Fine-tuning Data Curation for LLM-based Question-Answering Using LoRA/QLoRA for a Specific Domain

XIAOQIN FU, SUBHRAJEET GHOSH, and USAMA AHMED*, The University of Arizona, USA

Recent advances in AI and large language models (LLMs) have the ability to process, interpret, and generate text, making them effective for accessing relevant insights. However, foundational LLMs (e.g., GPT, Llama, T5, BERT) rely on general datasets that lack the specialized knowledge required for building energy. Fine-tuning these models with domain-specific data presents a promising solution, although challenges remain in addressing missing, duplicate, inconsistent, or unstructured information. Meanwhile, LoRA improves model efficiency by injecting trainable low-rank matrices into frozen model layers, significantly reducing training parameters and computational costs while maintaining or surpassing full fine-tuning performance. Furthermore, QLoRA builds on this by applying 4-bit quantization to the low-rank matrices, further cutting memory and resource use.

This paper proposes a training data generation and validation method for fine-tuning the foundation model to power question-answering systems specifically for a specific domain: building energy. Our method leverages various techniques and tools (e.g., PyPDF2, Hugging Face Hub) to query relevant data for preprocessing, incorporate it into an LLM, enhance contextual understanding, and provide more comprehensive insights through Question-Answer (QA) pairs for fine-tuning. From this processed content, we generate QA pairs and use LoRA & QLoRA to fine-tune LLMs focusing on the energy domain of buildings. The initial results show that fine-tuned Llama LLM versions outperform their foundational versions with the same parameters, and larger models generally perform better. The fine-tuned LLM can answer questions in a specific domain, such as building energy.

CCS Concepts: • **Computing methodologies** → **Neural networks**.

Additional Key Words and Phrases: fine-tuning, large language models, LoRA, QLoRA

ACM Reference Format:
Xiaoqin Fu, Subhrajeet Ghosh, and Usama Ahmed. 2025. Fine-tuning Data Curation for LLM-based Question-Answering Using LoRA/QLoRA for a Specific Domain.

1 INTRODUCTION

Buildings are responsible for approximately 39% of global energy-related carbon emissions, with 28% originating from operational emissions, such as the energy used to cool, heat, and power various buildings, and the remaining 11% from materials and construction processes [1]. Implementing effective energy strategies in buildings is essential to reduce greenhouse gas emissions and achieve global climate goals [5]. The building energy industry encompasses tons of specialized documents, including building codes, software technical manuals, and other technical publications. This vast pool of information presents significant challenges in locating, extracting, and learning from the knowledge contained within these documents. The main obstacles include: 1) **Data heterogeneity**: Information is often stored in various formats and structures, making it difficult to integrate and analyze effectively [7]; 2) **Data accessibility**:

Authors' address: Xiaoqin Fu, fuxiaoqin@arizona.edu; Subhrajeet Ghosh, subhrajeetghosh@arizona.edu; Usama Ahmed, usamaahmed@arizona.edu, The University of Arizona, 1200 E University Blvd, Tucson, Arizona, USA, 85721.

Access to comprehensive data sets can be hindered by proprietary restrictions, inconsistent data availability, and the need for specialized tools to interpret complex data [15]; and 3) **Knowledge management challenges**: The building energy sector faces difficulties in capturing and preserving explicit and tacit knowledge, which is essential for informed decision making [16].

Implementing robust knowledge management, standardizing data formats, and leveraging advanced question-answering systems can facilitate better access to and use of the wealth of available information. Moreover, there is no effective question-answering system from a community or user perspective, especially for general users, who often find it difficult to understand the knowledge and technologies of building energy.

In recent years, artificial intelligence (AI) and machine learning (ML) technologies, including large language models (LLMs), have made significant progress in natural language processing (NLP) and human-computer interaction. Trained with datasets, these models can use context learning to adapt to a wide range of tasks and applications. With the ability to understand, process, and generate various data types, including text, images, code, and multimodal inputs, LLMs offer versatile solutions in numerous fields. This adaptability brings new opportunities for building energy, opening doors to more efficient and sustainable solutions.

Fine-tuning is a pivotal method for integrating new (domain-specific) knowledge into LLMs, significantly enhancing their performance across various domains, including medical science and education. It involves further training base (foundational) models or pre-trained models with the domain-specific data (i.e., specialized data), adapting to specific tasks and fields, such as enabling the user to query and get the response, as shown in Figure 1.

