Scaling Law for Multimodal Large Language Model Supervised Fine-Tuning

Anonymous ACL submission

Abstract

The supervised fine-tuning (SFT) stage is crucial for multimodal large language models (MLLMs), yet a comprehensive scaling law to guide the optimal model-data configuration remains lacking. In this paper, we make an initial attempt to address this gap. First, we theoretically demonstrate that directly computing the optimal computation frontier for MLLM-SFT, as we can for traditional LLMs, is a challenging task. This complexity arises because MLLM-SFT is influenced by a broader range of factors, including model size, LLM pre-training tokens, and MLLM SFT tokens. To tackle this issue, we propose two scaling laws based on LLM paradigms: one applicable when training data volumes are well defined by researchers, and another for cases where models are sourced from open communities with unknown training data. Through theoretical modeling and approximations, we provide researchers with valuable recommendations for optimal resource allocation. Furthermore, we establish a strong correlation (R^2 = 0.98) between training loss and downstream performance, enabling accurate performance estimation without the need for exhaustive benchmarking. To validate our scaling laws, we construct a testbed of 60 models ranging from 50 million to 8 billion parameters, totaling 1,560 checkpoints. Each checkpoint is evaluated on than 10 MLLM benchmarks, ensuring robust fitting of our formulations.

1 Introduction

The rapid advancement of MLLMs (Liu et al., 2023b; Zhang et al., 2024a; Fu et al., 2024b) has unlocked unprecedented capabilities in understanding and reasoning across diverse modalities, including text, images, and structured data. A critical stage in developing these models is SFT, where pre-trained large language models (LLMs) are adapted to align with multimodal tasks through curated datasets. While existing works often prioritize scaling SFT data volumes to boost performance (Wang et al., 2024; Liu et al., 2024b), they largely overlook a fundamental question: *Given fixed model architectures or target performance levels, how can we determine the minimal SFT data required to achieve computational efficiency?* Blindly expanding data not only incurs prohibitive annotation costs but also risks diminishing returns, especially in resource-constrained scenarios.

Scaling laws, which quantitatively model relationships between computational resources, model parameters, and performance, have proven instrumental in guiding LLM pre-training (Hoffmann et al., 2022; Kaplan et al., 2020; Clark et al., 2022). However, existing laws are ill-suited for MLLM SFT due to its unique complexities. Unlike LLM training, MLLM SFT is governed by a broader set of factors: model size (N), LLM pre-training data (D_{pretrain}), the LLM's inherent capabilities (P_{base}), and multimodal SFT data (D_{SFT}). Moreover, as we theoretically demonstrate in our paper, deriving a concise "compute frontier" for MLLM SFT—analogous to the $N \propto D^{0.5}$ rule for LLMs—is inherently challenging.

To address this gap, we propose the first systematic framework for MLLM SFT scaling laws. Our approach introduces two complementary paradigms:

- **From-Scratch Scaling Law**: For scenarios with full control over LLM pre-training, we model performance as

$$P = A - B/N^{\alpha} - C/D_{\text{pretrain}}^{\beta} - E/D_{\text{SFT}}^{\gamma},$$

  capturing trade-offs between model size, pre-training, and SFT data.

- **Pre-Trained Model Scaling Law**: For widely adopted open-source LLMs (e.g., LLaMA, Qwen), we link downstream performance to the LLM's benchmark scores (P_{base}) via

$$P = F \cdot P_{\text{base}} - G/N^{\delta} - H/D_{\text{SFT}}^{\zeta},$$

where $A, B, C, E, \alpha, \beta$ are the coefficients and exponents that need to be fitted.

Additionally, we establish a robust correlation ($R^2 = 0.98$) between training loss and downstream accuracy, enabling performance prediction without exhaustive benchmarking. To validate these laws, we construct a testbed of 60 models (50M–8B parameters) and 1,560 checkpoints, rigorously evaluated across 10+ multimodal tasks. Our findings yield actionable insights for MLLM development:

- **Optimal Resources Allocation**: For LLMs trained from scratch, we provide the optimal pre-training token and SFT token numbers for various model sizes. For example, a 1B model is best pre-trained with 20.2B text tokens and fine-tuned with 9.2B image-text tokens. The relationship between D_{SFT} and D_{pretrain} follows a nearly linear growth trend, with $D_{\text{SFT}} \approx 0.48 \times D_{\text{pretrain}}^{0.98}$.

- **LLM Baseline Dominance**: For LLMs with opaque pre-training data, the pre-trained LLM's performance (P_{base}) contributes significantly more to downstream performance (P) than model size or SFT data. This highlights the importance of a strong baseline LLM.

- **Commonsense Reasoning Impact**: The LLM's commonsense reasoning capability has the greatest impact on MLLM performance after SFT. Next in importance is the model's reasoning ability, while capabilities related to Natural Language Inference (NLI) have smaller effect.

- **Task-Specific Dynamics**: Different multimodal tasks exhibit varying preferences for influencing factors. For example, OCR tasks rely heavily on D_{SFT} ($H = 146.2$), while real-world perception tasks benefit more from model scaling ($\delta = 0.13$).

- **Loss-Driven Prediction**: Cumulative training loss predicts downstream accuracy with strong correlation $R^2 = 0.98$, allowing for early stopping and efficient resource reallocation.

These results provide a principled foundation for optimizing MLLM SFT, balancing performance, cost, and practicality.

2 Related Works

Training of Multimodal Large Language Models: MLLMs are typically divided into three stages: pretraining (to bridge the modality gap), instruction tuning, and post-training (Zhang et al., 2025; Lu et al., 2025). Current research primarily focuses on the supervised fine-tuning (SFT) stage, which has been pivotal in enabling models to perform a wide range of multimodal tasks, including image-text alignment, reasoning, and instruction following. This stage also addresses the challenges associated with data fusion across various modalities. Recent open-source MLLMs such as mPLUG-Owl (Ye et al., 2023), LLaVA (Liu et al., 2023b), Qwen-VL (Bai et al., 2023b), Cambrian-1 (Tong et al., 2024), Mini-Gemini (Li et al., 2024b), MiniCPM-V 2.5 (Hu et al., 2024), DeepSeek-VL (Lu et al., 2024), SliME (Zhang et al., 2024a), and the VITA series (Fu et al., 2024a, 2025; Shen et al., 2025) have made significant contributions to the SFT stage, addressing some of the most fundamental challenges in multimodal AI. These include improving vision/audio-language alignment, reasoning, and instruction-following capabilities, thereby enabling more nuanced and context-aware interactions. Some of the most remarkable open-source models, such as InternLM-XComposer-2.5 (Zhang et al., 2023) InternVL-2 (Chen et al., 2023), and QwenVL-2.5 (Bai et al., 2025), have demonstrated impressive strides in multimodal understanding, closely rivaling proprietary models across a variety of multimodal benchmarks.

Neural scaling laws quantify the relationship between model size, dataset size, compute budget, and performance during the training of neural networks. Early works proposed unified formulas for scaling laws and provided practical guidelines for compute-optimal training, laying the foundation for understanding how model performance scales with increased computational resources (Hoffmann et al., 2022; Kaplan et al., 2020). These studies have since been extended to various domains and specialized architectures, offering insights into more specific scenarios. For example, the application of scaling laws to Mixture of Experts (MoE) models has been explored, demonstrating how sparse activation of model parameters introduces unique trade-offs between compute and performance (Clark et al., 2022). Similarly, the use of lower precision training, such as 16-bit floating point numbers, has been studied in the context of scaling laws to reduce memory consumption and computational overhead in large neural networks (Dettmers et al., 2022). Another line of research has conducted extensive experiments on scaling laws in the over-trained regime, addressing performance prediction benchmarks for neural networks and providing new insights into this underexplored area (Gadre et al., 2024b). Beyond the domain of LLMs, scaling laws have been applied in other fields. Image generation research has analyzed how model size and dataset size influence generative performance, providing actionable insights for tasks in computer vision (Henighan et al., 2020; El-Nouby et al., 2024). In the domain of acoustic models, scaling laws have been studied to understand their impact on automatic speech recognition tasks, showcasing their relevance in optimizing models for speech-based applications (Droppo and Elibol, 2021). Despite these advancements, there remains a significant gap in the literature concerning the SFT stage of MLLMs. Unlike the pre-training phase, the SFT stage involves adapting pre-trained models to multi-modal tasks using additional data and task-specific training objectives, making it a distinct and underexplored domain for scaling laws. Currently, no established scaling laws exist to characterize the relationship between model size, fine-tuning data volume, and computational budget during this critical stage.

3 Developing scaling laws for MLLM Supervised Fine-Tuning

In the context of MLLMs, SFT refers to the process of adapting a pre-trained LLM to multi-modal tasks by introducing additional data and training objectives. This stage builds upon the pre-trained LLM, extending its capabilities to understand and process multi-modal inputs such as text, images, and other modalities. SFT focuses on aligning the model's outputs with specific task objectives using curated datasets. The goal of SFT scaling laws is to determine the optimal training data volume required in the SFT stage to minimize computational costs while maximizing performance.

3.1 MLLM SFT Scaling Laws

The scaling laws describe the relationship between the amount of data used in the supervised fine-tuning stage and the model's performance, particularly in the multi-modal domain. To address the unique challenges in the multi-modal fine-tuning process, we define two types of scaling laws:

1. **From-Scratch Language Model Scaling Law**: This scaling law applies when the underlying language model is trained entirely from scratch. In such cases, the model parameters (N), the data volume used during the LLM pre-training phase (D_{pretrain}), and the fine-tuning data volume (D_{SFT}) are precisely known. The scaling law examines how these factors interact to influence the average performance ($P(N, D_{\text{pretrain}}, D_{\text{SFT}})$) of the multi-modal model on downstream tasks. We adopt a parametric form inspired by classical risk decomposition:

$$P(N, D_{\text{pretrain}}, D_{\text{SFT}}) = A - \frac{B}{N^{\alpha}} - \frac{C}{D_{\text{pretrain}}^{\beta}} - \frac{E}{D_{\text{SFT}}^{\gamma}}, \tag{1}$$

where A, B, C, E, α, β, and γ are fitted parameters. This formula captures the contribution of model parameters, pre-training data, and fine-tuning data to the final performance.

