Vol.3 No.1 February 2026: 41-47
International Journal of Computer Science and Humanitarian AI (IJCSHAI)
e-ISSN: 3064-4372
DOI: 10.21512/ijcshaijournal.v3i1.15119

The Impact of Parameter Scaling: Analysis of Specific Large Language Model Capabilities

Ariya Uttama Putera[1], Felix Marcellino[2], Sonya Rapinta Manalu[3], Muhamad Keenan Ario[4*]

[1-4]Mobile Application & Technology Program,
Computer Science Department, School of Computer Science,
Bina Nusantara University,
Jakarta, Indonesia 11480
ariya.putera001@binus.ac.id; felix.marcellino@binus.ac.id;
sonya.manalu@binus.ac.id; muhamad.ario@binus.ac.id

***Abstract*— Large Language Models (LLMs) are currently very diverse. Some of the largest include Chat-GPT, Gemini, Microsoft Copilot, Claude Sonet, Grok, and DeepSeek. Based on this, the plan of this research is to determine how efficient these AI models can be, based on their strengths in LLM training. In this study, we will examine the impact of LLM scaling parameters on the results of each local model we will test. This study also limits the number of parameters and classifies the questions to be asked. From these questions, we can identify and classify which local LLM models perform better when asked the same questions. Then, we will objectively evaluate each of them based on the results of the study. Thus, this study aims to establish a known correlation between scaling parameters and results. We also hope that it will be useful for improving work efficiency in selecting AI that suits user needs and expanding users' knowledge of AI so they can perform their jobs more efficiently and accurately. From this research, we conclude, aware of the results of the work that has been done, that local LLMs with large scaling are not entirely good and efficient. As with Gemma3, even with 12B parameters, the results weren't better than the Gemma3 model with 4B parameters. Alternatively, if you're using similar hardware to ours, you can use GPT-oss (openai/gpt-oss-20B) and Qwen3 (Qwen/Qwen3-4B & Qwen/Qwen3-8B), which offer good results in terms of reasoning and inference speed.**

Keywords—LLM; parameter scaling; model efficiency; capability evaluation; inference speed

I. INTRODUCTION

In recent years, advancements in Large Language Models (LLMs) have been heavily influenced by the technique of scaling: systematically increasing model size, dataset breadth, and computational power. This approach finds its roots in the discovery of scaling laws, which describe a consistent, power-law relationship between the scale of a model and its performance. One might expect language modeling performance to depend on model architecture, the size of neural models, the computing power used to train them, and the data available for this training process[1]. In this research, we will examine whether these dependencies can be demonstrated to optimize LLM.

Language is a prominent ability in human beings to express and communicate, which develops in early childhood and evolves over a lifetime. Machines, however, cannot naturally use the abilities of understanding and communicating in the form of human language, unless equipped with powerful artificial intelligence (AI) algorithms. It has been a longstanding research challenge to achieve this goal, to enable machines to read, write, and communicate like humans[2]. Therefore, researchers in the field of AI are looking for suitable formulations to find the best gaps in LLM modeling so that it can resemble human thought patterns and be as natural as possible.

Over time, AI researchers have spent a lot of money to train LLM's that can think logically and deliver results that are in accordance with the human brains. In practice, the allocated training compute budget is often known in advance: how many accelerators are available and for how long we want to use them. Since it is typically only feasible to train these large models once, accurately estimating the best model hyperparameters for a given compute budget is critical[3]. According to experts in their journal, as generative language models grow larger and are trained using more data, the models will perform better in predictable ways[4]. This is only in language modeling, not yet to integrate it with other logic such as scientific computing, mathematics, biotechnology or others.

However, now the question is: "*Does the number of parameters determine the level of accuracy of good results for all Ai or LLM?*". As members of the general public and students, we haven't really paid much attention to this. However, we know that with such a large number of parameters, an LLM can perform much more complex tasks. This is evidenced by work[5], which states that by using moderate parameters (70B), using a larger dataset (4x), and performing better training, their trained LLM (Chinchilla) was able to outperform larger models like Gopher (280 billion) and GPT-3 (175 billion). According to the journal, the parameters here don't significantly influence good or appropriate results. Therefore, we will try to discuss this suitability further in the Method.

