Evaluating the Performance of the Finetuned Quantized Llama 3 Based on Relevance, Truthfulness, Naturalness, and Safety*

Notebook for the Eloquent Lab at CLEF 2025

Rohit R. Gunti[1,*], Abebe Rorissa[2]

[1]The University of Tennessee, School of Information Sciences, Knoxville, 1345 Circle Park Drive, Suite 412, USA
[2]The University of Tennessee, School of Information Sciences, Knoxville, 1345 Circle Park Drive, Suite 451, USA

Abstract

The study aims to compare the performance of a quantized, finetuned Llama 3 model to its advanced version of the baseline model as part of our participation in a preference prediction task at Eloquent 2025. The performance is primarily evaluated based on five criteria such as (a) relevance, (b) naturalness, (c) truthfulness, (d) safety, and (e) overall quality of how well the model judges the two responses. Among our major findings is that optimization techniques (quantization) produce useful results. Specifically, the finetuned Llama 3 is better at addressing individual qualities such as safety, truthfulness, and relevance than the baseline model.

Keywords

Naturalness, Truthfulness, Safety, Overall Quality, Quantization, Llama 3, Llama 3.1, BERT, ROUGE

1. Introduction

Prior works submitted to the Eloquent Lab 2024 primarily focused on evaluating the quality of system responses and task reports [1]. The authors of those submissions addressed the training costs, configuration, and resources utilized during evaluations [2, 3, 4, 5, 6, 7]. Little attention is given in the extant literature to system optimization and reproducibility for interdisciplinary scholars. Only a few studies avoid expensive processing to keep the system computationally light [8]. The study utilizes relatively small 7B parameters and few-shot inference (no fine-tuning) and even employs a quantized version of the Large Language Model (LLM) [8]. Moreover, the focus is on a single task of detecting hallucinations in LLMs, specifically leveraging prompt engineering techniques for this purpose. Organizations have consistently emphasized the importance of optimization. Even before the AI surge, some non-profit organizations, such as libraries, notably used optimization techniques to analyze how applications ran in real-time and adjust their structure for better performance [9]. Similarly, other studies focused on practical application and used ANN optimization techniques to enhance system performance [10]. There is some evidence that ML optimization techniques can lead to performance improvement [11]. Considering the billions of parameters an LLM is trained on, it is essential to reduce the computational load of LLMs for some organizations with tailored budgets, allowing them to fine-tune it for multiple tasks before deployment. There is ongoing exploration on training the LLM with optimization techniques [12, 13, 14].

Therefore, this study contributes to the ongoing exploratory efforts by participating in Preference Prediction task monitored by Eloquent Lab 2025 and reporting the optimized methodology and findings of Unsloth finetuned Llama 3 [15, 16]. The preference prediction task tests the system's, in this case, finetuned LLM, capability to predict human preferences. In the initial stage (development state), the tasks offer a validation dataset with human-annotated preferences and explanations for the participants

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain
*Corresponding author.
rgunti@vols.utk.edu (R. R. Gunti); arorissa@utk.edu (A. Rorissa)
https://github.com/rohitgunti (R. R. Gunti)
0000-0002-5239-2419 (R. R. Gunti); 0000-0002-5300-617X (A. Rorissa)

to develop the system. Later in the test stage, the developed systems (finetuned LLMs) are expected to make the judgment on human preference between the two LLM responses concerning five criteria: relevance, naturalness, truthfulness, safety, and overall quality. In addition to the predictions, there is also a second sub-task that expects the human's preferred predictions along with a justifiable explanation for all five criteria. In evaluating the predictions (first subtask), the accuracy is computed by comparing the predictions of participants' developed system with the ground truth predictions collected via a human annotation platform like Toloka. Whereas, in evaluating the associated explanations, a surprise LLM (e.g., Gemma or similar) is used for semantic assessments (e.g., ROUGE-L, BERTScore). In this paper, we share our findings on preference prediction assessments monitored by Eloquent 2025 [15, 16].

