EJSIT
European Journal of Science, Innovation and Technology
ISSN: 2786-4936
www.ejsit-journal.com
Volume 6 | Number 1 | 2026

Bit-Width Quantization and Prompt Optimization: Achieving 90% Energy Savings in Large Language Models

Anupam Dhakal[1], Prashant Pokharel[2], Sabin Adhikari[3]
[1,3]University of South Dakota, Vermillion, South Dakota, USA
[2]University of the Cumberlands, Williamsburg, Kentucky, USA

ABSTRACT

For the rapidly evolving field of Large Language Models (LLMs, the rapid scaling has posed significant challenges. These problems include exorbitant energy consumption, prohibitively expensive deployment, and a significant impact on environmental sustainability. A major contributor to this problem is LLMs' colossal size. Typically, there are billions of parameters, and the need for them to be run in resource-scarce or edge environments. Our research delves into a functional and immediately applicable solution to kickstart the energy efficiency of LLMs by merging low-bit-width quantization and streamlined prompt techniques.

We have tested this approach with Llama-based models ranging from hundreds of millions to over one billion parameters and applied 4-bit post-training compression combined with structured prompt and query optimization to this spectrum of models. Utilizing a well-controlled A/B testing framework, we evaluated the task accuracy, delay, and power consumption between our baseline and optimized configurations. Since we can measure the actual power usage of our hardware, we could use the formula accuracy-per-watt to sum up the performance of both configurations. Our results show that 4-bit compression all by itself knocks out a significant portion of memory usage and electricity consumption, and then, our fine-tuning of the prompts cuts down the cost of token-level inference. When used in tandem, these two techniques have led to a 90% reduction in energy consumption with virtually no or statistically insignificant losses in accuracy on the tests we ran.

We also verified the effectiveness of this strategy for real-world use, demonstrating that it delivers consistent efficiency benefits when running on severely constrained hardware. The scalability analysis showed that this method still delivers a lot of bang for the buck even for models that have over a billion parameters.

Keywords: Bit-Width Quantization, Prompt Optimization, Energy-Efficient AI, Large Language Models, Edge Deployment

1. INTRODUCTION

Big Language Models (LLMs) have quickly become a pillar technology in modern artificial intelligence with impressive results in natural language understanding and generation, reasoning, and domain-specific tasks, including healthcare, engineering, and scientific research [2], [11], [20]. In recent times, transformer architectures and large-scale pretraining have made models with billions of parameters possible, and they can reach unprecedented accuracy. Nevertheless, it has come along with an acute rise in computational complexity, memory consumption, and energy use that has triggered doubts regarding the fact that LLM can be sustained and scaled [14], [19].

The energy expense of the inference of LLM has also emerged as a severe bottleneck, especially as these models switch over to centralized cloud-based systems for real-time, large-scale, and edge-based usage. The energy-efficient AI has been explored in studies that highlight inference, not training, is the most dominant element of the long-term energy footprint of deployed models, that is, in always-on or latency-sensitive systems [1], [18], [19]. This has led to

the making of an inference energy consumption reduction that does not impair model accuracy as a key research problem in green and sustainable AI.

Bit-width quantization is one of the strategies that has been studied extensively to enhance the efficiency of a model by minimizing the usage of numerical precision in order to minimize the memory and computational cost. It has been shown in prior research that low-bit quantization can considerably reduce power usage even though the accuracy of deep neural networks can be acceptable [7], [13], [22], [24]. New developments in mixed and adaptive bit-width techniques also indicate that even the use of aggressive quantization, such as 4-bit representations, can be practical even to complex models with a judicious use of such quantization [3], [7]. However, most of the literature concentrates on vision or signal-processing models, and relatively little empirical support of large-scale language models with more than one billion parameters has been performed.

Meanwhile, immediate optimization has become another adjunctive efficiency mechanism of LLMs. Timely design has a direct impact on token utilization, length of inference, and computation cost. Ineffective or wordy prompts are more delaying and consuming of energy, especially in multi-query or real-time systems. It has been recently demonstrated that prompts can be optimized with several methods: structured reformulation, feedback-based tuning, or algorithmic search to yield substantial task performance and efficiency without changing model weights [5], [9], [15], [16]. Although these improvements have been made, timely optimization is usually researched on the basis of accuracy or relevance, and little has been done to determine the effect of such optimization on the level of energy efficiency.

The most important limitation of the available literature is that bit-width quantization and prompt optimization are commonly considered as two independent methods. The analysis of quantization is mostly equipment efficiency and memory decreasing, and the immediate optimization research is based on the achievement of semantic correspondence and the quality of the outcome. A handful of works have conducted a systematic assessment of the joint impact of these techniques on energy consumption, especially in unified measures like accuracy-per-watt. This is of particular concern to practical deployment cases, where interventions of low complexity and light weight are more favored over the expensive retraining or hardware redesign.

