Dynamics of Instruction Fine-Tuning for Chinese Large Language Models

Chiyu Song[1,2] **Zhanchao Zhou**[1,2] **Jianhao Yan**[1,2]
Yuejiao Fei[1,2] **Zhenzhong Lan**[2,*] **Yue Zhang**[2,3,*]
[1]Zhejiang University [2]School of Engineering, Westlake University
[3] Institute of Advanced Technology, Westlake Institute for Advanced Study
songchiyu@westlake.edu.cn

Abstract

Instruction tuning is a burgeoning method to elicit the general intelligence of Large Language Models (LLMs). While numerous studies have examined the impact of factors such as data volume and model size on English models, the scaling properties of instruction tuning in other languages remain largely unexplored. In this work, we systematically investigate the effects of data quantity, model size, and data construction methods on instruction tuning for Chinese LLMs. We utilize a newly curated dataset, *DoIT*, which includes over 40,000 high-quality instruction instances covering ten underlying abilities, such as creative writing, code generation, and logical reasoning. Our experiments, conducted on models ranging from 7b to 33b parameters, yield three key findings: (i) While these factors directly affect overall model performance, some abilities are more responsive to scaling, whereas others demonstrate significant resistance. (ii) The scaling sensitivity of different abilities to these factors can be explained by two features: Complexity and Transference. (iii) By tailoring training strategies to their varying sensitivities, specific abilities can be efficiently learned, enhancing performance on two public benchmarks.

1 Introduction

Large Language Models (LLMs) have shown impressive capabilities across diverse tasks (Brown et al., 2020; Touvron et al., 2023; Chowdhery et al., 2022; Almazrouei et al., 2023; Wang et al., 2022a; Wei et al., 2022b; Zhao et al., 2021; Wei et al., 2023; Ivison et al., 2022; Zhang et al., 2023c; Radford et al., 2019), demonstrating their potential for artificial general intelligence (Bubeck et al., 2023). A key contributor to this success is instruction tuning, a process involving supervised fine-tuning of LLMs on instruction-output pairs (Ouyang et al., 2022; Taori et al., 2023; Chiang et al., 2023; Iyer et al., 2022; Zhou et al., 2023).

Despite the recognition that various factors, such as data quantity, distribution, and construction methods, directly impact the performance of instruction tuning (Zhao et al., 2023; Zhang et al., 2023b; Wang et al., 2023), there remains an inconsistent understanding of their specific roles in shaping model capabilities. For instance, while some studies (Wei et al., 2022a; Sanh et al., 2021) argue that scaling data volume is crucial for the success of certain tasks, other results (Zhou et al., 2023) suggest that a limited number of instructions can be sufficient. These discrepancies highlight the complexity of instruction tuning and raise concerns about the generalizability of these conclusions to other languages. Moreover, existing studies predominantly focus on English datasets (Hestness et al., 2017; Zhang et al., 2024), with a notable lack of comparative analyses in other languages, such as Chinese.

To address the above issues, we introduce *DoIT*, a new Chinese dataset encompassing over 40,000 human-curated instruction instances that span ten distinct LLM abilities. Each data instance is rigorously revised by human annotators to ensure high-quality text and is categorized according to its specific ability. Based on this dataset, we are able to disentangle the effects of each factor by maintaining control over others, thereby providing a clearer understanding of how data volume, parameter size, and construction methods individually influence the development of different abilities. To achieve this, we also employ pre-trained Chinese models such as Chinese-LLaMA (Cui et al., 2023), Baichuan2 (Yang et al., 2023), and Qwen1.5 (Bai et al., 2023), resulting in a comprehensive set of instruction-tuned models with sizes ranging from 7 billion to 33 billion parameters.

The results reveal three primary findings:

* Corresponding authors.

Proceedings of the 31st International Conference on Computational Linguistics, pages 10345–10366
January 19–24, 2025.

1. Data quantity and parameter size significantly influence overall performance, but each ability develops at different paces during instruction tuning. Abilities such as Creative Writing are more responsive to these factors and can be well-trained with a small amount of data. In contrast, abilities like Ethics show resistance to these changes.

2. We investigate the reasons behind these discrepancies and identify two features, Complexity and Transference, which can be calculated in low-resource scenarios. These features indicate the potential for ability growth when scaling up data or model parameters.

3. By leveraging the different sensitivities of abilities to scaling, we can make models learn specific abilities more efficiently and achieve practical performance improvements on two comprehensive benchmarks, CMMLU (Li et al., 2023) and AGIEval (Zhong et al., 2023).

We open-source our codebase, dataset, and model checkpoints for reproducibility and future research[1].

2 Related Work

The influence of data factors on instruction tuning has been widely studied, but some conclusions remain inconsistent. Some studies (Wei et al., 2022a; Chung et al., 2022; Zhang et al., 2024) suggest that larger datasets improve model performance, while others (Zhou et al., 2023) indicate that a smaller, high-quality dataset can suffice. Although there is evidence that instruction-tuned models generalize well (Sanh et al., 2021; Wei et al., 2022a), some argue that this generalization is limited to tasks heavily supported in the training data (Gudibande et al., 2023). Synthetic data has shown promise (Wang et al., 2022b; Yin et al., 2023), but the model's capability is constrained by imitating proprietary systems (Gudibande et al., 2023).

These discrepancies have motivated us to explore how various abilities develop during instruction tuning within a Chinese context. Our research, detailed in Section 4, reveals significant disparities in the impact of different factors on various abilities. This insight may help reconcile the differing conclusions drawn from prior studies.

[1]https://github.com/ChiyuSONG/dynamics-of-instruction-tuning

Instruction datasets are also crucial for the efficacy of instruction-tuned LLMs, and their construction methods can be broadly categorized into three types: Task-formatted datasets (Sanh et al., 2021; Muennighoff et al., 2022; Wei et al., 2022a; Chung et al., 2022; Mishra et al., 2021; Wang et al., 2022c) incorporate instances from diverse NLP tasks using human-crafted templates to enable multi-task training. While platforms like PromptSource(Bach et al., 2022) have been developed to expand these datasets, concerns about their alignment with real user requests (Ouyang et al., 2022; Zhao et al., 2023) have led to the exploration of alternative methods. Human-curated datasets (Ouyang et al., 2022; Zhou et al., 2023; Conover et al., 2023; Köpf et al., 2023) address the issue above using real-life tasks with human labeling, such as genuine user queries or examination questions. Proprietary models like ChatGPT (OpenAI, 2022) and GPT-4 (OpenAI, 2023) employ this data source for training. Unfortunately, these datasets are often not publicly available due to the high cost and effort required. Synthetic datasets (Honovich et al., 2022; Xu et al., 2023a,b) offer a cost-effective solution by semi-automating instruction generation. One approach is collecting user chats with proprietary models as in ShareGPT[2]. Self-Instruct (Wang et al., 2022b) is another representative approach, which bootstraps datasets from a small set of seed tasks. This approach has inspired open-source projects like Alpaca (Taori et al., 2023) and Vicuna (Chiang et al., 2023).

3 DoIT: A New Instruction Dataset

To systematically investigate the roles of data quantity, parameter size, and data construction methods in shaping a range of model abilities, it is necessary to rule out the influence of data quality and establish a controllable data distribution among different abilities. To fulfill these research needs, we introduce *DoIT*, a new human-curated dataset. This dataset contains over 40,000 quality-controlled Chinese instances, categorized into ten distinct ability classes, allowing for tailored experimental setups.