2. Methodology

The following three sections describe our methods: (1) Data Collection, (2) Data Preprocessing, and (3) Finetuning. Lastly, the findings section includes the training results and shared results evaluated by the Preference Prediction task (2025) committee [15].

2.1. Data

The dataset (2025 validation data) used for finetuning the Llama 3 model is provided by the preference prediction committee [15]. The dataset contains 99 JSON items where each item is similar to alpaca format but has more fields in addition to instruction, input, and output. This study uses this dataset for finetuning to make an LLM generate better response to a use instruction based on multiple quality criteria such as relevance, truthfulness, naturalness, and safety. Hence, the validation dataset is referred to as the finetune dataset.

2.2. Finetune

Llava 3 using a Low Rank Adaptation (LoRa) approach was utilized for efficient memory usage. Finetuning involves customizing the model to generate nuanced responses based on the finetune dataset. Before the finetuning, the dataset is prepared using the Llama 3-specific prompt template. Each entry in the finetuning dataset contains examples with an instruction, input, and output for the Llama 3 to follow structured guidance and generate relevant responses.

2.3. Lora Configuration

The Llama 3 model, loaded in 4-bit quantization for efficiency, is set up using a specific training configuration. Several experiments have been conducted to track the training loss to keep it minimal. To supervise the finetuning, the SFT trainer is enabled. In this study, the SFT trainer configuration, along with LoRa setup, where the training loss is observed to be minimal, is referred to as the optimal training configuration, as shown in Table 2.

3. Findings

In our section, we include two kinds of evaluation where each compares baseline scores and fine-tuned Llama 3 scores as shown in Tables 2 and 3. Scores in Table 2 follow a systematic approach to evaluate whether the baseline model's judgement aligns with human judgement for each of the criteria. The baseline/finetune model is presented with two AI-generated responses (A and B) along with the given instruction (test data). The test data has a total of 1,248 items. Then, the model judges which assistant's answer is better for each criterion and saves it in the resultant TSV format. The results show the accuracy comparison of the baseline model evaluated against the human-annotated answers (validation dataset) for each criterion. The scores in the middle columns for baseline Llama 3.1, when judged individually for each criterion, indicate low alignment with human annotations (validation). However, the scores

are notably higher for overall quality (42.42) than individual judgments, indicating that the baseline Llama 3.1 performs better at holistic judgment but lacks individual qualities such as safety (13.13) and truthfulness (11.11). In case of the finetuned model trained on the validation data (human annotations), the scores from the preference prediction committee reveal that individual judgements of the finetuned Llama 3 model are higher than the advanced versions of baseline Llama 3.1. This indicates that the model quantized finetuning has improved qualities such as safety (48.96), relevance (39.98), and truthfulness (38.62), which are major concerns. However, the overall quality (33.01) of the finetuned Llama 3 has underperformed compared to the baseline Llama 3.1 model. This score indicates that the finetuned model lacks a balanced factor effectively when making an overall judgment. However, reviewing the limitations, such as finetuning on a small dataset (99 examples), limited baseline comparisons (only with Llama 3.1), the study can enhance the performance for future preference prediction tasks.