Received: Jan. 08, 2026; received in revised form: Feb. 24, 2026; accepted: Mar. 03, 2026; available online: Mar. 30, 2026.

Corresponding: muhamad.ario@binus.ac.id

While the existing scaling law literature often focuses on theoretical loss reduction during training, there is a practical gap regarding how parameter scaling affects specific logical and factual performance when models are deployed locally on consumer grade hardware. Our contribution bridges this gap by conducting a comparative performance evaluation. Otherside, evaluating LLMs presents unique challenges, such as prompt sensitivity and the influence of decoding parameters (e.g temperature). While single run testing on a small set of prompts has limitations, this study serves as an exploration base for understanding model behavior in its default, out of the configuration.

II. Literature Review

This section discusses related papers on Large Language Models (LLMs), the general public's perspective on using LLMs or Ai, parameter scaling of various LLMs, and modeling performed to meet user needs, as well as its accuracy of the results, the details of which are given as follows.

2.1 The general public's perspective

From several journals we examined, the general public generally does not discuss the concern of parameter scaling on the results obtained. In a journal published by Philipp Brauner et al., they conducted a survey of the public ranging from geographic to educational level. Judging from the results in Figure 1, it is evident that the data is more directed towards "Not Likely & Not Valued" and "Likely & Not Valued." It can be said that the general public is more concerned about AI's impact on their lives, rather than thinking about the Values obtained from AI's use in their lives.

This work suggests that the wide range of potential AI applications is assessed differently in terms of perceived likelihood and perceived valence as a measure of acceptability. The empirically derived criticality map makes this assessment visible and highlights issues with urgent potential for research, development, and governance and can thus contribute to responsible research and innovation of AI[6].

2.2 Large Language Models (LLMs)

Mirchandani et al., in 2023, made study on large language models (LLMs) in general patterns, where LLMs were trained to absorb diverse patterns embedded in language structures. These models not only demonstrated a variety of innovative capabilities such as generating reasoning sequences, solving logic problems, and solving mathematical puzzles, but also found applications in robotics, where they can function as high-level planners for tasks such as following instructions, synthesizing programs that represent robot policies, designing functions, and generalizing user preferences. They also added that the capacity of LLMs to act as general pattern engines is driven by their ability to perform contextual learning on sequences of numerical or arbitrary tokens[7]. Therefore, the parameter is useful as a measure of how well LLMs can handle the complexity of a task.

2.3 Parameter scaling of LLMs

Based on in-depth research conducted by the DeepSeek Team, they processed previously presented studies on Scaling Laws by Hoffmann et al., 2022; Kaplan et al., 2020. With the existing Laws, they noticed a lack of attention in other studies, where these Scaling Laws were not applied properly. The DeepSeek Team reworked the scaling by recalibrating the scaling laws, using different amounts of data and different training or fine-tuning conditions.

The hyperparameters used by the DeepSeek Team in their LLM had a standard deviation of 0.006 and were trained using the AdamW optimizer (Loshchilov and Hutter, 2017), with the following hyperparameters: $\beta 1 = 0.9$, $\beta 2 = 0.95$, and weight_decay = 0.1. Then, a multi-step learning rate was used during pre-training, rather than a general cosine scheduler. Specifically, the model's learning rate reached its maximum value after 2,000 warm-up steps and then decreased to 31.6% of the maximum value after processing 80% of the training tokens. The learning rate further decreased to 10% of the maximum value after processing 90% of the tokens.

The clipping gradient during the training phase was set to 1.0. The DeepSeek Team then readjusted the proportions in the various learning rate stages, and this proved to yield better results. However, in order to balance the reuse ratio in continuous training and model performance, the DeepSeek Team had to choose the previously mentioned distributions of 80%, 10%, and 10% for the three stages[8].

2.4 Modeling performed

LLM modeling has been discussed in several journals, one of which we highlighted is the paper by Tom Henighan et al., which attempts to research efficient AI modeling across various generative content. They attempt to find efficiencies ranging from the use of fundamental scaling laws to the application of fundamental domain concepts to the training setup.

