SLM-Bench: A Comprehensive Benchmark of Small Language Models on Environmental Impacts

Nghiem Thanh Pham[1], Tung Kieu[2], Duc-Manh Nguyen[3,6], Son Ha Xuan[4], Nghia Duong-Trung[5,6], Danh Le-Phuoc[3]

[1]FPT University, Vietnam, [2]Aalborg University, Denmark,

[3]Technische Universität Berlin, Germany, [4]RMIT University, Vietnam,

[5]German Research Center for Artificial Intelligence, Germany, [6]HiveIntel GmbH, Germany

[1]nghiemptce160353@fpt.edu.vn, [2]tungkvt@cs.aau.dk, [3]{duc.manh.nguyen,danh.lephuoc}@tu-berlin.de,

[4]ha.son@rmit.edu.vn, [5]nghia_trung.duong@dfki.de

Abstract

Small Language Models (SLMs) offer computational efficiency and accessibility, yet a systematic evaluation of their performance and environmental impact remains lacking. We introduce SLM-Bench, the first benchmark specifically designed to assess SLMs across multiple dimensions, including accuracy, computational efficiency, and sustainability metrics. SLM-Bench evaluates 15 SLMs on 9 NLP tasks using 23 datasets spanning 14 domains. The evaluation is conducted on 4 hardware configurations, providing a rigorous comparison of their effectiveness. Unlike prior benchmarks, SLM-Bench quantifies 11 metrics across correctness, computation, and consumption, enabling a holistic assessment of efficiency trade-offs. Our evaluation considers controlled hardware conditions, ensuring fair comparisons across models. We develop an open-source benchmarking pipeline with standardized evaluation protocols to facilitate reproducibility and further research. Our findings highlight the diverse trade-offs among SLMs, where some models excel in accuracy while others achieve superior energy efficiency. SLM-Bench sets a new standard for SLM evaluation, bridging the gap between resource efficiency and real-world applicability.

1 Introduction

Recent advancements in Language Models (LMs) have profoundly influenced a wide range of domains, including finance (Theuma and Shareghi, 2024), healthcare (Yang et al., 2024a), and manufacturing (Li et al., 2024). These models have demonstrated exceptional capabilities in retaining vast amounts of knowledge (Zheng et al., 2023), solving highly complex tasks (Bursztyn et al., 2022), and conducting intricate reasoning processes (Yang et al., 2024b) that closely align with human-level intentions. Their ability to understand context, generate coherent text, and adapt to diverse applications has made them indispensable tools in both academic research and industry practices.

However, the impressive performance of LMs often comes at a significant cost. The large number of parameters that enable their functionality requires extensive computational resources (Lv et al., 2024), leading to prohibitively high operational expenses. Additionally, these models demand intensive energy consumption, which not only drives up costs but also contributes to substantial environmental challenges. The carbon footprint associated with training and deploying Large Language Models (LLMs) has raised concerns about their sustainability (Faiz et al., 2024), as they emit significant amounts of CO_2. These issues underline the pressing need for more efficient and environmentally conscious alternatives to LLMs, particularly in an era of growing awareness around climate change and resource optimization.

SLMs (Gao et al., 2023; Magister et al., 2023) have emerged as a promising solution to mitigate the negative impacts associated with LLMs. By significantly reducing the number of parameters, SLMs aim to lower computational costs, minimize energy consumption, and decrease the associated carbon emissions, making them a more sustainable alternative. Their potential has garnered increasing attention from both academic researchers and industry practitioners, positioning SLMs as a practical choice for applications requiring efficiency and scalability.

Despite their growing prominence, a notable gap exists in the systematic evaluation of SLMs. Currently, there is no dedicated benchmark that comprehensively assesses the performance of SLMs

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 21369–21392
November 4-9, 2025

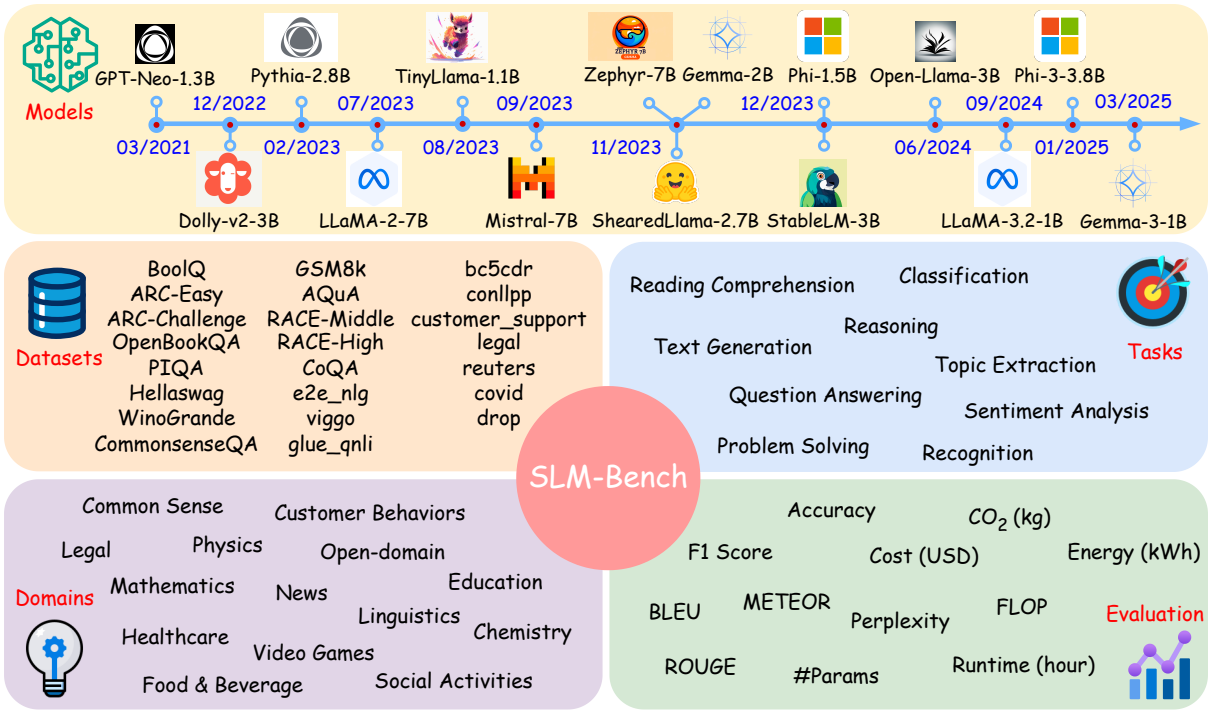


Figure 1: Overview of SLM-Bench

across diverse tasks while also quantifying their environmental impacts. This lack of standardized evaluation hinders a deeper understanding of their practical implications, particularly in resource-constrained environments where efficiency and sustainability are paramount. Addressing this gap is essential to unlock the full potential of SLMs and guide their development and deployment to balance performance with environmental responsibility.

In this paper, we aim to address the existing gap by introducing SLM-Bench (**S**mall **L**anguage **M**odel-**Bench**mark), a comprehensive benchmarking framework designed to evaluate SLMs across diverse settings with a focus on their environmental impacts. The insights gained from SLM-Bench will be instrumental in guiding the development and deployment of SLM-powered applications, particularly in resource-constrained environments. SLM-Bench is characterized by three primary features: (i) **Focus on SLMs:** Unlike existing benchmarks that predominantly emphasize LLMs (Myrzakhan et al., 2024), SLM-Bench evaluates SLMs that are computationally efficient and accessible to a wider range of users and systems. (ii) **Measurement of Environmental Impacts:** A unique aspect of SLM-Bench is its integration of metrics for energy consumption and CO_2 emissions. This allows for a holistic assessment of the sustainability of SLMs, addressing a crucial aspect often overlooked in benchmarking efforts. (iii) **Evaluation Across Diverse Settings:** SLM-Bench rigorously evaluates 15 SLMs on 9 tasks using 23 datasets from 14 domains. The evaluation is conducted on 4 hardware configurations, providing a comprehensive analysis of their performance across varied scenarios. This extensive evaluation ensures a deeper understanding of SLM capabilities and trade-offs.

SLM-Bench is illustrated in Figure 1, where we provide an overview of its key components, including the selected SLMs, the datasets used, the selected domains, the tasks evaluated, and the metrics employed for assessment. To the best of our knowledge, SLM-Bench represents the first systematic benchmarking framework dedicated to the evaluation of SLMs with a specific focus on their environmental impacts. We hope that SLM-Bench will drive construction through evaluation, facilitating the development of SLMs in many more applications. Aiming at reproducibility and transparency, the source code and homepage for SLM-Bench are provided at https://github.com/HiveIntel/SLM-Bench.git and

https://swarm.hiveintel.ai/leaderboard/, respectively, which will be continuously updated to catch up with state-of-the-arts. In summary, our contributions are as follows.