Following the literature reviewed in Section 2, our human-curated data are derived from real-life contexts, such as academic examinations, online platforms, and user queries. By referencing existing taxonomies (Chang et al., 2023; Huang et al., 2023), we focus on practical problems in diverse

[2]https://sharegpt.com/

Ability	Data Source	Data Size	
		1st Round	2nd Round
STEM - Biology	COIG - Exam (Zhang et al., 2023a)	1,200	1,242
Humanity - History	COIG - Exam (Zhang et al., 2023a)	1,200	2,093
Code Generation	Leetcode.cn	1,200	5,168
Creative Writing	User Queries from In-House Data	1,200	1,200
Chinese	COIG - Exam (Zhang et al., 2023a)	1,200	1,650
Dialogue Understanding	C3-D (Sun et al., 2020)	1,200	5,085
Role-play Chat	BELLE (Ji et al., 2023)	1,200	1,200
Logical Reasoning	LogiQA2.0 (Liu et al., 2023)	1,200	12,951
COT for Grad-Math	PRM800K (Lightman et al., 2023)	1,200	11,701
Ethics	COIG - Human Value (Zhang et al., 2023a)	1,200	1,200

Table 1: The data sources and data size after two rounds of human annotation for each ability category.

fields, including ethics, education, engineering, and creative generation, extending beyond basic language capabilities. The ten selected representative abilities encompass a broad spectrum of assessment: (1) STEM subject - Biology, (2) Humanities subject - History, (3) Code Generation, (4) Creative Writing, (5) Language Proficiency - Chinese, (6) Dialogue Understanding, (7) Role-play Chat, (8) Logical Reasoning, (9) Chain of Thought, and (10) Ethics.

To maintain consistent quality across all instances, we employ a three-stage annotation process:

1. **Standardization**: Data from diverse sources significantly differ in format, including raw web pages, exam papers, user inputs, and data pre-cleaned by other researchers to different extents. In this stage, we convert them into consistent instruction-output pairs, applying tailored rules for each category to extract relevant text and eliminate duplicates. Notably, the "Chain of Thought" data originated from PRM800K (Lightman et al., 2023) is the only non-Chinese source and is translated using the ChatGPT (OpenAI, 2022) API before human review.

2. **Human Filtering**: Each item is then reviewed by two independent annotators. They are required to (i) Check the correctness of the text. (ii) Control the diversity of instructions, such as filtering out high-frequency personas in Role-play Chat. (iii) Avoid potential ethical issues in the output, such as biased opinions in Creative Writing. Only items approved by both annotators are accepted, with pass rates ranging from 22.8% to 98.3% across different categories and an inter-annotator agreement (IAA) of 0.77.

3. **Human Revision**: To ensure adequate data for underrepresented or low-approval categories, we conduct human revision to ensure sufficient numbers for experiments. In this stage, each question is revised or answered by an annotator. Then the answer undergoes the same process as in stage 2, with two additional reviewers determining its validity.

All the hired annotators are native Chinese speakers, hold a bachelor's degree or higher, and dedicate over 1,000 labor hours to annotation. To meet the experimental requirements in Section 4, the first round of annotation produces 1,000 training data, 100 validation data, and 100 test data for each ability. We then expand the training set to 40k to compare different construction strategies in Section 5. The data sources and sizes for each ability category are outlined in Table 1, with examples provided in Appendix A.2.

4 Experiments

Employing the human-curated dataset proposed in Section 3, we study the abilities' development in response to alterations in data volume, parameter size, and construction methods. Experiments are conducted under both in-domain and out-of-domain conditions. This section outlines the process of model training, evaluation, and results analysis.

4.1 Experiment Setup

For quantity-based experiments, we uniformly sample data d_i of size n from each ability a_i within the ten categories $A = \{a_1, a_2, ..., a_{10}\}$ in our training set. The samples, combined as $D = \bigcup_{i=1}^{10}\{d_i\}$, are utilized for each model training. We increment the sample size from $n = 1$ logarithmically (base 4) to $n = 1000$ (totaling 10k instances). Regarding parameter sizes, we train models across a full range of 7b, 13b, and 33b scales. Each training session spans at least 15 epochs, with the corresponding checkpoint saved for evaluation after each epoch. We also compare our human-curated dataset, *DoIT*, with a synthetic dataset proposed by Peng et al. (2023) for instruction tuning. The synthetic dataset utilizes the Alpaca (Taori et al., 2023) instruction pool, created through the Self-Instruct (Wang et al., 2022b) framework, with responses generated by GPT-4 (OpenAI, 2023). By leveraging the cost-effectiveness of synthetic data to acquire a large and diverse set of instances, we can expand our experimental data volume to 41k on this dataset.

Taking into account all these factors, our study requires nearly 500 model checkpoints to draw systematic conclusions. To ensure the generalizability of our findings, we first analyze the scaling properties of different capabilities using the Chinese-LLaMA model (Cui et al., 2023), which maintains the straightforward architecture of LLaMA (Touvron et al., 2023) without any modifications. Subsequently, we employ more sophisticated foundation models such as Qwen1.5 (Bai et al., 2023) and Baichuan2 (Yang et al., 2023) to further validate how our insights can enhance model training. Detailed hyperparameter choices and training procedures are provided in Appendix A.1.

4.2 Evaluation

Selecting the optimal checkpoint for instruction-tuning is non-trivial. Prior studies (Ouyang et al., 2022; Zhou et al., 2023) note that training for more epochs can enhance the model's capabilities despite the risk of overfitting, and usually employ humans for evaluation. In contrast, automated evaluation is a more scalable solution but has long-lasting concerns about reliability in both statistical (Papineni et al., 2002; Lin, 2004; Banerjee and Lavie, 2005) and LLM-based (Luo et al., 2023; Shen et al., 2023; Chiang and Lee, 2023) metrics. Therefore, to efficiently and accurately scale the evaluation across hundreds of checkpoints, we employ a semi-automated approach to reduce the burden on human annotators.

There are two types of questions in our dataset that correspond to distinct evaluation approaches:

- Exact-match questions (e.g., multiple-choice, true/false, fill-in-the-blank) have one exclusive gold answer. Similar to other public benchmarks (Hendrycks et al., 2020; Li et al., 2023; Huang et al., 2023; Zhong et al., 2023), we automatically compute the accuracy by comparing generated answers to the ground truth.

- Open-ended questions, common in creative writing, role-play chat, and code generation abilities, lack standard answers. We introduce a semi-automated *comparison with distractors* method for these. This method creates distractors (examples shown in Appendix 9 and 10) by manually corrupting each ground truth in two ways: **Fine-grained corruption** subtly alter some numbers, operators, and terminologies to test the models' performance in modeling details. **Coarse-grained corruption** creates a distractor that disregards the given instruction but is textually error-free, testing the model's instruction understanding and adherence. A question scores 1 if the language modeling of ground truth g given the instruction i has a lower perplexity (PPL) than any distractor d_j, otherwise 0:

$$PPL(g|i) = e^{-\sum_{t=1}^{T} \log p(g_t|i, g_{<t})}$$

, t denotes the time series of tokens

$$Score = \begin{cases} 1, & \text{if } \min_j(PPL(d_j|i)) > PPL(g|i) \\ 0, & \text{otherwise} \end{cases}$$

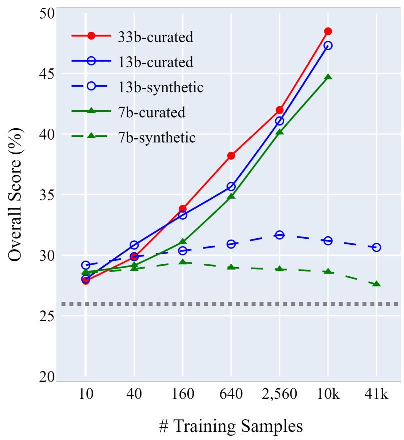


Figure 1: The impact of data volume, parameter scale, and construction method on the overall performance.

As outlined in Sections 3 and 4.1, we train 15 checkpoints for each factor setting and reserve 100 instances in both validation and test sets for evaluation. We select the highest-scoring checkpoint after 5 epochs using the validation set and then demonstrate its performance on the test set. Our observations and analysis are discussed in the next subsection.