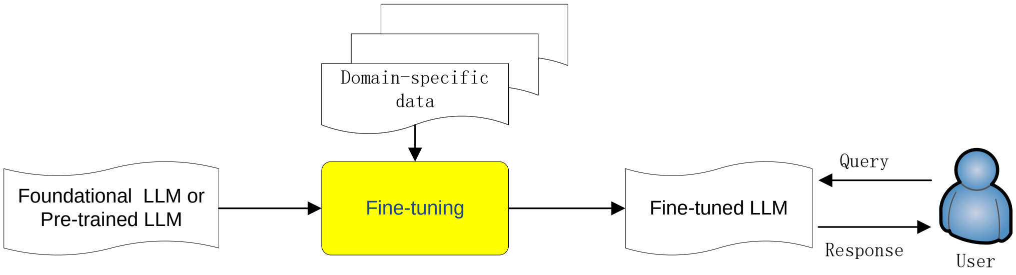


Fig. 1. Large language Model Fine-tuning Process.

The lack of fine-tuning in the building sector can be attributed to several factors: 1) **Data availability**: Access to high-quality and domain-specific datasets is crucial for effectively fine-tuning and the building industry may lack such curated datasets, hindering the application of fine-tuning techniques; 2) **Resource constraints**: Fine-tuning requires significant computational resources and expertise, which may not be readily available in the building sector; and 3) **Awareness and expertise**: There may be a lack of awareness or no understanding of the benefits and methodologies of fine-tuning LLMs within the industry. Addressing these challenges could unlock potential LLM applications in the building sector, leading to more effective, efficient, smart, and low-carbon buildings.

A critical research gap in fine-tuning LLMs for building energy is data curation. Specifically, determining the right quality and quantity of data for effective fine-tuning. Key unanswered questions include whether raw or processed data are more suitable, the optimal amount of data needed, which fine-tuning methods or algorithms are most effective, and the computational resources required for implementation. Addressing these issues is essential for advancing LLM-based solutions in smart building energy management and decision-making.

This paper addresses the research gap by effectively generating and comparing high-quality data for fine-tuning and applying Low-Rank Adaptation (LoRA) and Quantized LoRA (QLoRA) fine-tuning techniques for LLM-based smart question-answering systems. s The paper is organized as follows: We first give some related work. Then, we provide a detailed methodology for data preparation and fine-tuning LLMs used for building energy. Next, there are our experiments and the corresponding results. We also list relevant limitations. Lastly, we summarize our findings and future work in the conclusion section.

2 RELATED WORK

As an improved fine-tuning method, LoRA introduces low-rank matrices into the model's architecture, enabling efficient adaptation to new tasks with minimal computational overhead, as shown in Figure 2. It freezes the original weights and injects low-rank matrices, which are trainable, into each layer of the model. LoRA drastically reduces the amount of training parameters and the relevant computational burden. Empirical results demonstrate that LoRA matches or even outperforms the performance of full fine-tuning across multiple LLMs, despite updating far fewer parameters [6].

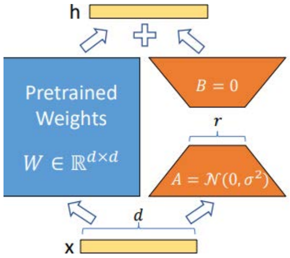
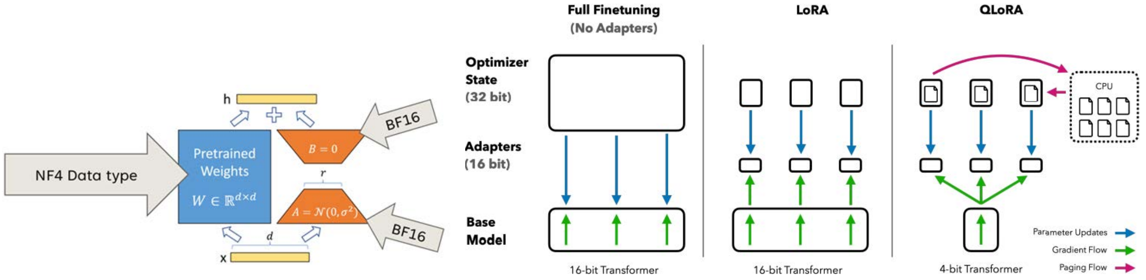

h
Pretrained Weights
$W \in \mathbb{R}^{d \times d}$
$B = 0$
r
$A = \mathcal{N}(0, \sigma^2)$
d
x

Fig. 2. LoRA [6].

Furthermore, QLoRA extends this by applying quantization to low-rank matrices, further reducing memory storage usage and computational resource requirements. It combines low-rank adaptation with model quantization, specifically 4-bit quantization, to reduce memory usage, achieving dramatic reductions in memory consumption and enabling fine-tuning of very large models. Its innovations include the introduction of "4-bit NormalFloat" (NF4) for higher precision at low bit-widths, "double quantization" to compress quantization constants, and "paged optimizers" to efficiently manage optimizer states in limited GPU memory. These techniques allow QLoRA to achieve similar performance with even smaller computational and memory usages than LoRA, as illustrated in Figure 3 [3]. Comparative evaluations show that QLoRA achieves parity or slight improvements over LoRA and other parameter-efficient fine-tuning methods, particularly in low-resource or high-model-size scenarios.



