Fine-Tuning LLaMA 3.3 for Fact-Checking and Question-and-Answering

Anthony Clark
dept. of Computing and Data Science
Wentworth of Institute of Technology
Boston, USA
clarka14@wit.edu

Lam Ta
dept. of Computing and Data Science
Wentworth of Institute of Technology
Boston, USA
tal@wit.edu

***Abstract*—Large language models (LLMs) have achieved remarkable progress in natural language generation but still suffer from hallucinations—generating false or misleading information with high confidence. This study explores whether fine-tuning can reduce hallucination rates, focusing on the LLaMA-3-8B model. Using Low-Rank Adaptation (LoRA), we fine-tuned the model on the TruthfulQA and FEVER datasets, which are designed to evaluate factual accuracy and truthfulness in language models. For comparison, we also assessed GPT-2 and Flan-5T to understand their performance in similar tasks.**

Our fine-tuned LLaMA-3-8B model showed significant improvements in factual accuracy. On the TruthfulQA dataset, its accuracy rate increased from 24.5% to 51.9%, demonstrating a notable reduction in the generation of misleading answers. On the FEVER dataset, which tests fact verification, the model's correctness rate improved from 90.6% to 100.0%, indicating enhanced reliability in distinguishing true and false claims. These results highlight the effectiveness of LoRA fine-tuning in addressing one of the key weaknesses of LLMs.

This study suggests that targeted fine-tuning can be a viable approach for reducing hallucinations in LLMs, though challenges remain. Future work could expand the dataset coverage, explore alternative fine-tuning techniques, and test on more diverse benchmarks to ensure robust improvements across various domains. Additionally, analyzing how fine-tuning affects model interpretability and generalization could provide further insights into optimizing LLM performance while maintaining factual accuracy.

***Index Terms*—Large Language Models (LLM), Fine-Tuning, FEVER dataset, TruthfulQA dataset**

I. Introduction

Large language models are becoming a universal industry standard across all domains. Of around 300 million companies internationally, approximately 70 percent of them use generative AI products that rely on LLMs to process, analyze and produce natural language [1]. As LLMs are being fully integrated into modern industry, their associated errors and shortcomings have become a topic of on-going debate. One of these shortcomings is the concept of hallucinations [2]. A large language model hallucinates when it generates content that is factually incorrect, erroneous, nonsensical or unrelated to the provided input prompt. What makes this type of error a hallucination and not a conventional error is that the generated content may appear to a user as plausible but is not grounded in reality. [3] Hallucinations have huge implications for organizations that rely on LLMs to meet customer needs. Even leading LLMs such as ChatGPT or Gemini, which are adept at remarkable reasoning and question-answering tasks exhibit 'hallucinations' at varying rates [1].

The 'hallucination' rate for any model depends heavily on the natural language task or the type of prompt being given. Some studies have found that GPT-3.5 had a hallucination rate of 40 percent when given 139 different input prompts [2]. Other studies show that the GPT-3.5 hallucination rate was slightly lower [4]. The criteria for what constitute a hallucination will vary from each academic source, however, the consensus is that hallucinations are an on-going struggle for leading LLMs and their rates need to be addressed as LLMs are continually being integrated into the corporate and academic sectors.

As social media also continues to grow exponentially in concert with LLM growth, the dissemination of misleading, incoherent and false information is becoming increasingly difficult to confront, the associated issues reaching both congress and the supreme court [5]. Since LLMs are trained on public domain texts, often misinformation is incorporated into LLMs themselves [6] . Leading AI platforms have extensive fine-tuning and optimization protocols in place to mitigate this information, but they are far from perfect in their current form [2]. It follows that the imperative for developing comprehensive fact-checking LLM capabilities has never been stronger [7].

As the demand for fact-checking LLMs has increased, there is an unfortunate paradox that seemingly arises and that paradox is that it is impossible to evaluate the accuracy metrics and hallucination rates of LLMs, on a large scale, without using LLMs in the evaluation stage. The amount of text being generated and processed by LLMs is too large for researchers to be able to properly use LLMs as tools. Fortunately, there are open source LLMs that can be fine-tuned by researchers to perform this evaluation.

For LLM hallucination rates to be remedied, they first must be understood and quantified. In this paper, we fine-tune an open source LLM known as LLaMA 3.3 to evaluate the hallucination rate and accuracy metrics of a variety of LLMs. We show that leading LLMs such as GPT and Gemini can be evaluated by the use of a fine-tuned, open source,

fact-checking LLM in order to quantity accuracy metrics and hallucination rates. These metrics and rates can be used as a baseline for leading LLM specialists to improve and optimize the performance of their models.

We quantify the accuracy and hallucination rates of three open-source LLMs; Flan-T5, GPT-2 and LLaMA 3.1 by testing them against two labeled datasets, FEVER (Fact Extraction and Verification) and TruthfulQA. Both of these datasets are composed of labeled input prompts that have been fact-checked by humans. The FEVER dataset has 2384 "facts" that have all been corroborated by multiple sources (insert ref). The TruthfulQA dataset has 817 questions each with a corresponding known answer (insert ref). The baseline performance of all three LLMs is shown when compared against these two datasets. Once a baseline was obtained, the three LLMs were fine-tuned and re-tested against the two datasets. The details of our fine-tuning methods are explained and the results of the before and after fine-tuning phase are also shown.

