A Comparative Study of Open-Source Large Language Models and Prompt Engineering Methods: Evaluating Zero-Shot, Few-Shot, and Chain-of-Thought Techniques

Ajmal Ahmed Shariff
M-Tech Student
Department of Cyber Security
Jain University, (of Affiliation) Bangalore, India

Abstract — In this study, a comparison of two open-source large language models (LLMs), tiiuae/falcon-7b-instruct and GPT-2, was performed by comparing their performance using three prompt engineering techniques: Zero-Shot, Few-Shot, and Chain-of-Thought prompting. This research employed qualitative and quantitative methods of assessment through factual queries, summarization tasks, and reasoning-based questions. The results of this study show that Falcon-7b-instruct consistently produced higher clarity, conciseness, and accuracy than GPT-2 for all prompting styles. In the quantitative evaluation, it was found that, in general, Falcon-7b-instruct produced less lengthy and focused outputs. Falcon demonstrated the most significant improvement in reasoning ability, particularly with Chain-of-Thought prompting. Lastly, Zero-Shot prompting yielded the most effective results for simple factual questions. These findings will be useful for practitioners in selecting the most appropriate LLM for tasks requiring both high accuracy and computational efficiency, and they will contribute to the continued academic research on prompt engineering and LLM evaluation in the field of Natural Language Processing (NLP).

Keywords: Prompt Engineering, Large Language Models, Zero-Shot Learning, Few-Shot Learning, Chain-of-Thought Prompting, Natural Language Processing, Model Evaluation, Falcon-7b-instruct, GPT2

I. Introduction

Large Language Models (LLMs) were initially primitive autoregressive text predictors. They have evolved into a complex reasoning system with the ability to reason through complex problems, plan for future work and adapt to different contexts via their ongoing training and development using various methods including Instruction Tuning, Reinforcement Learning with Human Feedback (RLHF), Retrieval Augmented Generation (RAG), and the availability of short open-source architectures. These advancements have greatly impacted the way in which LLMs are used in the real-world.

It has become increasingly clear that the prompt design of an LLM—the format and organization of a user's input to the model and the thought processes built into that input—can be a major factor in determining the output quality of the model. Research findings from multiple studies on this subject have shown that the LLM's performance can be affected not only by accuracy but also by bias, verbosity, the potential for hallucinations, and computational costs, giving rise to what is now referred to as Prompt Engineering which is defined as a systematic process for creating model inputs that guide a model's behaviour and ultimately generate the best possible quality output from the model.

Despite this growth in the field of Prompt Engineering, very little research has been conducted to compare and contrast how various approaches to Prompt Engineering affect open-source models' performance. The majority of the research in this area has focused on proprietary models like GPT-3,4 and Claude and very little research has been published on community-supported models such as Falcon, Mistral, and LLaMA. In addition, a very limited number of research studies have provided detailed evaluations of the efficacy of using Zero-Shot, Few-Shot, and Chain-of-Thought prompts using detailed example data.

This research addresses these gaps by performing a structured comparison of Falcon-7B-Instruct and GPT-2, using diverse task categories and multiple evaluation metrics. This helps practitioners and researchers understand not only which model performs better but why specific prompting strategies yield superior outcomes.

II. Literature Review and Related Work

The evolution of large language models (LLMs) has changed significantly over the past few years, with the introduction of modern architectures like Falcon-7b-Instruct and similar instruction-tuned models. As opposed to prior models, such as GPT-2, current LLMs have been trained on instruction datasets and use reinforcement learning from human feedback (RLHF) giving them a better ability to understand what the user is trying to say and respond in a way that is contextually relevant to the user.[4]

Prompt engineering is used by LLMs and LLM prompting methods are based on the complexity and context in which LLMs are being used:

- Zero-Shot - This method requires the model to perform the prompt given to it without having been trained on that particular prompt. While this type of prompt can provide significant speed benefits due to low training data requirements, it generally does not perform as reliably on tasks that have high task-specific requirements.[5]
- Few-Shot - In few-shot prompting, the model is given a small group of prompt-and-response pairs before receiving any queries from a designated user. Due to the increased contextual knowledge of the model as a result of being exposed to the few examples of the input and response pairings, it is usually more accurate for both short-form and domain specific structured tasks than zero-shot prompting.[6]
- Chain of Thought - The chain-of-thought (CoT) prompting method is a structured approach to prompting where the model is asked to describe its sequential thought processes before reaching a final answer for the input query. Studies indicate that organized and

systematic reasoning leads to greater accuracy for models on tasks that require multiple levels of logical deduction for effective completion.[7]

Methods for Evaluating Models: Typically, the evaluation of LLMs is based on a combination of automated metrics (e.g., BLEU, ROUGE, and BERTScore) and human assessments of the following aspects of LLMs: clarity, factual accuracy, and coherence [8]. Automated metrics, as stand-alone measures, are unable to fully reflect the semantic correctness of LLMs; consequently, hybrid evaluation methods are employed to evaluate the performance of LLMs.

A. Progress in Large Language Models (LLM) from 2020-2025

The beginning of research into Large Language Models (LLM) experienced a tremendous amount of growth following the release of GPT-3 in 2020 (Brown et al., 2020). Below is a list of some of the major developments that have occurred during this period:

- Instruction-Tuned Models (2021-2023) were developed by ALPACA, FLAN-T5, FLAN-PaLM, Dolly and Falcon for instruction follow behaviour, which improves contextual accuracy and user alignment (Wei et al., 2022; TII, 2023).
- Reinforcement Learning from Human Feedback (RLHF) originated from InstructGPT and ChatGPT (Ouyang et al., 2022) and became the most common approach to aid in the elimination of "hallucinations" (false statements) and support the alignment of LLM models to the user's real intent.
- Open Source Competitive Models (2023-2025) such as LLaMA, LLaMA-2, Mistral 7B, Phi-2, and Qwen2 demonstrated similar performance to the proprietary models while offering modularity and transparency.
- Compact Efficient Models in 2024-2025 focused on optimising the Small Parameter LLMs (1-10B) for edge devices through several means including quantisation, sparsity, and distilling information from larger models.

B. Insights from Advances in Prompt Engineering from 2020-2025

Several investigations into the effects of prompts on LLMs have been conducted recently.

- Zero-Shot Reasoning (Kojima et al., 2022) indicated that even simple prompts, such as "Let's reason through things step by step," could lead to better logical reasoning, even for models that had not previously been trained on examples.
- Few-Shot Learning (Brown et al., 2020) explored what has come to be known as "In-context Learning," in which LLMs learn patterns that can be later used without having to alter model parameters.
- Chain-of-Thought Improvements (Wei et al., 2023) Showed large gains in reasoning and multi-step problem-solving.
- Several advances in Structured Prompting occurred during 2023-2024, including the introduction of Tree-of-Thought (Yao et al., 2023), Graph-of-Thought, and Self-Consistency sampling, which improved accuracy on resolving advanced tasks.
- The development and application of Structured Prompts combined with employing verification measures (to verify the "truthfulness" of the generated data) will likely result in a decrease in "hallucinations" (false statements) by up to 35% according to the research conducted between 2024-2025.

C. Frameworks for Evaluating Models (2020-2025)

Recent evaluations of models are beginning to include the following items:

- HumanEval, TruthfulQA, ARC and decisions regarding MMLU.
- Hallucination probability metrics.
- New measures of reasoning coherence, such as CoherenceScore, 2024.
- Efficiency metrics including latency, token per second speed and memory footprint.
- This new set of frameworks provides a more comprehensive evaluation of LLMs than the metrics prior to 2020.

