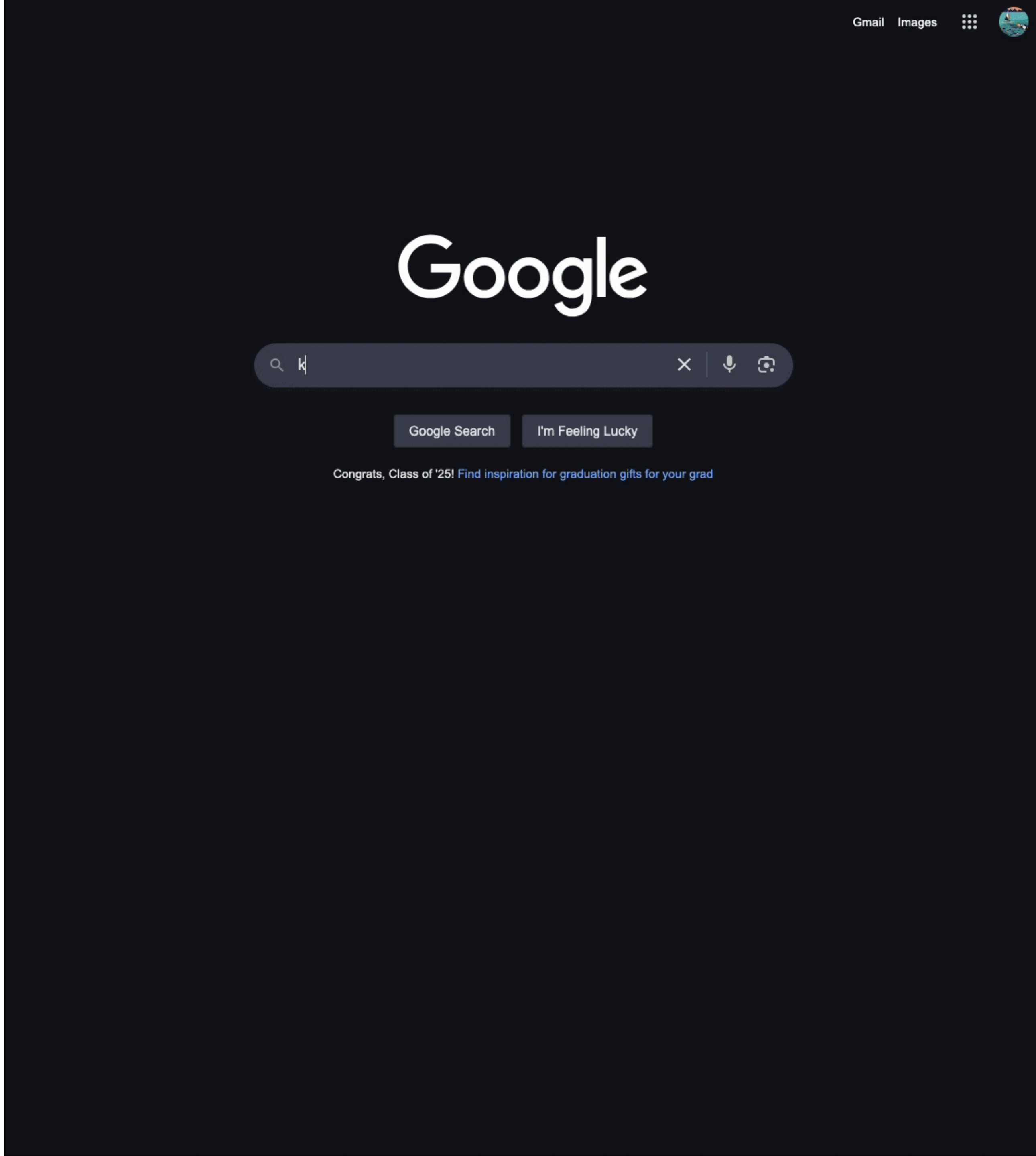
By Rahul Purswani

Department of EECS

University of Kansas

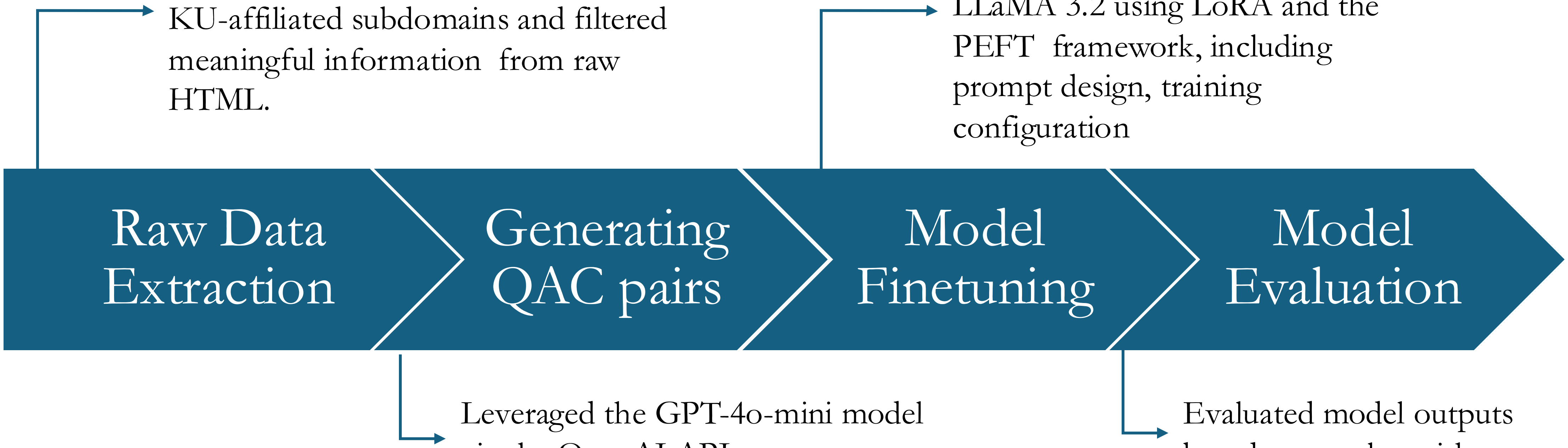
Motivation

- Students often need to click through multiple university webpages to find basic information like deadlines, forms, or policies.
- Large language models can help, but struggle with accuracy without being tailored to the domain.
- Fine-tuning an LLM on university-specific content enables faster, more reliable answers with less user friction.



Congrats, Class of '25! Find inspiration for graduation gifts for your grad

Project Overview



Raw Data Extraction

Scraped content from approximately ten KU-affiliated subdomains and filtered meaningful information from raw HTML.

Generating QAC pairs

Leveraged the GPT-4o-mini model via the OpenAI API to automatically generate question, answer, and context triplets from the extracted content.

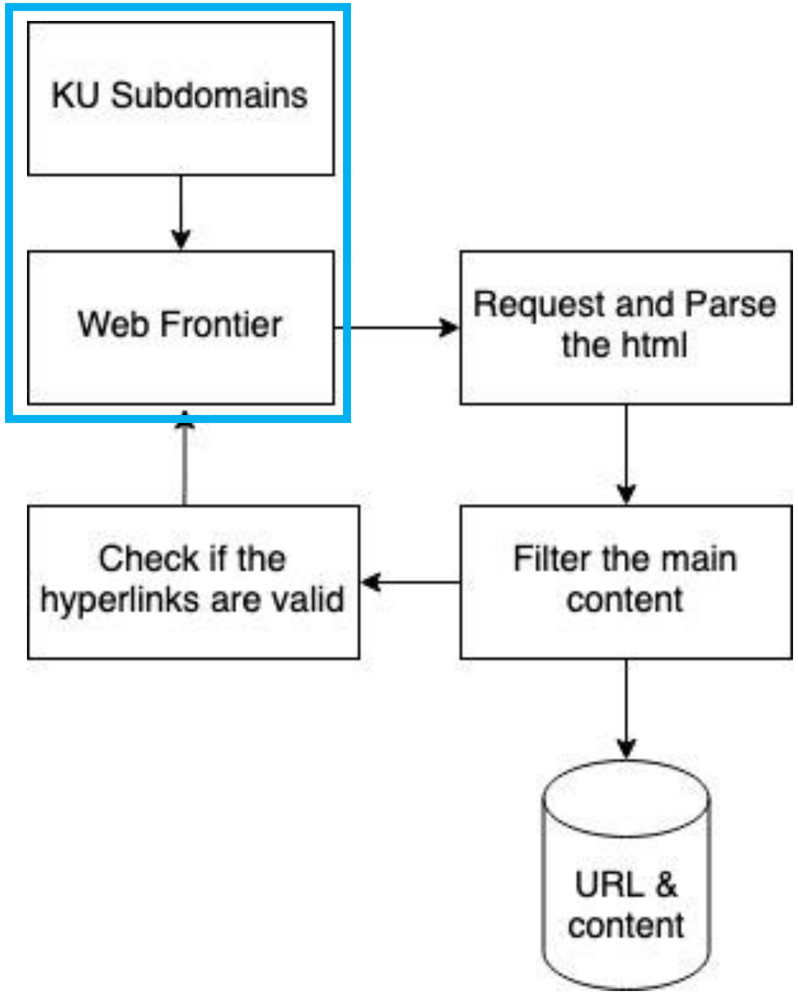
Model Finetuning

Training workflow of fine-tuning LLaMA 3.2 using LoRA and the PEFT framework, including prompt design, training configuration

Model Evaluation

Evaluated model outputs based on overlap with reference answers, fluency of phrasing, and semantic similarity.