- We introduce SLM-Bench, a comprehensive benchmark of SLMs, which are computationally efficient and more accessible for deployment in resource-constrained environments.
- We consider the environmental impacts, such as energy consumption and CO_2 emissions, enabling a holistic assessment of the sustainability of SLMs.
- We extensively evaluate 15 SLMs across 9 tasks using 23 datasets from 14 domains on 4 hardware configurations, offering a thorough understanding of their capabilities and limitations in various contexts.

2 Related Work

2.1 Small Language Models

LLMs, such as GPT-4 (OpenAI, 2023), LLaMA (Touvron et al., 2023), and PaLM (Vilar et al., 2023), have demonstrated exceptional capabilities across various tasks. However, their large parameter sizes lead to high computational costs (Sharir et al., 2020), energy demands (Strubell et al., 2020), and environmental impacts (Dhar, 2020). To address these challenges, SLMs have gained attention for their efficiency and scalability. Techniques like model pruning (Zhang et al., 2024b; Sun et al., 2024; Ma et al., 2023), knowledge distillation (Gu et al., 2024; Zhang et al., 2024a), and low-rank factorization (Hu et al., 2022; Xu et al., 2024) have enabled models like DistilBERT (Sanh et al., 2019) and TinyBERT (Jiao et al., 2020) to achieve strong performance with fewer parameters. Existing benchmarks, such as GLUE (Wang et al., 2019b) and SuperGLUE (Wang et al., 2019a), primarily evaluate LLMs and lack a focus on SLMs and their performance. This leaves a gap in understanding the trade-offs between efficiency, performance, and sustainability.

2.2 Emission-aware Benchmarking on Language Models

The development of LMs demands substantial computational resources, raising environmental concerns. Strubell et al. (Strubell et al., 2019) first quantified LMs' carbon footprint, and later work (Strubell et al., 2020; Patterson et al., 2021) underscored rising energy demands and sustainability trade-offs. While tools like CodeCarbon (Lottick et al., 2019) and ML CO_2 Impact(Lacoste et al., 2019) enable emissions tracking, benchmarks such as Big-Bench (Authors, 2023) focus largely on performance, neglecting energy impact. Recent work has addressed full-lifecycle emissions. For instance, Luccioni et al.(Luccioni et al., 2023) showed that BLOOM's total emissions (50.5 tCO_2) exceeded training emissions due to hardware and idle energy. Hardware choice matters too–T4 to A100 GPU migration can cut emissions by 83%(Liu and Yin, 2024). Fine-tuning also incurs significant cost (Wang et al., 2023). Gowda et al. (Gowda et al., 2023) proposed the Sustainable-Accuracy Metric to balance performance and efficiency. Poddar et al.(Poddar et al., 2025) benchmark LLM inference energy, while Singh et al.(Singh et al., 2024) explore sustainable training and deployment. However, these efforts mainly target LLMs and partial aspects of environmental impact. To our knowledge, SLM-Bench is the first benchmark focused specifically on SLMs with a comprehensive evaluation of both computational and environmental metrics.

3 Benchmarking Design

3.1 Data Collection

We extensively collect 23 datasets from 11 diverse domains, including common sense, mathematics, physics, news, and legal, among others. These datasets encompass 9 task types, covering reading comprehension, text classification, logical reasoning, sentiment analysis, and more. In total, our dataset collection comprises 799,594 samples, ensuring a well-rounded evaluation framework. Table 1 provides an overview of the collected datasets along with their associated characteristics. Due to space limitations, we provide the details and examples of these datasets in Appendices B and A, respectively. Our selection of these datasets is motivated by their widespread adoption in previous studies (Myrzakhan et al., 2024), ensuring compatibility and comparability with existing benchmarks. Furthermore, we aim to evaluate SLMs from multiple perspectives, assessing their generalization abil-

ity, reasoning capabilities, and robustness across diverse tasks and domains. Integrating datasets from various fields ensures a comprehensive and challenging benchmark that reflects real-world applications.

Dataset	#Samples	Domain	Task
BoolQ	15,432	Open-domain	Question Answering
ARC-Easy	5,876	Open-domain	Question Answering
ARC-Challenge	2,590	Open-domain	Question Answering
OpenBookQA	5,957	Open-domain	Question Answering
PIQA	16,113	Physics	Reasoning
Hellaswag	10,421	Common Sense	Reasoning
WinoGrande	44,321	Common Sense	Reasoning
CommonsenseQA	12,102	Common Sense	Reasoning
GSM8k	8,034	Mathematics	Problem Solving
AQuA	99,765	Mathematics	Problem Solving
RACE-Middle	24,798	Education	Reading Comprehension
RACE-High	26,982	Education	Reading Comprehension
CoQA	127,542	Open-domain	Question Answering
e2e_nlg	50,321	Food & Beverage	Text Generation
viggo	9,842	Video Games	Text Generation
glue_qnli	104,543	Linguistics	Question Answering
bc5cdr	20,764	Chemistry	Recognition
conllpp	23,499	Linguistics	Recognition
customer_support	14,872	Customer Behaviors	Classification
legal	49,756	Legal	Classification
reuters	9,623	News	Topic Extraction
covid	19,874	Healthcare	Sentiment Analysis
drop	96,567	Open-domain	Reasoning

Table 1: Datasets' Details

3.2 Model Selection

We select 15 SLMs for our benchmarking based on three key criteria as follows. (i) **Model Size**: The primary criterion for classification as an SLM is the number of parameters. The LMs must have fewer than 7 billion parameters to qualify as SLMs. (ii) **Popularity & Reputation**: The selected SLMs should be widely recognized and developed by well-known organizations to ensure the impact. (iii) **Open-Source Availability**: The models must be open-source, allowing for transparency, reproducibility, and further research. These criteria ensure a fair, representative, and reproducible benchmarking process, focusing on models that are both accessible and widely used in the research community. Table 2 provides an overview of the selected models. Due to space limitations, we provide the details of these models in Appendix C.

3.3 Evaluation

We evaluate an SLM using 11 metrics, covering various aspects of performance and efficiency. These metrics assess accuracy, such as *BLEU*, and computational performance, such as *Runtime*. Our evaluation focuses solely on the fine-tuning process. Since we do not have access to the necessary

Model	#Params (B)	Size (GB)	Year	Provider
`GPT-Neo-1.3B`	1.37	2.46	03/2021	EleutherAI
`Dolly-v2-3B`	3	5.8	12/2022	Databricks
`Pythia-2.8B`	2.8	5.5	02/2023	EleutherAI
`LLaMA-2-7B`	6.47	13	07/2023	Meta
`TinyLlama-1.1B`	1.1	2	08/2023	SUTD
`Mistral-7B`	7	13	09/2023	Mistral AI
`Zephyr-7B`	7	13.74	11/2023	WebPilot.AI
`ShearedLlama-2.7B`	2.7	5	11/2023	Princeton NLP
`Gemma-2B`	2	4.67	11/2023	Google
`Phi-1.5B`	1.42	2.84	12/2023	Microsoft
`StableLM-3B`	3	6.5	12/2023	Stability AI
`Open-LLaMA-3B`	3	6.8	06/2024	OpenLM
`Llama-3.2-1B`	1.24	2.47	09/2024	Meta
`Phi-3-3.8B`	3.82	2.2	01/2025	Microsoft
`Gemma-3-1B`	1	2	03/2025	Google

Table 2: SLMs' Details (sort by release time)

data for full training, we rely exclusively on pre-trained models. For inference, we did not include runtime comparisons in the paper, as the differences between models were insignificant. Additionally, to measure the resource consumption of SLMs, we incorporate three key metrics: *Cost*, CO_2 emissions, and *Energy* usage. This comprehensive evaluation ensures a balanced assessment of both the effectiveness of the model and its environmental and financial impact. To measure resource consumption, we use the APIs of the experimental server to measure both FLOPs and cost. Specifically, we conducted experiments by renting a **lightning.ai**[1] server and recorded the deducted amount from our account balance as the cost. Additionally, we used a built-in function of the same platform to measure FLOPs. To estimate CO_2 emissions, we use the `ML CO2` package[2], calling its built-in function. For energy consumption, we use the `Zeus` package[3] in a similar manner. It is important to note that FLOP, costs, CO_2 emissions, and energy consumption vary across different hardware configurations. If end-users run the models on a local server, FLOPs can be measured using the `calflops` library. The cost can be estimated by multiplying energy consumption, runtime, and electricity price. Table 3 provides an overview of the evaluation metrics. Due to space limitations, we provide the details of these metrics in Appendix D.

