Llama 3 Evaluation Based on Question Answering Dataset SQuAD

Zimo Peng
z5peng@ucsd.edu

1 Introduction

This study focuses on measuring the performance of the large language model, llama-3-8B (Hugo Touvron, 2023) (HuggingFace, 2024), in the downstream task of question answering. Question answering has become an essential task in LLM evaluations, thanks to the hype of LLM and its wide application in the tech field. There are many existing benchmarks to measure the performance of LLM in multiple different perspectives and fields. Due to hardware and time constraints, I picked the llama model with fewer parameters than the 40B version. Llama 3 is also one of the most powerful and up-to-date open-source LLM I can fine-tune locally.

I evaluate the llama-3-8B model through the SQuAD benchmark. I tested the question answering performance of the model against a subset of the SQuAD dataset. By measuring the exact match (EM) score, the percentage of predictions that match any one of the ground truth answers exactly, I was able to assess the model's performance.

The main limitations of llama-3-8B model identified in this study include the model's tendency to hallucinate, to produce redundant tokens in the generation process without restricting its response to the original tokens, and to struggle with maintaining logical consistency between connected entities, impacting the accuracy and reliability of its output.

Please note that the research paper for Llama 3 has not yet been released, so here I cited the paper of the base model Llama 2.

2 Dataset

The Stanford Question Answering Dataset (Rajpurkar et al., 2016), SQuAD, is a collection of question-answer pairs derived from Wikipedia articles. It mainly focuses on testing the reading comprehension, the ability to read text and then answer questions about it, of language models. For SQuAD, systems must select the answer from all possible spans in the passage. In other words, the correct answers of questions can be any sequence of tokens in the given text.

Because the questions and answers are produced by humans through crowdsourcing, SQuAD features a diverse range of range of question and answer types, which makes it challenging for this benchmark. The answers involve numerical types like Date or math numbers and non-numerical types like person, location, and other entities using NER tags.

SQuAD 1.1 contains 107,785 question-answer pairs on 536 articles. SQuAD2.0 (open-domain SQuAD, SQuAD-Open), the latest version, combines the 100,000 questions in SQuAD1.1 with over 50,000 un-answerable questions written adversarially by crowdworkers in forms that are similar to the answerable ones.

I picked SQuAD1.1 due to its sole focus on quetion answering without the distraction of un-answerable questions. SQuAD2.0 would be a nice future extension of this study. Also due to time constraints and hardware constraints, I was not able to evaluate the model against the entire dataset, rather I picked a subset of 1,000 questions as the benchmark. The subset is picked at random to preserve the general quality and property of the dataset. I preprocessed the dataset in a simple way by removing redundant columns such as the ID, the title, and the answer. The only remaining columns are the question, the answers, and the context. Samples of the SQuAD data are shown in Figure 1.

The questions are of various lengths, while the answers are typically short, spanning from one single word to a whole sentence. One thing worth

id string · lengths	title string · lengths	context string · lengths	question string · lengths	answers sequence
24 24	3 59	151 3.71k	1 25.7k	
5733849bd058e614000b5c56	University_of_Notre_Dame	In 1842, the Bishop of Vincennes, Célestine Guynemer de la Hailandière, offered land to…	In what year was Father Edward Sorin given two years to create a college?	{ "text": ["1842"], "answer_start": [3] }
5733849bd058e614000b5c57	University_of_Notre_Dame	In 1842, the Bishop of Vincennes, Célestine Guynemer de la Hailandière, offered land to…	Which individual offered land to Father Edward Sorin?	{ "text": ["Célestine Guynemer de la Hailandière"], "answer_start": [34] }
5733849bd058e614000b5c58	University_of_Notre_Dame	In 1842, the Bishop of Vincennes, Célestine Guynemer de la Hailandière, offered land to…	Which church was Father Edward Sorin representing?	{ "text": ["the Congregation of the Holy Cross"], "answer_start": [111] }

Figure 1: Sample data from SQuAD

mentioning is that the answer, or the ground truth, within the dataset is a list containing three chunks of phrases. Each chunk is an acceptable answer to the question. Thus for each given pair of question and context, there are 3 correct answers. Sometimes the three chunks are identical to each other.

I calculate and compare the F1 score and exact match score of llama-3-8B against the ground truth, evaluating the model's performance.