Raw Data Extraction



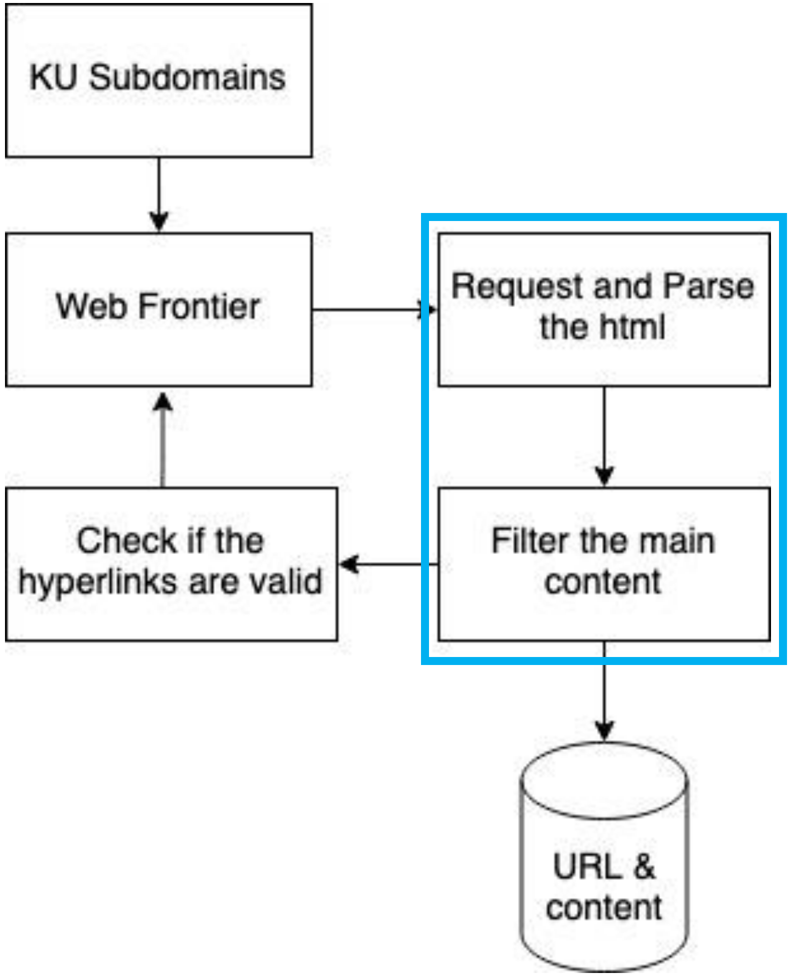
KU Subdomains and Web Frontier

- Selected some KU-affiliated subdomains containing information on engineering programs, international student services, research, etc.
- Used a curated list as the seed input for crawling.

Web Frontier

- Basically, a double ended queue that stores URLs to be visited
- Ensures breadth first crawl.

Raw Data Extraction

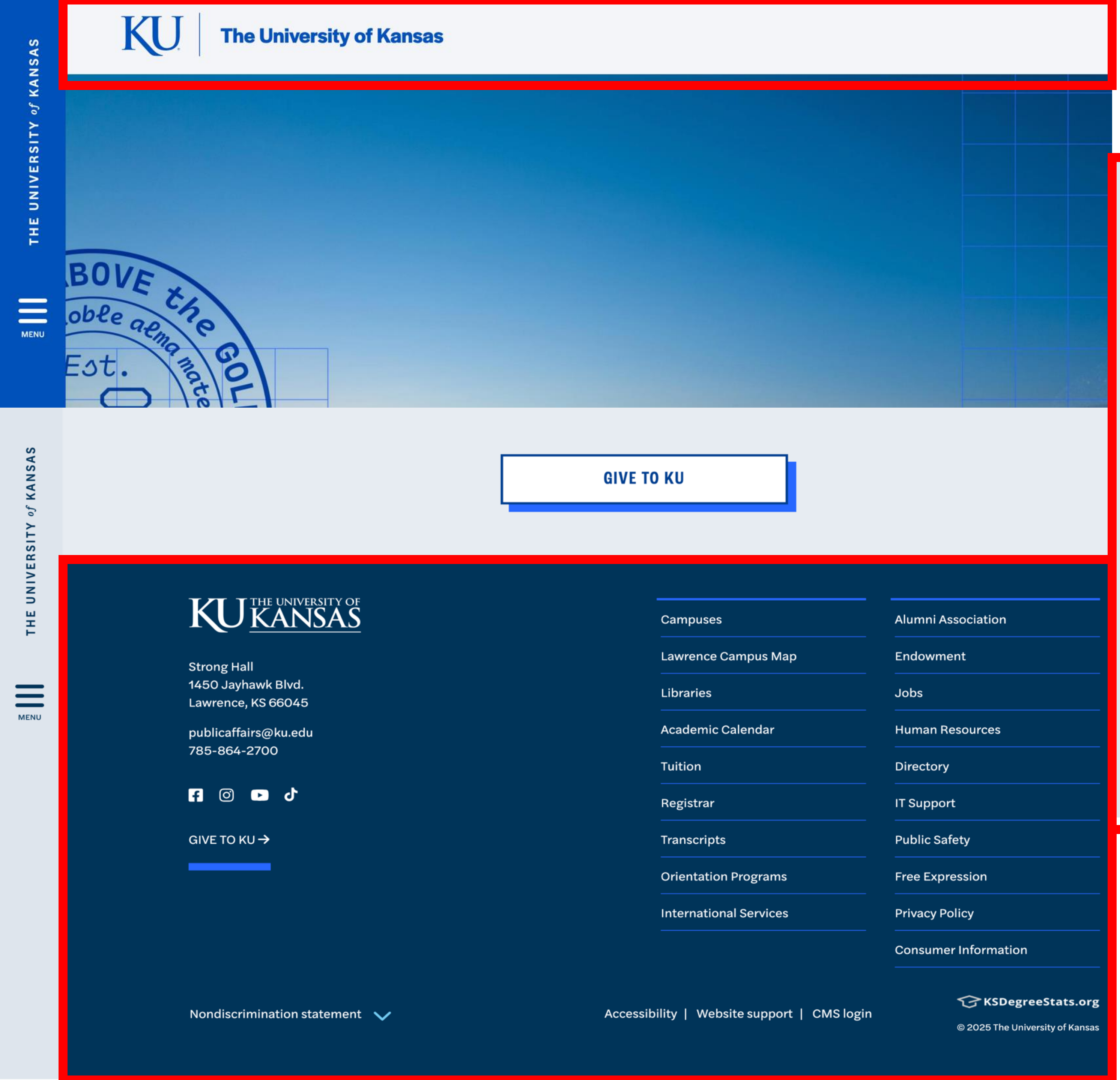


Request and Parse the HTML

- Send HTTP request and parse the HTML using BeautifulSoup to extract all the readable data.

Filter the main content

- Extract content only from the "main" section of the page.
- Clean the data – removing HTML tags, special characters, whitespace, etc.
- Count the num of tokens and store the cleaned data as URL, content, and num_tokens.



THE UNIVERSITY of KANSAS

MENU

KU | The University of Kansas

GIVE TO KU

KU THE UNIVERSITY OF KANSAS

Strong Hall
1450 Jayhawk Blvd.
Lawrence, KS 66045

publicaffairs@ku.edu
785-864-2700

GIVE TO KU →

- Campuses
- Lawrence Campus Map
- Libraries
- Academic Calendar
- Tuition
- Registrar
- Transcripts
- Orientation Programs
- International Services

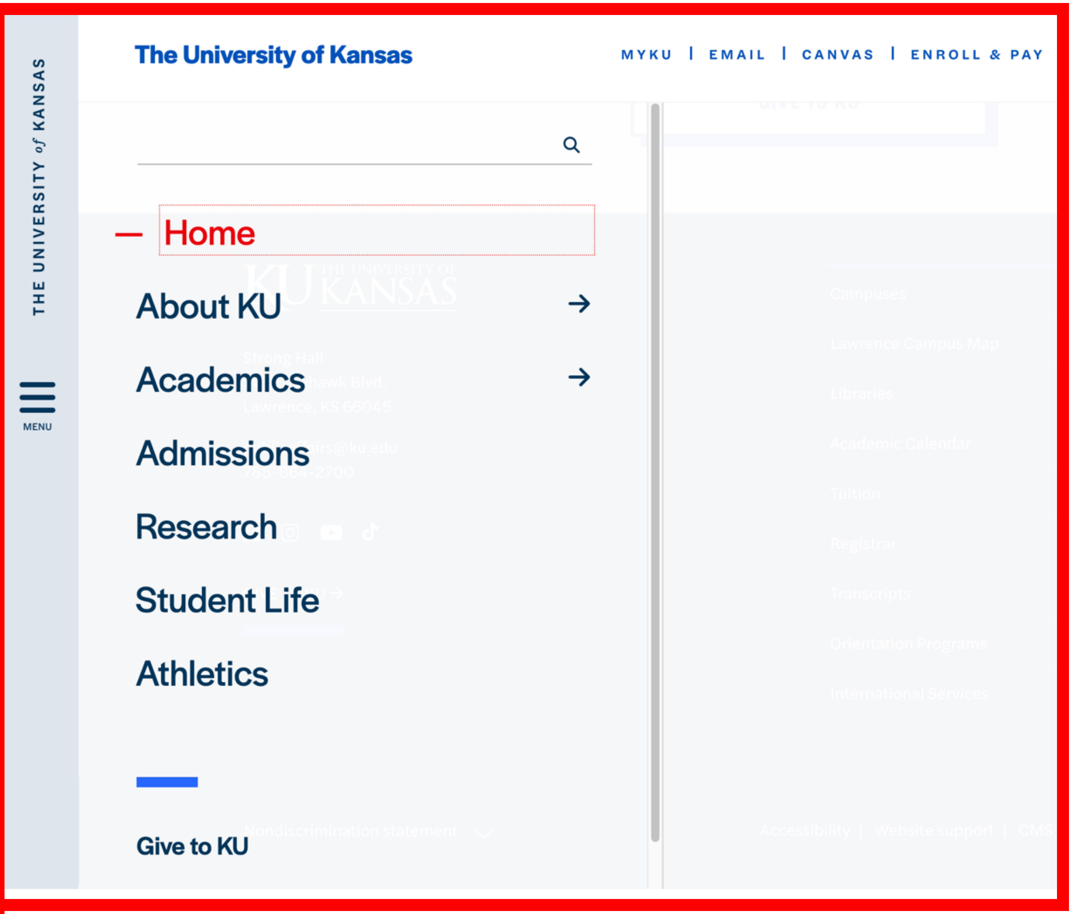
- Alumni Association
- Endowment
- Jobs
- Human Resources
- Directory
- IT Support
- Public Safety
- Free Expression
- Privacy Policy
- Consumer Information