[1] https://lightning.ai/
[2] https://mlco2.github.io/impact/
[3] https://ml.energy/zeus/

Metrics	Evaluation	Task	Dataset
Accuracy	Correctness	Question Answering Classification Recognition Reasoning Problem Solving Reading Comprehension	All except [e2e_nlg, viggo, reuters]
F1 Score	Correctness	Question Answering Classification Recognition Reasoning Problem Solving Reading Comprehension	All except [e2e_nlg, viggo, reuters]
BLEU	Correctness	Text Generation Topic Extraction	e2e_nlg, viggo, reuters
ROUGE	Correctness	Text Generation Topic Extraction	e2e_nlg, viggo, reuters
METEOR	Correctness	Text Generation Topic Extraction	e2e_nlg, viggo, reuters
Perplexity	Correctness	Text Generation Topic Extraction	e2e_nlg, viggo, reuters
Runtime	Computation	All	All
FLOP	Computation	All	All
Cost	Consumption	All	All
CO_2	Consumption	All	All
Energy	Consumption	All	All

Table 3: Metrics' Details

3.4 Benchmarking Pipeline

To enhance extensibility and reproducibility, we design a unified process pipeline for benchmarking, consisting of 7 key modules. The *Universal Data Loader* ensures consistency by converting datasets of different formats into a unified structure. The *Preprocessing* module then refines the data by trimming, removing special symbols, and applying necessary transformations. Once the data is prepared, the *Calling* module manages the execution of SLMs and tasks, enabling flexibility where a single SLM can be tested on multiple tasks and vice versa. After inference, the output undergoes further refinement through the *Post-processing* module before moving to the *Evaluation* module, which applies appropriate metrics to assess model performance. Finally, the *Report* module compiles results and visualizations, providing insights into the model's effectiveness. Additionally, the *Logging* module is integrated throughout the pipeline to record all events, enabling better traceability, debugging, and transparency. This modular design allows for scalability, flexibility, and ease of adaptation, making it well-suited for benchmarking and future extensions. Figure 2 illustrates the proposed pipeline.

3.5 Ranking Methodology

We propose a ranking method for SLMs based on their performance across multiple datasets and evaluation metrics. This method counts how often each model ranks first, second, or third in different experimental settings. Specifically, let K be the number of SLMs, N the number of datasets, T the number of tasks, and M the number of evaluation metrics. For each model k, we count the number of times it achieves the best (gold medal), second-highest (silver medal), and third-highest (bronze medal) performance across all $K \times N \times T \times M$ cases. This approach provides a straightforward yet effective way to assess overall model performance by considering rankings across diverse conditions rather than relying on a single aggregated metric. By aggregating the detailed benchmark results into a comprehensive summary, we aim to make insights more accessible to end-users, allowing users to quickly select models based on three major criteria: accuracy, efficiency, or energy consumption.

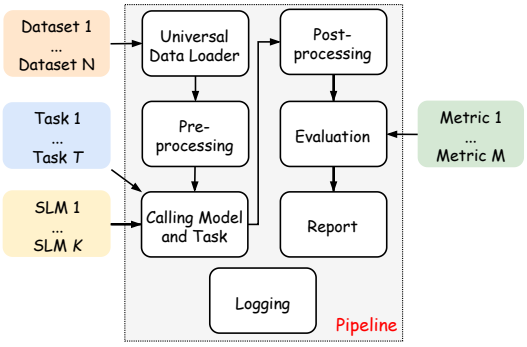


Figure 2: Benchmarking Pipeline

4 Experiments

4.1 Implementation Details

We conduct our experiments on `lightning.ai`, a cloud-based platform designed as a development studio for building, training, and deploying AI models. Lightning AI offers support for various hardware configurations, allowing flexibility in experimentation. We conduct evaluations across four hardware configurations: two server-grade and two edge-device setups. The server configurations use NVIDIA L4 and A10 GPUs, while the edge-device configurations use NVIDIA Jetson Orin AGX with 16GB and 64GB memory, respectively. Due to space constraints, the main paper reports results only on the NVIDIA L4 GPU. Results for the other configurations are provided on the extended ver-

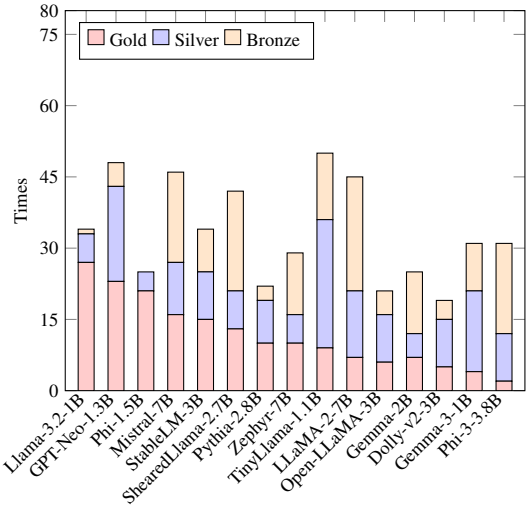


Figure 3: Overall Results

sion (Pham et al., 2025). Furthermore, while the magnitudes of the results vary on different hardware configurations, the relative ranking among the SLMs remains consistent. In addition, we implement the benchmarking pipeline (see Figure 2) by using `Python 3.10`, `Numpy 1.24`, `Pandas 1.5`, `PyTorch 2.0`, `Sklearn 1.2` and `Zeus-ML 0.7.0`.

4.2 Hyperparameter Settings

We select hyperparameters based on recommendations from various sources, including original research papers, official code repositories, and Hugging Face[4] model hubs, which often provide well-tuned configurations for strong performance. When specific values are unavailable for a given model-dataset pair, we apply a systematic fine-tuning strategy using a validation set to iteratively optimize performance. Our hyperparameter tuning process includes the following steps: (1) defining appropriate search ranges and (2) conducting a random search over the defined space. Specifically, we vary *learning_rate* from 1e-6 to 1e-4; *batch_size* among 2, 4, 8, 16, 32, 64, 128; fine-tuning *epochs* among 2, 4, 6, 8, 10; *LoRA_rank* among 2, 4, 6, 8, 10; and dropout probability among 0.1, 0.2, 0.3, 0.4, 0.5. This approach ensures a fair and consistent evaluation across all models and datasets.

4.3 Overall Results

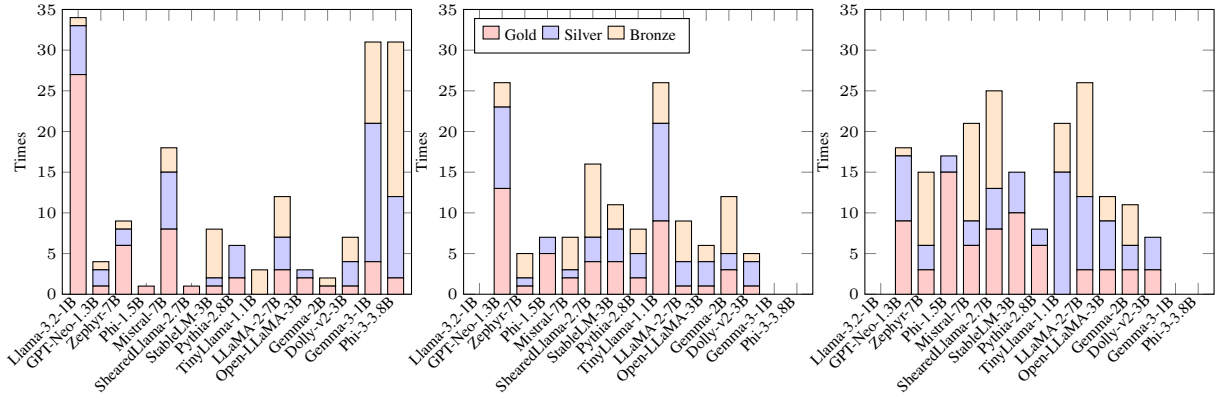
We present the overall results in Figure 3, which visualizes the ranking of each SLM across all datasets and evaluation metrics. The models are sorted from left to right based on the number of gold medals they have achieved. Our analysis reveals that `Llama-3.2-1B` outperforms the other SLMs, securing the highest number of gold medals. This indicates that it consistently delivers top-tier performance across a diverse range of evaluations. Additionally, `GPT-Neo-1.3B` emerges as the runner-up in terms of gold medal count, further demonstrating strong performance. Following closely, `Phi-1.5B` performs competitively, securing the third-highest number of gold medals. This model exhibit strong results, often excelling in specific datasets or metrics. Meanwhile, `Mistral-7B`, `TinyLlama-1.1B` and `LLaMA-2-7B`, while not securing the most gold medals, frequently appear in the top-three rankings. This suggests that these models maintain a high level of robustness and consistency across various tasks, even if they do not always achieve the top position.

