Optimizing Llama 2 for Knowledge-Grounded Dialogue: Navigating Trade-offs in Limited Resource Environments

Murad Huseynli[1], Kamran Gasimov[2], Junghun Park[3], Nurlykhan Kopenov[4], and Bryan Rakoto Dit Sedson[5]

***Abstract*— This research endeavors to bridge the gap in fine-tuning large language models (LLMs) for knowledge-grounded dialogue generation, particularly under the constraints of limited computational resources. We focus on enhancing the Llama 2 model using advanced techniques such as QLoRA and PEFT, aiming to improve the model's ability to generate contextually relevant dialogues. Faced with hardware limitations, our study highlights the complexities and trade-offs involved in optimizing LLMs in resource-constrained settings. Our comprehensive evaluation, combining manual and quantitative analysis, demonstrates substantial improvements in dialogue quality. However, it also reveals a crucial trade-off between model size and runtime efficiency. This paper contributes insights into fine-tuning LLMs within practical limitations, paving the way for future research in resource-efficient AI solutions.**

Keywords: Large Language Models, Llama 2, Knowledge-Grounded Dialogue, Fine-Tuning, QLoRA, PEFT, Computational Constraints, Model Optimization, AI Research

I. INTRODUCTION

The advent of artificial intelligence has revolutionized many aspects of technology, with dialogue systems being a prominent area of transformation. Knowledge-grounded dialogue generation, where conversational agents are enhanced with the ability to access and integrate external knowledge, has emerged as a critical research area. This development is not just a technical advancement but also aligns with the growing demand for more sophisticated and context-aware AI interactions in various applications, from customer service to personal assistants.

The pioneering work by Dinan et al. (2018) [1] introduced the Wizard of Wikipedia dataset, setting a new standard for knowledge-grounded dialogue systems. Their framework provided a platform for conversational agents to engage in more informative and contextually rich dialogues. However, as with any evolving technology, this area of research is not without its challenges. One of the persistent issues is the delicate balance between leveraging external knowledge and maintaining conversational coherence and relevance. While models have become adept at integrating vast amounts of information, they often falter in ensuring that the dialogue remains contextually grounded and coherent over multiple turns. This challenge is further compounded by the dynamic nature of human conversations, where contextual cues and implicit knowledge play a significant role.

Recent contributions in the field have aimed to address these challenges. For instance, Lewis et al. (2020) [2] introduced a retrieval-augmented generation framework for enhancing factual accuracy in language models, particularly targeting the issue of producing responses that may be factually incorrect or irrelevant. Similarly, Zhao et al. (2020) [3] proposed focusing on leveraging external knowledge in dialogue systems through a knowledge selection module combined with an unsupervised learning approach utilizing and fine-tuning various pre-trained models. Additionally, Touvron et al. (2023) [4] developed a series of pre-trained and fine-tuned large language models optimized for dialogue use cases, demonstrating superior performance in dialogue generation and introducing significant improvements in model safety and effectiveness. These studies represent crucial steps toward more reliable and contextually aware conversational agents.

Our research builds upon these foundations, with a focus on fine-tuning the Llama 2 model using the Wizard of Wikipedia dataset. We propose that advanced fine-tuning techniques such as QLoRA and PEFT can substantially enhance the model's ability to navigate the complexities of knowledge-grounded dialogues. By doing so, we aim to address the noted gap in the field – achieving a harmonious balance between factual accuracy and conversational naturalness, especially, under resource-constrained scenarios. This paper presents our methodology, experiments, and findings, contributing to the ongoing discourse in AI and dialogue systems.

The paper is structured as follows: Section II delves into related works, shedding light on both the achievements and the current challenges in the field. Section III outlines our methodology, detailing the steps taken in dataset preparation, model fine-tuning, and performance evaluation. Sections IV and V present our experimental results and discuss their implications, respectively. Finally, Section VI concludes the paper, offering insights and future research directions.