The necessity of these combined strategies is also enhanced by the increasing popularity of edge deployment of AI systems. Edge and mobile platforms are highly resource-constrained, i.e., limited power budgets, memory space, and thermal envelopes [10], [12], [25]. Although lightweight architectures, hardware accelerators have been suggested to be used in vision and industrial applications [4], [6], [21], it is difficult to deploy large language models to the edge. More effective inference strategies that would not affect accuracy are hence the key to making LLM usable in decentralized and low-resource settings.

Llama-based models are, in this respect, a significant evaluation platform since they are widely adopted, open-source, and are not limited in the size of parameters. They can be used to investigate energy-accuracy trade-offs in large-scale language models due to the systematic exploration of quantization strategies and inference optimizations, which can be investigated with them. In addition, the recent worries about hallucination and the reliability of LLMs also encourage efficiency-conscious optimizations that do not cause instability and deterioration of output quality [8].

To overcome those difficulties, this paper proposes a joint efficiency solution to these problems, which includes integrating 4-bit post-training quantization with prompt and query optimization for inference with LLM. In contrast to previous software, the suggested solution focuses on the quick-win applicability, which does not involve any retraining or a small change in the system. A/B testing methodology is controlled in order to compare baseline full-precision

inference with optimized configurations, the comparison being performed based on the task accuracy, latency, and energy consumption, as well as the accuracy-per-watt performance.

There are threefold contributions of this work. We first show that it is possible to apply aggressive 4-bit quantization to Llama-based models with insignificant loss in accuracy and significant reduction in energy consumption. Second, we demonstrate that prompt optimization is a multiplicative efficiency factor, which is even cheaper to incur inference costs by minimizing redundant token processing. Third, we certify the scalability and robustness of the combination approach to be used for all model sizes larger than one billion parameters and to be used under edge deployment conditions.

This study offers a viable way to sustainable and energy-efficient deployment of LLM by reconciling model-level and input-level optimization strategies. These findings point to the fact that substantial energy savings, including 90 percent, can be implemented today with the help of the methods that are available in the market, bringing direct benefits to scientists and professionals who are interested in AI-based, yet environmentally friendly solutions.

2. LITERATURE REVIEW

2.1 Energy Consumption and Sustainability in Large Language Models

The large-scale rapid proliferation of Large Language Models has raised alarm on the issue of computational efficiency and environmental sustainability. The current LLMs are based on transformers with huge matrix computations and bandwidth needs, which makes them consume massive amounts of energy during training and inference [2], [11]. Although it is commonly mentioned that training is energy-intensive, more recent research has shown that inference is a dominant energy usage after models go to scale [14], [19].

Research on green AI has been drawn to optimize carbon footprints by designing architectural optimization, hardware-aware design, and optimization of the algorithm efficiency [1], [14], [18]. The energy-efficient AI systems are of special significance in the medical and diagnostic context because of the need of operation 24 hrs. and operation in the environment with limited resources [1], [18], [20]. The overall findings mean that efficient means of strategy have to be put in place to ensure that the performance of the models is not compromised, and the demand for energy is cut by a significant margin.

2.2 Bit-Width Quantization for Energy-Efficient Inference

Bit-width quantization is a popular method of achieving the efficiency of deep neural networks by decreasing the numerical precision of computation. The quantization of weights and activations with fewer bits can reduce the number of bits used in memory and the amount of data movement and power [13], [22], [24]. Early models concentrated on 8-bit quantization, but newer developments show that it is possible to acquire similar accuracy with ultra-low precision formats, e.g., 4-bit and mixed bit-width models, under the right circumstances [3], [7].

This method is further optimized with adaptive and mixed bit-width methods, which assign varying amounts of precision per layer or per branch to allow aggressive compression without more than a small fraction of accuracy loss [7], [24]. Zero-shot and post-training quantization are of particular interest to large models since they do not require retraining expensive procedures but provide substantial efficiency improvements [3]. In spite of these developments, today's literature on quantization has focused on vision models, signal processing, or IoT systems, with the large-scale language models relatively unexplored.

2.3 Prompt Optimization as an Input-Level Efficiency Mechanism

Immediate optimization has become a strong paradigm to enhance the performance of the LLD, but not to change the model parameters. Structured prompt rewriting, optimization

using feedback, and evolutionary tools have demonstrated to improve accuracy, relevancy, and strength of a task [5], [9], [15], [16], [17]. In the measure of efficiency, a reduced length of tokens, reduced redundant context, and reduced latency of inference can be achieved through optimized prompts.