Furthermore, `TinyLlama-1.1B` and `LLaMA-2-7B` appear in the middle of the rankings, securing a moderate number of gold medals while frequently earning silver and bronze medals. Although they do not dominate in terms of gold medal achievements, their balanced performance across different evaluation criteria suggests that they remain competitive across various tasks. Their ability to accumulate a significant number of medals overall indicates that they can serve as reliable alternatives to the top-performing models, especially in scenarios where consistency across multiple benchmarks is valued. Meanwhile, `Dolly-v2-3B`, `Gemma-3-1B`, and `Phi-3-3.8B` rank toward the lower end of the leaderboard, achieving the fewest gold medals among the evaluated models. While their performance is less dominant, they still manage to secure some silver and bronze medals, indicating that they exhibit strengths in certain areas. These results highlight the varying capabilities of SLMs and the potential trade-offs when selecting a model for specific applications. Appendix E further discusses the metric contribution to medals in 15 models.

4.4 Evaluation Taxonomy

We categorize the evaluation metrics into three distinct types: correctness, computation, and consumption (see Table 3). Figure 4 presents the per-

[4]https://huggingface.co/



(a) Correctness (b) Computation (c) Consumption

Figure 4: Taxonomy Results

formance results for each evaluation type, highlighting the strengths of different SLMs across these dimensions.

In terms of correctness, `Llama-3.2-1B` earns the most gold medals, making it the most accurate SLM in our evaluation. `Mistral-7B` follows closely, with `Gemma-3-1B` and `Phi-3-3.8B` also frequently ranking in the top three, reflecting their consistent output quality.

For computational efficiency, `GPT-Neo-1.3B` leads with the highest number of gold medals, suggesting strong performance in processing speed and resource usage. `TinyLlama-1.1B` and `ShearedLlama-2.7B` also rank highly, demonstrating notable efficiency across tasks.

In terms of resource consumption, `Phi-1.5B` stands out as the most energy-efficient model. `StableLM-3B` and `GPT-Neo-1.3B` share the second-highest number of gold medals, while `LLaMA-2.7B` and `ShearedLlama-2.7B` frequently appear among the top performers, highlighting their sustainable design.

Although *computation and energy consumption are often correlated, they are not equivalent*. Some models use more energy to achieve faster runtimes, while others with similar compute may be less efficient due to non-parallelizable operations and architectural bottlenecks.

We observe that *correctness does not strongly correlate with computation or consumption*. This is because accuracy depends not only on model size, but also on the quality and diversity of pretraining data. Similarly, *consumption and computational cost are influenced by more than just model size*–factors like model architecture, parallelization efficiency, and computational complexity play a significant role. For example, `Llama-3.2-1B`, *despite its small size, achieves the highest accuracy but performs poorly in computation and energy consumption*–an unexpected yet insightful outcome.

4.5 Performance Trade-off

We visualize the performance trade-offs of SLMs using Kiviat charts. Our evaluation considers three key performance aspects: correctness, computation, and consumption. Initially, we focus on the number of gold medals each SLM has achieved. To enhance readability, we select five top-performing SLMs from previous experiments: `Llama-3.2-1B`, `GPT-Neo-1.3B`, `Zephyr-7B`, `Phi-1.5B`, and `Mistral-7B`. Figure 5 illustrates the performance trade-offs among these models.

Our observations reveal that `Llama-3.2-1B` excels in correctness but falls short in computation and consumption efficiency. In contrast, `Phi-1.5B` demonstrates the highest efficiency in consumption but lags in correctness. Next, `GPT-Neo-1.3B` achieves the best performance in computation but also lags in correctness. `Mistral-7B` maintains a balanced performance across all three metrics, making it a strong candidate for scenarios requiring an all-around solution.

Next, instead of solely considering the number of gold medals, we adopt a score-based evaluation approach that accounts for all gold, silver, and bronze medals. Specifically, we assign a weighted score to each medal type: each gold medal contributes 3 points, each silver medal contributes 2

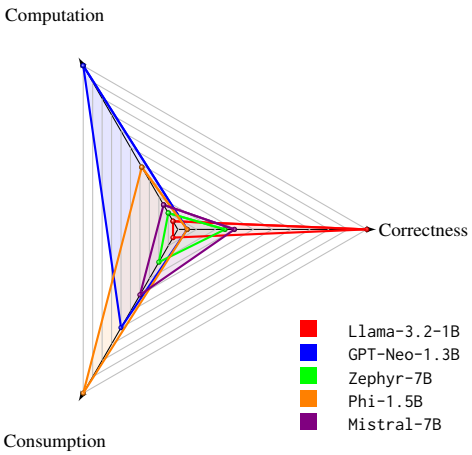


Figure 5: Performance Trade-off, only Gold Medals

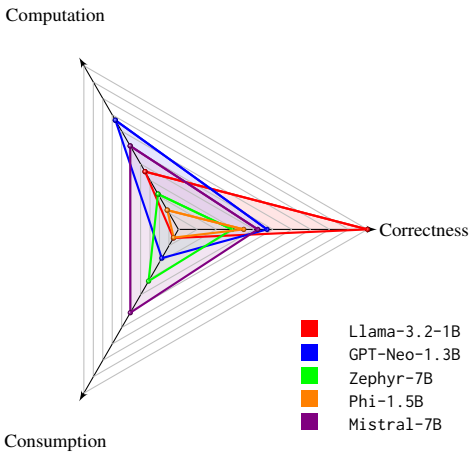


Figure 6: Performance Trade-off, Score-based Evaluation

points, and each bronze medal contributes 1 point. We then compute the total score for each SLM across the three evaluation categories by summing the respective scores. This method provides a more nuanced assessment of model performance, capturing relative strengths beyond just the highest achievements. To maintain consistency and readability, we again focus on the five top-performing SLMs identified in the previous experiment. Figure 6 illustrates the performance trade-offs among these models under the score-based evaluation framework.

Our observations remain consistent in that `Llama-3.2-1B` continues to excel in correctness while exhibiting weaker performance in computation and consumption efficiency. However, the score-based evaluation reveals additional insights. Notably, `Mistral-7B` outperforms `Phi-1.5B` and `Zephyr-7B` across all three evaluation metrics, indicating that it provides a more balanced trade-off between accuracy and efficiency. Furthermore, `GPT-Neo-1.3B` maintains well-rounded performance across correctness, computation, and consumption, reinforcing its suitability for scenarios where an all-purpose model is desirable.

4.6 Discussion

After conducting extensive experiments and analyzing the results, we can derive several key end-user recommendations when selecting an appropriate SLM. There are clear trade-offs among three key dimensions: correctness, computational efficiency, and resource consumption. If accuracy is the primary goal, `Llama-3.2-1B` is a strong choice, consistently ranking highest in correctness metrics. However, this comes with increased computational and energy costs. For scenarios requiring fast and efficient execution, `GPT-Neo-1.3B` offers superior runtime performance, making it suitable for latency-sensitive tasks. If energy efficiency and sustainability are critical, `Phi-1.5B` is the most resource-friendly option, ideal for deployment on low-power or edge devices. For applications needing a balance across all three criteria, `Mistral-7B` performs reliably and consistently, making it a versatile, well-rounded choice. Finally, the choice of SLM should be guided by specific application requirements, as different models offer varying strengths and weaknesses that may influence real-world deployment decisions.

5 Conclusion

We introduce `SLM-Bench`, a comprehensive benchmark for evaluating the environmental impact of SLMs. It provides critical insights to guide the selection, optimization, and deployment of models, helping researchers and practitioners balance accuracy, efficiency, and sustainability—especially in resource-constrained settings. To support future research, `SLM-Bench` includes an open-source, extensible pipeline for continuous integration. In future work, we plan to incorporate additional SLMs and explore a wider range of hardware settings. Additionally, we aim to enhance the accuracy of energy consumption and CO_2 emission measurements by utilizing sensor-equipped devices. We also plan to incorporate instruction–following capabilities as a distinct task dimension, extending the benchmark's coverage of real-world use cases.

6 Limitations

Although SLM-Bench provides a solid evaluation of SLMs, it has limitations. First, environmental metrics like energy consumption, CO_2 emissions, and runtime depend heavily on the hardware used. So far, we have only tested 4 configurations and plan to include more in future work. Second, energy consumption and CO_2 emissions are currently estimated using standardized formulas rather than real-time measurements. We aim to improve this by using sensor-equipped devices to capture actual energy usage and emissions.

7 Broader Impact

SLM-Bench promotes the development and deployment of efficient, sustainable, and accessible AI by focusing on SLMs. By evaluating models across accuracy, computation, and energy consumption, our benchmark helps users make informed, responsible choices—especially in resource-constrained environments. Our inclusion of environmental metrics raises awareness of AI's carbon footprint and supports more sustainable practices. By making the benchmark publicly available and regularly updated, we aim to drive progress in green AI and practical model deployment across both academia and industry.

8 Acknowledgements

This work was funded by the European Union's programme under grant agreement No.101092908 (SmartEdge), by the Chips Joint Undertaking (JU), European Union (EU) HORIZON-JU-IA, under grant agreement No. 101140087 (SMARTY) and by the German Research Foundation (DFG) under the COSMO project (ref. 453130567).