2. **Pre-Trained(PT) Language Model Scaling Law** : In practical scenarios, MLLMs often use publicly available pre-trained LLMs, such as Qwen or LLaMA, as foundational models. The training data volume (D_{pretrain}) for these models is typically unknown. Instead, this scaling law leverages the observed performance of the pre-trained LLM on specific benchmark tasks, such as NLI, Commonsense, and Reasoning. The baseline performance (P_{base}) of the pre-trained model is defined as a weighted combination of these tasks:

$$P_{\text{base}} = w_1 P_{\text{NLI}}^{k_1} + w_2 P_{\text{Commonsense}}^{k_2} + w_3 P_{\text{Reasoning}}^{k_3}, \tag{2}$$

where P_{NLI}, $P_{\text{Commonsense}}$, and $P_{\text{Reasoning}}$ are the performance scores on the respective tasks, and w_1, w_2, w_3, k_1, k_2, k_3 are task-specific weights and exponents that are fitted empirically. The scaling law models the relationship between P_{base}, model parameters (N), and fine-tuning data volume (D_{SFT}) to predict the multi-modal model's downstream performance ($P(N, P_{\text{base}}, D_{\text{SFT}})$):

$$P(N, P_{\text{base}}, D_{\text{SFT}}) = F * P_{\text{base}} - \frac{G}{N^{\delta}} - \frac{H}{D_{\text{SFT}}^{\zeta}}, \tag{3}$$

where F, G, δ, and ζ are fitted parameters. This formula accounts for the foundational model's initial capabilities and the incremental improvements from SFT. By empirically evaluating P_{base} across tasks, we provide a framework for optimizing fine-tuning data requirements without relying on unknown pre-training data sizes.

3. Training Loss and Downstream Performance Scaling Law: In addition to the two types of scaling laws classified by model type, we also investigate the scaling law between training loss and downstream performance. In LLMs, scaling laws are often studied in the context of loss. However, in MLLMs, researchers focus less on loss and more on downstream performance metrics. Whether there is a consistent relationship between training loss and downstream performance in MLLMs remains uncertain. To address this question, we hypothesize that the downstream performance P (defined as accuracy in this context) is related to the training loss L through a decaying relationship:

$$P(L) = P_{\text{min}} + \frac{(P_{\text{max}} - P_{\text{min}})}{1 + k \cdot L^{\gamma}}, \tag{4}$$

Specifically, P_{max} represents the maximum achievable performance when the loss L approaches zero, reflecting the model's best possible accuracy under ideal conditions. Conversely, P_{min} denotes the baseline performance when the loss becomes arbitrarily large, ensuring the formula remains bounded and realistic (e.g., reflecting random guessing or task-specific Bayes error). The parameter k controls the sensitivity of performance to changes in loss, allowing the formula to model how quickly performance degrades as loss increases. Finally, γ shapes the decay curve, with higher values resulting in a sharper decline at smaller losses, which aligns with empirical observations where small increases in loss can lead to disproportionately large drops in accuracy.

3.2 Model Fitting

To estimate the parameters in the scaling laws (e.g., A, B, C, E, α, β, γ, w_1, w_2, w_3, k_1, k_2, k_3, F, G, δ, ζ), we minimize the Huber loss between the predicted and observed log performance values using the L-BFGS algorithm (Hoffmann et al., 2022; Aghajanyan et al., 2023):

$$\min_{parameters} \sum_{i}^{\text{Runs}} \text{Huber}_{\delta}\left(\log \hat{P}(\cdot) - \log P_i\right), \tag{5}$$

where δ is a hyperparameter controlling the robustness to outliers. This optimization process accounts for potential local minima by selecting the best fit from a grid of initializations. The Huber loss, with $\delta = 10^{-3}$, is employed for its ability to handle outliers effectively, ensuring reliable predictive performance on held-out data points. To reduce the variance introduced by a single fitting attempt, we perform 100 independent fits and select the best five results based on performance.

In Section A, we discuss the significance of scaling laws in optimizing compute resource allocation, reducing data collection and computational costs, and predicting downstream task performance through stable training loss estimates.

4 Experimental Setup

In our experimental setup, we adopt the LLaVA 1.6 (Liu et al., 2024a) architecture, recognized for its simplicity and efficiency in integrating visual and textual modalities. The visual encoder is based on CLIP-ViT-L-336px, and we employ a dynamic high-resolution strategy for optimal image processing. For language models, we evaluate both models trained from scratch and pretrained models, encompassing a wide range of sizes and datasets—totaling 60 models and over 1560 checkpoints. The training corpus on SFT using the LLaVA-OneVision dataset, which consists of 3.7 million samples. Performance is assessed through benchmarks grouped into four categories: General Capabilities, Real-World (High-Resolution Perception), Chart and Document Understanding Tasks, and Optical Character Recognition (OCR) Tasks. We also evaluate core language model abilities, including reasoning, commonsense understanding, and natural language inference. For a comprehensive overview of the model architecture, training approach, and benchmark details, please refer to **Section B** in the Appendix. This setup provides a robust and comprehensive evaluation of MLLMs across a diverse range of tasks and modalities.

5 From-Scratch LLM Scaling Law

LLMs and MLLMs require efficient scaling of compute, pretraining data, and fine-tuning data to achieve optimal performance. Our objective is to maximize model performance $P(N, D_{\text{pretrain}}, D_{\text{SFT}})$ under a compute budget, while taking into account the constraints between model size, pretraining data, and fine-tuning data. The key constraint is (Kaplan et al., 2020; Hoffmann et al., 2022): FLOPs $= 6N(D_{\text{pretrain}} + D_{\text{SFT}})$

In the ideal case, we should be able to obtain the following expressions $x = C_x FLOPs^{\gamma_x}$, clearly showing the relationship between each variable and FLOPs, thus deriving the compute-optimal frontier. However, as we theoretically demonstrate

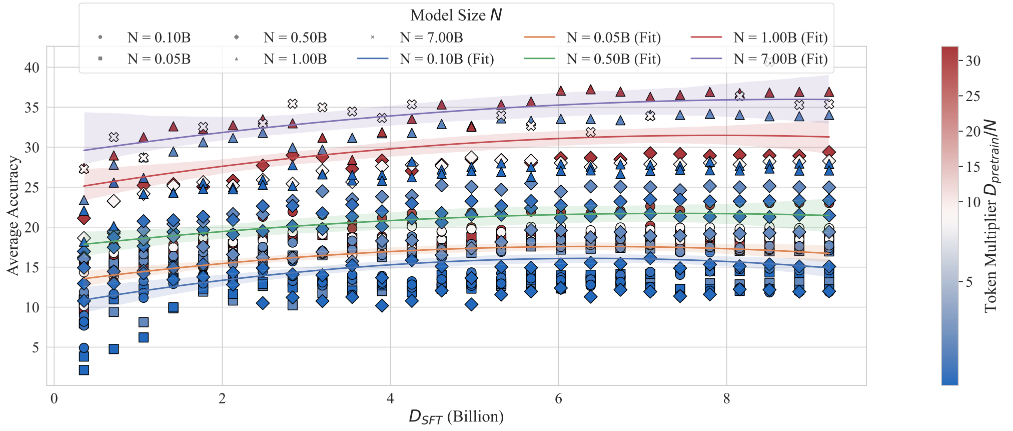


Figure 1: **Scaling law of model performance with respect to the size of D_{SFT} data**, where different colors represent the ratio of pretraining tokens to model parameters. Warmer colors (closer to red) indicate higher ratios, suggesting the model is closer to overtraining during the pretrain stage. Different model sizes are distinguished by varying markers, and quadratic polynomial fits are applied to illustrate the performance trends for each model size. All models except 7B size show signs of overtraining during the SFT stage, with their optimal D_{SFT} token counts aligning closely with the predicted value of from Table 1.

in Section C, deriving a closed-form solution for the multimodal scaling law is much more challednging for LLM scaling laws. Therefore, in the main text, we adopt an approximate fitting strategy. Specifically, we first study the optimal pretrain tokens for an LLM with model size N (for which there are already well-established results in existing works(Kaplan et al., 2020; Hoffmann et al., 2022)). Then, based on our fitting results, we investigate the optimal SFT tokens for the SFT phase, given N and pretrain tokens.

Firstly, in Table 1, we present the optimal pretrain tokens corresponding to different model sizes according to the existing LLM scaling law (Hoffmann et al., 2022). Secondly, in Appendix E, we derive the relationship between pretraining data and SFT data: $D_{\text{pretrain}} = \left(\frac{\beta C}{\gamma E}\right)^{\frac{1}{\beta+1}} D_{\text{SFT}}^{\frac{\gamma+1}{\beta+1}}$. Subsequently, leveraging the parameters fitted to our model:

$$A = 256.76,\ B = 143.75,\ C = 288.56,$$
$$E = 96.17, \alpha = 0.039,\ \beta = 0.054,\ \gamma = 0.074,$$

We determine approximate optimal numbers of Pretrain Tokens and SFT Tokens for a given model size. Figure 1 illustrates our experimental data, where each point represents a model checkpoint. The solid line represents our fitted curve, demonstrating that the results align closely with the approximate optimal solutions in Table 1, validating the efficacy of our approximation method.

Model Size	Pretrain Tokens	SFT Tokens	FLOPs
400 Million	8.0 Billion	3.7 Billion	2.81E+19
1 Billion	20.2 Billion	9.2 Billion	1.76E+20
10 Billion	205.1 Billion	89.5 Billion	1.77E+22
67 Billion	1.5 Trillion	631.0 Billion	8.57E+23
175 Billion	3.7 Trillion	1.5 Trillion	5.46E+24
280 Billion	5.9 Trillion	2.4 Trillion	1.39E+25
520 Billion	11.0 Trillion	4.4 Trillion	4.80E+25
1 Trillion	21.2 Trillion	8.4 Trillion	1.78E+26
10 Trillion	216.2 Trillion	82.9 Trillion	1.79E+28

Table 1: Estimated optimal training compute and tokens for various model sizes.

6 Pre-Trained LLM Scaling Law

Challenges in Deriving the Efficient Frontier: Unlike from-scratch LLM scaling laws, the PT Scaling Law faces a key issue: the pre-training process often involves proprietary or unknown data volume and compute, leaving the relationship between N and P_{base} unclear. Specifically:

- P_{base} often scales with N, but its growth typically saturates. For example, it may follow a saturating function such as:

$$P_{\text{base}}(N) = P_{\text{base,max}} \cdot \left(1 - \frac{1}{N^{\alpha}}\right). \quad (6)$$

- This dependency introduces a non-linear interaction between P_{base}, and N, complicating the optimization problem.

