Does Size Matter? Evaluating the Trade-Off Between Model Scale and Logical Reasoning Performance

Falkowski Nadav , Yom Tov Doron
Afeka College of Engineering

June 4, 2025

Abstract

The increasing availability of Large Language Models (LLMs) executable on local hardware presents new opportunities for applications requiring privacy and control, but raises questions about their cognitive capabilities compared to larger, cloud-based models. This study aims to evaluate and compare the binary (yes/no) logical reasoning abilities of the Llama 3.1 8B model, both its base version and a version fine-tuned using LoRA (Low-Rank Adaption) on 1,000 domain-specific logical questions, representing locally executable models, against GPT-4.1, an advanced cloud-based model. The evaluation used a distinct test set of 4,200 diverse logical questions, analyzing metrics such as accuracy, confidence scores, and uncertainty rates. Results indicate that GPT-4.1 achieved the highest accuracy (approx. 98%). Fine-tuning significantly improved the performance of Llama 3.1 8B on the test set, increasing its accuracy by 17.6% from 68% to 80%, thus narrowing the performance gap with GPT-4.1. However, local models, especially the base model without fine-tuning, exhibited lower certainty rates and a less optimal confidence score calibration compared to GPT-4.1, which demonstrated high and consistent confidence in its correct answers. These findings highlight the existing trade-offs between model accessibility and performance and underscore the potential of targeted fine-tuning. With a appropriate dataset to the task, we want the model to perform to substantially enhance the capabilities of local models, even if a performance disparity with leading cloud-based models remains.

Keywords: LLMs,Fine-Tune,Evaluation

1 Introduction

In recent years, the world has witnessed unprecedented advancements in the field of artificial intelligence,largely driven by the breakthrough of transformer architecture (Vaswani

et al., 2017)[6] and its implementation in Large Language Models (LLMs). These models, trained on vast amounts of text, demonstrate impressive capabilities across a wide range of linguistic tasks, from generating coherent text to translating languages and providing complex answers. Concurrently with the development of massive models requiring substantial computational resources and typically accessed via APIs, we are observing a growing trend in the development and dissemination of smaller, more efficient models, such as Meta's Llama series (Grattafiori et al. 2024)[1], which can be run on standard consumer hardware.

The ability to run LLMs locally ("on-device") holds significant advantages. It allows for a higher degree of data privacy and security, as data is not transmitted to external servers. Furthermore, it can lead to reduced long-term operational costs (compared to pay-per-use API services), grants users full control over the model and its customization for specific needs and enables offline usage. However, this accessibility raises a critical question: Are these "home-runnable" models, despite their size and resource limitations, capable of successfully tackling complex cognitive tasks, particularly those requiring logical understanding and reasoning, at a level comparable to the leading large-scale models in the market?

Despite these impressive capabilities, the field of testing and evaluating deep logical reasoning abilities in LLMs is still in its nascent stages. We believe that enhancing the logical understanding of these models is not merely an end in itself, but could lead to a general improvement in their capabilities across a broad spectrum of tasks that require a more profound comprehension of text and context. Recent studies have begun to indicate a strong correlation between logical abilities and improved performance in more complex tasks (Morishita et al., 2024)[4], highlighting the potential benefits of strengthening the foundational logical reasoning of these models.

The central question underpinning this research, therefore, is the assessment of the performance gap in logical reasoning between LLMs executable on local hardware and larger, cloud-based models, and an examination of the potential to mitigate this gap through targeted LoRA fine-tuning (Hu et, al. 2021)[3] Specifically, we focus on the ability of these models to correctly answer binary (yes/no) logical reasoning questions spanning a diverse range of logical domains and rules.

To this end, the objectives of this study are:

1. To quantitatively evaluate the logical reasoning performance of the base Llama 3.1 8B model, as a representative locally executable LLM.

2. To fine-tune the Llama 3.1 8B model using a dedicated dataset of 1,000 logical questions and subsequently evaluate the performance of this fine-tuned version.

3. To compare the performance of both Llama 3.1 8B versions (base and fine-tuned) against GPT-4.1, representing a state-of-the-art, large cloud-based model, using a

distinct test set of 4,200 logical questions

4. To analyze the relationship between the confidence scores generated by each model and their answer accuracy, and to examine uncertainty rates, aiming to understand model reliability in the context of local execution versus API-based access.

This research seeks to contribute to a deeper understanding of the capabilities and limitations of LLMs accessible to users with consumer-grade hardware for logical tasks. The findings may provide valuable insights for developers and researchers considering the trade-offs between accessibility, cost, privacy, and performance quality when selecting language models.