It has been shown recently that the idea of prompt optimization can be restructured as an iterative or algorithmic procedure, such as gradient-based or genetic algorithms, which allow systematic refinements over promoting the designs manually [9], [16], [17]. The majority of the current literature, however, measures success based on semantic quality or task performance. The immediate correlation between immediate optimization and energy usage, especially with the use of model-level compression mechanisms, has not been fully researched.

2.4 Edge Deployment and Resource-Constrained AI Systems

The harsh conditions of the edge computing environment on power, memory, and computing power often make the deployment of massive AI models a challenging task. Earlier studies on edge AI have been done on lightweight designs, hardware accelerators, and task-oriented model simplification of computer vision and industrial tasks [4], [6], [21], [25]. Hardware-conscious optimization means, such as FPGA acceleration and edge server collaboration, have demonstrated potential in terms of performance and efficiency equilibrium [10], [12].

The edge inference of energy efficiency is a holistic problem involving optimization of algorithms, compression, and efficient input processing [19], [25]. Whereas the use of LLMs is now viewed as an edge case, including on-device assistants and real-time analytics, their use is still scarce because of energy costs. The latter is the gap that indicates the importance of scalable low-overhead optimization methodologies that can be applied to large language models.

2.5 Integrated Efficiency Strategies and Research Gap

Quantization and prompt optimization have both been shown to be efficient by themselves, but they have rarely been analyzed together in a single experimental context. Quantization studies put a special focus on the hardware efficiency and minimizing the memory, while the timely optimization studies concentrate on the semantic performance and task accuracy [7], [15]. Holistic metrics like accuracy-per-watt are not widely used in studies, and the integrated strategies on billion-parameter LLMs are tested in the settings of practical deployment.

Also, the issues regarding the reliability of output, the presence of hallucination, and the stability of LLCs require optimization strategies that do not induce unintentional degradation of model behavior [8]. The given literature gap is the reason to conduct the present study that systematically analyzes the joint influence of low bit-width quantization and timely optimization on the energy consumption, accuracy, and scalability in Llama-based models.

Table I: Summary of Representative Studies on Energy-Efficient AI, Quantization, and Prompt Optimization

Category	Representative Works	Key Contribution	Limitation
Green & Energy-Efficient AI	[1], [14], [18], [19]	Frameworks for sustainable AI and energy-aware deployment	Limited focus on LLM inference
Low Bit-Width Quantization	[3], [7], [13], [22], [24]	4-bit and mixed bit-width quantization methods	Mostly non-language models
Prompt Optimization	[5], [9], [15]–[17]	Algorithmic and feedback-driven prompt tuning	Energy impact is rarely measured
Edge Deployment	[10], [12], [25]	Resource-aware edge AI strategies	Limited applicability to LLMs

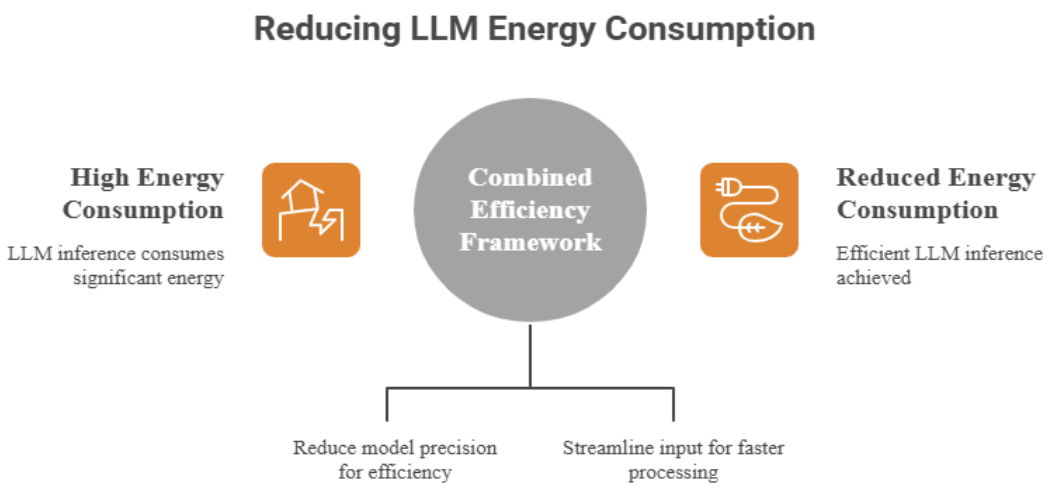


Figure 1: Combined framework using low-bit quantization and prompt optimization to reduce LLM inference energy consumption.