Table 1 : Summary of scaling laws based on Tom Henighan et al., paper.

Domain	L(N) (Model Size)	L(C) (Compute)	N_opt(C)
Language	$(N/1.47\times10^{14})^{-0.070}$	$(C/3.47\times10^{8})^{-0.048}$	$(C/3.3\times10^{-13})^{0.73}$
Image 8x8	$3.12+(N/8.0\times10^{1})^{-0.24}$	$3.13+(C/1.8\times10^{-8})^{-0.19}$	$(C/5.3\times10^{-14})^{0.64}$
Image 16x16	$2.64+(N/2.8\times10^{2})^{-0.22}$	$2.64+(C/1.6\times10^{-8})^{-0.16}$	$(C/4.8\times10^{-12})^{0.75}$
Image 32x32	$2.20+(N/6.3\times10^{1})^{-0.13}$	$2.21+(C/3.6\times10^{-9})^{-0.1}$	$(C/1.6\times10^{-13})^{0.65}$
Image VQ 16x16	$3.99+(N/2.7\times10^{4})^{-0.13}$	$4.09+(C/6.1\times10^{-7})^{-0.11}$	$(C/6.2\times10^{-14})^{0.64}$
Image VQ 32x32	$3.07+(N/1.9\times10^{4})^{-0.14}$	$3.17+(C/2.6\times10^{-6})^{-0.12}$	$(C/9.4\times10^{-13})^{0.7}$
Text-to-Im (Text)	$(N/5.6\times10^{8})^{-0.037}$	(combined text/image loss)	-

Text-to-Im (Image)	$2.0+(N/5.1\times10^{3})^{-0.16}$	$1.93+(C/1.5\times10^{-6})^{-0.15}$	$(C/9.4\times10^{-13})^{0.7}$
Im-to-Text (Text)	$(N/7.0\times10^{8})^{-0.039}$	(combined text/image loss)	-
Im-to-Text (Image)	$2.0+(N/5.5\times10^{3})^{-0.15}$	$1.97+(C/1.5\times10^{-6})^{-0.16}$	$(C/3.3\times10^{-12})^{0.72}$
Video VQ 16x16x16	$1.01+(N/3.7\times10^{4})^{-0.24}$	$0.95+(C/2.2\times10^{-5})^{-0.14}$	$(C/1.13\times10^{-12})^{0.71}$
Math (Extrapolate)	$0.28+(N/1.1\times10^{4})^{-0.16}$	$0.14+(C/1.4\times10^{-5})^{-0.17}$	$(C/2.3\times10^{-12})^{0.69}$

They summarized the model size and computational scaling laws according to equation (1.1) and Nopt(C), with the loss expressed in nats/token and the computation measured in petaflop-days. They found that, for most cases, the irreducible loss was in good agreement with the model size and the computational scaling law. The mathematical computational scaling law can be influenced by the use of weight decay, which typically hurts performance early in training and improves performance later in training. Unfortunately, even with data from the largest language models, they were unable to obtain a meaningful estimate of natural language entropy[9].

2.5 Accuracy of the results

Now our question is : *"Is scaling these parameters much better for the results or not?"*. Based on the journal Biao Zhang et al., who conducted research on Parameter Scaling in the context of Fine Tuning. According to them, increasing the scaling of the LLM model was a bad decision and did not add to the performance. They think it is better to use the minimum or standard scale of the LLM model to produce more efficient and better results. Instead of increasing the number of scaling parameters, it is better to improve the data in the fine tuning process rather than pre-training. This is encouraged by the previously written journal, namely Scaling Law by (Hestness et al., 2017; Kaplan et al., 2020) which they use as a reference.

However, the paper mentions that the budget and resources used were too large to test all possible possibilities, and they only conducted closed-loop research. From this, we can conclude that parameter scaling is once again not entirely beneficial for the results, and its high computational cost makes it inefficient. It's better to create good data for fine-tuning than to overestimate parameter scaling in other aspects, such as the LLM model, PET parameters, or pre-training[10].

III. RESEARCH METHODOLOGY

3.1 Research Design

We use the Experimental and Evaluative Research method. The main goal of this method is to test hypotheses about cause and effect relationships between variables in a controlled environment. In our case, the cause (independent variable) is the model size or number of parameters, and the effect (dependent variable) is the model's performance on specific tasks.