2 Related Work

2.1 Background: Logical Understanding and the Challenge for Large Language Models (LLMs)

Over the past few years, Large Language Models (LLMs) have significantly impacted various aspects of human life, demonstrating remarkable capabilities in natural language processing. This literature review aims to contextualize the current understanding of LLM performance in logical reasoning and recent developments in this domain. Despite the substantial advancements LLMs have made, there remains considerable room for improvement, particularly in their ability to genuinely understand and apply human logic. Human logic, as defined by (Yan et al. 2024)[9], is a core component of cognition, essential to comprehend, interact with, and influence our environment. This is echoed by (Parmar et al. 2024)[5], who describe it as a fundamental aspect of intelligence, and (Wan et al. 2024)[7], who characterize it as the cognitive process of using logic to draw conclusions from given facts. These studies collectively underscore that logic and reasoning are crucial facets of intelligence, yet their implementation in artificial systems remains a significant challenge. In light of these challenges ,our work explores whether fine-tuning of a locally executable model can lead to meaningful improvement in model logical understanding performance.

2.2 Current Approaches and Challenges of LLMs in Solving Logical Tasks

Current research suggests that LLMs often attempt to solve logic questions by relying on shortcuts, such as pattern recognition and memorization, rather than engaging in genuine, step-by-step logical reasoning (Wan et al., 2024)[7]. These models frequently leverage their exposure to vast training datasets, answering new logical questions by recalling similar

or identical examples encountered during training (Xu et al., 2024)[8]. This reliance on surface-level patterns, rather than an underlying grasp of logical principles, poses a limitation to their reasoning capabilities. This is particularly relevant when considering the trade-offs between locally executable models, which might inherently have access to or be effectively trained on smaller effective datasets (even if based on large pre-trained models), and their larger, cloud-based counterparts. Our research investigates whether this behavior is prevalent in a locally runnable Llama 3.1 8B model and if targeted fine-tuning can steer it towards more robust reasoning.

2.3 Methodologies for Evaluating Logical Capabilities of LLMs

Evaluating the logical reasoning capabilities of LLMs necessitates rigorous methodologies. (Gupta et al. 2023)[2], for instance, evaluated a Binary Choice Question (BCQ) model using exact match accuracy and pairwise accuracy to assess its ability to correctly identify true or false answers and understand feasibility. For Multiple Choice Question (MCQ) models, they employed exact match accuracy and recall to evaluate the model's proficiency in selecting all correct answer options. Similarly, (Parmar et al. 2024)[5] introduced LogicBench, a benchmark specifically designed to evaluate the logical reasoning capabilities of LLMs. LogicBench utilizes Binary Question Answering (BQA) and Multiple Choice Question Answering (MCQA) tasks to assess an LLM's ability to determine logical entailment and select the most appropriate conclusion from several options. Our study adopts a BQA approach, similar to those discussed, focusing on exact match accuracy for "yes/no" answers, which allows for a clear comparison of reasoning abilities between the locally-run Llama 3.1 8B model (base and fine-tuned) and the cloud-based GPT-4.1.

2.4 Strategies for Improving LLM Understanding of Logic

Several strategies have been proposed to enhance LLMs' understanding of logic. (Yan et al. 2024)[9] suggest that LLMs might improve their logical reasoning performance through in-context learning, although they caution that this method may primarily enhance task-specific performance rather than genuinely boosting the model's conceptual understanding. (Parmar et al. 2024)[5] advocate for the development of new, specialized datasets to train LLMs for improved logical performance. (Xu et al. 2024)[8] emphasize the criticality of developing pre-training or fine-tuning strategies specifically designed to enhance model reasoning abilities. A common thread across these studies is the suggestion that significant improvements in logical understanding are likely to stem from interventions at the pre-training or dedicated training/fine-tuning phases. Our research directly investigates the latter, examining the impact of a targeted, yet modest, fine-tuning process (using 1,000 specific logical questions) on the Llama 3.1 8B model, a strategy that is fea-

sible for users with local hardware, and assessing its efficacy in bridging the performance gap with larger models.

2.5 The Gap Between Local and Cloud-Based Models in Logical Capabilities

While extensive research has benchmarked large, proprietary models like those in the GPT series, and a growing body of work explores the capabilities of open-source, locally runnable models, direct comparisons focused on complex logical reasoning tasks, particularly considering the impact of accessible fine-tuning on these local models, are less common. Understanding the performance trade-offs between readily available local models and resource-intensive cloud-based models is crucial for practical applications where logical consistency and accuracy are important . This study aims to contribute to this understanding by specifically comparing the Llama 3.1 8B model, a prominent example of a locally deployable LLM, with GPT-4.1 in a structured logical reasoning benchmark.

2.6 Summarizing Gaps and Justifying the Current Research

The existing literature indicates that despite significant advancements, LLMs, including very large ones, still face challenges in fully understanding and applying the principles of human logic, often relying on pattern recognition and memorization. While various evaluation benchmarks and improvement strategies like specialized training or fine-tuning have been proposed, a clear understanding of how smaller, locally executable models perform against their larger counterparts, and to what extent accessible fine-tuning can enhance their logical capabilities, remains an active area of investigation. This review highlights the need for further research to better understand the capabilities and limitations of different classes of LLMs in the domain of human logic. Our study addresses this by providing a direct comparison between a locally run Llama 3.1 8B model (before and after fine-tuning on a modest dataset) and the powerful GPT-4.1, focusing on binary logical reasoning tasks. The findings aim to offer practical insights into the feasibility of leveraging local LLMs for tasks demanding logical acuity and the effectiveness of resource-efficient fine-tuning in this context.

