Energy Consumption of Large Language Models and the impact of maximizing prompt and output length

Sanne Eeckhout
University of Twente
s.eeckhout@student.utwente.nl

Aakash Huilgol
University of Twente
a.huilgol@student.utwente.nl

ABSTRACT

This study investigates the energy consumption patterns of three large language models (LLMs)—OPT, BLOOM, and FLAN-T5—during inference, focusing on how prompt and output lengths affect energy use and performance. Using the Alpaca dataset, prompt lengths (0-60 and 60-110 tokens) and output lengths (50, 100, 200 tokens) are evaluated to understand their impact on energy efficiency and model accuracy, reported by the MMLU Benchmark. Energy consumption was measured using the built-in `powermetrics` tool, with multiple execution runs to ensure robustness. Main findings reveal that FLAN-T5 consumed the least energy, with the highest reported, while OPT had the highest energy demands across all configurations. BLOOM performed moderately in terms of energy efficiency compared to the other two. Notably, longer output lengths significantly increased energy usage for all models, while prompt length showed a relatively minor impact. These insights underscore trade-offs between energy consumption, model architecture, and input/output configurations, offering guidance for environmentally sustainable LLM deployment.
This paper is accompanied by a replication package that can be accessed through: https://github.com/aakashhuilgol/UT-2024-llm-energy-consumption-rep-pkg

KEYWORDS

LLMs, Energy Consumption, Inference, OPT, BLOOM, FLAN-T5

1 INTRODUCTION

Natural Language Processing has emerged as one of the most innovative fields which has seen a lot of technological advancements in recent years. The ability to make machines understand and generate human-like text has been a main focus in this domain [15, 23, 26]. As a result, large language models (LLMs) have been improving rapidly, and there has been a significant rise in the number of models, with different characteristics, architectures and accuracy.

Recent studies [8, 13, 18] have started addressing the environmental cost of LLMs, but there is still a lack of comprehensive research on the energy consumption of the inference phase regarding several model architectures, which is crucial as it often is run in a continuous matter.

Our study aims to bridge this gap by analyzing the energy consumption of three LLMs—OPT (1.3B parameters), BLOOM (1.1B), and FLAN-T5 (783M)—during inference. The novelty of this research lies in examining how maximizing both input prompt and output lengths impacts energy usage across these models, providing insights into how these factors affect energy consumption. This research therefore aims to answer the following research question. ***How do variations in input prompt length and output length impact the energy consumption of OPT, BLOOM, and FLAN-T5 during inference, and how does this compare with each model's accuracy?***

We conducted controlled experiments using the built-in Powermetrics tool on MacOS [1], a tool to monitor the energy consumption of these LLMs during inference. We were able to systematically measure the power requirements of these models under different input and output configurations. The results will help provide insights into the trade-offs between model performance and energy efficiency, with a focus on potential improvements and more sustainable NLP.

The main findings indicate distinct energy consumption patterns across models, with longer output lengths contributing significantly to energy usage. FLAN-T5 consistently exhibited the lowest energy consumption along with the best accuracy, followed by BLOOM which was marginally more

October, 2024, University of Twente

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

efficient in terms of energy when compared to OPT. Additionally, shorter input prompts generally correlate with lower energy usage, though the differences are minimal. Overall, this study provides essential insights into the trade-offs between model performance and energy efficiency, highlighting the potential of FLAN-T5 to achieve high accuracy with the least energy consumption.

The document is structured as follows: we will first discuss the context in which this problem persists and provide a comprehensive review of previous work done in this field, discussing existing research on energy consumption in LLMs. Next, we provide a detailed overview of our methodology and the framework in which we conducted our experiments. We then present our results and identify any threats to validity. We conclude our research by discussing the next steps for this research.

2 CONTEXT

The advancements in LLMs come with significant energy consumption, both during training and during inference — the process by which trained models generate outputs based on inputs. This energy use has direct implications for carbon emissions, making the investigation of inference energy consumption a critical area of research [10, 20].

From a technical context, this investigation focuses on measuring and improving the energy efficiency of LLM inference. By using the selected models OPT, BLOOM and FLAN-T5, the project targets platforms that are both high in performance and accessible for experimentation. Each model presents unique architectural features, which will allow for diverse energy efficiency profiling. BLOOM and OPT use a decoder only architecture. in which tokens are generated one at a time based on previously generated tokens. BLOOM specifically focuses on being a multilingual model, and being able to run on distributed systems. On the other hand, OPT was modeled after GPT3, but integrates much better memory and energy optimization for better running. FLAN-T5 uses an encoder-decoder model, which helps in understanding the context by interpreting the entire input before the output is generated. These architectural differences can help us to determine which architecture provides the best energy efficiency.

