InfiR: Crafting Effective Small Language Models and Multimodal Small Language Models in Reasoning

Congkai Xie¹, Shuo Cai⁴, Wenjun Wang⁴, Pengxiang Li⁶, Zhijie Sang¹, Kejing Yang¹, Yiming Zhang², Zhen Li^{1,2}, Guanghao Zhu⁷, Zeyu Liu⁵, Yang Yu⁸, Yuhang Liu³, Su Lu², Baoyi He³, Qi Zhou⁵, Xiaotian Han⁹, Jianbo Yuan¹⁰, Shengyu Zhang³, Fei Wu³, Hongxia Yang^{1,2}

Reallm Labs, ² The Hong Kong Polytechnic University, ³ Zhejiang University,
 South China University of Technology, ⁵ Harbin Institute of Technology,

⁶ Dalian University of Technology, ⁷ University of Electronic Science and Technology of China,

⁸ Beijing University of Posts and Telecommunications, ⁹ TikTok, ¹⁰ Amazon

* Corresponding authors.

Abstract

Large Language Models (LLMs) and Multimodal Large Language Models (MLLMs) have made significant advancements in reasoning capabilities. However, they still face challenges such as high computational demands and privacy concerns. This paper focuses on developing efficient Small Language Models (SLMs) and Multimodal Small Language Models (MSLMs) that retain competitive reasoning abilities. We introduce a novel training pipeline that enhances reasoning capabilities and facilitates deployment on edge devices, achieving state-of-the-art performance while minimizing development costs. InfiR aims to advance AI systems by improving reasoning, reducing adoption barriers, and addressing privacy concerns through smaller model sizes. Resources are available at https://github. com/Reallm-Labs/InfiR.

1 Introduction

In recent years, Large Language Models (LLMs) (Touvron et al., 2023; DeepSeek-AI, 2024a,b; Qwen, 2023, 2024; Yang et al., 2024) and Multimodal Large Language Models (MLLMs) (Wang et al., 2024; Chen et al., 2024) have shown remarkable advancements in reasoning capabilities. However, these models, often consisting of hundreds of billions of parameters, pose significant challenges related to computational resources, deployment costs, and environmental impact. Their training and development demand substantial infrastructure investments, making them accessible primarily to major technology corporations. Additionally, these models usually require cloud deployment, which raises privacy concerns regarding user data.

Small models, typically comprising fewer than 2 billion parameters, have emerged as a promising solution to these challenges. These models strive to achieve an optimal balance between performance and efficiency. Their considerably lower training

and development costs increase accessibility for researchers and organizations. There is a key challenge for SLMs and MSLMs in enhancing reasoning capabilities while reducing the size of the model.

hongxia.yang@polyu.edu.hk

This paper focuses on: (1) developing efficient small language models with competitive reasoning capabilities and (2) extending these models to handle multimodal inputs while managing real-world operational system tasks. We explore novel training pipelines that enable these compact models to perform reasoning tasks while being deployable on edge devices. Our contributions to the field include the following:

- We propose a pre- and post-training pipeline for small models that enhances reasoning capabilities, completing training in under 6000 GPU hours.
- Our InfiR-1B-Base and InfiR-1B-Instruct models achieve state-of-the-art performance at the 1B parameter scale, with reasoningrelated average score improvements of 2.26x and 1.33x over Llama3.2-1B-Base and Llama3.2-1B-Instruct.
- Our InfiR-VL-1.6B model achieves the stateof-the-art performance in the Android World scenario, with an accuracy of 28% increment compared to the best SOTA among small models.

Through this research, we aim to advance the development of practical, efficient AI systems capable of performing reasoning tasks, while reducing barriers to AI system adoption and addressing user privacy requirements through model size reduction.

2 Small Language Model Pre-training

2.1 Pre-training Data

To construct effective SLMs with strong reasoning capabilities, high-quality pre-training data are es-

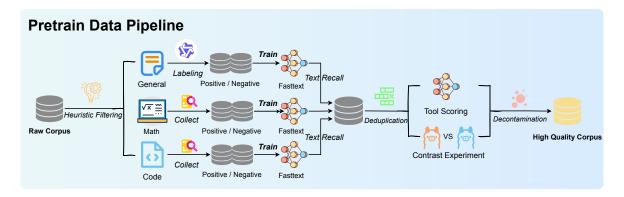


Figure 1: The pipeline of pretrain data drocesses: heuristic filtering, reasoning-oriented text recall, deduplication, quality assessment and decontamination. Comparative experiments on LLaMA3.2-1B with differently cleaned datasets validate the significance of data quality.

sential. We present a comprehensive pipeline for collecting and processing such data.

2.1.1 Data Collection

To develop SLMs with enhanced reasoning capabilities, we constructed a comprehensive dataset comprising two primary components: source code data and reasoning-oriented text data. This approach is designed to capture both programmatic logic and natural language reasoning patterns. For the source code component, we curated the dataset by combining several high-quality open-source resources, including The Stack v2 (Lozhkov et al., 2024b) and Starcoder (Li et al., 2023) and additional Pythoncentric repositories from GitHub. More details of the distribution of programming languages can be found in Figure 4a. To acquire reasoning-oriented text data, we assembled a diverse corpus encompassing web content, academic literature, books, and encyclopedic sources. This corpus was specifically curated to emphasize content that shows logical reasoning, analytical thinking, and structured argumentation.