References

Maged S. Al-Shaibani and Irfan Ahmad. 2023. Consonant is all you need: a compact representation of english text for efficient NLP. In *Findings of the Association for Computational Linguistics (EMNLP)*, pages 11578–11588.

Big-Bench Authors. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transaction of Machine Learning Research*, 2023.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2020. PIQA: reasoning about physical commonsense in natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 7432–7439.

Quentin Brabant, Gwénolé Lecorvé, and Lina Maria Rojas-Barahona. 2022. Coqar: Question rewriting on coqa. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, pages 119–126.

Victor S. Bursztyn, David Demeter, Doug Downey, and Larry Birnbaum. 2022. Learning to perform complex tasks through compositional fine-tuning of language models. In *Findings of the Association for Computational Linguistics (EMNLP)*, pages 1676–1686.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. Boolq: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 2924–2936.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the AI2 reasoning challenge. *CoRR*, abs/1803.05457.

Payal Dhar. 2020. The carbon impact of artificial intelligence. *Nature Machine Intelligence*, 2(8):423–425.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 2368–2378.

Ondrej Dusek, Jekaterina Novikova, and Verena Rieser. 2018. Findings of the E2E NLG challenge. In *Proceedings of the International Conference on Natural Language Generation (INLG)*, pages 322–328.

Ahmad Faiz, Sotaro Kaneda, Ruhan Wang, Rita Chukwunyere Osi, Prateek Sharma, Fan Chen, and Lei Jiang. 2024. Llmcarbon: Modeling the end-to-end carbon footprint of large language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Ze-Feng Gao, Kun Zhou, Peiyu Liu, Wayne Xin Zhao, and Ji-Rong Wen. 2023. Small pre-trained language models can be fine-tuned as large models via over-parameterization. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 3819–3834.

Shreyank N. Gowda, Xinyue Hao, Gen Li, Laura Sevilla-Lara, and Shashank Narayana Gowda. 2023. Watt for what: Rethinking deep learning's energy-performance relationship. *CoRR*, abs/2310.06522.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. 2024. Minillm: Knowledge distillation of large language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Rashmi Gupta, Tushar Agarwal, Aviral Mehlawat, Deeksha Mathur, Devendra Kumar Somwanshi, and Anil Kumar. 2023. SMER: A novel medical entity recognition technique based on scispacy (bc5cdr) and med7. In *Proceedings of the International Conference on Information Management & Machine Intelligence (ICIMMI)*, pages 119:1–119:6.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2020. TinyBERT: Distilling BERT for natural language understanding. In *Findings of the Association for Computational Linguistics (EMNLP)*, pages 4163–4174.

Juraj Juraska, Kevin Bowden, and Marilyn A. Walker. 2019. Viggo: A video game corpus for data-to-text generation in open-domain conversation. In *Proceedings of the International Conference on Natural Language Generation (INLG)*, pages 164–172.

Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. 2019. Quantifying the carbon emissions of machine learning. *CoRR*, abs/1910.09700.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard H. Hovy. 2017. RACE: large-scale reading comprehension dataset from examinations. In *Proceedings of the Conference on Empirical Methods in Natural Language Processin (EMNLP)*, pages 785–794.

David D. Lewis, Yiming Yang, Tony G. Rose, and Fan Li. 2004. RCV1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research*, 5:361–397.

Yiwei Li, Huaqin Zhao, Hanqi Jiang, Yi Pan, Zhengliang Liu, Zihao Wu, Peng Shu, Jie Tian, Tianze Yang, Shaochen Xu, Yanjun Lyu, Parker Blenk, Jacob Pence, Jason Rupram, Eliza Banu, Ninghao Liu, Linbing Wang, Wen-Zhan Song, Xiaoming Zhai, Kenan Song, Dajiang Zhu, Beiwen Li, Xianqiao Wang, and Tianming Liu. 2024. Large language models for manufacturing. *CoRR*, abs/2410.21418.

Bingbin Liu, Sébastien Bubeck, Ronen Eldan, Janardhan Kulkarni, Yuanzhi Li, Anh Nguyen, Rachel Ward, and Yi Zhang. 2023. Tinygsm: Achieving >80% on gsm8k with small language models. *CoRR*, abs/2312.09241.

Vivian Liu and Yiqiao Yin. 2024. Green AI: exploring carbon footprints, mitigation strategies, and trade-offs in large language model training. *Discovery Artificial Intelligence*, 4(1):49.

Kadan Lottick, Silvia Susai, Sorelle A. Friedler, and Jonathan P. Wilson. 2019. Energy usage reports: Environmental awareness as part of algorithmic accountability. *CoRR*, abs/1911.08354.

Alexandra Sasha Luccioni, Sylvain Viguier, and Anne-Laure Ligozat. 2023. Estimating the carbon footprint of bloom, a 176b parameter language model. *Journal of Machine Learning Research*, 24:253:1–253:15.

Kai Lv, Yuqing Yang, Tengxiao Liu, Qipeng Guo, and Xipeng Qiu. 2024. Full parameter fine-tuning for large language models with limited resources. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 8187–8198.

Xinyin Ma, Gongfan Fang, and Xinchao Wang. 2023. Llm-pruner: On the structural pruning of large language models. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS)*.

Lucie Charlotte Magister, Jonathan Mallinson, Jakub Adámek, Eric Malmi, and Aliaksei Severyn. 2023. Teaching small language models to reason. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1773–1781.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? A new dataset for open book question answering. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2381–2391.

Aidar Myrzakhan, Sondos Mahmoud Bsharat, and Zhiqiang Shen. 2024. Open-llm-leaderboard: From multi-choice to open-style questions for llms evaluation, benchmark, and arena. *CoRR*, abs/2406.07545.

Usman Naseem, Imran Razzak, Matloob Khushi, Peter W. Eklund, and Jinman Kim. 2021. Covidsenti: A large-scale benchmark twitter data set for COVID-19 sentiment analysis. *IEEE Transactions on Computational Social Systems*, 8(4):1003–1015.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

David A. Patterson, Joseph Gonzalez, Quoc V. Le, Chen Liang, Lluis-Miquel Munguia, Daniel Rothchild, David R. So, Maud Texier, and Jeff Dean. 2021. Carbon emissions and large neural network training. *CoRR*, abs/2104.10350.

Yuchen Peng, Ke Chen, Lidan Shou, Dawei Jiang, and Gang Chen. 2023. AQUA: automatic collaborative query processing in analytical database. *Proc. VLDB Endow.*, 16(12):4006–4009.

Nghiem Thanh Pham, Tung Kieu, Duc-Manh Nguyen, Son Ha Xuan, Nghia Duong-Trung, and Danh Le-Phuoc. 2025. Slm-bench: A comprehensive benchmark of small language models on environmental impacts–extended version. *Preprint*, arXiv:2508.15478.

Soham Poddar, Paramita Koley, Janardan Misra, Niloy Ganguly, and Saptarshi Ghosh. 2025. Towards sustainable NLP: insights from benchmarking inference energy in large language models. *CoRR*, abs/2502.05610.

Sumanth Prabhu, Aditya Kiran Brahma, and Hemant Misra. 2022. Customer support chat intent classification using weak supervision and data augmentation. In *Proceedings of the Joint International Conference on Data Science & Management of Data (CODS-COMAD)*, pages 144–152.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. Winogrande: An adversarial winograd schema challenge at scale. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 8732–8740.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108.

Or Sharir, Barak Peleg, and Yoav Shoham. 2020. The cost of training NLP models: A concise overview. *CoRR*, abs/2004.08900.

Dong Shu and Mengnan Du. 2024. Comparative analysis of demonstration selection algorithms for LLM in-context learning. *CoRR*, abs/2410.23099.

Aditi Singh, Nirmal Prakashbhai Patel, Abul Ehtesham, Saket Kumar, and Tala Talaei Khoei. 2024. A survey of sustainability in large language models: Applications, economics, and challenges. *CoRR*, abs/2412.04782.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*, pages 3645–3650.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2020. Energy and policy considerations for modern deep learning research. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 13693–13696.

Mingjie Sun, Zhuang Liu, Anna Bair, and J. Zico Kolter. 2024. A simple and effective pruning approach for large language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Adrian Theuma and Ehsan Shareghi. 2024. Equipping language models with tool use capability for tabular data analysis in finance. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 90–103.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

David Vilar, Markus Freitag, Colin Cherry, Jiaming Luo, Viresh Ratnakar, and George F. Foster. 2023. Prompting palm for translation: Assessing strategies and performance. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 15406–15427.

Lulu Wan, George Papageorgiou, Michael Seddon, and Mirko Bernardoni. 2019. Long-length legal document classification. *CoRR*, abs/1912.06905.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019a. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 3261–3275.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019b. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the International Conference on Learning Representations, (ICLR)*.