The availability of these models on widely-used platforms like cloud infrastructure further strengthens the relevance of this study, making it feasible to conduct repeatable and controlled experiments on real-world hardware configurations. Beyond the technical aspects, the societal context of this project is closely tied to sustainability concerns, particularly within the framework of the five-dimensional model of sustainability: environmental, economic, social, technical, and individual.

Environmental Sustainability. The project addresses the environmental impact of LLM inference by investigating energy consumption, aiming to evaluate the energy efficiency of AI deployments. This aligns with global efforts toward sustainable technology use.

Economic Sustainability. By identifying more energy-efficient LLMs, the project can help businesses reduce operational costs related to energy consumption, making AI technology more economically viable and accessible.

Social Sustainability. Energy-efficient AI models can democratize access to advanced technologies, promoting social equity by reducing the energy burden on underserved communities and regions with limited resources.

Technical Sustainability. The project contributes to Green Software Development by exploring how different LLMs perform in terms of energy efficiency, offering insights to optimize AI models for scalable, sustainable software engineering practices. This project will enable engineers to make better-informed decisions when designing new models on how to meet efficiency requirements, while at the same time considering accuracy.

The target audience of this study includes software developers, AI researchers, cloud platform providers, and policy makers. Developers and engineers will benefit from practical insights into how different models perform from an energy efficiency perspective, enabling them to make better choices when integrating LLMs into their applications. For AI researchers, this investigation could lead to further inquiries into optimizing the architectures of future models to minimize energy consumption.

At a broader societal level, the findings of this project will provide a pathway towards reducing the environmental and economic costs of AI, benefiting organizations and stakeholders that seek to balance innovation with sustainability. The scientific community will also benefit from a better understanding of the energy trade-offs involved in deploying these models, which will contribute to more informed debates on the role of AI in addressing global challenges like climate change.

3 RELATED WORK

Our research builds on the work of [3], which compares the energy efficiency of a few large language models, (Pythia, BLOOM, Dolly, RedPajama and LLaMa-2). While their study effectively examines energy consumption during the inference phase, we investigated additional models using a similar experimental framework. This allows us to replicate and extend the findings of the original work.

Recent studies [4, 5, 9, 21] have explored the energy consumption of large language models (LLMs) during both training and inference. Notably, Patil et al. [16] investigated energy metrics for several models, including CodeLlama-7B, Gemma-2B, and Llama3-8B, emphasizing the energy requirements for inference across these architectures. Similarly, Zhang et al. [27] examined Falcon-7B, Llama-2-7B, and Mistral-7B, contributing to a growing body of work that seeks to quantify the efficiency of these models.

Byun et al. [2] expanded on this theme by analyzing Falcon and LlaMA-2 models across a range of parameter sizes, specifically focusing on energy metrics for Falcon, Llama-2, and Mistral in the 7B parameter range. Argerich [3] introduced a framework for measuring energy consumption in models like Pythia, Bloom (3B), and DollyV2, further highlighting the importance of energy metrics in the development of LLMs.

In terms of larger models, Luccioni et al. [11] conducted a detailed analysis of energy consumption for Bloom (176B), GPT-3 (175B), OPT (175B), and Gopher (280B), demonstrating the correlation between their larger model size and energy usage. Wang et al. [24] specifically investigated the energy efficiency of BERT models, adding to the understanding of how different architectures impact energy consumption. Additionally, Patterson et al. [17] explored the carbon footprint associated with the training of the T5 model, shedding light on the broader implications of energy consumption in LLM development.

Other research has shown the effect of various factors affecting LLMs. Wilkins et al. [25] investigate the runtime implications of different input and output sizes, highlighting the relationship between token lengths and energy use during inference.The impact of prompt and output length on model reasoning has also been analyzed by Nayab et al. [14], with findings indicating performance trade-offs but without a focus on energy consumption. Meanwhile, Stojkovic et al. [19] explore multiple strategies for enhancing energy efficiency in LLM inference but do not specifically address the influence of varying prompt sizes.

