

Brand Engagement Network

NLP Engineer: Coding Assignment Report

H M QUAMRAN HASAN

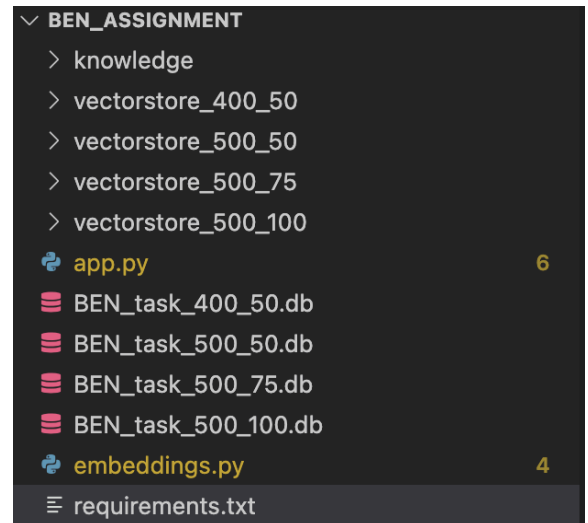
Abstract

In this project, I built a chatbot using the REST protocol for the assignment. The main idea was to utilize Retrieval Augmentation Generation and the target paper was “Llama 2: Open Foundation and Fine-Tuned Chat Models”. The model was also tested with multiple prompts and their output and latency were observed. Several chunking techniques were used to see the effect of segmentation on the performance.

CodeBase

The directory should contain the following files:

1. Requirements.txt: contains the necessary packages and libraries (pip install -r requirements.txt)
2. Vector_store: Contains the vector embeddings for the model.
Naming convention: vectorstore_CHUNK SIZE_ CHUNK OVERLAP
3. .db: stores the queries, responses, and latencies.
4. embeddings.py : run it to create the vector embeddings (python embeddings.py)
5. app.py: The code to run the application (python app.py)



Since the embeddings have already been created, the app.py file can be run directly and queries can be fed to the model. The DB to write to and the embeddings to use can be chosen in app.py (commented in the code). An extension like ThunderClient can be used:

Settings for ThunderClient:

- Method: POST
- URL: <http://localhost:5000/chat>
- Header: add content-type: application/json
- Feed the query in the Body window

The UI should look like something like the images below:

POST ⌵ http://localhost:5000/chat Send

Query Headers ³ Auth Body ¹ Tests Pre Run

HTTP Headers Raw

| | | |
|-------------------------------------|--------------|---|
| <input checked="" type="checkbox"/> | Accept | */* |
| <input checked="" type="checkbox"/> | User-Agent | Thunder Client (https://www.thunderclient.io) |
| <input checked="" type="checkbox"/> | Content-Type | application/json |
| <input type="checkbox"/> | header | value |

POST ⌵ http://localhost:5000/chat Send

Query Headers ³ Auth **Body ¹** Tests Pre Run

JSON XML Text Form Form-encode GraphQL Binary

JSON Content Format

```
1 {"query": "What are the four non-disjoint subset types defined in dataset contamination?"}
```

Experimental Setup

For the model I utilized TheBloke/Llama-2-7B-Chat-GGML which is available on huggingFace:

[llama-2-7b-chat.ggmlv3.q8_0.bin · TheBloke/Llama-2-7B-Chat-GGML at main \(huggingface.co\)](https://huggingface.co/TheBloke/Llama-2-7B-Chat-GGML)

For the parameters I chose a temperature of 0.5 to keep the responses less diverse and more focused. I also used the token size of 512.

