AI's Environmental Cost: Comparing Resource Consumption Between SLMs and LLMs Across Queries

Aryaanshi Sundaram[1], Sparsh Kamdar[1] and Shreyas Kumar[2]
[1]DiscoverSTEM, Plano, TX, USA
[2]Texas A&M University, USA

aryaanshi.sundaram@gmail.com
sparsh.kamdar@gmail.com
shreyas.kumar@tamu.edu

Abstract: As artificial intelligence becomes increasingly embedded in daily life, the environmental costs of its deployment remain underexplored. This study investigates the environmental footprint of both large language models (LLMs) and small language models (SLMs); specifically, ChatGPT, Gemini, Deepseek, and Claude, by associating their power draw and water use across queries of varying complexity. Building on evidence that AI services demand substantial resources, this paper asks: how do query complexity and type influence the energy and water consumption of SLMs versus LLMs, and at what threshold of complexity do SLMs become incapable of delivering accurate outputs? To address this, the experimental method categorizes queries into three complexity tiers based on logical steps, conceptual depth, and cognitive skills (recall, evaluation, creation), drawing from the College Board's question bank of SAT math, reading, and writing problems. Additionally, classic puzzles such as the Tower of Hanoi were selected. Each query was executed three times on the SLM and LLM versions of each commercial AI entity under identical hardware and software configurations. We recorded execution time, model version, and output accuracy. Using the average response time per query, we computed energy consumption and water usage per query. On average, SLMs consumed 60-70% less energy and water than their LLM counterparts, and in subjects such as Math and Reading, had the same level of accuracy as their respective LLMs. However, model performance declined as question difficulty increased, especially in abstract reasoning tasks such as Puzzles, where SLM accuracy dropped considerably. While LLMs were more resource-intensive, they maintained higher accuracy on these challenging queries. SLMs offer a significantly more environmentally sustainable option for simple tasks, but accuracy decreases as complexity increases. A dynamic approach, starting with SLMs and switching to LLMs only when needed, or vice versa, could reduce the environmental cost of AI while maintaining quality. These findings support the potential for context-aware AI deployment strategies that optimize environmental sustainability and accuracy. Future research should aim to quantify this breakpoint more accurately and look at the implementation of automatic query classification systems capable of efficiently switching between models to create more efficient AI models.

Keywords: AI environmental sustainability, Small language models, Large language models, Query complexity

1. Introduction

Eric Schmidt, former Google CEO, told Congress that "many people project demand for our industry will go from 3 percent to 99 percent of total generation" to power Superintelligent AI (Joe Wilkins, 2025). LLMs are now integral to daily life. "They have enabled digital assistants that can free people from the need to search, extract, and integrate information from multiple sources by offering straightforward answers in a single chat" (Wang et al., 2024). As global reliance on Artificial Intelligence (AI) increases, so does the strain on environmental resources. The rising demand for AI has led to exponential growth in the number and complexity of queries processed daily. Each query requires a computationally intensive process handled by data centers that consume large amounts of energy and water. These centers often use water-intensive cooling systems to prevent server overheating, compounding AI's environmental impact (Schäfer et al., 2024). This toll raises concerns about the efficiency of the models powering these queries.

The backbone of most AI systems is Language Models, typically classified as Large Language Models (LLMs) or Small Language Models (SLMs). LLMs are transformer models trained on vast datasets and defined by billions of parameters, enabling reasoning and abstraction. They perform tasks like masked language modeling and autoregressive prediction to generate human-like responses (Wang et al., 2024). In contrast, SLMs are leaner, trained on optimized parameters for specific use cases, offering better resource efficiency at the potential cost of capability (Magister et al., 2023).

2. Literature Review

2.1 AI Environmental Cost

The environmental implications of AI have drawn increasing attention as large-scale models consume unprecedented levels of computational resources. Strubell et. al. (2019) were among the first to quantify the

carbon footprint of model training, finding that developing a single NLP model could emit as much CO2 as five cars over their lifetimes. More recent analyses by Jegham et al. (2025) and Henderson et al. (2020) have expanded this focus, showing that the *inference* phase of AI, though individually lightweight, accounts for a far larger share of total energy use due to the scale and daily interactions. These findings highlight the growing urgency for sustainable inference models and energy-efficient architectures.

More recently, Vogginger et al. (2024) examined the prospect of neuromorphic hardware as an energy-efficient paradigm for data centers and reviewed its potential to reduce inference costs at scale. They argue that while inference is individually less energy-intensive than training, the cumulative cost (given the massive number of queries) is nontrivial, and that architectural innovations are needed to shift the efficiency frontier. Additionally, Alex de Vries (2023) reports that as an AI entity grows, the amount of data centers required to run the servers increases, therefore leading to an increase in energy consumption. He cites ChatGPT as an example, explaining it required over 3,500 of NVIDIA's HGX A100 servers.

Complimenting this view, Desislavov et al. (2023) address trends in AI energy consumption and highlight that inference scaling, such as deploying smaller or medium models, may surpass training in environmental relevance.

On the water usage side, data centers often employ water-based cooling systems, and their water use efficiency (WUE) becomes a critical metric, especially in arid regions. The Environmental and Energy Study Institute (EESI) has documented how cooling strategies can dramatically increase water consumption for power generation and data centers.

Together, these works underscore that both energy and water demand must be accounted for when assessing AI's environmental cost, and that sustainable architectures are urgently necessary.

2.2 Model Size, Reasoning, and Efficiency

The commonly held belief is that "bigger is better": larger models yield stronger performance. Kaplan et al. (2020) formalized this via scaling laws, showing that increasing parameters, compute, and data tend to improve performance (though with diminishing returns).