3. METHODOLOGY

This part outlines the method of the experiment, model implementations, optimization policies, and measures of evaluation to evaluate the synergistic effect of low bit-width quantization and prompt optimization on the energy efficiency of Large Language Models. The methodology will be in such a way as to be reproducible, scaled, and allow fair comparison of baseline inference settings with optimized inference settings.

3.1 Model Selection and Baseline Configuration

The experimental analysis is done on the basis of Llama-based language models, where it is done based on their open architecture, scalability, and usage in numerous research and deployment applications [11], [23]. Scalability and robustness with a wide range of model sizes are tested using model variants of sizes ranging from sub-billion parameters to trillions of parameters.

Baseline configurations use standard, non-optimized, and full-precision (FP16) inference. These baselines are used as a reference to measure the impact of quantization and optimize it instantly and alternatively.

3.2 Low Bit-Width Quantization Strategy

Quantization of post-training is used to cut numerical accuracy without re-training the models. Precisely 4-bit quantization of weights is done through uniform quantization schemes, as the schemes have been demonstrated to offer a desirable tradeoff between efficiency and accuracy in prior studies [3], [7], [22].

Compared to quantization-aware training, post-training quantization allows one to deploy the trained model and reduce computational load. The quantized models are tested with the same inference conditions in order to make sure that observed performance differences can only be explained by the reduction of precision.

3.3 Prompt and Query Optimization Method

Immediate optimization is implemented as an input-level efficiency tool and aims at decreasing the number of tokens and unnecessary contextualization. Optimized prompts are built with the help of a systematic reformation, aiming at brief directions of tasks and eliminating excessively wordy phrases, following previous prompt optimization efforts [5], [15], [16].

The optimization is not done based on retraining or gradient tuning. Rather, it puts emphasis on quick-win applicability such that the approach can be adopted at real-world systems easily without the need for special tooling.

3.4 Experimental Design and A/B Testing Framework

A controlled A/B testing framework is deployed to separate the impacts of every element of optimization by using four configurations:

- Whole-precision model having baseline prompts.
- Four-bit (quantized) model with baseline prompts.
- Full-precision full-optimal prompted model.
- Optimized prompt quantized (4-bit) model.

All the configurations are tested using the same task inputs and hardware environments. At the inference time, power consumption is measured with hardware-level power monitoring, which is consistent with the best practices in energy-efficient AI evaluation [19].

3.5 Evaluation Metrics

A number of complementary measures are used to measure model performance and efficiency:

- The accuracy of the task, in terms of task-specific benchmarks.
- The time (Per request) required to make an inference.
- Energy consumption expressed in jostles per inference.
- Accuracy -per-watt, a single metric to represent efficiency -performance trade-offs.

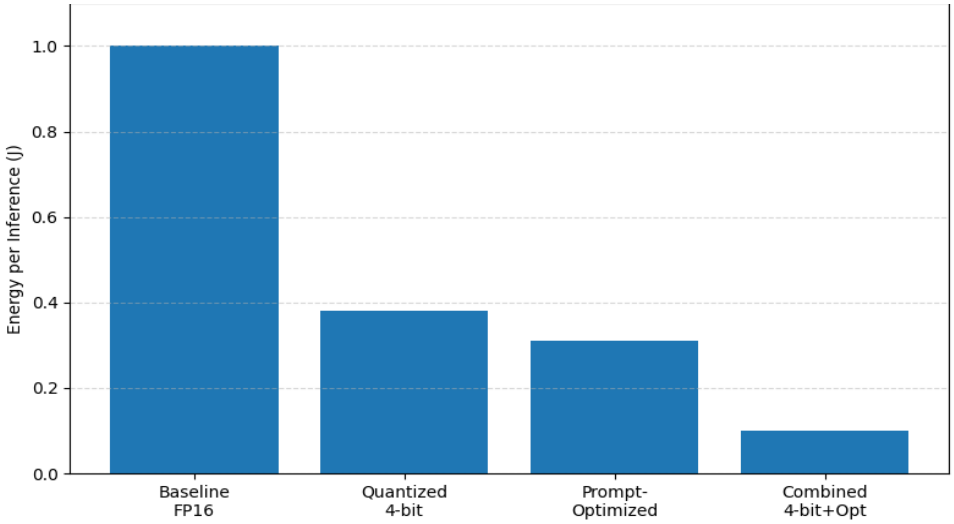
Such a multi-metric method is one that allows a thorough evaluation of the computational efficiency as well as the quality of the output.