Despite extensive research on large language models (LLMs), a notable gap exists regarding energy consumption during inference—particularly for models with around 1 billion parameters. This project addresses this gap by examining OPT, BLOOM, and FLAN-T5, making OPT a prime candidate as no prior work has evaluated its energy use in this context. A unique focus of this study is on how both input prompt length and output length affect energy consumption, an aspect largely unexplored in existing literature. Our findings, detailed with MMLU [7] benchmark scores in Table 1, offer comparative insights into each model's accuracy and energy efficiency. This work will contribute to the ongoing conversation about sustainable AI practices.

Model	Parameters	Accuracy
BLOOM	1.1B	26.7%
OPT	1.3B	24.9%
FLAN-T5	783M	45.1%

Table 1: Reported accuracy on the Massive Multitask Language Understanding (MMLU) benchmark [7]

4 EXPERIMENT PLANNING

The general goal of the study is defined as follows.

- *Analyze* the energy consumption of BLOOM, FLAN-T5 and OPT LLMs during inference.
- *For the purpose of* evaluations and comparison of energy efficiency and understanding environmental impacts.
- *With respect to* their performance in generating outputs (up to 100 tokens) in response to prompts of different sizes (0-60 and 60-110 tokens).
- *From the point of view of* AI practitioners, developers, eco-friendly engineers.
- *In the context of* growing concerns over the energy usage of AI technologies and the push for more sustainable AI development practices.

To achieve the above goals, we propose the following research questions that answer our main research question ***How do variations in input prompt length and output length impact the energy consumption of OPT, BLOOM, and FLAN-T5 during inference, and how does this compare with each model's accuracy?***.

For each research question, we formulate corresponding hypotheses. The null hypotheses (H_0) posit that there are no significant differences in energy consumption across the models or output lengths, while the alternative hypotheses (H_a) assert that significant differences do exist.

RQ1 ***How does the energy consumption of OPT, BLOOM and FLAN-T5 compare across multiple execution runs?***
H_0 There is no significant difference in energy consumption variability between OPT, BLOOM, and FLAN-T5 across multiple execution runs.
H_1 There is a significant difference in energy consumption variability between OPT, BLOOM, and FLAN-T5 across multiple execution runs.

RQ2 ***How does the energy consumption of OPT, BLOOM, and FLAN-T5 vary based on different lengths of prompts?***
H_0 There is no significant difference in the energy consumption of OPT, BLOOM, and FLAN-T5 when varying the lengths of prompts.
H_1 There is a significant difference in the energy consumption of OPT, BLOOM, and FLAN-T5 when varying the lengths of prompts.

RQ3 ***How does output length (50, 100, 200 tokens) affect the energy consumption and response quality of OPT, BLOOM, and FLAN-T5?***
H_0 There is no significant difference in energy consumption and response quality between the output lengths of 50 and 100 tokens for OPT, BLOOM, and FLAN-T5.
H_1 There is a significant difference in energy consumption and response quality between the output lengths of 50 and 100 tokens for OPT, BLOOM, and FLAN-T5.

RQ4 ***What is the overall accuracy of OPT, BLOOM, and FLAN-T5 as reported by benchmark evaluations, and how does it compare against the energy consumption of the models?***
H_0 There is no significant correlation between reported accuracy and energy consumption of the three models.
H_1 There is a significant correlation between reported accuracy and energy consumption of the three models.

4.1 Experimental Units and Variables

The experimental units for this study are the selected LLMs: OPT, BLOOM, and FLAN-T5. Each model undergoes testing under specific conditions to gather data on energy consumption during inference.

This study analyzes the following variables:

- *Independent Variables* The LLMs (OPT, BLOOM, and FLAN-T5), the input prompt token lengths (0-60 and 60-110) and the output token lengths (50, 100 and 200 tokens).
- *Dependent Variables* The energy consumption (measured in joules) of GPU and CPU of each model during inference tasks.

We employ the Alpaca dataset [22] from Hugging Face for generating prompts with sample size set for each combination to 100, utilizing three maximum output lengths: 50, 100 and 200 tokens. The choice of 100 tokens aligns with prior research in [3] while the inclusion of 50 and 200 tokens allows us to investigate its effect on energy consumption. We categorized our data in two different prompt lengths, 0-60 tokens and 60-110 in order to analyse the effect of prompt length on the consumption.