(a) LoRA + quantized linear layers [11]

(b) Full finetuning vs LoRA vs QLoRA [3]

Fig. 3. QLoRA.

In the building industry, LLMs have been utilized for tasks such as predicting energy consumption, optimizing building design, etc. For example, Zhang et al. [21] explored the role of LLMs in the field of building science, highlighted their potential applications, and mentioned their limitations. However, these applications often rely on general-purpose models without domain-specific fine-tuning. Although LLM fine-tuning has been extensively applied in various sectors, including medical science and education [21][4], its application in the building sector remains limited. Fine-tuning involves further training pre-trained models on specialized datasets to enhance their performance in specific tasks or domains. There are only two fine-tuning works [8][20], focusing on the building energy sector.

3 METHODOLOGY

We propose the workflow of fine-tuning data generation and validation for LLM-based smart systems that answer the questions related to building energy. Figure 4 depicts the overarching workflow, highlighting two closely connected parts: Data Preparation (Part 1) and LLM Execution (Part 2).

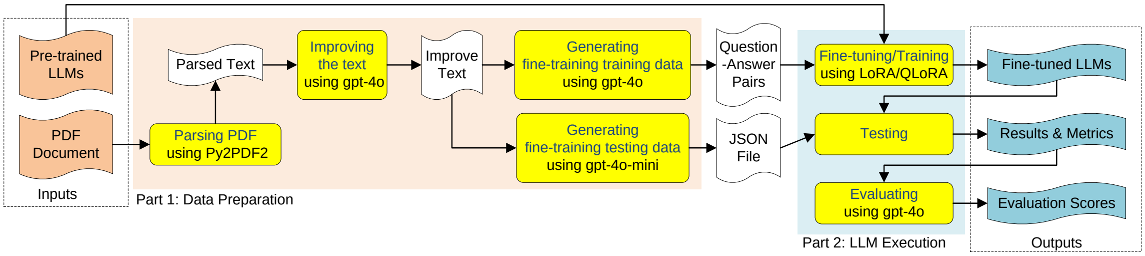


Fig. 4. The workflow for preparing data and fine-tuning LLMs for building energy.

Our system takes two types of inputs: the pre-trained LLMs and PDF document(s). In general, data quality plays a key role in fine-tuning LLMs, directly influencing the model's performance and accuracy. High-quality and suitable data enable LLMs to learn effectively, leading to improved LLMs' outcomes in various applications [?]. In Part 1, we first use the Python library PyPDF2 to parse the PDF document(s) to get the corresponding unstructured text document(s), as shown in Listing 1.

Listing 1. Example parsed text

```
EnergyPlus Version 22.2.0 Documentation
Engineering Reference
U.S. Department of Energy
September 28, 2022 .......
```

Using the OpenAI model gpt-4o (2024-11-20 version) and prompt engineering, we improved the parsed but unstructured text and then efficiently generated question-answer (QA) pairs as fine-tuning training data from the improved text, as shown in Listings 2 and 3. The corresponding prompts are described in Listings 4 and 5.

Listing 2. Example improved text

```
EnergyPlus Version 22.2.0 Documentation Engineering Reference provides a comprehensive guide
to understanding the modeling and simulation of energy systems in buildings.
This document is a product of collaboration among the U.S. Department of Energy, .......
```

Listing 3. Example QA pair for fine-tuning LLMs

```
[{ "question": "What is the purpose of the EnergyPlus Version 22.2.0 Documentation
Engineering Reference?",
"answer": "It provides a comprehensive guide to understanding the modeling and simulation
of energy systems in buildings." }, ...]
```

Listing 4. The Prompt for improving the parsed and unstructured text

```
    prompt = f"""
    Please generate meaningful and structured text from the given unstructured text using OpenAI API,
        no changing the text, remaining all meaningful contents, and just removing directories,
        references, and meaningless sentences:
    Text: "{text}"     """
```

Listing 5. The Prompt for generating the QA pairs as the training data for fine-tuning LLMs

```
prompt = f""" Generates meaningful question-answer pairs from the following text:
Text: "{text}"
Provide the output in JSON format that the question and answer are on the same line...] """
```

In addition, as the testing data for fine-tuned LLMs, the JSON file was generated using the OpenAI model gpt-4o-mini (2024-07-18 version) and the corresponding prompt, as shown in Listing 6.