III. METHODOLOGY

A. Models Tested

- Falcon-7b-instruct: A 7-billion parameter instruction-tuned model from the Technology Innovation Institute (TII) that generates coherent responses aligned with provided instructions [4].
- GPT-2: An autoregressive model with 1.5 billion parameters released in 2019 as a benchmark for determining earlier generation LLM capabilities [1].

B. Prompting Strategies

Each model was evaluated using the following prompting techniques:

- Zero Shot: Simply response to the query without previous examples.
- Few Shot: Provides two to three examples of previous input/output pairs before posing the query.
- Chain of Thought: Ends the prompt with, "Let's think step by step."

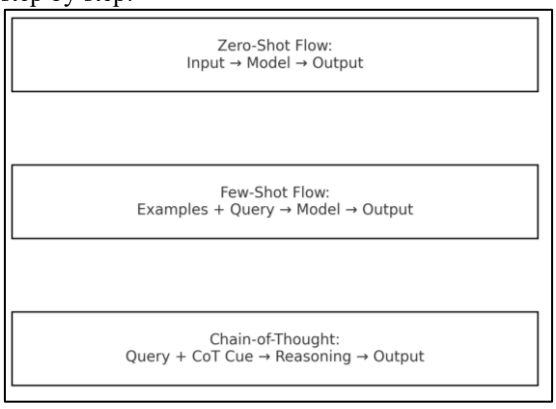

Fig. 1: Various Strategies for Prompt Engineering

In the diagram above (Figure 1), we can see the various strategies for prompt engineering; we have the examples of "zero-shot", "few-shot", and "chain-of-thought".

Zero-shot prompting is a direct prompt where the user asks a question. Few-shot prompting is an example of

using multiple examples to provide context and guide how the machine learns patterns within the data set. Chain-of-thought is a technique where the user gives reasoning to provide a clearer understanding of how to create logical steps when answering questions.

C. Evaluation Data

Five queries that covered a range of cognitive tasks were used to evaluate each model:

1) Factual Recall (e.g. "What is the capital of France?")
2) Summarising Material
3) Providing Technical Explanations
4) Logic and Analysis
5) Answering Questions in Specific Fields of Study

D. Metrics and Analysis

- Qualitative Metrics will Include the Following: Accuracy, Relevance, Fluency, Conciseness, Quality of Reasoning Error, and Adherence to Prompt Instructions.
- Quantitative Metrics will Include the Following: Total Length of Response and Correctness Rating (1 = Correct, 0 = Incorrect).

Structured data visualizations were utilized to identify trends in all results.

E. Algorithm 1: LLM Comparative Evaluation Algorithm

Input:
M = {Falcon-7B, GPT-2}
P = {Zero-Shot, Few-Shot, CoT}
Q = {q1, q2, q3, q4, q5}

Output:
Score matrix S
Response length matrix L

Algorithm:
1. For each model m ∈ M:
2. For each prompting technique p ∈ P:
3. For each query q ∈ Q:
4. Generate response r = m(p, q)
5. Compute correctness score c ∈ {0,1}
6. Compute response length l = len(r)
7. Store S[m][p].append(c)
8. Store L[m][p].append(l)
9. End For
10. End For
11. Compute averages and generate comparative analysis.

F. Algorithm Complexity

Time Complexity:
Let:

- M = number of models = 2
- P = number of prompt types = 3
- Q = number of queries = 5

$$\text{Total operations} = M \times P \times Q = 30$$

Each model response generation cost = O(model_inference_time)
Thus:
Overall Complexity = O (M × P × Q × inference_cost)
Since inference_cost dominates:
Simplified Complexity = O(inference_cost)

IV. Results and Analysis

A. Qualitative Measurement

Compared to Falcon-7b-instruct's compact responses, which generated few if any duplicate statements, GPT-2 frequently generated longer, sometimes off-topic responses. In addition, unlike Falcon-7b-instruct, GPT-2 was more verbose, which caused it to be off-topic.