4.2 Experiment Design and Analysis Procedure

This study utilizes a within-subjects design, where each model is subjected to all prompt and output lengths. Thus, it has a 3x3x2 design, with each model being tested at both token limits.

The collected data will be analyzed using several statistical methods. These statistical tests are described below.

- A one-way ANOVA will be employed for RQ1 to assess average energy consumption differences among models [6].
- A repeated measures ANOVA will be used for RQ2 to determine if there are significant interactions between input length and energy consumption across different models [12].
- Linear regression analysis will be performed for RQ3 to evaluate the effect of output length on energy consumption, analyzing if the increase in tokens significantly impacts energy usage for each model.
- For RQ4, a comparative analysis will be conducted by referencing benchmark-reported accuracy scores of OPT, BLOOM, and FLAN-T5. These accuracy benchmarks will be qualitatively compared to the energy consumption results derived from the statistical tests of RQ1, offering insights into the trade-offs between model accuracy and energy efficiency.

Statistical analyses will be conducted using Python, ensuring rigorous computation and data handling. Effect sizes will be calculated to quantify the strength of observed differences, and visualizations such as box plots and line graphs will be used to illustrate trends and distributions.

5 MEASUREMENT TOOLS, ENVIRONMENT AND PROCEDURE

The experiments were conducted on a MacBook M1, equipped with 16GB of RAM. To measure energy consumption during the inference tasks, we utilized the built-in tool `powermetrics` available on macOS. This tool provides insights into the energy usage of various processes, allowing us to capture accurate energy consumption data during model inference. The rationale for selecting `powermetrics` is its integration with the macOS environment and its capability to provide detailed energy metrics without requiring external hardware.

The experimental procedure involves executing each LLM (OPT, BLOOM, and FLAN-T5) on prompts from the Alpaca dataset, alternating two input prompt lengths of 0-60 and 60-110 tokens. Each input prompt length will be run against three maximum output lengths of 50, 100 and 200 tokens. Each execution run is preceded and succeeded by a 60-second baseline measurement to establish the energy consumption profile of the system in idle state. To measure the power consumption of the powermetrics tool itself, a 20-minute baseline was analyzed when no processes were run. This baseline helps us to mitigate the effects of any background processes on the energy measurements during the model inference.

To ensure the validity and reliability of our results, we conducted 3 executions per trial for each model, repeating the inference tasks to account for variability in energy consumption. The workloads are representative of real-world usage scenarios, as the Alpaca dataset comprises diverse prompts suitable for evaluating the performance of LLMs.

6 THREATS TO VALIDITY

Background processes. The experiment sets a baseline by capturing 60 seconds prior to and succeeding the process. However, background systems can fluctuate, affecting energy measurements. MacOS might launch some background tasks during the run, creating inconsistent data. We monitored the background tasks during the runs to ensure that the results are not impacted, as well as analysing three different runs per model to minimize the effect of background processes. To ensure that the energy measurement tool does not provide any additional overhead, we also monitored the energy usage of the tool itself for 20 min as a baseline, thus mitigating its effect on our results.

Hardware generalization. Since the experiment is run on a local machine, the results of the experiment cannot be applied to environments where LLMs are generally deployed (like an array of GPUs). There is a difference in the processor, and normal LLM deployment environments generally utilize power efficiency algorithms. However, our research focuses on analysing which model architectures are more energy efficient to encourage developers to make informed choices.

Real World Application. The experiment uses two input lengths (0-60 and 60-110 tokens) and three output lengths (50, 100 and 200 tokens). However real world cases of widely used LLMs are capable of processing and generating much longer texts. The results might not apply to LLMs like ChatGPT which have a lot more parameters and are capable of generating more than 100 tokens. These results are more applicable to short-text generation tasks, and our focus is on more accurate responses.

Execution Runs. The experiment analyses three execution runs per model per output length. This is a relatively small number of repetitions, that might not be enough to detect meaningful differences in energy consumption across models. However, since the models are tested under the same conditions, the comparison is reliable within the scope of this experiment.

7 RESULTS

The energy consumption patterns for the three large language models (LLMs) studied— OPT, FLAN-T5, and BLOOM— are illustrated across various conditions in this section. Figure 1 shows the overall average power consumption for each model across all runs, providing an overview of general energy demands among the models, whereas Table 3 details each model's power consumption across specific input-output configurations, facilitating a direct comparison of energy usage across varied prompt and output lengths for OPT, FLAN-T5, and BLOOM. Further, Figure 4 captures each model's power consumption over the course of three execution runs shown over time, averaging power demand across multiple input-output length combinations and revealing temporal patterns in power consumption.