Listing 6. The Prompt for generating the JSON file, including a few QA pairs, as the testing data

```
prompt = f"""  Generates one meaningful and best question-answer pair from the following text:
Text: "{text}"
Provide the output in JSON format that the question and answer are on the same line..."""
```

In Part 2, we selected Meta's Llama for fine-tuning. It is a parameter-efficient LLM that has demonstrated strong performance on various benchmarks. The choice of Llama was influenced by its open-source nature. Llama's parameter efficiency and cost-effectiveness make it a suitable candidate for fine-tuning in specialized domains. We used different sizes and versions of the above to compare their performance in learning external knowledge related to building energy. After fine-tuning, we obtain fine-tuned LLMs and use testing data to test them to generate results (i.e., the questions and their answers in the testing data) and relevant metrics, such as **BLEU** (i.e., bilingual evaluation understudy), **ROUGE** (i.e., recall-oriented understudy for gisting evaluation), and **BERTScore**.

BLEU measures the precision of n-grams (i.e., n words' sequences) in the generated text compared to reference texts, emphasizing the overlap of n-grams between the candidate and expected texts, penalizing shorter candidate texts to avoid overly concise outputs [12]. **ROUGE** evaluates the recall of n-grams, or word sequences/pairs between the generated text and reference text, including several sub-metrics: 1) **ROUGE-1** measuring the overlap of unigrams (single words), 2) **ROUGE-2** measuring the overlap of bigrams (consecutive word pairs), and 3) **ROUGE-L** considering the longest common subsequence and capturing sentence-level structure similarity [9]. For each of the three ROUGE sub-metrics, three essential statistical measures are: 1) **Precision (p)**, the ratio of overlapping n-grams between the generated and reference texts to the total n-grams in the generated text, reflecting the proportion of relevant content produced; 2) **Recall (r)**, the ratio of overlapping n-grams to the total n-grams in the reference text, reflecting how much of the reference content was captured in the generated text; and 3) **F1 score (f)**, which balances precision and recall to account for both false positives and false negatives.

For contextual embeddings from pre-trained BERT models, **BERTScore** assesses the similarity between the generated text and reference text. It is used to compute cosine similarity between token embeddings, providing a more nuanced evaluation that captures semantic similarity beyond surface-level n-gram overlap [22]. Three components, **Precision (P)**, **Recall (R)**, and **F1 score (F1)** of BERTScore, are similar to r, p, and f of ROUGE.

Listing 7. The Prompt for grading the answers generated by fine-tuned LLMs

```
prompt = f"""You are an expert evaluator.
Your task is to grade the correctness of an answer to a question on a scale from 0 to 10, where:
- 10 means that the answer is completely correct and fully explains the concept.
- 0 means that the answer is completely wrong.
- between 0 to 10 means that the answer is partly correct and partly wrong.
Provide a numerical score to evaluate the answer of the question, based on the following text:
Question: {question}
Answer: {provided_answer}
Text: "{text}" """
```

Furthermore, we use the OpenAI model gpt-4o (2024-11-20 version), which can be regarded as a human evaluator to assess the quality of the textual generations of other LLMs and can obtain competitive evaluation scores with human evaluators [19]. The corresponding prompt is given for evaluating the correctness of answers (in our results) to questions (in our testing data) ranging from 0 to 10, as shown in Listing 7.

Finally, there are three types of outputs: fine-tuned LLMs, testing results and metrics, and evaluation scores.

4 EXPERIMENTS AND RESULTS

Our experiments aim to demonstrate the effectiveness of the developed workflow. We selected the EnergyPlus Engineering Reference document as our input due to its comprehensive and detailed documentation of the EnergyPlus simulation application, which serves as a strong example of professional software documentation important for building energy. This reference offers in-depth insights into the principles, mathematical models, and algorithms that underpin EnergyPlus, which are essential for accurate simulations. From this authoritative source, our question-answering system is built on validated and reliable methodologies, thus enhancing the credibility and precision of its responses [18].

Developed by the U.S. Department of Energy (DoE), EnergyPlus is a type of comprehensive building energy simulation software. It models cooling, heating, lighting, ventilation, and other energy flows, enabling detailed analysis of building energy performance. EnergyPlus integrates the best features of two earlier building energy simulation tools, BLAST and DOE-2, while introducing advanced capabilities for modeling complex building systems and innovative technologies. It employs a modular system and plant integrated solution manager, allowing users to simulate a wide range of HVAC configurations and control strategies. The software supports sub-hourly time steps for system dynamics and offers flexible input and output options, facilitating detailed and customizable simulations. EnergyPlus has been utilized in various research studies, including the development and validation of HVAC controllers for residential and small commercial buildings, as well as the optimization of building energy models through auto-tuning techniques. Its continuous development and validation make it a reliable tool for architects, engineers, researchers, and other stakeholders, aiming to design energy-efficient and low-carbon buildings [2].