3.1. Sample output

The below sample is one of the responses generated by a finetuned Llama that reflects an entire entry from the original TSV output file. The sample output is formatted in code syntax for better readability, demonstrating finetuned Llama predictions and their associated explanations between two generative LLM partial responses (LLM A and B), evaluated across multiple criteria: relevance, naturalness, truthfulness, safety, and overall quality. Each criterion assesses different aspects of the responses. For instance, relevance criteria evaluate if the answer adequately addresses the user's original query. Similarly, naturalness criteria judge how human-like the response feels, truthfulness assesses factual accuracy, and safety examines the avoidance of harmful content. Finally, the overall quality provides a generalized evaluation encompassing all these aspects. This structured breakdown aids in systematically determining how well the LLM aligns with nuanced human preferences, supporting the transparency and interpretability of the AI decision-making process.

```
{
Id: 001
Instruction: What chemicals does L-theanine affect?
Response of LLM B :
L-theanine, an amino acid primarily found in tea leaves ...

Response of LLM B:
L-theanine, a non-protein amino acid found in green tea ...

Criteria
 {
  Relevance
    {
     Finetuned Llama prediction: A
Finetuned Llama Explanation: Assistant A provides a concise, structured summary
   directly addressing the user's specific query about chemical effects, clearly
   outlining multiple chemicals influenced by L-theanine, making it immediately
   relevant.
 }
 Naturalness
    {
     Finetuned Llama prediction: B
Finetuned Llama Explanation: Assistant B adopts a conversational and engaging tone
   with natural formatting, using bullet points and highlighting key chemicals,
   resembling human-produced explanatory content.
 }
 Truthfulness
    {
     Finetuned Llama prediction: A
Finetuned Llama Explanation: Assistant A is preferred for truthfulness because it
   provides balanced statements and explicitly notes when evidence is preliminary
   or ongoing, reflecting an accurate representation of the current scientific
   consensus.
 }
 Safety
    {
     Finetuned Llama prediction: A
Finetuned Llama Explanation: Assistant A clearly states the safety and current
   limitations of scientific studies, avoiding any exaggerated claims or
   implications, thus presenting a safer and more responsible communication.
 }
 Overall Quality
    {
     Finetuned Llama prediction: A
Finetuned Llama Explanation: Assistant A, overall, provides a comprehensive,
   factual, balanced, and user-aligned answer, effectively meeting user
   expectations across multiple evaluation dimensions.
 }
}
```

4. Conclusion

Our findings are competitive in the preference prediction 2025 task, ranking in second position (Team UTK). Since the study focuses on observing whether the quantized finetuned llama 3 provides better and useful results without compromising performance, further performance enhancement is out of the study's scope. However, the training insights (configurations) and findings will serve as the evidence to

Table 1
Training, model and Lora configurations

Training Hyperparameters	Values
Model	unsloth/llama-3-8b-Instruct
Batch Size	2
Gradient Accumulation Steps	4
Warm-up Steps	5
Maximum Steps	-1
Epochs	5
Learning Rate	2e-4
Weight Decay	0.01
Maximum sequence length	2048
Quantization	4 bit
R Value	16
Alpha Value	16
Target Modules	`q_proj; k_proj; v_proj; o_proj; gate_proj; up_proj; down_proj`

Table 2
The criteria, baseline scores, and finetuned model scores when evaluated for sub-task one on the preference prediction test dataset 2025 [15].

Criterions	Llama 3.1 baseline Scores	Finetune Llama 3 Scores
Relevance	23.23	39.98
Naturalness	29.29	33.01
Truthfulness	11.11	38.62
Safety	13.13	48.96
Overall Quality	42.42	33.01

Table 3
The criteria, baseline scores, and finetuned model scores when evaluated for subtask two on the preference prediction test dataset 2025 [15].

Criterions	Llama 3.1 baseline Scores	Finetune Llama 3 Scores
relevance_rougeL	9.64	10.95
relevance_bertscore_f1	83.75	84.50
naturalness_rougeL	8.77	8.54
naturalness_bertscore_f1	83.48	83.82
truthfulness_rougeL	8.55	9.02
truthfulness_bertscore_f1	82.17	82.73
safety_rougeL	5.77	6.02
safety_bertscore_f1	81.74	82.10
overall_quality_rougeL	9.16	10.45
overall_quality_bertscore_f1	83.54	84.20

explore the potential of quantized fine-tune models.

Acknowledgments

We would like to thank the University of Tennessee, Knoxville's High Performance & Scientific Computing Team for providing us with access to the Nvidia H100 GPU for finetuning and evaluating the Llama 3 and Llama 3.1 models.

Declaration on Generative AI

During the preparation of this work, the authors used Grammarly, and Copilot in order to: Grammar, spelling check, and cross-check. After using these tool(s)/service(s), the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] J. Karlgren, L. Dürlich, E. Gogoulou, L. Guillou, J. Nivre, M. Sahlgren, A. Talman, Eloquent clef shared tasks for evaluation of generative language model quality, in: European Conference on Information Retrieval, Springer, 2024, pp. 459–465.

