ISSN 1816-6075 (Print), 1818-0523 (Online)
Journal of System and Management Sciences
Vol. 14 (2024) No. 10, pp. 254-266
DOI:10.33168/JSMS.2024.1015

Methods and Applications of Fine-Tuning Llama-2 and Llama-Based Models: A Systematic Literature Analysis

Vincencius Christiano Tjokro, Samuel Ady Sanjaya*

Department of Information Systems, Faculty of Engineering and Informatics, Universitas Multimedia Nusantara, Banten, Indonesia

vincencius@student.umn.ac.id, samuel.ady@umn.ac.id (Corresponding Author)

Abstract. This Systematic Literature Review (SLR) scrutinizes 20 papers focusing on methods and applications for fine-tuning the open-source Llama language model. The findings underscore the efficacy of instruction-based tuning, demonstrating high performance across diverse domains, including medicine and psychology. The fine-tuned version Llama-2-Chat model, aligned with human preferences, emerges as a preferred base for subsequent tuning efforts. Despite its promise, scalability hurdles persist due to demanding computational requirements. A critical gap in the existing literature lies in the absence of a balanced evaluation of trade-offs among various fine-tuning approaches. Moreover, ethical considerations, particularly addressing bias and associated risks, demand greater attention in the deployment of tuned models. In conclusion, while instruction tuning holds significant promise for specializing Llama variants, overcoming limitations related to resource constraints, safety, and transparency is imperative for responsible real-world impact. This refined understanding emphasizes the need for a comprehensive assessment of fine-tuning methods and a conscientious approach to ethical considerations in the evolving landscape of large language models.

Keywords. Fine-Tuning, Llama, Llama-2, Large Language Model

1. Introduction

Chatbots are crafted to identify user input through pattern matching and retrieve information from a database to furnish a suitable response. Frequently employed to aid human tasks in disseminating information, chatbots find applications in various domains such as virtual learning systems and customer service (Richardson & Wicaksana, 2022). An example of chatbot technology involves the utilization of Large Language Models (LLMs). Large language model (LLM) refers to a language model built upon a Transformer by training huge corpus data and containing billions of parameters. This ability is inherited by LLM upon leveraging huge text training data and billions of parameters related to the neural scaling law theory (Zhao et al., 2023). Neural scaling law theory explains that the performance of LLM is highly related to the size of pre-training data, model parameters, and computational capability involved while training the model (Aghajanyan et al., 2023; Young, 2021). The advancement of LLM has experienced exponential growth in the past 5 years, marked by the emergence of ChatGPT by OpenAI in November 2022, reaching one million users within just 5 days (De Angelis et al., 2023). Instruction LLM such as ChatGPT inhibits an impressive generalization ability to comprehend instructions and produce human-like responses even on unseen cases or tasks (Wang, Liu, et al., 2023). This ability is inherited by ChatGPT due to its enormous 175B parameters, this comes with a limitation that makes it computationally expensive, furthermore, it is only accessible through API provided by OpenAI (*close-source*) (Ray, 2023). As an alternative to address those limitations, smaller LLM has emerged, marked by the introduction of Llama as the state-of-the-art (SOTA) open-source LLM.

Fine-tuning refers to the process of further training pre-trained models on specific domain data. There are several approaches of fine-tuning including transfer learning, instruction fine-tuning, alignment fine-tuning, and parameter-efficient fine-tuning. Transfer learning is an approach of fine-tuning pre-trained on specific and smaller task-specific data. Whereas, the instruction fine-tuning objective is to train a pre-trained model to respond based on prompt and user input by leveraging an instruction formatted dataset (consisting of instruction and input-output). On the other hand, alignment fine-tuning is a fine-tuning approach to align LLM based on human intentions and ethical values to prevent bias and harmful text by employing RLHF. Finally, LLM has huge parameters which make it computationally extensive for training, to address this limitation parameter efficient fine-tuning was introduced as a fine-tuning technique that works by either adding existing parameters or new ones while the rest of the model held frozen during training (Naveed et al., n.d.).

Understanding the potential inherited by Llama as an open-source SOTA LLM and the ability to fine-tune to train the pre-trained model on a specific task, this systematic literature review aim to further understand methods and applications of fine-tuning leveraging Llama-based model on various domain downstream tasks.

2. Methods

In order to assess further various methods and applications of fine-tuning Llama, a systematic literature review is being conducted. Systematic literature review (SLR) is a process of reviewing previous studies to identify new research opportunities. In this study Search, Appraisal, Synthesis, and Analysis (SALSA) are being implemented as a methodology framework for performing SLR. Search is the first phase in which previous studies are gathered from the literature database, whereas Appraisal is a phase in which collected papers are assessed and screened based on the relevancy to research questions, and Synthesis and Analysis is the last phase in which data are being extracted from selected papers and conclusions are being drawn. In the subsequent section, a more detailed breakdown of the methodologies employed will be presented (García-Holgado et al., 2020).