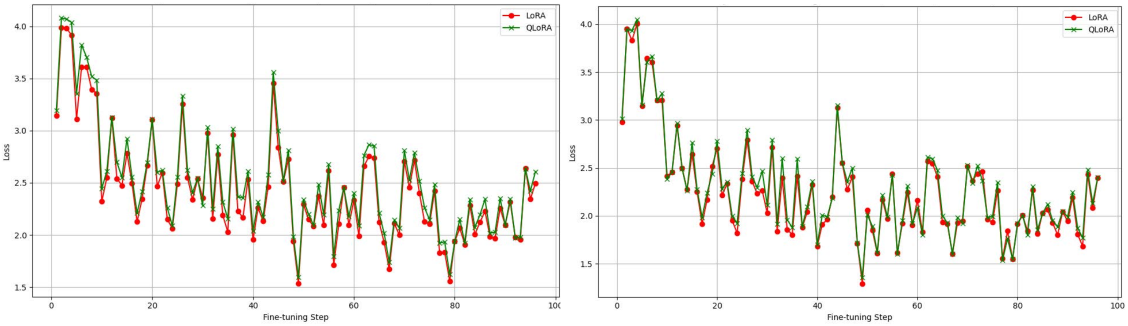
Developed by Meta AI, Llama (Large Language Model Meta AI) is a series of LLMs for translating languages, generating human-like text, and tackling a wide range of creative and informative tasks. The architecture of Llama models is based on autoregressive decoder-only transformers, utilizing SwiGLU activation functions, rotary positional embeddings (RoPE), and RMSNorm for layer normalization. Llama's open-source nature has facilitated widespread research and development, leading to various fine-tuned versions and applications across different domains [17]. Llama 3.2, released in September 2024, marks a significant advancement in LLMs by introducing multimodal capabilities, enabling the processing of both textual and visual data. This development facilitates the creation of advanced AI applications, such as augmented reality tools, visual search engines, and comprehensive document analysis systems. Unlike earlier versions, Llama 3.2 is designed to process both text and images. This multimodal functionality enables applications in augmented reality, visual search engines, and document analysis, broadening its utility beyond text-based tasks [14]. The release of Llama 3.2 underscores Meta's commitment to advancing AI technology by providing open-source models that are both versatile and accessible, promoting innovation across various platforms and applications.

Llama 3.2 offers two lightweight text-only models with 1 billion parameters and 3 billion parameters. These smaller models are optimized for mobile devices, functioning efficiently on hardware platforms such as Qualcomm and MediaTek, thereby extending advanced AI capabilities to mobile applications. These lightweight text-only models are specifically designed to operate on mobile hardware, facilitating the development of AI applications that are both powerful and portable [14]. Furthermore, Meta has augmented Llama 3.2 with voice capabilities, enabling AI to interpret and generate audio data. This feature supports functionalities such as live translation, automatic video dubbing, and personalized voice interactions, including the use of celebrity voices such as Dame Judi Dench and John Cena. Llama 3.2 two lightweight text-only models with 1 billion and 3 billion parameters were exploited in our experiments.

The machine used in our experiments was an Ubuntu 22.04.5 LTS workstation equipped with two Intel® Xeon® Gold 6438Y+ CPUs, 1007 GB of DIMM memory, 15 TB hard disks, and 4 GPUs, including three 80 GB NVIDIA H100 and one 4GB NVIDIA T400.

4.1 Results and Discussion

Our experiments showcase building energy question-answering applications using various LLMs and fine-tuning techniques. Figure 5 presents the loss curves for fine-tuning Llama 3.2 1B and 3B models using LoRA and QLoRA. The curves exhibit similar patterns, all showing a general downward trend with some fluctuations.



(a) 1B (b) 3B

Fig. 5. The Loss Curves of Fine-tuning Llama 3.2 1B and 3B versions, including LoRA and QLoRA methods

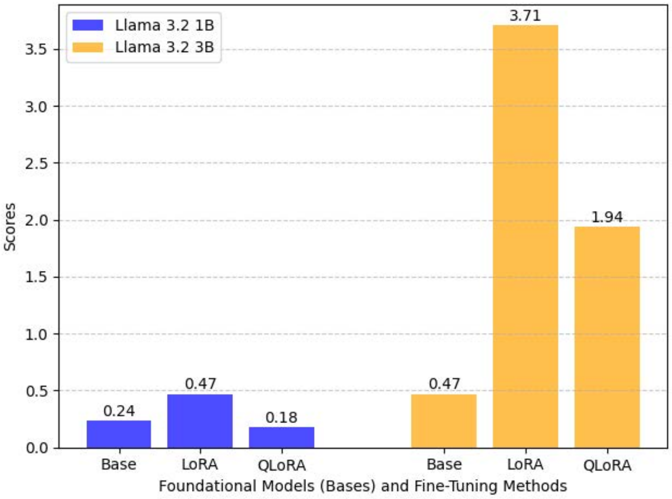


Fig. 6. The Evaluation Scores of Llama 3.2 1B (left) and 3B (right) Versions.