II. RELATED WORKS

The evolution of open-domain dialogue generation has been significantly influenced by the integration of external knowledge sources. Addressing a prevalent challenge in open-domain conversations, where knowledge was not effectively utilized, one of the landmark advancements in

[2]Department of Computer Science, KAIST, kirito233 at kaist.ac.kr
[1]Department of Computer Science, KAIST, kamran.qasimoff at kaist.ac.kr
[2]Department of Aerospace Engineering, KAIST, jokjeb2 at kaist.ac.kr
[2]Department of Computer Science, KAIST, knurlykhan at kaist.ac.kr
[2]Department of Computer Science, KAIST, b.rakotoditsedson at kaist.ac.kr

this domain was introduced by Dinan et al. (2018) [1], where they developed a model to effectively utilize Wikipedia as a knowledge source for dialogue generation. Their creation of the Wizard of Wikipedia dataset not only enabled benchmarking in this area but also served as a key resource for training dialogue systems to generate more informative and contextually rich conversations. This dataset forms a crucial part of our study, where we utilize it to fine-tune the Llama 2 model, aiming to enhance its ability to generate contextually rich and informative dialogues.

Subsequently, pre-trained language models capable of generating contextually relevant responses and fine-tuning for specific tasks were employed. Nevertheless, despite the progress in knowledge integration, pre-trained language models often struggle with generating engaging and specific responses, particularly when tackling niche topics in open-domain dialogues. This limitation was addressed by Zhao et al. (2020) [3], who developed methods for enabling models to adapt to diverse topics, resulting in generating more precise and topic-relevant responses. Their contributions in knowledge-grounded conversations utilizing various pre-trained large language models, significantly influence our approach to fine-tuning Llama 2, as we aim to enhance the model's ability to produce context-specific dialogues.

Nevertheless, pre-trained language models faced challenges in manipulating knowledge. This resulted in the introduction of non-parametric memory in dialogue models which marked another significant advancement in this domain. Lewis et al. (2020) [2] proposed a retrieval-augmented generation model, which demonstrated improved performance in generating specific, diverse, and factually accurate responses in open-domain question-answering tasks. This concept of retrieval-augmented generation is particularly relevant to our methodology, where we explore advanced fine-tuning techniques to enrich the knowledge grounding of the dialogues generated by Llama 2.

Furthermore, Touvron et al. (2023) [4] released the Llama 2 model comprising various parameter configurations. With that, the landscape of language models witnessed a significant leap forward. The Llama 2-chat, optimized for dialogue use cases, exhibited commendable performance in terms of helpfulness and safety among closed-source models. Its training on public sources has contributed immensely to the open-source community, aligning with our research goals of enhancing dialogue quality in open-domain scenarios and making advanced models more accessible.

However, the increasing complexity and size of language models present challenges in their practical application, especially for resource-constrained users. To address this, the Parameter-Efficient Fine-Tuning (PEFT) method was introduced by Liu et al. (2022) [5], allowing for fine-tuning of models with a small number of parameters being updated. In addition to that, Dettmers et al. (2023) [6] introduced the QLoRA technique, which facilitates efficient fine-tuning of large models. Their work, featuring innovations such as 4-bit NormalFloat and Double Quantization, aligns with our approach of employing PEFT and QLoRA to fine-tune the parameter-rich Llama 2 model, optimizing for computational efficiency.

In summary, our study is positioned at the intersection of these significant developments. By leveraging the Wizard of Wikipedia dataset for fine-tuning Llama 2 and employing advanced techniques such as PEFT and QLoRA, we aim to contribute to the ongoing evolution of knowledge-grounded dialogue systems. Our goal is to address the challenges of integrating detailed, relevant knowledge into dialogues while maintaining conversational coherence and efficiency, especially, in resource-constrained environments.

III. METHODOLOGY

A. Dataset Description

The Wizard of Wikipedia (WoW) dataset [1] serves as the foundation for this study. It is a large-scale dataset designed specifically for the tasks of "Knowledge-grounded Open Dialogue Generation". The dataset contains 22,311 dialogues with 201,999 conversational turns across 2,437 diverse topics. The dataset's structure is intricate, involving multiple components:

- **Topic:** The central theme of the dialogue.
- **Persona:** A corresponding persona motivating the conversation's topic.
- **Wizard Evaluation:** Performance evaluation of the wizard by the apprentice at the dialogue's conclusion.
- **Dialogue:** A list of turns within the dialogue.
- **Topic Passage:** Sentences from Wikipedia corresponding to the chosen topic, used by the wizard to inform responses.

Each entry in the "Dialogue" component comprises:

- **Speaker:** Identified as either "wizard" or "apprentice".
- **Text:** The actual text written by the speaker.
- **Retrieved Topics:** Topics retrieved for that particular utterance.
- **Checked Sentence (Wizard only):** A dictionary entry mapping the topic to the sentence chosen by the wizard.
- **Checked Passage (Wizard only):** A dictionary entry mapping the topic to the chosen passage by the wizard.