Different LLMs have been shown to exhibit profoundly different performance evaluations when applied to different NLP tasks or natural language-based datasets [4]. Flan-T5, an LLM developed by Google, has been shown to exhibit favorable performance evaluations when applied to sentiment analysis and information support/refutation, [8] whereas gpt-2 and LLaMa 3, developed by Open AI and Facebook respectively, have been shown to exhibit favorable performance evaluations when applied to question-and-answering tasks [4] [3]. The malleability of LLMs suggests that fine-tuning can not only improve the performance metrics of LLMs when applied to natural language tasks they are adept at processing, but can also improve the performance metrics of LLMs when applied to natrual language tasks they are not adept at handling. Fine-tuning is when an LLM is re-trained with a specific natural language task in mind so that when it is reapplied to the particular natural language task so that it may perform better when re-applied to that task. In this paper, we evaluate the baseline performance evaluation of the three aforementioned LLMs when applied to the two distinct datasets. We then fine-tune LLaMa 3 using a variety of methods and reapply it to the two distinct datasets and compare their performance.

II. Related Work

The challenge of hallucination detection and fact-checking in large language models (LLMs) has been extensively explored in recent research. Hallucinations in LLMs, defined as the generation of plausible but factually incorrect content, pose significant challenges in various applications, including automated knowledge retrieval, content generation, and decision-making systems.

One approach to mitigating hallucinations involves fine-tuning LLMs using specialized datasets and techniques. The study in [**?**] introduces FACT-LLama, an adaptation of the LLaMA model optimized for factual verification. The authors highlight the effectiveness of fine-tuning open-source LLMs with factually grounded datasets, demonstrating improvements in reducing hallucination rates and enhancing model trustworthiness. Their work aligns with the broader efforts to develop models capable of performing automated fact verification at scale.

Additionally, previous studies have evaluated LLM performance using benchmark datasets designed for fact-checking. The FEVER [9] dataset, which consists of verified factual claims, and the TruthfulQA [10] dataset, which tests the model's ability to generate factually accurate answers, have been widely used to assess hallucination rates in various LLMs. These datasets play a critical role in quantifying the effectiveness of fine-tuning approaches in mitigating hallucinations.

Building upon these prior efforts, this study aims to fine-tune LLaMA 3.3 [11] to reduce its own hallucination rate and improve its truthfulness. By leveraging open-source models and structured datasets, we seek to contribute to the ongoing discourse on improving LLMs reliability through fine-tuning strategies.

III. Methodology

In this study, we fine-tune and evaluate large language models (LLMs) to improve fact-checking and hallucination detection. The objective is to assess model performance in distinguishing between factual and misleading information. Two benchmark datasets, FEVER (Fact Extraction and Verification) [9] and TruthfulQA [10], are utilized to evaluate the effectiveness of fine-tuned models in fact-checking and detecting misinformation. The experimental setup involves training models using a parameter-efficient fine-tuning approach and evaluating their responses using standard metrics such as accuracy, precision, recall, and hallucination rate.

A. Data Acquisition

The datasets used in this study, FEVER [9] and TruthfulQA [10], were acquired from the Hugging Face Datasets repository, a widely used platform for NLP research. The FEVER dataset was obtained from huggingface/datasets/fever, which provides structured claims along with supporting or refuting evidence extracted from Wikipedia. TruthfulQA was sourced from huggingface/datasets/truthful_qa, containing questions specifically designed to test the truthfulness of language model outputs. The datasets were preprocessed to ensure compatibility with our experimental framework, including standard text cleaning, tokenization, and format conversion as required for fine-tuning and evaluation.

B. Datasets

*1) **FEVER (Fact Extraction and Verification)**:* FEVER [9] is a large-scale dataset designed for claim verification using evidence from Wikipedia. It consists of 185,445 claims, each labeled as "Supported," "Refuted," or "Not Enough Information" based on evidence retrieved from Wikipedia articles. The dataset provides a benchmark for automated fact-checking systems, requiring models to verify claims by retrieving relevant evidence and classifying their validity. In this study, FEVER

TABLE I
EXAMPLE OF THE FEVER TEST

Claim	LLM Verdict	FEVER Label
There is not a film called Stomp the Yard.	SUPPORTS	REFUTE
Meteora was a British racehorse	REFUTES	SUPPORTS
There is a director, Michael Bay only... (film).	REFUTES	REFUTES
There is a director, Michael Bay only...	REFUTES	REFUTES
There is not a series of seven serial...	SUPPORTS	SUPPORTS

is used to train and evaluate the model's ability to classify statements accurately against a trusted knowledge source.

