Towards Optimal NLP Solutions: Analyzing GPT and LLaMA-2 Models Across Model Scale, Dataset Size, and Task Diversity

Ankit Kumar
Department of Computer Science, University of Delhi, India
akumar@cs.du.ac.in

Richa Sharma
Department of Computer Science, University of Delhi, India
richasharma@keshav.du.ac.in (corresponding author)

Punam Bedi
Department of Computer Science, University of Delhi, India
pbedi@cs.du.ac.in

Received: 16 March 2024 | Revised: 30 March 2024 | Accepted: 2 April 2024

 DOI: https://doi.org/10.48084/etasr.7200

ABSTRACT

This study carries out a comprehensive comparison of fine-tuned GPT models (GPT-2, GPT-3, GPT-3.5) and LLaMA-2 models (LLaMA-2 7B, LLaMA-2 13B, LLaMA-2 70B) in text classification, addressing dataset sizes, model scales, and task diversity. Since its inception in 2018, the GPT series has been pivotal in advancing NLP, with each iteration introducing substantial enhancements. Despite its progress, detailed analyses, especially against competitive open-source models like the LLaMA-2 series in text classification, remain scarce. The current study fills this gap by fine-tuning these models across varied datasets, focusing on enhancing task-specific performance in hate speech and offensive language detection, fake news classification, and sentiment analysis. The learning efficacy and efficiency of the GPT and LLaMA-2 models were evaluated, providing a nuanced guide to choosing optimal models for NLP tasks based on architectural benefits and adaptation efficiency with limited data and resources. In particular, even with datasets as small as 1,000 rows per class, the F1 scores for the GPT-3.5 and LLaMA-2 models exceeded 0.9, reaching 0.99 with complete datasets. Additionally, the LLaMA-2 13B and 70B models outperformed GPT-3, demonstrating their superior efficiency and effectiveness in text classification. Both the GPT and LLaMA-2 series showed commendable performance on all three tasks, underscoring their ability to handle a diversity of tasks. Based on the size, performance, and resources required for fine-tuning the model, this study identifies LLaMA-2 13B as the most optimal model for NLP tasks.

Keywords-natural language processing; large language models; GPT series; LLaMA-2 series; fine tuning

I. INTRODUCTION

Foundational research across diverse domains has significantly influenced the transformative evolution of Natural Language Processing (NLP), showcasing its potential in complex problem-solving and decision-making scenarios [1-4]. The introduction of large pre-trained language models, such as the Generative Pre-trained Transformer (GPT) series by OpenAI and Meta AI's LLaMA-2 models, marks a pivotal moment in this evolution. The GPT series has set new benchmarks in NLP by demonstrating exceptional performance in a myriad of language tasks. The first in the series, GPT-1 [5], introduced with 110 million parameters, laid the groundwork for generating coherent and contextually rich text. Based on this, GPT-2 [6] was proposed with 1.5 billion parameters, significantly enhancing the quality of text generation. Observing power laws relating the model size to performance, researchers proposed another model, GPT-3 [7], with 175 billion parameters, offering unparalleled depth in language understanding and generation, capable of handling a broad array of NLP tasks with minimal task-specific training. The series continued to evolve with studies introducing a chat-based version, GPT-3.5 [8], in 2022 and a larger multimodal model, GPT-4 [9], in 2023. Each iteration refined and expanded upon its predecessors' capabilities and pushed the boundaries of what is possible in NLP.

Similarly, LLaMA [10] models, including the subsequent LLaMA-2 [11] series, were introduced as formidable contenders in the realm of Large Language Models (LLMs),

offering competitive performance while simultaneously being notably efficient, accessible, and especially beneficial in resource-constrained settings. The LLaMA series initially introduced models with 7, 13, 33, and 65 billion parameters, focusing on efficiency and scalability. The LLaMA-2 series presented models with 7, 13, and 70 billion parameters, with each model demonstrating impressive results that rival the performance of GPT-3 models, making the former a primary focus for this comparative study due to their open-source accessibility and efficiency. LLaMA-2 models differentiate themselves with the adoption of a grouped multi-query self-attention mechanism and root mean square layer normalization, diverging from the standard self-attention and layer normalization used in GPT models. These modifications, along with the incorporation of SwiGLU activation in their feedforward blocks, reduce computational complexity while offering nuanced enhancements to model efficiency and processing capabilities. Figure 1 illustrates the architectural distinctions between the two model series.