Nondiscrimination statement

Accessibility | Website support | CMS login

KSDegreeStats.org

© 2025 The University of Kansas



THE UNIVERSITY of KANSAS

MENU

The University of Kansas

MYKU | EMAIL | CANVAS | ENROLL & PAY

— Home

About KU →

Academics →

Admissions

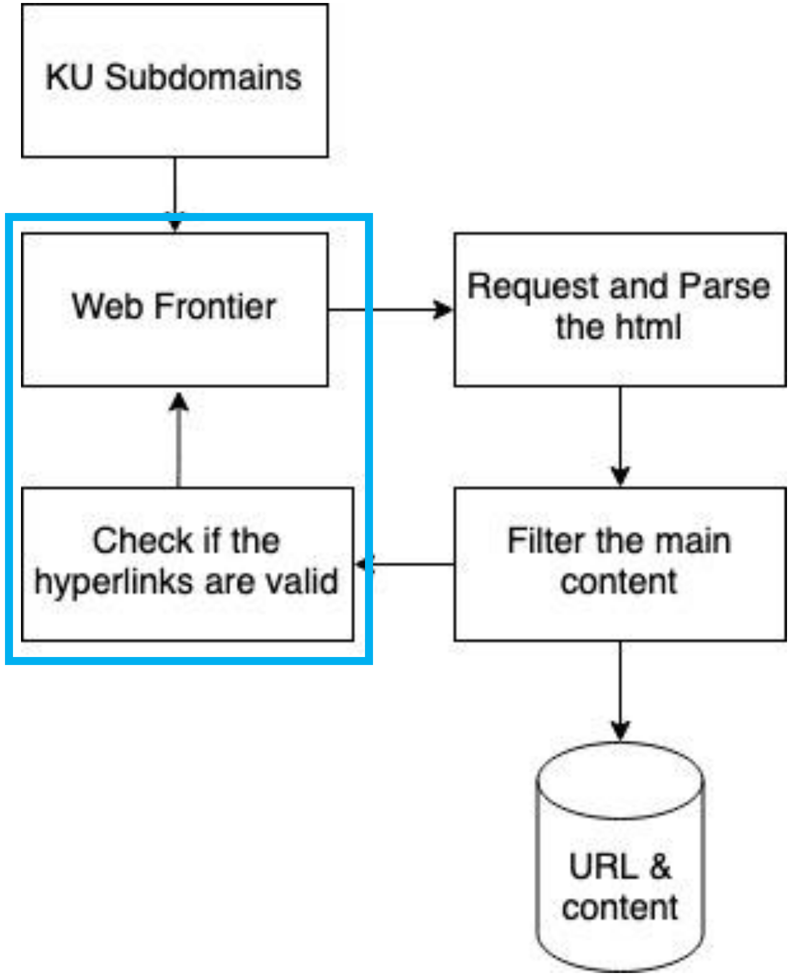
Research

Student Life

Athletics

Give to KU

Raw Data Extraction



Check for hyperlinks

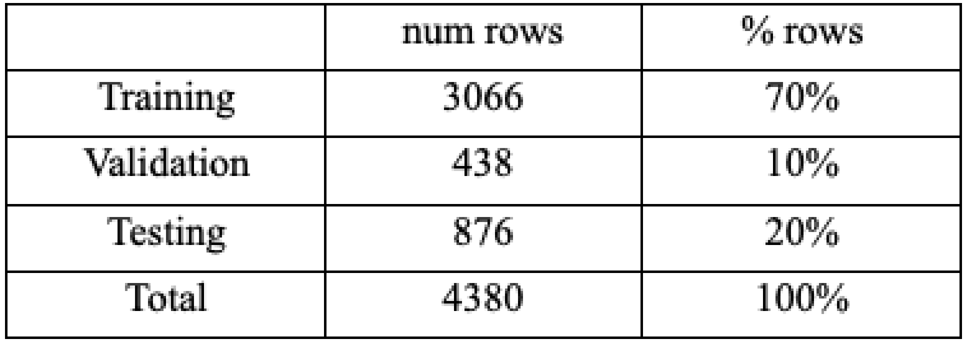
- Extract all the hyperlinks from the HTML.
- Check if the links are valid – KU subdomain, not previously visited, not a redirect (login, returnto, etc.)
- Append links to the web frontier.

Raw Data Example

```
{
    "url": "https://iss.ku.edu/f-1-cap-gap",
    "text": "Cap Gap Extension The cap gap extension is a period in which an eligible F-1
student\u2019s status is ...",
    "num_tokens": 886},
{
    "url": "https://iss.ku.edu/f-1-economic-hardship",
    "text": "Employment Based on Severe Economic Hardship F-1 students may be eligible if
they have demonsrated severe economic hardship (SEH), to apply ...",
    "num_tokens": 887},
{
    "url": "https://app.eecs.ku.edu/mill/emailform.php",
    "text": "This Page has moved! This page has moved! Please update your links ...
    "num_tokens": 62},
```

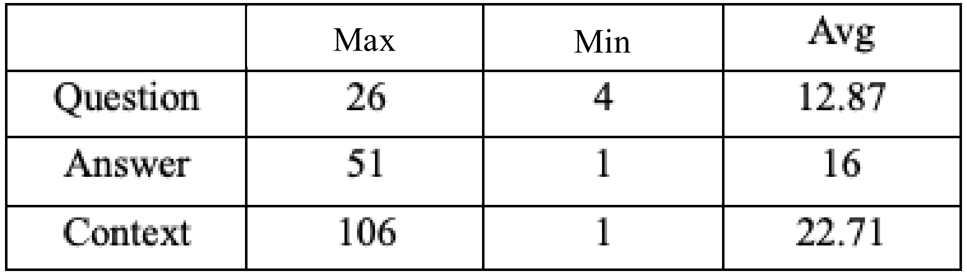
Generating QAC pairs and Final Dataset

- Converted raw data into pairs of Question, Answer, and Context.
- Used GPT-4o-mini via OpenAI API to automatically generate question–answer–context (QAC) triples.
- Initial prompt versions led to issues:
  - Duplicate or irrelevant questions
  - Invalid JSON formatting
  - Answers not grounded in context
- Refined the prompt iteratively to improve quality and formatting.
- Generated ~4,380 QAC pairs from ~5,600 pages; excluded low-value pages (e.g., people.ku.edu).
- Final dataset contains good-quality, source-linked QAC pairs, essential for effective fine-tuning.



	num rows	% rows
Training	3066	70%
Validation	438	10%
Testing	876	20%
Total	4380	100%

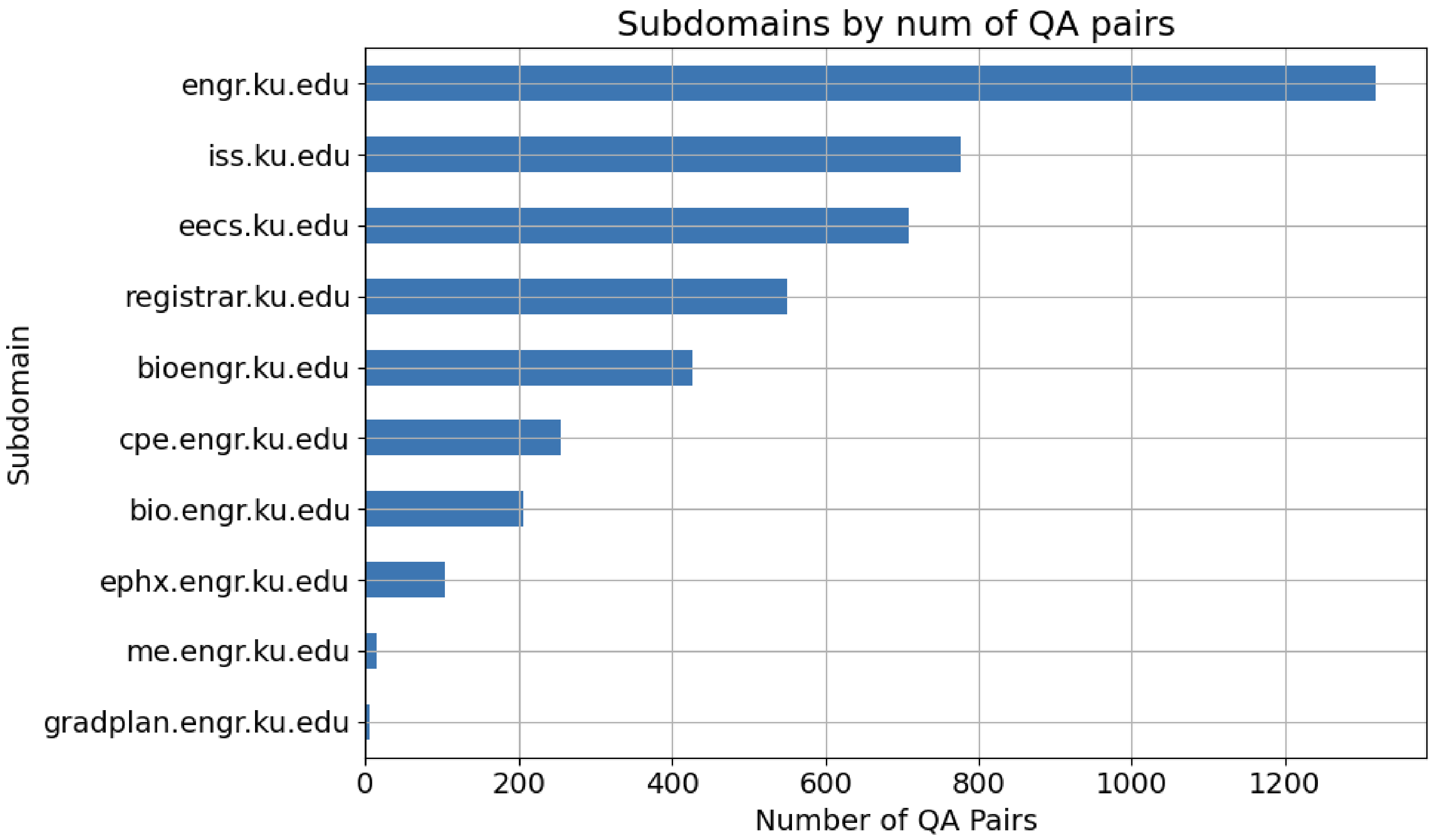
Table 1: Train-test split of the dataset