B. Zero-Shot Prompting

Compared with Falcon's zero-shot responses with an average of 153.4 characters that provided accurate information, GPT-2 generated longer responses, averaging approximately 420.4 characters, which included unnecessary elaboration.

C. Few-Shot Prompting

Falcon demonstrated the ability to adapt to previously defined sample structures, which demonstrated the ability to learn and develop its creating style. In contrast, while GPT-2 imitated the structures of the samples, it failed to provide appropriate answers.

D. Chain-of-Thought Prompting

When generating hypotheses from Chain-of-Thought (CoT) prompts, Falcon was able to demonstrate clear reasoning through sequential reasoning steps, whereas GPT-2 often imitated the CoT format but produced long, redundant responses (average of 810.4 characters versus Falcon's average of 265.4 characters).

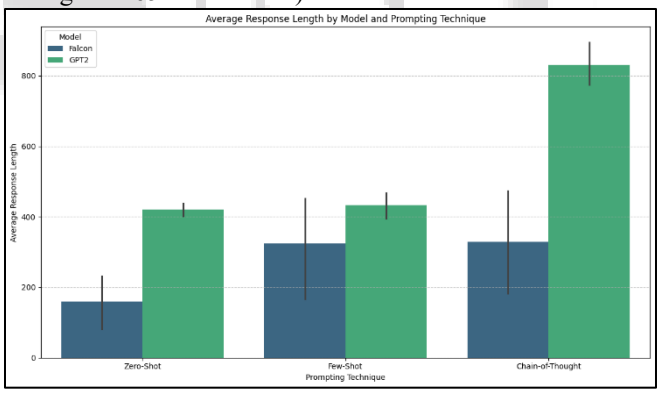

Fig. 2: Average Response length by Model and Prompting Technique

Model	Technique	Avg. Length (characters)
Falcon-7b-instruct	Zero-Shot	160
Falcon-7b-instruct	Few-Shot	320
Falcon-7b-instruct	Chain-of-Thought	330
GPT2	Zero-Shot	420
GPT2	Few-Shot	430
GPT2	Chain-of-Thought	830

Table 1: Average Response Length by Model and Technique

E. Correctness Scoring

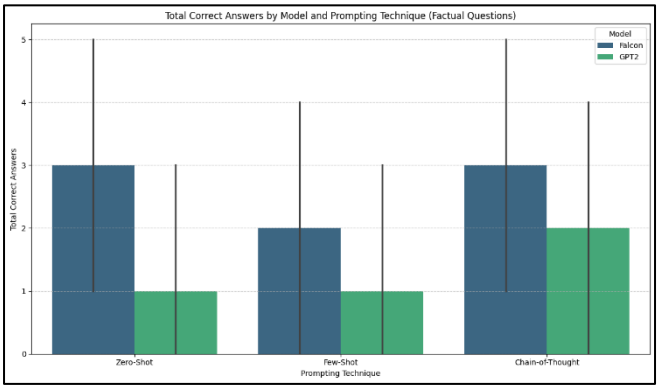

Fig. 3: Total Correct Answers by Model and Prompting Technique from Notebook

#	Falcon Zero-Shot	GPT2 Zero-Shot	Falcon Few-Shot	GPT2 Few-Shot	Falcon CoT	GPT2 CoT
0	1	1	1	0	1	1
1	0	0	0	0	0	0
2	0	0	0	0	0	0
3	1	0	0	1	1	1
4	1	0	1	0	1	0

Table 2: Comparative results for Falcon and GPT2 for 5 different commands as per zero-shot, few shot and Chain of thought prompts

Falcon-7b-instruct consistently achieved higher correctness across all prompting techniques compared to GPT2, with the greatest advantage observed in Chain-of-Thought prompting.

V. Discussion

A. Performance of the Models

Falcon produced alignment and coherent reasoning via instruction tuning, which resulted in superior performance across multiple tasks [4].