However, more recent empirical findings suggest that reasoning and generalization do not scale indefinitely with size. Schaeffer et al. (2024) demonstrate that scaling laws often fail to predict downstream reasoning capacity accurately, implying that blindly increasing model size may be inefficient. Vogginger and team also contend that neuromorphic or brain-inspired architectures may provide better efficiency-per-inference, especially for models or tasks that can exploit sparsity and event-driven dynamics.

Further research explores how smaller or optimized models can achieve near LLM performance with far less computational demand. Li et al. (2020) propose training large models and subsequently compressing them for efficient inference, showing that model distillation and pruning can drastically reduce energy use without major accuracy loss. Their findings reveal that compressed models retain much of the accuracy of their larger counterparts while cutting energy usage by up to an order of magnitude. This introduces an important nuance: efficiency gains may be best realized after large-scale pretraining, rather than avoiding large models altogether.

Tan and Le (2019) approach the efficiency problem from a structural perspective with EfficientNet, a scaling method that optimizes width, depth, and resolution simultaneously rather than arbitrarily enlarging one dimension. While their work focuses on vision models, the implications extend to NLP, showing that balanced scaling delivers better performance-per-compute ratios than sheer parameter expansion. Together Li et al. (2020) and Tan & Le (2019) present complementary strategies: Li et al. emphasize post-training optimization, while Tan & Le optimize architecture scaling at design time.

Collectively, these findings converge on a critical insight: achieving sustainability in AI requires coordinated improvements across architecture and reasoning algorithms, not merely shrinking or scaling models in isolation.

2.3 Identified Research Gap

Although the literature contains substantial work on training-time carbon footprints and scaling behavior, fewer studies focus on per-query inference-level environmental impact. In addition, the literature does not clearly define the breakpoint at which smaller models lose reliability under greater task complexity (or when neuromorphic or hybrid architectures begin to outperform purely digital models).

This paper addresses that gap by evaluating how variations in query type and complexity impact the environmental cost and output accuracy of large and small language models used in commercial AI products such as ChatGPT, Claude, and Deepseek. It aims to identify the optimal intersection of efficiency and capability by quantifying energy and water consumption per query and correlating them with performance outcomes.

Accordingly, this study explores two research questions:

RQ1: How does query complexity affect the power and water consumption of Small and Large Language Models during inference?

RQ2: At what level of complexity do SLMs begin to lose accuracy, making LLMs necessary despite their higher environmental cost?

H1: Query complexity is positively correlated with both energy and water consumption in inference.

H2: Smaller or neuromorphic-inspired models maintain higher energy and water efficiency than large models up to a complexity threshold, beyond which their accuracy degrades.

3. Method

Step	Description	Tool Used	Output
1	Select SAT and Puzzle questions	College Board/ Online	Datasets of 48 queries
2	Execute each on SLM and LLM	ChatGPT, Claude, DeepSeek	Response times and accuracy
3	Compute power and water	Formulas (Eq. 1, Eq. 2)	Pre-query metrics
4	Visualize data	Matplotlib	Graphs
5	Compare results	Accuracy vs. Resource use	Insights

To evaluate AI responses across different query types and reasoning tasks, an experimental method using four question categories was conducted. The four categories were math, reading, writing, and puzzles. The math, reading, and writing questions were sourced from CollegeBoard's SAT question bank using its difficulty filter to select easy, medium, and hard questions. Math included four types: Problem Solving, Algebra, Advanced Math, and Geometry (College Board, n.d.). For each type, three questions (one per difficulty level) were selected, totaling twelve math questions. The questions were selected randomly through a random number generator. The questions were also selected to ensure no visual prompts, such as graphs or charts. This decision was made to avoid the need to upload images, as recognizing them would increase the response time. Reading featured two types: Information and Ideas, and Expression of Ideas. Four questions were chosen for each difficulty level, totaling twelve reading questions. Writing followed the same structure, using Craft and Structure, and Standard English Conventions as its two types. All math, reading, and writing questions were multiple choice.

There were four types of puzzles: Tower of Hanoi, Go, Checkers, and River Crossing.

The Tower of Hanoi is a puzzle involving three pegs and *n* disks of different sizes stacked on the first peg (largest at the bottom). The goal is to move all disks to the third peg, following rules: only one disk may be moved at a time, only the top disk can be taken from any peg, and a larger disk cannot be placed on a smaller one. Difficulty is controlled by the number of disks, with the minimum number of required moves being $2^n - 1$. As question difficulty increases, so does the number of disks and moves required (Shojaee et al., 2025).

River Crossing is a constraint satisfaction puzzle involving *n* people who must cross a river using a boat. The boat holds at most *k* individuals and cannot travel empty. Invalid states occur when certain groups are left alone. The task's complexity is controlled by adjusting *n* and *k* (Shojaee et al., 2025).

Checkers is a one-dimensional puzzle with red and blue checkers and a single space arranged in a line. The objective is to swap red and blue checkers, mirroring the initial setup. Valid moves include sliding into an adjacent space or jumping over one opposing checker into a space. Backward moves are not allowed. Each question was presented as a multiple-choice item asking the AI for the best move and its reasoning (Shojaee et al., 2025).

GO is a traditional Chinese board game played on a 19×19 grid. Two players alternate placing black or white stones on board intersections to enclose the largest possible area while blocking the opponent. Since stones differ only in color, the key challenge is spatial positioning. Each GO question was multiple-choice, asking for the best sequence of moves. As difficulty increased, the required sequence length and board size grew (Chen et al., 2003).