Xiaorong Wang, Clara Na, Emma Strubell, Sorelle Friedler, and Sasha Luccioni. 2023. Energy and carbon considerations of fine-tuning BERT. In *Findings of the Association for Computational Linguistics (EMNLP)*, pages 9058–9069.

Yuhui Xu, Lingxi Xie, Xiaotao Gu, Xin Chen, Heng Chang, Hengheng Zhang, Zhengsu Chen, Xiaopeng Zhang, and Qi Tian. 2024. Qa-lora: Quantization-aware low-rank adaptation of large language models.

In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Kailai Yang, Tianlin Zhang, Ziyan Kuang, Qianqian Xie, Jimin Huang, and Sophia Ananiadou. 2024a. MentaLLaMA: Interpretable mental health analysis on social media with large language models. In *Proceedings of the ACM on Web Conference (WWW)*, pages 4489–4500.

Sohee Yang, Elena Gribovskaya, Nora Kassner, Mor Geva, and Sebastian Riedel. 2024b. Do large language models latently perform multi-hop reasoning? In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 10210–10229.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*, pages 4791–4800.

Lihui Zhang and Ruifan Li. 2023. Knowledge prompting with contrastive learning for unsupervised commonsenseqa. In *Neural Information Processing - Proceedings of the International Conference on Neural Information Processing (ICONIP)*, pages 27–38.

Songming Zhang, Xue Zhang, Zengkui Sun, Yufeng Chen, and Jinan Xu. 2024a. Dual-space knowledge distillation for large language models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 18164–18181.

Yingtao Zhang, Haoli Bai, Haokun Lin, Jialin Zhao, Lu Hou, and Carlo Vittorio Cannistraci. 2024b. Plug-and-play: An efficient post-training pruning method for large language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Junhao Zheng, Qianli Ma, Shengjie Qiu, Yue Wu, Peitian Ma, Junlong Liu, Huawen Feng, Xichen Shang, and Haibin Chen. 2023. Preserving commonsense knowledge from pre-trained language models via causal inference. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 9155–9173.

Appendix

A Datasets Examples

Figures A1–A23 provide representative examples from the benchmark datasets used in this study. We randomly sample two instances from each dataset, showing the variety of question formats, input structures, and answer types present in these benchmarks. These examples provide insight into the challenges posed by each dataset and highlight the differences in their underlying tasks.

Example 1

Question: Can dogs see color?
Passage: Dogs can see colors, but not as vividly as humans. Their color vision is similar to that of a person with red-green color blindness. They primarily see shades of blue and yellow.
Answer: true

Example 2

Question: Is Mount Everest the tallest mountain in the world?
Passage: Mount Everest, located in the Himalayas, is the highest mountain above sea level, with a peak reaching 8,848.86 meters. However, if measuring from base to peak, Mauna Kea in Hawaii is technically the tallest mountain.
Answer: true

Figure A1: Examples from **BoolQ**

Example 1

Question: Which planet is known as the Red Planet?
Options: [Mercury, Venus, Earth, Mars]
Answer: Mars

Example 2

Question: What do plants absorb from the soil?
Options: [Oxygen, Nutrients, Carbon dioxide, Sunlight]
Answer: Nutrients

Figure A2: Examples from **ARC-Easy**

Example 1

Question: Why do stars appear to twinkle in the night sky?
Options: [Because stars themselves are flickering, Due to the movement of air in the Earth's atmosphere, Because of their distance from Earth, Due to changes in their brightness]
Answer: Due to the movement of air in the Earth's atmosphere

Example 2

Question: Which of the following substances is a poor conductor of electricity?
Options: [Silver, Copper, Rubber, Iron]
Answer: Rubber

Figure A3: Examples from **ARC-Challenge**

Example 1

Question: What is the primary function of the root in a plant?
Options: [To store sunlight for later use, To absorb water and nutrients, To produce oxygen, To perform photosynthesis]
Answer: To absorb water and nutrients

Example 2

Question: Why do objects fall to the ground when dropped?
Options: [Because of magnetism, Due to the force of gravity, Because the air pushes them down, Due to Earth's rotation]
Answer: Due to the force of gravity

Figure A4: Examples from **OpenBookQA**

Example 1

Question: What is the best way to move a heavy box across a smooth floor?
Options: [Push it while keeping it flat on the ground, Lift it and carry it, Drag it using a rope tied to the bottom, Kick it repeatedly until it moves]
Answer: Push it while keeping it flat on the ground

Example 2

Question: How can you efficiently dry wet clothes indoors?
Options: [Lay them flat on a table, Hang them near a heat source or a fan, Roll them into a ball and leave them overnight, Put them in a plastic bag]
Answer: Hang them near a heat source or a fan

Figure A5: Examples from **PIQA**

Example 1

Context: A person is slicing a tomato on a cutting board.
Ending options: [The person places the tomato slices on a plate., The person throws the tomato slices into the trash.]
Correct ending: The person places the tomato slices on a plate.

Example 2

Context: A man is playing a guitar on stage during a concert.
Ending options: [The audience claps and cheers after his performance., The man stops playing and leaves the stage silently.]
Correct ending: The audience claps and cheers after his performance

Figure A6: Examples from **Hellaswag**

Example 1

Sentence: The trophy doesn't fit into the brown suitcase because it is too large.
Question: What is too large? **Options**: [trophy, suitcase]
Answer: trophy

Example 2

Sentence: The cat chased the mouse because it was hungry.
Question: What was hungry? **Options**: [cat, mouse]
Answer: cat

Figure A7: Examples from **WinoGrande**

Example 1

Question: Where would you find a chandelier?
Options: [ceiling, floor, wall, table, window]
Answer: ceiling

Example 2

Question: What do people use to keep their hair dry while swimming?
Options: [swim cap, goggles, flippers, towel, sunscreen]
Answer: swim cap

Figure A8: Examples from **CommonsenseQA**

Example 1

Question: If a train travels at a speed of 60 miles per hour, how long will it take to travel 180 miles?
Answer: 3 hours

Example 2

Question: A bookstore sold 120 books in a week. If they sold 30 books each day from Monday to Thursday, how many books did they sell on Friday?
Answer: 0 books

Figure A9: Examples from **GSM8k**

Example 1

Question: If a car's fuel efficiency is 25 miles per gallon and the car has a 15-gallon tank, how far can the car travel on a full tank?
Options: [300 miles, 350 miles, 375 miles, 400 miles]
Answer: 375 miles

Example 2

Question: A rectangle has a length of 10 cm and a width of 5 cm. What is the area of the rectangle?
Options: [15 cm^2, 30 cm^2, 50 cm^2, 100 cm^2]
Answer: 50 cm^2

Figure A10: Examples from **AQuA**

Example 1

Article: There is not enough oil in the world now. As time goes by, it becomes less and less, so what are we going to do when it runs out? Perhaps we will have to go back to using horses, carriages and bicycles.
Question: According to the passage, which of the following statements is TRUE?
Options: [There is more petroleum than we can use now., Trees are needed for some other things besides making gas., We got electricity from ocean tides in the old days., Gas wasn't used to run cars in the Second World War.]
Answer: Gas wasn't used to run cars in the Second World War.

Example 2

Article: Schoolgirls have been wearing such short skirts at Paget High School in Branston that they've been ordered to wear trousers instead.
Question: The girls at Paget High School are not allowed to wear skirts because.
Options: [short skirts give people the impression of sexualisation, short skirts are too expensive for parents to afford, the headmaster doesn't like girls wearing short skirts, the girls wearing short skirts will be at the risk of being laughed at]
Answer: short skirts give people the impression of sexualisation.

Figure A11: Examples from **RACE-Middle**

Example 1

Article: Next eliminate most of their planets; they are either too far from or too close to their suns. Then eliminate all those planets which are not the same size and weight as the earth. Finally, remember that the proper conditions do not necessarily mean that life actually does exist on a planet. It may not have begun yet, or it may have already died out. This process of elimination seems to leave very few planets on which earthlike life might be found.
Question: What is the main idea of the passage?
Options: [The universe is full of planets., Life is rare in the universe., Earth is unique in the universe., The conditions for life are complex.]
Answer: Life is rare in the universe.

Example 2

Article: Some kids listen to music, watch TV or use the phone while doing their homework. 'It's important to make sure that you can stop and concentrate on one thing deeply,' says Rideout. With new and exciting devices hitting stores every year, keeping technology use in check is more important than ever. 'Kids should try,' adds Rideout. 'But parents might have to step in sometimes.'
Question: Which of the following is an example of multitasking according to the passage?
Options: [Watching TV while using the computer., Talking on the phone while staying with others., Playing video games on the Internet., Listening to music while relaxing.]
Answer: Watching TV while using the computer.