A closer look at the impact of input prompt length on energy usage is shown in Figure 2, where models are grouped by prompt ranges of 0–60 and 60–110 tokens. Here, the effect of prompt length is observed with each model's power consumption averaged within these two ranges. Meanwhile, Figure 3 examines the influence of output length by displaying energy consumption across output lengths of 50, 100, and 200 tokens for each model, revealing how extended outputs correlate with increased power demands.

Table 2: ANOVA Results for Energy Consumption Analysis of Models by Prompt and Output Lengths

Source	*df*	*F*	*p-value*	*Effect Size*
Two-Way ANOVA on the models and Prompt Lengths				
Model	2, 12	18.65	0.0002	
Prompt Length	1, 12	0.018	0.894	
Model × Prompt Length	2, 12	0.0049	0.995	
Repeated Measures ANOVA on Output Lengths				
Output Length	2, 4	18.19	0.0098	0.198

Lastly, Table 2 summarizes the ANOVA results, analyzing variations in energy consumption based on model type, prompt length, and output length. A two-way ANOVA explores the interactions between model type and prompt length, while a repeated measures ANOVA evaluates the effect of increasing output length.

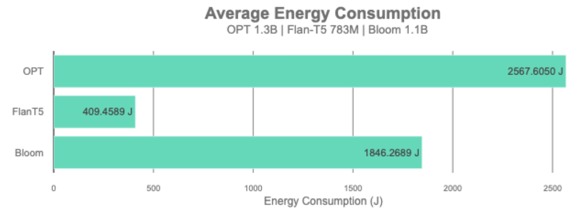

Figure 1: Combined Energy Consumption of the three LLMs OPT, FLAN-T5 and BLOOM averaged over all execution runs

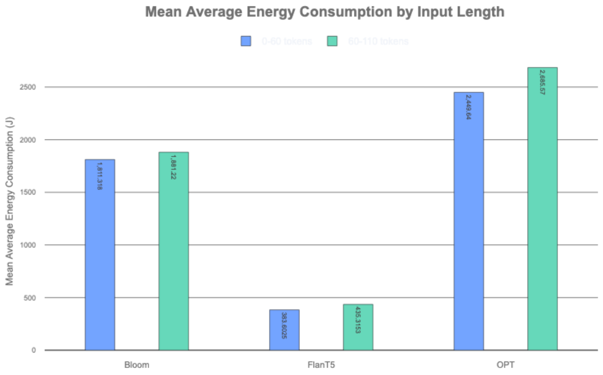

Figure 2: Combined Energy Consumption of OPT, FLAN-T5 and BLOOM averaged over all execution runs for two different prompt lengths, 0-60 tokens and 60-110 tokens

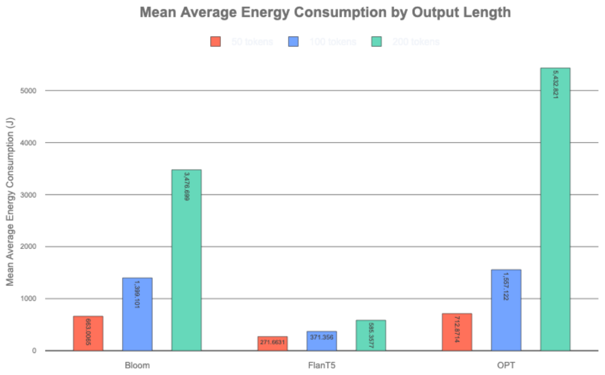

Figure 3: Combined Energy Consumption of OPT, FLAN-T5 and BLOOM averaged over all execution runs when maximizing the output length by 50, 100 and 200 tokens

8 DISCUSSION

This section analyzes the energy consumption patterns of three large language models (LLMs)—BLOOM, FLAN-T5, and OPT—during inference, focusing on the effects of varying input prompt lengths, output token lengths, and energy consumption over time.

Effect of Output Token Length on Energy Consumption. Increasing the maximum output token length significantly impacts energy consumption for each model, as shown in Table 2 All three models show a clear increase in energy consumption as output length increases from 50 to 200 tokens, depicted in Figure 3. This finding suggests that concise outputs are more energy-efficient, with models expending increasingly more energy to generate longer outputs. This trend is particularly relevant for applications where energy efficiency is prioritized, as selecting shorter output lengths can yield substantial energy savings.