I also added some code to keep the source documents in the output. But in the DB I just store the query, the response, and the latency. A sample output of the UI is below:

POST ⌵ http://localhost:5000/chat Send

Query Headers ³ Auth **Body ¹** Tests Pre Run

JSON XML Text Form Form-encode GraphQL Binary

JSON Content Format

```
1 {"query": "What are the four non-disjoint subset types defined in dataset contamination?"}
```

Status: 200 OK Size: 1.3 KB Time: 35.26 s

Response Headers ⁵ Cookies Results Docs {} ⋮

```
1 {
2   "latency": "35.2598651512146",
3   "query": "What are the four non-disjoint subset types defined in dataset
4   contamination?",
5   "response": "The four non-disjoint subset types defined in dataset contamination
6   are \"Clean\", \"Not clean\", \"Not dirty\", and \"Dirty\".",
7   "source_documents": [
8     "have an overall worse average score than their complement
9     , and the \"dirtiest\" samples have an overall better average score than their
10    complement. It is insufficient evidence for contamination if only one of
11    these were true. To this end, we define four (non-disjoint) subset types
12    as follows:
13    \"Clean\" samples, with less than 20% token contamination
14    ,
15    \"Not clean\" samples, with greater than (or equal to) 20% token
16    contamination,
17    \"Not dirty\" samples, with less than 80% token
18    contamination,
19    \"et al. (2022), modified to work on a PySpark cluster
20    (effectively without random access to disk).
21    Given the (nearly) parallel nature of the task, we are able to find all such 10-
22    grams (and their full lengths) in our entire dataset in around seven hours
23    (including time to tokenize), utilizing an estimated 1,500 cores.
24    As there are many confounding factors at play when determining whether
25    dataset contamination has contributed to evaluation performance
```

The prompt I used was:

```
custom_prompt_template = """Use the following information to answer the questions asked.
If you don't know the answer, just say that you don't know the answer.
Keep your answers short and within 50 words.

Context: {context}
Question: {question}

Only return the helpful answer below and nothing else.
Helpful answer:
"""
```

I kept the parameters and prompt the same for all the experiments. I ran the experiments for four conditions:

1. Chunk size = 500, chunk overlap = 50
2. Chunk size = 500, chunk overlap = 75
3. Chunk size = 500, chunk overlap = 100
4. Chunk size = 400, chunk overlap = 50

The results are collected in separate DBs which are also named following the convention: BEN_task_CHUNK SIZE_CHUNK OVERLAP. A sample DB is shown below:

| | query | response | latency |
|---|------------------------|--|---------------------|
| | Search column... | Search column... | Search column... |
| 1 | How many variants ... | "There are three variants of Llama 2-... | 33.7014200687408... |
| 2 | What hyperparamet... | "The following hyperparameters were ... | 22.84371781349182 |
| 3 | What are the three ... | "The three key dimensions of LM safe... | 30.11897611618042 |

Questions asked

I tried three categories of questions:

- Type A: Factual. Less open-ended. Usually a few words for the answer
- Type B: Not exactly factual. A bit descriptive, but the main points should stay the same
- Type C: More descriptive. Can use wordplay to paraphrase and describe the answers.

The following questions were asked in each experiment. The screenshots also show the expected answers.

1. How many variants of Llama 2-chat were released and what were their parameters? **(Type A)**

but are not releasing."

2. **LLAMA 2-CHAT**, a fine-tuned version of **LLAMA 2** that is optimized for dialogue use cases. We release variants of this model with 7B, 13B, and 70B parameters as well.

2. What hyperparameters were used for training the Llama 2 model? **(Type A)**

Hyperparameters. We trained using the AdamW optimizer (Loshchilov and Hutter, 2017), with $\beta_1 = 0.9$, $\beta_2 = 0.95$, $\text{eps} = 10^{-5}$. We use a cosine learning rate schedule, with warmup of 2000 steps, and decay final learning rate down to 10% of the peak learning rate. We use a weight decay of 0.1 and gradient clipping of 1.0. Figure 5 (a) shows the training loss for LLAMA 2 with these hyperparameters.

3. What are the three key dimensions of LM safety? **(Type B)**

Safety Benchmarks for Pretrained Models. We evaluate the safety capabilities of LLAMA 2 on three popular automatic benchmarks, pertaining to three key dimensions of LM safety.

1. **Truthfulness**, referring to whether a language model produces known falsehoods due to misconceptions or false beliefs. We employ **TruthfulQA** (Lin et al., 2021) to measure how well our LLMs can generate reliable outputs that agree with factuality and common sense.
2. **Toxicity**, defined as the tendency of a language model to generate toxic, rude, adversarial, or implicitly hateful content. We choose **ToxiGen** (Hartvigsen et al., 2022) to measure the amount of generation of toxic language and hate speech across different groups.
3. **Bias**, defined as how model generations reproduce existing stereotypical social biases. We use **BOLD** (Dhamala et al., 2021) to study how the sentiment in model generations may vary with demographic attributes.