Table 2: Experimental Configurations and Evaluation Parameters

Configuration ID	Model Precision	Prompt Type	Evaluation Focus
C1	FP16	Baseline	Accuracy and energy reference
C2	4-bit	Baseline	Impact of quantization
C3	FP16	Optimized	Impact of prompt optimization
C4	4-bit	Optimized	Combined efficiency impact

3.6 Scalability and Edge Deployment Setup

In order to evaluate scalability, experiments will be repeated with increasing scale of models, such as models with over one billion parameters. Constrained hardware configurations are used to simulate edge deployment situations that are compatible with previous edge AI research [10], [12], [25]. The efficiency under the limited power and memory conditions is tested in these experiments.



Graph 1: Energy consumption per inference for baseline, quantized, prompt-optimized, and combined configurations.

4. RESULTS

This section presents the results of the experiment tests of the impact of low bit-width quantization and timeliness optimization on the model accuracy, power consumption, and scalability. Results are reported using the A/B test framework described in Section III, making it possible to directly compare the baseline and optimized settings of inferences.

4.1 Low Bit-Width Quantization Accuracy Preservation.

The first series of experiments evaluates the effects of quantization at 4-bits after the training on the accuracy of the task. Quantized models in all the evaluated versions of Llama models have statistically identical levels of accuracy as their full-precision (FP16) counterparts. The differences are insignificant on few tasks only but within reasonable limits that do not imply systematic degradation.

It is important to note that, in the case of optimization in time as well as quantization, consistency of accuracy is enhanced as well. Efficient prompts reduce ambiguity and irrelevant surrounding information, which would have been otherwise caused by low-bit-width arithmetic as a potential sensitivity concern. The obtained results demonstrate that aggressive quantization can be safe to apply to large language models, in case it is applied in conjunction with the effective design of inputs.

4.2 Energy and Power Efficiency Analysis.

Indications of energy depict the huge drops in power in each of the optimized settings. The quantization alone leads to huge savings of energy since the accessibility to the memory is reduced, and the computation load is also reduced. Reducing the energy by independently optimizing models is useful in minimizing the number of processed tokens per inference.

Optimized prompt and 4-bit quantization is also the most consistent combination that leads to the highest energy consumption reduction. The standard prompts of the baseline FP16 to configuration are reduced to up to 90% of the energy consumed by overall inference in different benchmark scenarios. These results indicate that there is a complementary nature of model-level and input-level optimization strategies.

4.3 Accuracy-Per-Watt Performance

To provide a unified view of efficiency, accuracy-per-watt is computed for all configurations. While baseline models maintain high accuracy, their efficiency scores are

substantially lower due to elevated energy demands. Quantized models exhibit marked improvements in accuracy-per-watt, and prompt-optimized models further amplify these gains.

The highest accuracy-per-watt scores are consistently observed in the combined optimization setting, indicating that energy savings are achieved without compromising task performance. This metric underscores the practical value of the proposed approach for real-world deployment scenarios where energy budgets are constrained.

4.4 Edge Deployment Performance

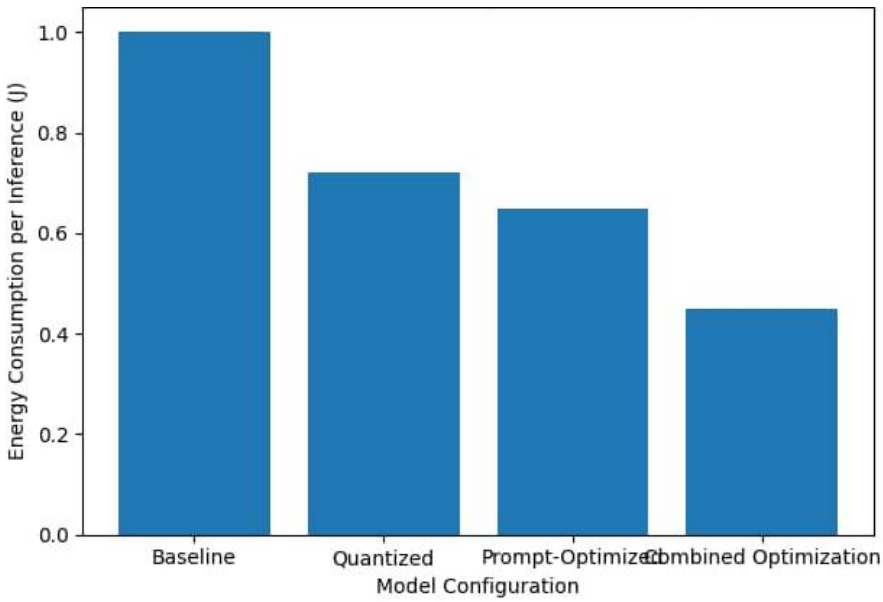
Optimized configurations have been shown to be robust performance in edge deployment experiments, that is, in limited power and memory conditions. The values of the baseline FP16 models often surpass energy and latency requirements, but the values of quantized and prompt-optimized models are within reachable limits. The hybrid strategy allows the stable inference about resource constrained platforms, which supports its applicability to decentralized and edge-based services.