Figure A12: Examples from **RACE-High**

Example 1

Passage: Once upon a time, there was a little girl named Goldilocks. She went for a walk in the forest. Pretty soon, she came upon a house. She knocked and, when no one answered, she walked right in.
Turns: [*Question*: What was the girl's name? *Answer*: Goldilocks , *Question*: Where did she go for a walk? *Answer*: In the forest]

Example 2

Passage: John took his dog Max to the vet because Max was not feeling well. The vet examined Max and gave him some medicine.
Turns: [*Question*: Why did John take Max to the vet? *Answer*: Because Max was not feeling well , *Question*: What did the vet do? *Answer*: Examined Max and gave him some medicine]

Figure A13: Examples from **CoQA**

Example 1

Meaning representation: [*name*: The Eagle, *eat_type*: restaurant, *food*: French, *price_range*: high, *customer_rating*: 5 out of 5, *near*: Riverside]
Human reference: The Eagle is a highly rated French restaurant near Riverside with a high price range.

Example 2

Meaning representation: [*name*: The Golden Curry, *eat_type*: pub, *food*: Indian, *price_range*: low, *customer_rating*: 3 out of 5, *family_friendly*: yes]
Human reference: The Golden Curry is a family-friendly Indian pub with a low price range and a customer rating of 3 out of 5.

Figure A14: Examples from **e2e_nlg**

Example 1

Meaning representation: [*name*: Super Mario World, *release_year*: 1990, *esrb*: E (for Everyone), *rating*: excellent, *genres*: [platformer], *player_perspective*: [side view], *has_multiplayer*: yes, *platforms*: [Nintendo]]
Human reference: Super Mario World is an excellent platformer game that is played in the side view. It was released in 1990 on Nintendo and is rated E (for Everyone). It can be played multiplayer.

Example 2

Meaning representation: [*name*: The Elder Scrolls V: Skyrim, *release_year*: 2011, *esrb*: M (for Mature), *genres*: [adventure, role-playing]]
Human reference: The Elder Scrolls V: Skyrim is an adventure RPG that came out in 2011. It's rated M (for Mature).

Figure A15: Examples from **viggo**

Example 1

Question: What is the capital of France?
Sentence: Paris is the capital and most populous city of France.
Label: entailment

Example 2

Question: Who wrote 'Pride and Prejudice'?
Sentence: 'Pride and Prejudice' is a novel by Jane Austen.
Label: entailment

Figure A16: Examples from **glue_qnli**

Example 1

Text: Aspirin is commonly used to reduce fever.
Entities: [*entity*: Aspirin, *type*: Chemical, *start*: 0, *end*: 7]

Example 2

Text: Hypertension can lead to serious health complications.
Entities: [*entity*: Hypertension, *type*: Disease, *start*: 0, *end*: 11]

Figure A17: Examples from **bc5cdr**

Example 1

Sentence: Barack Obama was born in Hawaii.
Entities: [[*entity*: Barack Obama, *type*: PERSON, *start*: 0, *end*: 12], [*entity*: Hawaii, *type*: LOCATION, *start*: 25, *end*: 31]]

Example 2

Sentence: Apple Inc. is headquartered in Cupertino, California.
Entities: [[*entity*: Apple Inc., *type*: ORGANIZATION, *start*: 0, *end*: 9], [*entity*: Cupertino, *type*: LOCATION, *start*: 29, *end*: 38], [*entity*: California, *type*: LOCATION, *start*: 40, *end*: 50]]

Figure A18: Examples from **conllpp**

Example 1

Query: How can I reset my password?
Category: Account Management

Example 2

Query: What is the status of my order #12345?
Category: Order Tracking

Figure A19: Examples from **customer_support**

Example 1

Text: The defendant was found guilty of theft under Section 378 of the Penal Code.
Category: Criminal Law

Example 2

Text: The contract was terminated due to a breach of the confidentiality clause.
Category: Contract Law

Figure A20: Examples from **legal**

Example 1

Headline: Oil prices rise as OPEC agrees to cut production.
Topics: [Economy, Energy]

Example 2

Headline: Tech stocks rally amid strong quarterly earnings.
Topics: [Technology, Markets]

Figure A21: Examples from **reuters**

Example 1

Text: The new vaccine has shown promising results in early trials.
Sentiment: positive

Example 2

Text: Lockdown measures have been extended due to rising case numbers.
Sentiment: negative

Figure A22: Examples from **covid**

Example 1

Passage: In 1995, the population of the city was 1.2 million. By 2005, it had increased to 1.5 million.
Question: What was the increase in population from 1995 to 2005?
Answer: 300,000

Example 2

Passage: The team scored 24 points in the first half and 30 points in the second half.
Question: What was the total score for the team?
Answer: 54 points

Figure A23: Examples from **drop**

B Datasets Details

Table A1 presents a comprehensive overview of the benchmark datasets used in this study, including their key characteristics such as dataset size, dataset domains, venues, and task objectives. This detailed comparison helps contextualize their relevance and challenges in the broader scope of our analysis

C Models

Table A2 provides detailed information on SLMs in this study, including their providers, licenses, parameter sizes, model sizes, training time, and training performance. We observe that `Llama-3.2-1B`, `Phi-3-3.8B`, and `Gemma-3-1B` possess extended context lengths and were pre-trained on more extensive corpora, which appears to strongly influence their correctness.

D Metrics Details

Table A3 provides detailed explanations of the metrics used in our evaluation framework. We categorize our metrics into three main groups: correctness evaluation, computation evaluation, and consumption evaluation.

E Expanded Analysis of Metric Contributions to Medals

For each model, gold medals indicate the best performance using specific metrics. Some models show dominance in a single metric, suggesting specialization, while others maintain a balanced distribution of Gold medals across multiple metrics, indicating versatility. Models with more evenly distributed Gold, Silver, and Bronze medals across different criteria tend to be more well-rounded, offering strong performance without extreme specialization.

The Silver and Bronze medal distributions further refine this understanding. A model that consistently earns Silver in multiple metrics suggests strong overall performance but is slightly outperformed by one or more competitors in specific cases. Conversely, a model with many Bronze medals may still be efficient but is frequently ranked below two other models. This suggests that while it is competitive, it does not lead to any single criterion.

Among the expanded selection of models, differences in trade-offs become more apparent. Some models score highly in cost efficiency but rank lower in computational complexity, indicating that they are affordable but require significant processing resources. Others achieve substantial results in energy efficiency while being slightly less cost-effective, making them preferable in energy-conscious environments. The expanded dataset also provides insights into the impact of computational complexity, where models that rank lower in this criterion might still perform well in other categories, balancing out the trade-offs.

By examining a more extensive set of models, this analysis helps refine decision-making for different use cases. Whether the goal is to prioritize training speed, minimize cost, reduce energy consumption, or optimize computational efficiency, this breakdown provides clear guidance on which models align best with specific operational priorities. The broader dataset ensures that model selection is not limited to a few top-performing options. Instead, it considers a wider range of competitive alternatives, each offering different strengths based on their medal distributions.

F Inference Cost

After deploying SLMs, inference is performed orders of magnitude more frequently than fine-tuning, often by many users concurrently. As a result, even small differences in inference efficiency can accumulate into significant computational and energy costs—often exceeding those of fine-tuning. To quantify this, we conducted experiments to measure the computation time and energy consumption required to generate 1,000 tokens during the inference phase. The results, presented in Table A4, were obtained using an NVIDIA L4 GPU, as described in Section 4.1. Among the evaluated models, `Phi-1.5B` demonstrates the best performance in both computation efficiency and energy consumption.