4. What are the four non-disjoint subset types defined in dataset contamination? **(Type B)**

have an overall word average score that their complement, and the cleanest samples to have an average score than their complement. It is insufficient evidence for contamination if only one true. To this end, we define four (non-disjoint) subset types as follows:

- "Clean" samples, with less than 20% token contamination,
- "Not clean" samples, with greater than (or equal to) 20% token contamination,
- "Not dirty" samples, with less than 80% token contamination,
- "Dirty" samples, with greater than (or equal to) 80% token contamination.

5. Tell me about the annotator selection process described in the paper. **(Type C)**

and that the research community is important to

A.5.4 Annotator Selection

To select the annotators who could work on our different data collection tasks, we conducted a multi-step assessment process where we tested their understanding of our guidelines, the alignment with our quality assessment criteria, the alignment with our sensitive topics guidelines and their reading and writing skills.

The process included 4 tests:

6. What Optimizer was used for training the Llama 2 model? **(Type A)**

Answer: AdamW

Experimental Results

The results are shown below. They are also present in the .db files.

| chunk size = 500, chunk overlap = 50 | | |
|--|---|-------------|
| Question | Response | Latency |
| How many variants of Llama 2-chat was released and what were their parameters? | "Llama 2-Chat was released with 7B, 13B, and 70B parameters." | 31.40920615 |
| What hyperparameters were used for training the Llama 2 model? | "\u03b21=0.9, \u03b22=0.95, eps=10\u22125" | 26.15722084 |
| What are the three key dimensions of LM safety? | "The three key dimensions of LM safety are:\n1. Truthfulness, referring to whether a language model produces known falsehoods due to misconceptions or false beliefs." | 26.9634912 |
| What are the four non-disjoint subset types defined in dataset contamination? | "The four non-disjoint subset types defined in dataset contamination are \"clean samples\" with less than 20% token contamination, \"dirty samples\" with more than 60% token contamination, \"clean but not clean enough\" with between 20% and 60% token contamination, and \"dirty but not dirty enough\" with between 40% and 70% token contamination." | 26.79381394 |
| Tell me about the annotator selection process described in the paper. | "The authors describe a multi-stage annotation process where annotators are first asked to write a prompt and then choose between two sampled model responses based on provided criteria." | 24.06116605 |
| What Optimizer was used for training the Llama 2 model? | "AdamW optimizer." | 19.93243003 |

| chunk size = 500, chunk overlap = 75 | | |
|--|---|-------------|
| Question | Response | Latency |
| How many variants of Llama 2-chat was released and what were their parameters? | "The paper releases 34 variants of Llama 2-Chat, with 7B, 13B, and 70B parameters." | 33.42840075 |
| What hyperparameters were used for training the Llama 2 model? | "The hyperparameters used for training the Llama 2 model include:\n* Learning rate: 0.01\n* Number of epochs: 5\n* Batch size: 4\n* Discount factor: 0.97\n* Exploration-exploitation trade-off: 0.6" | 30.92569089 |

| | | |
|---|--|-------------|
| What are the three key dimensions of LM safety? | "The three key dimensions of LM safety are:\n1. Truthfulness, referring to whether a language model produces known falsehoods due to misconceptions or false beliefs.\n2. Adversarial robustness, referring to the ability of a language model to resist adversarial attacks designed to exploit vulnerabilities in the model.\n3. Demographic coverage, referring to the range of demographic categories that the model can handle without bias or errors." | 30.29359794 |
| What are the four non-disjoint subset types defined in dataset contamination? | "The four non-disjoint subset types defined in dataset contamination are:\n\u2022 Clean\u201dsamples, with less than 20% token contamination,\n\u2022 Dirty\u201dsamples, with more than 20% token contamination,\n\u2022 Mixed\u201dsamples, with both clean and dirty tokens,\n\u2022 Control\u201d samples, with the same number of clean and dirty tokens as in the original dataset." | 26.68041801 |
| Tell me about the annotator selection process described in the paper. | "The annotators were selected based on their expertise in natural language processing and their ability to understand complex prompts." | 23.31403232 |
| What Optimizer was used for training the Llama 2 model? | "The Adagrad optimizer was used for training the Llama 2 model." | 27.1806531 |