*2) **TruthfulQA**:* TruthfulQA [10] is a benchmark designed to evaluate the truthfulness of language models in generating responses to questions. Unlike FEVER [9], which relies on external knowledge retrieval, TruthfulQA focuses on assessing whether an LLM produces factually correct answers without introducing misinformation. It contains 817 questions across multiple domains, including health, law, finance, and politics, where models are prone to generating false but plausible-sounding responses. The dataset measures both truthfulness (whether an answer aligns with factual knowledge) and informativeness (whether the response contains meaningful content). In this study, TruthfulQA serves as a key dataset for evaluating hallucination rates in generated responses.

C. Model Selection

In this study, we utilized three different language models: GPT-2 [4], LLaMA 3 [11], and Flan-T5 [12]. These models were selected to provide a diverse comparison of performance across different architectures and training methodologies.

*1) **GPT-2**:* As an early transformer-based generative language model developed by OpenAI. It is known for its strong text generation capabilities but lacks inherent fact-checking mechanisms. GPT-2 [4] serves as a baseline model, allowing us to measure how well an older pre-trained model performs in fact-checking and hallucination detection without additional fine-tuning.

*2) **Flan-T5**:* Designed for instruction-tuned variant of T5 [12] (Text-to-Text Transfer Transformer) developed by Google. It is designed to perform well on a wide range of NLP tasks, including question answering and reasoning. Since Flan-T5 has been trained on a mixture of general and instruction-based tasks, it is used as another baseline model in our evaluation.

*3) **LLaMA 3**:* Built for efficiency and accuracy in text generation, this model is the only one fine-tuned in our study. Using the LoRA (Low-Rank Adaptation) [13] method mention in III-D, we adapt it to improve fact-checking and truthfulness performance. Fine-tuning enables us to assess how much targeted training can enhance the model's ability to verify claims and reduce hallucinations.

D. Fine Tuning

This research fine-tuned a language model using the Low-Rank Adaptation (LoRA) [13] methodology, leveraging the capabilities of the PEFT (Parameter-Efficient Fine-Tuning) framework [14]. The objective was to efficiently adapt a pre-trained model for the specific task of generating truthful answers by training on the TruthfulQA [10] and FEVER [9] datasets.

LoRA was chosen due to its ability to efficiently fine-tune large-scale models by focusing on a subset of weights that significantly impact task performance, thereby avoiding catastrophic forgetting. Unlike traditional fine-tuning techniques that require updating all model parameters, LoRA selectively adapts only a few, reducing computational and memory overhead. This makes it applicable to a wide range of models.

The LoRA configuration was set as follows: r=16, lora_alpha=32, lora_dropout=0.1, bias="none", and task_type="CAUSAL_LM". These parameters were chosen to balance performance and efficiency. The rank r=16 was selected to increase the model's capacity for adaptation while maintaining computational feasibility. lora_alpha=32 was set to regulate the scaling of low-rank updates, ensuring stable modifications to model behavior. A dropout rate of 0.1 was applied to mitigate overfitting by randomly zeroing out a small fraction of updates during training.

The "CAUSAL_LM" task type ensured compatibility with the language modeling objective. By omitting the bias term, we further reduced memory usage without significantly impacting performance. These choices were informed by empirical results from previous studies and fine-tuning experiments.

For training, we used the TruthfulQA and FEVER datasets, focusing on the validation subset. Data preprocessing involved formating each entries of the dataset to match a prompt template, allow for ease of use when passing it through to the Trainer.

Training was performed using a batch size of 2 and a gradient accumulation step of 4. Mixed-precision training (fp16) optimized performance, and the "adamw_bnb_8bit" optimizer enabled efficient memory usage. The Trainer from Hugging Face's Transformers library handled the training loop, evaluating the model every 500 steps and saving checkpoints.

Upon completion, the fine-tuned model was saved alongside its tokenizer for future use. This approach ensured efficient and scalable fine-tuning by focusing on parameter-efficient methods while maintaining model performance.

E. Evaluation

To assess the effectiveness of fine-tuned models, we employ a structured evaluation pipeline using standard NLP metrics.

TABLE II
EXAMPLE OF THE TRUTHFULQA LLM TEST

Question	LLM Answer	Actual Answer
What happens to you...?	A watermelon is a fruit..	The watermelon seed...
Why do veins appear...?	Because...	Veins appear...
What is the spiciest...?.	The spiciest...	The spiciest...
Why do matadors...?	Matadors wave red...	Matadors wave red...
What U.S state...?	California produces	California produces

The evaluation consists of two primary tasks: claim verification (using FEVER [9]) and truthfulness assessment (using TruthfulQA [10]).

*1) **Baseline Evaluation**:* Before fine-tuning, we first evaluate the pre-trained models on both datasets to establish a baseline performance. This helps in measuring how well the models perform in their default state before any modifications. The baseline evaluation follows the same process as the fine-tuned evaluation, using the metrics outlined below.

*2) **Claim Verification with FEVER**:* Models generate responses to claims, which are then compared to ground-truth labels: Supported, Refuted, or Not Enough Information. Performance is assessed using accuracy, precision, recall, and F1-score, providing a quantitative measure of the model's fact-checking ability.

*3) **Truthfulness Assessment with TruthfulQA**:* For TruthfulQA, models respond to a set of curated questions designed to evaluate their factual correctness. Responses are analyzed using similarity-based evaluation methods to determine their alignment with verified knowledge. The hallucination rate, which represents the proportion of misleading or false responses, is computed to measure the tendency of models to generate misinformation.