Impact of Input Prompt Length. Table 3 and Figure 2 reveal that input prompt length has a noticeable but comparatively smaller effect on energy consumption than output length.

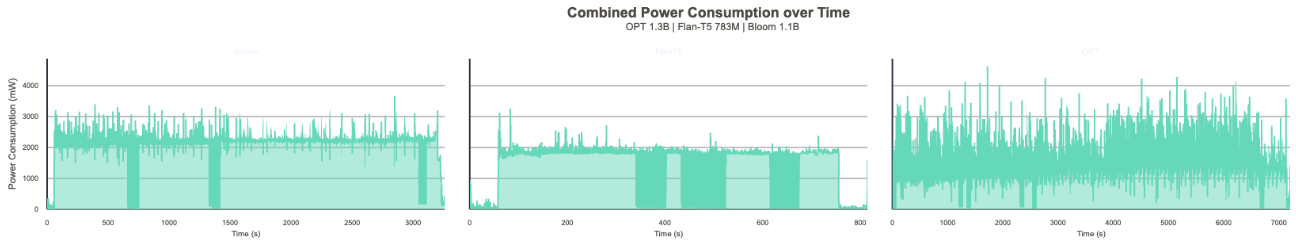

Figure 4: Combined Power Consumption of OPT, FLAN-T5 and BLOOM over the course of 3 execution runs, averaged over all combinations of input-output length

Table 3: Combined Energy Consumption for OPT, Bloom and FLAN-T5 by Input Length and Output Length (Energy in J, Time in s)

Model	Input Length (tokens)	Output Length (tokens)			Time		
		50	100	200	50	100	200
Bloom	0-60	645.88	1393.73	3394.33	361.5	695.5	1554.5
	60-110	680.13	1404.46	3559.06	375.5	708.5	1631
FLAN-T5	0-60	274.50	357.25	519.04	200.5	247	338
	60-110	268.82	385.45	651.66	201.5	263	408
OPT	0-60	722.27	1545.77	5080.86	635	1188	3602.5
	60-110	631.82	1568.46	5784.77	635	1301.5	3362.5

Shorter prompts (0-60 tokens) consume slightly less energy across all models compared to longer prompts (60-110 tokens), though this difference is minimal and not statistically significant in some cases, shown by Table 2. This minor increase could be attributed to the model's requirement to process additional contextual information in longer prompts, which might slightly increase the computational load and thus the energy consumed.

Energy Consumption over Time. The energy consumption patterns over time for the three models, displayed in **??**, reveal distinct behaviors. Notably, OPT's energy consumption fluctuates more significantly than BLOOM and FLAN-T5, which display relatively steady levels. OPT's high variability might stem from its adaptive mechanisms or optimization strategies that adjust processing power based on the computational requirements at different inference stages, leading to frequent fluctuations. This could indicate that OPT leverages dynamic load balancing to conserve energy during less complex processing stages.

Another interesting observation in the time-based figure is the periodic dip in energy consumption across all models, where consumption briefly drops close to zero. This pattern might reflect moments of lower activity or processing delays within the model, possibly due to internal buffering, batch processing, or other factors that temporarily reduce computational demands. Such dips could represent idle moments or efficiency strategies that briefly halt processing between inference tasks, helping to reduce the average energy footprint.

Overall Energy Consumption. The data presented in Table 3 and Figure 1 shows that among the three models, FLAN-T5 consistently consumes the least energy, followed by BLOOM, with OPT consuming the most energy across different input and output lengths. This ranking suggests that FLAN-T5's encoder-decoder architecture is optimized for efficiency, allowing it to handle tasks with lower energy expenditure compared to BLOOM, which has a larger model size and less efficient design features. The architecture of OPT, while more efficient in some respects, leads to higher overall energy consumption likely due to its complexity and the specific computational strategies it employs.

The higher reported accuracy on the MMLU Benchmark in Table 1 indicates that FLAN-T5 may leverage its architectural advantages to deliver superior performance on benchmark tasks despite its lower energy consumption. This finding is particularly relevant to the research question (RQ) regarding how variations in input prompt length and output length impact energy consumption and accuracy across different models. It highlights a critical trade-off between energy efficiency and performance, suggesting that more efficient models may not always achieve the highest accuracy, as seen with OPT.