2.1. Research Questions

Understanding Llama-2 as state of art (SOTA) open-source large language model developed Meta and at the moment this paper is written there is no systematic review that specifically discusses fine-tuning

llama or its derivative. Therefore with an objective to understand various methods and applications in fine-tuning llamas, we formed several research questions as mentioned:

a. (RQ1) Which Llama-based model should be used for fine-tuning purposes?
b. (RQ2) What are the methods for fine-tuning Llama based model?
c. (RQ3) What framework was used for fine-tuning Llama based model effectively?
d. (RQ4) What are the applications of a fine-tuned Llama-based model?
e. (RQ5) How to evaluate fine-tuned Llama-based model performance?

2.2. SALSA Application

A systematic literature screening process is being conducted following the SALSA framework. A flowchart of systematic literature can be seen in Fig 1. The search phase is done by implementing an academic literature retrieval software, Publish or Perish (POP), with Google Scholar as the literature database. Since Llama is being released in the first quarter of 2023, the year of paper published is limited to only 2023. The following keyword strings are being used to properly select paper: (*'Large language model' OR 'LLM' AND 'Fine-Tuning' AND 'Llama-2'*). 500 papers are successfully being gathered in the Search phase. The search phase is done by implementing an academic literature retrieval software, Publish or Perish (POP), with Google Scholar as the literature database. Since Llama is being released in the first quarter of 2023, the year of paper published is being limited to only 2023. The following keyword strings are being used to properly select paper: (*'Large language model' OR 'LLM' AND 'Fine-Tuning' AND 'Llama-2'*). 500 papers are successfully being gathered in the Search phase.

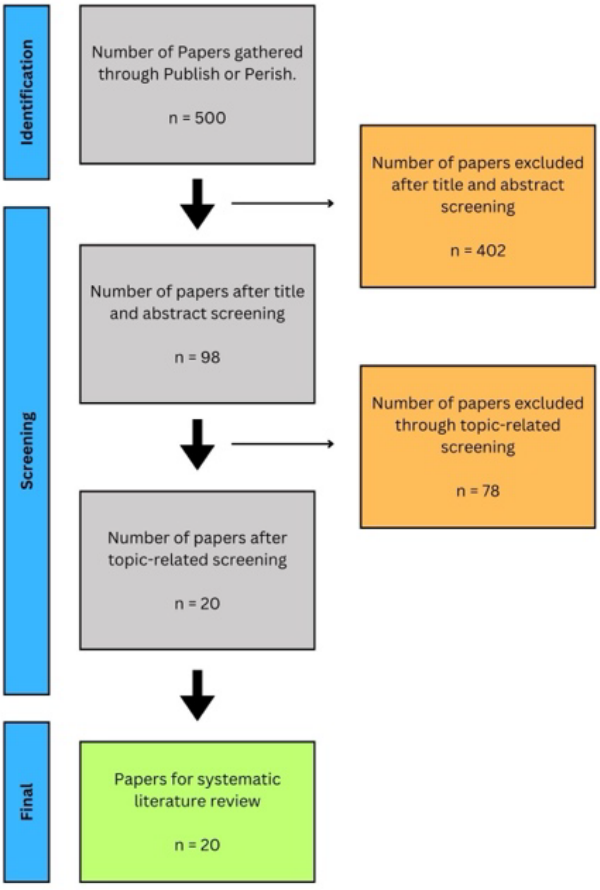

Fig. 1: Flowchart of Systematic Literature Review

The Appraisal Phase represents a crucial step in the systematic review process, serving as the initial filter for the gathered papers. In this meticulous screening process, the selection criteria prioritize the

relevance of titles and abstracts to the specific focus on Llama and its derivatives. The inclusion criteria for the research necessitated the explicit mention of 'Llama' in either the paper title or abstract. As a result of applying these criteria, a total of 402 papers were excluded during the initial screening phase due to their divergence from the specified subject matter. The second screening, involving a comprehensive full-text review, further refines the selection by evaluating the papers against predetermined research questions. This stage specifically encompasses papers focusing on the model, framework, fine-tuning method, and evaluation technique to address predetermined research questions. With a total exclusion of 78 additional papers during this phase, a refined collection of 20 papers is attained, showcasing significant relevance for the subsequent phases of Synthesis and Analysis. In the subsequent Synthesis pPhase, the 20 meticulously chosen papers undergo a thorough examination with the explicit goal of extracting pertinent data to address the established research questions. This phase involves a comprehensive review of the methodologies, results, and conclusions of each paper. The synthesized findings from these studies are then thoughtfully summarized and meticulously presented in Table 1. This synthesis not only facilitates a coherent overview of the selected papers but also lays the foundation for the subsequent analytical phase, where the extracted data will be subjected to in-depth analysis and interpretation.