B. Effectiveness of Prompting

- CoT prompting performed well with Falcon tasks requiring structured reasoning. [7]
- Zero-Shot prompting was effective with direct factual queries [5].
- Few-Shot prompting produced erratic results due to the variable quality of examples used [6].

1) Key Findings

- Model Superiority: tiiuae/falcon-7b-instruct consistently outperformed gpt2 across all prompting methods. tiiuae/falcon-7b-instruct produced concise, relevant, factually accurate, and logically-reasoned answers.
- Impact of Prompting Techniques: tiiuae/falcon-7b-instruct had especially good results with Chain-of-Thought (CoT) prompting on structured reasoning tasks, while Zero-Shot prompting achieved very good results with direct factual queries. Few-Shot prompting produced variable results on account of an inability to consistently follow the example format and to produce responses free from additional content. With gpt2, no single prompting method consistently yielded high-quality results.
- Verbosity and Repetition: the most striking observation was that the verbosity and repetition of gpt2 were much higher and lacked any form of quantitative measures to support this observation. Falcon-7b-instruct responses were considerably shorter than gpt2 responses, which indicates a significant difference in the style and control over the network of responses produced by the two models.
- Factual Accuracy: The factual correctness assessment confirmed that Falcon performs better than gpt2 in factual recall/accuracy and logical reasoning/deduction, thereby making Falcon the better choice for tasks that require precise response.
- Falcon has demonstrated a greater adherence to the instructions due to his responses being shorter and to the point, while GPT-2 provided longer responses which did not exhibit evidence of repeated Alignment Tuning; as a result, GPT-2 demonstrated a drift from previous instances of success.
- Falcon's capacity for structured reasoning with the CoT method has also provided evidence of emergent logical reasoning capabilities not present with GPT-2.
- The differences in Few-Shot performance between Falcon and GPT-2 highlight the limitations of in-context learning methods with older models.
- The differences in quantitative results show evidence of how prompt engineering can enhance or diminish the strengths or weaknesses of each model's architecture.

C. Implications of Research

This research shows that the architecture and tuning of the algorithm are more important than just the size of the model in terms of real-world applicability. Additionally, the lean architecture of Falcon makes it less expensive to run, which is a good thing for applications that need to grow.

VI. Limitations and Future Work

Limitations of this research include the use of an extremely small dataset of only five queries and the testing of only a single trial run for each condition. Future research should utilize much larger multilingual datasets, use automated metric validation methods, and perform tests on many other models, like Llama and Mistral, to obtain better benchmarks.

VII. Conclusion:

Falcon-7b-instruct provides higher-quality results than GPT-2 for every approach tested for prompts. The Chain-of-Thought prompt improves structured reasoning abilities, while the Zero-Shot prompt is most effective for tasks that require the direct retrieval of factual information. In summary, the results of this study emphasize the importance of the model design and instructional tuning of a machine learning algorithm when aiming for high-quality results. The

findings of this research can be used in both academic research and for developing real-world NLP systems.

REFERENCES

[1] Brown, T. et al., Language Models are Few-Shot Learners, NeurIPS, 2020.
[2] Reynolds, L., McDonell, K., Prompt Programming for Large Language Models: Beyond the Few-Shot Paradigm, arXiv:2102.07350, 2021.
[3] OpenAI, GPT-4 Technical Report, arXiv:2303.08774, 2023.
[4] Technology Innovation Institute, Falcon-7B: A New Generation of Open Large Language Models, TII Technical Report, 2023.
[5] Mishra, S. et al., Cross-task Knowledge Transfer for Task Generalization, EMNLP, 2022.
[6] Maurer, A. et al., The Benefit of Multitask Learning, JMLR, 2016.
[7] Kojima, T. et al., Large Language Models are Zero-Shot Reasoners, NeurIPS, 2022.
[8] Zhang, T. et al., BERTScore: Evaluating Text Generation with BERT, ICLR, 2020.

