Assessing LLaMA's Zero and Few-Shot Performance across diverse tasks and prompts

Dimple Naresh Nachnani
Department of Computer Science
Purdue University
West Lafayette, IN 47906
dnachnan@purdue.edu

Sree Sai Ankit Rao Pittala
Department of Computer Science
Purdue University
West Lafayette, IN 47906
pittala@purdue.edu

Abstract

Traditional NLP models, including earlier GPT versions, used to demand substantial amounts of specialized training data to excel in specific tasks. However, with the arrival of large language models like LLaMA and GPT-3, which possessed unique few-shot capabilities, the laborious and time-consuming process of fine-tuning these models for various applications became unnecessary. There has been extensive discussion on enhancing the performance of such Large Language Models (LLMs) like GPT, focusing on prompting techniques such as zero-shot, few-shot prompting, prompt engineering, and the chain of thought. When comparing zero-shot and few-shot techniques, some results suggest that zero-shot performance may outshine few-shot for particular tasks that don't require explicit examples, especially when the models already possess inherent knowledge gained during pre-training. We investigated this aspect in our project, where we understood how prompts and specific tasks influence the zero-shot versus few-shot performance of LLaMA, a relatively new LLM developed by Meta, which hasn't been thoroughly explored yet. Github - Code (2023)

1 Introduction and motivation

In July 2023, Meta made headlines by unveiling LLaMA-2, its open-source Large Language Model, short for Large Language Model Meta AI-2. This was a modification of the LLaMA model released earlier in February. Recognizing that smaller models are more accessible to the research community due to various computational and resource constraints, Meta introduced LLaMA-2 in various sizes, ranging from 7 billion to 70 billion parameters. This approach stands in contrast to other models like GPT-3, which boasts a massive 175 billion parameters and is proprietary, restricting its usage through an API. In contrast, LLaMA's largest model comprises 70 billion parameters, and importantly, all of its models are open source. Moreover, LLaMA-2 outperforms GPT-3 on a range of tasks. This characteristic makes LLaMA-2 highly appealing for researchers looking to delve into the world of Large Language Models and conduct their own investigations. Furthermore, due to its recent release, much of its capabilities have not been explored, making it a highly suitable model for our project's exploration and experimentation.

Earlier works on Large Language models like GPT-3 have explored various techniques to improve the performance of Large Language models primarily through prompt engineering. Prompt engineering refers to the technique of carefully crafting the prompts or instructions that are input to large language models like LLaMA in order to get better, more useful outputs. A prominent technique used for prompt engineering is few-shot learning. Few-shot learning refers to the ability of certain machine learning models to learn new concepts from just a few examples. A special case of few-shot learning is when we do not provide any examples, instead, we just specify the task description and the input test prompt. This is also referred to as zero-shot learning. Significant approaches for improving the quality of prompts in few-shot learning are judicious selection of few-shot examples through methods like KATE[Liu et al. ([n. d.])], calibration of these examples[Zhao et al. (2021)], usage of prompt templates[Gao et al. (2021)], trying out zero-shot, one-shot and varied number of examples in few-shot learning, etc. The outcomes of these studies have demonstrated a wide range of results and findings. In our project, we aim to conduct a more in-depth investigation into a particular observation regarding GPT-3: its ability to perform well without any examples (zero-shot performance) surpassing its performance when given a few examples (few-shot performance) for specific tasks.

2 Problem Statement

In this project, our goal was to evaluate and compare the performance of LLaMA-2 in zero-shot versus few-shot scenarios, seeking to determine if its performance aligns with the findings reported in a previous study referenced as Reynolds and McDonell (2021). Additionally, we also investigated the specific tasks and their corresponding datasets where these findings are applicable and the efficiency of various prompt engineering techniques.