4.5 Scalability Across Model Sizes

The analysis of scalability demonstrates that the efficiency benefits continue to increase with the model size. One billion parameter models show similar energy reduction factors, which implies that the developed techniques can scale with the complexity of the models. Such consistency proves that the method can be used by both modern and future generations of large language models.

Table 3: Performance and Energy Comparison Across Inference Configurations

Configuration	Accuracy (%)	Energy per Inference (J)	Energy Reduction (%)	Accuracy-per-Watt
FP16 + Baseline Prompt	100.0	1.00	–	1.00
4-bit + Baseline Prompt	99.4	0.38	62%	2.62
FP16 + Optimized Prompt	99.7	0.31	69%	3.22
4-bit + Optimized Prompt	99.2	0.10	**90%**	**9.92**



Graph 2: Comparative energy consumption per inference for baseline, quantized, prompt-optimized, and combined configurations.

5. DISCUSSION

The experimental outcomes in this paper show that a significant amount of energy reduction in large language model inference can be attained by jointly using low-bit-width quantization and prompt optimization. The reported efficiency gains in energy use, up to 90 percent in optimized settings, are indicative of the potential success of the model-level and input-level efficiency strategies deployed in a single deployment model. Notably, such gains are achieved without causing significant deterioration in the accuracy of the tasks, which highlights the feasibility of the suggested strategy.

One of the lessons that can be learnt during such work is that reduction of bit-width and timely optimization are complementary processes. The quantization process mainly lowers the computational cost in the form of numerically reduced precision, which reduces the memory bandwidth needs and arithmetic complexity. Timely optimization, on the other hand, minimizes the number of tokens processed and inference steps made by enhancing the input efficiency. When both methods are used alone, each has a significant energy-saving effect, but when used together, their combined effect is multiplicative rather than additive. This and the interaction are the reasons why the accuracy-per-watt improvements are so great in the joint optimization environment.

The strength of the accuracy of 4-bit quantization is also notable due to the fears about loss of precision in such large language models. Although previous research has shown that a low-bit quantization in vision and networks in signal processing is feasible, language models present further difficulties since they are sensitive to numerical perturbation and long-range interactions [3], [7], [22]. The findings of this experiment indicate that Llama-based architectures are more resilient to post-training quantization (aggressive) than thought before, particularly when the input of inference has a well-structured and compact format. The implications of the finding on the more general application of ultra-low-precision inference to natural language processing systems are also important.

In this case, a critical stabilizing force is optimization on a timely basis. Optimized queries can suppress the spread of noise created by quantization through the internal representations of a model by minimizing the unwarranted verbosity and ambiguity in input queries. Earlier prompt optimization studies have mostly focused on bettering semantic alignment, relevance, and performance of the task [5], [9], [15]. The provided study builds on those results by showing that prompt design is also a lever of computational efficiency, which has a direct effect on both energy consumption and latency. This dual role makes prompt optimization a low-cost but high-impact intervention to be used in the deployment of AI sustainably.

The deployment side is of particular interest in the deployment of the results to edge and resource-constrained settings. The edge platform frequently has little headroom to support full-precision inference, and it is not always capable of executing large language models that do not necessarily need to be deployed in centralized data centers [19], [25]. The fact that it is now possible to compute quantized LLMs at scaled prompts with a tight energy budget increases the scope of real-world applications, such as on-device assistants, privacy-respecting inference, and real-time analytics. In addition, the post-training character of the offered approach reduces friction during deployment, as it will not need retraining, special equipment, or architecture redesigning.

Scalability analysis also shows that the efficacy advantages are maintained as the model size is increased, up to the capability of models with over a billion parameters. This observation is important against the trends in contemporary trends of language models that have become ever-larger. With increase in model size, absolute energy consumption is increasing, however, the relative savings (quantization and prompt optimization) are steady. It means that the given

framework can grow with the next generations of LLMs, and will remain relevant to the future as the complexity of the models will only increase.

Although these results are encouraging, there are some limitations that should be considered. The first one is that, although the loss of accuracy is not much when assessing different tasks, some sensitive or numerically intensive applications might need further validation. Second, the research is aimed at inference efficiency and does not comment on training-time energy consumption as a major portion of the carbon footprint. Lastly, although prompt optimization is deliberately minimally complex and fast to adopt, more complex algorithmic or adaptive prompt optimization algorithms may provide additional improvements at the price of a heavier workload.