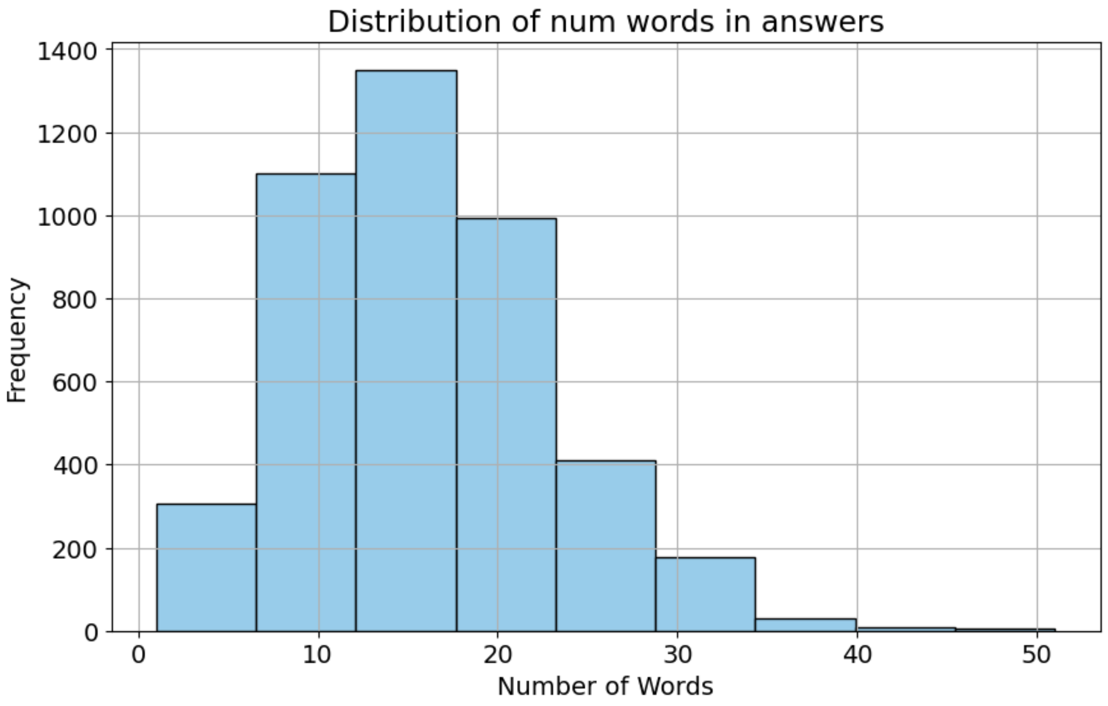


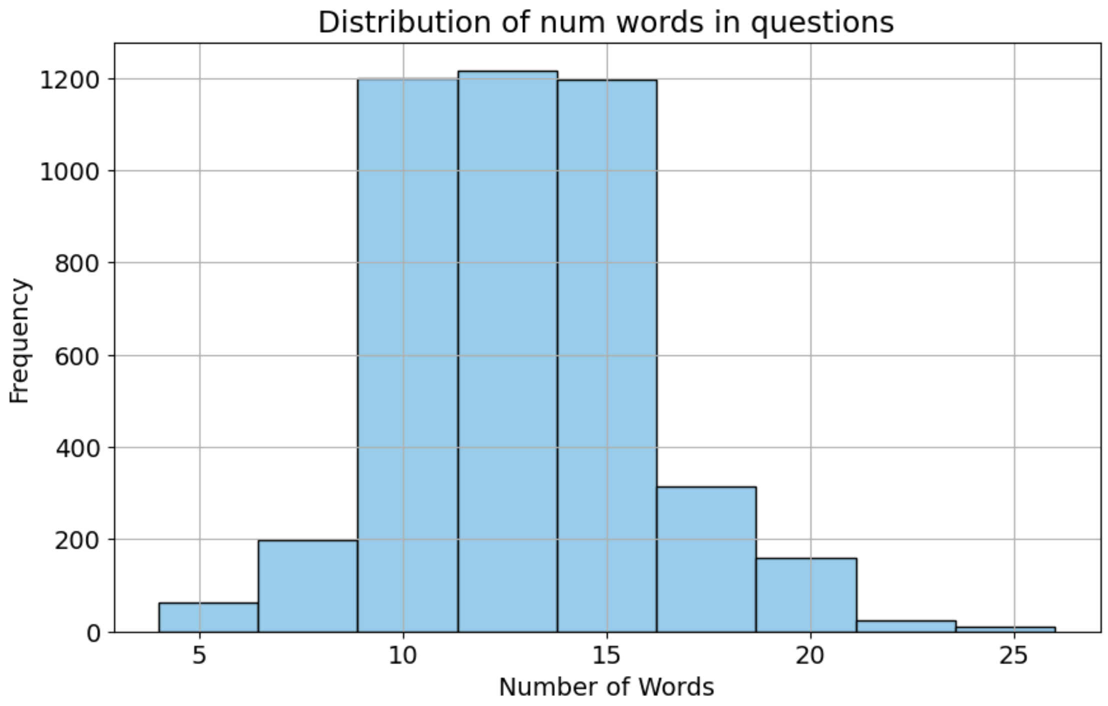
	Max	Min	Avg
Question	26	4	12.87
Answer	51	1	16
Context	106	1	22.71