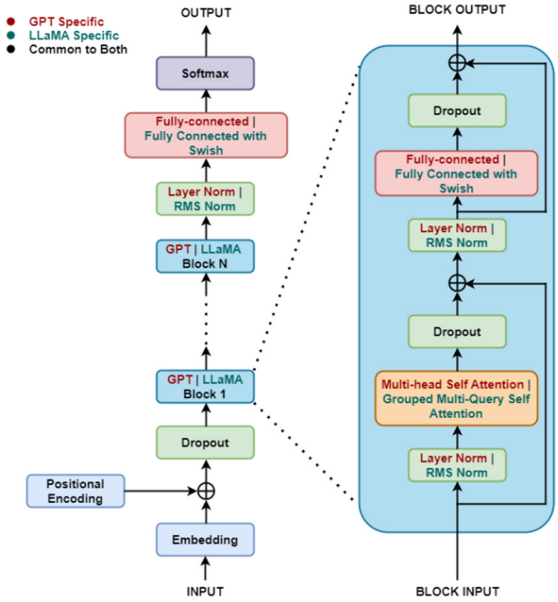

Fig. 1. Basic architecture of GPT and LLaMA models with differences.

Addressing the literature gap on a comprehensive comparison between the GPT and LLaMA-2 models in text classification, this study performs a systematic analysis of fine-tuned GPT-2, GPT-3, GPT-3.5, and LLaMA-2 (7B, 13B, and 70B) models in a variety of classification tasks. Through this exploration, this study aims to discern the performance dynamics of the particular models across different dataset sizes, examine the impact of model size on learning rates, and evaluate their effectiveness in specific classification scenarios. This investigation focuses on key research questions that delve into the adaptability and robustness of the models to provide a nuanced guide for choosing optimal models for NLP tasks. This analysis is performed by positing two hypotheses. The null hypothesis (H0) does not suggest a significant difference in performance between the fine-tuned GPT and LLaMA-2 models, whereas the alternative hypothesis (HA) indicates a significant difference.

Furthermore, this study compares the performance of the fine-tuned GPT and LLaMA-2 models against baseline models in the Hate Speech and Offensive Language (HSOL) [12] and Stanford Sentiment Treebank 2 (SST-2) [13] datasets. For the HSOL dataset, the BERT-HateXplain baseline integrates human rationale into training, improving performance metrics and reducing biases, yet it exhibits a trade-off between performance and explainability. For the SST-2 dataset, baseline models include BERT Single Task [14], which leverages bidirectional contexts for high performance across NLP benchmarks but is trained specifically for single tasks, XLNet single task [14], which utilizes an autoregressive approach for a nuanced understanding of language. and XLNet multi-task [14], which enhances XLNet's performance by training on multiple NLP tasks simultaneously. A baseline for the Fake News Classification (FNC) dataset for text classification was not available and was not included.

II. METHODOLOGY

This study was designed to comprehensively analyze and compare the performance of various LLMs, specifically the GPT and LLaMA-2 series, in the context of dataset size, model scale, and diversity of tasks.

A. Model Selection

The models were selected based on several criteria, focusing on their relevance, representativeness, and the feasibility of obtaining meaningful comparative data. The following models were chosen for the comparative analysis:

- GPT-2, GPT-3, and GPT-3.5 from the GPT series: These models represent significant milestones in the evolution of the GPT series and are indicative of the advancements in LLM capabilities over time. GPT-2 represents a major leap from its predecessor, demonstrating substantial improvements in language understanding and generation. GPT-3, with its unprecedented scale, marks a significant step in the field, displaying its ability to perform a variety of language-related tasks. GPT-3.5, while not constituting a leap as GPT-3, it refines and improves the capabilities of GPT-3, offering enhanced performance in nuanced language understanding.
- LLaMA-2 7B, 13B, and 70B from the LLaMA-2 series: The LLaMA-2 series models were chosen for their efficiency and scalability. Each of these models (7B, 13B, and 70B) represents different points in the spectrum of computational efficiency and language processing capability, which makes them suitable for comparison with the GPT series models. The LLaMA-2 models are particularly relevant in scenarios where computational resources are a constraint.