4.3 Results and Analysis

We analyze the effect of data volume, parameter size, and construction method. Their impact on overall model performance is illustrated in Fig 1, where the x-axis represents changes in data volume and the y-axis represents the average scores across ten in-domain evaluations plus three out-of-domain abilities. Lines of different colors and symbols represent models with different parameter sizes. We also have a grey dotted line representing the score of random guesses. When scaling the number of training instances, there is a substantial discrepancy

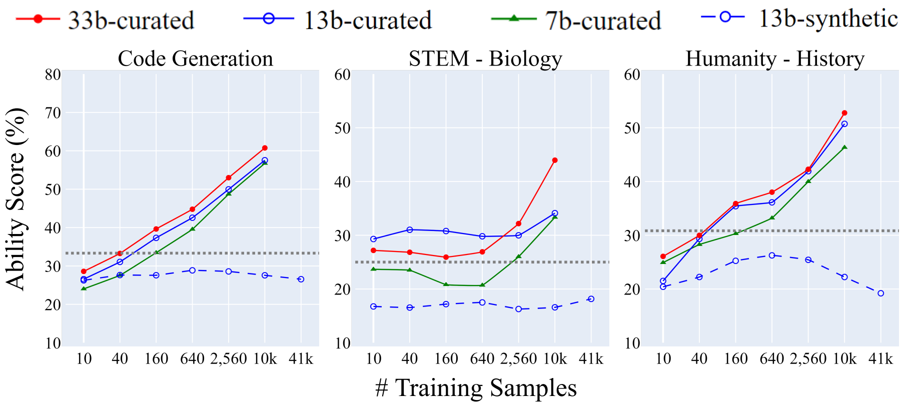


Figure 2: Abilities that are responsive to the data quantity and parameter scale in the human-curated dataset, also comparing the data efficiency of different construction methods with synthetic data.

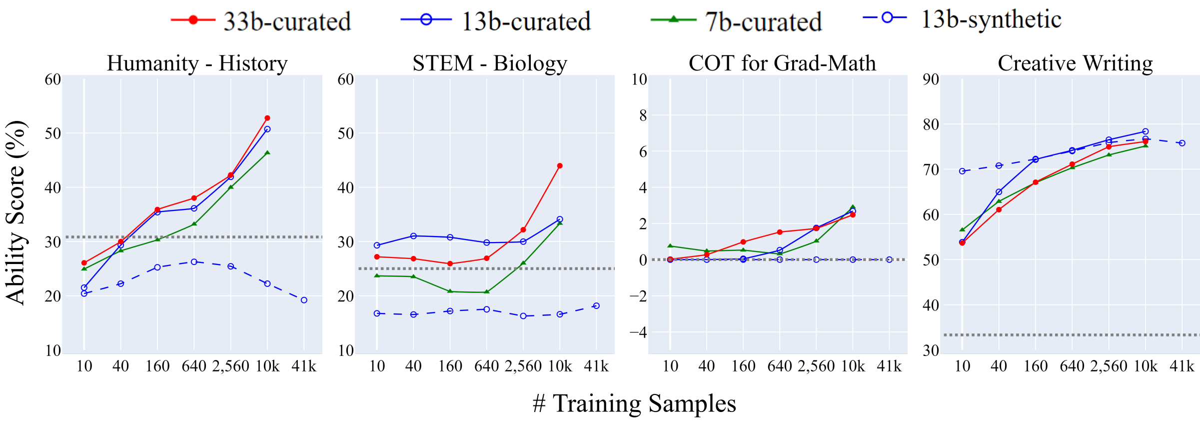


Figure 3: Comparison of abilities with varying sensitivities to data scaling in the human-curated dataset, also comparing the data efficiency of different construction methods with synthetic data.

on the performance of models trained on human-curated data (depicted by solid lines) and synthetic data (depicted by dashed lines).

Moreover, the overall trend is not universally applicable to different abilities when we observe them in the next section. Subsequently, we quantify the scaling sensitivity of each ability by investigating the relationship between its task accuracy and the factors above. We further analyze two interpretable features that potentially forge different scaling sensitivities across these abilities.

4.3.1 Disparities in ability growth trajectories

We present the empirical results for each ability in this section, exhibiting their distinct growth paces when facing factor changes:

Abilities responsive to scaling: Some abilities such as Code Generation, STEM-Biology, and Humanity-History are responsive to factor changes. As illustrated in Fig 2, they show clear upward trends with the growth of data volume and parameter scale.

Varying sensitivities to data scaling: As depicted in Fig 3, the rate of improvement is not uniform across abilities. This figure reveals varying degrees of data scaling sensitivity, with Creative Writing being a notable case. The slope of its curve gradually disappears, indicating a plateau with limited data expansion.

Varying sensitivities to parameter scaling: Fig 4 demonstrates that the impact of parameter size scaling also varies among abilities. From left to right in the figure, their curves for different model sizes become increasingly intertwined, indicating the insensitivity to this change.

Abilities resistant to scaling: As shown in Fig 5, certain abilities like Ethics and Role-play Chat appear to resist both factors and maintain stagnant scores across all changes. This lack of progress implies that supervised fine-tuning (SFT) alone may not effectively advance these abilities, warranting the investigation of approaches beyond it, such as reinforcement learning from human feedback

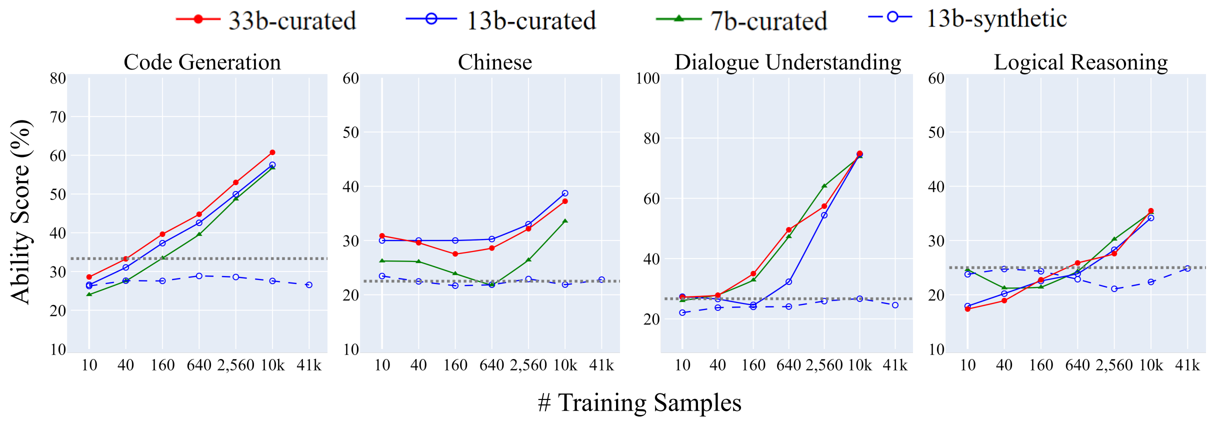


Figure 4: Comparison of abilities with varying sensitivities to parameter size scaling in the human-curated dataset, also comparing the data efficiency of different construction methods with synthetic data.

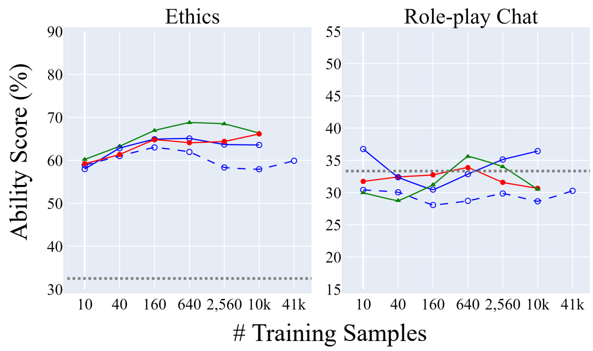


Figure 5: Abilities that are resistant to the data quantity and parameter scale in the human-curated dataset, also comparing the data efficiency of different construction methods with synthetic data.