3 Analysis Approach

Given the subset of SQuAD, I constructed the pipeline for llama-3-8B to make predictions on each single pair of questions and context in the subset and then compared the predicted result with the ground truth so that I could calculate the EM score and F1 score. The higher the EM score and F1 score, the better the performance of Llama 3 on the SQuAD dataset.

One challenging aspect is that I must limit the model's predictions solely to the tokens from the original question, otherwise even though llama-3-8B is able to produce the correct answers, judged by my human eyes, these answers are always written in different formats and lengths apart from the tokens in the original context, leading to a poor exact match score of 0.00% and an F1 score of 0.21. Samples of these predictions are shown in Table 1. A possible solution to this problem is that I can switch the evaluation metrics from F1 and EM score to more comprehensive metrics like BLEU and ROUGE, or even feed the answer to a second language model like GPT4 as a judge.

Another little challenging aspect is that the output of the model needs to be parsed properly, so that it does not include redundant tokens from the question part or the context part, solely containing the answer.

Once I add the limit, the model results in a higher exact match score average of 18% and an F1 score average of 0.33%. Please note that the model produces different predictions or inference

Table 1: Examples of Unmatched Result

Prediction	Ground Truth
Newton was 26	['26', '26', '26']
Independence Day: Resurgence	['Resurgence', 'Resurgence', 'Resurgence']
The theme of Super Bowl 50 was "Golden"	['golden anniversary', 'gold-themed', 'golden anniversary']

results based on the same dataset across different rounds of predictions, thus the evaluation results vary from time to time. To understand the range of possible EM scores and F1 scores, I ran 10 rounds of predictions based on the same dataset and calculated the average of these scores. The results are shown in Figure 2.

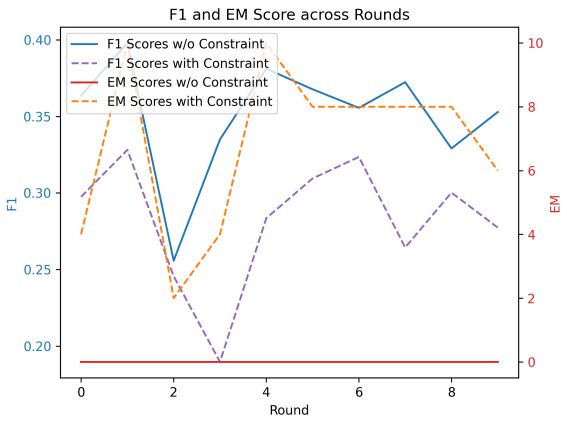

Figure 2: Comparison of F1 and EM scores across different rounds.

Even though the metrics have increased, they are still far from being good. The performance of Llama-3-8B on SQuAD should be 76.4, according to the HuggingFace scoreboard (HuggingFace, 2024). The potential reasons of this difference

could be I did not run through the entire SQuAD benchmark, I did not use 1-shot learning as indicated by the scoreboard, or the instructions that I provided to my llama-3 model to only use tokens from the original context differ from the HuggingFace's.

I also manually went through some question-answer pairs that indicate errors in the inference process, gathered them into a data file, and analyzed them in the following section.

4 Error Analysis and Categorization

Thanks to the low EM score and F1 score that the model produces, it was not hard for me to identify the errors and the specific examples. Observing the set of incorrect answers, I identified some recurring patterns including frequent question types that the model failed to generate a correct answer on. Before showing the error samples, let me introduce the different types of reasoning and their corresponding data samples covered in the SQuAD dataset. The reasoning is divided into 5 types: lexical variation (synonymy), lexical variation (world knowledge), syntactic variation, multiple sentence reasoning, and ambiguous, as shown in Table 2. The table is originally from the SQuAD paper.

From my observation, the majority of the reasoning types that the model fails to create exact answers for are ambiguous reasoning and syntactic variation. Also even though I managed to ask the model to only use sequences of tokens from the given context, llama-3-8B does still use random combinations of tokens from the context. Some sample errors are shown in Table 3.

Due to the inherent, manual, and laborious process of identifying the reasoning type that each problem belongs to, I was not able to accumulate enough statistics that I can present as accurate descriptions of the frequencies of the occurrences of these reasoning type-related questions.