Table 2: Number of words in QAC pairs

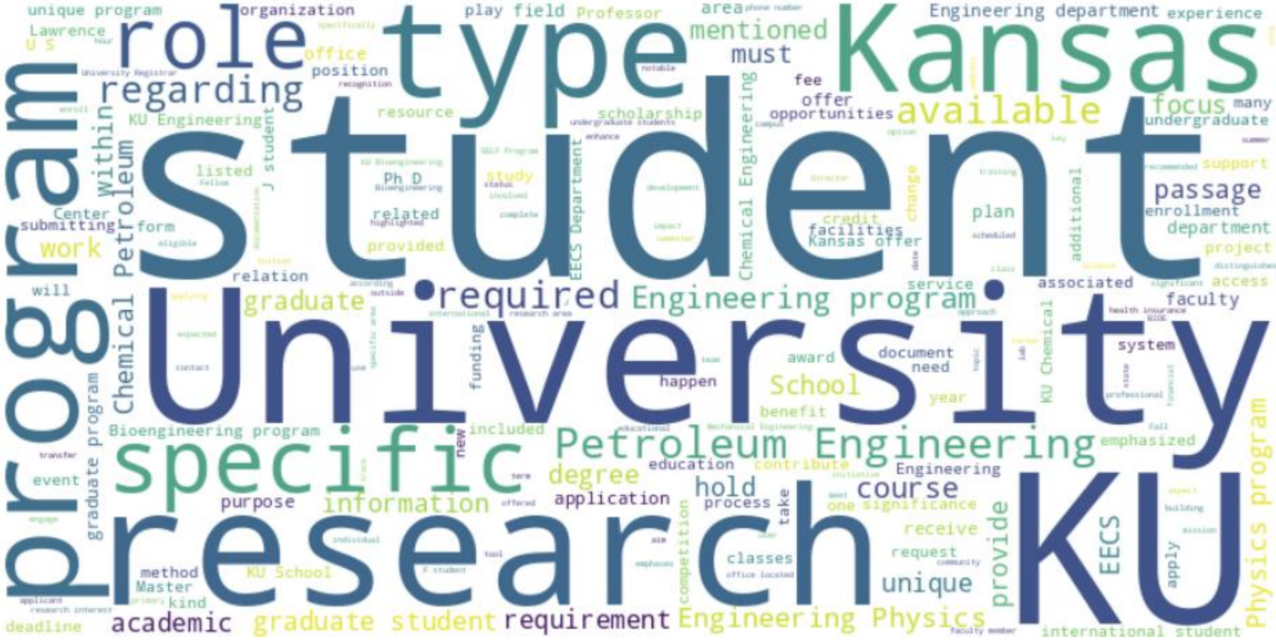
Understanding the Dataset



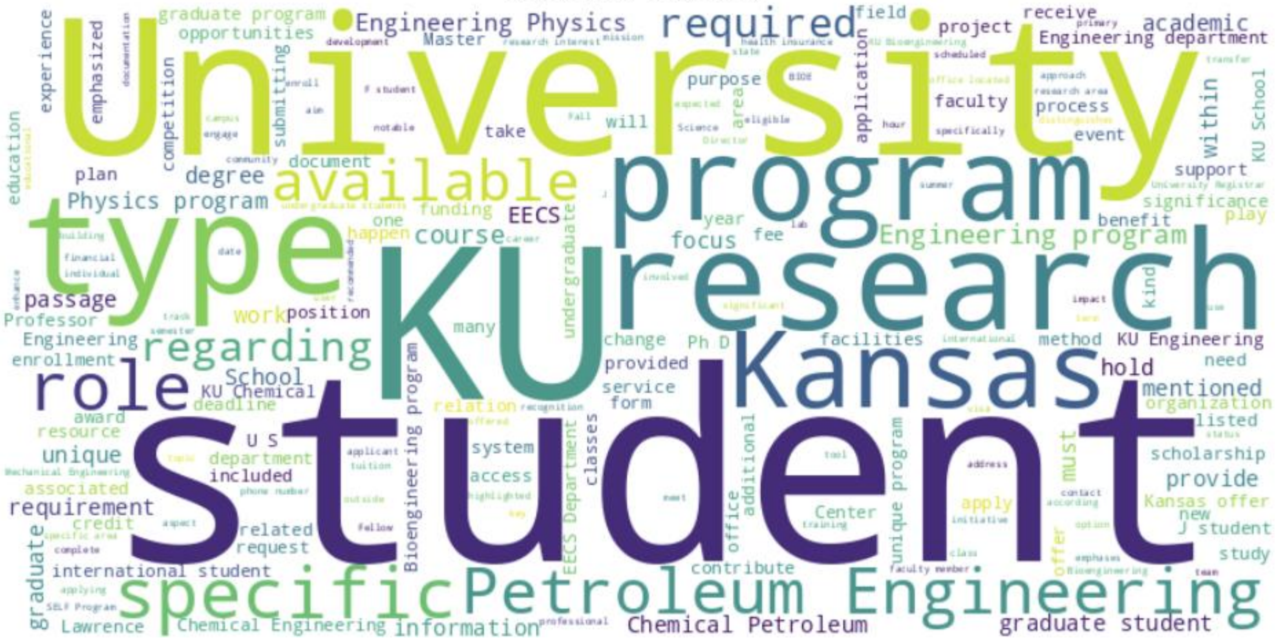

Distribution of num words in answers
Frequency
1400
1200
1000
800
600
400
200
0
0
10
20
30
40
50
Number of Words


Distribution of num words in questions
Frequency
1200
1000
800
600
400
200
0
5
10
15
20
25
Number of Words

Word Cloud of Questions



Word Cloud of Answers



Some good examples of QAC pairs

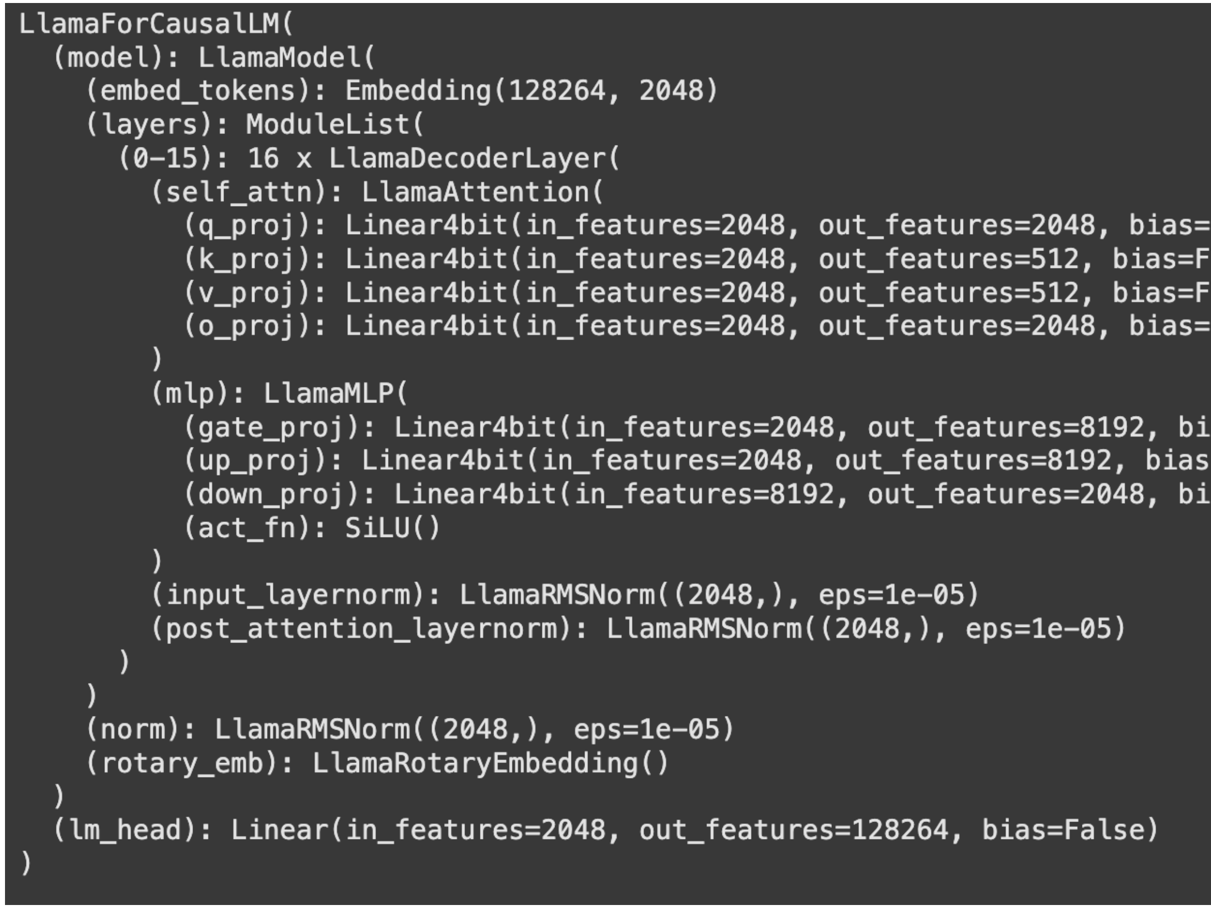
```
 {

    "question": "What specific requests can be reviewed during a Drop-In Advising session?",

    "answer": "During Drop-In Advising, the following iHawk requests can be reviewed and
processed if they are completely submitted: Reduced Course Load (F-1 or J-1), Add Dependents (F
or J), Financial Update on I-20, among others.",

    "context": "The following iHawk requests can be reviewed and processed during drop-ins/walk-
ins...",

    "source": "https://iss.ku.edu/advising"},

 {

    "question": "What does 'D/S' on your I-94 signify?",

    "answer": "'D/S' stands for 'Duration of Status', meaning you are admitted to the U.S. for as
long as you are maintaining your status.",

    "context": "You should have the letters 'D/S' on your I-94, which means that you are admitted
to the U.S. for the 'Duration of Status', or for as long as your are maintaining status - this is
typically the time that it takes for you to complete your degree/program.",

    "source": "https://iss.ku.edu/i94-newly-admitted"

  }
```

