RETHINKING DATA SELECTION AT SCALE: RANDOM SELECTION IS ALMOST ALL YOU NEED

Tingyu Xia[1,3†] **Bowen Yu**[2*] **Kai Dang**[2] **An Yang**[2] **Yuan Wu**[1,3*] **Yuan Tian**[1,3]
Yi Chang[1,3,4] **Junyang Lin**[2]
[1]School of Artificial Intelligence, Jilin University
[2]Alibaba Group
[3]Engineering Research Center of Knowledge-Driven Human-Machine Intelligence, MOE, China
[4]International Center of Future Science, Jilin University
xiaty21@mails.jlu.edu.cn, yubowen.ybw@alibaba-inc.com, dangkai.dk@alibaba-inc.com
ya235025@alibaba-inc.com, yuanwu@jlu.edu.cn, yuantian@jlu.edu.cn
yichang@jlu.edu.cn, junyang.ljy@alibaba-inc.com

ABSTRACT

Supervised fine-tuning (SFT) is crucial for aligning Large Language Models (LLMs) with human instructions. The primary goal during SFT is to select a small yet representative subset of training data from the larger pool, such that fine-tuning with this subset achieves results comparable to or even exceeding those obtained using the entire dataset. However, most existing data selection techniques are designed for small-scale data pools, which fail to meet the demands of real-world SFT scenarios. In this paper, we replicated several self-scoring methods—those that do not rely on external model assistance—on two million-scale datasets, and found that nearly all methods struggled to significantly outperform random selection when dealing with such large-scale data pools. Moreover, our comparisons suggest that, during SFT, diversity in data selection is more critical than simply focusing on high-quality data. We also analyzed the limitations of several current approaches, explaining why they perform poorly on large-scale datasets and why they are unsuitable for such contexts. Finally, we found that filtering data by token length offers a stable and efficient method for improving results. This approach, particularly when training on long-text data, proves highly beneficial for relatively weaker base models, such as Llama3. The code is available at `https://github.com/xiatingyu/SFT-DataSelection-at-scale`.

1 INTRODUCTION

With the advent of large language models (LLMs) such as ChatGPT, we have observed significant advancements in tasks involving instruction following (Wang et al., 2023b), intent comprehension (Lu et al., 2023), and text generation (Zhao et al., 2023). One of the primary objectives of developing LLMs is to harness their potential for generalizing to unseen natural language processing (NLP) tasks. To achieve this aim, many LLMs focus on precisely aligning with human instructions.

Recent studies indicate that supervised fine-tuning (SFT) can customize LLMs for specific domains, tasks, or applications by utilizing well-crafted data. According to the study in Zhou et al. (2024a), it is feasible to fine-tune a pre-trained language model with a relatively small set of examples. Building on this insight, several papers have explored data selection strategies for SFT of LLMs (Wang et al., 2024; Qin et al., 2024), emphasizing the importance of enhancing the quality of instruction tuning (IT) data or increasing data diversity. These strategies can be classified into two primary categories: (1) Extenral-scoring methods, which require support from more sophisticated external models like GPT-4 to score the data for the subsequent selection (Lu et al., 2023; Chen et al., 2023; Du et al., 2023; Liu et al., 2023; Zhou et al., 2024b); (2) Self-scoring methods, which leverage LLMs them-

[†]Work done during the author's internship at the Alibaba Group
[*]Corresponding authors

selves as data scorers (Zhou et al., 2023a; Li et al., 2023d;b; Liu et al., 2024; Xia et al., 2024; Yin et al., 2024).

Existing SFT data selection methodologies, both external-scoring and self-scoring, are primarily assessed using several widely recognized IT datasets, such as alpaca-GPT4 (Peng et al., 2023), Dolly (Conover et al., 2023), FLAN (Longpre et al., 2023), WizardLM (Xu et al., 2024), and ShareGPT (Chiang et al., 2023). These datasets are limited in size and originate from a single source. Meanwhile, in practical applications, to fully leverage the inherent knowledge of LLMs, the SFT process frequently necessitates a substantial volume of varied data, ideally in the range of millions or more. This discrepancy creates a gap between the present SFT data selection strategies and real-world applications. In order to observe the impact brought by the dataset size on the performance of different selection strategies, we analyze the difference in outcomes between existing SFT data selection methods and random selection within source datasets ranging from 10K-30K to 1M on Llama3-8B (AI@Meta, 2024). As shown in Figure 1, when the scale of the datasets increases to 1M, these data selection methods yield suboptimal performance compared with random selection.

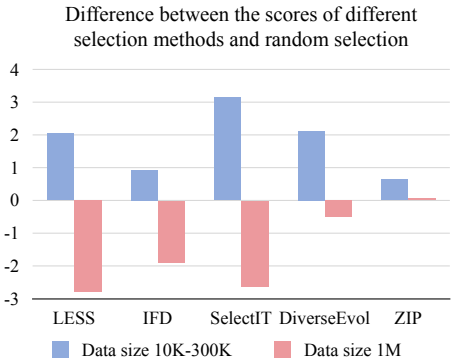

Figure 1: The discrepancy between each methods and random selection on BBH benchmark. The Y-axis represents the differential score, which is computed by subtracting the random selection score from the scores obtained using various methods.

Inspired by this finding, we rethink whether SFT data selection methods can work when they are required to handle large-scale IT datasets. For external-scoring approaches, it is impractical to apply them to tackle vast amounts of IT data due to the substantial costs (Liu et al., 2023), we hence focus on the self-scoring methods. For self-scoring approaches, we refer to the article Qin et al. (2024) to categorize the techniques into two types: data quality-based methods and data diversity-based methods. To explore how self-scoring methods influence LLMs' performance when dealing with large-scale IT data, we evaluate several recent methods on two benchmarks that contain millions of instances. The findings from our experiments reveal three main points:

- Most self-scoring data selection techniques do not significantly outperform random selection on large-scale datasets. Even though these self-scoring methods can achieve significant gains on small-scale datasets, their effectiveness will be greatly reduced when the data size increases and the data sources become complex.
- Data diversity holds more significance than data quality during the SFT phase. Data quality-based selection methods are more effective than data diversity-based methods when dealing with a small-scale dataset from a single source. However, when tackling multi-source data, only considering data quality is far from enough.
- Through a comparative empirical analysis of two IT datasets, we find that it is useful to utilize token length as a criterion to conduct data filtering, yielding stable and efficient results for SFT when dealing with large-scale IT data. Previous work (Liu et al., 2023) has demonstrated the benefit of long texts training for models on subjective evaluation tasks such as MTbench (Zheng et al., 2023) and AlpacaEval (Li et al., 2023c), we have further confirmed the positive effect of long texts training on objective evaluation tasks, such as Big-Bench-Hard (Suzgun et al., 2022). While utilizing token length in SFT may not yield optimal outcomes on every language model, it is highly beneficial for applying it in training with long texts, especially on a relatively weak BASE language model, like Llama3-8B.

2 Related Work

External-scoring Method. Lu et al. (2023) introduced an open-set instruction tagging method called INSTAG, which employed ChatGPT to generate detailed tags to measure and examine the variety and intricacy of human instructions for LLMs during SFT. Chen et al. (2023) presented the