5 Discussion

One major reason for the error is the hallucination of LLM. This has been a long-lasting problem since the birth of LLM, yet it is still unable to be completely avoided in the current status of the LLM development. Hallucination refers to the incidance where the model generates incorrect or unverifiable information that looks trustworthy, which has been widely documented in studies such as "TruthfulQA: Measuring How Models Mimic Human Falsehoods" (Lin et al., 2021). As described in the paper, the hallucination of LLM often stems from the model's training on noisy data or its attempt to generalize from the patterns observed in the training dataset.

Another major reason is the problematic format of the generated predictions. Examining a lot of them, I realize that even tho the prediction is accurate when compared to the ground truth, there are always extra contents that should be removed if I have a tighter manipulation over the model's output. This often reflects the model's tendency to generate redundant tokens or include additional context that the model finds relevant. This could be due to the influence of the diverse and extensive information the model has been trained on, as suggested by Bender et al. (2021) in "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?".

Another potential reason that I discovered is that sometimes the model has a hard time dealing with the inherent logic connections between different entities in the given tokens. For example, when there is a one-directional logic connection between two entities A and B, say A is Y of B. This one-directional logic connection usually indicates a reverse logic connection from B to A, say B is X of A. Existing LLMs do not capture this kind of logic connection very accurately. This limitation can be traced back to the model's inability to understand deep semantic relationships or to maintain consistency in logical reasoning across contexts. The paper "Evaluating Commonsense in Pre-trained Language Models" (Zhou et al.) discussed this phenomenon. The challenge lies in the inherently probabilistic nature of language modeling, which may not always align with the logical or factual correctness in the original settings.

The complexity and diversity of SQuAD datasets also lead to the prediction errors of Llama-3-8B.

6 The way forward

To improve the model's performance on the SQuAD benchmark, many existing popular approaches can be applied. For example, fine-tuning the model with another subset of the SQuAD dataset should significantly increase the evaluation metrics (EM score and F1 score). PEFT and LoRA come in handy in fine-tuning the model.

Table 2: Reasoning Categories of SQuAD questions

Reasoning	Description	Question Example
Lexical variation (synonymy)	Major correspondences between the question and the answer sentence are synonyms	Q: What is the Rankine cycle sometimes called? Sentence: The Rankine cycle is sometimes referred to as a practical Carnot cycle.
Lexical variation (world knowledge)	Major correspondences between the question and the answer sentence require world knowledge to resolve.	Q: Which governing bodies have veto power? Sen.: The European Parliament and the Council of the European Union have powers of amendment and veto during the legislative process.
Syntactic variation	After the question is paraphrased into declarative form, its syntactic dependency structure does not match that of the answer sentence even after local modifications.	Q: What Shakespeare scholar is currently on the faculty? Sen.: Current faculty include the anthropologist Marshall Sahlins, ..., Shakespeare scholar David Bevington.
Multiple sentence reasoning	There is anaphora, or higher-level fusion of multiple sentences is required.	Q: What collection does the V&A Theatre& Performance galleries hold? Sen.: The V&A Theatre & Performance galleries opened in March 2009. ... They hold the UK's biggest national collection of material about live performance.
Ambiguous	We don't agree with the crowdworkers' answer, or the question does not have a unique answer.	Q: What is the main goal of criminal punishment? Sen.: Achieving crime control via incapacitation and deterrence is a major goal of criminal punishment.

Also fine-tuning the model with other datasets besides SQuAD could also increase the model's performance on reading comprehension tasks. As proved by the public scoreboard of Llama-3-8B instruct, the systematically fine-tuned version of Llama-3-8B performs significantly better than the base model.

It may not be necessary to change the evaluation metrics, as EM score and F1 score are sufficient in evaluating the model's performance. I just need to give llama-3 clear instructions on how to limit its predictions based on the existing tokens in the given context. But also note that this freedom in choosing the right instructions may lead do different evaluation results. As discussed in Part 2, it may be a good idea to use other metrics like BLEAU or ROUGE to compensate for this limitation, providing a more comprehensive evaluation. Additionally, the data should not be adjusted, as the meaning of SQuAD is to evaluate the LLM's reading comprehension. We are not trying to improve the benchmark so that the LLM can perform better, rather the LLM should adopt other techniques to improve its performance on the existing benchmark.