(RLHF) (Ouyang et al., 2022; Nakano et al., 2021).
Out-of-domain evaluation: Beyond in-domain abilities, Figure 6 shows model performance on three out-of-distribution (OOD) tasks from the C-Eval datasets (Huang et al., 2023): Teacher Qualification, Physician Qualification, and Urban and Rural Planning. The observed growth trends suggest robust cross-ability generalization. Similar to in-domain evaluations, these OOD tasks exhibit diverse responses to variations in data quantity and parameter scale.
Human-curated vs. Synthetic: Figures 2-6 also present the results from models trained on synthetic data generated by GPT-4 (Peng et al., 2023). We evaluate both 7B and 13B models, which yield analogous conclusions. For simplicity, only the results for the 13B model are plotted, with the 7B results included in Appendix 11. Our findings align with previous studies (Gudibande et al., 2023), indicating that synthetic data is only effective for learning partial abilities. Additionally, Figure 1 demonstrates that increasing the volume of synthetic data does not continuously improve model performance. We further empirically demonstrate in Appendix A.3 that even when synthetic data is combined with human-curated data, its effectiveness still has an upper limit. Consequently, in subsequent experiments, we focus solely on exploring the scaling properties of human-curated data.

4.3.2 Understanding Diverse Scaling Behaviors

To understand the varying scaling properties of abilities, a notable observation from Section 4.3.1 is that abilities tied to professional (academic) knowledge are more sensitive to parameter scaling. We define such common feature as **Complexity**, indicating these abilities are inherently *"more challenging for language modeling"* and *"benefit less from the training of other abilities"*. We hypothesize that Complexity is associated with how different abilities respond to model size changes.

To examine the relationship between Complexity and parameter sensitivity, we first quantify the sensitivities of individual abilities. Adopting the scaling law function similar to Kaplan et al. (2020), we model the task score ACC (averaged across varying data volume) as a function of model size N for each ability i:

$$exp(\text{ACC}_i) = (exp(c_i) \cdot N)^{\alpha_i}, c_i \text{ is constant} \quad (1)$$

Here, the exponent α_i represents the rate of accuracy improvement with increasing model size, indicating the *scaling sensitivity*.

$$\text{ACC}_i = \alpha_i \cdot \log(N) + \alpha_i \cdot c_i \quad (2)$$

We further demonstrate that Complexity can be measured in a low-resource setting by fine-tuning

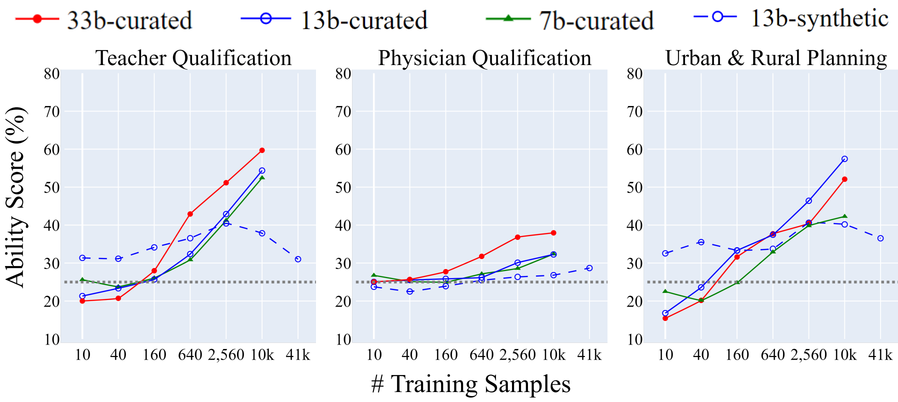


Figure 6: Growth paces of out-of-domain abilities that not included in the human-curated dataset, also comparing the data efficiency of different construction methods with synthetic data.

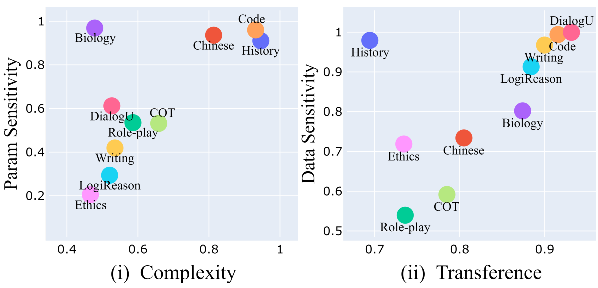


Figure 7: Two interpretable features, (i) Complexity and (ii) Transference, of different abilities demonstrate linear relationships with their sensitivities to scaling in parameter size and data volume. These features can infer the growth of abilities after scale-up, as discussed in Section 5. Sensitivity values are normalized to (0, 1) for visualization.

separate 7b models with only 64 data points per ability. According to its definition, Complexity is calculated as a weighted sum of the test loss (L) for each model trained individually on ability i, along with its accuracy achieved by training on ability j data, $Acc(j, i)$, improved over the foundation model's performance $Acc(f, i)$:

$$\text{Complexity}_i = w_1 \cdot \text{L}_i - w_2 \cdot \sum_{\text{j}\neq\text{i}} (\text{Acc}(j, i) - \text{Acc}(f, i))$$

Results depicted in Figure 7 (i) show a clear linear relationship between Complexity and the sensitivity to parameter scale, indicating that even with minimal resources (7b model size, 64 data points), we can forecast how abilities will develop with increased model size.

Correspondingly, we have also computed the **Transference** for each ability, which reflects *"how well ability i data enhances other abilities j"* via this formula:

$$\text{Transference}_i = w \cdot \sum_{\text{j}\neq\text{i}} (\text{Acc}(i, j) - \text{Acc}(j, j))$$

By substituting model size (N) with data volume (D) in Equation (1), we can also evaluate each ability's sensitivity to data volume. Figure 7 (ii) shows that Transference is linearly related to data scaling sensitivity. This confirms that using checkpoints trained with limited data can also infer the impact of enlarged data volume on ability development.

5 Guidance on Model Training

After understanding that *"abilities react differently to factor changes"* and learning *"how to estimate the sensitivities of different abilities to scaling,"* we delve deeper into whether these varying sensitivities can enhance the efficiency of learning specific or overall model abilities. At this stage, we employ two more advanced foundation models, Qwen1.5(Bai et al., 2023) and Baichuan2(Yang et al., 2023), to conduct our experiments.

5.1 Learning a Specific Ability

When learning a new ability, due to its unique sensitivity to different factors, the efficiency of improvement can vary significantly depending on how resources are allocated between computational power and data annotation. We select two representative abilities to test the effectiveness of model training according to their distinct sensitivities:

Logical Reasoning: As demonstrated in Figure 7, this ability is data-biased, meaning it is more sensitive to data scaling than to parameter scaling.

Novel Generation[3]: This OOD task involves continuing a novel from a given starting point. According to the calculations described in Section 4.3.2, Novel Generation is parameter-biased with a parameter sensitivity of 0.92 and a data sensitivity of 0.54. This indicates that increasing the model's

[3] https://huggingface.co/datasets/zxbsmk/webnovel_cn

	Parameter Scale	Novel Generation		Logical Reasoning	
		Data Size	Score	Data Size	Score
Qwen1.5	7B	2,000	53.0	2,000	26.0
	7B	10,000	52.0	4,000	**37.0**
	14B	2,000	**62.0**	2,000	34.0
Baichuan2	7B	2,000	37.0	2,000	25.0
	7B	10,000	38.0	4,000	**35.0**
	13B	2,000	**57.0**	2,000	**36.0**

Table 2: Parameter-biased ability (Novel Generation) shows significantly greater performance gains from increasing model parameters compared to adding more data. Conversely, data-biased ability (Logical Reasoning) benefits more from additional labeled data, as discussed in Section 5.1.

parameter size is more effective for enhancing this ability than increasing the data volume.

Table 2 illustrates the performance of Qwen1.5 and Baichuan2 when trained with varying amounts of data and parameters, following the evaluation methodology outlined in Section 4. The results show that for the data-biased ability, Logical Reasoning, increasing the data volume from 2000 to 4000 can yield performance that matches or exceeds that of models with 13/14B parameters. This suggests that annotating more data can effectively reduce computational resource requirements. Conversely, for the parameter-biased ability, Novel Generation, even a fivefold increase in data volume fails to match the performance of larger models. Therefore, for such abilities, increasing the model's parameter size is the more effective strategy.