Some bad examples of QAC pairs

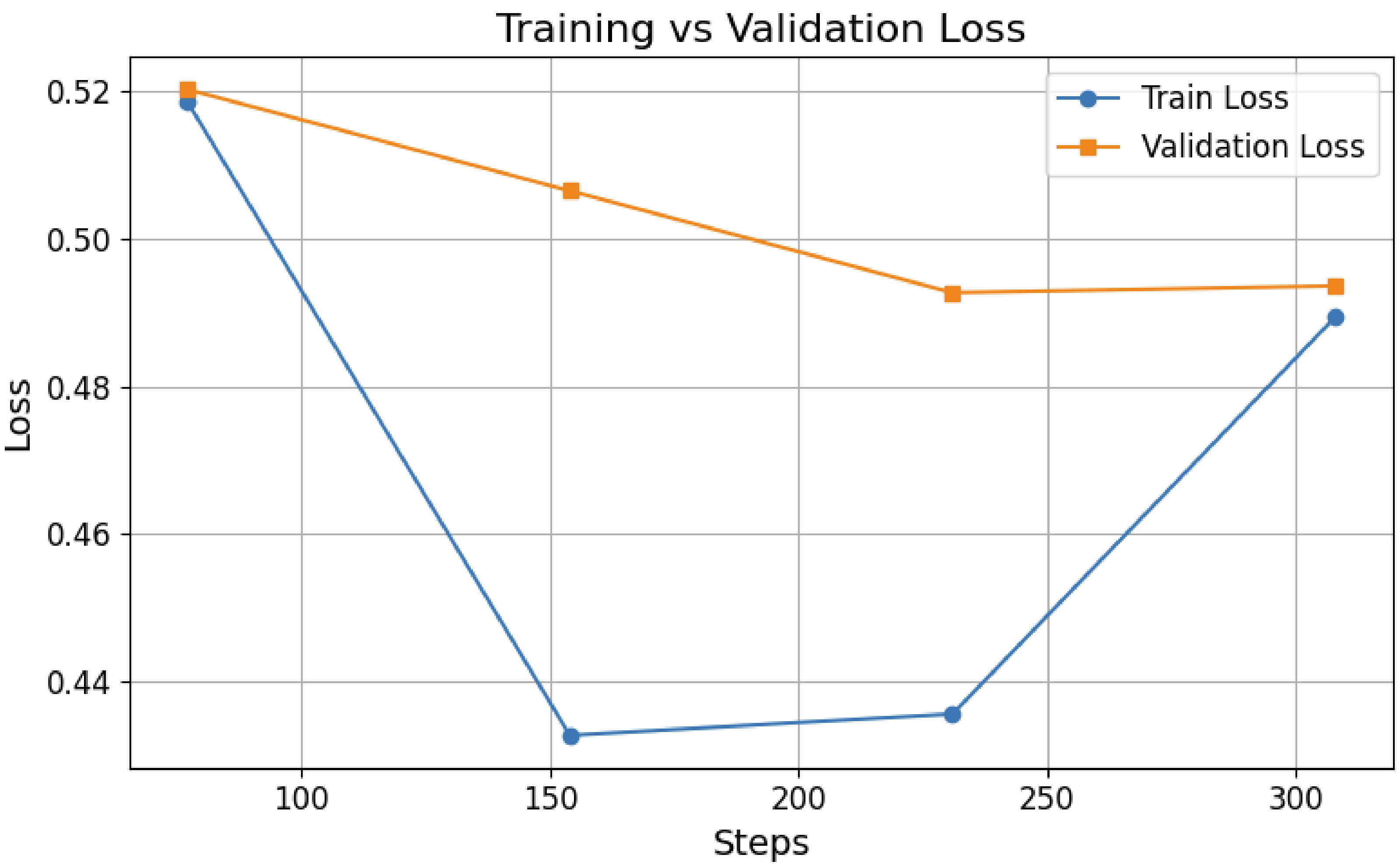
```
 {
    "question": "How many new international students arrived at KU this fall?",
    "answer": "Over 400 new international students arrived this fall.",
    "context": "400+ new international students arrived this fall",
    "source": "https://iss.ku.edu"},
 {
    "question": "What is required to change, release, or withhold certain information aside from contact
details?",
    "answer": "A form must be submitted to the office to change, release, or withhold other information.",
    "context": "...the change, release, or withholding of other information requires a form be submitted to our
office.",
    "source": "https://registrar.ku.edu/student-records"},
 {
    "question": "What type of bank account should your ATM card be tied to?",
    "answer": "ATM cards should be tied to a checking account, not a savings account.",
    "context": "ATM cards should be tied to a checking (not savings) account.",
    "source": "https://iss.ku.edu/finances-newly-admitted"},
```

Training Workflow

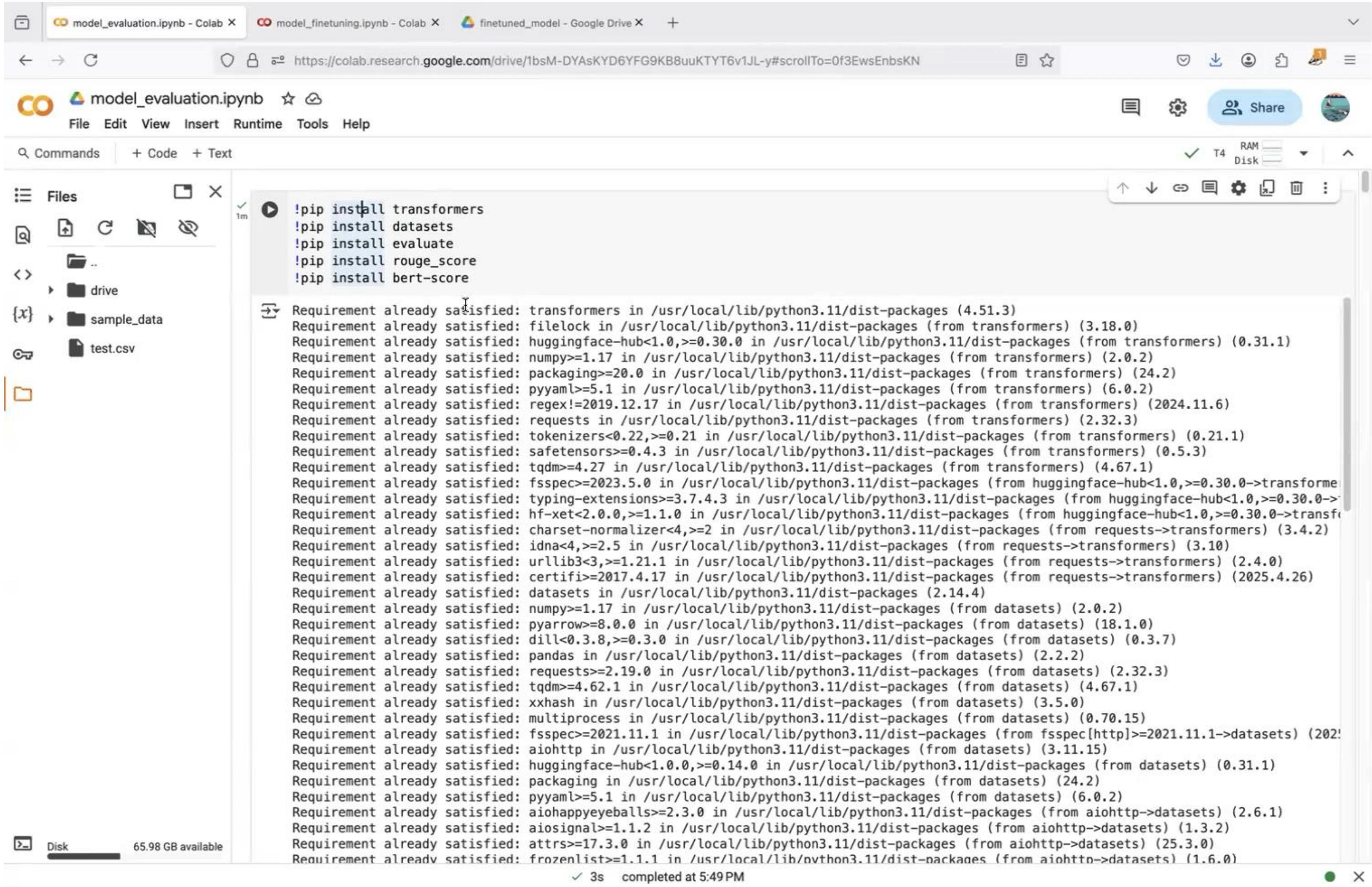
- Model – meta-llama/Llama-3.2-1B-Instruct from Hugging Face
- Used Parameter-Efficient Fine-Tuning (PEFT) with Low-Rank Adaptation (LoRA) to update only 1.79% (~22.5M) of model parameters.
- Targeted layers for LoRA:
  - Self-attention: q_proj, k_proj, v_proj, o_proj
  - MLP: gate_proj, up_proj, down_proj
- Applied 4-bit quantization (nf4 precision, bfloat16 compute) via BitsAndBytes for memory efficiency—trained on a single T4 GPU.
- Used Hugging Face's AutoModelForCausalLM with resized token embeddings and automatic device mapping.
- Used the SFTTrainer from trl library to train the model.
- Loss function: Cross-Entropy, applied only to the model's answer using DataCollatorForCompletionOnlyLM



```
LlamaForCausalLM(
  (model): LlamaModel(
    (embed_tokens): Embedding(128264, 2048)
    (layers): ModuleList(
      (0-15): 16 x LlamaDecoderLayer(
        (self_attn): LlamaAttention(
          (q_proj): Linear4bit(in_features=2048, out_features=2048, bias=
          (k_proj): Linear4bit(in_features=2048, out_features=512, bias=F
          (v_proj): Linear4bit(in_features=2048, out_features=512, bias=F
          (o_proj): Linear4bit(in_features=2048, out_features=2048, bias=
        )
        (mlp): LlamaMLP(
          (gate_proj): Linear4bit(in_features=2048, out_features=8192, bi
          (up_proj): Linear4bit(in_features=2048, out_features=8192, bias
          (down_proj): Linear4bit(in_features=8192, out_features=2048, bi
          (act_fn): SiLU()
        )
        (input_layernorm): LlamaRMSNorm((2048,), eps=1e-05)
        (post_attention_layernorm): LlamaRMSNorm((2048,), eps=1e-05)
      )
    )
    (norm): LlamaRMSNorm((2048,), eps=1e-05)
    (rotary_emb): LlamaRotaryEmbedding()
  )
  (lm_head): Linear(in_features=2048, out_features=128264, bias=False)
)
```


Training vs Validation Loss
Train Loss
Validation Loss
Loss
Steps
0.52
0.50
0.48
0.46
0.44
100
150
200
250
300