3 Methodology

3.1 Overview

This section present the methodological framework used to evaluate and compare the logical reasoning capabilities of three language models:

- Pre-trained LLaMA 3.1 8B
- Fine-tuned LLama 3.1 8B
- GPT-4.1

The evaluation focused on binary logic questions that require clear "yes"/"no" response. The goal is to determine wheter targeted fine-tuning on a modest dataset of logical reasoning tasks can significantly improve the performance of a locally runnable model, and how such improvements compare to a state-of-the-art cloud-based model. The following subsections describe the dataset,models, and the evaluation setup

3.2 Dataset

The evaluation of the models in this study was based on a publicly available dataset of logical reasoning questions sourced from the LogicAsker project (Wan et al., 2024)[7]. The complete dataset contains 5,200 unique questions, formulated to require the model to perform binary logical inference and produce a "yes" or "no" response. The questions in the dataset span two main types of formal logic: propositional logic and predicate logic, This distinction enables a better assessment of the models ability to handle different levels of logical problems. For the fine-tuning of the Llama 3.1 8B model, a portion of this dataset, 1,000 questions, was dedicated as the training set. The remaining 4,200 questions from the dataset constituted a distinct and non-overlapping evaluation set, upon which all models (base Llama 3.1 8B, fine-tuned Llama3.1 8B, and GPT-4.1) were assessed. Each model was evaluated in a single run across all 4,200 of these questions, and the performance metrics reported in this study refer to this full evaluation

3.3 Evaluated Models

Three Large Language Models (LLMs) were examined in this study, representing two primary categories: models executable locally on standard user hardware, and an advanced cloud-based model requiring API access.

3.3.1 Locally Executable Models: Llama 3.1 8B

The Llama 3.1 8B model was selected as representative of a locally executable LLM due to its open-source availability and relatively high performance for its size. Two versions of this model were used for the study:

Base Version (Llama 3.1 8B - Base) - This is the standard pre-trained version of the model, as obtained from its official distribution source.

Fine-tuned Version (Llama 3.1 8B - Fine-tuned) - This version was based on the same Llama 3.1 8B base model, which underwent a fine-tuning process specifically for logical reasoning tasks using the first 1,000 questions from the dataset (as described in Section 3.2) Using LoRA Fine-tuning method. LoRA is an efficient fine-tuning technique that significantly reduces the number of trainable parameters by injecting smaller, adaptable matrices into the existing layers of a pre-trained model, allowing for effective specialization on new tasks with lower computational cost.

3.3.2 Cloud-Based Model: GPT-4.1

To provide a benchmark representing a large, state-of-the-art language model not typically executable locally by most users, GPT-4.1 was selected. Access to this model was facilitated through the official OpenAI API.The inclusion of this model aims to assess the potential performance gap between accessible local models and leading-edge models requiring extensive computational resources.

3.4 Evaluation Procedure

The evaluation of each model involved presenting each question from the test set to the model and obtaining a binary answer ("yes" or "no") along with a confidence score (between 0.5 and 1).

3.4.1 Confidence Score and Uncertainty Calculation

We set the confidence threshold to 0.2. This means that if the model assigns a confidence score of 0.6 or higher to a prediction, the prediction is considered confident and is retained for further analysis. Predictions with confidence scores below this threshold are marked as uncertain.

3.5 Fine-tuning Configuration Selection

Prior to the full comparative evaluation, a preliminary experiment was conducted to determine an optimal training dataset size for the fine-tuning of the Llama3.1 8B model, within the available resources and the allocated 1,000-question training pool. The Llama3.1 8B model was separately fine-tuned using different-sized subsets of this training pool: 100 questions, 250 questions, 500 questions, and the full 1,000 questions. The fine-tuning process itself (model loading, LoRA configuration, training arguments) was identical to that described in Section 3.3.1 for the finally selected fine-tuned version. Each of these four fine-tuned model versions was then evaluated on the full test set (4,200 questions). Their performance was compared using overall Accuracy, Precision, Recall, and F1-scores (specific to "yes" and "no" answers), as will be detailed in the Results section 1 .The

version fine-tuned on 1,000 questions yielded the highest overall accuracy (80.4%) among all tested fine-tuned variants and was consequently selected to represent the fine-tuned local model in the main comparisons of this study

Model	Precision by Answer		Recall by Answer		F1 Score by Answer		Accuracy
	Precision By Yes.	Precision By No.	Recall By Yes.	Recall By No.	F1 Score by Yes.	F1 Score by No.	
100 Questions Fine-tuned	48.6 %	93.7 %	90.5%	60.8%	63.3%	72.8%	68.7%
250 Questions Fine-tuned	57.4 %	85.2 %	67.7%	78.7%	62.1%	81.8%	75.4%
500 Questions Fine-tuned	59.1 %	77.8 %	40.9%	88%	48.3%	82.6%	74%
1000 Questions Fine-tuned	65.4 %	87.6 %	72.2%	83.8%	68.6%	85.6%	80.4%

Table 1: Fine Tuned Accuracy Matrix

3.6 Evaluation Metrics

The models' performance was assessed using a range of standard quantitative metrics pertinent to machine learning and language model evaluation, focusing on the binary classification task (yielding "yes" or "no" answers). For the calculation of specific metrics, the following standard confusion matrix components were considered:

- **True Positives - TP yes** : Instances where the expected answer was "yes" and the model predicted "yes".
- **True Negatives - TN yes** : Instances where the expected answer was "no" and the model predicted "no".
- **False Positives - FP yes** : Instances where the expected answer was "no" but the model predicted "yes".
- **False Negatives - FN yes** : Instances where the expected answer was "yes" but the model predicted "no".