Each puzzle type had three questions, with increasing difficulty, making for a total of twelve questions. The Tower of Hanoi and River Crossing questions were open-ended, while the Checkers and Go questions were multiple choice.

A Google Drive folder was created to organize the questions and responses. Inside, a "Question Bank" subfolder contained three folders: one for each AI system. Each system folder had two subfolders for LLM and SLM responses. A document listing all questions, correct answers, and Question IDs (organized by subject) was also included. Model responses were uploaded into the appropriate folders based on system and model type.

To organize the data, a Google Sheet was created inside the Google Drive folder. Inside the spreadsheet, separate columns were created for the index of the question, model type (LLM or SLM), AI commercial entity (ChatGPT, Claude, Deepseek), question difficulty (easy, medium, hard), question subject (Math, Reading, Writing, Puzzle), specific question type (Algebra, Geometry, Standard English Conventions, etc.), the response time for each repetition of the trial, the average response time of the three repetitions, the power and water consumption, and output accuracy.

The power and water consumption columns were computed using system-specific formulas based on the average response time of each trial.

The power was measured in kilowatts and was calculated using this formula:

$$E_{query}\ (kWh) = \frac{P_{critical} \times\ t\ (s) \times\ PUE}{3600}$$

This equation represents the energy consumed per query depending on the time taken for the model to respond to the query. The $P_{critical}$ is how much power is consumed from the model's GPU per hour, and it is multiplied by the *PUE* and the time taken for the system to give a final answer to the query. Because *t* is meant to be measured in hours, the equation was divided by 3600 as the response times were recorded in seconds (Jegham et al., 2025).

The water was measured in liters and was calculated using this formula:

$$W_{query}(L) = (WUE_{onsite}\ +\ WUE_{offsite}) \times\ E_{query}$$

The WUE_{onsite} and $WUE_{offsite}$ account for the water used inside and outside the data center. This is then multiplied by the energy used by the system to answer a query, resulting in the water consumption for the query (Jegham et al., 2025).

The 48 questions were fed into each AI entity's LLM and SLM, with response times recorded for each. The questions included multiple-choice reading, writing, and math, as well as multiple-choice or open-ended puzzle questions. All models were tested under identical conditions: similar laptops, Google Chrome browser, and a testing window from 2:00 pm to 5:00 pm. To ensure consistency, each question was entered three times. Response times for each trial were captured using the browser's developer tools. In total, 96 trials were conducted for each AI system. The average response time was calculated and used to estimate the power and water consumption of the models.

To begin data collection, ChatGPT was opened in the browser, and the GPT-4o model (ChatGPT's LLM) was selected. A temporary chat session was initiated to prevent the model from retaining any information. This was done by clicking the "Temporary Chat" button. The browser's developer console was then opened using Fn + F12 (or Command + Option + I for Mac).

Once setup was complete, questions were entered individually, and response times were recorded using the developer console. The tab was refreshed and the console cleared after each entry. Each question was submitted three times, with all response times logged. After the third attempt, a screenshot of the response was taken, named after the question's index, and uploaded to the corresponding AI system's folder.

This process was repeated for each question until all 48 trials were completed. Following this, the model was switched to GPT-4o-mini (ChatGPT's SLM), and the same procedure was followed to collect the SLM's data.

The process for Claude was similar, with slight modifications. Claude was opened in a browser tab, and the Sonnet 4 model (Claude's LLM) was selected. Since Claude does not offer a temporary chat option, the "New Chat" button was used, and the tab was refreshed after each response to eliminate information retention. The same input and timing process was followed, except the completion command was used instead of conversation. This was because every time Claude received a query, it looked at it as a task to complete, whereas ChatGPT looked at the query as a conversation to be had, hence the different command names between AI entities. After completing the 48 trials with Sonnet 4, the model was switched to Claude 3.5-Haiku (Claude's SLM), and the procedure was repeated for another 48 trials to capture SLM data.

The process for DeepSeek was similar to Claude's. DeepSeek's SLM (vl2-Tiny), however, cannot be launched on a browser and is hardware-dependent. This means that when the model is selected, its response time will vary depending on the hardware it is running on. Because of this, the SLM chosen for DeepSeek was V3. This model is not a reasoning model, nor is it dependent on computer hardware. For the LLM, DeepSeek's R1 model was chosen as it is a reasoning model and is made for more reasoning and comprehension-heavy tasks. DeepSeek was opened in a browser tab, and the R1 model was selected. Since DeepSeek does not provide a temporary chat option, the "New Chat" button was used, and the tab was refreshed after each response to eliminate information retention. The same input and timing process was followed as Claude. The "completion" command was timed, and after 48 trials, the model was switched to V3. The procedure was repeated for another 48 trials to capture SLM data.

The collected data was analyzed from multiple viewpoints. To maintain consistency, it was organized by subject. All coding and visualizations were performed in Jupyter Notebook using Matplotlib to generate linear graphs. In these graphs, the y-axis represented resource consumption (Power or Water), while the x-axis represented individual questions per subject. Questions increased in difficulty across the x-axis, starting with four easy, followed by four medium, and ending with four hard questions.

Each data point on the graph was color-coded: green indicated a correct response, and red an incorrect one, allowing resource usage to be interpreted in the context of system accuracy. Separate lines were plotted for SLMs and LLMs to enable direct comparisons. For instance, comparisons were made between an SLM and an LLM from the same AI entity on a single subject (ChatGPT Reading SLM vs. ChatGPT Reading LLM).

Additionally, graphs were generated to compare different systems using the same model type within a subject area (Math LLMs from Deepseek, ChatGPT, and Claude). Finally, combined graphs were created to show overall resource consumption across all three AI systems and both model types (SLM and LLM) per subject.