3 Related work

During our literature survey, we came across the paper - Prompt Engineering for Large Language Models: Beyond the Few-Shot Paradigm[Reynolds and McDonell (2021)] which discusses the limitations of present methods for evaluating large language models and proposes new approaches to prompt engineering. The researchers used GPT-3 as their model to conduct a case study to show that zero-shot prompts can outperform few-shot prompts while analyzing the features and capabilities of the model. During their core experiments, the researchers explored a single, illustrative example, a French-to-English translation task. They found that zero-shot prompts can match and even surpass few-shot performance. The zero-shot accuracy can be improved greatly with minor prompt engineering. They suggested the usage of "Simple Colon" and "Masterful translator" prompts for performing this. Overall, the paper enhances the rise of prompt programming by investigating its theory and offers how these various methods can be used in the evaluation and application of language models

4 Proposed approach

Our research experiments were performed and evaluated across two dimensions.

1. Prompt Engineering
   - In this study, we assessed how LLaMA-2 performs under different prompting methods. These methods encompass both straightforward techniques like using colons and more intricate strategies like the master approach.
   - Additionally, we investigated how the inclusion of encouraging, discouraging, or neutral language when describing the master in the prompt influenced the model's performance.
   - We used 5 prompt techniques - Simple Colon 0-shot, Master Method, Master Encouraged, Master Discouraged, and Simple Colon 1-shot. The Simple Colon and Master methods for prompting were proposed in Reynolds and McDonell (2021). The Master Encouraged and Master Discouraged are modifications of the original Master method where we describe the master with positive or negative adjectives [“extremely efficient masterful translator”, “incompetent masterful translator”] respectively.
2. Task
   - Reynolds and McDonell (2021) presents findings regarding zero-shot versus few-shot learning using the GPT-3 model, focusing exclusively on a single task—French to English translation. Notably, it lacks any information about the applicability of these results to other tasks. This has driven our interest in examining the generalizability of these findings across various tasks.

- Upon careful deliberation, we decided to compare performance in three different task domains: the original French-English translation task, tasks involving natural language generation such as summarization, and tasks related to classification, such as sentiment analysis.

Task details:

Task	Dataset	Source	Proposed evaluation
Translation	English (EN) to French (FR)	Bojar et al. (2014)	BERT Score, BLEU Score
Translation	English (EN) to Czech (CS)	Bojar et al. (2016)	BERT Score
Classification-Sentiment Analysis	IMDB	Maas et al. (2011)	0/1 Accuracy Score
Text Generation-Summarization	CNN Daily Mail	See et al. (2017)	BERT Score

5 Hypothesis

With our existent knowledge, initially, we had hypothesised the following outcomes.

Tasks where Zero Shot Prompting is better than Few shot.

1. In the case of Classification, and Translation zero-shot should be better than one-shot.
2. Translation by LLaMA-2 should be less efficient as compared to GPT-3 since LLaMA's training data lacks much of multilingual text.
3. For Text Generation, it is possible that the model learns the importance while exploring examples so few-shot can be better as compared to zero-shot.

Effects of Prompt Engineering

1. Simple Colon and Master Encouraged should be better as compared to the other prompt techniques because of their simplicity and positive words respectively.
2. Master Discouraged should be comparable to neutral since we did not explicitly use strong negative words.

6 Prompts and Results

Below are the results obtained for each of the tasks where Prompt 1 - Simple Colon 0-Shot, Prompt 2 - Simple Colon 1-Shot, Prompt 3 - Master, Prompt 4 - Master Encouraged, Prompt 5 - Master Discouraged

Task	**Prompt 1**	**Prompt 2**	**Prompt 3**	**Prompt 4**	**Prompt 5**
French to English (BERT Score)	0.87	0.90	0.76	0.76	0.75
Czech to English (BERT Score)	0.83	0.83	0.75	0.75	0.74
Summarization (BERT Score)	0.89	0.85	0.90	0.89	0.89
Classification (0/1 Accuracy)	0.92	0.71	0.78	0.67	0.71
Classification (N/A)	35 N/A	318 N/A	36 N/A	13 N/A	41 N/A