The metrics were collected and organized into three main groups, as will be presented in the results tables

3.6.1 Overall and Answer-Specific Performance Metrics

This first set of metrics provides a comprehensive overview of each model's classification performance, distinguishing its ability to correctly predict each of the possible answers. These metrics include:

- **Precision by Answer:**
  - **Precision by Yes:** $\frac{TP}{TP+FP}$
  - **Precision by No:** $\frac{TN}{TN+FN}$

- **Recall by Answer:**
  - **Recall by Yes:** $\frac{TP}{TP+FN}$
  - **Recall by No:** $\frac{TN}{TN+FP}$
- **F1 Score by Answer:**
  - **F1 Score by Yes:** $2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision}+\text{Recall}}$
  - **F1 Score by No:** $2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision}+\text{Recall}}$
- **Accuracy (Overall Accuracy):** $\frac{TP+TN}{TP+TN+FP+FN}$

Model	Precision by Answer		Recall by Answer		F1 Score by Answer		Accuracy
	Precision By Yes.	Precision By No.	Recall By Yes.	Recall By No.	F1 Score by Yes.	F1 Score by No.	
Base Llama 3.1 8B	4.7%	95.4%	70.2%	30.5%	8.8%	46.2%	68.4%
Fine-tuned Llama 3.1 8B	65.4 %	87.6 %	72.2%	83.8%	68.6%	85.6%	80.4%
GPT-4.1	96.4%	98%	95.4%	98.4%	95.8 %	98.1%	97.5%

Table 2: Accuracy Matrix

3.6.2 Performance Analysis by Question Characteristics

The second set of metrics examines model accuracy across different sub-categories of questions within the dataset, as defined by the questions' metadata fields. These metrics include accuracy for:

- **Logic:** Propositional (Prop.), Predicate (Pred.).
- **Category:** Inference(Infer.), Equivalence(Equiv.), Fallacy(Fall.).
- **Problem:** Inference(Infer.), Unrelated(Unrel.), Contradiction (Contr.)

Model	Accuracy by Logic		Accuracy by Category			Accuracy by Problem		
	Prop.	Pred.	Infer.	Equiv.	Fall.	Infer.	Unrel.	Contr.
Base Llama 3.1 8B	67.2%	68.7%	65.1%	66%	95.2%	27.7%	96.4%	94.4%
Fine-tuned Llama 3.1 8B	72%	82.3%	81.8%	83%	63.7%	70%	87.1%	87.3%
GPT-4.1	98.9%	97.2%	97.1%	98.3%	96.9%	96.5%	97.7%	98.7%

Table 3: Performance comparison across different type of questions

3.6.3 Performance Analysis by Answer Type and Confidence Levels

The third set of metrics investigates the relationship between model confidence and performance, and the distribution of confidence across answer types. These metrics include:

- **Avg. Confidence by Answer:**

  - **Confidence by Yes:** The average confidence score of the model when the expected answer was "yes."
  - **Confidence by No:** The average confidence score of the model when the expected answer was "no."

- **Accuracy by Confidence:**
  - **Accuracy with Low Confidence:** Model accuracy on predictions where its confidence score was low (confidence score < 0.6).
  - **Accuracy with High Confidence:** Model accuracy on predictions where its confidence score was high (confidence score $>= 0.6$).

- **Avg. Confidence (Overall Average Confidence):** The mean confidence score across all predictions made by the model.

Model	Avg. Confidence by Answer		Accuracy by Confidence		Avg. Confidence
	Confidence by Yes	Confidence by No	Accuracy with Low Confidence	Accuracy with High Confidence	
Base Llama 3.1 8B	50.2%	50.3%	68.5%	51.7%	50.3%
Fine-tuned Llama 3.1 8B	74.9%	65.6%	78.7%	81.8%	68.3%
GPT-4.1o	99.7%	99.7%	66.7%	97.5%	99.7%

Table 4: Performance comparison across different type of Answer and confidence

3.6.4 Performance Analysis by ROC-AUC

To evaluate the abilities of the models, we employed the Receiver Operating Characteristic (ROC) curve and its corresponding Area Under the Curve (AUC) metric. The ROC curve plots the true positive rate (TPR) against the false positive rate (FPR) at various classification thresholds, providing a threshold-independent evaluation of model performance.