Figure 7 shows that several metrics—BLEU and ROUGE-1/2/L—are similar across the base, LoRA fine-tuned, and QLoRA fine-tuned versions of Llama 3.2 1B. However, the LoRA fine-tuned version achieves the highest BERTScores, while the base version yields the lowest, even negative, scores. For Llama 3.2 3B, the LoRA fine-tuned model is equal to or outperforms the QLoRA fine-tuned model that is better than the base, as shown in Figure 8. And from Figure 6 scores, we also know the order of the evaluation scores: LoRA fine-tuned version > QLoRA fine-tuned version > foundational version (base). And Llama 3.2 3B versions outperform the counterparts of Llama 3.2 1B. In summary, for the same small-scale LLMs (e.g., 1 billion or 3 billion parameters), LoRA fine-tuning achieves better or similar performance than QLoRA, despite QLoRA's lower memory usage. Additionally, within the same model type, larger parameter sizes generally lead to the performance improvement.

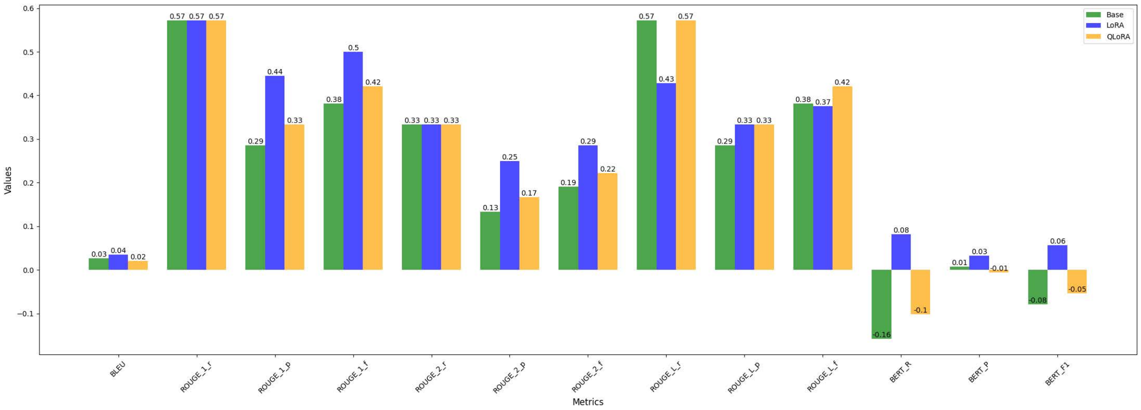


Fig. 7. The comparison of LLama 3.2 1B Base, LoRA Fine-tuning, and QLoRA Fine-tuning Metrics.

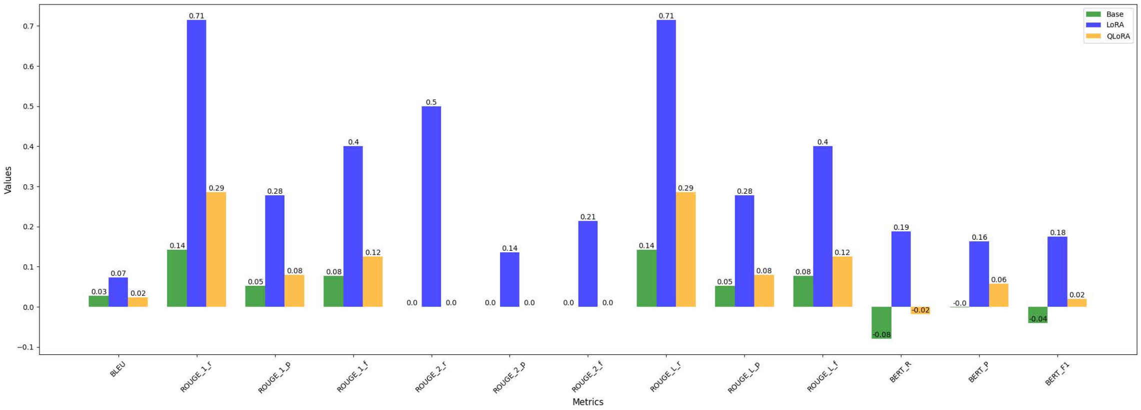


Fig. 8. The comparison of LLama 3B Base, LoRA Fine-tuning, and QLoRA Fine-tuning Metrics.

5 LIMITATIONS

Despite the encouraging results and promising performance of the fine-tuned Llama models, this project is subject to several limitations that should be acknowledged. First, the dataset used for fine-tuning and evaluation was derived exclusively from a single document, the EnergyPlus Engineering Reference. This limited source base reduces the overall diversity and coverage of domain-specific knowledge, potentially constraining the model's ability to generalize to other topics within building energy systems.

Secondly, the study was confined to the Meta Llama 3.2 models, specifically the 1B and 3B parameter variants. Although these models demonstrated significant improvements after fine-tuning, the investigation did not extend to other language models such as Falcon, BLOOM, and larger versions of Llama, which may have yielded additional insights or performance gains. Moreover, resource constraints prevented experimentation with models beyond 3 billion parameters, leaving open the question of how performance scales with significantly larger architectures.