GPT-1 and GPT-4, along with the original LLaMA models, were excluded from this analysis. GPT-1 was omitted due to its outdated capabilities [15] in relation to the newer iterations, which offer a more current representation of LLM advancements. The exclusion of GPT-4 was due to its recent introduction and the scarcity of detailed performance data. The

initial LLaMA models, being foundational and without fine-tuning capabilities, were also not considered.

B. Datasets

Various text classification datasets were employed to evaluate the performance of fine-tuned GPT and LLaMA-2 models. The HSOL dataset includes tweets classified into hate speech, offensive language, or neither. The FNC dataset, from the 2017 Fake News Challenge [16], contains articles labeled as agree, disagree, discuss, or unrelated, testing the ability to identify misinformation. The SST-2 dataset, known for sentiment analysis, features movie reviews with positive or negative sentiments. Table I summarizes these datasets.

TABLE I. SUMMARY OF THE DATASETS

Dataset	Train samples	Validation samples	Test samples	Classes
HSOL multi-class	20,000	2,500	5,000	3
HSOL binary-class	20,000	2,500	5,000	2
FNC multi-class	14,000	1,500	3,000	4
FNC multi-class	14,000	1,500	3,000	2
SST-2	67,349	872	1,821	2

C. Baselines

This study compared the performance of the fine-tuned GPT and LLaMA-2 models against established baselines on the SST-2 and HSOL datasets. Table II details the accuracies of these baseline models, highlighting their performance benchmarks.

D. Data Preparation

Diverse datasets were used to evaluate the models across various domains and complexities, preprocessing them to remove extraneous characters and convert text into numerical formats suitable for models. Tokenization was performed deploying the pre-training tokenizers of the GPT and LLaMA-2 models to ensure consistency.

TABLE II. PERFORMANCE OF BASELINE MODELS ON HSOL AND SST-2 DATASETS

Model	Dataset	Accuracy
BERT-HateXplain [12]	HSOL (multi-class)	0.698
BERT Single task [14]	SST-2 (binary-class)	0.932
XLNet Single task [14]	SST-2 (binary-class)	0.956
XLNet Multitask [14]	SST-2 (binary-class)	0.968

1) Data Size for Fine-Tuning

Considering the substantial capabilities of both models, this study delved into their adaptability and performance across varying dataset sizes to closely examine how the performance of the GPT and LLaMA-2 models scales with the volume of data provided during the fine-tuning phase. Except for applying the complete datasets, experiments were performed by training these models using 500, 1000, and 2500 examples per class. This resulted in the creation of 24 different models for each task, tailored to each of the GPT and LLaMA-2 model types under investigation. The specific analysis seeks to illuminate the trade-off between the dataset size and model performance, offering valuable insights into the capacity of both models to effectively learn from limited labeled examples.

2) Handling Multi-Class and Binary Classification

The HSOL dataset is categorized into three classes, and the FNC dataset into four classes. To explore binary and multiclass classification capabilities, these datasets were simplified by merging classes: in HSOL, hate speech and offensive language were combined into one class, while in FNC, discusses and agrees were merged, and unrelated and disagrees formed another. This allowed the evaluation of model effectiveness across binary and multi-class tasks, providing insights into how the two models adapt to different classification challenges.

3) Fine-Tuning Process

The GPT-2 experiments were performed on a computer with NVIDIA RTX 3060 GPUs. LLaMA-2 experiments were conducted on Google Colab [17] with T4 and A100 GPUs. Each model was separately fine-tuned on the prepared datasets.