4. Results

The following section presents the results of our experiments evaluating the environmental footprint of large and small language models. Drawing from controlled testing across subjects and complexity levels, we report the power and water consumption of ChatGPT, Claude, and DeepSeek under identical conditions. These findings highlight key efficiency differences between SLM and LLM configurations, offering insight into the resource demands associated with model scale and task complexity.

Table 1: Data Analysis of Power Consumption Table

			Commercial Entity									
			ChatGPT			*Claude*			*DeepSeek*			
ID	**Subject**	**Complexity**	**SLM-Power**	**LLM-Power**	**Difference Factor**	**SLM-Power**	**LLM-Power**	**Difference Factor**	**SLM-Power**	**LLM-Power**	**Difference Factor**	**Highest Factor**
1	Math	Easy	0.0070	0.0129	1.853	0.0145	0.0160	1.104	0.0539	0.2249	4.170	DeepSeek
2	Math	Medium	0.0075	0.0225	3.004	0.0140	0.0166	1.187	0.0581	0.2881	4.961	DeepSeek
3	Math	Hard	0.0090	0.0252	2.805	0.0199	0.0238	1.192	0.0987	0.3752	6.461	DeepSeek
4	Reading	Easy	0.0079	0.0162	2.050	0.0255	0.0289	1.137	0.0503	0.1444	2.868	DeepSeek
5	Reading	Medium	0.0064	0.0147	2.282	0.0289	0.0298	1.030	0.0596	0.2145	3.600	DeepSeek
6	Reading	Hard	0.0069	0.0198	2.883	0.0277	0.0286	1.033	0.0541	0.2335	3.918	DeepSeek
7	Writing	Easy	0.0123	0.0342	2.787	0.0239	0.0276	1.155	0.0457	0.1655	3.618	DeepSeek
8	Writing	Medium	0.0100	0.0344	3.439	0.0386	0.0306	0.794	0.0491	0.3272	6.665	DeepSeek
9	Writing	Hard	0.0091	0.0371	4.085	0.0304	0.0320	1.052	0.0571	0.2388	4.864	DeepSeek
10	Puzzles	Easy	0.0314	0.0313	0.996	0.0595	0.0238	0.400	0.2183	0.5076	2.325	DeepSeek
11	Puzzles	Medium	0.0073	0.0126	1.731	0.0339	0.0287	0.847	0.2759	1.1713	4.246	DeepSeek
12	Puzzles	Hard	0.0192	0.0245	1.278	0.0793	0.0171	0.215	0.3801	1.6624	6.027	DeepSeek

The difference factor refers to the factor of how much more or less the LLM consumes in relation to the SLM. The Highest Factor column refers to the model with the biggest Difference Factor.

In *Table 1,* DeepSeek shows the highest Difference Factor between the SLM and LLM's power consumption across all subjects, often exceeding 4 and peaking at 6.665. Claude demonstrates the lowest power difference and has the most instances of the SLM consuming more than the LLM. Some tasks, like hard puzzles, showed an extremely small Difference Factor of 0.215, indicating efficient scaling. ChatGPT's power difference is moderate

and consistently above 1 in almost all query subjects. The highest Difference Factor (4.085) was observed in a hard writing query, suggesting that task complexity has a clear impact on power consumption efficiency.

Accuracy is a key component of a model's success. Looking at the queries submitted in the study, the models tend to miss query subjects such as Writing and Puzzles. Each model missed several different questions, of which the question IDs are listed below.

Questions Missed (Refer to Data Collection Sheet for question ID):

ChatGPT LLM: TR-MExp1, TR-MExp2, TR-MSta1, TR-EGo, TR-MGo, TR-HCheck, TR-MCross, TR-HCross
ChatGPT SLM: TR-HExp1

Claude LLM: TR-HExp2, TR-EGo, TR-MGo, TR-MCheck

Claude SLM: TR-HExp2, TR-HSta1, TR-MHanoi, TR-HHanoi, TR-EGo, TR-MGo, TR-ECheck, TR-HCross

DeepSeek LLM: TR-EGo, TR-MGo, TR-HGo

DeepSeek SLM: TR-HExp2, TR-MSta1, TR-HSta1, TR-EGo, TR-MGo, TR-HCross

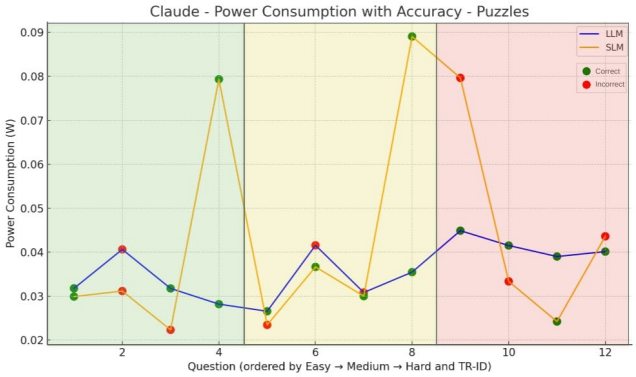

Figure 1: Claude Power Consumption Puzzles

Note. Applicable for Figures 1 - 9. X-axis is Query, and Y-axis is Power Consumption. The portion of the graph in green refers to easy questions, yellow refers to medium questions, and red refers to hard questions.

For the Puzzle category, Claude's LLM's power consumption standard deviation was 0.00604 kW, which is relatively stable compared to the SLM's standard deviation of 0.0244 kW. There are data points where the LLM consumed less power than the SLM, such as the fourth easy question (TR-ECross), the fourth medium question (TR-MCross), the first hard question (TR-HHanoi), and the fourth hard question (TR-HCross). The LLM got 66% of the questions correct while the SLM got 50% of the questions correct.