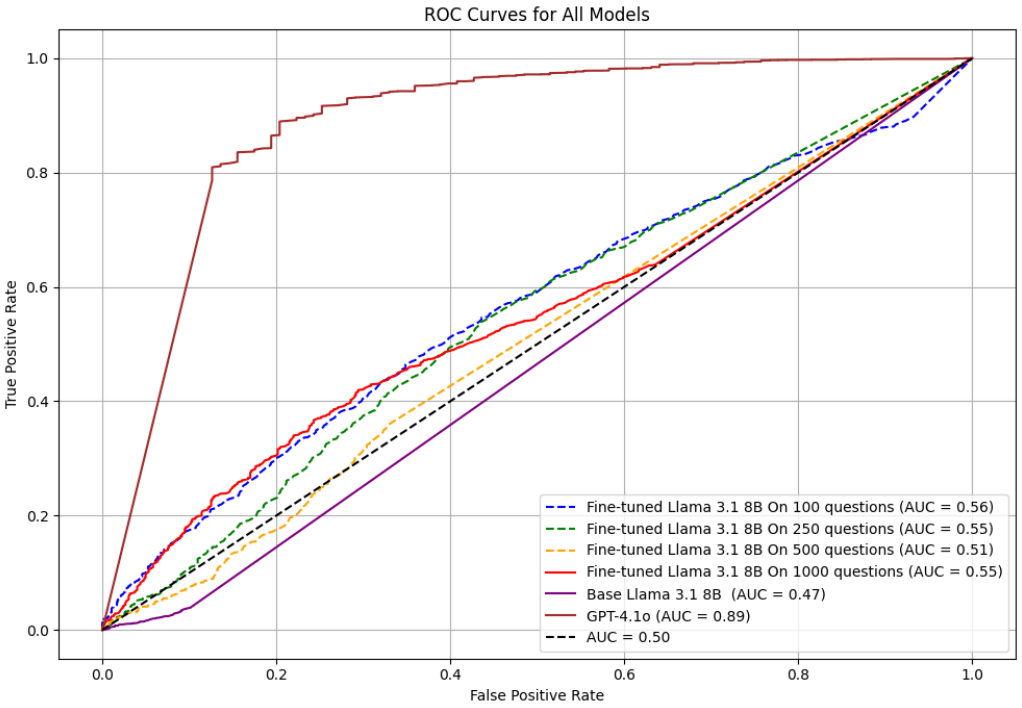

Figure 1: ROC Curves for Llama and GPT-4 models, evaluated on the binary classification task.

4 Results

In this study , we evaluated the performance of three LLM's models : a pre-trained Llama 3.1 8B , a fine-tuned version of the Llama 3.1 8B model and GPT-4.1 The evaluation focused on their accuracy across various question types and their confidence levels in responding. Our findings from the experiments are detailed below.

4.1 Overall Model Performance

As shown in Table 2 ,which presents the overall accuracy matrix , there were clear distinctions in performance between the models .

4.1.1 Base Llama 3.1 8B Overall performance

The pre-trained base Llama 3.1 8B model achieved an overall accuracy of 68.4% When looking at its ability to correctly identify "Yes" and "No" answers, it showed a precision of 4.7% for "Yes" answers and 95.4% for "No" answers. Its recall was 70.2% for "Yes" and 30.5% for "No", leading to F1 scores of 8.8% for "Yes" and 46.2% for "No".

4.1.2 Fine-tuned Llama 3.1 8B Overall performance

The fine-tuned Llama 3.1 8B model demonstrated a notable improvement over the base version, with its overall accuracy rising to 80.4%. This enhancement was reflected in its precision for "Yes" answers (72.2%) and "No" answers (83.8%), as well as its recall (65.4% for "Yes" and 87.6% for "No"). Consequently, the F1 scores also improved to 68.6% for "Yes" and 85.6% for "No".

4.1.3 GPT-4.1 Overall performance

The GPT-4.1 model preformed the among the chosen models, achieving a striking overall accuracy of 97.5%. It exhibited very high precision for both "Yes" (96.4%) and "No" (98%) answers. Similarly, its recall was strong at 95.4% for "Yes" and 98.4% for "No", resulting in impressive F1 scores of 95.8% for "Yes" and 98.1% for "No".

4.2 Performance across different question types

We further analyzed the models' performance across various types of questions , as detailed in Table 3.

4.2.1 Models performance across different logic questions

In logical reasoning, specifically Accuracy by Logic, the base Llama 3.1 8B model scored 67.2% on propositional logic (Prop.) and 68.7% on predicate logic (Pred.). The fine-tuned Llama 3.1 8B model showed improvements, reaching 72% for propositional and 82.3% for predicate logic. GPT-4.1 again outperformed the others with 98.9% in propositional and 97.2% in predicate logic.

4.2.2 Models performance across different Category questions

When examining Accuracy by Category, the base Llama 3.1 8B model achieved 65.1% for category inference (Infer.), 66% for equivalence (Equiv.), and a higher 95.2% for fallacy detection (Fall.). The fine-tuned Llama 3.1 8B model showed a mixed pattern of improvement, scoring 81.8% for category inference, 83% for equivalence, but a lower 63.7% for fallacy detection compared to its base counterpart. GPT-4.1 consistently performed well, with 97.1% for category inference, 98.3% for equivalence, and 96.9% for fallacy detection.