- GPT series: At first, the pre-trained version from Hugging Face was used to fine-tune the GPT-2 model. For fine-tuning, the dataset was prepared for straightforward input and output processing, allowing direct learning from the tailored dataset. The fine-tuning process for GPT-3 and GPT-3.5 involved uploading custom JSONL files for the requirements of each model to the OpenAI portal [18]. Due to their exclusivity, these models can only be fine-tuned through OpenAI's dedicated platform, designed to efficiently manage and train on custom datasets. After uploading the data, the training began employing the robust computational resources of the portal to effectively adapt the models to the datasets. After completing training, the performance of the fine-tuned models was evaluated, certifying that they met the research objectives.
- LLaMA-2 series: The fine-tuning process for all LLaMA-2 models utilized the QLoRA [19] technique. The LLaMA-2 models and datasets were initialized with precise configurations for quantization and precision. Employing QLoRA, a method adept at optimizing neural network weights through low-rank matrices, the LLaMA-2 7B and 13B models were fine-tuned on Google Colab with the free T4 GPU and the 70B with the A100 GPU. After fine-tuning, the performance of the models was evaluated through inference and merging the fine-tuned model with adapter weights, thus ensuring a detailed and efficient enhancement of the capabilities of LLaMA-2 models.

4) Evaluation Metrics

Standard evaluation metrics, such as accuracy and F1-score were engaged to assess the effectiveness of fine-tuned models. For multiclass classification tasks, a weighted version of these metrics was used. These metrics provide a detailed performance evaluation of these models in classifying the text.

5) Statistical Analysis

A one-way Analysis of Variance (ANOVA) test was performed to rigorously compare the mean performance scores

of the fine-tuned GPT and LLaMA-2 models. This statistical test can determine whether the observed performance differences are statistically significant.

III. RESULTS AND ANALYSIS

The findings are structured to reflect the impacts of dataset size, model size, and task diversity on model performance, and conclude with insights into the selection of the optimal model for NLP tasks and comparison with the baseline models. Table III summarizes the performance of the models on the HSOL and SST-2 datasets, and Table IV presents the results for the FNC dataset. Figure 2 and Figure 3 portray the performance of GPT and LLaMA-2 models on the HSOL and FNC datasets, respectively, across binary and multiclass tasks, whereas Figure 4 displays the performance on the SST-2 datasets for binary classification tasks.

TABLE III. PERFORMANCE OF FINE-TUNED GPT AND LLAMA-2 MODELS ON HSOL AND SST-2 DATASET

Model	HSOL (Binary-class)		HSOL (Multi-class)		SST-2 (Binary-class)	
	Acc	***F1***	***Acc***	***F1***	***Acc***	***F1***
GPT-2 FT 500	0.7832	0.7754	0.6961	0.6902	0.7441	0.7268
GPT-3 FT 500	0.9192	0.9121	0.8066	0.7896	0.8784	0.8698
GPT-3.5 FT 500	0.9694	0.9625	0.9624	0.9577	0.9001	0.8908
LLaMA-2 7B FT 500	0.9150	0.9028	0.8092	0.7953	0.8832	0.8785
LLaMA-2 13B FT 500	0.9249	0.9187	0.8475	0.8408	0.8906	0.8836
LLaMA-2 70B FT 500	0.9259	0.9235	0.9278	0.9249	0.8979	0.8948
GPT-2 FT 1000	0.8206	0.8056	0.7470	0.7301	0.8873	0.8722
GPT-3 FT 1000	0.9298	0.9137	0.8513	0.8403	0.9087	0.8940
GPT-3.5 FT 1000	0.9784	0.9733	0.9709	0.9692	0.9284	0.9246
LLaMA-2 7B FT 1000	0.9301	0.9211	0.8624	0.8542	0.9111	0.9048
LLaMA-2 13B FT 1000	0.9382	0.9300	0.8939	0.8882	0.9136	0.9084
LLaMA-2 70B FT 1000	0.9526	0.9502	0.9613	0.9595	0.9221	0.9198
GPT-2 FT 2500	0.8354	0.8276	0.7729	0.7661	0.9223	0.9092
GPT-3 FT 2500	0.9336	0.9286	0.9061	0.8892	0.9437	0.9357
GPT-3.5 FT 2500	0.9797	0.9735	0.9743	0.9706	0.9745	0.9708
LLaMA-2 7B FT 2500	0.9364	0.9274	0.9184	0.9118	0.9587	0.9496
LLaMA-2 13B FT 2500	0.9514	0.9447	0.9364	0.9303	0.9598	0.9516
LLaMA-2 70B FT 2500	0.9753	0.9710	0.9735	0.9701	0.9721	0.9683
GPT-2 FT C	0.9266	0.9129	0.9011	0.8914	0.9757	0.9624
GPT-3 FT C	0.9821	0.9743	0.9800	0.9728	0.9812	0.9772
GPT-3.5 FT C	**0.9875**	**0.9807**	**0.9829**	**0.9790**	**0.9942**	**0.9910**
LLaMA-2 7B FT C	0.9832	0.9762	0.9814	0.9745	0.9824	0.9783
LLaMA-2 13B FT C	0.9862	0.9782	0.9803	0.9759	0.9863	0.9802
LLaMA-2 70B FT C	*0.9866*	*0.9793*	*0.9816*	0.9773	*0.9923*	*0.9885*