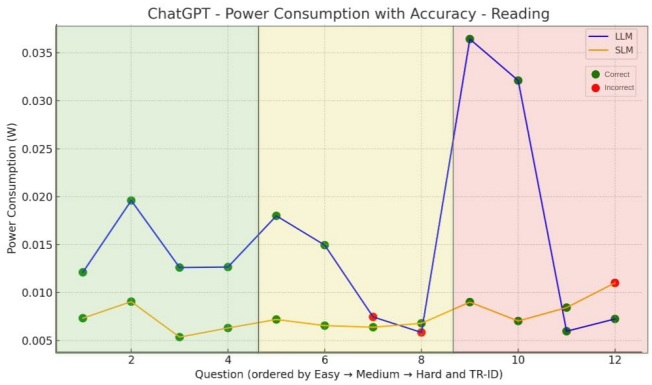

Figure 2: ChatGPT Power Consumption Reading

For the Reading questions, ChatGPT's LLM's power consumption varied more than its SLM's. The LLM had a standard deviation of 0.0099 kW, far greater than the SLM's standard deviation of 0.0015 kW. There were points where the LLM consumed less power than the SLM, such as the fourth medium question (TR-MExp2) and the last two hard questions (TR-HExp1, TR-HExp2). The LLM's power consumption peaked at first and second hard questions (TR-HInfo1, TR-HInfo2) with values of 0.0364 kW and 0.0321 kW, respectively. The LLM got 75% of the reading questions correct while the SLM got 91.67% of the questions correct.

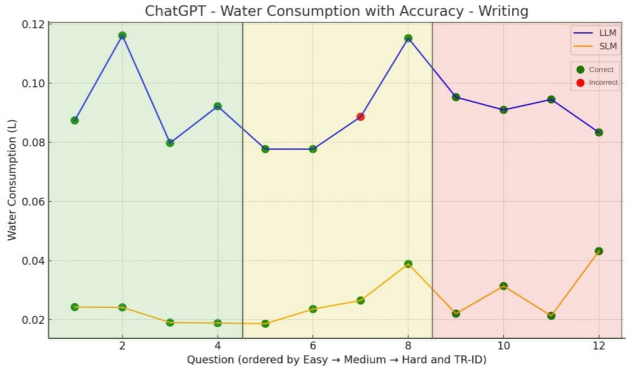

Figure 3: ChatGPT Water Consumption Writing

For the Writing questions, ChatGPT's LLM's average water consumption of 0.092 L was higher than its SLM's average water consumption of 0.026 L. The LLM's water consumption peaked at the third easy question (TR-ESta1) with a value of 0.116 L, which was greater than the SLM's peak of 0.0432 L at the fourth hard question (TR-HSta2). The LLM was 91.67% accurate in terms of correct answers, while the SLM was 100% accurate.

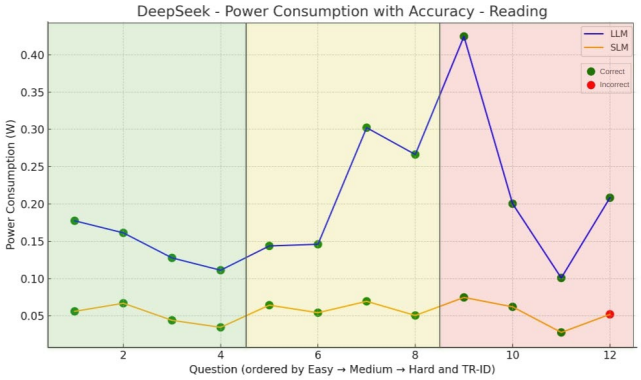

Figure 4: DeepSeek Power Consumption Reading

For the reading questions, DeepSeek's LLM's power consumption was higher on average compared to the SLM's power consumption. DeepSeek's LLM's average power consumption of 0.197 kW was much higher than its SLM's average power consumption of 0.0547 kW. The LLM's power consumption peaked at 0.425 kW at the first hard question (TR-HInfo1). The SLM's power consumption also peaked at the first hard question (TR-HInfo1) with a value of 0.0747 kW. The LLM was 100% accurate while the SLM was 91.67% accurate.

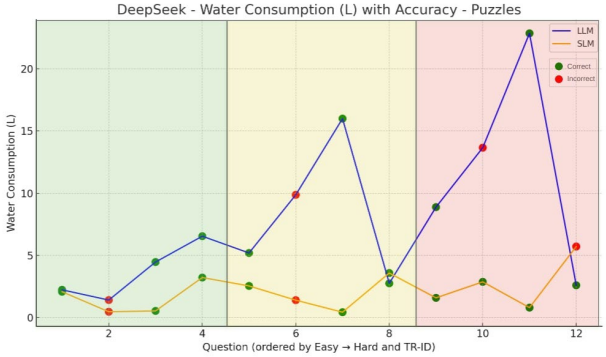

Figure 5: DeepSeek Water Consumption Puzzles

For the Puzzles questions, DeepSeek's LLMs' water consumption was higher on average than the SLM's. The LLM's average water consumption was 8.037 L, far greater than the SLM's average water consumption of 2.103 L. There were points where the LLM consumed less water than the SLM, such as the fourth medium question (TR-MCross) and fourth hard question (TR-HCross). The LLM's water consumption peaked at 22.85 L for the third hard question (TR-HCheck), while the SLM's water consumption peaked at 5.712 L for the fourth hard question (TR-HCross). The LLM got 75% of the questions correct, and the SLM also got 75% of the questions correct.