4.2.3 Models performance across different problems type questions

Finally, looking at Accuracy by problem type, the base Llama 3.1 8B model had varied results: 27.7% for inference (Infer.), 96.4% for unrelated (Unrel.), and 94.4% for contradiction (Contr.). The fine-tuned Llama 3.1 8B model significantly improved in inference (70%), but saw a decrease in accuracy for unrelated (87.1%) and contradiction (87.3%)

questions. GPT-4.1 demonstrated robust performance across these rule-based questions, scoring 96.5% for inference, 97.7 for unrelated, and 98.7% for contradiction.

4.3 Performance by answer confidence

Table 4 provides insights into the models' confidence levels and how confidence correlated with accuracy.

4.3.1 Base Llama 3.1 8B model confidence analysis

The base Llama 3.1 8B model showed an average confidence of 50.2% for "Yes" answers and 50.3% for "No" answers, with an overall average confidence of 50.3%. Its accuracy was 68.5% when its confidence was low and 51.7% when its confidence was high.

4.3.2 Fine-tuned Llama 3.1 8B Model confidence analysis

The fine-tuned Llama 3.1 8B model displayed higher average confidence levels, with 74.9% for "Yes" answers and 65.6% for "No" answers, resulting in an overall average confidence of 68.3%. Interestingly, this model was more accurate when its confidence was high (81.8%) compared to when its confidence was low (78.7%).

4.3.3 GPT-4.1 Model confidence analysis

GPT-4.1 exhibited very high confidence across the board, with an average confidence of 99.7% for both "Yes" and "No" answers, and consequently, an overall average confidence of 99.7%. Its accuracy was substantially higher when its confidence was high (97.5%) compared to when its confidence was low (66.7%), though instances of low confidence were likely rare given its high average.

4.4 Comparison of Models Based on ROC-AUC

Figure 1 shows the ROC curves for several variants of the Llama 3.1 8B model, fine-tuned on different numbers of labeled questions (100, 250, 500, and 1000), as well as the base model and GPT-4. The AUC values provide a quantitative summary of performance, where a value of 0.5 indicates random guessing and a value of 1.0 indicates perfect classification.

The results reveal that:

- The base Llama 3.1 8B model performs close to random (AUC = 0.47), indicating limited discriminative power without fine-tuning.
- Fine-tuning improves performance modestly, with AUC scores ranging from 0.51 to 0.56 as more training data is used.

- GPT-4 significantly outperforms all Llama variants, achieving an AUC of 0.89.

5 Discussion

This chapter discusses the significance of the results presented in Table 2,Table 3 and Table 4 interprets the models' performance in the context of the research objectives, compares them to prior studies (as reviewed in Section 2), acknowledges the study's limitations, and suggests implications and insights.

5.1 Interpretation and Comparison of Model Performances

The data presented in Table 2 show clear distinctions in logical reasoning capabilities among the three evaluated models, as well as a significant impact of the fine-tuning process on the Llama 3.1 8B model

- Superiority of the GPT-4.1 Model: The GPT-4.1 model consistently demonstrated the highest performance across all key metrics shown in Table 1, achieving an overall accuracy of 97.5%. This superiority was also evident in its Precision (96.4% for "Yes" and 98.0% for "No"), Recall (95.4% for "Yes" and 98.4% for "No"), and F1-scores (95.8% for "Yes" and 98.1% for "No"). These results, suggest robust and consistent logical inference capabilities. The significant performance gap between GPT-4.1 and the Llama 3.1 8B models, even post-fine-tuning, highlights the potential influence of model scale, the extent of original training data, and possibly more advanced architectures present in larger cloud-based models.

- Further examination of the models' discrimination ability: as depicted by the ROC curves and their corresponding AUC values (Figure 1), provides additional insights into the impact of fine-tuning and training dataset size on the Llama3.1 8B model. GPT-4.1o achieves a notably high AUC of 0.89, indicating excellent discriminative power between "yes" and "no" answers based on its confidence scores. In contrast, the base Llama 3.1 8B model shows a low AUC of 0.47, close to random guessing (AUC=0.50), reinforcing the conclusion about its limited discrimination capabilities without fine-tuning. Fine-tuning the Llama3.1 8B model demonstrates an improvement in AUC values compared to the base model. Interestingly, fine-tuning on 100 questions yielded the highest AUC among the fine-tuned Llama versions (0.56). Increasing the training dataset size to 250 questions (AUC=0.55), 500 questions (AUC=0.51), and even 1000 questions (AUC=0.55) did not lead to a linear or monotonic increase in AUC, with a slight dip observed for the 500-question set. This result might suggest that a very small and highly focused dataset (like 100 questions) can be effective for an initial improvement in discrimination as measured by

AUC. However, further increases in training data size may require different strategies, or the dataset itself might contain complexities leading to this non-monotonic behavior. It is also possible that the relatively low AUC values (around 0.5-0.56) for all fine-tuned Llama versions indicate that their confidence scores, even after tuning, do not optimally separate the classes, despite improvements in overall accuracy (as seen in Table 1, where fine-tuning on 1000 questions yielded the highest accuracy).