Another limitation lies in the prompt engineering approach used to generate improved instruction–question–answer examples. Although the prompts were carefully designed to reflect realistic and informative tasks, they were not optimized through iterative testing or fine-grained tuning. A more sophisticated prompt design could improve both the quality of the training data and the resulting model performance.

The evaluation methodology also presents constraints. It relied primarily on metrics (i.e., BLEU, ROUGE, and BERTScore) to assess the similarity between the model's outputs and the reference answers. Although these metrics provide useful quantitative indicators, they do not fully capture the nuance or contextual relevance of the generated data. Furthermore, the test set was drawn from the same source document as the training data, which limits the ability to assess how well the model generalizes to unseen materials or real-world application scenarios.

Additionally, the performance of fine-tuned models (Llama 3.2 1B and 3B) using LoRA/QLoRA can be unstable and imprecise. For example, once uploaded to and executed from the Hugging Face Hub, the models often produce varying execution metrics, results, and evaluation scores. Some of their outputs are non-sensical or unhelpful. Fine-tuning larger LLMs could improve both the model quality and the coherence of their responses.

In summary, while the project shows that fine-tuning compact LLMs with domain-specific data can yield meaningful performance improvements, the limited training data, evaluation scope, and model instability and inaccuracy constrain the generalizability and scalability of the results.

6 CONCLUSION AND FUTURE WORK

We propose a systematic approach to data preparation, fine-tuning, and evaluation of LLMs for building energy applications. We began by extracting raw text from the "EnergyPlus Engineering Reference" PDF and then enhanced the text using OpenAI's GPT-4o and GPT-4o-mini to generate QA pairs for both training and evaluation datasets in structured JSON format.

We fine-tuned two parameter-scale variants (1B and 3B) of Meta's open-source Llama 3.2 using LoRA and QLoRA techniques. Experimental results show that models with more parameters generally achieve better performance, and LoRA fine-tuning consistently outperforms QLoRA, despite the latter's lower memory requirements. These findings offer meaningful insights into how fine-tuned LLMs can support question-answering tasks in the domain of building energy, while also providing a replicable workflow for data preparation, model selection, and fine-tuning strategy.

Looking ahead, we plan to expand our work across multiple domains (e.g., healthcare, education), and incorporate additional reference documents (e.g., ASHRAE Fundamentals) as the domain-specific data to enrich the domain-specific knowledge base. Future experiments will explore the integration of both QA pairs and improved or parsed textual data in fine-tuning processes. We are also considering advanced fine-tuning techniques like "Weight-Decomposed Low-Rank Adaptation (DoRA)", which improves model adaptability by decomposing weights into directional and magnitude components [10, 13].

Furthermore, we intend to apply other LLMs, including OpenAI (e.g., GPT-4.5), BLOOM (with 176 billion parameters), Falcon (with 7 billion, 40 billion, or 180 billion parameters), and Llama (along with their new versions, such as Llama 4.0 and Llama 3.2 vision model with 90 billion parameters), to our new question-answering system and upcoming experiments across different scenarios. Moreover, to further enhance the quality of training and evaluation data, we will refine our prompt engineering strategy by defining clearer prompt structures, specifying expected formats and contextual constraints, including examples of ideal inputs/outputs, and employing step-by-step instruction patterns for LLM guidance.

REFERENCES

[1] World Green Build Council. 2019. Bringing embodied carbon upfront. https://worldgbc.org/advancing-net-zero/embodied-carbon/

[2] D. E. Crawley, M. J. Witte, and J. Glazer. 2001. EnergyPlus: Creating a New-Generation Building Energy Simulation Program. *Energy and Buildings* 33, 4 (2001), 319–331.

[3] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient Finetuning of Quantized LLMs. In *Advances in Neural Information Processing Systems*, Vol. 36. https://arxiv.org/abs/2305.14314

[4] Lin Gao, Jing Lu, Zekai Shao, Ziyue Lin, Shengbin Yue, Chokit Ieong, Yi Sun, Rory James Zauner, Zhongyu Wei, and Siming Chen. 2025. Fine-Tuned Large Language Model for Visualization System: A Study on Self-Regulated Learning in Education. *IEEE Transactions on Visualization and Computer Graphics* 31, 1 (2025), 1353–1367. https://doi.org/10.1109/TVCG.2024.3456145

[5] Taryn Holowka. 2024. Building decarbonization to combat climate change. https://www.usgbc.org/articles/building-decarbonization-combat-climate-change

[6] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. 2022. LoRA: Low-Rank Adaptation of Large Language Models. *arXiv preprint arXiv:2106.09685* (2022).

[7] D. Hugo, J. McCulloch, A. Hameed, W. Borghei, M. Grimeland, V. Felstead, and M. Goldsworthy. 2023. A smart building semantic platform to enable data re-use in energy analytics applications: the Data Clearing House.