The conversation flow in the WoW dataset begins with either the wizard or apprentice selecting a topic. The wizard, upon receiving a message from the apprentice, is shown relevant Wikipedia knowledge and chooses a sentence to construct their response. The dialogue continues with this pattern until it concludes after a minimum of 4 or 5 turns each.

B. Pre-processing of the WoW Dataset

The pre-processing of the WoW dataset involved several stages:

1) Data Extraction and Initial Processing: The dataset was accessed via the ParlAI toolkit. Each dialogue and its associated Wikipedia articles were extracted from the JSON file. The raw data was then transformed into a structured format suitable for the subsequent processing steps.

2) Conversion for Llama 2 Model Compatibility: For the Llama 2 model to effectively process the dataset, each dialogue needed to be structured as a single "text" attribute. The WoW dataset, originally featuring multiple attributes with separate utterances, was reorganized. Each dialogue was transformed into a continuous string following a specific template. This template included user messages, model answers, and optional instructions. These instructions, crucial for contextual understanding, guide the model's responses, especially in multi-turn dialogues where utterances are concatenated in sequence.

3) Incorporating Instruction-based Tuning: To further adapt the dataset for Llama 2, instructional context was integrated based on the behavior of the wizard as described in the original WoW paper. This involved embedding Wikipedia articles, originally used to generate utterances, directly into the instruction set of each dialogue string. The inclusion of these articles was vital, as we did not provide an external knowledge source, thus ensuring that each dialogue instance was grounded in relevant information.

4) Compilation into Training-Ready Format: Finally, the reformatted dialogues were compiled into a new JSON file. This file represented the WoW dataset now fully optimized and formatted for training the Llama 2 model. Each object in the dataset was converted into a singular string, suitable for the fine-tuning process of the Llama 2 model, ensuring compatibility and readiness for the subsequent training phase.

C. Model Description - Llama 2

The Llama 2 model, with its 7 billion parameters, was the chosen architecture for this study. This version was selected due to its balance between computational efficiency and performance capability, considering our hardware constraints (GPUs' memory amount. Llama 2 is acclaimed for its exceptional natural language processing abilities [4], particularly in understanding and generating contextually rich dialogues. This makes it an ideal candidate for tasks involving complex conversational dynamics and knowledge integration, such as those presented by the WoW dataset.

D. Fine-Tuning Techniques

1) Default Llama 2 Fine-Tuning: In the default fine-tuning approach, Llama 2 was trained directly on the pre-processed WoW dataset. This process involved adjusting the model's parameters to better align with the dataset's unique conversational structure and knowledge elements. The aim was to enhance the model's capability to generate responses that are not only contextually relevant but also rich in factual content.

2) Llama 2 Fine-Tuning with QLoRA and PEFT: For more advanced fine-tuning, we employed the QLoRA and PEFT techniques. QLoRA (Query-based Learning of Representational Arrays) is a method focused on improving the model's ability to handle query-specific knowledge. PEFT (Progressive Embedding Fine-Tuning) is aimed at incrementally adjusting the model's embedding layers to better capture the nuances of knowledge-grounded dialogue. These techniques were specifically chosen to bolster Llama 2's proficiency in generating responses that are both contextually appropriate and informationally accurate given the initial constraints.

E. Baseline Model Establishment

For our study, the baseline was established using the default, unmodified Llama 2 model. This baseline model served as a critical reference for comparative analysis. It was used to objectively evaluate the improvements and enhancements achieved through the fine-tuning processes detailed earlier. Specifically, this baseline was compared against the two fine-tuned versions of Llama 2: the first fine-tuned exclusively on the pre-processed WoW dataset and the second fine-tuned using the QLoRA and PEFT techniques. The comparison focused on evaluating dialogue coherence, contextual relevance, and the accuracy of knowledge integration in each model. This approach provided a comprehensive understanding of the fine-tuning's effectiveness in enhancing the model's performance in knowledge-grounded dialogue generation.

F. Performance Assessment of Fine-tuned Models

To thoroughly evaluate the performance of the fine-tuned Llama 2 models, our study employed both quantitative and qualitative assessment methods:

1) Quantitative Assessment: The quantitative analysis centered around a meticulously constructed confusion matrix, alongside other key metrics, to objectively assess language understanding and generation capabilities of the models. This approach provided valuable statistical insights, allowing for a comprehensive comparison of the models' performances against the baseline. The analysis highlighted the nuanced improvements achieved through our fine-tuning processes, despite the hardware constraints faced.