| chunk size = 500, chunk overlap = 100 | | |
|--|--|-------------|
| Question | Response | Latency |
| How many variants of Llama 2-chat was released and what were their parameters? | "There are three variants of Llama 2-Chat with different parameter sizes:\n\u2022 Llama 2-Chat-7B: This variant has 7 billion parameters.\n\u2022 Llama 2-Chat-13B: This variant has 13 billion parameters.\n\u2022 Llama 2-Chat-70B: This variant has 70 billion parameters." | 33.70142007 |
| What hyperparameters were used for training the Llama 2 model? | "The following hyperparameters were used for training the Llama 2 model:\n* Weight decay of 0.1\n* Gradient clipping of 1.0" | 22.84371781 |

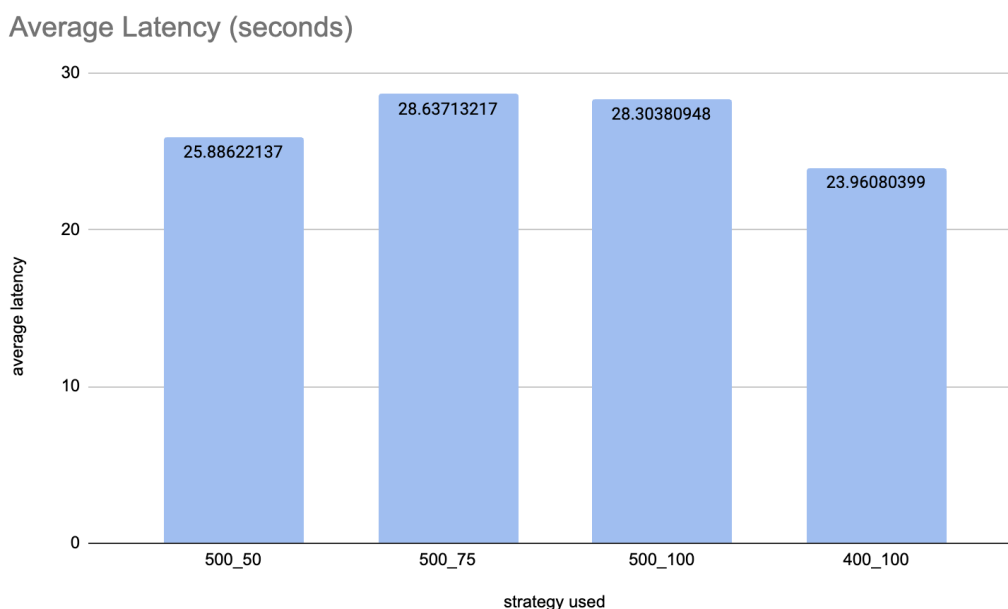
| | | |
|---|--|-------------|
| What are the three key dimensions of LM safety? | "The three key dimensions of LM safety are:\n1. Truthfulness - referring to whether a language model produces known falsehoods due to misconceptions or false beliefs.\n2. Safety - referring to the likelihood that the language model will generate unsafe or harmful content, such as hate speech or offensive material.\n3. Benchmark evaluation - referring to the process of evaluating the safety capabilities of a language model using standardized and measurable benchmarks." | 30.11897612 |
| What are the four non-disjoint subset types defined in dataset contamination? | "The four non-disjoint subset types defined in dataset contamination are \"Clean\", \"Not clean\", \"Not dirty\", and \"Dirty\"." | 32.82251596 |
| Tell me about the annotator selection process described in the paper. | "The annotators were selected based on their experience with natural language processing (NLP) tasks, including text classification, sentiment analysis, and named entity recognition." | 23.76174402 |
| What Optimizer was used for training the Llama 2 model? | "AdamW optimizer (Loshchilov and Hutter, 2017)" | 26.57448292 |