In addition, the performance formula is inherently coupled. For example, increasing N im-

Subset	Task Weights			Task Exponents			Scaling Parameters			Scaling Exponents		R^2
	w_1	w_2	w_3	k_1	k_2	k_3	F	G	H	δ	ζ	
overall	0.2512	0.7018	0.0470	0.4841	0.8895	1.0045	4.4104	34.4322	99.9371	0.0016	0.0350	0.9129
chart & document	0.0232	0.9195	0.0573	0.0272	0.7483	0.7269	7.8389	92.6541	88.2644	0.0050	0.0651	0.8931
general knowledge	0.2169	0.2982	0.4849	0.4703	0.9200	0.5539	6.4711	35.1324	132.0660	0.0085	0.0665	0.8223
ocr	0.0072	0.8320	0.1608	0.2094	0.9861	0.9326	3.0259	27.1967	146.2460	0.0675	0.0208	0.9267
real world	0.5323	0.3813	0.0864	0.7043	1.0221	0.5556	2.6240	46.5741	75.1569	0.1342	0.0220	0.8340

Table 2: **Summary of best fitted parameters across different subsets.** The parameters w_1, w_2, and w_3 represent the weights assigned to NLI, commonsense, and reasoning tasks, respectively, while k_1, k_2, and k_3 are their corresponding exponents. F, G, and H describe the influence of the baseline performance, model size (N), and fine-tuning data (D_{SFT}), respectively. δ and ζ are scaling exponents for N and D_{SFT}. R^2 represents the goodness of fit for the scaling law.

proves P_{base}, but this also reduces marginal returns due to the penalty term $-G/N^{\delta}$. At the same time, increasing D_{SFT} reduces the penalty $-H/D_{\text{SFT}}^{\zeta}$, but its impact is influenced by the starting value of P_{base}. These interactions make it infeasible to derive the efficient frontier analytically without precise knowledge of the functional form of $P_{\text{base}}(N)$. Therefore, in this section, we only provide valuable analysis and suggestions for training MLLMs from pretrained LLMs.

6.1 Analysis of Scaling Parameters and Recommendations

The proposed scaling law offers a holistic perspective on how model size (N), the LLM baseline performance (P_{base}), and fine-tuning data (D_{SFT}) influence downstream performance. Based on the fitted parameters shown in Table 2, the following key insights have been derived:

- **Dominance of LLM Baseline Performance (P_{base}):** The scaling law reveals that the baseline performance of the pre-trained LLM is the primary determinant of downstream performance, as reflected by the relatively high value of $F = 4.3585$. While both model size (N) and fine-tuning data volume (D_{SFT}) make meaningful contributions, their impact is secondary to the inherent capabilities of the pretrained LLM.

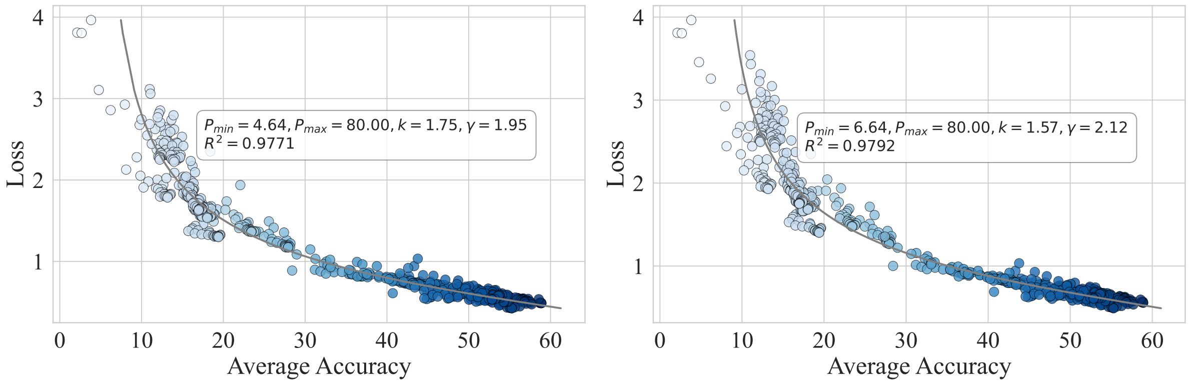
- **Task Contributions within P_{base}:** The task-specific weights (w_1, w_2, w_3) highlight that commonsense reasoning ($w_2 = 0.7598$) is the most critical component of P_{base}, followed by reasoning ($w_3 = 0.0996$) and natural language inference (NLI) ($w_1 = 0.1404$). This underscores that *LLM performance on commonsense reasoning and general reasoning tasks is crucial for developing robust multimodal models.*

- **Impact of Model Size (N):** $G = 64.3677$ and $\delta = 0.0032$ indicate that while increasing model size contributes positively to downstream performance, the marginal gains decrease significantly as the model size grows. This diminishing return suggests that after a certain scale, increases in model size yield limited improvements.

- **Significance of Fine-Tuning Data Volume (D_{SFT}):** The parameters $H = 170.2884$ and $\zeta = 0.0995$ emphasize the critical role of fine-tuning data volume in enhancing downstream performance. Compared to model size, D_{SFT} emerges as the second most influential factor, following P_{base}. The combination of a relatively large H and a moderate ζ suggests that increasing fine-tuning data volume can yield substantial improvements, particularly for smaller or moderately sized LLMs.

7 Average Downstream Performance Scales as a Function of loss.

Reconciling Results from FS-Scaling and PT-Scaling Laws: It is worth noting that the findings presented here appear to differ from those of the FS-scaling law, where model size (N) was identified as more significant than fine-tuning data volume (D_{SFT}). However, this apparent discrepancy can be explained by considering the implicit dependencies within P_{base}. As stated earlier in this section, P_{base} inherently encapsulates the contributions of model size. Therefore, the role of N here reflects the marginal impact of increasing model size *given a fixed baseline LLM performance*, rather than its absolute contribution to the multi-modal downstream performance.



(a)

(b)

Figure 2: **Loss and Accuracy Correlation under Different Loss Calculation Strategies.** The left plot evaluates the model every 1000 steps, using the average loss of those 1000 steps, while the right plot evaluates the model every 1000 steps but uses the cumulative average loss from the beginning of training up to the current step. Both approaches reveal a strong correlation (above 0.977) between loss and accuracy, demonstrating that either loss calculation strategy can effectively reflect the model's performance.

Efficient SFT from Pretrained LLM

1. Investing in pre-trained LLMs with strong commonsense and reasoning capabilities provides the most efficient foundation.
2. Prioritize fine-tuning data scaling, particularly for mid-sized models, to achieve balanced performance gains without over-reliance on massive model sizes.
3. While increasing model size is critical for improving the baseline LLM performance, its contribution to enhancing multi-modal understanding capabilities is relatively limited.

These two sections are thus complementary rather than contradictory. The FS-scaling law highlights the fundamental importance of model size, whereas the PT-scaling law reveals that the primary role of increasing model size lies in improving the LLM baseline performance P_{base}, rather than directly enhancing multi-modal understanding capabilities. By contrast, for multi-modal understanding, fine-tuning data volume (D_{SFT}) emerges as the more significant factor.

We conduct a task-specific scaling law analysis across four key categories: General Knowledge, OCR, Chart and Document Understanding, and Real-World tasks. This analysis reveals distinct trends in how different factors—such as LLM performance (reasoning, commonsense understanding, and NLI) and scaling parameters (model size and fine-tuning data)—contribute to task performance. For instance, general knowledge tasks are heavily influenced by reasoning capabilities, while OCR tasks benefit significantly from fine-tuning data augmentation. Detailed findings, including specific scaling parameters and takeaways for efficient training strategies, are provided in Appendix F.

First, since evaluating the model at every step is computationally prohibitive, we evaluate it every 1000 steps and record the loss at those intervals. However, using the loss of a single point (e.g., at step 1000) as input for fitting yields unstable results due to high variance. This variance arises because the loss at a single step is influenced not only by the model's inherent ability but also by the specific data in the current batch, making it unreliable.

To address this issue, two alternative strategies were adopted, as shown in Figure 2:

1.**Average Loss over 1000 Steps:** This strategy calculates the mean loss over the 1000-step interval before each evaluation to reduce variance. The fitted parameters from this strategy ($P_{\text{min}} = 4.64, P_{\text{max}} = 80.00, k = 1.75, \gamma = 1.95$) indicate a clear relationship between the average loss and downstream performance. The slightly sharper decay ($\gamma = 1.95$) suggests that performance is more sensitive to loss reductions in this setup.

2. **Cumulative Average Loss:** This strategy uses the cumulative average loss from the beginning of training up to the evaluation point. By incorporating a longer history of training performance, this

Task-Specific Loss and Accuracy Predictions

General Knowledge: Loss is predictive of accuracy due to the lower sensitivity ($\gamma = 1.41$) and slower degradation. While loss reductions improve performance, fine-tuning beyond a certain point yields diminishing returns.

Chart & Document Understanding: Training loss is highly predictive of downstream performance ($R^2 = 0.974$), with high sensitivity to low-loss improvements ($\gamma = 2.50$). Fine-tuning for minimal loss is critical, as even small reductions can yield significant accuracy gains.

OCR: Loss and accuracy are strongly correlated ($R^2 = 0.9831$), with the sharpest decay ($\gamma = 2.75$). This task benefits the most from loss reduction, making loss a reliable metric.

Real-World Tasks: Loss is a reasonably strong predictor ($R^2 = 0.9296$). Moderate sensitivity ($\gamma = 1.90$) suggests that loss reductions improve performance but with less drastic gains compared to OCR or Chart tasks. A task-specific approach is recommended.

method reduces the influence of outliers and captures training dynamics more effectively. The fitted parameters ($P_{\text{min}} = 6.64, P_{\text{max}} = 80.00, k = 1.57, \gamma = 2.12$) reveal a slightly higher baseline performance ($P_{\text{min}} = 6.64$) and a steeper decay ($\gamma = 2.12$). This indicates that cumulative averaging is more robust to noise and provides a smoother estimate of the training trajectory.

While the overall scaling law provides a general relationship between training loss and downstream performance, specific tasks exhibit unique sensitivities and dependencies on loss, which require a more granular analysis. To better understand these variations, we summarize the key findings for task-specific scaling laws in the appendix. Below, we highlight the primary takeaways, emphasizing the nuances of using training loss to predict downstream accuracy for different task types. This complements the overall observations and provides actionable insights tailored to specific tasks.

8 Conclusion and Future Work

This work presents the first principled framework for understanding scaling laws in MLLM-SFT. We systematically model the interplay between model size (N), pre-training data (D_{pretrain}), fine-tuning data (D_{SFT}), and the inherent capabilities of pre-trained LLMs (P_{base}). Our findings offer valuable insights into the optimal configuration of these factors for efficient training.

Although this study lays a foundation for optimizing MLLM performance, there are several avenues for future research and aspects not addressed in this work:

1. Exploring Alternative Theoretical Modeling Approaches: As discussed in article, while various approximation methods have been attempted, none lead to a theoretically optimal computational Pareto frontier. In future, we intend to explore alternative modeling and approximation strategies.