5.2 Learning Comprehensive Abilities

If an ability exhibits low sensitivity to both data and parameter variations, it implies that this ability cannot be effectively developed through instruction tuning. We investigate whether reducing the amount of such insensitive data while increasing the proportion of other, more responsive data can enhance the efficiency of model training.

We evaluate three data-mixing strategies:

Baseline: We use the model trained on 1,000 instances per ability (totaling 10,000 instances) from Section 4 as the baseline.

Reconstruct: Ethics and Role-play Chat, which show low sensitivity to all factor changes in Section 4.3, are reduced to 64 instances each (corresponding to their relatively higher points in Figure 5). The total number of instances is maintained at 10,000 by uniformly increasing the data volume of other categories. Creative Writing remains at 1,000 instances, as Figure 3 indicates that this is sufficient for it to reach its performance plateau.

Models	Data Quantity	AGIEval - 0shot			CMMLU - 0shot		
		CLLaMA	Baichuan2	Qwen1.5	CLLaMA	Baichuan2	Qwen1.5
Baseline	10k	34.64	42.15	69.08	36.75	52.60	72.71
Reconstruct	10k	35.43↑	45.09	69.56↑	36.85↑	53.00↑	73.07↑
Maximum	40k	37.61↑	46.59↑	69.21↑	37.28↑	56.50↑	72.33
		AGIEval - 5shot			CMMLU - 5shot		
		CLLaMA	Baichuan2	Qwen1.5	CLLaMA	Baichuan2	Qwen1.5
Baseline	10k	31.01	47.03	70.12	35.14	54.87	71.97
Reconstruct	10k	32.37↑	48.46↑	70.97↑	35.89↑	55.00↑	71.82
Maximum	40k	33.57↑	53.12↑	70.96↑	37.16↑	58.02↑	71.18

Table 3: Comparing the performance of three data-mixing strategies on two benchmarks, evaluated using checkpoints at epoch 10 with a parameter size of 7b. Scores superior to the baseline are marked with ↑.

Maximum: We further expand the data volume based on the same insights, keeping Ethics and Role-play Chat at 64 instances each and Creative Writing at 1,000 instances. Other abilities are scaled up following the procedures outlined in Section 3, with specific quantities detailed in Table 1. This expansion results in an unbalanced dataset due to the varying difficulty in annotating each ability.

For evaluation, we use two comprehensive benchmarks: AGIEval (Zhong et al., 2023) and CMMLU (Li et al., 2023). AGIEval (Zhong et al., 2023) assesses the general capabilities of LLMs in tasks related to human cognition and problem-solving, focusing specifically on the multiple-choice questions within its three Chinese subsets. CMMLU (Li et al., 2023), akin to MMLU (Hendrycks et al., 2020), evaluates LLMs' knowledge and reasoning capabilities within a Chinese cultural context, spanning 67 diverse subjects from elementary to advanced professional levels.

Models with 7b parameters are trained for each data-mixing strategy and their performance is tested at epoch 10 on both benchmarks under 0-shot and 5-shot settings. Table 3 marks the results that show improvement over the baseline with ↑. In our experiments with three distinct foundation models, both new strategies generally outperform the "Baseline" approach. Specifically, the "Reconstruct" strategy improves AGIEval scores by 1%-3% without increasing the data volume. Moreover, the "Maximum" strategy further enhances performance across all evaluated abilities.

6 Conclusion

This research introduces a novel, human-curated Chinese dataset comprising over 40,000 instruction instances across ten ability categories. This dataset facilitates the investigation of how the growth of large language model (LLM) abilities is influenced by data volume, parameter size, and data construction methods during instruction tuning. We are the first to disentangle the effects of these factors by examining a comprehensive set of over 500 model checkpoints, ranging from 7 billion to 33 billion parameters. Our findings reveal that the impact of increasing data volume and model size varies across different abilities. We identify two key features, Complexity and Transference, which can predict ability growth in low-resource scenarios. Guided by these findings, we enhance the effectiveness of learning specific tasks and achieve better comprehensive abilities on benchmarks such as CMMLU and AGIEval. We provide open access to our codebase, dataset, and model checkpoints to enhance reproducibility and contribute to future research.

References

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Merouane Debbah, Etienne Goffinet, Daniel Heslow, Julien Launay, Quentin Malartic, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. Falcon-40B: an open large language model with state-of-the-art performance.

Stephen H. Bach, Victor Sanh, Zheng Xin Yong, Albert Webson, Colin Raffel, Nihal V. Nayak, Abheesht Sharma, Taewoon Kim, M Saiful Bari, Thibault Févry, Zaid Alyafeai, Manan Dey, Andrea Santilli, Zhiqing Sun, Srulik Ben-David, Canwen Xu, Gunjan Chhablani, Han Wang, Jason Alan Fries, Maged S. Al-shaibani, Shanya Sharma, Urmish Thakker, Khalid Almubarak, Xiangru Tang, Mike Tian-Jian Jiang, and Alexander M. Rush. 2022. Promptsource: An integrated development environment and repository for natural language prompts. *ArXiv*, abs/2202.01279.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Kaijie Zhu, Hao Chen, Linyi Yang, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2023. A survey on evaluation of large language models. *arXiv preprint arXiv:2307.03109*.

Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? *arXiv preprint arXiv:2305.01937*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam M. Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Benton C. Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier García, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Díaz, Orhan Firat, Michele Catasta, Jason Wei, Kathleen S. Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways. *ArXiv*, abs/2204.02311.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Y. Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *CoRR*, abs/2210.11416.

Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. 2023. Free dolly: Introducing the world's first truly open instruction-tuned llm.

Yiming Cui, Ziqing Yang, and Xin Yao. 2023. Efficient and effective text encoding for chinese llama and alpaca. *arXiv preprint arXiv:2304.08177*.

Arnav Gudibande, Eric Wallace, Charlie Snell, Xinyang Geng, Hao Liu, Pieter Abbeel, Sergey Levine, and Dawn Song. 2023. The false promise of imitating proprietary llms. *arXiv preprint arXiv:2305.15717*.

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*.

Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory Diamos, Heewoo Jun, Hassan Kianinejad, Md Mostofa Ali Patwary, Yang Yang, and Yanqi Zhou. 2017. Deep learning scaling is predictable, empirically. arxiv. *arXiv preprint arXiv:1712.00409*.

Or Honovich, Thomas Scialom, Omer Levy, and Timo Schick. 2022. Unnatural instructions: Tuning language models with (almost) no human labor. *arXiv preprint arXiv:2212.09689*.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, Yao Fu, Maosong Sun, and Junxian He. 2023. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *arXiv preprint arXiv:2305.08322*.

Hamish Ivison, Akshita Bhagia, Yizhong Wang, Hannaneh Hajishirzi, and Matthew E. Peters. 2022. Hint: Hypernetwork instruction tuning for efficient zero-shot generalisation. *ArXiv*, abs/2212.10315.

Srinivasan Iyer, Xi Victoria Lin, Ramakanth Pasunuru, Todor Mihaylov, Daniel Simig, Ping Yu, Kurt Shuster, Tianlu Wang, Qing Liu, Punit Singh Koura, Xian Li, Brian O'Horo, Gabriel Pereyra, Jeff Wang, Christopher Dewan, Asli Celikyilmaz, Luke Zettlemoyer, and Ves Stoyanov. 2022. OPT-IML: scaling language model instruction meta learning through the lens of generalization. *CoRR*, abs/2212.12017.

Yunjie Ji, Yong Deng, Yan Gong, Yiping Peng, Qiang Niu, Baochang Ma, and Xiangang Li. 2023. Belle: Be everyone's large language model engine. `https://github.com/LianjiaTech/BELLE`.

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*.

Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi-Rui Tam, Keith Stevens, Abdullah Barhoum, Nguyen Minh Duc, Oliver Stanley, Richárd Nagyfi, et al. 2023. Openassistant conversations–democratizing large language model alignment. *arXiv preprint arXiv:2304.07327*.

Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. 2023. Cmmlu: Measuring massive multitask language understanding in chinese. *arXiv preprint arXiv:2306.09212*.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let's verify step by step. *arXiv preprint arXiv:2305.20050*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Hanmeng Liu, Jian Liu, Leyang Cui, Zhiyang Teng, Nan Duan, Ming Zhou, and Yue Zhang. 2023. Logiqa 2.0—an improved dataset for logical reasoning in natural language understanding. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Zheheng Luo, Qianqian Xie, and Sophia Ananiadou. 2023. Chatgpt as a factual inconsistency evaluator for abstractive text summarization. *arXiv preprint arXiv:2303.15621*.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2021. Cross-task generalization via natural language crowdsourcing instructions. *arXiv preprint arXiv:2104.08773*.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, et al. 2022. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. 2021. Webgpt: Browser-assisted question-answering with human feedback. *ArXiv preprint*, abs/2112.09332.

OpenAI. 2022. Introducing chatgpt. *OpenAI Blog*.

OpenAI. 2023. GPT-4 Technical Report. *arXiv e-prints*, arXiv:2303.08774.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *CoRR*, abs/2203.02155.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. 2021. Multitask prompted training enables zero-shot task generalization. *arXiv preprint arXiv:2110.08207*.

Chenhui Shen, Liying Cheng, Yang You, and Lidong Bing. 2023. Are large language models good evaluators for abstractive summarization? *arXiv preprint arXiv:2305.13091*.

Chiyu Song, Hongliang He, Haofei Yu, Pengfei Fang, Leyang Cui, and Zhenzhong Lan. 2023. Uni-encoder: A fast and accurate response selection paradigm for generation-based dialogue systems. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6231–6244.

Kai Sun, Dian Yu, Dong Yu, and Claire Cardie. 2020. Investigating prior knowledge for challenging chinese machine reading comprehension. *Transactions of the Association for Computational Linguistics*, 8:141–155.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca.

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

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Huai hsin Chi, and Denny Zhou. 2022a. Self-consistency improves chain of thought reasoning in language models. *ArXiv*, abs/2203.11171.

Yizhong Wang, Hamish Ivison, Pradeep Dasigi, Jack Hessel, Tushar Khot, Khyathi Raghavi Chandu, David Wadden, Kelsey MacMillan, Noah A. Smith, Iz Beltagy, and Hanna Hajishirzi. 2023. How far can camels go? exploring the state of instruction tuning on open resources. *ArXiv*, abs/2306.04751.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2022b. Self-instruct: Aligning language model with self generated instructions. *arXiv preprint arXiv:2212.10560*.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Anjana Arunkumar, Arjun Ashok, Arut Selvan Dhanasekaran, Atharva Naik, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Gary Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Maitreya Patel, Kuntal Kumar Pal, M. Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit Verma, Ravsehaj Singh Puri, Rushang Karia, Shailaja Keyur Sampat, Savan Doshi, Siddharth Deepak Mishra, Sujan Reddy, Sumanta Patro, Tanay Dixit, Xudong Shen, Chitta Baral, Yejin Choi, Noah A. Smith, Hanna Hajishirzi, and Daniel Khashabi. 2022c. Super-naturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks. In *Conference on Empirical Methods in Natural Language Processing*.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022a. Finetuned language models are zero-shot learners. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022b. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

Xiang Wei, Xingyu Cui, Ning Cheng, Xiaobin Wang, Xin Zhang, Shen Huang, Pengjun Xie, Jinan Xu, Yufeng Chen, Meishan Zhang, et al. 2023. Zero-shot information extraction via chatting with chatgpt. *arXiv preprint arXiv:2302.10205*.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. 2023a. Wizardlm: Empowering large language models to follow complex instructions.

Canwen Xu, Daya Guo, Nan Duan, and Julian McAuley. 2023b. Baize: An open-source chat model with parameter-efficient tuning on self-chat data. *arXiv preprint arXiv:2304.01196*.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, et al. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

Da Yin, Xiao Liu, Fan Yin, Ming Zhong, Hritik Bansal, Jiawei Han, and Kai-Wei Chang. 2023. Dynosaur: A dynamic growth paradigm for instruction-tuning data curation. *arXiv preprint arXiv:2305.14327*.

Biao Zhang, Zhongtao Liu, Colin Cherry, and Orhan Firat. 2024. When scaling meets llm finetuning: The effect of data, model and finetuning method. *arXiv preprint arXiv:2402.17193*.

Ge Zhang, Yemin Shi, Ruibo Liu, Ruibin Yuan, Yizhi Li, Siwei Dong, Yu Shu, Zhaoqun Li, Zekun Wang, Chenghua Lin, Wen-Fen Huang, and Jie Fu. 2023a. Chinese open instruction generalist: A preliminary release. *ArXiv*, abs/2304.07987.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, et al. 2023b. Instruction tuning for large language models: A survey. *arXiv preprint arXiv:2308.10792*.

Yue Zhang, Leyang Cui, Deng Cai, Xinting Huang, Tao Fang, and Wei Bi. 2023c. Multi-task instruction tuning of llama for specific scenarios: A preliminary study on writing assistance. *ArXiv*, abs/2305.13225.

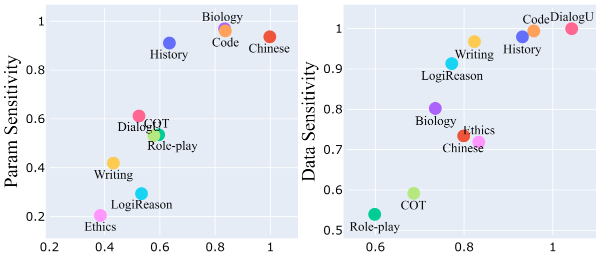
Tony Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *International Conference on Machine Learning*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2023. Agieval: A human-centric benchmark for evaluating foundation models. *arXiv preprint arXiv:2304.06364*.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. 2023. Lima: Less is more for alignment. *arXiv preprint arXiv:2305.11206*.

A Appendix



(i) Complexity (ii) Transference

Figure 8: Two interpretable features, (i) Complexity and (ii) Transference, of different abilities demonstrate linear relationships with their sensitivities to scaling in parameter size and data volume. These features are calculated using 1000 data points instead of 64 in Section 5. Sensitivity values are normalized to (0, 1) for visualization.

A.1 Training Procedures

Our hyperparameter choices are generally in line with Zhou et al. (2023), using AdamW optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.95$ and $weight\_decay = 0.1$. The peak learning rate is set to $1e$-5 and linearly decays to $1e$-6 by the end of training. The batch size is 64, with inputs exceeding 2048 tokens trimmed. Referring to the protocol in Song et al. (2023), we also incorporate two speaker tokens, $<$ user $>$ and $<$ assistant $>$, to segment the utterances from instruction and output for training and inference.

A.2 Examples of each ability category

Ability	Example
Creative writing	**Prompt:** Write a copywriting for the 'Wangshan Hawthorn'. **Response:** Folded hopes' branches, Lightly swaying within hawthorn's embrace from within. A gentle touch on the tongue, Tartness tracing the intricate pathways of taste buds, Unfolding and spreading deep into the recesses of lips and teeth. Having savored the bittersweet fruit, Have you ever witnessed the bloom of its origin? Remember every tangy journey, Never forget those times when the flowers were in full bloom. #GoodHopeWater #Wangshan Hawthorn #AdvertisingCopy
Dialogue Understanding	**Prompt:** Man: What kind of animal do you like? Woman: Elephants. They have long trunks, big ears, and they look so clumsy, which makes them adorable. Man: Elephants aren't clumsy, though. They often help humans with work. Woman: I said they "look" clumsy, you know. Why does the woman like elephants? A. Elephants are big. B. Elephants are clumsy. C. Elephants are adorable. D. Elephants are smart. **Response:** C

Continued on next page

Ability	Example
Role-play Chat	**Prompt:** Ali is the CEO of a young startup company, possessing an innovative mindset and a pursuit of excellence. He is actively seeking professional talent to assist him in realizing his business vision. John, on the other hand, is an experienced HR manager with profound expertise and extensive experience in talent acquisition. Generate a short conversation happening between them. **Response:** Ali: Hey John, thank you for taking the time to chat with me. I'm really interested in learning more about your career experience, and I'm quite impressed by your expertise and experience in HR. John: Thank you for the invitation, Ali. I'm very interested in your company and its business vision. Can you share your perspective on talent recruitment and development within the company? Ali: In the early stages of our startup, we're in great need of visionary and passionate talents. We're looking for individuals who can contribute to the company's growth and help us achieve our vision together. We'd like to better understand how to attract, select, and nurture these talents to ensure the ongoing success of our company. John: Your thoughts align well with our priorities. I believe attracting high-caliber talent is a crucial mission for any startup. There are various ways our company can achieve these goals, such as through recruitment, building reputation, and talent development. However, I think the most critical aspect is having a tailored talent management and development plan to ensure that individuals who fit our company's needs can grow within the organization and continually enhance their skills and knowledge. Ali: That sounds excellent. I think we can delve deeper into this project.