Table. 1. Research Papers Summaries

Nr.	Ref.	Application	Model	Framework	Method	Evaluation
A	(J. M. Liu et al., 2023)	Psychology	Vicuna-7b-v1.3		Instruction fine-tuning	Auto Evaluation
B	(X. Xu et al., 2023)	Psychology	Alpaca-7b, FLAN-T5		Instruction fine-tuning	Accuracy
C	(Pavlyshenko, 2023a)	News	Llama-2-7B-Chat-Hf	LoRA	Instruction fine-tuning	Human Evaluation
D	(Luu & Buehler, 2023)	Biology	Llama-2-13B-Chat-Hf	LoRA	Instruction fine-tuning	Human Evaluation
E	(Zhong et al., 2023)	Medical	Llama-2-7B, Llama-2 -7B-Chat-Hf, Llama-2-7B-Chinese-Chat, ChatGLM2-6B	LoRA	Instruction fine-tuning	ROGUE, Human Evaluation
F	(Yang, Zhang, Kuang, Xie, & ..., 2023)	Psychology	Llama-2-7B, Llama-2-7B-Chat-Hf, Llama-2-13B-Chat-Hf		Completion fine-tuning, Instruction fine-tuning	Accuracy, Weighted F1, Bart-Score
G	(H. Xu et al., 2023)	Translation	Llama-2-7B Llama-2-13B	LoRA	Instruction fine-tuning	BLEU dan COMET
H	(Wang, Liu, et al., 2023)	Medical	Llama-7B		Instruction fine-tuning	Human Evaluation
I	(Nguyen et al., 2023)	Sexual Assault	Llama-2-7B	LoRA	Instruction fine-tuning	Accuracy, F1
J	(Pavlyshenko, 2023b)	Financial	Llama-2-7B-Chat-Hf	LoRA	Instruction fine-tuning	Human Evaluation
K	(Roziere et al., 2023)	Coding	Llama-2-7B, Llama-2-13B, Llama-2-34B		Instruction fine-tuning	HumanEval, MBPP, APPS
L	(Choi et al., 2023)	Financial	Llama-2-7B	LoRA	Instruction fine-tuning	Accuracy, precision
M	(Khan et al., 2023)	Medical	Llama-7B	Alpaca-LoRA	Instruction fine-tuning	Human Evaluation
N	(Zhang et al., 2023)	Psychology	Llama-2-7B	LoRA	Instruction fine-tuning	F1, accuracy

O	(Wang, Gao, et al., 2023)	Medical	Llama-7B	LoRA	Instruction fine-tuning	F1, Accuracy
P	(Nayak & Timmapathini, 2023)	knowledge Bases	Llama-2-13B-Chat-Hf, StableBeluga-13B	LoRA	Instruction fine-tuning	Precision, Recall, F1
Q	(Tan et al., 2023)	Medical	Baichuan-7B		Instruction fine-tuning	BLEU, GLEU, ROUGE
R	(Yu et al., 2023)	Math	Llama-2-7B, Llama-2-13B, Llama-2-70B	QLoRA (70B model)	Instruction fine-tuning	Accuracy
S	(Z. Liu et al., 2023)	Medical	Llama-2	LoRA	Instruction fine-tuning	ROGUE
T	(Zheng et al., 2023)	Transportation	Llama-7B		Instruction fine-tuning	BLEU, ROUGE, BERTScore, BLEURT

3. Result and Discussions

The analysis Phase of the SALSA framework is being conducted in this section to address predetermined research questions based on the data extracted and summarized presented at Table 2.

a. (RQ1) Which Llama-based model should be used for fine-tuning purposes?

Determining which Llama-based model should be used for further fine-tuning depends on several aspects including commerciality, computational resources, and performance. Commerciality refers to the ability of a fine-tuned model implemented as a tool to generate profit. Llama-1 as a predecessor was published as a non-commercial open-source model, which differs from Llama-2 which was later announced as available for commercial purposes (Zheng et al., 2023). Understanding the commerciality of a project, the Llama-2-based model would be preferable compared to the Llama-1-based model. Apart from that, the availability of computational resources might also be considered in the process to determine which Llama-based model should be used.

Table 2 below shows details of computational resources allocated in each research. Higher model parameters tend to require more computational resource allocation. As a solution to address this limitation, Google provides accessible graphics processing unit (GPU) infrastructure through Google Colaboratory (Colab) which is also being implemented in research C and J. Understanding this circumstance, 7B model parameters would be more preferrable and accessible for further fine-tuning process compared to 13B or 70B model parameters provide by Llama which require more extensive and expensive computational resources.

Table. 2. Computational Resources Allocated

Nr.	Model Params.	Google Collab	GPU	Num. of GPU
F	7B, 13B		Nvidia Tesla A100 80GB	4
L	7B		Nvidia A100	1
G	7B, 13B		Nvidia M1200	16
T	7B		Nvidia Titan RTX	8
M	7B		Nvidia A100	1
Q	7B		Nvidia A800	8
J	7B	✓		
P	13B		Nvidia V100	2
K	7B, 13B, 34B		Nvidia A100	1
B	7b		Nvidia A100 80GB	8
A	7b		Nvidia A100 40GB	8

O	7B		Nvidia RTX A6000 48GB	1
N	Llama-2-7B		Nvidia A100 40GB	1
C	7B	✓		
R	7B, 13B, 70B		Nvidia A100	8

Performance is the last aspect to be considered on selecting the right Llama based model for further fine-tuning process. Table 3 below shows the summarized comparison of each model on several evaluation benchmark (Touvron et al., 2023). It is shown that Llama-2 outperforms Llama-1 performance on all benchmark of evaluations. The fine-tuned version of Llama-2, known as Llama-2-Chat, exhibits a noteworthy enhancement in safety benchmarks, demonstrating an increased capability to generate truthful and informative responses as measured by TruthfulQA. Additionally, it shows a near-zero percentage of toxic response generation, as assessed by the Toxigen benchmark. Research F further indicates that the implementation instruction fine-tuning on Llama-2-7B-Chat-Hf surpasses the performance of Llama-2-7B in terms of correctness and explanation quality generated by the model, as highlighted in the same research. This improved performance is attributed to the RLHF process, aligning with human preferences and implemented by Llama-2-7B-Chat-HF.