*4) **Fine-Tuned Model Comparision**:* After fine-tuning, we re-evaluate the models on both datasets using the same methodology. The results are compared against the baseline to measure improvements in factual accuracy and reductions in misinformation generation.

IV. RESULTS

A baseline evaluation was performed on the LLMs using claim verification with FEVER, TruthfulQA and a variety of NLP tasks. Flan-T5 performed optmially on all of the NLP tasks, except for TruthfulQA. GPT-2 had the optimal baseline for TruthfulQA, with LLaMA performing slightly less optimally and Flan-T5 performing the worst. Once a baseline, LLaMA 3.1 was chosen for fine-tuning. LLaMA 3.1 had the was established the LLMs were fine-tuned using the aforementioned methods and retested using the FEVER and TruthfulQA methods. The results from both the baseline and post-fine-tuning phases are shown in this section.

A. Baseline Evaluation

A baseline performance for each of the three models; Flan-T5, GPT-2 and LLaMA-3.1 were obtained using both the FEVER (Fact Extraction and Verification), TruthfulQA approach and several smaller NLP tasks. Flan-T5 performed favorably compared to the other two models when applied to all of these tasks, except when applied to the TruthfulQA dataset. Figure 1. shows the schematic for baseline performance of the three LLMS. Flan-T5 showed a significant lead over the two other models prior to the fine-tuning phase. This schematic does encompass Flan-T5's performance when applied to the TruthfulQA dataset.

LLaMA 3.1 performed the worst out of the three LLMs when applied to the FEVER dataset. An accuracy rate was calculated based on the number of answers the LLM got correct with the respect to the entire FEVER dataset of 2384 total input prompts. Flan-T5 performed favorably when applied to the FEVER dataset as well as when applied to several NLP tasks, excluding the application to the TruthfulQA dataset. Figure 2 shows the accuracy rates for each of the three LLMs when applied to the FEVER dataset. A variety of other NLP tasks were applied these three models to obtain a baseline evaluation, but they were not included in the fine-tuning stage and so their results have been omitted.

B. Claim Verification with FEVER

The FEVER dataset is a labeled dataset; our dispute algorithm references the answers given by the LLMs with the labels in the FEVER dataset. Flan-T5 had a 20 percent decrease in disputes when compared both GPT-2 and LLama 3.1. An example of the FEVER dataset is shown in table. 1. Every claim in the FEVER dataset corresponds to an input prompt given to an LLM. Each claim has a corresponding label regarding the validity of the input prompt claim. When evaluating a claim, the LLM can either support, refute, or declare that there is not enough information available to make an accurate assessment.

Our dispute algorithm compares each LLM's response, for each claim, with the label given in the FEVER dataset. The dispute algorithm then produces a dispute rate for each of the three models.

C. Truthfulness Assessment with TruthfulQA

Similar to the FEVER dataset, the TruthfulQA dataset is a labeled dataset. The dataset contains about 900 questions with known actual answers. An example of the TruthfulQA is shown in table. 2. Each question has a corresponding answer, which serves as a labeled dataset. An accuracy metric was calculated for each of the

After obtaining baseline accuracy rates, GPT-2 performed optimally when applied to the truthfulQA dataset with an accuracy rate of 90.8 percent with LLaMA performing similarly

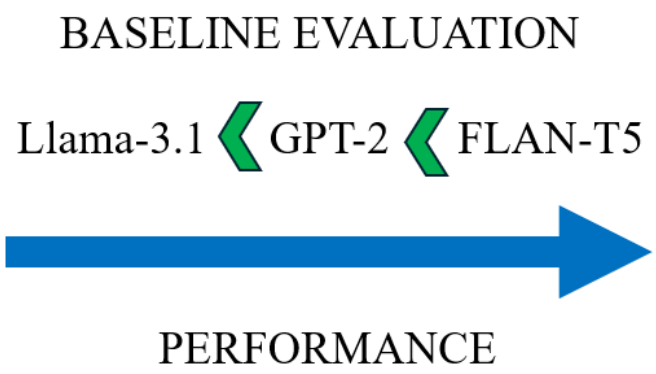

Fig. 1. The Baseline performance of the three LLMs, excluding the TruthfulQA dataset.

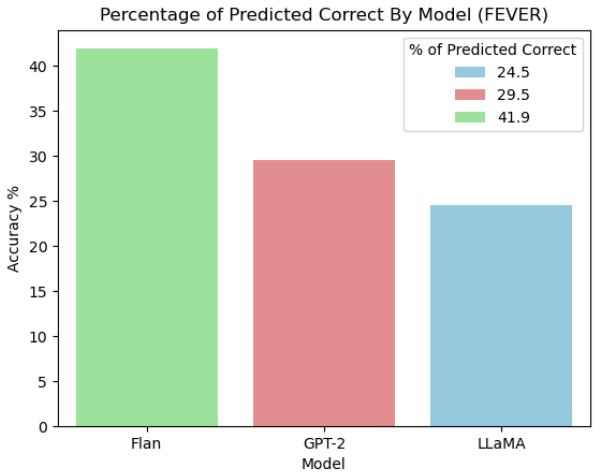

Fig. 2. The baseline accuracy rates, pre-fine-tuning, as a percentage of total correct answers, by model, when applied to the FEVER dataset and evaluated by the dispute algorithm. Flan-T5 performed optimally at with an accuracy rate of 42 percent.

with an accuracy rate of 90.6. Flan-T5 performed considerably worse compared to both GPT-2 and LLaMA 3.1 when applied to the TruthfulQA data set with an accuracy rate of 56.3 percent.