- Impact of Fine-tuning on Llama 3.1 8B: The fine-tuning of the Llama 3.1 8B model led to a considerable improvement in its performance, as evident from Table 1. The overall accuracy rose from 68.4% in the base version to 80.4% in the fine-tuned version. This enhancement was particularly notable in metrics related to "Yes" answers: Precision increased from 4.7% to 72.2%, and the F1-score rose from 8.8% to 68.6%. The Recall for "Yes" changed from 70.2% in the base version to 65.4% in the fine-tuned version (a slight decrease). For "No" predictions, Precision no decreased from 95.4% to 83.8%, but Recall no significantly increased from 30.5% to 87.6%,leading to a higher F1_No (from 46.2% to 85.6%). These results suggest that fine-tuning helped the model to better balance its overall ability.

An analysis of performance across different question types Table 3 reveals a more complex picture. While accuracy on "Inference" tasks ("Accuracy by Problem: Infer.") dramatically improved after fine-tuning (from 27.7% to 70%), and in predicate logic tasks ("Accuracy by Logic: Pred.") (from 68.7% to 82.3%), a surprising decrease in accuracy was observed for "Fallacy detection" ("Accuracy by Category: Fall.") (from 95.2% to 63.7%) and also performance drops for "Unrelated" (from 96.4% to 87.1%) and "Contradiction" (from 94.4% to 87.3%) questions under the "Accuracy by Problem" categorization.

- Performance of the Base Llama 3.1 8B Model: The base version of Llama 3.1 8B exhibited the lowest performance among the three models (overall accuracy of 68.4%). The particularly low Precision for "Yes" answers (4.7%) combined with a relatively high Recall for "Yes" (70.2%, per Table 1), versus high Precision for "No" (95.4%) but very low Recall (30.5%), suggests that the base model tends to predict "No" very frequently. This bias results in high precision for "No" but significantly impairs its ability to correctly identify "Yes" instances, as reflected in the very low F1 Score for "Yes" (8.8%).

5.2 Analysis of Confidence Scores and Model Calibration

Table 4 provides insights into the relationship between model confidence and prediction accuracy.

- GPT-4.1: Displayed a very high overall average confidence (99.7%), with similar average confidence scores for expected "Yes" and "No" answers (99.7% for both).

Its accuracy was substantially higher (97.5%) when its confidence was classified as "High" (confidence score ¿= 0.6), compared to lower accuracy (66.7%) in the rare instances of "Low" confidence (confidence score ¡ 0.6). This indicates good calibration.

- Llama 3.1 Fine-tuned: Showed an increase in overall average confidence to 68.3% compared to the base model. Average confidence for expected "Yes" answers was 74.9% and for expected "No" answers 65.6%. This model was also more accurate when its confidence was high (81.8%) than when it was low (78.7%), suggesting improved calibration over the base model.

- Llama3.1 Base: Exhibited a very low overall average confidence (50.3%), with similar averages for expected "Yes" and "No" answers (50.2% and 50.3% respectively). Indicating poor calibration, its accuracy was surprisingly higher when its confidence was classified as "Low" (68.5%) than when classified as "High" (51.7%). This finding underscores that the base model's confidence score is not a reliable indicator of answer correctness.

These findings suggest that the larger model (GPT-4.1) is not only more accurate but also better calibrated in terms of its confidence. Fine-tuning improved the calibration of Llama 3.1 8B, but the base model shows a problematic relationship between confidence and accuracy.

5.3 Comparison with Prior Work

The findings of this study align with several existing insights in the literature concerning the logical understanding of large language models, while also presenting important nuances regarding locally executable models and the impact of fine-tuning.

- The base Llama 3.1 8B model's relatively low performance on pure inference tasks (Table 3: 27.7% in "Accuracy by Problem: Infer.") and its poorly calibrated confidence (Table 4) are consistent with observations from (Wan et al.2024)[7] and (Xu et al. 2024)[8], who suggest that many LLMs, without specific adaptation, tend to rely on superficial pattern recognition rather than deep logical reasoning. Our findings demonstrate this phenomenon persists even in relatively modern models like Llama 3.1 8B.

- The substantial improvement in Llama 3.1 8B's performance after fine-tuning on only 1,000 examples (e.g., overall accuracy increased to 80.4%, Table 2) highlights the effectiveness of fine-tuning strategies, as emphasized by (Xu et al. (2024)[8]. This aligns with the direction suggested by (Parmar et al. 2024)[5] regarding the

need for specialized training data. This demonstrates that significant gains can be achieved with modest, accessible fine-tuning efforts on local hardware.

- GPT-4.1's superior performance across all evaluated aspects is in line with the general consensus on the capabilities of leading large-scale proprietary models. Our study quantifies this gap for specific logical tasks and shows that while fine-tuning can narrow it, a substantial difference remains.

- Our evaluation methodology, employing detailed class-specific metrics and performance breakdowns by question characteristics, align with the approach of modern benchmarks like LogicBench (Parmar et al., 2024)[5], which advocate for systematic and granular assessment beyond overall accuracy.