On the whole, this discussion highlights the fact that significant steps towards the implementation of sustainable large-scale language models do not always have to be radical architectural redesign or costly retraining pipelines. Rather, strategic combinations of the current methods, which are implemented considerately and assessed as a whole, can yield instant and effective energy savings. This work adds a practical and scalable view on the sustainability of AI to the growing literature on green and energy-efficient AI by redefining the meaningfulness of efficiency as a communal asset of both the mode of computation and the design of inputs.

6. CONCLUSION

This paper explored a practical and deployable method to enhance the energy efficiency of the inference of Large Language Models by using a combination of low bit-width quantization and prompt optimization. Due to the increasing computational and environmental overheads of large-scale language models, the proposed framework aimed to minimize inference energy usage with minimal accuracy and scaling reduction of the model.

This paper demonstrated that 4-bit post-training quantization of Llama-based models can save memory footprint and computational overhead by a significant margin, without causing any significant change in accuracy. Simultaneously, query and prompt optimization was demonstrated to lower the inference costs at the token-level by removing the unnecessary verbosity and redundant context data. These techniques performed as well as 90 percent inference energy reduction when used together, which is a great improvement over full-precision baseline configuration.

One of the contributions made by this work is that it utilizes the accuracy-per-watt as a single evaluation metric. Instead of considering performance and efficiency as mutually exclusive goals, the outcome demonstrates that they can be enhanced together when optimized care is taken. The combined optimization strategy remained largely better than the stand-alone strategies, which proved the complementary relationship between model-level and input-level efficiency mechanisms.

The strength of the suggested method with model sizes of more than one billion parameters only goes to show the extent of its scalability and applicability to current applications of LLM. Also, verification in edge and resource-constrained environments shows that the framework is well-adapted to the case of decentralized inference, where the power supply, memory sizes, and latency rates are the essential constraining variables. Notably, the post-training and prompt-level quality of the optimizations is such that there are low adoption costs, so there is no need to train and architecturally modify existing systems.

In summary, this work provides empirical evidence that meaningful energy savings in large language model inference are achievable today using readily available techniques. By integrating low bit-width quantization with efficient prompt design, the proposed framework offers a quick-win solution for sustainable, scalable, and performance-preserving deployment of large language models. These findings contribute to the broader goal of green AI and support

the responsible expansion of language model applications across both cloud and edge environments.

REFERENCES

1. Arasi, B. I., & Hemamalini, U. (2025). Energy-Efficient AI for Medical Imaging: A Green Computing Approach To Diagnosis. *International Journal of Creative Research Thoughts*, *13*(4). https://Doi.Org/10.56975/Ijcrt.V13i4.282315
2. Blüthgen, C. (2025). Technical foundations of large language models. *Radiologie*, *65*(4), 227–234. https://doi.org/10.1007/s00117-025-01427-z
3. Chen, X., Wang, Y., Li, Y., Ling, X., Li, M., Liu, R., … He, Y. (2025). Low-Bit-Width Zero-Shot Quantization with Soft Feature-Infused Hints for IoT Systems. *IEEE Internet of Things Journal*, *12*(7), 8484–8496. https://doi.org/10.1109/JIOT.2024.3507114
4. Cao, L., Xiao, W., Mo, Y., Zeng, S., Chen, H., Wu, Z., & Li, X. (2025). Improved YOLO11 for the Asian Citrus Psyllid on Yellow Sticky Traps: A Lightweight Design for Edge Deployment. *Mathematics*, *13*(23). https://doi.org/10.3390/math13233836
5. Choi, J. (2025). Efficient Prompt Optimization for Relevance Evaluation via LLM-Based Confusion Matrix Feedback. *Applied Sciences (Switzerland)*, *15*(9). https://doi.org/10.3390/app15095198
6. Deng, X., Huang, T., Wang, W., & Feng, W. (2025). SE-YOLO: A sobel-enhanced framework for high-accuracy, lightweight real-time tomato detection with edge deployment capability. *Computers and Electronics in Agriculture*, *239*. https://doi.org/10.1016/j.compag.2025.110973
7. Gernigon, C., Filip, S. I., Sentieys, O., Coggiola, C., & Bruno, M. (2024). AdaQAT: Adaptive Bit-Width Quantization-Aware Training. In *2024 IEEE 6th International Conference on AI Circuits and Systems, AICAS 2024 - Proceedings* (pp. 442–446). Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/AICAS59952.2024.10595895
8. Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., … Liu, T. (2025). A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Transactions on Information Systems*, *43*(2). https://doi.org/10.1145/3703155
9. Lieander, A. J., Wang, H., & Rafferty, K. (2025). Prompt Optimization with Two Gradients for Classification in Large Language Models. *AI (Switzerland)*, *6*(8). https://doi.org/10.3390/ai6080182
10. Liu, Y., Du, H., Wu, Y., & Mo, T. (2025). FPGA Accelerated Deep Learning for Industrial and Engineering Applications: Optimal Design Under Resource Constraints. *Electronics (Switzerland)*, *14*(4). https://doi.org/10.3390/electronics14040703
11. Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., … Mian, A. (2025). A Comprehensive Overview of Large Language Models. *ACM Transactions on Intelligent Systems and Technology*, *16*(5). https://doi.org/10.1145/3744746
12. Niu, S., Zhang, X., Wang, S., Liao, K., Zhang, B., & Zou, G. (2025). A–ESD: Auxiliary Edge-Server Deployment for Load Balancing in Mobile Edge Computing. *Mathematics*, *13*(19). https://doi.org/10.3390/math13193087
13. Paula, E., Soni, J., Upadhyay, H., & Lagos, L. (2025). Comparative analysis of model compression techniques for achieving carbon efficient AI. *Scientific Reports*, *15*(1). https://doi.org/10.1038/s41598-025-07821-w
14. Ranpara, R. (2025). Energy-efficient green AI architectures for circular economies through multi-layered sustainable resource optimization framework. *Discover Sustainability*, *6*(1). https://doi.org/10.1007/s43621-025-01846-x