2) Qualitative Assessment: Complementing the quantitative analysis, qualitative assessment played a vital role in our evaluation strategy. It involved in-depth human evaluation (done by the research team) of a subset of dialogues generated by the models. Reviewers focused on aspects such as coherence, relevance, and conversational naturalness. This assessment was crucial in gauging the practical applicability of the fine-tuned models, especially in scenarios that closely mimic real-world conversational settings. The qualitative feedback provided insights into the models' performance, beyond what could be captured through quantitative metrics alone. These combined assessment methods ensured a holistic evaluation of the fine-tuned models, enabling us to draw comprehensive conclusions about their performance in knowledge-grounded dialogue generation.

IV. Evaluation

A. Context and Challenges

In this study, evaluating the Llama 2 large language models (LLMs) presented unique challenges, particularly due to our team's limited hardware resources. The substantial size of the Llama 2 model during training phases far exceeded

the memory capacities of our available GPUs, necessitating innovative approaches to effectively assess the model's performance under these constraints.

B. Evaluation Methodology

Our evaluation strategy was adapted to the nature of our training dataset, which primarily consisted of open-domain dialogues without mathematical or coding content. We assessed the model's performance on a carefully chosen set of 200 questions covering 10 randomly chosen topics from the training dataset. This methodological choice was pivotal, ensuring the evaluation's relevance to the model's training and intended application. A comparison with the baseline model was also included to gauge the improvements achieved through fine-tuning.

C. Manual Assessment and Confusion Matrix

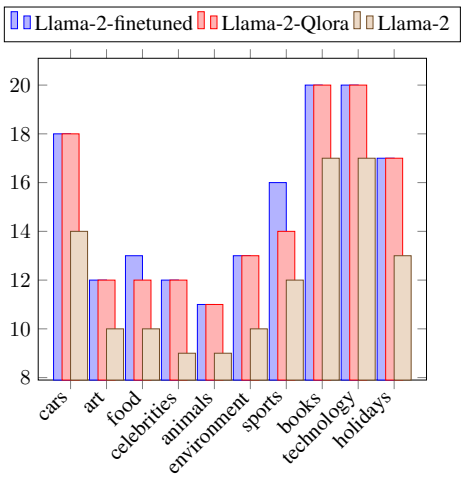
To accurately assess the model's response quality, we employed a manual evaluation approach. This was essential in identifying instances where the model either hallucinated or failed to respond appropriately. Automated metrics based solely on generated token numbers would have been inadequate, potentially misclassifying hallucinations as correct responses. Instead, our manual approach classified responses into three categories: "correct generation," "wrong generation," and "ambiguous generation," enabling the construction of a detailed confusion matrix, offering a quantifiable insight into the model's performance.

TABLE I
MODEL GENERATION OUTPUT EVALUATION

	Correct Generation	Wrong Generation	Ambiguous Generation
Llama-2-7b	122	**68**	**10**
Wiki-Llama-2-w/o QLORA	**152**	45	3
Wiki-Llama-2	149	48	3

D. Bar Chart Analysis

The bar chart provides a visual representation of the models' performance across various topics. It illustrates the differences in the number of correct predictions, shedding light on each model's strengths and weaknesses in specific content areas.



V. DISCUSSIONS

A. Analysis of Results

The confusion matrix and bar chart collectively offer a comprehensive view of each model's performance. The fine-tuned models, particularly the Wiki-Llama-2 with QLoRA, exhibit improved accuracy in generating contextually relevant responses. The manual evaluation approach was instrumental in capturing the nuanced capabilities and limitations of each model variant.

B. Model Size and Runtime Efficiency

Our fine-tuning efforts, especially the application of quantization techniques, resulted in a significant reduction in the model size. The Wiki-Llama-2 model, for instance, decreased from approximately 12GB to about 60MB, marking a reduction of nearly 99.5%. However, this dramatic decrease in size was accompanied by a notable increase in runtime, with the Wiki-Llama-2 model requiring substantially more time for both initialization and response generation compared to the baseline model. The Wiki-Llama-2-w/o QLORA model, while also experiencing an increase in runtime, was less affected than its QLoRA-utilizing counterpart. These findings highlight a critical trade-off between model size and computational efficiency.