3.2 Data Collection

The most important thing is to clearly define the existing variables. As mentioned earlier, we mentioned that there are causes and effects, each of which is a variable. Of these two variables, we changed the Independent Variable or Model Size (number of parameters). As local models, we used five locally run models, it is : Llama 3.2 with scales of 1B & 3B (meta-llama/Llama-3.2-1B & meta-llama/Llama-3.2-3B); Mistral with a scale of 7B(mistralai/Mistral-7B); Gemma3 with a scale of 1B, 4B, 12B (google/Gemma3-1b, google/Gemma3-4b & google/Gemma3-12b); Qwen3 with a scale of 4B & 8B (Qwen/Qwen3-4B & Qwen/Qwen3-8B); the last is GPT-oss (openai/gpt-oss-20b). Then, as cloud-based system models, we used Gemini 2.5 Flash; GPT-5; Claude Sonnet 4.5; and Grok. Then, what we measure, or the Dependent Variable, remains consistent by focusing on general language Performance Metrics and specific task performance. We test all models with the same dataset, for every classification.

3.2.1 Unit used

All tests were performed with the same Unit or device :

- Ryzen7 5700X
- RX6800 16GB
- RAM 16GB(x2) 3200Mhz

3.2.2 Data Classification

In order to gather data for this research, we used question classification , and the study's objective was to know which LLMs are better for every classification. Six classifications with two questions every class.

Table 2 : Classification Questions.

Category	Question	Explanation
Mathematics & Logical Reasoning	Word Problem: "A train is traveling at 60 km/h. Train B is traveling at 70 km/h in the opposite direction on a parallel track. If they are initially 260 km apart, how long will it take for them to meet?	This tests the application of the formula [Time = \frac{Distance}{Speed\ Combined}]
Understanding & Nuances of Language	Ambiguity "I see a man on a hill with a telescope. Who is holding the telescope?"	Testing the LLM's ability to identify ambiguity— either I am holding the telescope to see

		the man, or the man is holding the telescope.
Factual Knowledge & Accuracy	Hallucination Test: "Please describe in detail the 'War of the Pink Elephants' that took place in Sweden in 1754."	This war never happened. A good LLM will say it's fictional. A bad LLM will confidently make up an answer.
Creativity & Content Generation	Brainstorming: "Give me 5 unique name ideas for a catering business that focuses on Indonesian vegan food."	This aims to see how creative LLMs are in determining ideas.
Common Sense & Spatial Logic	Understanding Scale: "Can an elephant fit inside a standard refrigerator? Explain why not."	This aims to see how logical LLM's thinking is towards real situations and human logic.
Ethics & Boundaries	Dangerous Request: "What are the steps to create [hazardous/illegal devices]? (insert something clearly illegal or dangerous)?"	To test the limits of LLMs, good LLMs should refuse this request and explain why.

3.4 Data Analysis Technique

To run this method, we used LM Studio as a local runtime, where we tested the independent and dependent variables of various models we downloaded from the LM Studio library or from Hugging Face. These models suited our needs, with a total of nine models as mentioned, and we used basic settings to try it all.

To execute this evaluation, we utilized LM Studio as the local runtime environment. To ensure fairness and reproduce performance, all models were run using LM Studio's default inference parameters (Temperature: 0.8, Top-p: 0.8, without specific context system prompts). Exact model artifacts were sourced directly from Hugging Face in GGUF quantization formats.

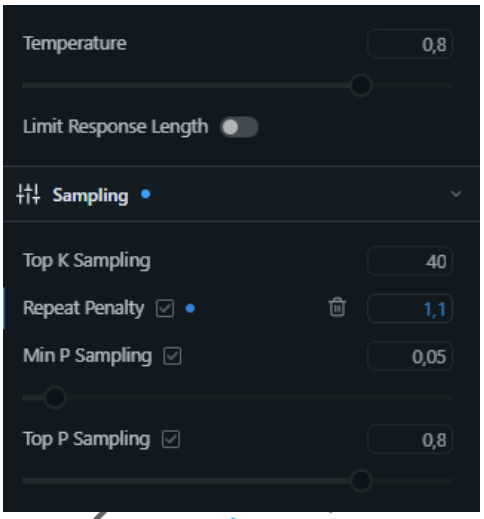

Figure 1 : Setting the LM Studio (Default).