No	Dataset	#Samples	#Tokens	Domain	Task	Venues	Description
1	**BoolQ**	15,432	3M	Common Sense	Question Answering	NAACL 2019	A yes/no question-answering dataset where each question requires reasoning over a short passage (Clark et al., 2019).
2	**ARC-Easy**	5,876	1.8M	Common Sense	Question Answering	arXiv	A subset of the AI2 Reasoning Challenge with straightforward grade-school science questions (Clark et al., 2018).
3	**ARC-Challenge**	2,590	1.5M	Common Sense	Question Answering	arXiv	A more difficult subset of ARC with questions requiring reasoning and external knowledge (Clark et al., 2018).
4	**OpenBookQA**	5,957	1M	Common Sense	Question Answering	EMNLP 2018	A multiple-choice question-answering dataset that requires knowledge from a small "open book" of science facts (Mihaylov et al., 2018).
5	**PIQA**	16,113	1.6M	Physics	Reasoning	AAAI 2020	A dataset for physical commonsense reasoning, testing knowledge of how objects are used in everyday situations (Bisk et al., 2020).
6	**Hellaswag**	10,421	10M	Common Sense	Reasoning	ACL 2019	A dataset for commonsense reasoning and next-sentence prediction with adversarially mined, plausible distractors (Zellers et al., 2019).
7	**WinoGrande**	44,321	5M	Common Sense	Reasoning	AAAI 2020	A large-scale dataset for pronoun resolution tasks, inspired by the Winograd Schema Challenge (Sakaguchi et al., 2020).
8	**CommonsenseQA**	12,102	1.8M	Common Sense	Reasoning	ICONIP 2023	A multiple-choice dataset testing broad commonsense knowledge using questions based on ConceptNet (Zhang and Li, 2023).
9	**GSM8k**	8,034	2M	Math	Problem Solving	arXiv	A dataset of grade-school math word problems designed to evaluate problem-solving abilities (Liu et al., 2023).
10	**AQuA**	99,765	3M	Math	Problem Solving	VLDB 2023	A dataset with multiple-choice questions Algebraic Word Problems requiring reasoning over numeric expressions (Peng et al., 2023).
11	**RACE-Middle**	24,798	7M	Education	Reading Comprehension	EMNLP 2017	A subset of the RACE dataset with English reading comprehension questions from middle school exams (Lai et al., 2017).
12	**RACE-High**	26,982	10M	Education	Reading Comprehension	EMNLP 2017	A subset of RACE with more challenging reading comprehension questions from high school exams (Lai et al., 2017).
13	**CoQA**	127,542	5M	Common Sense	Question Answering	LREC 2022	A conversational question-answering dataset where each question depends on the context of prior dialogue (Brabant et al., 2022).
14	**e2e_nlg**	50,321	600K	Food & Beverage	Text Generation	INLG 2018	A dataset for end-to-end natural language generation for restaurant-related dialogue (Dusek et al., 2018).
15	**viggo**	9,842	500K	Video Games	Text Generation	INLG 2019	A dataset for video game-related natural language generation, mapping structured input into natural language descriptions (Juraska et al., 2019).
16	**glue_qnli**	104,543	6M	Linguistics	Question Answering	arXiv	A question-answering dataset reformulated as a binary classification task for sentence pair entailment (Shu and Du, 2024).
17	**bc5cdr**	20,764	5M	Chemistry	Recognition	ICIMMI 2023	A biomedical dataset for disease and chemical entity recognition, derived from PubMed articles (Gupta et al., 2023).
18	**conllpp**	23,499	1.2M	Linguistics	Recognition	EMNLP 2023	An enhanced version of the CoNLL-2003 dataset for named entity recognition (Al-Shaibani and Ahmad, 2023).
19	**customer_support**	14,872	300K	Customer Behaviors	Classification	CODS 2022	A dataset of customer support interactions for intent classification and response generation (Prabhu et al., 2022).
20	**legal**	49,756	5M	Legal	Classification	arXiv	A legal domain dataset for natural language processing tasks such as legal document classification and contract analysis (Wan et al., 2019).
21	**reuters**	9,623	2M	News	Topic Extraction	JMLR 2004	A classic dataset for text classification, often used in topic modeling and news categorization (Lewis et al., 2004).
22	**covid**	19,874	3M	Healthcare	Sentiment Analysis	TCSS 2021	A dataset containing COVID-19-related texts from social media posts, typically used for sentiment analysis (Naseem et al., 2021).
23	**drop**	96,567	4M	Common Sense	Reasoning	NAACL 2019	A reading comprehension dataset requiring discrete reasoning over paragraphs, such as arithmetic operations (Dua et al., 2019).

Table A1: Dataset Details

No	Model	Provider	License	#Params (B)	Context Length	Training Time	Size (GB)	Throughput (Tokens/s)	Latency (ms)
1	GPT-Neo-1.3B	EleutherAI	Apache 2.0	1.37	2,048	10 days (32 GPUs)	2.46	1,500	50
2	Dolly-v2-3B	DataBricks	Apache 2.0	3.00	2,048	10 days (32 GPUs)	5.8	1,250	55
3	Pythia-2.8B	EleutherAI	Apache 2.0	2.80	2,048	12 days (32 GPUs)	5.5	1,350	50
4	LLaMA-2-7B	Meta	LLAMA 2 L	6.47	4,096	21 days (64 GPUs)	13.0	1,200	62
5	TinyLlama-1.1B	Hugging Face	Apache 2.0	1.10	2,048	8 days (16 GPUs)	2.0	1,600	45
6	Mistral-7B	Mistral AI	Apache 2.0	7.00	8,192	15 days (128 GPUs)	13.0	1,400	55
7	Zephyr-7B	Hugging Face	Apache 2.0	7.00	8,192	20 days (64 GPUs)	13.74	1,300	52
8	ShearedLlama-2.7B	Hugging Face	Apache 2.0	2.70	2,048	12 days (32 GPUs)	5.0	1,300	54
9	Gemma-2B	Google	Proprietary	2.00	2,048	14 days (32 GPUs)	4.67	1,450	54
10	Phi-1.5B	Microsoft	Proprietary	2.70	2,048	12 days (32 GPUs)	2.45	1,400	52
11	StableLM-3B	Stability AI	Apache 2.0	3.00	2,048	14 days (64 GPUs)	6.5	1,250	50
12	Open-LLaMA-3B	OpenLM	Apache 2.0	3.00	4,096	18 days (64 GPUs)	6.8	1,300	60
13	Llama-3.2-1B	Meta	Llama 3.2	1.24	128,000	30 days (512 GPUs)	2.47	1,350	55
14	Phi-3-3.8B	Microsoft	MIT	3.82	128,000	7 days (512 GPUs)	2.2	1,300	55
15	Gemma-3-1B	Google	Proprietary	1.00	32,000	Unknown	2	1,400	50

Table A2: SLMs' Details (sort by release time)

No	Metric	Evaluation	Task	Datasets	Description
1	*Accuracy*	Correctness	Question Answering Classification Recognition Reasoning Problem Solving Reading Comprehension	All except [e2e_nlg, viggo, reuters]	This metric is used for tasks where a model's correctness can be measured by its ability to predict a correct answer, such as in Question Answering (QA) and Classification tasks.
2	*F1 Score*	Correctness	Question Answering Classification Recognition Reasoning Problem Solving Reading Comprehension	All except [e2e_nlg, viggo, reuters]	A balanced metric that combines precision and recall, ideal for tasks like Named Entity Recognition (NER) and toxicity detection where both false positives and false negatives are crucial.
3	*BLEU*	Correctness	Text Generation Topic Extraction	e2e_nlg, viggo, reuters	Used primarily for evaluating the quality of text generation tasks, such as translation and natural language generation, by comparing generated outputs against reference texts.
4	*ROUGE*	Correctness	Text Generation Topic Extraction	e2e_nlg, viggo, reuters	Commonly used for summarization tasks, measuring the overlap of n-grams between the generated text and reference summaries.
5	*METEOR*	Correctness	Text Generation Topic Extraction	e2e_nlg, viggo, reuters	Evaluates generated text quality with consideration for synonyms and paraphrasing based on the harmonic mean of unigram precision and recall, with recall weighted higher than precision.
6	*Perplexity*	Correctness	Text Generation Topic Extraction	e2e_nlg, viggo, reuters	It is a measurement of uncertainty in the predictions of a language model. In simpler terms, it indicates how surprised a model is by the actual outcomes.
7	*Runtime*	Computation	All	All	Measures computational efficiency in terms of running time.
8	*FLOP*	Computation	All	All	Measures computational efficiency in terms of number of computation per second.
9	*Cost*	Consumption	All	All	The total cost of fine-tuning, calculated in USD.
10	CO_2	Consumption	All	All	An estimate of the environmental impact in terms of CO_2 emission of model fine-tuning.
11	*Energy*	Consumption	All	All	The total energy consumed during the training process, measured in kilowatt-hours (kWh)

Table A3: Metric Details

Model	Cost (USD)	Energy (kWh)	CO2 Emissions (kg)	FLOP	Runtime (h)
GPT-Neo-1.3B	0.2237	0.0186	0.011	421,420	0.1417
Phi-1.5B	0.1632	0.0136	0.008	1,780,810	0.1033
Open-LLaMA-3B	0.3158	0.0263	0.0155	1,850,000	0.2
LLaMA-2-7B	0.2434	0.0203	0.0119	1,890,000	0.154
Mistral-7B	0.4211	0.0351	0.0206	5,407,500	0.2667
Zephyr-7B	0.3158	0.0263	0.0155	3,097,500	0.2
TinyLlama-1.1B	0.2763	0.023	0.0135	636,170	0.175
StableLM-3B	0.2368	0.0197	0.0116	855,000	0.15
ShearedLlama-2.7B	0.3684	0.0307	0.0181	1,955,250	0.2333
Dolly-v2-3B	0.2171	0.0181	0.0106	740,000	0.1375
Pythia-2.8B	0.2632	0.0219	0.0129	833,000	0.1667
Gemma-2B	0.3421	0.0285	0.0168	1,435,000	0.2167
Llama-3.2-1B	0.4342	0.0362	0.0213	7,083,330	0.275
Phi-3-3.8B	0.4079	0.034	0.02	6,000,000	0.2583
Gemma-3-1B	0.4474	0.0373	0.0219	5,666,670	0.2833

Table A4: SLMs' Inference Cost