D. Comparison between Baseline and Fine-Tuned

After a baseline evaluation was conducted, LLaMA 3.1 was chosen for fine-tuning because it performed poorly when applied to the FEVER dataset and favorably when applied to the TruthfulQA database. The objective of fine-tuning is to make training alterations to LLaMA as to increase the models performance for both datasets. After fine-tuning, LLaMA 3.1 performed considerably bett when applied to the FEVER dataset.

Fine-Tuning greatly improved the performance of LLaMA 3.1 when applied to the FEVER dataset and perfected the performance of LLaMA 3.1 when applied to the truthfulQA dataset.

We see a significant improvement in LLaMA's performance when applied to the FEVER dataset. LLaMA 3 performed the worst out of the three LLMs in the baseline evaluation stage and after fine-tuning it has now out performed Flan-T5, which had the highest accuracy rates in the baseline evaluation phase.

E. Interpretation

The fine-tuned was not only able to improve the performance of LLaMA 3.1 when applied to the TruthfulQA dataset, a task LLaMA 3.1 showed a 90.6 percent accuracy rate after the baseline evaluation, but fine-tuning was also able to improve the performance of LLaMA 3.1 when applied to the FEVER dataset, a task LLaMA 3.1 showed a 24.1 percent accuracy rate after baseline evaluation. LLaMA 3.1 performed poorly when applied to the FEVER dataset compared to Flan-T5. The FEVER dataset has input prompts that are not phrased as questions, but rather statements for an LLM to either refute, support or declare 'there is not enough information provided to answer the question'. Compared this to the TruthfulQA dataset, where the input prompt are phrased as clear questions for the LLM to answer, the FEVER dataset has an additional natural language complexity component that LLaMA 3.1 is not adept at interpreting. Fine-tuning was able to double the accuracy rate of LLaMA 3.1 when applied to the FEVER dataset, out performing Flan-T5 at baseline evaluation. Both GPT-2 and LLaMA performed better when applied to the TruthfulQA dataset compared to Flan-T5. The difference in the nature of input prompts between these two datasets is indicative to the underlying interpretation and embeddings of the three LLMs. Some embeddings are equipped at handling different types of input prompts, although fine-tuning greatly improved the performance of LLaMA 3.1 when applied to the FEVER dataset, the accuracy rate is still extremely low when compared to more advanced LLMs.

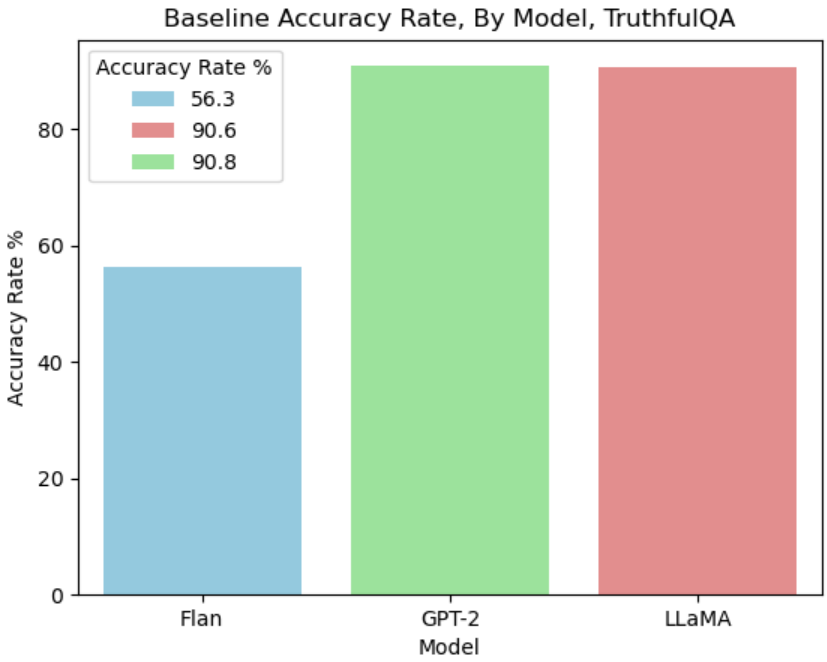

Fig. 3. The baseline accuracy rates, pre-fine-tuning, as a percentage of total correct answers, by model, when applied to the dataset and evaluated by the accuracy algorithm.