A. Impact of Dataset Size on the Performance of the Models

The results demonstrate a clear correlation between dataset size and improved performance across all models. As dataset size increases, there is a notable enhancement in both accuracy and F1 scores, a trend consistent across the HSOL, SST-2, and FNC datasets. The GPT-3.5 and LLaMA-2 70B models show the most substantial improvements. For instance, the F1-score for the binary classification task in the HSOL, FNC, and SST-2 datasets increased from 0.9625 to 0.9807, from 0.9688 to 0.9832, and from 0.8908 to 0.9910, accordingly, for the GPT-3 model. Similarly, for multiclass tasks, the F1-score increased from 0.9577 to 0.9790 on HSOL and from 0.9600 to 0.9801 on FNC, indicating that all models' performance improved as the dataset size increased, with LLaMA-2 70B closely following. Additionally, the performance on multiclass tasks closely matches that of binary tasks for both the HSOL and FNC datasets. This performance improvement with larger dataset sizes highlights the learning efficacy of models and emphasizes the importance of ample and diverse training data for optimizing outcomes.

B. Impact of Model Size on the Learning Rate of the Models

The transition from GPT-2 to GPT-3.5, as well as the scaling of LLaMA-2 models from 7B to 70B, underscores the profound impact of model size on learning capabilities. In particular, larger models, such as GPT-3.5 and LLaMA-2 70B show superior performance across datasets and tasks, which can be attributed to their advanced architectural complexity and heightened parameter counts. These models exhibit an accelerated learning rate, which enables them to more effectively grasp complex patterns and subtleties within the text compared to their smaller counterparts. For instance, the accuracy for GPT-2 on the HSOL dataset with a full dataset size for the binary classification task is 0.9266, which significantly improves to 0.9821 for GPT-3 and further increases to 0.9875 for GPT-3.5. Similarly, for LLaMA-2 models, the accuracy for the 7B model is 0.9832, which increases to 0.9862 for the 13B model and further increases to 0.9866 for the 70B model. Furthermore, the performance of all models in multiclass classification tasks is closely aligned with that observed in binary classification tasks. This trend emphasizes that architectural and parameter enhancements in larger models are pivotal for boosting their abilities in NLP tasks.