Table 1: Evaluation

1. Zero shot is better than few shot for all tasks except French to English
2. For classification, the result is very significant, with zero shot having better performance and less ambiguous results(N/A)
3. Master method and its variants are always less efficient than Simple Colon
4. Adding more words (positive/negative) to Master method does not help.
5. Comparison of BLEU scores with original paper denotes LLaMA-2 is not as efficient as GPT-3 for translation (as expected)

Below are the BLEU score results obtained for the French-to-English task - GPT-3 vs LLaMA-2

French to English	GPT-3	LLaMA-2
Simple Colon-0 Shot	33.3	20.51
Simple Colon-1 Shot	27.6	27.16
Master	32.9	9.51

Table 2: BLEU Score - GPT-3 vs LLaMA-2

To further test our results and based on Professor Dan's feedback, we randomly selected 50 small review samples from the IMDB Classification dataset that had ambiguous results for Simple Colon 1 Shot (N/A). Our task was to run a new type of prompt template - 3 Shot Simple Colon. Below are the results obtained:

Classification	N/A values	Accuracy(0/1 score)
Simple Colon-3 Shot	22	0.96

Table 3: Simple Colon-3 shot (Classification - 50 samples)

7 Challenges

Roadblock	Solution
13B model was heavy and computationally expensive	Using GPU-accelerated GGML model of LLaMA-2-13B on Google Colab with T4 GPU
Time constraints and compute unit constraints	Sampled test dataset each having 500 records for the 4 tasks. Testing took >335 hrs
Issues in classification task output	Cleaning manually to decide the meaning of output. Cases when it was ambiguous, introduction of N/A
Token limit in case of Text Generation Task	Filtered records that fit inside the token limit and then sampled

Table 4: Challenges and Solutions

8 Conclusion

The tasks outlined in the original proposal have been completed, including working with various tasks and prompting techniques. Additionally, few-shot learning techniques were employed with more examples to further improve the models' capabilities. After the successful completion of the project, we believe that Zero Shot is better than Few Shot in most cases. Few-shot(Especially 1-shot) leads to inconsistent results since it confuses smaller models like LLaMA-2. 3-shot has shown better performance than 1-shot by removing its ambiguities and even achieving better accuracy. Further, comparison of our results with that of Reynolds and McDonell (2021), we see that GPT-3 performs better than LLaMA-2 for the French to English translation task. We believe this is because of the lack of multilingual training data in LLaMA-2. The future scope of this project will be to compare multiple variants of the LLaMA-2 model (with the 7B and 70B versions) to determine differences in performance across numerous test dataset sizes, and few-shot configurations. We believe we explored the capabilities of LLaMA-2, understood and overcame the challenges in an LLM-based research project and learned a lot along the way. The inconsistent behaviors of the model while testing various prompts before finalizing the wordings especially surprised us. Overall, we believe that significant progress and learning have been made in line with the goals of the project.

References

Ondrej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Ale s Tamchyna. 2014. Findings of the 2014 Workshop on Statistical Machine Translation. , 12–58 pages. `http://www.aclweb.org/anthology/W/W14/W14-3302`

Ond rej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurelie Neveol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. 2016. Get To The Point: Summarization with Pointer-Generator Networks. , 131–198 pages. `http://www.aclweb.org/anthology/W/W16/W16-2301`

Code. 2023. LLM Group Project. `https://github.com/ssarp9/MS-CS59200-LLM-GroupProject`.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making Pre-trained Language Models Better Few-shot Learners.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. [n. d.]. What Makes Good In-Context Examples for GPT-3?

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning Word Vectors for Sentiment Analysis. , 142–150 pages. `http://www.aclweb.org/anthology/P11-1015`

Laria Reynolds and Kyle McDonell. 2021. Prompt Programming for Large Language Models: Beyond the Few-Shot Paradigm.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get To The Point: Summarization with Pointer-Generator Networks. , 1073–1083 pages. `https://doi.org/10.18653/v1/P17-1099`

Tony Z. Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate Before Use: Improving Few-Shot Performance of Language Models.