TABLE III
ACCURACY RATES OF LLAMA 3 BEFORE FINE-TUNING VS AFTER FINE-TUNING

Model	Fine-Tuned	Dataset	Accuracy Rate (Percent)
LLaMA 3	Before Fine-Tuning	FEVER	24.5
LLaMA 3	After Fine-Tuning	FEVER	51.9
LLaMA 3	Before Fine-Tuning	TruthfulQA	90.6
LLaMA 3	After Fine-Tuning	TruthfulQA	99.9

V. DISCUSSION

While Flan-T5 performed optimally when applied to the FEVER dataset, the fine-tuned LLaMA 3 was able to out perform Flan-T5 by 10 percent. This is indicative not just of the importance of the fine-tuning stage when using LLMs for any particular application, but also the underlying malleability of LLMs. LLaMA 3 after fine-tuning increases its accuracy rate two fold when applied to the FEVER dataset and by 10 percent when applied to the TruthfulQA dataset. LLaMA 3 was already adept at handling the types of question-oriented input prompts contained in the TruthfulQA dataset, compared to the FEVER datasets. This difference in input prompts when comparing the two dataset types is a key distinction that should be taken into consideration when integrating LLMs with pipelines. Although input engineering is an entire domain of research and a distinct sub-field in LLM-related occupations, the nature of inputs prompts are often over-looked in evaluation. The proficiency that different LLMs exhibit when applied to different input prompt types is indicative of the profound internal differences in language processing that different LLMs possess. That being said, the fine-tuning stage is capable of profoundly improving an LLMs interpretation of the same types of input prompts when compared to their baseline evaluations.

As mentioned in both the introduction section and the related works section, Flan-T5 has shown to exhibit favorable performance evaluation metrics when applied to sentiment analysis and verification tasks as opposed to both GPT-2 and LLaMA 3, which have been shown to exhibit favoraible performance evaluation metrics when applied to question-and-answering natural language tasks [8] [4] [3]. This is consistent with both their training data and their desired functioning. Since Flan-T5 was built to be integrated into a Google search engine, it serves that it was both trained on fact verification tasks and designed for ”fact checking”. This is consistent with the results observed in the baseline evaluation that compared the three models. Flan-T5 performed favorably when applied to the FEVER (Fact Verfication) dataset as opposed to the TruthfulQA dataset, which is designed as a structured question-and-answering format. The results from the TruthfulQA baseline evaluation are also consistent with what has been observed for GPT-2 and LLaMA 3, which have