TABLE IV. PERFORMANCE OF FINE-TUNED GPT AND LLAMA-2 MODELS ON THE FNC DATASET

Model	FNC (Binary-class)		FNC (Multi-class)	
	Acc	***F1***	***Acc***	***F1***
GPT-2 FT 500	0.8043	0.7920	0.7146	0.7017
GPT-3 FT 500	0.8435	0.8365	0.8396	0.8284
GPT-3.5 FT 500	0.9711	0.9688	0.9649	0.9600
LLaMA-2 7B FT 500	0.8695	0.8499	0.8441	0.8356
LLaMA-2 13B FT 500	0.8723	0.8586	0.8489	0.8376
LLaMA-2 70B FT 500	0.9707	0.9632	0.9637	0.9593
GPT-2 FT 1000	0.8461	0.8311	0.8458	0.8302
GPT-3 FT 1000	0.9252	0.9134	0.9001	0.8756
GPT-3.5 FT 1000	0.9751	0.9694	0.9747	0.9738
LLaMA-2 7B FT 1000	0.9268	0.9078	0.9178	0.9067
LLaMA-2 13B FT 1000	0.9295	0.9158	0.9207	0.9023
LLaMA-2 70B FT 1000	0.9750	0.9682	0.9740	0.9653
GPT-2 FT 2500	0.9393	0.9125	0.9073	0.8809
GPT-3 FT 2500	0.9722	0.9641	0.9387	0.9115
GPT-3.5 FT 2500	0.9817	0.9734	0.9785	0.9688
LLaMA-2 7B FT 2500	0.9746	0.9592	0.9417	0.9269
LLaMA-2 13B FT 2500	0.9792	0.9680	0.9497	0.9406
LLaMA-2 70B FT 2500	0.9804	0.9760	0.9741	0.9672
GPT-2 FT C	0.9792	0.9642	0.9333	0.9128
GPT-3 FT C	0.9839	0.9731	0.9796	0.9680
GPT-3.5 FT C	**0.9898**	*0.9832*	**0.9847**	**0.9801**
LLaMA-2 7B FT C	0.9852	0.9736	0.9798	0.9679
LLaMA-2 13B FT C	0.9861	0.9781	0.9821	0.9772
LLaMA-2 70B FT C	*0.9883*	**0.9844**	*0.9838*	*0.9798*

C. Impact of Dataset Diversity on the Efficiency of the Models

The diversity of tasks represented by the HSOL, SST-2, and FNC datasets provides a comprehensive view of model performance across varied NLP challenges. The results reveal that both the GPT and LLaMA-2 models excel at adapting to different tasks, with the largest models of both series demonstrating the most remarkable effectiveness. Specifically, for the HSOL binary classification task, the highest F1 scores achieved were 0.9807 for GPT-3.5 and 0.9793 for LLaMA-2 70B. In the multiclass scenario, these models obtained F1 scores of 0.9790 and 0.9773, respectively. For the FNC dataset, in binary classification, GPT-3.5 attains an F1 score of 0.9832, and LLaMA-2 70B scores 0.9844, whereas in multiclass tasks, the scores are 0.9801 and 0.9798, correspondingly. On the SST-2 dataset, GPT-3.5 and LLaMA-2 70B achieved F1 scores of 0.9910 and 0.9885, underscoring their prowess in binary and multiclass classification tasks. This adaptability to dataset diversity underscores the versatility and robustness of the models, confirming their ability to generalize effectively across different contexts and domains, accomplishing high levels of performance regardless of the task at hand.

Fig. 2. Comparison of the performance of GPT and LLaMA-2 models on the HSOL dataset for binary and multiclass classification tasks.

Fig. 3. Comparison of the performance of GPT and LLaMA-2 models on the FNC dataset for binary and multi-class classification tasks.

D. Selection of Optimal Model for NLP Tasks

The results disclose that model size critically influences the training requirements, learning capabilities, and effectiveness of the language models. Within the GPT series, GPT-3 and GPT-3.5 outperformed GPT-2, with GPT-3.5 showing remarkable efficiency even from smaller datasets but at the cost of higher computational resources. In the LLaMA-2 series, LLaMA-2 13B performs better than LLaMA-2 7B, whereas LLaMA-2 70B performs better than both the LLaMA-2 7B and 13B models. The LLaMA-2 13B model, in particular, offers a superior balance between computational efficiency and effectiveness, outperforming GPT-2 and GPT-3 and closely rivaling GPT-3.5. Given its performance across various tasks and its moderate computational demands, LLaMA-2 13B emerges as the optimal choice for NLP tasks.