Table. 3. Llama Based Model Evaluation

Benchmark	Llama-1			Llama-2			Llama-2-Chat		
	7B	13B	65B	7B	13B	70B	7B	13B	70B
Code	14.1	18.9	30.7	16.8	24.5	37.5	-	-	-
Commonsense Reasoning	60.8	66.1	70.7	63.9	66.9	71.9	-	-	-
World Knowledge	46.2	52.6	60.5	48.9	55.4	63.6	-	-	-
Reading Comprehension	58.5	62.3	68.6	61.3	65.8	69.4	-	-	-
Math	6.95	10.9	30.8	14.6	28.7	35.2	-	-	-
MMLU	35.1	46.9	63.4	45.3	54.8	68.9	-	-	-
BBH	30.3	37.0	43.5	32.6	39.4	51.2	-	-	-
AGI Eval	23.9	33.9	47.6	29.3	39.1	54.2	-	-	-
TruthfulQA	27.42	41.74	48.71	33.29	41.86	50.18	57.04	62.18	64.14
ToxiGen	23.00	23.08	21.77	21.25	26.10	24.60	0.00	0.00	0.01

In summary, considering factors such as commercial viability, allocation of computational resources, and overall model performance, Llama-2-7B-Chat-Hf emerges as a preferable choice for further fine-tuning processes. The Llama-2 model proves valuable for commercial applications, surpassing Llama-1 in terms of performance. Meanwhile, the 7B parameters demand fewer and more accessible computational resources, making them suitable for utilization with platforms like Colab. The fine-tuned version, Llama-2-Chat, inherits the RLHF process, contributing to enhanced model performance when compared to the baseline Llama-2-7B model. It's crucial to note that the model's performance demonstrates a linear alignment with the number of parameters, implying that as parameter count increases, so does the generated performance by the model.

b. (RQ2) What are the methods for fine-tuning the Llama-based model?

Two methods of fine-tuning are being introduced as shown in Table 1, instruction and completion fine-tuning. Completion fine-tuning refers to the process of fine-tuning by implementing completion-based data, completion-based data structured of prompt and response (Yang, Zhang, Kuang, Xie, Ananiadou, et al., 2023). Whereas, instruction fine-tuning implements instruction-based data, structured of

instruction, and input-output pair (Wang, Liu, et al., 2023). Yet, Research F shows that utilizing completion fine-tuning on Llama-2 is inefficient in both capability and cost, compared to effective improvement of explanation quality generation capability through instruction fine-tuning on the Llama-2-Chat model, outperforming Llama-2 as a vanilla model.

Moreover, the implementation of instruction fine-tuning on Llama based model demonstrated by research I showed an outstanding capability for the model to learn Urdu and Roman Urdu as a non-English language utilizing only a small sample of data, outperforming traditional text classification approach such as CNN, LSTM, LogitBoost, etc. Further research marked an incredible result in addressing specific domain limitations faced by LLM especially Llama through the implementation of instruction fine-tuning. First research T proposed TrafficSafetyGPT, a fine-tuned Llama-based model on transportation safety domain. Second research D proposed BioInspiredLLM, Llama-based conversational LLM fine-tuned to specialize in structural biological materials, outperforming base model and showing promising capability to accelerate researcher in this field either by generating datasets or grouping/clustering tasks. Another notable example of domain-specific task fine-tuning demonstrated in research E which proposes ChatRadio-Valuer, Llama-based LLM fine-tuned specifically for radiology report generation and surpassing other SOTA LLM performance. However, the diversity of instruction datasets used for fine-tuning is critical to ensure the model's capability to generate responses on unseen tasks (Wang, Liu, et al., 2023). Research M strengthens that a handful of diverse instruction data even synthetically generated one could accomplish robustness comparable or exceeded to ChatGPT performances.