5.4 Study Limitations

- **Dataset:** The LogicAsker dataset, while substantial, may not represent the full spectrum of logical problems and might have inherent biases. The specific 1,000 training samples could have selectively influenced the fine-tuned model's performance

- **Model Selection:** The study focused on specific versions of Llama 3.1 8B and GPT-4.1. Other local models or different versions could yield different results.

- **Temperature:** The study focused on singular temperature value (0) different value might yield better results

- **Fine-tuning Process:** The chosen LoRA and training parameters are one configuration; others might yield different outcomes.

- **Binary Evaluation:** The study focused on "yes/no" accuracy, not the quality of explanations or reasoning chains.

5.5 Implications and Contributions

This research offers several valuable insights. Firstly, it quantifies the performance gap in logical reasoning between leading cloud-based models and smaller, locally executable open-source models, even post-fine-tuning. Secondly, it demonstrates that targeted fine-tuning with a modest dataset can dramatically improve certain aspects of logical performance in local models but also highlights risks of over-specialization or detriment to other capabilities. This is crucial for developers considering fine-tuning. Thirdly, confidence score analysis reveals significant calibration differences, with the base local model showing particularly poor calibration, underscoring the importance of not blindly trusting

confidence scores from smaller models. Overall, the study contributes to a practical understanding of trade-offs between accessibility and logical performance, and the effectiveness and challenges of enhancing local models via accessible fine-tuning.

6 Conclusions and Future Work

6.1 Main Conclusions

This research compared the logical reasoning abilities of the Llama 3.1 8B model (base and fine-tuned) against GPT-4.1. The main findings are:

1. GPT-4.1 demonstrated clear superiority in overall accuracy and most class-specific performance metrics compared to both Llama 3.1 8B versions.

2. Fine-tuning Llama 3.1 8B on 1,000 logical examples significantly improved its overall accuracy and its ability on certain inference tasks, but concurrently degraded its performance on other question types, such as fallacy detection.

3. GPT-4.1 exhibited better confidence calibration, with high confidence scores generally aligning with high accuracy. In contrast, the base Llama 3.1 model showed poor confidence calibration, and the fine-tuned version, while improved, was still less consistent than GPT-4.1.

4. These results suggest that while local models can be improved via fine-tuning, achieving the performance and reliability levels of leading large-scale cloud models remains a significant challenge, requiring careful attention not only to accuracy but also to potential unintended consequences of the tuning process.

6.2 Recommendations for Future Work

- **Expansion and Diversification of Fine-tuning Data:** Investigate the impact of larger and more diverse training datasets for fine-tuning, covering a broader range of logical rules and problem types, to potentially mitigate over-specialization and improve performance across the full spectrum of logical challenges.

- **Optimization of the Fine-tuning Process:** Explore different LoRA parameters and other fine-tuning techniques to maximize improvements in logical capabilities

- **Qualitative Error Analysis:** Conduct an in-depth qualitative analysis of the types of errors models make, particularly in cases where fine-tuning negatively impacted performance, to better understand the underlying reasons.

- **Temperature Analysis:** Find the optimal value which yields the best result

- **Evaluation on Additional Datasets:** Test the models on other logical reasoning benchmarks to assess the generalizability of the findings.
- **Exploration of Other Local Models:** Compare with a wider array of locally executable models of varying sizes and architectures.
- **Enhancing Confidence Calibration:** Investigate techniques specifically aimed at improving the confidence calibration of local models to make their outputs more reliable for end-users.

References

[1] Aaron Grattafiori et al. The llama 3 herd of models, 2024.

[2] Himanshu Gupta, Neeraj Varshney, Swaroop Mishra, Kuntal Kumar Pal, Saurabh Arjun Sawant, Kevin Scaria, Siddharth Goyal, and Chitta Baral. "john is 50 years old, can his son be 65?" evaluating nlp models' understanding of feasibility, 2023.

[3] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021.

[4] Terufumi Morishita, Gaku Morio, Atsuki Yamaguchi, and Yasuhiro Sogawa. Enhancing reasoning capabilities of llms via principled synthetic logic corpus, 2024.

[5] Mihir Parmar, Nisarg Patel, Neeraj Varshney, Mutsumi Nakamura, Man Luo, Santosh Mashetty, Arindam Mitra, and Chitta Baral. Logicbench: Towards systematic evaluation of logical reasoning ability of large language models, 2024.

[6] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023.

[7] Yuxuan Wan, Wenxuan Wang, Yiliu Yang, Youliang Yuan, Jen tse Huang, Pinjia He, Wenxiang Jiao, and Michael R. Lyu. Logicasker: Evaluating and improving the logical reasoning ability of large language models, 2024.

[8] Fangzhi Xu, Qika Lin, Jiawei Han, Tianzhe Zhao, Jun Liu, and Erik Cambria. Are large language models really good logical reasoners? a comprehensive evaluation and beyond, 2024.

[9] Junbing Yan, Chengyu Wang, Jun Huang, and Wei Zhang. Do large language models understand logic or just mimick context?, 2024.