Continued on next page

Ability	Example
STEM - Biology	**Prompt:** Please choose the appropriate option based on the content of the question. Regarding enzymes, which statement is correct? A. Enzymes provide the necessary activation energy for the reaction. B. Changes in enzyme activity are unrelated to changes in the enzyme's environment. C. Changes in enzyme structure can result in the loss of its active site or its entire activity. D. Enzyme molecules are immediately degraded into amino acids after catalyzing a reaction. **Response:** C
Chinese	**Prompt:** Reading the following poem and anwsering the question, Moon Song of Mount Emei by Li Bai On Emei Mountain, a crescent moon graces autumn's sky, Its reflection dances upon the tranquil water of the Pingqiang River nearby. By night, I set forth on the clear stream towards the Three Gorges afar, Thoughts of you linger during my journey, for you are not in sight in Yuzhou. Please describe the scene painted by the lines "On Emei Mountain, a crescent moon graces autumn's sky, Its reflection dances upon the tranquil water of the Pingqiang River nearby" in vivid words. **Response:** Lofty Emei Mountain stands tall and grand, with a crescent autumn moon hanging amidst its peaks, resembling the mountain itself spilling the moonlight. The pristine moonbeam is mirrored on the surface of the tranquil Pingqiang River, which flows gently, carrying the moon's reflection along its course.

Continued on next page

Ability	Example
Ethics	**Prompt:** Some people admire Western culture and believe that Chinese culture is conservative and outdated. How should you respond to them? Drawing on your knowledge, share your perspective. **Response:** 1. Different cultures have different values, and it's not appropriate to judge other cultures solely by Western standards. 2. Despite being relatively conservative, Chinese culture possesses unique charm and profound depth. 3. Rather than excluding other cultures, it's important to maintain an open-minded attitude and learn from, as well as incorporate, the beneficial aspects of other cultures. 4. One should avoid unilaterally idealizing any single culture and, instead, choose and integrate various cultures based on practical circumstances.
Logical Reasoning	**Prompt:** DM advertising, also known as direct mail advertising, involves delivering promotional materials to consumers through mailing, gifting, or other means, directly to their homes or places of business. According to the definition provided, which of the following is not considered as DM advertising? A. Samples of sunscreen lotion given away at cosmetics counters in department stores. B. Flyers promoting a commercial real estate project distributed on crowded streets. C. Coupons for a large chain of fast-food restaurants received in an email inbox. D. Two screen protectors received as a gift when purchasing a mobile phone at an electronics store. **Response:** D

Continued on next page

Ability	Example
COT for Grad-Math	**Prompt:** Solve the inverse matrix of $$\begin{pmatrix} 2 & 3 \\ -1 & 7 \end{pmatrix}$$. Output a zero matrix if the inverse matrix doesn't exists **Response:** To solve the inverse matrix of a 2×2 matrix, we can use the following formula $$\begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} = \frac{1}{ad - bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix},$$ and the condition holds as long as its determinant $ad - bc$ is non-zero. In this example, $2 \times 7 - 3 \times (-1) = 14 + 3 = 17$, which is not zero, thus there exsits an inverse matrix. Apply the formula, and we will have: $$\begin{pmatrix} 2 & 3 \\ -1 & 7 \end{pmatrix}^{-1} = \frac{1}{17} \begin{pmatrix} 7 & -3 \\ 1 & 2 \end{pmatrix}.$$ The final anwser is $$\frac{1}{17} \begin{pmatrix} 7 & -3 \\ 1 & 2 \end{pmatrix}.$$
Humanity - History	**Prompt:** True or False: The three major battles of Liaoshen, Huaihai, and Crossing the Yangtze River essentially annihilated the main forces of the Nationalist Army, accelerating the overall victory of the Liberation War nationwide. ( ) **Response:** False

Continued on next page

Ability	Example
Code Generation	**Prompt:** You are given 2 integers left and right, indicating interval [left, right], return the bitwise AND results of all numbers within this interval (including the left and right endpoints). Example 1: Input: left = 5, right = 7 Output: 4 Example 2: Input: left = 0, right = 0 Output: 0 Example 3: Input: left = 1, right = 2147483647 Output: 0 Hint: 0 <= left <= right <= 231 - 1 Please use c# to finish the problem. **Response:**

```
The following are reference implementation options:
“‘cs
public class Solution {
    public int RangeBitwiseAnd(int left, int right) {
        while (left < right) {
            right &= (right - 1);
        }
        return right;
    }
}
“‘
```

Models	Quantity	AGIEval - 0shot			CMMLU - 0shot		
		ep05	ep10	ep15	ep05	ep10	ep15
Maximum+0	40k+0	36.41	37.61	38.85	**37.74**	37.28	37.53
Maximum+2.56k	40k+2.56k	**37.08**	**39.21**	**39.88**	37.30	**37.74**	**37.74**
Maximum+41k	40k+41k	32.69	34.43	34.38	33.98	36.20	35.34
		AGIEval - 5shot			CMMLU - 5shot		
		ep05	ep10	ep15	ep05	ep10	ep15
Maximum+0	40k+0	33.37	33.57	33.35	**37.02**	**37.16**	**37.13**
Maximum+2.56k	40k+2.56k	**34.11**	**34.07**	**34.00**	36.91	36.87	36.46
Maximum+41k	40k+41k	30.06	31.65	31.41	34.07	35.06	35.17

Table 4: Comparing the performance of three mixing strategies with synthetic data on two benchmarks, evaluated using checkpoints at epochs 5, 10, and 15 with a parameter size of 7b. Highest performance under each setting is in bold.

A.3 Mix up with Syhtnetic Data

Synthetic data is a rich open resource, but Section 4.3 indicates that "*Synthetic data does not consistently enhance model performance with increased volume.*" Investigating the optimal use of synthetic data alongside human-curated data is crucial for practical applications. We thus utilize the "Maximum" construction from Section 5.2 as our baseline and then integrate varying quantities (0, 2560, 41k) of synthetic data to train 7b models.

Table 4 presents the efficacy of three mixing strategies at epochs 5, 10, and 15, evaluated on two benchmarks in both 0-shot and 5-shot scenarios. For AGIEval, a modest addition of 2,560 synthetic instances shows performance gains. In contrast, for CMMLU, peak performance is achieved with no synthetic data or a similar modest addition. Notably, further incorporating 41k synthetic instances does not yield further benefits and instead results in a marked decrease in performance, consistent with our earlier observations in Section 4.

Prompt:
A: A young music enthusiast who is passionate about listening to music by actors and female singers in the pop genre.

B: **An older music enthusiast who enjoys listening to classical and jazz music**.

Generate a snippet of their conversation.