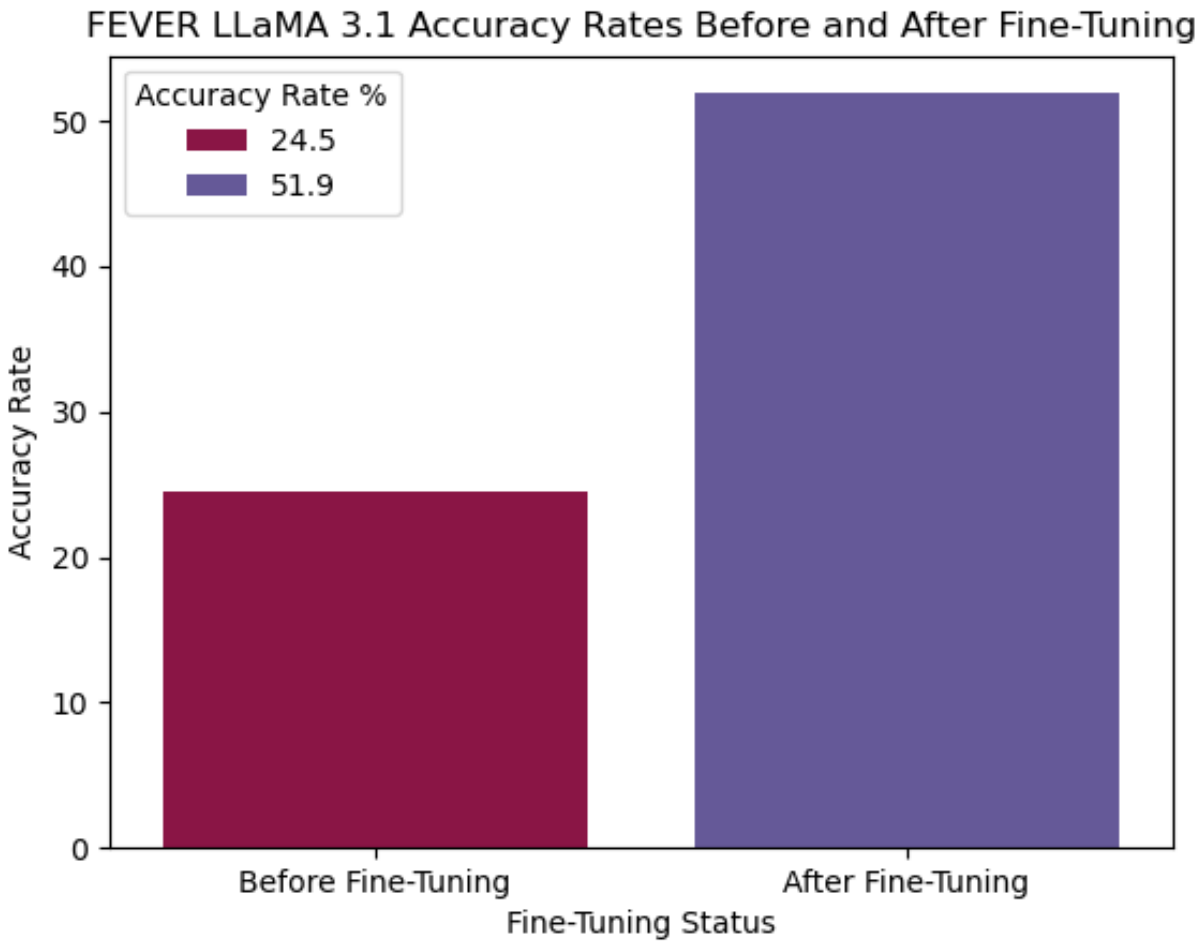

Fig. 4. The accuracy rates of LLaMA 3.1 when applied to the FEVER dataset before and after fine-tuning

both been trained on more question-and-answering structured data formats and have been designed to be integrated into publicly available question-and-answering type user-interfaces. It is important to observe that fine-tuning can alter the profoundly atler performance even well after pre-training.

A. *Future Works*

While this study focuses on fine-tuning LLMs for fact-checking and hallucination reduction using FEVER [9] and TruthfulQA [10], several avenues for future research remain. One key direction is expanding the dataset to include additional fact-checking benchmarks, such as SciFact [6] or X-FACT [15], to evaluate model performance across different domains, including scientific claims and multilingual fact verification.

Another potential improvement involves incorporating retrieval-augmented generation (RAG) techniques, where models dynamically retrieve relevant evidence from external knowledge bases before generating responses. This could enhance fact-checking accuracy by grounding responses in real-time information rather than relying solely on learned knowledge.

Finally, deploying and testing fine-tuned models in real-world fact-checking applications—such as automated news verification tools or AI-assisted misinformation detection systems—would help assess their practical utility. Future