Fig. 4. Comparison of the performance of GPT and LLaMA-2 models on the SST-2 dataset for binary classification tasks.

The narrowing performance gap between LLaMA-2 13B and GPT-3.5 with increasing the dataset size, coming within 1% of GPT-3.5's performance while consistently surpassing that of the GPT-3 model, further underscores the LLaMA-2 13B's efficiency. Its consistent performance, closely mirroring LLaMA-2 70B and GPT-3.5 across the full dataset, combined with the feasibility of fine-tuning on freely accessible platforms, such as Google Colab's T4 GPU or training on A100 GPU, showcases its utility and accessibility. This is in contrast to the LLaMA-2 70B model, which demands more A100 GPUs, and GPT-3.5, which is restricted to fine-tuning via its exclusive portal, presenting barriers to wider accessibility. Therefore, despite the top-tier performance of GPT-3.5 and LLaMA-2 70B, the LLaMA-2 13B model emerges as the most practical and efficient choice, especially when resource limitations are considered. With just 13 billion parameters, it provides an ideal compromise between high-level performance, task adaptability, and manageable resource requirements, which establishes it as an excellent option for NLP tasks.

The substantiation of these observations comes from the results of a one-way ANOVA test, which revealed an F-statistic of 14.96 and a p-value of 0.0043, indicating a statistically significant difference in performance between the GPT and LLaMA-2 models. This significant variance, notably with LLaMA-2 models surpassing GPT models by more than 5% for the 13B and 70B variants compared to GPT-2 and more than 3% against GPT-3, leads to the rejection of the null hypothesis. This statistical confirmation emphasizes the advantage of LLaMA-2 models and further solidifies the position of LLaMA-2 13B as the optimal selection for NLP tasks.

IV. CONCLUSION

This study presented a detailed comparative analysis of the latest GPT and LLaMA-2 models, focusing on their adaptability and performance across various dataset sizes, model scales, and classification tasks. This investigation

highlights the nuanced impact of dataset size, model scale, and task diversity on model performance, with a keen emphasis on identifying the most efficient model for NLP tasks. The results revealed a clear correlation between the size of the dataset and the improvement in model performance, with larger datasets significantly increasing both accuracy and F1 scores across all models. For example, GPT-3.5 reached accuracies of 0.9694, 0.9784, 0.9797, and 0.9875 for dataset sizes 500, 1000, 2500, and complete, respectively, on the HSOL dataset for the binary classification task. This emphasizes the critical role of data volume in optimizing model outcomes. Additionally, the study highlighted the impact of model size on learning capabilities, with larger models, such as GPT-3.5 and LLaMA-2 70B demonstrating superior performance due to their increased architectural complexity and parameter counts. For instance, the smaller LLaMA-2 7B model achieved an accuracy of 0.9832, the medium-sized LLaMA-2 13B model achieved 0.9862, while the larger LLaMA-2 70B model reached 0.9866 on the HSOL dataset for the binary classification task. Additionally, this analysis across a broad spectrum of NLP challenges confirmed the adaptability of both the GPT and LLaMA-2 models, especially the largest models, to effectively handle a diverse range of tasks. Furthermore, the performance remained high for both binary and multiclass classification tasks, highlighting the adaptability of the models. Among the models evaluated, the LLaMA-2 13B emerged as the most balanced option for NLP tasks, striking an optimal balance between computational efficiency, performance, and adaptability. This model can be fine-tuned on freely available GPUs, such as T4 on Google Colab, with techniques like QLoRA, which renders it an attractive choice for researchers and developers in NLP.