| chunk size = 400, chunk overlap = 50 | | |
|--|--|-------------|
| Question | Response | Latency |
| How many variants of Llama 2-chat was released and what were their parameters? | "There were three variants of Llama 2-Chat released with different hyperparameters and the exact hyperparameter configurations for each are provided in Appendix A of the paper." | 27.28497005 |
| What hyperparameters were used for training the Llama 2 model? | "\u03b22 = 0.95, eps = 10\u22125, learning rate schedule = cosine." | 22.63216209 |
| What are the three key dimensions of LM safety? | "The three key dimensions of LM safety are:\n1. Content (Unsafe): This refers to the content of the text that is potentially harmful or offensive, such as hate speech, graphic violence, or adult content.\n2. User-generated content (Unsafe): This refers to content created by users that may be unsafe, such as hate speech, graphic violence, or content promoting terrorist acts.\n3. Behavior (Safe): This refers to the behavior of users when interacting with AI language models, including their ability to use the technology responsibly and ethically." | 26.46206617 |
| What are the four non-disjoint subset types defined in dataset contamination? | "The four non-disjoint subset types defined in dataset contamination are:\nA. Textual. B. Named. C. Labelled. D. Multipurpose." | 19.47297573 |

| | | |
|---|--|-------------|
| Tell me about the annotator selection process described in the paper. | "We first ask potential annotators to complete a series of tasks that assess their understanding of our guidelines, their alignment with our quality assessment criteria, and their ability to read and write clearly. We evaluate these results before selecting annotators for our data collection tasks." | 23.66398001 |
| What Optimizer was used for training the Llama 2 model? | "The optimizer used for training the Llama 2 model is stochastic gradient descent (SGD)." | 24.24866986 |

Discussion

Response Latency:



From the figure we can see that the configuration of 400_100 shows the least latency. This is expected because smaller chunks can be processed more efficiently. For the 500_x configurations, we can see a gradual increase in latency when the overlap is increased. This aligns with the expectation because when the overlap increases more time is needed for processing.

Response Quality: good ■ average ■ bad ■

| Question | 500_50 | 500_75 | 500_100 | 400_50 |
|----------|--|---|---|--|
| 1 | "Llama 2-Chat was released with 7B, 13B, and 70B parameters." | "The paper releases 34 variants of Llama 2-Chat, with 7B, 13B, and 70B parameters." | "There are three variants of Llama 2-Chat with different parameter sizes:\n\u2022Llama 2-Chat-7B: This variant has 7 billion parameters.\n\u2022Llama 2-Chat-13B: This variant has 13 billion parameters.\n\u2022Llama 2-Chat-70B: This variant has 70 billion parameters." | "There were three variants of Llama 2-Chat released with different hyperparameters and the exact hyperparameter configurations for each are provided in Appendix A of the paper." |
| 2 | "\u03b21=0.9, \u03b22=0.95, eps=10\u2022125" | "The hyperparameters used for training the Llama 2 model include:\n* Learning rate: 0.01\n* Number of epochs: 5\n* Batch size: 4\n* Discount factor: 0.97\n* Exploration-exploitation trade-off: 0.6" | "The following hyperparameters were used for training the Llama 2 model:\n* Weight decay of 0.1\n* Gradient clipping of 1.0" | "\u03b22 = 0.95, eps = 10\u2022125, learning rate schedule = cosine." |
| 3 | "The three key dimensions of LM safety are:\n1. Truthfulness, referring to whether a language model produces known falsehoods due to misconceptions or false beliefs." | "The three key dimensions of LM safety are:\n1. Truthfulness , referring to whether a language model produces known falsehoods due to misconceptions or false beliefs.\n2. Adversarial robustness, referring to the ability of a language model to resist adversarial attacks designed to exploit vulnerabilities in the model.\n3. Demographic coverage, referring to the range of demographic categories that the model can handle without bias or errors." | "The three key dimensions of LM safety are:\n1. Truthfulness - referring to whether a language model produces known falsehoods due to misconceptions or false beliefs.\n2. Safety - referring to the likelihood that the language model will generate unsafe or harmful content, such as hate speech or offensive material.\n3. Benchmark evaluation - referring to the process of evaluating the safety capabilities of a language model using standardized and measurable benchmarks." | "The three key dimensions of LM safety are:\n1. Content (Unsafe): This refers to the content of the text that is potentially harmful or offensive, such as hate speech, graphic violence, or adult content.\n2. User-generated content (Unsafe): This refers to content created by users that may be unsafe, such as hate speech, graphic violence, or content promoting terrorist acts.\n3. Behavior (Safe): This refers to the behavior of users when interacting with AI language models, including their ability to use the technology responsibly and ethically." |