C. Implications and Future Directions

The increased runtime, particularly for the Wiki-Llama-2 model, although is fair enough for resource-constrained environments, still raises important considerations for practical applications, especially in scenarios requiring real-time responses. The balance between model size, runtime efficiency, and performance accuracy is a key takeaway from our evaluation. Future research may explore optimizing this balance, potentially through alternative fine-tuning methods or advanced model compression techniques, to achieve more efficient and effective language models for various applications.

VI. CONCLUSION AND FUTURE WORK

A. Concluding Remarks

This research embarked upon the rigorous endeavor of fine-tuning the Llama 2 model for the purpose of enhancing knowledge-grounded dialogue generation. Confronted with significant constraints in time and computational resources, particularly GPU memory limitations, this study necessitated innovative approaches in both model optimization and performance evaluation. Our empirical findings indicate a notable enhancement in the model's ability to generate contextually accurate and relevant dialogues, as substantiated by our manual evaluation methodology and quantitative analyses via confusion matrices and bar charts. The fine-tuned Wiki-Llama-2 models, incorporating QLoRA, have demonstrated marked improvement in generating coherent and context-appropriate responses given the constrained resources.

Nonetheless, the study also brought to light the intricate trade-offs encountered in model optimization, particularly

between the reduction in model size and the efficiency of runtime performance. The quantization techniques employed successfully diminished the model's size, yet concurrently resulted in an increased operational time, posing challenges for applications necessitating real-time responsiveness.

B. Reflections on Limitations

The constraints of limited GPU memory and time not only posed technical challenges but also provided valuable insights into conducting high-level AI research under resource constraints. These limitations highlighted the critical balance between model sophistication and operational feasibility, underscoring the importance of resource-efficient AI solutions in practical settings.

C. Contributions to the Field

Our research may potentially contribute to the field of AI and language modeling in several areas:

- Demonstrating the potential of advanced fine-tuning techniques, such as QLoRA, to enhance LLMs within the confines of limited computational resources.
- Bringing to light the intricate trade-offs between model size, runtime efficiency, and performance accuracy in the context of constrained hardware environments.
- Proposing an adaptable framework for evaluating LLMs that can be applied in settings with similar resource limitations.

D. Future Research Directions

The findings of this study may unlock several avenues for future research, especially in optimizing language models within resource-limited environments:

- **Advancements in Optimization Strategies:** Future research could explore innovative fine-tuning and model compression methods that refine the equilibrium between computational efficiency, model size, and linguistic accuracy.
- **Enhancing Scalability and Accessibility:** Investigating approaches to democratize access to advanced language models, making them feasible for use in environments with limited computational resources.
- **Focus on Real-Time Application Efficiency:** Developing strategies to curtail the operational time of fine-tuned models, thereby facilitating their deployment in real-time conversational systems.
- **Incorporating a Diverse Range of Data:** Expanding the training datasets to encompass a broader spectrum of topics and data types, thereby examining the model's adaptability and robustness across various contexts.
- **Development of Refined Automated Evaluation Techniques:** Creating more sophisticated automated metrics capable of capturing the complexities and nuances inherent in natural language generation for resource-constraint environments.

In conclusion, our study, while constrained by limited resources, has attempted to contribute to the ongoing development of knowledge-grounded dialogue systems and highlighted the importance of resource efficiency in AI research. The challenges faced have provided a unique perspective on the trade-offs and considerations essential in the field of AI, potentially paving the way for future innovations in language model optimization.

REFERENCES

[1] E. Dinan, S. Roller, K. Shuster, A. Fan, M. Auli, and J. Weston, "Wizard of wikipedia: Knowledge-powered conversational agents," *arXiv preprint arXiv:1811.01241*, 2018.

[2] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, et al., "Retrieval-augmented generation for knowledge-intensive NLP tasks," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 9459-9474.

[3] X. Zhao, W. Wu, C. Xu, C. Tao, D. Zhao, and R. Yan, "Knowledge-Grounded Dialogue Generation with Pre-trained Language Models," Nov. 2020, doi: https://doi.org/10.18653/v1/2020.emnlp-main.272.

[4] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, et al., "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023.

[5] H. Liu, D. Tam, M. Muqeeth, J. Mohta, T. Huang, M. Bansal, and C. A. Raffel, "Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning," in *Advances in Neural Information Processing Systems*, vol. 35, 2022, pp. 1950-1965.

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