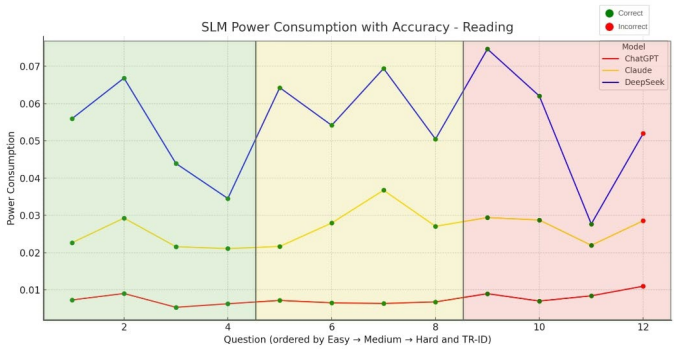

Figure 6: SLM Power Consumption Reading

For the Reading questions, the power consumption of Claude and ChatGPT SLMs did not show much variation, while DeepSeek's SLM varied more. ChatGPT's SLM had a standard deviation of 0.00538 kW. Similarly, Claude's SLM had a standard deviation of 0.00555 kW, far less than DeepSeek's SLM's standard deviation of 0.0141 kW. DeepSeek's SLM peaked at 0.0747 kW for the first hard question, Claude's SLM peaked at 0.0368 kW for the third medium question, and ChatGPT's SLM peaked at 0.011 kW for the fourth hard question. All of the SLMs were 91.67% accurate as they all answered incorrectly on the fourth hard question.

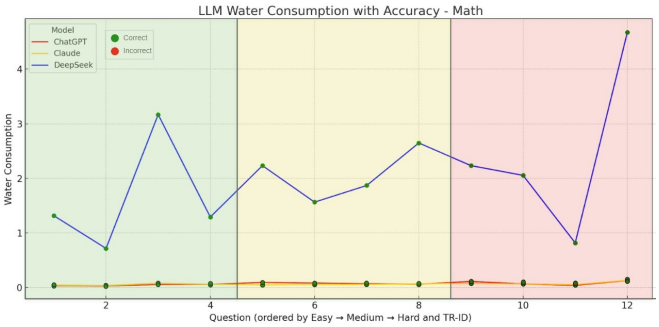

Figure 7: LLM Water Consumption Math

For the Math questions, the LLMS of Claude and ChatGPT had similar water consumption levels throughout each question, while DeepSeek had significantly higher water consumption levels per question. ChatGPT's and Claude's LLMs had average water consumptions of 0.0694 L and 0.0647 L, respectively. This was far less than DeepSeek's LLM's average water consumption of 2.049 L. DeepSeek's LLM also had a standard deviation of 1.097 L, far higher than ChatGPT's and Claude's (0.0302 L and 0.0245 L, respectively). DeepSeek's LLM peaked at 4.674 L on the fourth hard question. Claude's LLM peaked at 0.131 L on the fourth hard question. ChatGPT's LLM peaked at 0.128 L on the fourth hard question. All models were 100% accurate in responses.

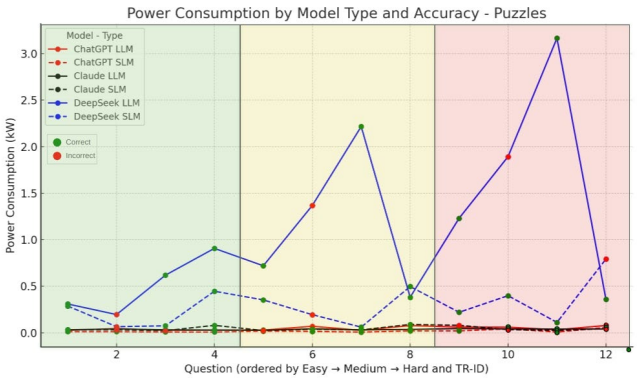

Figure 8: ALL Power Consumption Puzzles

For the Puzzles questions, the power consumption of ChatGPT and Claude throughout the models was relatively lower and did not show much difference between the SLM and LLM versions. On the other hand, DeepSeek's SLM and LLM were both continuously higher in power consumption compared to all the other models.

DeepSeek's models were 75% accurate. Claude's LLM was 58.33% accurate, and its SLM was 50% accurate. ChatGPT's LLM was 58.33% accurate, and its SLM was 100% accurate.

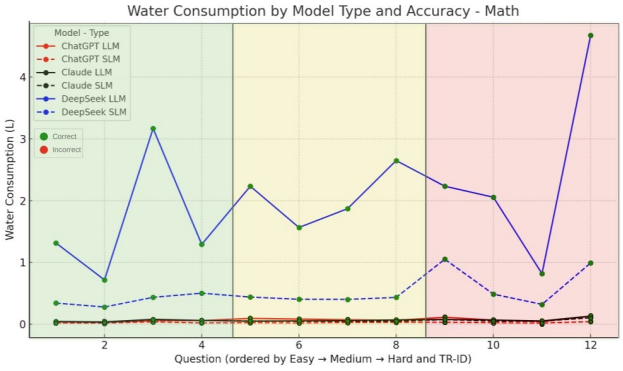

Figure 9: ALL Water Consumption Math

For the Math questions, DeepSeek's LLM had the highest average power consumption, and its SLM had the second highest. All the other models exhibited minimal differences. All models had 100% accuracy in responses.