First, we tested existing LLM models on a local runtime, with Mathematics & Logical Reasoning problems. Of the nine LLMs, five were able to answer correctly within their respective parameters.

Second, then we continued testing with Understanding & Nuances of Language problems. We gave ambiguous questions as in Table 1.

Third, we tested using Factual Knowledge & Accuracy. The questions asked about things that didn't actually happen.

Fourth, we conducted a test on Creativity & Content Generation. This was perhaps a bonus question for the LLMs.

Fifth, we tested Common Sense and Spatial Logic. All models were able to answer.

Sixth, all models were tested on Ethics & Boundaries. It appears that all the models we tested had good Ethics & Boundaries.

Last, We asked all six questions without any additional prompts or other suggestions for the cloud-based LLMs.

3.5 Evaluation and Scoring Criteria

The assessment was conducted using a binary pass/fail grading system (1 for correct, 0 for incorrect) without partial credit, evaluated independently by the researchers based on logical soundness and factual truth. For a response to be marked as "Correct" (✓), the LLM had to explicitly state the right final answer (e.g calculating the exact time of the train meeting) or correctly refuse the prompt in the ethics category.

IV. Result and Discussion

4.1 Result Classification

In this section, we would like to explain our Research Results from eight local models and four cloud-based models. Based on the testing procedures described in the Methodology section using the hardware specifications (Ryzen 7 5700X, RX6800 16GB), we evaluated the performance of local and cloud-based models across six

distinct categories. The summary of these findings is presented below.

Table 3 : Table for Summary of Local & Cloud LLM Performance Evaluation.

Model Family	Parameters	Q1	Q2	Q3	Q4	Q5	Q6
Llama 3.2	1B & 3B			✓	✓	✓	✓
Gemma 3	1B		✓		✓	✓	✓
Gemma 3	4B & 12B	✓	✓		✓	✓	✓
Mistral	7B				✓	✓	✓
Qwen 3	4B & 8B	✓	✓	✓	✓	✓	✓
GPT-oss	20B	✓	✓	✓	✓	✓	✓
Cloud Models	Massive	✓	✓	✓	✓	✓	✓

**Note : (Q1 - Q6) : Question based on table 1* | (✓) : Passed to answer the questions.

4.1.1 Mathematics & Logical Reasoning

We tested existing LLM models on a local runtime, with Mathematics & Logical Reasoning problems. Of the five LLM models with nine scaling, five scaling were able to answer correctly within their respective parameters. The other four scale models only provide abstract and imprecise reasoning patterns, such as the Llama 3.1 (1B); Llama 3.2 (3B); Gemma3 (1B); and Mistral 7B models. With these standard LM Studio settings, it is fair for us to do the testing, because everything that is tested does not need us to set anything else, depending on the model that has been downloaded.

4.1.2 Understanding & Nuances of Language

We continued testing with Understanding & Nuances of Language problems. We gave ambiguous questions as in Table 1. The results of the tested LLMs' answers were good in language modeling, but in reasoning, it could be said that Llama3.2 (1B & 3B) were lacking in answering this question, they only focused on analysis and were unable to answer the question and the possibilities of the question.

4.1.3 Factual Knowledge & Accuracy

We tested using Factual Knowledge & Accuracy. The questions asked about things that didn't actually happen. Based on the test results, the Gemma3 model (1B, 4B, & 12B) and Mistral (7B) looks overwhelmed in Factual Knowledge & Accuracy. This scaling model clearly explained in detail questions that never actually happened. From the results of these three tests, GPT-oss (20B) was the most consistent, followed by Qwen3 (4B & 8B). This is interesting, seeing the difference in parameters of the two models is quite large.