Effectives of instruction fine-tuning crucially relies on the quality of instruction data, this arise several limitations. First the development or collection of every possible instruction in specific task of domain could be extensive and expensive due to huge possibilities of instruction sequences, resulting in incomplete optimization of instruction. Finally, instruction fine-tuned model might have struggled to answer tasks which lacks a definitive right answer, especially in creative tasks where providing adequate examples of instruction is difficult or impossible (Ouyang et al., 2022). Additionally, instruction fine-tuning carries risks like unintentional biases and ethical concerns. These challenges may compromise the model's reliability and trustworthiness, and there's also a threat of malicious manipulation. This emphasized the need for strong safeguards and ethical guidelines during both the development and deployment of instruction fine-tuned models. Especially for AI developers who aimed to release the weights of fine-tuned model publicly, due to irreversibility of this action (Lermen et al., 2023). In conclusion, while instruction fine-tuning demonstrates significant capabilities for the Llama-based model across specific domain applications and non-English tasks, it faces challenges. Issues such as the extensive and expensive collection of diverse instructions, particularly in creative tasks, pose optimization limitations. Additionally, there are risks of unintentional biases and ethical concerns, highlighting the need for robust safeguards and ethical guidelines throughout the development and deployment of instruction fine-tuned models.

c. (RQ3) What framework is used for fine-tuning Llama based model effectively?

Three frameworks for fine-tuning are being introduced in Table 1 including Low-Rank Adaptation (LoRA), Alpaca-LoRA, and Quantized Low-Rank Adaptation (QLoRA). These three frameworks are LoRA-based fine-tuning frameworks categorized as Parameter Efficient Fine-Tuning (PEFT). PEFT approach fine-tuning only a small number of parameters, resulting in more optimum computational resource allocation and cost compared to full fine-tuning while preserving model performance (Pavlyshenko, 2023c). As a PEFT approach, LoRA is inspired by the structure-aware intrinsic dimension method. Pretrained model parameter weights are frozen during training, which are later combined with a product of two low-rank or trainable matrices after training is conducted. Utilization of trainable matrices allows weights adaptation on new data without the necessity of computing pre-trained weights, thereby resulting in fast and efficient fine-tuning (Nguyen et al., 2023).

Research L demonstrated LoRA's notable capability to exceed other PEFT additive performance (p-tuning and adapter) by shorter inference time and reaching the accuracy of 90% by leveraging the Llama-2-13B model on 3000 training data. As shown in Table 2, research C and J expand on the LoRA framework as a solution for computational resource constraints. Implementation of LoRA enables to perform fine-tuning with Colab leveraging 4bit or 8bit quantization on Llama-2-7B-Chat-Hf and small text data. Furthermore, more research D shows that implementing LoRA as a fine-tuning strategy could prevent catastrophic forgetting of model original knowledge while learning new tasks and time-effectively perform fine-tuning. Strengthen by research S that by employing LoRA as a framework inhibits stability and reliability in instruction fine-tuning.

In conclusion, LoRA as a PEFT approach shows a notable accuracy and shorter inference time performance compared to other PEFT additive approaches. LoRA implementation could also address computational resource constraints since it requires less memory. Moreover, LoRA could prevent catastrophic forgetting while learning new tasks and time-effectively perform fine-tuning. Lastly, LoRA is stable and reliable when employed with instruction fine-tuning.

d. (RQ4) What are the applications of a fine-tuned Llama-based model?

Fig. 2 below shows a pie chart diagram of several applications of Llama as a based model for further fine-tuning on specific downstream tasks. 30% of research is conducted on medical knowledge including radiology (research E), a medical assistant in Chinese (research H and Q), Covid-19 literature assistant (research M), prediction of diagnosis-related on the patient (research O), and even radiation oncology (research S). While 20% of research is conducted in the psychology field such as mental health counseling (research A), mental health analysis (research B and F), and emotion recognition in conversation (research N). Further other domains could also be applied by fine-tuning Llama based model which includes financial analysis and information retrieval (research J and L), fake news detection (research C), bio-inspired materials (research D), language translation (research G), sexual harassment detection (research I), code generation (research K), knowledge bases construction (research P), math problem-solving (research R), to transportation safety information retrieval (research T).

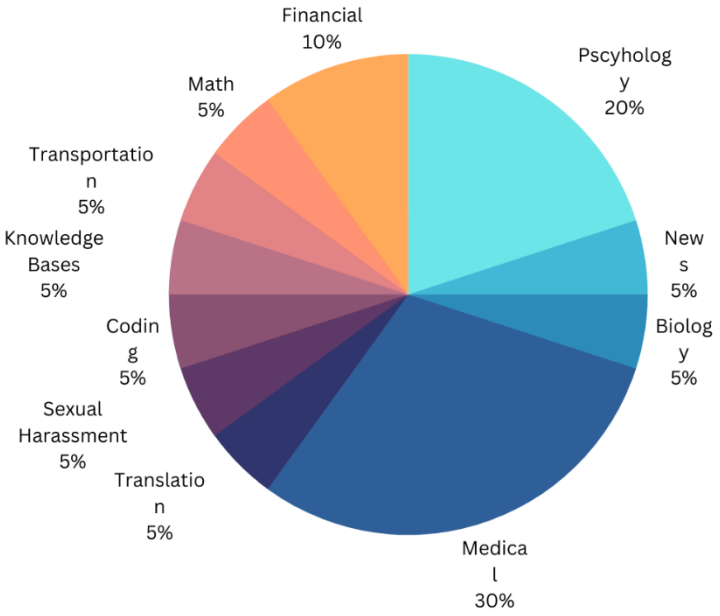

Fig. 2. Pie Chart of Fine-tuned Llama-Based Model Application

Since Llama is trained on mostly English corpus data, further fine-tuning shows model capabilities

to perform well in various languages including Chinese Mandarin (research H), Urdu, and Roman Urdu (research I), even on monolingual translation purposes on: German, Czech, Chinese, and Russian to and from English (research G). Despite having a huge range of domain applications, is it important to acknowledge that LLM might generate harmful advice and biased output in real-life applications (Yang, Zhang, Kuang, Xie, Ananiadou, et al., 2023). Especially in the medical domain which might lead to bad negative effects (Wang, Liu, et al., 2023).