Inference Demo


model_evaluation.ipynb - Colab
model_finetuning.ipynb - Colab
finetuned_model - Google Drive
https://colab.research.google.com/drive/1bsM-DYAsKYD6YFG9KB8uuKTYT6v1JL-y#scrollTo=0f3EwsEnbsKN
model_evaluation.ipynb
File Edit View Insert Runtime Tools Help
Share
Commands
+ Code
+ Text
T4
RAM
Disk
Files
..
drive
sample_data
test.csv
Disk
65.98 GB available
1m

```
!pip install transformers
!pip install datasets
!pip install evaluate
!pip install rouge_score
!pip install bert-score
```

```
Requirement already satisfied: transformers in /usr/local/lib/python3.11/dist-packages (4.51.3)
Requirement already satisfied: filelock in /usr/local/lib/python3.11/dist-packages (from transformers) (3.18.0)
Requirement already satisfied: huggingface-hub<1.0,>=0.30.0 in /usr/local/lib/python3.11/dist-packages (from transformers) (0.31.1)
Requirement already satisfied: numpy>=1.17 in /usr/local/lib/python3.11/dist-packages (from transformers) (2.0.2)
Requirement already satisfied: packaging>=20.0 in /usr/local/lib/python3.11/dist-packages (from transformers) (24.2)
Requirement already satisfied: pyyaml>=5.1 in /usr/local/lib/python3.11/dist-packages (from transformers) (6.0.2)
Requirement already satisfied: regex!=2019.12.17 in /usr/local/lib/python3.11/dist-packages (from transformers) (2024.11.6)
Requirement already satisfied: requests in /usr/local/lib/python3.11/dist-packages (from transformers) (2.32.3)
Requirement already satisfied: tokenizers<0.22,>=0.21 in /usr/local/lib/python3.11/dist-packages (from transformers) (0.21.1)
Requirement already satisfied: safetensors>=0.4.3 in /usr/local/lib/python3.11/dist-packages (from transformers) (0.5.3)
Requirement already satisfied: tqdm>=4.27 in /usr/local/lib/python3.11/dist-packages (from transformers) (4.67.1)
Requirement already satisfied: fsspec>=2023.5.0 in /usr/local/lib/python3.11/dist-packages (from huggingface-hub<1.0,>=0.30.0->transforme
Requirement already satisfied: typing-extensions>=3.7.4.3 in /usr/local/lib/python3.11/dist-packages (from huggingface-hub<1.0,>=0.30.0->
Requirement already satisfied: hf-xet<2.0.0,>=1.1.0 in /usr/local/lib/python3.11/dist-packages (from huggingface-hub<1.0,>=0.30.0->transf
Requirement already satisfied: charset-normalizer<4,>=2 in /usr/local/lib/python3.11/dist-packages (from requests->transformers) (3.4.2)
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.11/dist-packages (from requests->transformers) (3.10)
Requirement already satisfied: urllib3<3,>=1.21.1 in /usr/local/lib/python3.11/dist-packages (from requests->transformers) (2.4.0)
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.11/dist-packages (from requests->transformers) (2025.4.26)
Requirement already satisfied: datasets in /usr/local/lib/python3.11/dist-packages (2.14.4)
Requirement already satisfied: numpy>=1.17 in /usr/local/lib/python3.11/dist-packages (from datasets) (2.0.2)
Requirement already satisfied: pyarrow>=8.0.0 in /usr/local/lib/python3.11/dist-packages (from datasets) (18.1.0)
Requirement already satisfied: dill<0.3.8,>=0.3.0 in /usr/local/lib/python3.11/dist-packages (from datasets) (0.3.7)
Requirement already satisfied: pandas in /usr/local/lib/python3.11/dist-packages (from datasets) (2.2.2)
Requirement already satisfied: requests>=2.19.0 in /usr/local/lib/python3.11/dist-packages (from datasets) (2.32.3)
Requirement already satisfied: tqdm>=4.62.1 in /usr/local/lib/python3.11/dist-packages (from datasets) (4.67.1)
Requirement already satisfied: xxhash in /usr/local/lib/python3.11/dist-packages (from datasets) (3.5.0)
Requirement already satisfied: multiprocess in /usr/local/lib/python3.11/dist-packages (from datasets) (0.70.15)
Requirement already satisfied: fsspec>=2021.11.1 in /usr/local/lib/python3.11/dist-packages (from fsspec[http]>=2021.11.1->datasets) (202
Requirement already satisfied: aiohttp in /usr/local/lib/python3.11/dist-packages (from datasets) (3.11.15)
Requirement already satisfied: huggingface-hub<1.0.0,>=0.14.0 in /usr/local/lib/python3.11/dist-packages (from datasets) (0.31.1)
Requirement already satisfied: packaging in /usr/local/lib/python3.11/dist-packages (from datasets) (24.2)
Requirement already satisfied: pyyaml>=5.1 in /usr/local/lib/python3.11/dist-packages (from datasets) (6.0.2)
Requirement already satisfied: aiohappyeyeballs>=2.3.0 in /usr/local/lib/python3.11/dist-packages (from aiohttp->datasets) (2.6.1)
Requirement already satisfied: aiosignal>=1.1.2 in /usr/local/lib/python3.11/dist-packages (from aiohttp->datasets) (1.3.2)
Requirement already satisfied: attrs>=17.3.0 in /usr/local/lib/python3.11/dist-packages (from aiohttp->datasets) (25.3.0)
Requirement already satisfied: frozenlist>=1.1.1 in /usr/local/lib/python3.11/dist-packages (from aiohttp->datasets) (1.6.0)
```

✓ 3s completed at 5:49 PM

Evaluation Setup

- Base model is the original model – not trained on KU's dataset.
- Finetuned model is basically base model trained on KU's dataset.
- Both the models were tested in two settings – without context and with context.
- Apart from manual checks, we evaluated the model outputs on –
  - BERT Score
  - ROUGE Score
  - BLEU Score

Testing Prompts Examples

With Context

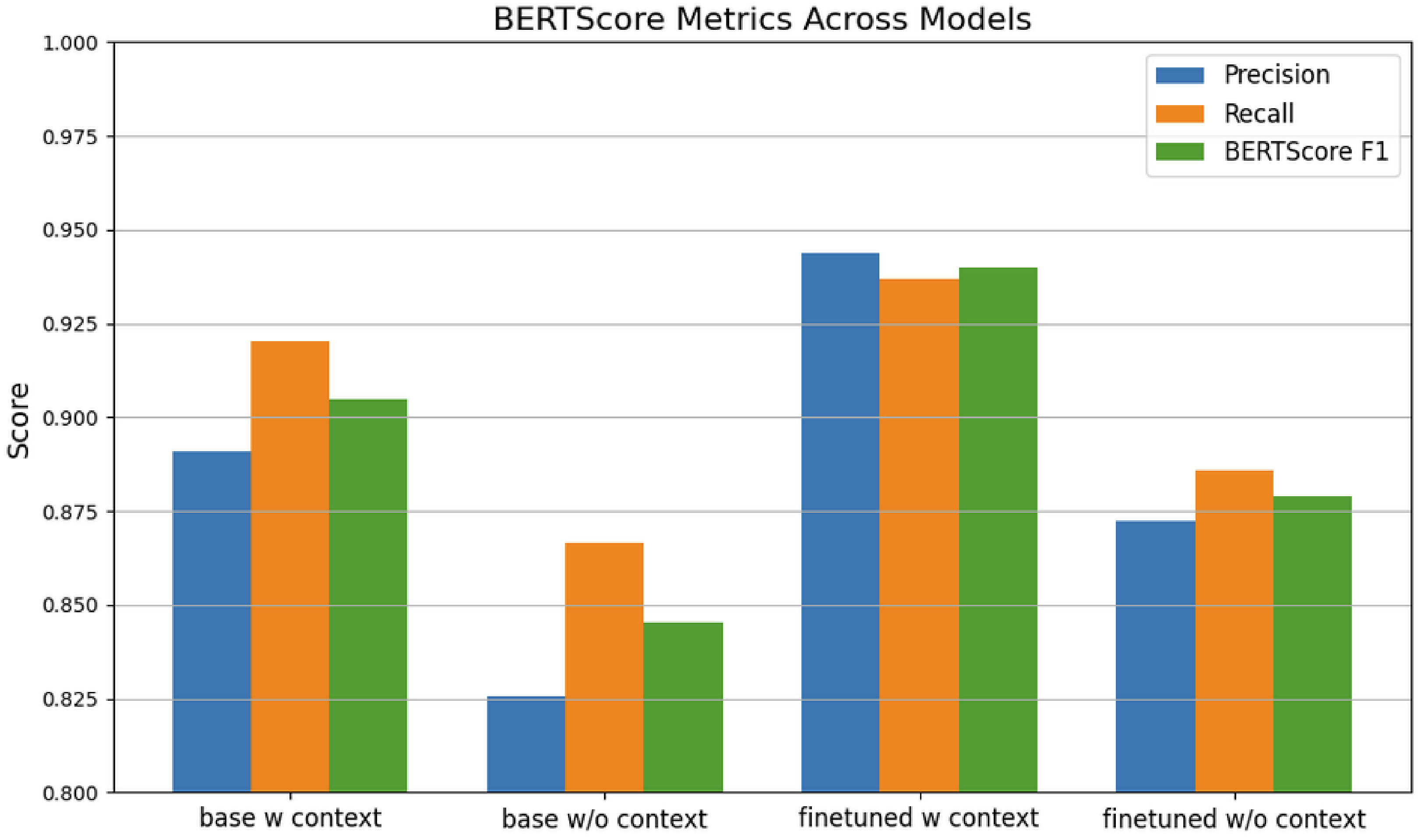
````
...Use only the information to answer the question<|eot_id|>

<|start_header_id|>user<|end_header_id|> What should international students with F-1 or J-1 status do if they need to change their name on university records?

Information: ``` International students with an F-1 or J-1 status must contact International Student Services at (785) 864-3617 to change the name on your I-20 or DS-2019. ```<|eot_id|><|start_header_id|>assistant<|end_header_id|>
````

Without Context

```
...You are a helpful assistant<|eot_id|>

<|start_header_id|>user<|end_header_id|> What should international students with F-1 or J-1 status do if they need to change their name on university records?

<|eot_id|><|start_header_id|>assistant<|end_header_id|>
```