2. Interaction Between Model Size and Fine-Tuning Data Volume: We quantitatively model the interaction between model size (N) and fine-tuning data volume (D_{SFT}), establishing a relationship that captures the combined impact on performance:

$$P(N, D_{\text{SFT}}, P_{\text{base}}) = P_{\text{base}} \cdot K + 1 - \frac{F}{(N \cdot D_{\text{SFT}})^{\gamma}},$$

where the parameter γ captures the joint effect of model size and fine-tuning data volume. This model aids in understanding the trade-offs between computational resources and training effectiveness for different configurations.

3. Nonlinear Combination of Tasks: We also demonstrate that the baseline performance (P_{base}) of an LLM can be modeled as a nonlinear combination of the model's capabilities across various tasks, such as NLI, commonsense and general reasoning. The relationship is expressed as:

$$P_{\text{base}} = \left(w_1 P_{\text{NLI}}^{k_1} + w_2 P_{\text{Commonsense}}^{k_2} + w_3 P_{\text{Reasoning}}^{k_3} \right)^{\gamma},$$

where γ controls the degree of nonlinearity, emphasizing the complex interdependencies between different task capabilities.

4. Noise and Uncertainty Modeling: To enhance performance prediction, we incorporate a noise term that accounts for variance in model performance, modeled as:

$$P(N, P_{\text{base}}, D_{\text{SFT}}) = \text{Original Formula} + \epsilon,$$

where $\epsilon \sim \mathcal{N}(0, \sigma^2)$ represents the uncertainty in performance. This addition provides a more robust and reliable prediction framework for MLLM development.

Limitations

This study has several limitations. Firstly, there is a lack of a theoretically optimal computational Pareto frontier, indicating the need to explore alternative theoretical modeling methods. Secondly, the relationship between model size and fine-tuning data volume is not yet fully understood, necessitating the establishment of a quantitative model to analyze its impact on performance. Additionally, while baseline performance is modeled as a nonlinear combination of task capabilities, the complex interdependencies between tasks, such as natural language reasoning, common sense, and inference tasks, require further exploration. Lastly, the current model insufficiently accounts for performance variance, and the study suggests incorporating noise terms to enhance the model's robustness and reliability.

References

2019. Winogrande: An adversarial winograd schema challenge at scale.

Armen Aghajanyan, Lili Yu, Alexis Conneau, Wei-Ning Hsu, Karen Hambardzumyan, Susan Zhang, Stephen Roller, Naman Goyal, Omer Levy, and Luke Zettlemoyer. 2023. Scaling laws for generative mixed-modal language models. In *International Conference on Machine Learning*, pages 265–279. PMLR.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023a. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023b. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Xingyu Chen, Zihan Zhao, Lu Chen, Danyang Zhang, Jiabao Ji, Ao Luo, Yuxuan Xiong, and Kai Yu. 2021. Websrc: A dataset for web-based structural reading comprehension. *arXiv preprint arXiv:2101.09465*.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. 2023. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*.

Wei-Lin Chiang, Zhuohan Li, Ziqing Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3):6.

Aidan Clark, Diego de Las Casas, Aurelia Guy, Arthur Mensch, Michela Paganini, Jordan Hoffmann, Bogdan Damoc, Blake Hechtman, Trevor Cai, Sebastian Borgeaud, et al. 2022. Unified scaling laws for routed language models. In *International conference on machine learning*, pages 4057–4086. PMLR.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. 2022. Gpt3. int8 (): 8-bit matrix multiplication for transformers at scale. *Advances in Neural Information Processing Systems*, 35:30318–30332.

Jasha Droppo and Oguz Elibol. 2021. Scaling laws for acoustic models. *arXiv preprint arXiv:2106.09488*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Alaaeldin El-Nouby, Michal Klein, Shuangfei Zhai, Miguel Angel Bautista, Alexander Toshev, Vaishaal Shankar, Joshua M Susskind, and Armand Joulin. 2024. Scalable pre-training of large autoregressive image models. *arXiv preprint arXiv:2401.08541*.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. 2023. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*.

Chaoyou Fu, Haojia Lin, Zuwei Long, Yunhang Shen, Meng Zhao, Yifan Zhang, Xiong Wang, Di Yin, Long Ma, Xiawu Zheng, et al. 2024a. Vita: Towards open-source interactive omni multimodal llm. *arXiv preprint arXiv:2408.05211*.

Chaoyou Fu, Haojia Lin, Xiong Wang, Yi-Fan Zhang, Yunhang Shen, Xiaoyu Liu, Yangze Li, Zuwei Long, Heting Gao, Ke Li, et al. 2025. Vita-1.5: Towards gpt-4o level real-time vision and speech interaction. *arXiv preprint arXiv:2501.01957*.

Chaoyou Fu, Yi-Fan Zhang, Shukang Yin, Bo Li, Xinyu Fang, Sirui Zhao, Haodong Duan, Xing Sun, Ziwei Liu, Liang Wang, et al. 2024b. Mme-survey: A comprehensive survey on evaluation of multimodal llms. *arXiv preprint arXiv:2411.15296*.

Samir Yitzhak Gadre, Georgios Smyrnis, Vaishaal Shankar, Suchin Gururangan, Mitchell Wortsman, Rulin Shao, Jean Mercat, Alex Fang, Jeffrey Li, Sedrick Keh, Rui Xin, Marianna Nezhurina, Igor Vasiljevic, Jenia Jitsev, Alexandros G. Dimakis, Gabriel Ilharco, Shuran Song, Thomas Kollar, Yair Carmon, Achal Dave, Reinhard Heckel, Niklas Muennighoff, and Ludwig Schmidt. 2024a. Language models scale reliably with over-training and on downstream tasks. *arXiv preprint*. https://arxiv.org/abs/2403.08540.

Samir Yitzhak Gadre, Georgios Smyrnis, Vaishaal Shankar, Suchin Gururangan, Mitchell Wortsman, Rulin Shao, Jean Mercat, Alex Fang, Jeffrey Li, Sedrick Keh, et al. 2024b. Language models scale reliably with over-training and on downstream tasks. *arXiv preprint arXiv:2403.08540*.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *CVPR*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Tom Henighan, Jared Kaplan, Mor Katz, Mark Chen, Christopher Hesse, Jacob Jackson, Heewoo Jun, Tom B Brown, Prafulla Dhariwal, Scott Gray, et al. 2020. Scaling laws for autoregressive generative modeling. *arXiv preprint arXiv:2010.14701*.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. 2022. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*.

Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, et al. 2024. Minicpm: Unveiling the potential of small language models with scalable training strategies. *arXiv preprint arXiv:2404.06395*.

Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. *CVPR*.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. 2016. A diagram is worth a dozen images. In *ECCV*.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024a. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

Yanwei Li, Yuechen Zhang, Chengyao Wang, Zhisheng Zhong, Yixin Chen, Ruihang Chu, Shaoteng Liu, and Jiaya Jia. 2024b. Mini-gemini: Mining the potential of multi-modality vision language models. *arXiv preprint arXiv:2403.18814*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024a. Llava-next: Improved reasoning, ocr, and world knowledge.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*.

Yuliang Liu, Zhang Li, Biao Yang, Chunyuan Li, Xucheng Yin, Cheng-lin Liu, Lianwen Jin, and Xiang Bai. 2023c. On the hidden mystery of ocr in large multimodal models. *arXiv:2305.07895*.

Zhijian Liu, Ligeng Zhu, Baifeng Shi, Zhuoyang Zhang, Yuming Lou, Shang Yang, Haocheng Xi, Shiyi Cao, Yuxian Gu, Dacheng Li, et al. 2024b. Nvila: Efficient frontier visual language models. *arXiv preprint arXiv:2412.04468*.

Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Yaofeng Sun, et al. 2024. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*.

Jinda Lu, Junkang Wu, Jinghan Li, Xiaojun Jia, Shuo Wang, YiFan Zhang, Junfeng Fang, Xiang Wang, and Xiangnan He. 2025. Damo: Data-and model-aware alignment of multi-modal llms. *arXiv preprint arXiv:2502.01943*.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*.

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. 2021. Docvqa: A dataset for vqa on document images. In *WACV*.

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 *EMNLP*.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The refinedweb dataset for falcon llm: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *2011 AAAI Spring Symposium Series*.

Yunhang Shen, Chaoyou Fu, Shaoqi Dong, Xiong Wang, Peixian Chen, Mengdan Zhang, Haoyu Cao, Ke Li, Xiawu Zheng, Yan Zhang, et al. 2025. Long-vita: Scaling large multi-modal models to 1 million tokens with leading short-context accuray. *arXiv preprint arXiv:2502.05177*.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. In *CVPR*.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, et al. 2024. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *arXiv preprint arXiv:2406.16860*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Maurice Weber, Daniel Y. Fu, Quentin Anthony, Yonatan Oren, Shane Adams, Anton Alexandrov, Xiaozhong Lyu, Huu Nguyen, Xiaozhe Yao, Virginia Adams, Ben Athiwaratkun, Rahul Chalamala, Kezhen Chen, Max Ryabinin, Tri Dao, Percy Liang, Christopher Ré, Irina Rish, and Ce Zhang. 2024. Redpajama: an open dataset for training large language models. *NeurIPS Datasets and Benchmarks Track*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. 2024a. Qwen2 technical report. *Preprint*, arXiv:2407.10671.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024b. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. 2023. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*.

Pan Zhang, Xiaoyi Dong Bin Wang, Yuhang Cao, Chao Xu, Linke Ouyang, Zhiyuan Zhao, Shuangrui Ding, Songyang Zhang, Haodong Duan, Hang Yan, et al. 2023. Internlm-xcomposer: A vision-language large model for advanced text-image comprehension and composition. *arXiv preprint arXiv:2309.15112*.

Yi-Fan Zhang, Qingsong Wen, Chaoyou Fu, Xue Wang, Zhang Zhang, Liang Wang, and Rong Jin. 2024a. Beyond llava-hd: Diving into high-resolution large multimodal models. *arXiv preprint arXiv:2406.08487*.

Yi-Fan Zhang, Tao Yu, Haochen Tian, Chaoyou Fu, Peiyan Li, Jianshu Zeng, Wulin Xie, Yang Shi, Huanyu Zhang, Junkang Wu, et al. 2025. Mm-rlhf: The next step forward in multimodal llm alignment. *arXiv preprint arXiv:2502.10391*.

Yi-Fan Zhang, Huanyu Zhang, Haochen Tian, Chaoyou Fu, Shuangqing Zhang, Junfei Wu, Feng Li, Kun Wang, Qingsong Wen, Zhang Zhang, et al. 2024b. Mme-realworld: Could your multimodal llm challenge high-resolution real-world scenarios that are difficult for humans? *arXiv preprint arXiv:2408.13257*.