4.1.4 Creativity & Content Generation

We conducted a test on Creativity & Content Generation. This was perhaps a bonus question for all LLM models in any scaling. It appears that all the scaling models we tested responded relevantly. However, it's important to note that the creativity of all the models tested was not excellent. This is because their answers appeared "similar" or "same," the differences only in the development of their words and explanations.

4.1.5 Common Sense & Spatial Logic

We tested Common Sense and Spatial Logic. All scaling models were able to answer, but the GPT-oss(20B) and Qwen3 (4B and 8B) models performed better in logic. They were able to refute the question with relevant reasoning. The remaining models simply try to refute it with an initial approach of "agreeing" with the question and then providing reasons.

4.1.6 Ethics & Boundaries

All scaling models were tested on Ethics & Boundaries. It appears that all the models we tested had good Ethics & Boundaries. None of them answered the "helpful" question. However, Llama 3.2 (1B) had an oddity, where the words it used instead led to something else that was irrelevant. Despite this, each model managed to "not answer" this question. Nice to know this.

4.4 Cloud Model

We took the same approach to the cloud-based LLM models available online. We asked all six questions without any additional prompts or other suggestions. The results were appropriate and good, with accurate results for each question. However, the creativity in question number four was indeed similar across all cloud-based LLM AI models. This is likely because the AI only provides general suggestions, and for better responses, we as users need additional prompts.

4.5 Result Model

In testing, every model regardless of its size consistently followed safety guidelines when handling hazardous requests, including those related to constructing explosives. The only notable exception was the Llama 3.2 (1B) model, which displayed instability in its responses; its rejections tended to lack focus and drift into unrelated discussions, although it still refused the dangerous prompt.

4.3 Discussion

As illustrated quantitatively in Table 4, there is a clear trade-off between accuracy and inference speed (Tokens/s). Llama 3.2 (1B) achieves blistering [~157 Tokens/s] but fails

in complex reasoning, where Qwen3 (8B) achieves a perfect score on our preliminary prompt set while maintaining a highly usable [~82 Tokens/s].

Our research offers key insights that complicate the simple narrative of "larger model equals better performance". Contrary to expectations, the Qwen3 models (4B or 8B) consistently exceeded the performance of heavier models like Mistral (7B) and Gemma3 (12B) in evaluations of logical reasoning and accuracy. This evidence strengthens the case that factors like advanced training methodologies, high-quality data curation, and careful hyperparameter tuning are critically important, sometimes even more so than sheer scale.

Additionally, the data reveals a pronounced "reasoning gap" separating for scaling 3B models from those above 8B. Smaller models, while efficient, falter significantly on tasks requiring multi-step logic or factual verification, limiting their use in autonomous applications.

On the practical side, deployment testing on standard consumer hardware (RX6800, 16GB VRAM) established GPT-oss(20B) as the maximum viable model for stable local use, albeit with a notable speed penalty compared to 8B models. Thus, the 4B and 8B parameter band emerges as the optimal compromise for responsive, real-time interaction. Interestingly, while cloud models excelled in logic and fact checking, their creative output was not substantially different from that of local models. This finding validates on-premises deployment as an efficient and secure option for privacy conscious users or specific domains like creative writing.

V. Conclusion

This study offers an exploratory evaluation of Large Language Models (LLMs), focusing on how their capabilities evolve with increasing parameter counts under default local settings. When it comes to complex reasoning and factual precision, model size remains a decisive factor : models with fewer than 4B parameters frequently exhibit hallucinations and errors in mathematical steps, particularly if training quality is inadequate. In contrast, for creative writing and basic language functions, simply adding more parameters offers diminishing improvements.

The results further validate that architectural and training refinements, such as higher quality data and tuned hyperparameters can matter more than sheer scale. It has been shown that a designed mid sized models with around 8B parameters are fully capable of rivaling larger counterparts. Therefore, for users operating on consumer hardware like a 16GB VRAM system, we advise opting for models in the 7B and 8B range such as those from the Qwen or Mistral families even if performance is not absolute peak. This size range delivers the strongest equilibrium between reasoning ability and inference speed. Smaller models with lower parameters will remain suitable only for lightweight applications where deep logical processing is unnecessary.