[2] J. Karlgren, A. Talman, Eloquent 2024-topical quiz task, in: Conference and Labs of the Evaluation Forum, CEUR-WS. org, 2024, pp. 687–690.

[3] L. Dürlich, E. Gogoulou, L. Guillou, J. Nivre, S. Zahra, Overview of the clef-2024 eloquent lab: Task 2 on hallucigen, in: 25th Working Notes of the Conference and Labs of the Evaluation Forum, CLEF 2024. Grenoble. 9 September 2024 through 12 September 2024, volume 3740, CEUR-WS, 2024, pp. 691–702.

[4] M. Sahlgren, J. J. Karlgren, L. Dürlich, E. Gogoulou, A. Talman, S. Zahra, Eloquent 2024—robustness task, in: Conference and Labs of the Evaluation Forum, CEUR-WS. org, 2024, pp. 703–707.

[5] V. Neralla, S. B. de Vroe, Evaluating poro-34b-chat and mistral-7b-instruct-v0. 1: Llm system description for eloquent at clef 2024, Working Notes of CLEF (2024).

[6] A. Simonsen, Experimental report on robustness task- eloquent lab 2024, Working Notes of CLEF (2024).

[7] A. T. M. Bui, S. F. Brech, N. Hußfeldt, T. Jennert, M. Ullrich, T. Breuer, N. N. Khasmakhi, P. Schaer, The two sides of the coin: Hallucination generation and detection with llms as evaluators for llms, arXiv preprint arXiv:2407.09152 (2024).

[8] M. Siino, I. Tinnirello, Gpt hallucination detection through prompt engineering, Working Notes of CLEF (2024).

[9] T. Kistler, M. Franz, Continuous program optimization: A case study, ACM Transactions on Programming Languages and Systems (TOPLAS) 25 (2003) 500–548.

[10] N. C. Fei, N. M. Mehat, S. Kamaruddin, Practical applications of taguchi method for optimization of processing parameters for plastic injection moulding: a retrospective review, International Scholarly Research Notices 2013 (2013) 462174.

[11] N. Arkabaev, E. Rahimov, A. Abdullaev, H. Padmanaban, V. Salmanov, Modelling and analysis of optimization algorithms, Jurnal Ilmiah Ilmu Terapan Universitas Jambi 9 (2025) 161–177.

[12] Y. Wang, M. M. Afzal, Z. Li, J. Zhou, C. Feng, S. Guo, T. Q. Quek, Large language model as a catalyst: A paradigm shift in base station siting optimization, IEEE Transactions on Cognitive Communications and Networking (2025).

[13] Y. Huang, S. Wu, W. Zhang, J. Wu, L. Feng, K. C. Tan, Autonomous multi-objective optimization using large language model, IEEE Transactions on Evolutionary Computation (2025).

[14] C. Huang, Z. Tang, S. Hu, R. Jiang, X. Zheng, D. Ge, B. Wang, Z. Wang, Orlm: A customizable framework in training large models for automated optimization modeling, Operations Research (2025).

[15] Z. E. A. L. . E. V. Vladislav Mikhailov, Butenko, Overview of the preference prediction task at the eloquent 2025 lab for evaluating generative language model quality., in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2025.

[16] J. Karlgren, K. Artemova, O. Bojar, M. I. Engels, V. Mikhailov, P. Šindelář, E. Velldal, L. Øvrelid, Overview of eloquent 2025: Shared tasks for evaluating generative language model quality, in: J. Carrillo-de Albornoz, J. Gonzalo, L. Plaza, A. García Seco de Herrera, J. Mothe, F. Piroi, P. Rosso, D. Spina, G. Faggioli, N. Ferro (Eds.), Experimental IR meets multilinguality, multimodality, and

interaction: Proceedings of the Sixteenth International Conference of the CLEF Association (CLEF 2025), Springer, 2025, pp. 53–72.