

DeepSeek LLM

Scaling Open-Source Language Models with Longtermism

Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, Huazuo Gao, Kaige Gao, Wenjun Gao, Ruiqi Ge, Kang Guan, Daya Guo, Jianzhong Guo, Guangbo Hao, Zhewen Hao, Ying He, Wenjie Hu, Panpan Huang, Erhang Li, Guowei Li, Jiashi Li, Yao Li, Y.K. Li, Wenfeng Liang, Fangyun Lin, A.X. Liu, Bo Liu, Wen Liu, Xiaodong Liu, Xin Liu, Yiyuan Liu, Haoyu Lu, Shanghao Lu, Fuli Luo, Shirong Ma, Xiaotao Nie, Tian Pei, Yishi Piao, Junjie Qiu, Hui Qu, Tongzheng Ren, Zehui Ren, Chong Ruan, Zhangli Sha, Zhihong Shao, Junxiao Song, Xuecheng Su, Jingxiang Sun, Yaofeng Sun, Minghui Tang, Bingxuan Wang, Peiyi Wang, Shiyu Wang, Yaohui Wang, Yongji Wang, Tong Wu, Y. Wu, Xin Xie, Zhenda Xie, Ziwei Xie, Yiliang Xiong, Hanwei Xu, R.X. Xu, Yanhong Xu, Dejian Yang, Yuxiang You, Shuiping Yu, Xingkai Yu, B. Zhang, Haowei Zhang, Lecong Zhang, Liyue Zhang, Mingchuan Zhang, Minghua Zhang, Wentao Zhang, Yichao Zhang, Chenggang Zhao, Yao Zhao, Shangyan Zhou, Shunfeng Zhou, Qihao Zhu, Yuheng Zou *

***DeepSeek-AI**

Abstract

The rapid development of open-source large language models (LLMs) has been truly remarkable. However, the scaling laws described in previous literature presents varying conclusions, which casts a dark cloud over scaling LLMs. We delve into the study of scaling laws and present our distinctive findings that facilitate the scaling of large scale models in two prevalent used open-source configurations, 7B and 67B. Guided by the scaling laws, we introduce DeepSeek LLM, a project dedicated to advancing open-source language models with a long-term perspective. To support the pre-training phase, we have developed a dataset that currently consists of 2 trillion tokens and is continuously expanding. We further conduct supervised fine-tuning (SFT) and direct preference optimization (DPO) on DeepSeek LLM Base models, resulting in the creation of DeepSeek Chat models. Our evaluation results demonstrate that DeepSeek LLM 67B surpasses LLaMA-2 70B across a range of benchmarks, especially in the domains of code, mathematics, and reasoning. Furthermore, open-ended evaluations reveal that our DeepSeek LLM 67B Chat exhibits superior performance compared to GPT-3.5.

*Authors are ordered alphabetically by the last name.

Contents

1 Introduction	3
2 Pre-Training	4
2.1 Data	4
2.2 Architecture	5
2.3 Hyperparameters	5
2.4 Infrastructures	6
3 Scaling Laws	7
3.1 Scaling Laws for Hyperparameters	8
3.2 Estimating Optimal Model and Data Scaling	9
3.3 Scaling Laws with Different Data	12
4 Alignment	12
5 Evaluation	13
5.1 Public Benchmark Evaluation	13
5.1.1 Base Model	14
5.1.2 Chat Model	14
5.2 Open-Ended Evaluation	17
5.2.1 Chinese Open-Ended Evaluation	17
5.2.2 English Open-Ended Evaluation	18
5.3 Held-Out Evaluation	18
5.4 Safety Evaluation	19
5.5 Discussion	20
6 Conclusion, Limitation, and Future Work	23
A Appendix	30
A.1 Acknowledgments	30
A.2 Different Model Scale Representations	30
A.3 Benchmark Metrics Curves	31
A.4 Comparison with Code or Math Specific Models	32
A.5 Benchmark Results w/ DPO Stage	32
A.6 Evaluation Formats	32

1. Introduction

Over the past few years, Large Language Models (LLMs) based on decoder-only Transformers (Vaswani et al., 2017) have increasingly become the cornerstone and pathway to achieving Artificial General Intelligence (AGI). By predicting the next word in continuous text, LLMs undergo self-supervised pre-training on massive datasets, enabling them to achieve various purposes and possess many abilities, such as novel creation, text summarization, code completion, and more. Subsequent developments like supervised fine-tuning and reward modeling have enabled Large Language Models (LLMs) to better follow user intentions and instructions. This has endowed them with more versatile conversational capabilities and rapidly expanded their influence.

This wave is sparked with closed products, such as ChatGPT (OpenAI, 2022), Claude (Anthropic, 2023), and Bard (Google, 2023), which are developed with extensive computational resources and substantial annotation costs. These products have significantly raised the community’s expectations for the capabilities of open-source LLMs, consequently inspiring a series of work (Bai et al., 2023; Du et al., 2022; Jiang et al., 2023; Touvron et al., 2023a,b; Yang et al., 2023). Among these, the LLaMA series models (Touvron et al., 2023a,b) stand out. It consolidates a range of works to create an efficient and stable architecture, building well-performing models ranging from 7B to 70B parameters. Consequently, the LLaMA series has become the de facto benchmark for architecture and performance among open-source models.