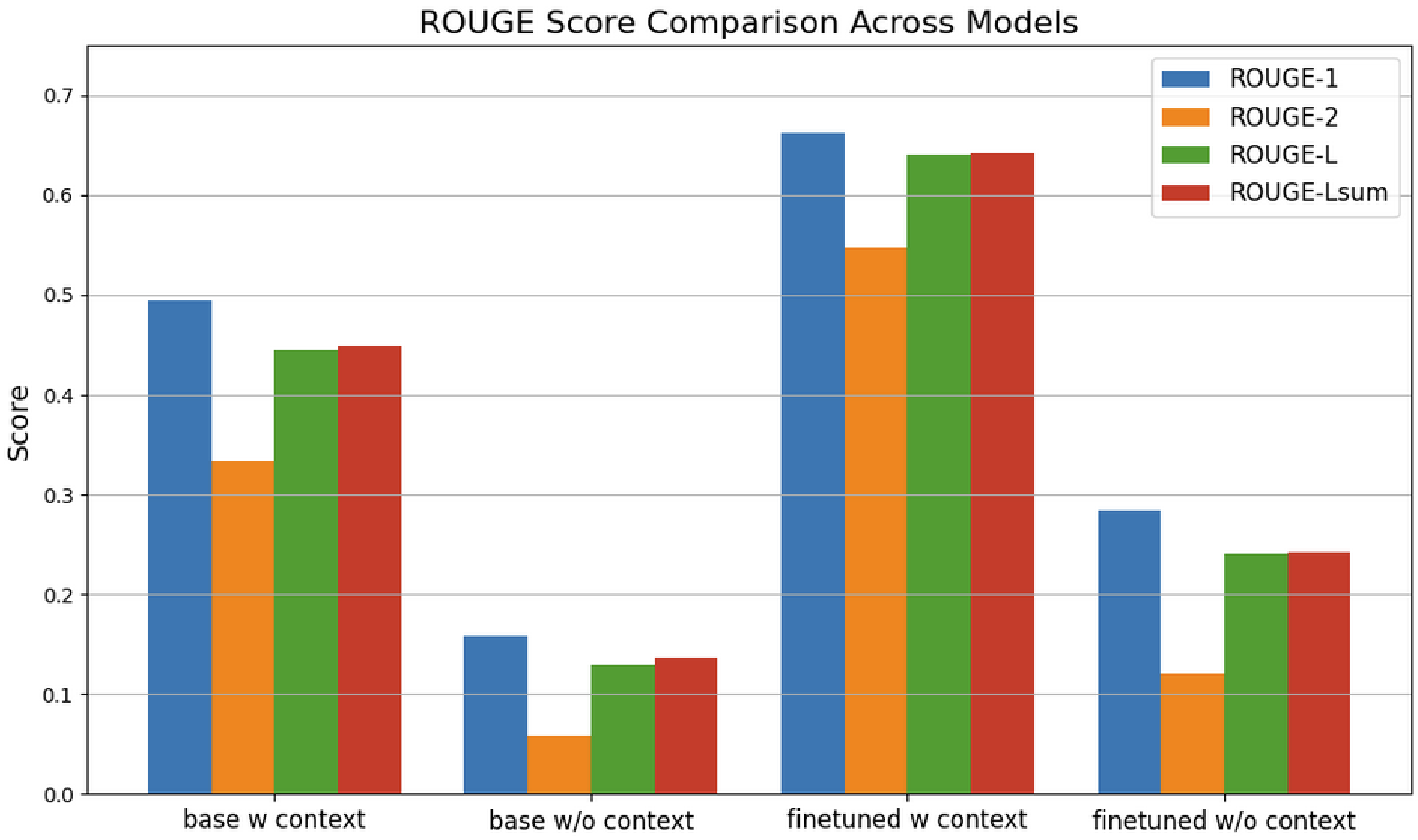
BERT Score – Semantic Similarity

- BERTScore, a more recent metric, uses contextual embeddings from pre-trained language models like BERT to compute semantic similarity between the generated answer and the actual answer.
- Without context –
  - The finetuned model significantly outperforms the base model by about 5%.
- With context –
  - The finetuned model achieves about 4% better F1 score than the base model, indicating that it produces answers semantically closest to the actual answers.



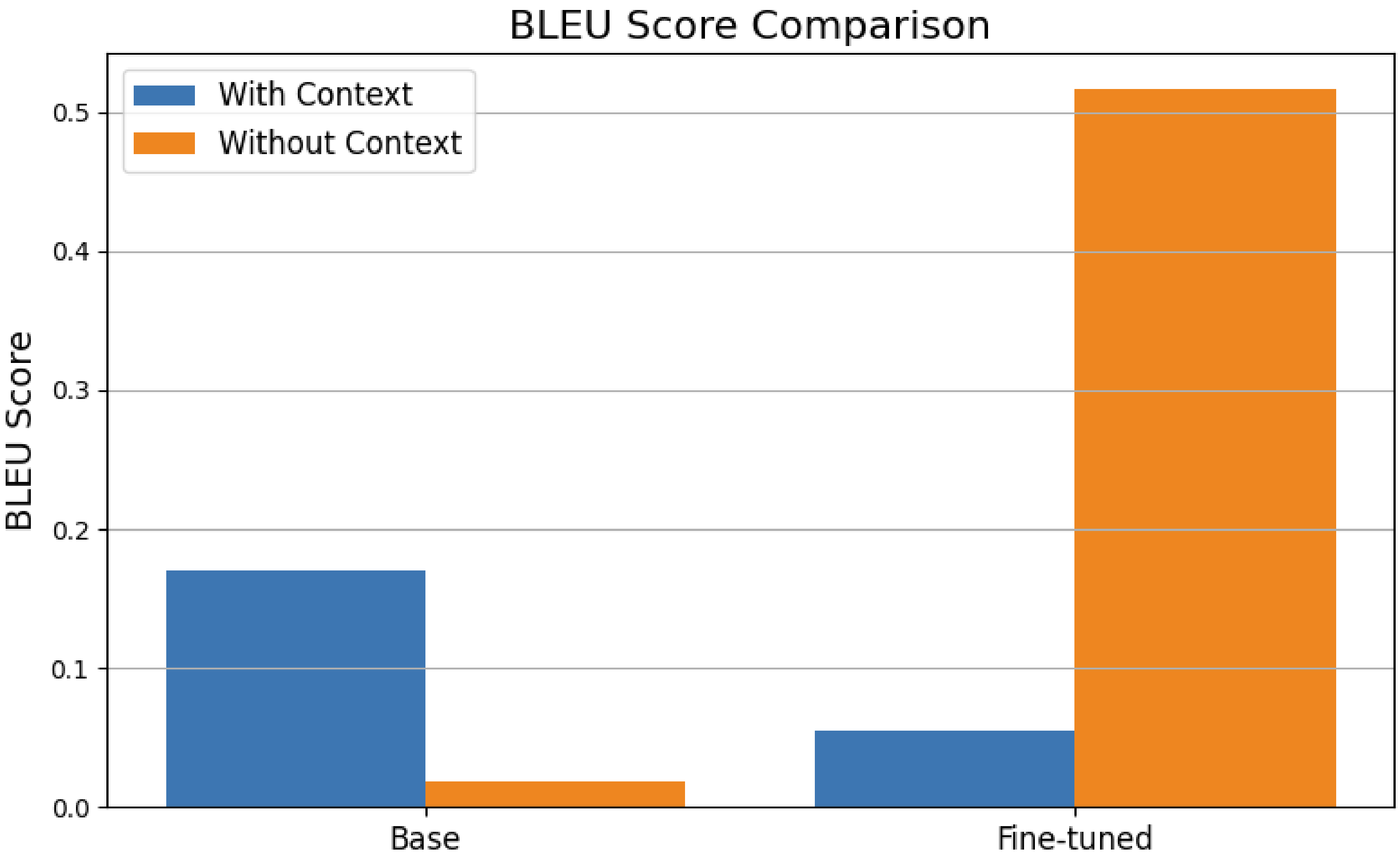
ROUGE Score – recall based n-gram overlap

- It measures how much of the actual answer matches the generated answer.
- Recall based - penalizes missing information from the answer.
- Without context –
  - Finetuned model outperforms base model by about 2x across all score.
- With context
  - Finetuned model performs significantly better than the base model.



BLEU Score – precision based n-gram overlap

- It measures how much of the generated answer matches the actual answer.
- Precision based - penalizes extra information in the answer.
- Without context –
  - Finetuned model outperforms the base model by a much large factor
- With context
  - Base model outperforms the finetuned model by almost 3x.



Conclusion

- Developed a KU-specific question-answering system by extracting data from KU webpages and finetuning Llama 3.2.
- Observed significant improvements in the model's ability to generate accurate responses, especially in no-context settings.
- Finetuned model mostly preformed better across BLEU, ROUGE, and BERTScore metrics showed consistent gains across both lexical and semantic metrics.
- Demonstrated that even small, high-quality domain-specific datasets can effectively enhance language model performance.

Limitations and Future Work

- One major limitation we observed during this project was that the presence of low-quality or noisy QAC pairs in the dataset.
    - This could be improved by adding a data quality check to the pipeline based on filtering techniques based on token length, semantic similarity, or sentence embeddings to improve alignment in QAC pairs.
- It would also be better to have a domain specific benchmark dataset for evaluation than cross validation.
- A larger number of high-quality QAC pairs would likely improve model performance and generalization.

- Automate the data pipeline using tools like Airflow, AWS EC2, and S3 for scalable and repeatable data extraction.
- Evaluate the model in user-facing applications to assess real-world usability and feedback.

Thank you!