Overall, previous research demonstrated the capabilities of the Llama-based model to perform well on various domains downstream tasks and other languages outside English, outperforming its fine-tuned-based model and other similar domain open-source models. However, it is important to notice ethical considerations while fine-tuning LLM including Llama based model, especially in medical and psychology domains as the major applications.

e. (RQ5) How to evaluate fine-tuned Llama-based model performance?

Evaluation of Llama-based model performance post-fine-tuning is grouped into 4 categories: text generation evaluation, classification evaluation, human evaluation, and auto evaluation. Details of evaluation techniques are displayed in Table 4. The decision to determine which evaluation technique(s) should be implemented is based on the fine-tuning objectives. For example, to evaluate text generation, research K leveraging HumanEval, MBPP, and APPS as evaluation benchmarks to assess code generation ability after fine-tuning. Research G implements COMET and BLEU to evaluate monolingual translation performance due to their ability to quantify the similarity between generated and referenced text.

Other than text generation evaluation, traditional classification evaluation could also be implemented to evaluate LLM fine-tuned on detection or answer generation domains that have either true or false class as an objective. Research R proposed MetaMath, Llama-based fine-tuned for solving mathematical problems, measure accuracy to evaluate performance by leveraging GSM8K and MATH data as a benchmark. Furthermore, Research I implement accuracy and F1 as evaluation metrics to evaluate model performance post-fine-tuning on detecting sexual and abusive chats. Lastly, Research L implements accuracy and precision to measure fine-tuned model performance in retrieving financial information.

Table. 4. Evaluation Techniques

Category	Techniques	Explanation
Text Generation Evaluation	BLEU	Quantify similarity between generated and referenced text (Tan et al., 2023).
	GLEU	Evaluation of sentence fluency (Tan et al., 2023).
	ROGUE	Evaluation of generated and referenced summary based on overlap n-grams (Tan et al., 2023)(Zheng et al., 2023).
	BERTScore	Computation of token similarity, to assess information coverage (Zheng et al., 2023).
	BLEURT	Evaluation of generated text fluency and paraphrase level. (Zheng et al., 2023)
	COMET	Neural based multilingual machine translation evaluation (Rei et al., 2020).
	BART-Score	Evaluation of response quality including fluency, accuracy, and effectiveness (Yuan et al., 2021).
	HumanEval	164 programming problems to evaluate performance on code (Chen et al., 2021).
	MBPP	974 python to evaluate model code generation ability on Python (Austin et al., 2021)
	APPS	Benchmark to evaluate model code, task understanding, and algorithm performance (Hendrycks et al., 2021).

Classification Evaluation	Accuracy	Accuracy refers to the calculation of correctly classified dividied by total samples (Hicks et al., 123 C.E.).
	Precision	Precision refers to the caculation of correctly classified divided by total samples in assigned class (Hicks et al., 123 C.E.).
	F1	F1 is a harmonic mean of recall and precision, low F1 indicate model poor performance on specific class (Hicks et al., 123 C.E.).
Human Evaluation	Determined by researcher based on needs.	Evaluation of response generated conducted by annotators in relevant background on determined metrics. (Wang, Liu, et al., 2023)
Auto Evaluation	GPT-4	Evaluation automatically assessed by prompting GPT-4 to judge answer on several benchmark evaluation (J. M. Liu et al., 2023)

Evaluation on output generation could also be assessed manually through human evaluation. Research H proposed HuaTuo, a Llama-based LLM fine-tuned on Chinese medical knowledge. To evaluate HuaTuo's performance compared to other LLMs, five medical-based annotators are being assembled to assess the model based on 3 criteria: Safety, Usability, and Smoothness (SUS). This human evaluation approach is used to ensure model safety in generating medicine recommendations, as well as usability and smoothness of response which couldn't be obtained through BLEU or ROUGE evaluations. In contrast to human evaluation, auto evaluation utilized superior LLM to evaluate generated output. Research A implements GPT-4 to evaluate ChatCounselor, LLM fine-tuned for mental health support leveraging 229 questions based on seven perspectives.

In conclusion, to evaluate fine-tuning performance on Llama-based LLM depends on the objectives of fine-tuning. Text generation evaluations such as BLUE, ROGUE, and GLUE are well-suited to compare responses on ground-truth information. Whereas classification evaluation metrics such as accuracy, precision, and F1 are well-suited to model fine-tuned for detection, answering, and prediction. Lastly, the human-evaluation approach is suitable for evaluating LLM performance by customized metrics such as safety, knowledge recall ability, etc., as a substitute of human evaluation, auto-evaluation by leveraging a superior large language model such as GPT-4 could be implemented while still conducting evaluation on customized metrics.