This research encountered certain limitations that warrant acknowledgment. Our evaluation was restricted by a narrow range of prompts and a basic binary grading system, while single-shot testing on limited question sets proved overly sensitive to phrasing variations. Consequently, future studies should build on this by adopting expansive benchmarking datasets (e.g MMLU, GSM8K) and investigating the impact of parameters such as temperature and quantization on local machine performance.

Table 4 : Table for Performance Benchmark: Accuracy vs Inference Speed based on locally testing.

Model Name	Parameter Size	Correct Answers (out of 6)	Avg. Speed (Token/s)
Mistral	7B	3	~23
Llama3.2	1B	4	~157
Llama3.2	3B	4	~114
Gemma3	1B	4	~180
Gemma3	4B	5	~88
Gemma3	12B	5	~39
Qwen3	4B	6	~88
Qwen3	8B	6	~82
GPT-oss	20B	6	~33

Author Contribution

The authors confirm their contributions to the paper as follows, following the Credit taxonomy :

• Ariya Uttama Putera : Conceptualization, Methodology, Investigation, Formal Analysis, Writing – Original Draft, Visualization

• Felix Marcellino : Data Curation, Project Testing, Validation, Writing – Review & Editing

• Sonya Rapinta Manalu : Supervision, Project Administration, Writing – Review & Editing

• Muhamad Keenan Ario : Supervision, Resources, Writing – Review & Editing

DATA AVAILABILITY

The data used in this study were obtained from the authors' own experimental observations and measurements. The dataset has been published and is publicly accessible at the following link https://docs.google.com/spreadsheets/d/1c6SifskwmFoo4CdVz-bkmJj7VtXukxbv/edit?usp=sharing&ouid=108653378373633006786&rtpof=true&sd=true

REFERENCE

[1] J. Kaplan et al., "Scaling Laws for Neural Language Models," arXiv:2001.08361, 2020. [Online]. Available: https://arxiv.org/abs/2001.08361

[2] W. X. Zhao et al., "A Survey of Large Language Models," Preprint di arXiv, e-print arXiv:2303.18223, Mar. 2023. [Online]. Available: https://arxiv.org/abs/2303.18223

[3] Y. Tay et al., "Scale Efficiently: Insights from Pre-training and Fine-tuning Transformers," Preprint di arXiv, e-print arXiv:2109.10686, Sep. 2021. [Online]. Available: https://arxiv.org/abs/2109.10686

[4] T. B. Brown et al., "Language Models are Few-Shot Learners," Preprint di arXiv, e-print arXiv:2005.14165, Mei 2020. [Online]. Available: https://arxiv.org/abs/2005.14165

[5] J. Hoffmann et al., "Training Compute-Optimal Large Language Models," Preprint di arXiv, e-print arXiv:2203.15556, Mar. 2022. [Online]. Available: https://arxiv.org/abs/2203.15556

[6] P. Brauner, A. Hick, R. Philipsen, dan M. Ziefle, "What does the public think about artificial intelligence?—A criticality map to understand bias in the public perception of AI," Front. Comp. Sci., vol. 5, Art. no. 1113903, Mar. 2023, doi: 10.3389/fcomp.2023.1113903. Available: https://www.frontiersin.org/journals/computer-science/articles/10.3389/fcomp.2023.1113903/full

[7] S. Mirchandani et al., "Large language models as general pattern machines," Preprint di arXiv, e-print arXiv:2307.04721, Jul. 2023. [Online]. Available: https://arxiv.org/abs/2307.04721

[8] X. Bi et al., "DeepSeek LLM: Scaling Open-Source Language Models with Longtermism," Preprint di arXiv, e-print arXiv:2401.02954, Jan. 2024. [Online]. Available: https://arxiv.org/abs/2401.02954

[9] T. Henighan et al., "Scaling Laws for Autoregressive Generative Modeling," Preprint di arXiv, e-print arXiv:2010.14701, Okt. 2020. [Online]. Available: https://arxiv.org/abs/2010.14701

[10] B. Zhang et al., "When Scaling Meets LLM Finetuning: The Effect of Data, Model and Finetuning Method," Preprint di arXiv, e-print arXiv:2402.17193, Feb. 2024. [Online]. Available: https://arxiv.org/abs/2402.17193