A The Role of Scaling Laws in Multi-Modal Model Training

Scaling laws are essential for understanding and optimizing the SFT process of MLLMs. These laws quantitatively model the relationships between key factors such as model size (N), fine-tuning data volume (D_{SFT}), pre-training data volume (D_{pretrain}), and the performance of downstream tasks (P). By identifying these relationships, scaling laws provide actionable insights for achieving efficient compute allocation and performance optimization in multi-modal tasks. Specifically, scaling laws serve the following purposes:

1. **Compute-Optimal Allocation**: Scaling laws enable researchers to determine the ideal distribution

of compute resources between model size (N) and the data volume used in SFT (D_{SFT}) to optimize performance. By modeling the loss and performance trade-offs, scaling laws provide a framework to achieve compute-optimal configurations.

2. **Practical Optimization of SFT Data Volume**: Scaling laws are particularly useful for multi-modal SFT as they provide a systematic way to determine the optimal fine-tuning data volume (D_{SFT}) required to achieve a desired level of performance. In many practical scenarios, collecting or annotating large-scale multi-modal datasets is expensive and time-consuming. By using scaling laws, researchers can estimate the minimum necessary D_{SFT} to achieve a target performance, reducing computational and data collection costs. This is especially valuable when leveraging pre-trained LLMs where D_{pretrain} is often unknown or fixed.

3. **Performance Prediction from Training Loss**: As MLLM benchmarks continue to grow in diversity, comprehensively evaluating model performance across all downstream tasks has become increasingly challenging. Scaling laws relating training loss to downstream performance provide a powerful tool for addressing this issue. By modeling the relationship between a model's final convergence loss and its performance (P), researchers can predict performance ranges directly from loss without requiring exhaustive evaluations on every benchmark. This capability simplifies the evaluation process, enabling efficient comparison of models and configurations while reducing the reliance on costly benchmark runs.

In summary, scaling laws provide a critical framework for the compute-efficient design of MLLMs during the SFT stage. By balancing model size, fine-tuning data, and computational resources, these laws ensure that training and fine-tuning processes are both cost-effective and performance-optimized.

B Experimental Setup

B.1 Model Structure

We primarily follow the architecture design of LLaVA 1.6 (Liu et al., 2024a), which is one of the most widely adopted and efficient architectures for MLLMs. This architecture is known for its simplicity and effectiveness. Specifically, our model processes visual information and establishes connections between the visual and textual modalities using the following approaches. By default, we adopt CLIP-ViT-L-336px[1] as the visual encoder. To handle image inputs, we utilize the dynamic high resolution strategy, which is a mainstream approach for image splitting and encoding. This method employs a grid configuration of {2×2, 1×{2,3,4}, {2,3,4}×1} and selects the optimal configuration for splitting and encoding images. Subsequently, the image features are mapped to the textual feature space using a two-layer MLP. The resulting image tokens are concatenated with the text tokens, and the combined tokens are passed into the LLM for further processing.

B.2 Language Model

For the language models trained from scratch, we used 45 models from OpenLM (Gadre et al., 2024a). These models are divided into four different sizes (50M, 0.1B, 0.5B, 1B, 7B), and each size was trained with different datasets and training data ratios. During the SFT phase, we evaluated every 1000 steps, resulting in over one thousand checkpoints for performance evaluation. For models where the data sources and pretraining data volumes are less clear, we selected 15 representative models, including various model sizes from 0.5B to 8B. Similarly, in the SFT phase, performance at every 1000-step evaluation is recorded as a checkpoint. It is worth noting that the checkpoints from language models trained from scratch can also be used for fitting the scaling law in this phase. Specifically, the pretrain LLM scaling law was fitted using 1560 checkpoints. The model sizes and datasets used for both "training from scratch" and "training from pretrain" are summarized in Table 3.

B.3 Training Corpus and Strategy

As our primary focus is not on the pre-training stage of MLLMs, all experiments use the pretraining data from LLaVA-1.5 (Liu et al., 2023a), which consists of 558K samples. The first-stage training is not counted in the total token count. For the SFT stage, we utilize the single-image training dataset from LLaVA-OV (Li et al., 2024a), comprising a total of 3.7M training samples. The average image + text tokens per sample is 2041.7.

B.4 Benchmarks

To comprehensively evaluate the performance of MLLMs on downstream tasks, we categorize the

[1] https://huggingface.co/openai/clip-vit-large-patch14-336

	N	$D_{pretrain}/N$	Pretrain Dataset
From Scratch	50M	0.25, 4, 32	C4, RPJ (Weber et al., 2024), RW (Penedo et al., 2023)
	0.1B	0.25, 4, 32	C4, RPJ (Weber et al., 2024), RW (Penedo et al., 2023)
	0.5B	0.25, 4, 16	C4, RPJ (Weber et al., 2024), RW (Penedo et al., 2023)
	1B	0.25, 1, 4, 16	C4, RPJ (Weber et al., 2024), RW (Penedo et al., 2023)
	7B	1, 4	C4, RPJ (Weber et al., 2024), RW (Penedo et al., 2023)
	N		**Model Name**
From Pretrain	0.5B		Qwen1.5 (Bai et al., 2023a), Qwen2 (Yang et al., 2024a), Qwen2.5 (Yang et al., 2024b)
	1.5/1.8B		Qwen1.5 (Bai et al., 2023a), Qwen2 (Yang et al., 2024a), Qwen2.5 (Yang et al., 2024b)
	3/4B		Qwen1.5 (Bai et al., 2023a), Qwen2 (Yang et al., 2024a)
	7/8B		LLaMA2 (Touvron et al., 2023), LLaMA3 (Dubey et al., 2024), LLaMA3.1 (Dubey et al., 2024), Qwen2 (Yang et al., 2024a), Vicuna1.1 (Chiang et al., 2023), Vicuna1.3 (Chiang et al., 2023), Vicuna1.5 (Chiang et al., 2023)

Table 3: Pretraining datasets and model sizes for language models trained from scratch and pretrained models

MLLM evaluation benchmarks into four groups: 1. *General Capabilities:* This includes benchmarks such as MME (Fu et al., 2023), GQA (Hudson and Manning, 2019), and VQAv2 (Goyal et al., 2017), which assess overall multi-modal performance. 2. *Real-World (High-Resolution Perception):* Benchmarks include RealWorld-QA[2] and MME-RealWorld-CN (Zhang et al., 2024b), targeting high-resolution perception tasks and understanding fine-grained real-world details. 3. *Chart and Document Understanding Tasks:* Benchmarks like ChartQA (Masry et al., 2022), AI2D (Kembhavi et al., 2016), and DocVQA (Mathew et al., 2021) are used to assess the model's capability in understanding structured data and visual information in charts and documents. 4. *Optical Character Recognition (OCR) Tasks:* This includes OCRBench (Liu et al., 2023c), TextVQA (Singh et al., 2019), and WebSRC (Chen et al., 2021), focusing on extracting text information and reasoning over textual content.

To evaluate the foundational performance of the underlying LLM, we assess three key abilities: 1. *Reasoning:* Benchmarks include MMLU (Hendrycks et al., 2020), SciQ[3], and ARC-Easy[4] to test logical and problem-solving abilities. 2. *Commonsense Understanding:* Benchmarks include Winogrande (ai2, 2019) and OpenBookQA (Mihaylov et al., 2018) to evaluate the model's grasp of general world knowledge and commonsense reasoning. 3. *Natural Language Inference (NLI):* Benchmarks such as COPA (Roemmele et al., 2011) and RTE[5] are used to test the model's ability to infer relationships between statements. These benchmarks provide a holistic evaluation of the MLLM's performance, encompassing both its multi-modal and foundational language model capabilities. By covering a diverse set of tasks, we ensure that the scaling laws are applicable to a wide range of real-world use cases.

C Direct Calculation of the Efficient Frontier is Extremely Challenging

Assume the optimal solution follows a power-law form, i.e., there exist constants k_1, k_2, k_3 and exponents a, b, and c, such that $N = k_1\,\text{FLOPs}^a$, $D_{\text{pretrain}} = k_2\,\text{FLOPs}^b$, $D_{\text{SFT}} = k_3\,\text{FLOPs}^c$.

Substituting into the Compute Constraint: Given the compute constraint

$$6\,N\,(D_{\text{pretrain}} + D_{\text{SFT}}) = \text{FLOPs},$$

substitute the assumptions into the equation:

$$6\left(k_1\,\text{FLOPs}^a\right)\left(k_2\,\text{FLOPs}^b + k_3\,\text{FLOPs}^c\right) = 6\,k_1\left(k_2\,\text{FLOPs}^{a+b} + k_3\,\text{FLOPs}^{a+c}\right) = \text{FLOPs}.$$

For the equation to hold for all values of FLOPs, the highest FLOPs exponent in the two terms must be exactly 1. A common assumption is that D_{pretrain} and D_{SFT} are "balanced" in terms of resource allocation, i.e., $b = c$. This gives us:

$$6\,k_1(k_2 + k_3)\,\text{FLOPs}^{a+b} = \text{FLOPs}.$$

This implies that

$$a + b = 1, \quad \text{or equivalently,} \quad b = c = 1 - a.$$

In practical scenarios, the data sizes for MLLM SFT and LLM pretrain are generally not on the same scale. In other words, **there is a gap between the theoretical and actual results**! This gap arises

[2] https://x.ai/blog/grok-1.5v

[3] https://huggingface.co/datasets/allenai/sciq

[4] https://huggingface.co/datasets/allenai/ai2_arc

[5] https://aclweb.org/aclwiki/Recognizing_Textual_Entailment

from our modeling approach. In reality, the size of D_{SFT} should be closely related to N and D_{pretrain}, rather than being independent of them (i.e., there should be a nonlinear relationship among the three; for example, an LLM trained with a sufficiently large D_{pretrain} intuitively converges faster than a freshly initialized LLM).

However, there are two difficulties in modeling this dependency:

1. First, we do not know what kind of nonlinear dependency this would be.

2. As shown in Section D, even without considering additional inequality or equality constraints, it is extremely difficult to perform mathematical analysis to obtain the Efficient Frontier for our problem, let alone more complex modeling approaches.

Therefore, in the main text, we adopt an approximate fitting strategy. Specifically, we first study the optimal pretrain tokens for an LLM with model size N (for which there are already well-established results in existing works). Then, based on our fitting results, we investigate the optimal SFT tokens for the SFT phase, given N and pretrain tokens.

D Failure Case 1: Challenges in Approximating the Efficient Frontier

In this section, we explore an approximation strategy to derive the Efficient Frontier, ultimately we successfully describe the dependence of D_{pretrain}, D_{SFT}, and N on $FLOPs$. However, this approximation may incur significant errors, leading to poor performance in practical applications.