Following LLaMA, the open-source community has primarily focused on training fixed-size (7B, 13B, 34B, and 70B), high-quality models, often neglecting research exploration into LLM scaling laws (Hoffmann et al., 2022; Kaplan et al., 2020). Nonetheless, research on scaling laws is of utmost importance, considering that the current open-source models are merely at the initial stage of Artificial General Intelligence (AGI) development. In addition, early works (Hoffmann et al., 2022; Kaplan et al., 2020) reached varying conclusions on the scaling of model and data with increased compute budgets and inadequately addressed hyperparameter discussions. In this paper, we extensively investigate the scaling behavior of language models and apply our findings in two widely used large-scale model configurations, namely 7B and 67B. Our study aims to lay the groundwork for future scaling of open-source LLMs, paving the way for further advancements in this domain. Specifically, we first examined the scaling laws of batch size and learning rate, and found their trends with model size. Building on this, we conducted a comprehensive study of the scaling laws of the data and model scale, successfully revealing the optimal model/data scaling-up allocation strategy and predicting the expected performance of our large-scale models. Additionally, during development, we discovered that the scaling laws derived from different datasets show significant differences. This suggests that choice of dataset remarkably affects the scaling behavior, indicating that caution should be exercised when generalizing scaling laws across datasets.

Under the guidance of our scaling laws, we build from scratch open-source large language models, and release as much information as possible for community reference. We collect 2 trillion tokens for pre-training, primarily in Chinese and English. At the model level, we generally followed the architecture of LLaMA, but replaced the cosine learning rate scheduler with a multi-step learning rate scheduler, maintaining performance while facilitating continual training. We collected over 1 million instances for supervised fine-tuning (SFT) (Ouyang et al., 2022) from diverse sources. This paper shares our experiences with different SFT strategies and findings in data ablation techniques. Additionally, we have utilized direct preference optimization (DPO) (Rafailov et al., 2023) to improve the conversational performance of the model.

We conduct extensive evaluations using our base and chat models. The evaluation results demonstrate that DeepSeek LLM surpasses LLaMA-2 70B across various benchmarks, particularly in the fields of code, mathematics, and reasoning. Following SFT and DPO, the DeepSeek 67B chat model outperforms GPT-3.5 in both Chinese and English open-ended evaluations. This highlights the superior performance of DeepSeek 67B in generating high-quality responses and engaging in meaningful conversations in both languages. Furthermore, the safety evaluation indicates that DeepSeek 67B Chat can provide harmless responses in practice.

In the rest of this paper, we first introduce our pre-training basic concepts of DeepSeek LLM in Section 2, including the composition of data, model architecture, infrastructure, and hyperparameters. In Section 3, we provide a detailed explanation of the scaling laws we have discovered and its implications. Additionally, we discuss the rationale behind our selection of pre-training hyperparameters, taking into account the insights gained from the scaling laws analysis. In Section 4, we discuss our fine-tuning methodology, encompassing the composition of fine-tuning data and specific methods during the SFT and DPO stages. We then present the detailed evaluation results of DeepSeek LLM in Section 5, covering both the base and chat models, as well as their performance in open-ended evaluations and safety evaluations. Finally, we discuss the current limitations and future directions of DeepSeek LLM in Section 6.

2. Pre-Training

2.1. Data

Our main objective is to comprehensively enhance the richness and diversity of the dataset. We have gained valuable insights from reputable sources such as (Computer, 2023; Gao et al., 2020; Penedo et al., 2023; Touvron et al., 2023a). To achieve these goals, we have organized our approach into three essential stages: deduplication, filtering, and remixing. The deduplication and remixing stages ensure a diverse representation of the data by sampling unique instances. The filtering stage enhances the density of information, thereby enabling more efficient and effective model training.

We adopted an aggressive deduplication strategy, expanding the deduplication scope. Our analysis revealed that deduplicating the entire Common Crawl corpus results in higher removal of duplicate instances compared to deduplicating within a single dump. Table 1 illustrates that deduplicating across 91 dumps eliminates four times more documents than a single dump method.

Dumps Used	1	2	6	12	16	22	41	91
Deduplication Rate (%)	22.2	46.7	55.7	69.9	75.7	76.3	81.6	89.8

Table 1 | Deduplication ratios for various Common Crawl dumps.

In the filtering stage, we focus on developing robust criteria for document quality assessment. This involves a detailed analysis incorporating both linguistic and semantic evaluations, providing a view of data quality from individual and global perspectives. In the remixing phase, we adjust our approach to address data imbalances, focusing on increasing the presence of underrepresented domains. This adjustment aims to achieve a more balanced and inclusive dataset, ensuring that diverse perspectives and information are adequately represented.

For our tokenizer, we implemented the Byte-level Byte-Pair Encoding (BBPE) algorithm based on the tokenizers library (Huggingface Team, 2019). Pre-tokenization was employed to