This study bridges a crucial gap in NLP research, providing a thorough comparative analysis of the adaptability and efficiency of the latest LLMs in text classification. The significance of this study lies in pinpointing the LLaMA-2 13B model as the optimal choice for a broad spectrum of NLP tasks, underscoring a paradigm shift towards models that meld high performance with resource efficiency. This insight is pivotal for effectively harnessing advanced NLP models, particularly in contexts with constrained computational resources. Future research directions include exploring hybrid architectures, leveraging instruction tuning, and evaluating model performance on cross-lingual tasks. Addressing computational challenges through techniques like pruning and knowledge distillation can enhance the efficiency and accessibility of NLP technology, paving the way for broader applications and impacts of AI and machine learning in NLP.

REFERENCES

[1] E. Yilmaz and O. Can, "Unveiling Shadows: Harnessing Artificial Intelligence for Insider Threat Detection," *Engineering, Technology & Applied Science Research*, vol. 14, no. 2, pp. 13341–13346, Apr. 2024, https://doi.org/10.48084/etasr.6911.

[2] A. Kazm, A. Ali, and H. Hashim, "Transformer Encoder with Protein Language Model for Protein Secondary Structure Prediction," *Engineering, Technology & Applied Science Research*, vol. 14, no. 2, pp. 13124–13132, Apr. 2024, https://doi.org/10.48084/etasr.6855.

[3] R. Sharma, S. Deol, U. Kaushish, P. Pandey, and V. Maurya, "DWAEF: a deep weighted average ensemble framework harnessing novel indicators for sarcasm detection 1," *Data Science*, vol. 6, no. 1–2, pp. 17–44, Jan. 2023, https://doi.org/10.3233/DS-220058.

[4] K. A. Aldriwish, "Empowering Learning through Intelligent Data-Driven Systems," *Engineering, Technology & Applied Science Research*, vol. 14, no. 1, pp. 12844–12849, Feb. 2024, https://doi.org/10.48084/etasr.6675.

[5] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving Language Understanding by Generative Pre-Training." [Online]. Available: https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf.

[6] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language Models are Unsupervised Multitask Learners," OpenAI, San Francisco, CA, USA.

[7] T. Brown *et al.*, "Language Models are Few-Shot Learners," in *Advances in Neural Information Processing Systems*, 2020, vol. 33, pp. 1877–1901.

[8] "OpenAI Platform: GPT-3.5 Turbo." https://platform.openai.com/docs/models/gpt-3-5-turbo.

[9] "GPT-4 Technical Report," OpenAI, San Francisco, CA, USA, 2023. [Online]. Available: https://cdn.openai.com/papers/gpt-4.pdf.

[10] H. Touvron *et al.*, "LLaMA: Open and Efficient Foundation Language Models." arXiv, Feb. 27, 2023, https://doi.org/10.48550/arXiv.2302.13971.

[11] H. Touvron *et al.*, "Llama 2: Open Foundation and Fine-Tuned Chat Models." arXiv, Jul. 19, 2023, https://doi.org/10.48550/arXiv.2307.09288.

[12] B. Mathew, P. Saha, S. M. Yimam, C. Biemann, P. Goyal, and A. Mukherjee, "HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 17, pp. 14867–14875, May 2021, https://doi.org/10.1609/aaai.v35i17.17745.

[13] R. Socher *et al.*, "Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank," in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, Seattle, WA, USA, pp. 1631–1642.

[14] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le, "XLNet: generalized autoregressive pretraining for language understanding," in *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, Vancouver, Canada, Sep. 2019, pp. 5753–5763.

[15] C. Bekar, K. Carlaw, and R. Lipsey, "General purpose technologies in theory, application and controversy: a review," *Journal of Evolutionary Economics*, vol. 28, no. 5, pp. 1005–1033, Dec. 2018, https://doi.org/10.1007/s00191-017-0546-0.

[16] "Fake News Challenge." http://www.fakenewschallenge.org/.

[17] "Google Colaboratory." https://colab.research.google.com/.

[18] "Fine-tuning - OpenAI API." https://platform.openai.com/docs/guides/fine-tuning.

[19] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "QLoRA: Efficient Finetuning of Quantized LLMs." arXiv, May 23, 2023, https://doi.org/10.48550/arXiv.2305.14314.