D.1 Assumptions

1. **Performance Function**: The performance P is decomposed into contributions from model size N, pretraining data D_{pretrain}, and fine-tuning data D_{SFT}:

$$P(N, D_{\text{pretrain}}, D_{\text{SFT}}) = A - \frac{B}{N^{\alpha}} - \frac{C}{D_{\text{pretrain}}^{\beta}} - \frac{E}{D_{\text{SFT}}^{\gamma}}$$

2. **Compute Constraint**: Compute resources are consumed as:

$$6N(D_{\text{pretrain}} + D_{\text{SFT}}) = \text{FLOPs}$$

D.2 Efficiency Frontier Derivation

We use the Lagrange multiplier method to incorporate the compute constraint:

$$\mathcal{L}(N, D_{\text{pretrain}}, D_{\text{SFT}}, \lambda) = -P(N, D_{\text{pretrain}}, D_{\text{SFT}}) + \lambda\left(\text{FLOPs} - 6N(D_{\text{pretrain}} + D_{\text{SFT}})\right)$$

Taking partial derivatives and solving, we find:

1. For N:

$$\frac{\partial \mathcal{L}}{\partial N} = \frac{\alpha B}{N^{\alpha+1}} - 6\lambda(D_{\text{pretrain}} + D_{\text{SFT}}) = 0$$

2. For D_{pretrain}:

$$\frac{\partial \mathcal{L}}{\partial D_{\text{pretrain}}} = \frac{\beta C}{D_{\text{pretrain}}^{\beta+1}} - 6\lambda N = 0$$

3. For D_{SFT}:

$$\frac{\partial \mathcal{L}}{\partial D_{\text{SFT}}} = \frac{\gamma E}{D_{\text{SFT}}^{\gamma+1}} - 6\lambda N = 0$$

4. For λ:

$$\frac{\partial \mathcal{L}}{\partial \lambda} = \text{FLOPs} - 6N(D_{\text{pretrain}} + D_{\text{SFT}}) = 0$$

Since there are many variables involved, it is very difficult to directly obtain the closed-form solution for each variable with respect to FLOPs. Therefore, for simplicity, we introduce an auxiliary variable

$$\psi \equiv 6\lambda N.$$

In section. D.3, we show

$$\boxed{\psi \approx \frac{C_1}{FLOPs^{1/p}}},$$

where

$$p = \frac{1}{\alpha} + \frac{\alpha - 1}{\alpha} \cdot \frac{1}{1 + \min(\beta, \gamma)},$$

and

$$C_1 = 6^{1/p} (\alpha\, B)^{\frac{1}{\alpha\, p}} \left[(\beta\, C)^{\frac{1}{\beta+1}} + (\gamma\, E)^{\frac{1}{\gamma+1}} \right]^{\frac{\alpha-1}{\alpha\, p}}.$$

From this, we can step by step obtain the dependence of D_{pretrain}, D_{SFT}, and N on $FLOPs$ based on the optimization problem and the approximate solution for ψ.

1. First-order Optimality Conditions for the Data Size: Take the partial derivatives of D_{pretrain} and D_{SFT}, and set them to zero to obtain:

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial D_{\text{pretrain}}} &= -\beta\,\frac{C}{D_{\text{pretrain}}^{\beta+1}} + 6\lambda N = 0 \\
\implies \quad D_{\text{pretrain}}^{\beta+1} &= \frac{\beta\,C}{6\lambda N}, \\
\frac{\partial \mathcal{L}}{\partial D_{\text{SFT}}} &= -\gamma\,\frac{E}{D_{\text{SFT}}^{\gamma+1}} + 6\lambda N = 0 \\
\implies \quad D_{\text{SFT}}^{\gamma+1} &= \frac{\gamma\,E}{6\lambda N}.
\end{aligned}$$

Thus, the above expressions can be written as

$$\begin{aligned}
D_{\text{pretrain}} &= \left(\frac{\psi}{\beta\,C}\right)^{-1/(\beta+1)} \\
&= (\beta\,C)^{1/(\beta+1)}\,\psi^{-1/(\beta+1)}, \\
D_{\text{SFT}} &= \left(\frac{\psi}{\gamma\,E}\right)^{-1/(\gamma+1)} \\
&= (\gamma\,E)^{1/(\gamma+1)}\,\psi^{-1/(\gamma+1)}.
\end{aligned}$$

2. Optimality Condition for N: Taking the partial derivative of the objective function with respect to N, we get:

$$-\alpha\,\frac{B}{N^{\alpha+1}} + 6\lambda\Big(D_{\text{pretrain}} + D_{\text{SFT}}\Big) = 0.$$

Remembering that $\psi = 6\lambda N$, we can rearrange this to obtain:

$$\frac{\alpha\,B}{N^{\alpha}} = \psi\Big(D_{\text{pretrain}} + D_{\text{SFT}}\Big).$$

Substitute the expressions for D_{pretrain} and D_{SFT} from earlier. Note that

$$D_{\text{pretrain}} + D_{\text{SFT}} = \\ (\beta\,C)^{1/(\beta+1)}\,\psi^{-1/(\beta+1)} + (\gamma\,E)^{1/(\gamma+1)}\,\psi^{-1/(\gamma+1)}.$$

Thus, we get

$$\frac{\alpha\,B}{N^{\alpha}} = \\ (\beta\,C)^{1/(\beta+1)}\,\psi^{1-\frac{1}{\beta+1}} + (\gamma\,E)^{1/(\gamma+1)}\,\psi^{1-\frac{1}{\gamma+1}}.$$

Solving for N, we obtain

$$N = \\ \left\{\frac{\alpha\,B}{(\beta\,C)^{1/(\beta+1)}\,\psi^{1-\frac{1}{\beta+1}} + (\gamma\,E)^{1/(\gamma+1)}\,\psi^{1-\frac{1}{\gamma+1}}}\right\}^{1/\alpha}$$

The original computational budget requirement is:

$$6N\Big(D_{\text{pretrain}} + D_{\text{SFT}}\Big) = FLOPs.$$

Substituting the earlier expressions for D_{pretrain} and D_{SFT}, we can rewrite it as

$$N\left[(\beta C)^{1/(\beta+1)}\,\psi^{-1/(\beta+1)} + (\gamma E)^{1/(\gamma+1)}\,\psi^{-1/(\gamma+1)}\right] \\ = \frac{FLOPs}{6}.$$

(From now on, we use F to represent $FLOPs$, which is simply an adjustment of the constant factor.)

Expressing the Dependence of Variables on F (i.e., $FLOPs$):

(1) Data Size Using the expressions obtained earlier:

$$\begin{aligned}
D_{\text{pretrain}} &= (\beta C)^{1/(\beta+1)}\psi^{-1/(\beta+1)} \\
&= (\beta C)^{1/(\beta+1)} C_1^{-1/(\beta+1)} F^{1/(p(\beta+1))}, \\
D_{\text{SFT}} &= (\gamma E)^{1/(\gamma+1)}\psi^{-1/(\gamma+1)} \\
&= (\gamma E)^{1/(\gamma+1)} C_1^{-1/(\gamma+1)} F^{1/(p(\gamma+1))}.
\end{aligned}$$

This can be written as:

$$\boxed{\begin{aligned}
D_{\text{pretrain}} &= \Big[(\beta C)^{1/(\beta+1)} C_1^{-1/(\beta+1)}\Big]\, F^{\frac{1}{p(\beta+1)}}, \\
D_{\text{SFT}} &= \Big[(\gamma E)^{1/(\gamma+1)} C_1^{-1/(\gamma+1)}\Big]\, F^{\frac{1}{p(\gamma+1)}}.
\end{aligned}}$$

(2) Model Size N

Recalling the expression for N:

$$N = \\ \left\{\frac{\alpha B}{(\beta C)^{1/(\beta+1)}\psi^{1-\frac{1}{\beta+1}} + (\gamma E)^{1/(\gamma+1)}\psi^{1-\frac{1}{\gamma+1}}}\right\}^{1/\alpha}.$$

In the limit of large F (i.e., small ψ), assuming that pretraining data dominates (i.e., $\beta \leq \gamma$, so $1/(\beta+1) > 1/(\gamma+1)$), the first term dominates, and we approximate:

$$N \approx (\beta C)^{1/((\beta+1)\alpha)}\psi^{-1/(\alpha(\beta+1))}(\alpha B)^{1/\alpha}.$$

Substituting $\psi \approx C_1/F^{1/r}$, we get

$$N \approx \\ (\alpha B)^{1/\alpha}(\beta C)^{-1/((\beta+1)\alpha)} C_1^{-1/(\alpha(\beta+1))} F^{\frac{1}{p\alpha(\beta+1)}}.$$

Similarly, if fine-tuning data dominates ($\gamma < \beta$), we have

$$N \approx \\ .(\alpha B)^{1/\alpha}(\gamma E)^{-1/((\gamma+1)\alpha)} C_1^{-1/(\alpha(\gamma+1))} F^{\frac{1}{p\alpha(\gamma+1)}}.$$

Thus, we can summarize the conditional expression as:

$$N = \begin{cases} (\alpha B)^{1/\alpha} (\beta C)^{-1/((\beta+1)\alpha)} C_1^{-1/(\alpha(\beta+1))} F^{\frac{1}{p\alpha(\beta+1)}}, \\ \beta \leq \gamma, \\ (\alpha B)^{1/\alpha} (\gamma E)^{-1/((\gamma+1)\alpha)} C_1^{-1/(\alpha(\gamma+1))} F^{\frac{1}{p\alpha(\gamma+1)}}, \\ \gamma < \beta. \end{cases}$$

where

$$p = \frac{1}{\alpha} + \frac{\alpha - 1}{\alpha} \frac{1}{1 + \min(\beta, \gamma)}.$$

D.3 Derivation of the optimal ψ

We aim to address the following constrained optimization problem

$$\begin{aligned} \min_{N, D_{\text{pretrain}}, D_{\text{SFT}} > 0} \ & \frac{B}{N^\alpha} + \frac{C}{D_{\text{pretrain}}^\beta} + \frac{E}{D_{\text{SFT}}^\gamma} \\ \text{s. t.} \quad & 6N(D_{\text{pretrain}} + D_{\text{SFT}}) = FLOPs \end{aligned}$$

Using the Lagrangian multiplier λ, it becomes minimax optimization problem, i.e.

$$\begin{aligned} \max_{\lambda \geq 0} \min_{N, D_{\text{pretrain}}, D_{\text{SFT}} > 0} \ & \frac{B}{N^\alpha} + \frac{C}{D_{\text{pretrain}}^\beta} + \frac{E}{D_{\text{SFT}}^\gamma} \\ & + \lambda \left(6N(D_{\text{pretrain}} + D_{\text{SFT}}) - FLOPs \right) \end{aligned}$$