7 Implementation

To improve upon the base model Llama-3-8B's performance on the subset of SQuAD, I implemented the fine-tuning approach. Despite the initial model usage info that I found on HuggingFace, I added a lot of code related to SQuAD evaluation and fine-tuning. I completed a naive pipeline to fine-tune the model based on a subset of SQuAD data pairs that are not included in the evaluation set. I imported the pre-trained base model from existing libraries, then I added the fine-tune module and SQuAD evaluation module. The only code file that I submitted on gradescope covers the entire pipeline.

I developed the evaluation pipeline using Google Colab Pro. Initially, I ran everything on the standard Google Colab, but I soon ran out of computing units, the 15GB GPU RAM limit of Tesla T4 is also often exceeded during the evaluation process. I experimented with L4 and A100

Table 3: Samples of Errors

Reason	Question	Ground Truth	Prediction
Incorrect format	Which entity did Notre Dame hire to build a parking structure outside of Eddy Street Commons?	[”the City of South Bend”]	”South Bend”
Hallucination	When did Destiny’s Child end their group act?	[”June 2005”]	”July 2004”
Logic connection	Who managed the Destiny’s Child group?	[”Mathew Knowles”]	””
Ambiguous	What is the name of the service that gets local businesses contract chances	[”Business Connect”]	”Stadium”

GPU, which are significantly faster and contain bigger GPU RAM. Using L4 GPU, it took slightly more than 1 hour to finish one round of inference.

Adding the finetune module did not significantly increase the model’s performance measure by F1 and EM scores, which could be due to the small size of the extra data pairs that I used to finetune the model.

8 Conclusion

This study evaluated the performance of Lhama-3-8B model over a chosen subset of the SQuAD 1.1 question answering dataset. Despite achieving a lower than expected EM score and F1 score compared to the established scoreboard, the analysis identified key issues that led to this result, including the model’s tendency to hallucinate, to generate tokens that are out of the required format, and to poorly recognize the semantic relationship between entities.

Through further fine-tuning using other sample QA pairs from SQuAD, Lhama-3-8B shows a higher performance in the increase of EM score and F1 score. This study evaluated the reading comprehension ability of Lhama-3-8B and demonstrated the effectiveness of LLM fintune techniques like LoRA and PEFT, exploring the technique of using additional datasets to efficiently enhance the model’s reading comprehension ability when answering questions from SQuAD benchmark.

9 Acknowledgements

I would like to thank Google Colab, HuggingFace, Meta, Unsloth, OpenAI, David Ondrej’s video (Ondrej, 2024), the Standford researchers who created SQuAD, and the entire teaching staff of CSE 256 for this precious opportunity to explore NLP tasks and research beyond the class.

Using GenAI tools to boost the study efficiency, I used ChatGPT4 in running the code evaluation and debugging. The changes I made to my code based on ChatGPT suggestions are minor changes. Since I already wrote the main framework of the evaluation pipeline and the fine-tuning pipeline. When writing the report, I did use ChatGPT4 to proof-read some of the paragraphs, mainly to reduce the amount of grammar mistakes. I made minor changes to the report contents based on ChatGPT suggestions.

References

HuggingFace (2024). Meta-llama-3-8b.

Hugo Touvron, Louis Martin, K. S. P. A. A. A. Y. B. N. B. S. B. P. B. S. B. D. B. L. B. C. C. F. M. C. G. C. D. E. J. F. J. F. W. F. B. F. C. G. V. G. N. G. A. H. S. H. R. H. H. I. M. K. V. K. M. K. I. K. A. K. P. S. K. M.-A. L. T. L. J. L. D. L. Y. L. Y. M. X. M. T. M. P. M. I. M. Y. N. A. P. J. R. R. R. K. S. A. S. R. S. E. M. S. R. S. X. E. T. B. T. R. T. A. W. J. X. K. P. X. Z. Y. I. Z. Y. Z. A. F. M. K. S. N. A. R. R. S. S. E. T. S. (2023). Llama 2: Open foundation and fine-tuned chat models.

Ondrej, D. (2024). Lhama 3 for my use case.

Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. (2016). Squad: 100,000+ questions for machine comprehension of text.