Overall, these findings indicate that while increasing output length impacts energy usage consistently across models, prompt length effects are minor. Additionally, each model exhibits unique patterns in energy consumption over time, as well as averaged energy consumption, which can be attributed to their architectures

9 CONCLUDING REMARKS

This study provided a comprehensive analysis of the energy consumption associated with large language models (LLMs) during inference, specifically focusing on OPT, BLOOM, and FLAN-T5, all within the 1 billion parameter range. By utilizing the Alpaca dataset and measuring energy consumption on a MacBook M1, we examined how variations in prompt length (0–60 and 60–110 tokens) and output length (50, 100, and 200 tokens) impact energy efficiency and performance.

Our results demonstrated distinct energy consumption patterns among the models. FLAN-T5 was the most energy-efficient across most configurations, while OPT exhibited significantly higher energy demands. In comparison, BLOOM exhibited significantly higher energy demands, placing it between the other two models in terms of energy usage. Despite consuming more energy, OPT showed lower accuracy compared to FLAN-T5, which also seemed to be the best performing model of all three. Notably, output length played a substantial role in determining energy usage, with clear increases in consumption as output length extended. In contrast, prompt length had a comparatively minor effect on energy consumption, though it still influenced the models' energy profiles. These findings underscore the complexity of optimizing LLM inference, as both input and output configurations must be carefully managed to achieve desired performance and energy efficiency.

The study's contributions are threefold. First, it adds new insights into the energy-performance trade-offs of commonly used LLMs, especially those that are smaller in size, allowing for insights into which architectures are more energy-efficient. Second, it highlights the impact of input length on energy usage and lastly, the study explores the relationship between output length and energy consumption, highlighting how generating longer responses significantly affects the energy demands of LLMs.

Future research could extend these findings by exploring energy consumption across a more extensive range of models, prompt- and output length. Additionally, further studies could include a larger range of samples and classifying the samples into types of prompt. Another Another promising direction involves focus on the directly scaled versions of the three LLMs, in order to investigate the energy efficiency as model sizes increase. Ultimately, our work contributes to a broader understanding of sustainable AI development and encourages continued exploration into methods that can reduce the carbon footprint of advanced language technologies.

REFERENCES

[1] Apple Inc. 2023. *powermetrics Manual Page*. https://developer.apple.com/documentation/powermetrics.

[2] Juhee Byun, Hyoungseok Minn, Gauri Joshi, Irena Laskowski, and Marta Sarzynska. 2024. Offline Energy-Optimal LLM Serving: Workload-Based Energy Models for LLM Inference on Heterogeneous Systems. *arXiv preprint arXiv:2407.04014* (2024). https://arxiv.org/abs/2407.04014

[3] Mauricio Fadel Argerich and Marta Patiño-Martínez. 2024. Measuring and Improving the Energy Efficiency of Large Language Models Inference. *IEEE Access* (2024).

[4] R. K. Gupta, S. D. Varma, and R. Kumar. 2021. Energy-Efficient Training of Neural Networks: A Survey. *ACM Computing Surveys (CSUR)* 54, 6 (2021), 1–34.

[5] P. G. P. S. Han and et al. 2021. Green AI: Energy-Efficient Training of Deep Learning Models. In *Proceedings of the 38th International Conference on Machine Learning*. 9112–9122.

[6] Richard Harris. 2001. One-way ANOVA. *The Annals of Statistics* 29, 1 (2001), 29–36.

[7] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring Massive Multitask Language Understanding. *CoRR* abs/2009.03300 (2020). arXiv:2009.03300 https://arxiv.org/abs/2009.03300

[8] John Hogan and Nabilah King. 2023. Environmental Impact of Large Language Models. *Cutter Consortium* (2023). https://www.cutter.com/article/environmental-impact-large-language-models

[9] A. Lacoste and et al. 2020. Energy and Climate Implications of the Deep Learning Ecosystem. *arXiv preprint arXiv:2004.03906* (2020).