5. Discussion

As per the data collection sheet, the energy and water consumption of DeepSeek surpasses both ChatGPT and Claude throughout all subjects, query complexities, and model types. As expected, LLMs utilized more resources than SLMs because of their higher computational requirements. The LLMs of DeepSeek demonstrated the highest levels of consumption in all cases. During Math queries, DeepSeek's LLM consumed 0.2839 kW on average, yet ChatGPT used 0.0202 kW and Claude used 0.0188 kW. The same pattern appeared in water usage measurements. The power and water consumption of DeepSeek exceeded the other models by at least nine times throughout all subjects and query complexities. ChatGPT's SLM proved to be the most energy-efficient model for Math, Reading, and Writing tasks. However, for Writing tasks, Claude's LLM surprisingly outperformed ChatGPT's LLM in efficiency. In the Puzzles category, no model consistently outperformed the other, consumption varied, suggesting near-equivalent efficiency depending on the complexity and type of the query.

The high resource usage stems from longer system latency, which likely results from server location differences. DeepSeek operates its infrastructure from China, but the testing took place in the United States. The measurement of energy consumption relied on response duration, which led to increased power and water usage because of network latency.

Even with its higher environmental impact, DeepSeek did not outdo other models in performance. The accuracy analysis revealed that ChatGPT's LLM had the best average scores, including a perfect score in Math (100.0%), and high scores in Writing (91.67%) and Reading (83.33%). DeepSeek's accuracy was generally lower. Claude offered a balanced profile, with moderate accuracy and the lowest resource use, so it's less resource-intensive than DeepSeek but not as accurate as ChatGPT.

This indicates a tradeoff between resource efficiency and the performance of models. For simpler tasks, Claude's or ChatGPT's SLMs may be the best choice due to their minimal resource use and no impact on accuracy. In Math, these models used an average of 0.0202 kW compared to the LLMs (0.0271 kW), while maintaining accuracy. But for more complex skills, like puzzles, the LLMs were more accurate. In such contexts, the higher resource utilization may be justified by the better accuracy and capability.

While SLMs showed high accuracy and resource efficiency in algorithmic subjects such as Math, their accuracy quickly diminished when faced with queries calling for reasoning and comprehension, such as Puzzles, and in some cases, Reading and Writing. For example, Claude's SLM achieved perfect accuracy in Math with an average of just 0.01618 kW consumed per query. However, its accuracy dropped to 50% in Puzzle-related queries despite an increase in resource consumption (0.0436 kW).

This suggests a subject-oriented breaking point for SLM accuracy. However, for more complex tasks that require thought and reasoning, ChatGPT's LLM proved to have performed markedly better than the other models. This suggests a subject-specific threshold where SLMs fail to maintain response accuracy with queries in reasoning and comprehensive subjects.

The graphs revealed spikes in resource consumption among LLMs for certain questions, typically correlating with an increase in complexity or response times. Interestingly, LLMs presented unpredictable resource consumption patterns, especially when faced with mid-to-high complexity queries in Writing or Puzzles. Some instances showcased LLMs consuming less than their SLM counterparts for the same question. These anomalies were seen most frequently in Claude as the SLM and LLM consumption seemed to be close together and interchanging throughout all subjects and complexities. This means it cannot be assumed that SLMs are more resource-efficient at all times. Real-world performance may vary based on multiple interacting factors.

These findings parallel Li et al. (2020), who demonstrated that compressed or distilled models can retain much of the performance of larger models while reducing energy use by nearly an order of magnitude. However, while their results showed that post-training optimization preserves accuracy even for complex tasks, the present study suggests that commercially deployed SLMs, most of which rely on architectural downsizing rather than distillation, still experience a notable drop in reasoning accuracy at higher complexities. This contrast implies that sustainable model design cannot rely solely on smaller parameter counts; it must also integrate structural optimization and intelligent compression strategies to maintain efficiency without sacrificing reasoning ability. Thus, Li et al's findings highlight a pathway to bridge the gap observed in this study, where smaller models remain efficient but falter in complex cognitive domains.

These observations highlight the importance of implementing context-aware model switching strategies. For routine or fact-based queries, systems should utilize their SLM models, which offer lower resource consumption and minimal to no accuracy tradeoff. However, as task complexity increases, particularly in subjects requiring comprehension or reasoning, SLM responses tend to be inadequate, necessitating switching to more resource-intensive LLMs. In settings that don't necessitate high-quality, reasoning responses, implementing a model-switching system will reduce resource consumption significantly while preserving acceptable accuracy.

However, this study does not go without its limitations. Deepseek's higher resource consumption correlates with network-based latency. Deepseek's server is located in China, and the testing took place in the U.S., causing a possible delay in response times and subsequently inflated calculated resource consumption. Additionally, the study excluded image and OCR (Optical Character Recognition) based queries, meaning the findings can only be extrapolated to text-based questions within the test subjects. Although testing environments were standardized, internet traffic or server load during certain times of day could slightly vary response times. Lastly, system-specific differences in treatment, such as the use of temporary chat or refreshing the browser, could have caused a slight difference in performance across queries.

6. Future Directions

In summary, the findings emphasize the importance of balancing environmental impact and system performance when utilizing AI models in real-world scenarios. The drop observed in the accuracy of SLMs when answering queries on Puzzles and reasoning demonstrates a threshold for when context-aware model-switching should occur. Future research should aim to quantify this breakpoint more accurately and look at the implementation of automatic query classification systems capable of efficiently switching between models to create more efficient AI models. This approach will allow for better scalability and environmental sustainability of AI models for the future.