We first fix N and λ. In this case, we can redfine $\psi = 6\lambda N$ and write the solutions for D_{pretrain} and D_{SFT} in terms of ψ,

$$D_{\text{pretrain}} = \left(\frac{\psi}{\beta C} \right)^{-1/(\beta+1)}, \quad D_{\text{SFT}} = \left(\frac{\psi}{\gamma E} \right)^{-1/(\gamma+1)}$$

Then, the equation for N becomes

$$\frac{\alpha B}{N^\alpha} = \frac{\psi^{\beta/(\beta+1)}}{(\beta C)^{-1/(\beta+1)}} + \frac{\psi^{\gamma/(\gamma+1)}}{(\gamma E)^{-1/(\gamma+1)}}$$

and therefore

$$N = \left(\frac{\psi^{\beta/(\beta+1)}}{(\beta C)^{-1/(\beta+1)}} + \frac{\psi^{\gamma/(\gamma+1)}}{(\gamma E)^{-1/(\gamma+1)}} \right)^{-1/\alpha} (\alpha B)^{1/\alpha}$$

Putting all these together, according to $6N(D_{\text{pretrain}} + D_{\text{SFT}}) = FLOPs$, we have

$$6\psi^{-1/\alpha} \left(\left(\frac{\psi}{\beta C} \right)^{-1/(\beta+1)} + \left(\frac{\psi}{\gamma E} \right)^{-1/(\gamma+1)} \right)^{(\alpha-1)/\alpha} = FLOPs \, (\alpha B)^{-1/\alpha}$$

We begin by examining the summation term within the parentheses:

$$S(\psi) = \left(\frac{\psi}{\beta \, C} \right)^{-1/(\beta+1)} + \left(\frac{\psi}{\gamma \, E} \right)^{-1/(\gamma+1)}.$$

Rewriting this in an equivalent form:

$$\begin{aligned} S(\psi) = & (\beta \, C)^{1/(\beta+1)} \, \psi^{-1/(\beta+1)} \\ & + (\gamma \, E)^{1/(\gamma+1)} \, \psi^{-1/(\gamma+1)}. \end{aligned}$$

Note that when FLOPs are large, in order to meet the budget constraint, ψ must become very small. In this case, the powers of ψ in the two terms are $-1/(\beta + 1)$ and $-1/(\gamma + 1)$. Clearly, as $\psi \to 0$, the term with the more "negative" exponent (i.e., the larger value) will dominate the sum.

Notice that

$$\frac{1}{\beta + 1} \quad \text{and} \quad \frac{1}{\gamma + 1}$$

the larger of these can be written as

$$\frac{1}{1 + \min(\beta, \gamma)},$$

because if $\min(\beta, \gamma)$ is smaller, the corresponding term $1/(1 + \min(\beta, \gamma))$ will be larger than the other. Therefore, when ψ is small, we have the approximation

$$S(\psi) \sim \left[(\beta \, C)^{1/(\beta+1)} + (\gamma \, E)^{1/(\gamma+1)} \right] \psi^{-1/(1+\min(\beta,\gamma))}.$$

Substituting the Dominant Term into the Original Equation: The left-hand side of the original equation is

$$LHS = 6\psi^{-1/\alpha} \Big[S(\psi) \Big]^{(\alpha-1)/\alpha}.$$

Substituting the approximate form of $S(\psi)$, we get

$$LHS \sim 6\psi^{-1/\alpha} \cdot \left\{ \left[(\beta \, C)^{1/(\beta+1)} + (\gamma \, E)^{1/(\gamma+1)} \right] \psi^{-1/(1+\min(\beta,\gamma))} \right\}^{(\alpha-1)/\alpha}.$$

By separating the constants from the powers of ψ, we have

$$LHS \sim 6 \left[(\beta \, C)^{1/(\beta+1)} + (\gamma \, E)^{1/(\gamma+1)} \right]^{(\alpha-1)/\alpha} \cdot \psi^{-1/\alpha - \frac{\alpha-1}{\alpha} \cdot \frac{1}{1+\min(\beta,\gamma)}}.$$

Thus, the overall power of ψ is

$$-\left(\frac{1}{\alpha}+\frac{\alpha-1}{\alpha}\cdot\frac{1}{1+\min(\beta,\gamma)}\right).$$

Expressing the Right-hand Side and Solving for ψ: The right-hand side of the budget constraint is

$$RHS = FLOPs\,(\alpha\,B)^{-1/\alpha}.$$

We approximate the two sides as equal (typically equality is taken at the optimal solution), and we write

$$\left[(\beta\,C)^{1/(\beta+1)}+(\gamma\,E)^{1/(\gamma+1)}\right]^{(\alpha-1)/\alpha}\cdot \psi^{-\left[\frac{1}{\alpha}+\frac{\alpha-1}{\alpha}\cdot\frac{1}{1+\min(\beta,\gamma)}\right]} \approx FLOPs/6 * \ (\alpha\,B)^{-1/\alpha}.$$

Rearranging the above expression into a form for ψ, we get

$$\psi^{\frac{1}{\alpha}+\frac{\alpha-1}{\alpha}\cdot\frac{1}{1+\min(\beta,\gamma)}} \approx \frac{6(\alpha\,B)^{-1/\alpha}}{FLOPs}\left[(\beta\,C)^{1/(\beta+1)}+(\gamma\,E)^{1/(\gamma+1)}\right]^{(\alpha-1)/\alpha}.$$

Taking the reciprocal and extracting the appropriate powers, we obtain

$$\psi \approx \frac{C_1}{FLOPs^{1/p}} \quad \text{where} \quad p=\frac{1}{\alpha}+\frac{1}{1+\min(\beta,\gamma)}.$$

Here, we approximate the exponent by

$$p \approx \frac{1}{\alpha}+\frac{\alpha-1}{\alpha}\frac{1}{1+\min(\beta,\gamma)},$$

which simplifies the description of the scaling relationship between FLOPs and ψ.

Deriving the Complete Expression for C_1: From the equation above, we write

$$\psi^{p} \approx \frac{6(\alpha\,B)^{-1/\alpha}}{FLOPs}\left[(\beta\,C)^{1/(\beta+1)}+(\gamma\,E)^{1/(\gamma+1)}\right]^{(\alpha-1)/\alpha}.$$

That is,

$$\psi \approx \left\{\frac{6(\alpha\,B)^{-1/\alpha}}{\left[(\beta\,C)^{1/(\beta+1)}+(\gamma\,E)^{1/(\gamma+1)}\right]^{-(\alpha-1)/\alpha}}\right\}^{1/p}\cdot\frac{1}{FLOPs^{1/p}}.$$

For simplicity, we define

$$C_1 = 6^{1/p}\cdot\left[(\alpha\,B)^{-1/\alpha}\left((\beta\,C)^{1/(\beta+1)}+(\gamma\,E)^{1/(\gamma+1)}\right)^{-(\alpha-1)/\alpha}\right]^{-1/p}.$$

Or equivalently,

$$C_1 = 6^{1/p}(\alpha\,B)^{1/(\alpha\,p)}\left[(\beta\,C)^{1/(\beta+1)}+(\gamma\,E)^{1/(\gamma+1)}\right]^{\frac{\alpha-1}{\alpha\,p}},$$

That is, we have

$$\boxed{\psi \approx \frac{C_1}{FLOPs^{1/p}}},$$

where

$$p=\frac{1}{\alpha}+\frac{\alpha-1}{\alpha}\cdot\frac{1}{1+\min(\beta,\gamma)},$$

and

$$C_1 = 6^{1/p}(\alpha\,B)^{\frac{1}{\alpha\,p}}\left[(\beta\,C)^{\frac{1}{\beta+1}}+(\gamma\,E)^{\frac{1}{\gamma+1}}\right]^{\frac{\alpha-1}{\alpha\,p}}.$$

From this, we can obtain the dependence of D_{pretrain}, D_{SFT}, and N on $FLOPs$.

E Connection between D_{SFT} and $D_{pretrain}$

We begin with the following system of equations:

1. For N:

$$\frac{\alpha B}{N^{\alpha+1}} = 6\lambda(D_{\text{pretrain}}+D_{\text{SFT}})$$

2. For D_{pretrain}:

$$\frac{\beta C}{D_{\text{pretrain}}^{\beta+1}} = 6\lambda N$$

3. For D_{SFT}:

$$\frac{\gamma E}{D_{\text{SFT}}^{\gamma+1}} = 6\lambda N$$

From the equations for D_{pretrain} and D_{SFT}, we can derive the relationship:

$$\frac{\beta C}{D_{\text{pretrain}}^{\beta+1}} = \frac{\gamma E}{D_{\text{SFT}}^{\gamma+1}}$$

Rearranging, we get:

$$D_{\text{pretrain}}^{\beta+1} = \frac{\beta C}{\gamma E}D_{\text{SFT}}^{\gamma+1}$$

Thus, we have:

$$D_{\text{pretrain}} = \left(\frac{\beta C}{\gamma E}\right)^{\frac{1}{\beta+1}} D_{\text{SFT}}^{\frac{\gamma+1}{\beta+1}}$$

F Task-Specific Scaling Law Analysis

While the overall scaling law provides a broad understanding, task-specific subsets reveal distinct trends. Below, we analyze the scaling parameters with a focus on the contributions of LLM performance ($w_1, w_2, w_3, k_1, k_2, k_3$) and scaling impacts for each subset.

General Knowledge (MME, VQA v2, GQA):

- **LLM Performance:** $w_1 = 0.2169$, $w_2 = 0.2982$, $w_3 = 0.4849$, $k_1 = 0.4703$, $k_2 = 0.9200$, $k_3 = 0.5539$: Reasoning tasks ($P_{\text{Reasoning}}$) contribute the most to LLM performance, followed by commonsense ($P_{\text{Commonsense}}$), while NLI (P_{NLI}) plays a smaller role. Exponentially scaling commonsense yields the strongest effect on task performance.

- **Scaling Impact:** $F = 6.4711$, $G = 35.1324$, $\delta = 0.0085$, $H = 132.0660$, $\zeta = 0.0665$: LLM performance dominates general knowledge tasks. Model size has steep diminishing returns beyond 7B, and fine-tuning data provides secondary contributions.

OCR (OCRBench, TextVQA):

- **LLM Performance:** $w_1 = 0.0072$, $w_2 = 0.8320$, $w_3 = 0.1608$, $k_1 = 0.2094$, $k_2 = 0.9861$, $k_3 = 0.9326$: OCR tasks are predominantly driven by commonsense ($P_{\text{Commonsense}}$), with reasoning and NLI playing secondary roles. Commonsense and reasoning have the strongest exponential impacts.