4. Conclusion

In summary, this SLR provides valuable insight into current techniques for customizing Llama models, especially instruction-based tuning. But there are clear gaps in transparent reporting of fine-tuning data and procedures which could contribute to issues with fairness and accountability down the line. Furthermore, computational constraints pose challenges to operationalization that must be addressed through efficient methods like LoRA. There remains much open investigation to comprehensively benchmark fine-tuned Llama against state-of-the-art models on real-live applications, while proactively preventing harm. Evaluations focused narrowly on performance metrics provide an incomplete picture. Scholars and practitioners must emphasize ethical considerations fundamental to responsible development and deployment of ever-more powerful tuned language models.

Acknowledgments

We would like to express our gratitude for the assistance provided by the Institution of Research and Community Services at Universitas Multimedia Nusantara in supporting this research. Our thanks also go to our peers at the Information Systems Department's Big Data Laboratory at Universitas Multimedia Nusantara for their valuable contributions and expertise, which significantly enhanced the quality of this study.

References

Aghajanyan, A., Yu, L., Conneau, A., Hsu, W. N., Hambardzumyan, K., Zhang, S., Roller, S., Goyal, N., Levy, O., & Zettlemoyer, L. (2023). Scaling Laws for Generative Mixed-Modal Language Models. *Proceedings of Machine Learning Research*, *202*, 265–279.

Austin, J., Odena, A., Nye, M., Bosma, M., Michalewski, H., Dohan, D., Jiang, E., Cai, C., Terry, M., Le, Q., & Sutton, C. (2021). *Program Synthesis with Large Language Models*. http://arxiv.org/abs/2108.07732

Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H. P. de O., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., Ray, A., Puri, R., Krueger, G., Petrov, M., Khlaaf, H., Sastry, G., Mishkin, P., Chan, B., Gray, S., … Zaremba, W. (2021). *Evaluating Large Language Models Trained on Code*. http://arxiv.org/abs/2107.03374

Choi, S., Gazeley, W., Wong, S. H., & Li, T. (2023). Conversational Financial Information Retrieval Model (ConFIRM). *ArXiv Preprint ArXiv:2310.13001*.

De Angelis, L., Baglivo, F., Arzilli, G., Privitera, G. P., Ferragina, P., Tozzi, A. E., & Rizzo, C. (2023). ChatGPT and the rise of large language models: the new AI-driven infodemic threat in public health. *Frontiers in Public Health*, *11*. https://doi.org/10.3389/fpubh.2023.1166120

García-Holgado, A., Marcos-Pablos, S., & García-Peñalvo, F. (2020). Guidelines for performing Systematic Research Projects Reviews. *International Journal of Interactive Multimedia and Artificial Intelligence*, *6*(2), 9. https://doi.org/10.9781/ijimai.2020.05.005

Hendrycks, D., Basart, S., Kadavath, S., Mazeika, M., Arora, A., Guo, E., Burns, C., Puranik, S., He, H., Song, D., & Steinhardt, J. (2021). *Measuring Coding Challenge Competence With APPS*. http://arxiv.org/abs/2105.09938

Hicks, S. A., Strümke, I., Thambawita, V., Hammou, M., Riegler, M. A., Halvorsen, P., & Parasa, S. (123 C.E.). *On evaluation metrics for medical applications of artificial intelligence*. https://doi.org/10.1038/s41598-022-09954-8

Khan, Y. A., Hokia, C., Xu, J., & Ehlert, B. (2023). covLLM: Large Language Models for COVID-19 Biomedical Literature. *ArXiv Preprint ArXiv:2306.04926*.

Lermen, S., Rogers-Smith, C., & Ladish, J. (2023). *LoRA Fine-tuning Efficiently Undoes Safety Training in Llama 2-Chat 70B*. https://doi.org/https://doi.org/10.48550/arXiv.2310.20624

Liu, J. M., Li, D., Cao, H., Ren, T., Liao, Z., & Wu, J. (2023). ChatCounselor: A Large Language Models for Mental Health Support. *ArXiv Preprint ArXiv:2309.15461*.

Liu, Z., Wang, P., Li, Y., Holmes, J., Shu, P., Zhang, L., & ... (2023). RadOnc-GPT: A Large Language Model for Radiation Oncology. *ArXiv Preprint ArXiv …*.

Luu, R. K., & Buehler, M. J. (2023). BioinspiredLLM: Conversational Large Language Model for the Mechanics of Biological and Bio-inspired Materials. *ArXiv Preprint ArXiv:2309.08788*.

Naveed, H., Ullah Khan, A., Qiu, S., Saqib, M., Anwar, S., Usman, M., Akhtar, N., Barnes, N., & Mian, A. (n.d.). *A Comprehensive Overview of Large Language Models*.

Nayak, A., & Timmapathini, H. P. (2023). LLM2KB: Constructing Knowledge Bases using instruction tuned context aware Large Language Models. *ArXiv Preprint ArXiv:2308.13207*.