[10] Q. Liu, W. Wu, and Y. Zhang. 2021. The Carbon Footprint of Machine Learning: A Review. *IEEE Transactions on Neural Networks and Learning Systems* (2021), 1–14. https://doi.org/10.1109/TNNLS.2021.3082921

[11] Alexandra Sasha Luccioni, Sylvain Viguier, and Anne-Laure Ligozat. 2022. Estimating the Carbon Footprint of BLOOM, a 176B Parameter Language Model. arXiv:2211.02001 [cs.LG] https://arxiv.org/abs/2211.02001

[12] N Muhammad Lutfiyya. 2023. Guidelines for repeated measures statistical analysis approaches with basic science research considerations. *The Journal of Clinical Investigation* 133, 11 (2023), e171058. https://doi.org/10.1172/JCI171058

[13] Hyoungseok Minn, Amirmohammad Bahrami, Sharad Purohit, Marta Sarzynska, Dongkeun Kang, Kaifeng Sun, Utsha Majumder, Gauri Joshi, and Irena Laskowski. 2023. From Words to Watts: Benchmarking the Energy Costs of Large Language Model Inference. *arXiv preprint arXiv:2310.03003* (2023). https://arxiv.org/abs/2310.03003

[14] Sania Nayab, Giulio Rossolini, Giorgio Buttazzo, Nicolamaria Manes, and Fabrizio Giacomelli. 2024. Concise thoughts: Impact of output length on llm reasoning and cost. *arXiv preprint arXiv:2407.19825* (2024).

[15] Deborah W Otter, Julian R Medina, and Jugal K Kalita. 2020. A survey of the usages of deep learning for natural language processing. *IEEE Transactions on Neural Networks and Learning Systems* 32, 2 (2020), 604–624.

[16] Ananya Patil, Amirmohammad Bahrami, Hyoungseok Minn, Gauri Joshi, and Irena Laskowski. 2023. The Price of Prompting: Profiling Energy Use in Large Language Model Inference. *arXiv preprint arXiv:2309.01456* (2023). https://arxiv.org/abs/2309.01456

[17] David Patterson, Joseph Gonzalez, Quoc Le, Chen Liang, Lluis-Miquel Munguia, Daniel Rothchild, David So, Maud Texier, and Jeff Dean. 2021. Carbon emissions and large neural network training. *arXiv preprint arXiv:2104.10350* (2021).

[18] Matthias C. Rillig, Marlene Ågerstrand, Mohan Bi, Kenneth A. Gould, and Uli Sauerland. 2023. Risks and Benefits of Large Language Models for the Environment. *Environmental Science & Technology* 57, 9 (2023), 3464–3466. https://doi.org/10.1021/acs.est.3c01106 arXiv:https://doi.org/10.1021/acs.est.3c01106 PMID: 36821477.

[19] Jovan Stojkovic, Esha Choukse, Chaojie Zhang, Inigo Goiri, and Josep Torrellas. 2024. Towards Greener LLMs: Bringing Energy-Efficiency to the Forefront of LLM Inference. *arXiv preprint arXiv:2403.20306* (2024).

[20] E. Strubell, A. Ganesh, and A. McCallum. 2019. Energy and Policy Considerations for Deep Learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 3645–3650. https://doi.org/10.18653/v1/P19-1355

[21] M. M. Strubell, A. Ganesh, and A. McCallum. 2019. The Carbon Footprint of Machine Learning. *arXiv preprint arXiv:1906.02243* (2019).

[22] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Alpaca: A strong, replicable instruction-following model. *Stanford Center for Research on Foundation Models. https://crfm. stanford. edu/2023/03/13/alpaca. html* 3, 6 (2023), 7.

[23] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*. 5998–6008.

[24] Xiaorong Wang, Clara Na, Emma Strubell, Sorelle Friedler, and Sasha Luccioni. 2023. Energy and Carbon Considerations of Fine-Tuning BERT. *arXiv preprint arXiv:2311.10267* (2023).

[25] Grant Wilkins, Srinivasan Keshav, and Richard Mortier. 2024. Offline energy-optimal llm serving: Workload-based energy models for llm inference on heterogeneous systems. *arXiv preprint arXiv:2407.04014* (2024).

[26] Tom Young, Devamanyu Hazarika, Soujanya Poria, and Erik Cambria. 2018. Recent trends in deep learning based natural language processing. *IEEE Computational Intelligence Magazine* 13, 3 (2018), 55–75.

[27] Haoyu Zhang, Dongkeun Kang, Marta Sarzynska, Kaifeng Sun, Gauri Joshi, and Irena Laskowski. 2024. Hybrid Heterogeneous Clusters Can Lower the Energy Consumption of LLM Inference Workloads. *arXiv preprint arXiv:2402.01234* (2024). https://arxiv.org/abs/2402.01234