[8] Gang Jiang, Zhihao Ma, Liang Zhang, and Jianli Chen. 2024. EPlus-LLM: A Large Language Model-Based Computing Platform for Automated Building Energy Modeling. *Applied Energy* 367 (2024), 123431. https://doi.org/10.1016/j.apenergy.2024.123431

[9] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*. Association for Computational Linguistics, Barcelona, Spain, 74–81. https://aclanthology.org/W04-1013 WAS 2004: Workshop on Text Summarization Branches Out.

[10] Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. 2024. DoRA: Weight-Decomposed Low-Rank Adaptation. In *Proceedings of the 41st International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 235)*. PMLR, 32100–32121. https://proceedings.mlr.press/v235/liu24bn.html Oral Presentation.

[11] NVIDIA. 2024. NeMo QLoRA Guide. https://docs.nvidia.com/nemo-framework/user-guide/24.07/sft_peft/qlora.html

[12] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*. Association for Computational Linguistics, 311–318. https://doi.org/10.3115/1073083.1073135

[13] Venkatesh Balavadhani Parthasarathy, Ahtsham Zafar, Aafaq Khan, and Arsalan Shahid. 2024. The Ultimate Guide to Fine-Tuning LLMs from Basics to Breakthroughs: An Exhaustive Review of Technologies, Research, Best Practices, Applied Research Challenges and Opportunities. https://arxiv.org/abs/2408.13296. (2024). arXiv preprint arXiv:2408.13296.

[14] Kylie Robison. 2024. Meta Releases Its First Open AI Model That Can Process Images. https://www.theverge.com/2024/9/25/24253774/meta-ai-vision-model-llama-3-2-announced. Accessed: 2025-04-29.

[15] Marta Schantz. 2024. Unlocking Whole-Building Energy Data: Commercial Building Owner Challenges and Industry Solutions. https://urbanland.uli.org/resilience-and-sustainability/unlocking-whole-building-energy-data-commercial-building-owner-challenges-and-industry-solutions

[16] Pradeepa somasundaram. 2024. Overcoming Barriers to Knowledge Management Adoption in the Energy Industry. https://document360.com/blog/knowledge-management-in-energy-sector

[17] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and Aurelien Rodriguez. 2023. LLaMA: Open and Efficient Foundation Language Models. *arXiv preprint arXiv:2302.13971* (2023). https://arxiv.org/abs/2302.13971

[18] U.S. Department of Energy. 2010. *EnergyPlus Engineering Reference: The Reference to EnergyPlus Calculations*. U.S. Department of Energy, Washington, D.C. https://energyplus.net/documentation.

[19] Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui Sun, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023. Is ChatGPT a Good NLG Evaluator? A Preliminary Study. In *Proceedings of the 4th New Frontiers in Summarization Workshop*. Association for Computational Linguistics, Singapore, 1–11. https://doi.org/10.18653/v1/2023.newsum-1.1

[20] Jian Zhang, Chaobo Zhang, Jie Lu, and Yang Zhao. 2025. Domain-specific Large Language Models for Fault Diagnosis of Heating, Ventilation, and Air Conditioning Systems by Labeled-Data-Supervised Fine-Tuning. *Applied Energy* 377 (2025), 124378. https://doi.org/10.1016/j.apenergy.2024.124378

[21] Liang Zhang, Zhelun Chen, and Vitaly Ford. 2024. Advancing Building Energy Modeling with Large Language Models: Exploration and Case Studies. *Energy and Buildings* 323 (2024), 114788. https://doi.org/10.1016/j.enbuild.2024.114788

[22] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In *Proceedings of the 8th International Conference on Learning Representations (ICLR)*. OpenReview.net. https://openreview.net/forum?id=SkZp-AcKPS

APPENDIX:

A ONLINE RESOURCES

The code and data to reproduce the results can be found at https://github.com/fuArizona/INFO621Final.

In particular, the videos demonstrating the execution of programs (i.e., LoRAQLoRA_FinetuneLlama_1B, LoRAQLoRA_FinetuneLlama_3B, Fine-tuned_Llama_1B_results, and Fine-tuned_Llama_3B_results) on the server are provided in the repository folder https://github.com/fuArizona/INFO621Final/tree/main/videos.

The following fine-tuned Llama 3.2 models are available on Hugging Face:

Llama 3.2 1B fine-tuned with LoRA: xfu20/Building_energy_Llama_1B_LoRA
Llama 3.2 1B fine-tuned with QLoRA: xfu20/Building_energy_Llama_1B_QLoRA
Llama 3.2 3B fine-tuned with LoRA: xfu20/Building_energy_Llama_3B_LoRA
Llama 3.2 3B fine-tuned with QLoRA: xfu20/Building_energy_Llama_3B_QLoRA