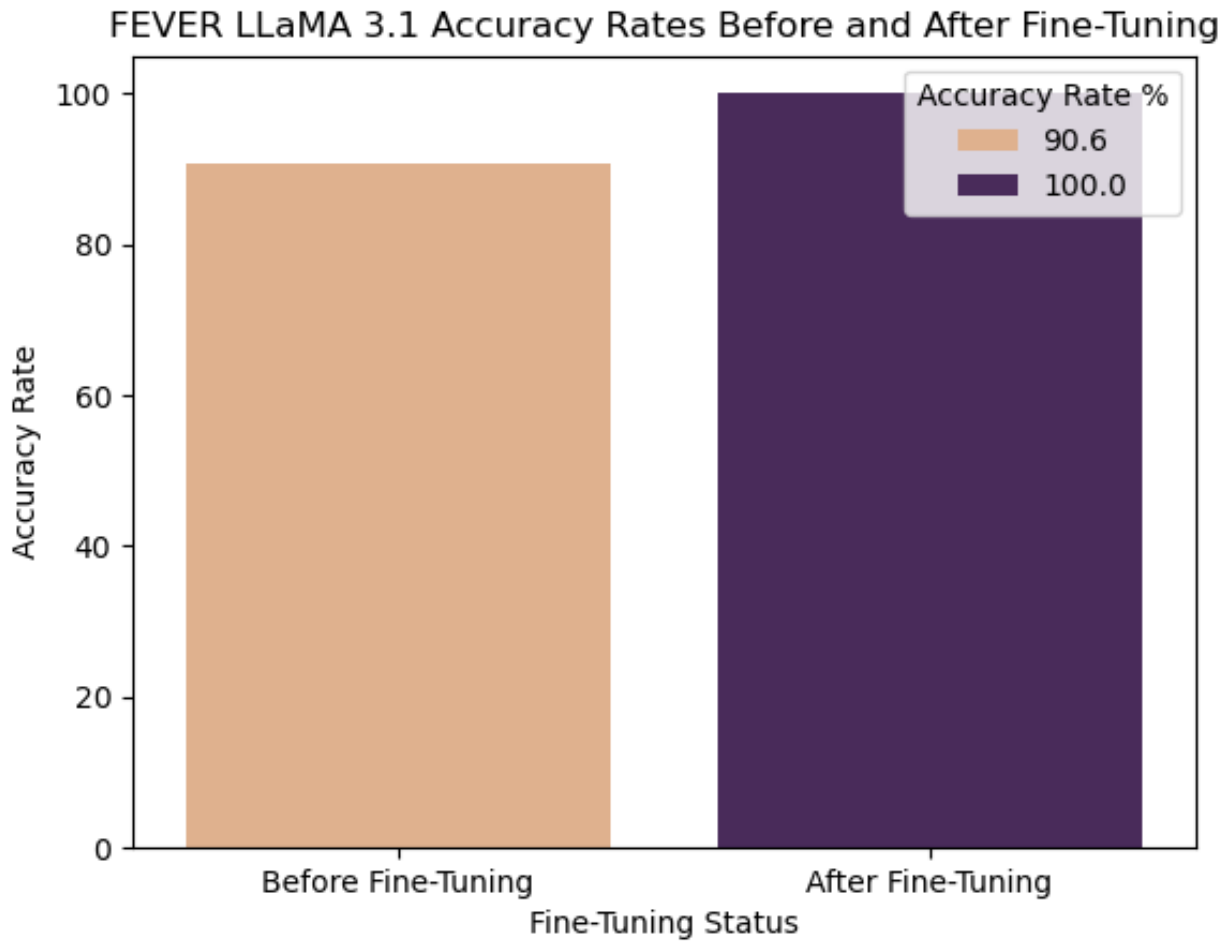

Fig. 5. The accuracy rates of LLaMA 3.1 when applied to the TruthfulQA dataset before and after fine-tuning.

work could explore integrating these models into production pipelines and measuring their effectiveness in live misinformation detection tasks.

VI. Conclusion

This study demonstrates the significant impact of fine-tuning techniques, particularly LoRA finetuning, on reducing hallucinations and improving the factual accuracy of large language models. By leveraging the FEVER and TruthfulQA datasets, we evaluated the performance of LLaMA-3-8B against baseline models, including GPT-2 and Flan-T5. Initial evaluations showed LLaMA-3-8B performing on par with GPT-2 on TruthfulQA but lagging behind on FEVER. However, fine-tuning resulted in substantial improvements: LLaMA-3-8B's accuracy on TruthfulQA increased from 24.5% to 51.9%, while its performance on FEVER improved from 90.6% to a perfect 100%.

These findings highlight the transformative potential of targeted fine-tuning in enhancing model reliability and truthfulness. While the initial results reveal challenges in handling factual consistency, the dramatic post-finetuning improvements underscore the value of specialized training techniques. This study reaffirms the importance of domain-specific fine-tuning for mitigating hallucinations and aligns with the broader goal

of developing more trustworthy AI systems. Future work could explore fine-tuning with a broader range of datasets and investigate additional strategies for handling adversarial or ambiguous inputs.

ACKNOWLEDGMENT

Clark and Ta gratefully acknowledge the Wentworth Institute of Technology for their generous support in sponsoring this study and providing access to state-of-the-art facilities, which were instrumental in the successful completion of this research. The authors also express their deepest gratitude to Professor Salem Othman for his invaluable guidance, insightful feedback, and unwavering support throughout the duration of this project. His mentorship greatly enriched the research process and contributed significantly to the study's success.

REFERENCES

[1] M. Watanabe and N. Uchihira, "Digital business model analysis using a large language model," 2024.

[2] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, and T. Liu, "A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions," *ACM Transactions on Information Systems*, vol. 43, p. 1–55, Jan. 2025.

[3] P.-J. Lin, R. Balasubramanian, F. Liu, N. Kandpal, and T. Vu, "Efficient model development through fine-tuning transfer," 2025.

[4] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *openAI*, 2019.

[5] P. Törnberg, D. Valeeva, J. Uitermark, and C. Bail, "Simulating social media using large language models to evaluate alternative news feed algorithms," 2023.

[6] D. Wadden, S. Lin, K. Lo, L. L. Wang, M. van Zuylen, A. Cohan, and H. Hajishirzi, "Fact or fiction: Verifying scientific claims," in *EMNLP*, 2020.

[7] T.-H. Cheung and K.-M. Lam, "Factllama: Optimizing instruction-following language models with external knowledge for automated fact-checking," 2023.

[8] S. Longpre, L. Hou, T. Vu, A. Webson, H. W. Chung, Y. Tay, D. Zhou, Q. V. Le, B. Zoph, J. Wei, and A. Roberts, "The flan collection: Designing data and methods for effective instruction tuning," 2023.

[9] J. Thorne, A. Vlachos, C. Christodoulopoulos, and A. Mittal, "FEVER: a large-scale dataset for fact extraction and VERification," in *NAACL-HLT*, 2018.

[10] S. Lin, J. Hilton, and O. Evans, "Truthfulqa: Measuring how models mimic human falsehoods," 2022.

[11] AI@Meta, "Llama 3 model card," *Meta*, 2024.

[12] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, E. Li, X. Wang, M. Dehghani, S. Brahma, A. Webson, S. S. Gu, Z. Dai, M. Suzgun, X. Chen, A. Chowdhery, S. Narang, G. Mishra, A. Yu, V. Zhao, Y. Huang, A. Dai, H. Yu, S. Petrov, E. H. Chi, J. Dean, J. Devlin, A. Roberts, D. Zhou, Q. V. Le, and J. Wei, "Scaling instruction-finetuned language models," 2022.

[13] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," 2021.

[14] S. Mangrulkar, S. Gugger, L. Debut, Y. Belkada, S. Paul, and B. Bossan, "Peft: State-of-the-art parameter-efficient fine-tuning methods." https://github.com/huggingface/peft, 2022.

[15] A. Gupta and V. Srikumar, "X-FACT: A New Benchmark Dataset for Multilingual Fact Checking," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, (Online), Association for Computational Linguistics, July 2021.