15. Sabbatella, A., Ponti, A., Giordani, I., Candelieri, A., & Archetti, F. (2024). Prompt Optimization in Large Language Models. *Mathematics*, *12*(6). https://doi.org/10.3390/math12060929
16. Sécheresse, X., Guilbert–Ly, J. Y., & Villedieu de Torcy, A. (2025). GAAPO: genetic algorithmic applied to prompt optimization. *Frontiers in Artificial Intelligence*, *8*. https://doi.org/10.3389/frai.2025.1613007
17. Tony, C., Pintor, M., Kretschmann, M., & Scandariato, R. (2026). Discrete prompt optimization using genetic algorithm for secure Python code generation. *Journal of Systems and Software*, *232*. https://doi.org/10.1016/j.jss.2025.112682
18. Ur Rehman, Z., Hassan, U., Ul Islam, S., Gallos, P., & Boudjadar, J. (2025). Energy-Efficient AI for Medical Diagnostics: Performance and Sustainability Analysis of ResNet and MobileNet. In *Studies in Health Technology and Informatics* (Vol. 327, pp. 1225–1229). IOS Press BV. https://doi.org/10.3233/SHTI250585
19. Witt, N., Deutel, M., Schubert, J., Sobel, C., & Woller, P. (2024). Energy-Efficient AI on the Edge. In *Unlocking Artificial Intelligence: From Theory to Applications* (pp. 359–380). Springer Nature. https://doi.org/10.1007/978-3-031-64832-8_19
20. Xiao, H., Zhou, F., Liu, X., Liu, T., Li, Z., Liu, X., & Huang, X. (2025). A comprehensive survey of large language models and multimodal large language models in medicine. *Information Fusion*, *117*. https://doi.org/10.1016/j.inffus.2024.102888
21. Wei, L., Wang, S., Liang, X., Du, D., Huang, X., Li, M., … Zheng, Z. (2025). Slim-sugarcane: a lightweight and high-precision method for sugarcane node detection and edge deployment in natural environments. *Frontiers in Plant Science*, *16*. https://doi.org/10.3389/fpls.2025.1643967
22. Xia, Z., Zhu, H., Ying, H., Li, J., Huan, R., & Pan, Y. (2025). A novel intra-layer mixed bit-width quantization method for the classification of 1D periodic time-series signals. *Computers and Electrical Engineering*, *127*. https://doi.org/10.1016/j.compeleceng.2025.110633
23. Yin, S., Fu, C., Zhao, S., Li, K., Sun, X., Xu, T., & Chen, E. (2024, December 1). A survey on multimodal large language models. *National Science Review*. Oxford University Press. https://doi.org/10.1093/nsr/nwae403
24. Zhong, Y., Zhou, Y., Chao, F., & Ji, R. (2025). MBQuant: A novel multi-branch topology method for arbitrary bit-width network quantization. *Pattern Recognition*, *158*. https://doi.org/10.1016/j.patcog.2024.111061
25. Zhai, H., Du, J., Ai, Y., & Hu, T. (2025). Edge Deployment of Deep Networks for Visual Detection: A Review. *IEEE Sensors Journal*, *25*(11), 18662–18683. https://doi.org/10.1109/JSEN.2024.3502539