| | | | | |
|---|---|---|---|--|
| 4 | "The four non-disjoint subset types defined in dataset contamination are \"clean samples\" with less than 20% token contamination, \"dirty samples\" with more than 60% token contamination, \"clean but not clean enough\" with between 20% and 60% token contamination, and \"dirty but not dirty enough\" with between 40% and 70% token contamination." | "The four non-disjoint subset types defined in dataset contamination are:\n\u2022 Clean\u201d samples, with less than 20% token contamination,\n\u2022 Dirty\u201d samples, with more than 20% token contamination,\n\u2022 Clean but not clean enough\u201d samples, with between 20% and 60% token contamination,\n\u2022 Dirty but not dirty enough\u201d samples, with between 40% and 70% token contamination, with the same number of clean and dirty tokens as in the original dataset." | "The four non-disjoint subset types defined in dataset contamination are \"Clean\", \"Not clean\", \"Not dirty\", and \"Dirty\"." | "The four non-disjoint subset types defined in dataset contamination are:\nA. Textual. B. Named. C. Labelled. D. Multipurpose." |
| 5 | "The authors describe a multi-stage annotation process where annotators are first asked to write a prompt and then choose between two sampled model responses based on provided criteria." | "The annotators were selected based on their expertise in natural language processing and their ability to understand complex prompts." | "The annotators were selected based on their experience with natural language processing (NLP) tasks, including text classification, sentiment analysis, and named entity recognition." | "We first ask potential annotators to complete a series of tasks that assess their understanding of our guidelines, their alignment with our quality assessment criteria, and their ability to read and write clearly. We evaluate these results before selecting annotators for our data collection tasks." |
| 6 | "AdamW optimizer." | "The Adagrad optimizer was used for training the Llama 2 model." | "AdamW optimizer (Loshchilov and Hutter, 2017)" | "The optimizer used for training the Llama 2 model is stochastic gradient descent (SGD)." |

Question 1:

Both 500_50 and 500_100 provided correct info, however, 500_100 was more on the descriptive side. 500_75 also gave an almost correct response but with an addition of 34 variants which was absent in the document. 400_50 provided the worst response.

Question 2:

For this one, all of them gave satisfactory answers. However, none of them gave the entire list of parameters. They all gave subsets. However, 500_75 for some reason gave a value of 0.1 which was not present in the info.

Question 3:

When the model faced a Type B question, the suffering was visible. I was particularly looking for the three keywords: Truthfulness, Toxicity, and Bias. The response by 400_50 was very different from the correct answer. The response by 500_50 did contain truthfulness, but it was not able to talk about the other two. Although the other two configurations did not also get all three, but they got two of them correct. Although they paraphrased the statements quite a lot, the main idea was there.

Question 4:

For the fourth question only the 500_100 configuration gave an answer that had all correct subsets. The other configurations gave some correct and some wrong subsets and the 400_50 configuration gave all wrong subsets.

Question 5:

In this question, the 500_100 configuration somehow fell behind and gave an answer that was not present in the gold standard answer. Surprisingly the 400_100 configuration gave the closest answer to the gold standard.

Question 6:

This was a very simple question and both 500_50 and 500_100 configurations gave correct and precise answers. The other configurations due to some reason came up with optimizers that were nowhere to be seen.

Overall, we can conclude that the 500_50 and 500_100 configurations performed much better than the other configurations. We also see that Type C is the hardest for the models, which was expected, as it is more open-ended and allows the model to play with the answers.

Conclusion

This project allowed me to build my first chatbot to answer questions relevant to the Meta paper. After trying several configurations and question types, I was able to see how LLMs suffer in answering certain question types and how the configuration also plays a big role in response generation. Further experiments can be done with proprietary LLMs like GPT-4 which is much improved and has better performance and also allows a much bigger token size.