- **Scaling Impact:** $F = 3.0259$, $G = 27.1967$, $\delta = 0.0675$, $H = 146.2460$, $\zeta = 0.0208$: Fine-tuning data volume overwhelmingly drives OCR performance, with significant returns even for smaller models (0.5B–3B).

Chart and Document Understanding (ChartQA, AI2D, DocVQA):

- **LLM Performance:** $w_1 = 0.0232$, $w_2 = 0.9195$, $w_3 = 0.0573$, $k_1 = 0.0272$, $k_2 = 0.7483$, $k_3 = 0.7269$: Commonsense ($P_{\text{Commonsense}}$) dominates, reflecting the importance of structured knowledge in document-related tasks. Reasoning ($P_{\text{Reasoning}}$) has a moderate impact, while NLI plays a minimal role.

- **Scaling Impact:** $F = 7.8389$, $G = 92.6541$, $\delta = 0.0050$, $H = 88.2644$, $\zeta = 0.0651$: Model size and fine-tuning data contribute more equally compared to other subsets, reflecting the need for a balanced scaling strategy.

Real-World (High-Resolution Perception, RealWorld-QA, MME-RealWorld):

- **LLM Performance:** $w_1 = 0.5323$, $w_2 = 0.3813$, $w_3 = 0.0864$, $k_1 = 0.7043$, $k_2 = 1.0221$, $k_3 = 0.5556$: NLI (P_{NLI}) becomes the dominant contributor, reflecting the importance of logical reasoning and textual entailment for real-world tasks. Commonsense plays a secondary role, while reasoning has limited impact.

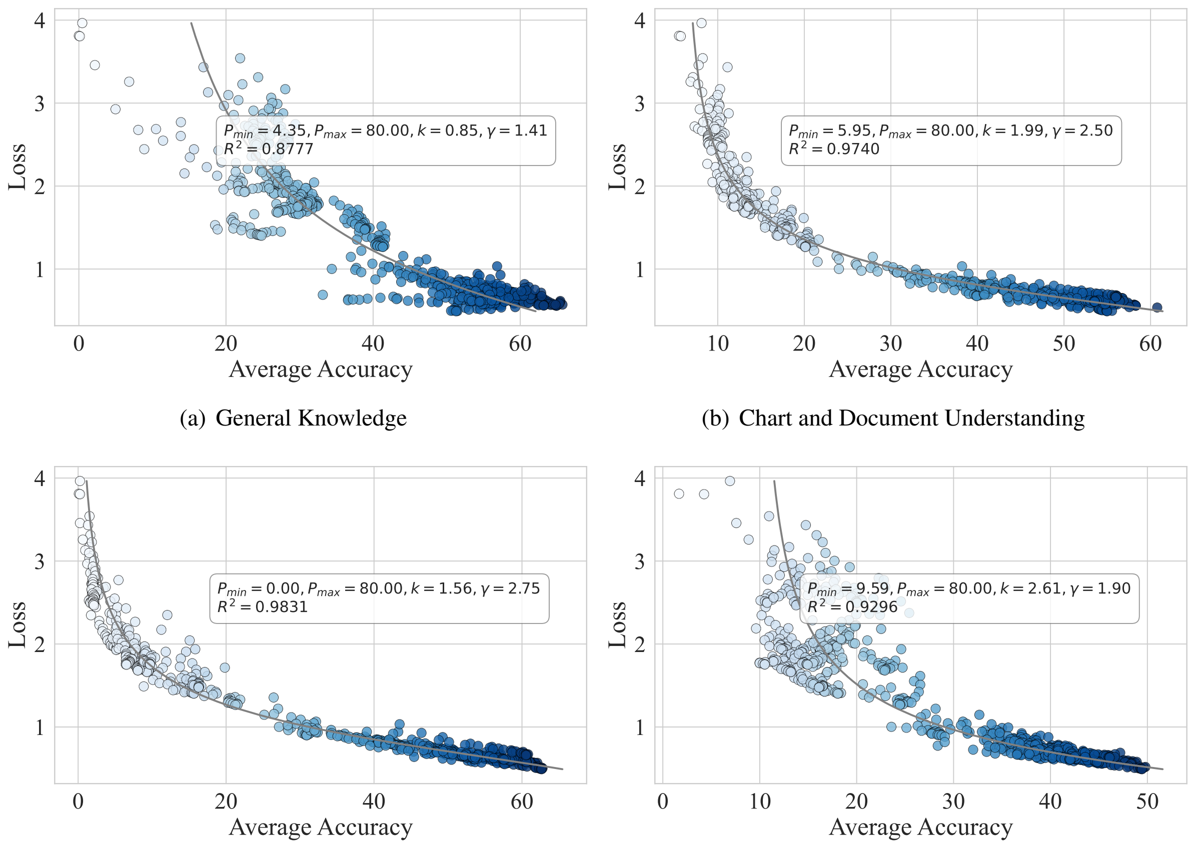
- **Scaling Impact:** $F = 2.6240$, $G = 46.5741$, $\delta = 0.1342$, $H = 75.1569$, $\zeta = 0.0220$: Model size significantly impacts real-world tasks, with slower diminishing returns compared to general knowledge. Fine-tuning data remains important but secondary.

Efficient Training of Specific Tasks

General Knowledge: Pre-trained LLMs with strong reasoning and commonsense capabilities are essential. Scaling model size beyond 7B yields limited gains.
OCR: Focus on fine-tuning data augmentation, as smaller models paired with robust datasets can achieve competitive performance.
Chart & Document Understanding: A balanced strategy scaling both model size and fine-tuning data volume is critical.
Real-World Tasks: Prioritize scaling model size to handle task complexity. Fine-tuning data quality should take precedence over quantity.

G Task-Specific Analysis of Performance-Loss Scaling Laws

Because our computations use model sizes smaller than 13B, the optimal average performance or task-specific performance generally does not exceed 60. Under such circumstances, the scaling law tends to select a smaller P_{max} to optimize the fitting loss, which limits its ability to extrapolate to better-performing models. To address this, we set a minimum value for P_{max} at 80 to ensure that the scaling law retains the ability to extrapolate for



(a) General Knowledge

(b) Chart and Document Understanding

(c) OCR

(d) Real-World

Figure 3: **Task-Specific Loss and Accuracy Correlation.** This figure illustrates the performance-loss scaling laws for specific tasks, including General Knowledge, Chart and Document Understanding, OCR, and Real-World Tasks. For each task, the model is evaluated every 1000 steps using the cumulative average loss from the beginning of training to the current step. Across all tasks, a strong correlation between training loss and accuracy is observed, demonstrating that cumulative loss is an effective metric for predicting task-specific downstream performance.

smaller losses. Even with this hard constraint, our experiments demonstrate that training loss remains strongly correlated with task-specific performance, which is shown in Figure. 3.

G.1 General Knowledge (MME, VQA v2, GQA)

Prediction Results: $P_{\text{min}} = 4.35, P_{\text{max}} = 80.00, k = 0.85, \gamma = 1.41, R^2 = 0.8777$

Analysis: General Knowledge tasks exhibit a shallow decay ($\gamma = 1.41$), suggesting that accuracy is less sensitive to variations in training loss. The low $k = 0.85$ indicates slower performance degradation as loss increases, and the moderately high $R^2 = 0.8777$ shows that training loss is a reasonable, though not perfect, predictor. The lower correlation is likely due to the reliance of these tasks on multimodal reasoning, commonsense, and instruction following capability[6], which may not be directly reflected in training loss.

Key Insight: Loss is moderately predictive of performance, but the relationship is weaker compared to other tasks. Fine-tuning for loss reductions has limited benefits beyond a certain point.

Difference from Overall Scaling Law: General Knowledge tasks deviate significantly from the overall scaling law due to their lower sensitivity (γ) and slower degradation (k).

G.2 Chart and Document Understanding

Prediction Results: $P_{\text{min}} = 5.95, P_{\text{max}} = 80.00, k = 1.99, \gamma = 2.50, R^2 = 0.974$

Analysis: These tasks show a steep decay ($\gamma = 2.50$), highlighting high sensitivity to changes in training loss. The higher $k = 1.99$ reflects faster degradation in accuracy as loss increases. The high $R^2 = 0.974$ indicates a strong correlation between loss and performance, suggesting that training loss

[6] The MME benchmark needs the model to directly answer 'yes' or 'no', which is hard for small scale models.

is a reliable predictor for these tasks, which require precise feature extraction and structured reasoning.

Key Insight: Training loss is highly predictive of downstream performance, especially in the low-loss region where small improvements yield significant accuracy gains.

Difference from Overall Scaling Law: Compared to the overall scaling law, these tasks show much higher sensitivity ($\gamma = 2.50$) and a more pronounced dependency on loss.

G.3 OCR (OCRBench, TextVQA)

Prediction Results: $P_{\text{min}} = 0.00, P_{\text{max}} = 80.00, k = 1.56, \gamma = 2.75, R^2 = 0.9831$

Analysis: OCR tasks demonstrate the sharpest decay ($\gamma = 2.75$) among all tasks, indicating extreme sensitivity to small loss reductions. The baseline $P_{\text{min}} = 0.00$ reflects the absence of meaningful performance from random guessing. The very high $R^2 = 0.9831$ shows that training loss is an excellent predictor for OCR tasks, where precise text recognition is critical.

Key Insight: Fine-tuning to achieve minimal loss is essential for OCR tasks, as even small improvements in loss yield significant performance gains.

Difference from Overall Scaling Law: OCR tasks show a much stronger dependency on low losses and sharper decay than the overall scaling law, emphasizing the importance of fine-grained loss optimization.

G.4 Real-World Tasks (High-Resolution Perception, RealWorld-QA, MME-RealWorld)

Prediction Results: $P_{\text{min}} = 9.59, P_{\text{max}} = 80.00, k = 2.61, \gamma = 1.90, R^2 = 0.9296$

Analysis: Real-World tasks exhibit moderately steep decay ($\gamma = 1.90$) and a higher baseline performance ($P_{\text{min}} = 9.59$), suggesting these tasks retain some accuracy even with higher losses. The moderately high $R^2 = 0.9296$ shows that training loss is a reasonably strong predictor for these tasks, though less so than for OCR or Chart tasks. The higher $k = 2.61$ indicates faster performance degradation.

Key Insight: While training loss is predictive, the relationship is less sharp than in OCR or Chart Understanding tasks, suggesting that other task-specific factors may play a larger role.

Difference from Overall Scaling Law: Real-World tasks align more closely with the overall scaling law but exhibit a higher baseline and faster degradation.

H Broader impacts

By lowering costs and providing clearer guidelines (a "principled basis for optimizing MLLM SFT"), this research can make the development of advanced MLLMs more accessible to a wider range of researchers and organizations, including those with limited resources. This could foster broader innovation and application in the field.