Ground Truth:	**Fine-grained corruption:**	**Coarse-grained corruption:**
[history conversations] A: I see. Do you have any other music recommendations? B: I've been listening to **Billie Holiday's jazz music** lately. Are you interested in giving it a try?	[history conversations] A: I see. Do you have any other music recommendations? B: I've been listening to **Justin Bieber's jazz music** lately. Are you interested in giving it a try?	A: Hello, I'm A. Lately, I've gotten into cooking and I especially enjoy trying out various new recipes. Are you interested in cooking? B: Hello, I'm B. Actually, I'm quite interested in cooking too, especially baking. Recently, I've learned to make a delicious chocolate cake. [folded conversations]

Prompt:
[...]
According to foreign media reports, Microsoft announced that it would lay off 10,000 employees, about **5%** of its total workforce, by the end of March. On January 20th, Google suddenly announced a layoff of 12,000 people, accounting for **6%** of its workforce. On January 26th, The Wall Street Journal reported that IBM had joined the ranks of layoffs and would cut about 3,900 people...
[...]
Write a news article based on the given information.

Ground Truth:	**Fine-grained corruption:**	**Coarse-grained corruption:**
Title: After the Chinese New Year Holiday, Wave of Layoffs Hits - Cherishing Work Becomes the Greatest Sense of Security for Adults Article: [...] According to foreign media reports, Microsoft has announced plans to lay off 10,000 employees by the end of March, accounting for approximately **5%** of its total workforce. On January 20th, Google made a sudden announcement to lay off 12,000 people, representing **6%** of its total workforce. On January 26th, The Wall Street Journal reported that IBM has also joined the layoff trend, planning to cut approximately 3,900 jobs. [...]	Title: The Chinese New Year Holiday Ends, Most Companies Have Ceased Operations Article: [...] According to foreign media reports, Microsoft plans to lay off 1,000 employees by the end of March, accounting for approximately **10%** of its total workforce. On January 20th, Google made a sudden announcement to lay off 12,000 people, representing **60%** of its total workforce. The Wall Street Journal reported on January 26th that IBM has also joined the wave of layoffs, intending to terminate approximately 3,900 employees. [...]	According to the latest reports, several cities in China have recently experienced snow disasters and cold wave weather, causing significant disruptions to people's lives and travel. It has been reported that heavy snowfall has led to the closure of many major transportation routes, resulting in a large number of vehicles stranded on the roads and many people trapped in traffic jams. Additionally, power outages in many cities due to the snow disaster have inconvenienced residents' daily lives.

Figure 9: Examples of distractors derived from manually corrupted Role-play Chat and Creative-writing output.

Prompt:

Given a binary array "nums", you need to remove one element from it. Return the length of the longest non-empty subarray containing only "1" in the resulting array. If there is no such subarray, return "0". [folded examples] Please use python3 to finish it.

Ground Truth:

```
The following is a reference implementation for your consideration:
"'python
class Solution:
     def longestSubarray(self, nums: List[int]) -> int:
          n = len(nums)
          left = [0] * n
          right = [0] * n
          for i in range(1, n):
               if nums[i - 1] == 1:
                    left[i] = left[i - 1] + 1
          for i in range(n - 2, -1, -1):
               if nums[i + 1] == 1:
                    right[i] = right[i + 1] + 1
          return max(a + b for a, b in zip(left, right))
"'
```

Fine-grained corruption:

```
The following is a reference implementation for your consideration:
"'python
class Solution:
     def longestSubarray(self, nums: List[int]) -> int:
          n = len(nums)
          left = [0] * n
          right = [0] * n
          for i in range(1, n):
               if nums[i - 1] == 1:
                    left[i] = left[i - 1] + 1
          for i in range(-1, n-2, 1):
               if nums[i + 1] == 1:
                    right[i] = right[i + 1] + 1
          return max(a + b for a, b in zip(left, right))
"'
```

Coarse-grained corruption:

```
The following is a reference implementation for your consideration:
"'python
def find_max_sum(nums):
     max_sum = float('-inf')
     curr_sum = 0
     for num in nums:
          curr_sum += num
          if curr_sum > max_sum:
               max_sum = curr_sum
          if curr_sum < 0:
               curr_sum = 0
     return max_sum

def find_longest_subarray(nums):
     n = len(nums)
     max_len = 0
     for i in range(n):
          for j in range(i, n):
               subarray = nums[i:j+1]
               if find_max_sum(subarray) == len(subarray):
                    max_len = max(max_len, len(subarray))
     return max_len
"'
```

Figure 10: Examples of distractors derived from manually corrupted Code Generation output.

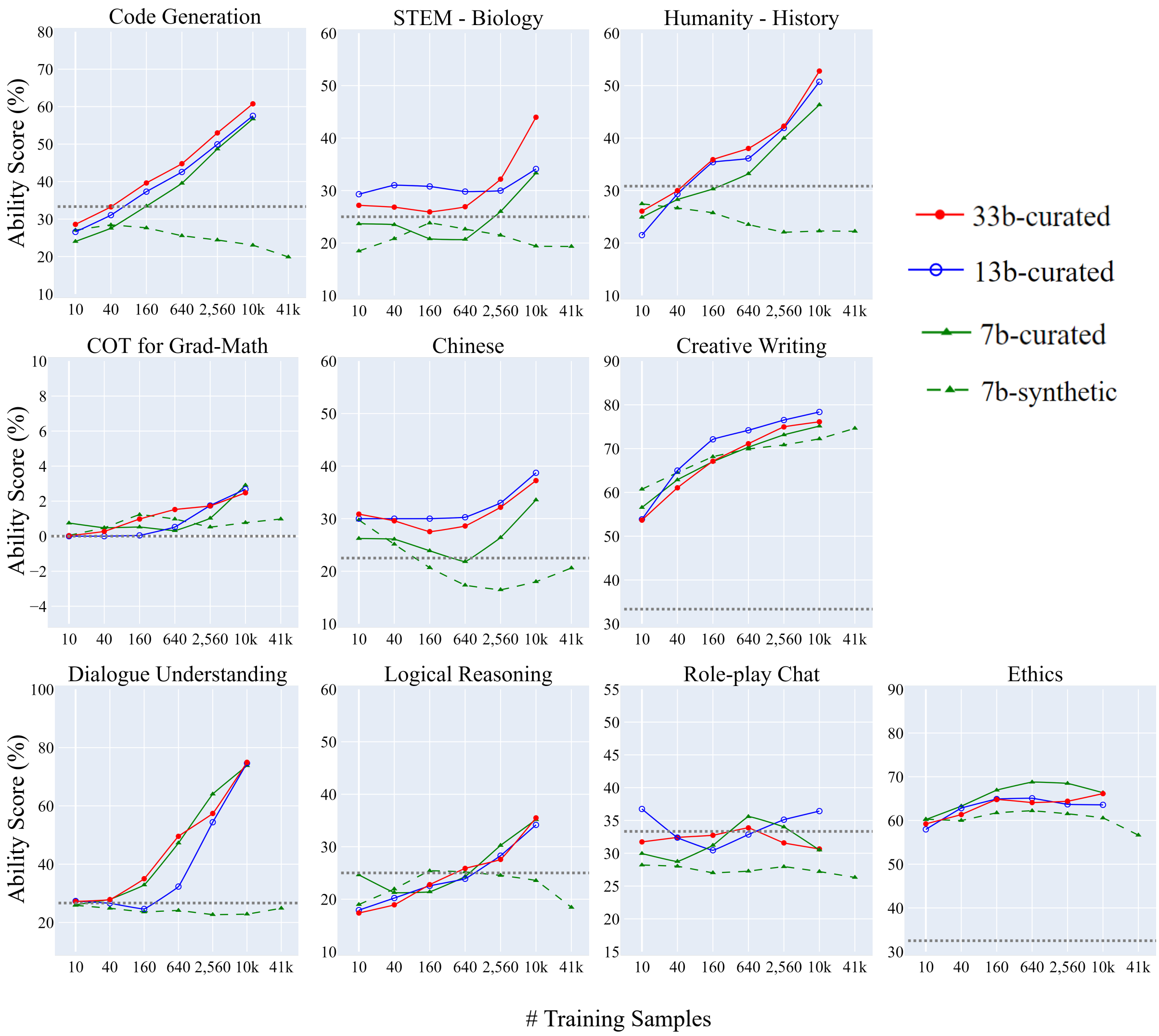


Figure 11: Evaluations of 7b models trained on synthetic data, yielding analogous conclusions as 13b models.