7. Conclusion

The study examined the environmental impact of SLMs and LLMs while analyzing their performance across different levels of task complexity. The research showed that SLMs mostly require less energy and water than LLMs, yet their **accuracy** deteriorates when dealing with **more** complex queries, especially when reasoning and comprehension are required. **The relationship between query complexity and resource consumption showed little correlation for SLMs, however LLMs' resource consumption generally increased as query complexity increased.** The results showed LLMs maintained better performance on complex queries, even though they sometimes produced incorrect answers. However, they are much more resource-intensive, **consuming more power and water than SLMs**. The research demonstrates the benefits of using models based on context. AI models should start with SLMs before moving to LLMs, and vice versa, for specific tasks. The proposed approach merges environmental sustainability benefits with accuracy levels to create a workable model for AI implementation. The research demonstrates the need to match computational resources with task requirements, which will create a sustainable and efficient AI future.

Acknowledgements

We thank the DiscoverSTEM Innovation and Research Lab for hosting our research. We thank our mentors Mirza Faizan and Sheik Ahamed (DiscoverSTEM Innovation and Research Lab, Texas, United States) for guiding us through research design and experimentation. We thank Dr. Zisis Kozlakidis (International Agency for Research on Cancer - World Health Organization, Lyon, France) for advising us on the best practices on manuscript authorship.

Ethical Declaration: This paper did not require ethical clearance.

AI Declaration: AI tools were not used to write this paper.

References

Chen, X., Zhang, D., Zhang, X., Li, Z., Meng, X., He, S. and Hu, X. (2003) 'A functional MRI study of high-level cognition', Cognitive Brain Research, 16(1), pp. 32–37. https://doi.org/10.1016/s0926-6410(02)00206-9

College Board (n.d.) SAT Practice Questions. Available at: https://satsuitequestionbank.collegeboard.org/

De Vries, A. (2023) 'The growing energy footprint of artificial intelligence', Joule, 7(10), pp. 2191–2194. https://doi.org/10.1016/j.joule.2023.09.004

Desislavov, R., Martínez-Plumed, F. and Hernández-Orallo, J. (2023) 'Trends in AI inference energy consumption: Beyond the performance-vs-parameter laws of deep learning', Sustainable Computing: Informatics and Systems, 38, 100857. https://doi.org/10.1016/j.suscom.2023.100857

Environmental and Energy Study Institute (EESI) (n.d.) Data centers and water consumption | Article | EESI. Available at: https://www.eesi.org/articles/view/data-centers-and-water-consumption

Henderson, P., Hu, J., Romoff, J., Brunskill, E., Jurafsky, D. and Pineau, J. (2020) 'Towards the systematic reporting of the energy and carbon footprints of machine learning', Journal of Machine Learning Research, 21(248), pp. 1–43. Available at: https://jmlr.csail.mit.edu/papers/volume21/20-312/20-312.pdf

Jegham, N., Abdelatti, M., Elmoubarki, L. and Hendawi, A. (2025) 'How Hungry is AI? Benchmarking Energy, Water, and Carbon Footprint of LLM Inference', arXiv.org, 14 May. Available at: https://arxiv.org/abs/2505.09598

Li, Z., Wallace, E., Shen, S., Lin, K., Keutzer, K., Klein, D. and Gonzalez, J.E. (2020) 'Train large, then compress: Rethinking model size for efficient training and inference of transformers', arXiv.org, 26 February. Available at: https://arxiv.org/abs/2002.11794

Magister, L.C., Mallinson, J., Adamek, J., Malmi, E. and Severyn, A. (2022) 'Teaching small language models to reason', arXiv.org, 16 December. Available at: https://arxiv.org/abs/2212.08410

Parshin, S., Mirzadeh, I., Alizadeh, K., Horton, M., Bengio, S., Farajtabar, M. and Apple (n.d.) The illusion of thinking: Understanding the strengths and limitations of reasoning models via the lens of problem complexity. Apple.

Schaeffer, R., Miranda, B. and Koyejo, S. (2023) 'Are emergent abilities of large language models a mirage?', arXiv.org, 28 April. Available at: https://arxiv.org/abs/2304.15004

Schaeffer, R., Schoelkopf, H., Miranda, B., Mukobi, G., Madan, V., Ibrahim, A., Bradley, H., Biderman, S. and Koyejo, S. (2024) 'Why has predicting downstream capabilities of frontier AI models with scale remained elusive?', arXiv.org, 6 June. Available at: https://arxiv.org/abs/2406.04391

Strubell, E., Ganesh, A. and McCallum, A. (2019) 'Energy and policy considerations for deep learning in NLP', arXiv.org, 5 June. Available at: https://arxiv.org/abs/1906.02243

Tan, M. and Le, Q.V. (2019) 'EfficientNet: Rethinking model scaling for convolutional neural networks', arXiv.org, 28 May. Available at: https://arxiv.org/abs/1905.11946

Vogginger, B., Rostami, A., Jain, V., Arfa, S., Hantsch, A., Kappel, D., Schäfer, M., Faltings, U., Gonzalez, H.A., Liu, C., Mayr, C. and Maaß, W. (2024) 'Neuromorphic hardware for sustainable AI data centers', arXiv.org, 4 February. Available at: https://arxiv.org/abs/2402.02521

Wang, Y., Wang, M., Manzoor, M.A., Liu, F., Georgiev, G., Das, R.J. and Nakov, P. (2024) 'Factuality of large language models: A survey', arXiv.org, 4 February. Available at: https://arxiv.org/abs/2402.02420

Wilkins, J. (2025b) 'Former Google CEO tells Congress that 99 percent of all electricity will be used to power superintelligent AI', Futurism, 12 April. Available at: https://futurism.com/google-ceo-congress-electricity-ai-superintelligence