Nguyen, T. T., Wilson, C., & Dalins, J. (2023). *Fine-Tuning Llama 2 Large Language Models for Detecting Online Sexual Predatory Chats and Abusive Texts*. http://arxiv.org/abs/2308.14683

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). *Lecture Lecture 4: Learning from Human Feedback*. Advances in Neural Information Processing Systems.

Pavlyshenko, B. M. (2023a). Analysis of Disinformation and Fake News Detection Using Fine-Tuned Large Language Model. *ArXiv Preprint ArXiv:2309.04704*.

Pavlyshenko, B. M. (2023b). Financial News Analytics Using Fine-Tuned Llama 2 GPT Model. *ArXiv Preprint ArXiv:2308.13032*.

Pavlyshenko, B. M. (2023c). *Analysis of Disinformation and Fake News Detection Using Fine-Tuned Large Language Model*. http://arxiv.org/abs/2309.04704

Ray, P. P. (2023). ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems*, *3*, 121–154. https://doi.org/10.1016/j.iotcps.2023.04.003

Rei, R., Stewart, C., Farinha, A. C., & Lavie, A. (2020). COMET: A Neural Framework for MT Evaluation. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2685–2702. https://doi.org/10.18653/v1/2020.emnlp-main.213

Richardson, B., & Wicaksana, A. (2022). Comparison of Indobert-Lite and Roberta in Text Mining for Indonesian Language Question Answering Application. *International Journal of Innovative Computing, Information and Control*, *18*(6), 1719–1734. https://doi.org/10.24507/ijicic.18.06.1719

Roziere, B., Gehring, J., Gloeckle, F., Sootla, S., Gat, I., & ... (2023). Code llama: Open foundation models for code. *ArXiv Preprint ArXiv …*.

Tan, Y., Li, M., Huang, Z., Yu, H., & Fan, G. (2023). MedChatZH: a Better Medical Adviser Learns from Better Instructions. *ArXiv Preprint ArXiv:2309.01114*.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., & Lample, G. (2023). *LLaMA: Open and Efficient Foundation Language Models*. https://doi.org/https://doi.org/10.48550/arXiv.2302.13971

Wang, H., Gao, C., Dantona, C., Hull, B., & Sun, J. (2023). DRG-LLaMA: Tuning LLaMA Model to Predict Diagnosis-related Group for Hospitalized Patients. *ArXiv Preprint ArXiv …*.

Wang, H., Liu, C., Xi, N., Qiang, Z., Zhao, S., Qin, B., & Liu, T. (2023). *HuaTuo: Tuning LLaMA Model with Chinese Medical Knowledge*.

Xu, H., Kim, Y. J., Sharaf, A., & Awadalla, H. H. (2023). A Paradigm Shift in Machine Translation: Boosting Translation Performance of Large Language Models. *ArXiv Preprint ArXiv:2309.11674*.

Xu, X., Yao, B., Dong, Y., Yu, H., Hendler, J., Dey, A. K., & ... (2023). Leveraging large language models for mental health prediction via online text data. *ArXiv Preprint ArXiv …*.

Yang, K., Zhang, T., Kuang, Z., Xie, Q., & ... (2023). MentalLLaMA: Interpretable Mental Health Analysis on Social Media with Large Language Models. *ArXiv Preprint ArXiv …*.

Yang, K., Zhang, T., Kuang, Z., Xie, Q., Ananiadou, S., & Huang, J. (2023). *MentaLLaMA: Interpretable Mental Health Analysis on Social Media with Large Language Models*. http://arxiv.org/abs/2309.13567

Young, A. A. S. J. C. (2021). Embedding from Language Models (ELMos)- based Dependency Parser for Indonesian Language. *International Journal of Advances in Soft Computing and Its Applications*, *13*(3), 2–11. https://doi.org/10.15849/IJASCA.211128.01

Yu, L., Jiang, W., Shi, H., Yu, J., Liu, Z., Zhang, Y., & ... (2023). Metamath: Bootstrap your own mathematical questions for large language models. *ArXiv Preprint ArXiv …*.

Yuan, W., Neubig, G., & Liu, P. (2021). *BARTScore: Evaluating Generated Text as Text Generation*. http://arxiv.org/abs/2106.11520

Zhang, Y., Wang, M., Tiwari, P., Li, Q., Wang, B., & ... (2023). DialogueLLM: Context and Emotion Knowledge-Tuned LLaMA Models for Emotion Recognition in Conversations. *ArXiv Preprint ArXiv …*.

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., … Wen, J.-R. (2023). *A Survey of Large Language Models*. 1–124. http://arxiv.org/abs/2303.18223

Zheng, O., Abdel-Aty, M., Wang, D., Wang, C., & ... (2023). TrafficSafetyGPT: Tuning a Pre-trained Large Language Model to a Domain-Specific Expert in Transportation Safety. *ArXiv Preprint ArXiv …*.

Zhong, T., Zhao, W., Zhang, Y., Pan, Y., Dong, P., & ... (2023). ChatRadio-Valuer: A Chat Large Language Model for Generalizable Radiology Report Generation Based on Multi-institution and Multi-system Data. *ArXiv Preprint ArXiv …*.