2.1.2 Data Pipeline

From a technical perspective, as the model size decreases, its information storage capacity diminishes proportionally. To develop a SLM specifically targeted for reasoning tasks, it is essential to construct a sophisticated data pipeline that can effectively filter out noise while preserving reasoning-related information. The workflow of the entire pipeline architecture is depicted in Figure 1.

Heuristic filtering In the initial phase, we employ heuristic filters to perform preliminary filtering on the raw corpus, thereby reducing noise in the data. We employed heuristic filters from

FineWeb (Penedo et al., 2024) to extract high-quality text from web pages. For code data, we selectively filtered and processed repositories written in widely used languages such as Python, JavaScript, Java, and C to ensure relevance and quality. Rule-based filters (Huang et al., 2024b) were applied to remove files that were contaminated or significantly deviated from standard patterns.

Reasoning-oriented text recall Reasoningoriented data, unlike code, lack explicit patterns that can be identified through rule-based ap-Instead, model-based methods are proaches. typically required to extract potential reasoning relationships from large volumes of text. Our reasoning-oriented data retrieval strategy is categorized into three components. First, we prioritize the retrieval of mathematics-related samples, as mathematical texts often exhibit strong reasoning characteristics. Second, we recognize that a significant portion of code-related content in internet data contains reasoning elements. Finally, we aim to retrieve reasoning-related texts from various other domains.

We developed a standard text recall data pipeline, drawing inspiration from DeepSeekmath (Shao et al., 2024) and OpenCoder (Huang et al., 2024a). The pipeline begins with establishing seed datasets. For mathematics-related content, we utilize high-quality mathematical datasets such as OpenWebMath (Paster et al., 2023) and InfiMM-WebMath (Han et al., 2024) as seed data. For coderelated text, we employ StackOverflow as the seed dataset. For other domains, we utilize Qwen2.5-7B-Instruct (Yang et al., 2024) to annotate text URLs and titles, while also incorporating responses syn-

the sized by large language models from datasets like Infinity Instruct as seed data.

Following seed data acquisition, we train domainspecific fasttext models using positive samples from the seed data and random negative samples from web pages and books. These fasttext models are then used to recall relevant content from the remaining corpus.

Deduplication An excessive amount of homogeneous data can be detrimental to small language models. therefore, it is crucial to ensure data diversity. To optimize data efficiency while maintaining semantic diversity, we implemented a global Min-Hash algorithm to efficiently detect and eliminate near-duplicate documents.

Quality assessment To ensure comprehensive data quality across diverse domains, we establish a two-step evaluation framework. First, we employ a domain-specific quality assessment tool: (1) For web content, we leverage FineWeb-eduscorer (Penedo et al., 2024) to evaluate document quality through multiple dimensions; (2) For mathematical content, we utilize a model-based classifier following (Lozhkov et al., 2024a) that scores reasoning and deduction capabilities on a 1-5 scale, filtering to retain only high-quality samples; (3) For code data quality validation, we utilize a static analysis tool to validate syntactic correctness and identify potential structural issues. We also perform comprehensive ablation experiments through continued pre-training on LLaMA 3.2-1B, comparing the performance between models trained on raw and filtered datasets. The model trained on filtered data exhibits better performance across multiple code-related benchmarks compared to its counterpart trained on the raw dataset. Moreover, the filtering process facilitates accelerated convergence, enabling the model to achieve desired performance metrics with fewer training steps.

Decontamination To ensure fairness of comparison, we implemented a token-level 10-gram decontamination algorithm to remove potentially contaminated content in standard benchmarks.

2.2 Annealing Data

After training on 900 billion tokens, we aimed to further enhance reasoning abilities and bridge the gap between the pretraining and supervised finetuning stages. To achieve this, we constructed an annealing dataset comprising additional high-quality reasoning data and continued training the

checkpoint from the previous stage on 40 billion annealing data to obtain the final base model.

Original data During the annealing phase, we maintained the original proportion of source code. For reasoning-oriented text, we removed most web page data, retaining only a small portion related to mathematics and code.

Open source annealing data We collected annealing data used by Dolmino (OLMo, 2024) and OpenCoder. We also include a high-quality coderelated datasets, incorporating training samples from APPS (Hendrycks et al., 2021a) and Code Contest (Li et al., 2022). For each programming problem across the above datasets, we retain a single solution and prioritize Python implementations. These high-quality datasets were utilized in comparative experiments, which demonstrated a significant improvement in the base model's performance on reasoning benchmarks in few-shot settings.

Synthetic data We further leveraged a large language model to generate a batch of synthetic data, implementing stringent selection mechanisms to ensure high quality. For reasoning-oriented data, we employed a reward model with rejection sampling to enhance logical coherence and correctness. For code-related data, we validated its syntax and functionality within a sandbox environment, ensuring robustness and reliability.

2.3 Offline Evaluations for Data-mixing2.3.1 Pre-training Stage

During the pre-training phase, the model compresses knowledge into its parameters. We prioritize the model's **recall** capability, primarily using Negative Log Likelihood (NLL) to assess performance. NLL evaluations are conducted on two types of data:

Sample from webpage During pretraining, we utilize a large amount of web page data. To measure model convergence, we sample a validation set from the web data. Although the web data has been cleaned, there might still be meaningless text in the samples. Therefore, we use existing large language models to calculate the NLL of the text and rank them separately. When a text ranks in the bottom 20% across all rankings, we consider it meaningless and discard it. The NLL for both types is defined as:

$$NLL = -\sum_{i=1}^{n} \log P(t_i|t_{< i}), \tag{1}$$

where t_i represents the i-th token in the webpage text sequence, and $P(t_i|t_{< i})$ is the model's predicted probability for that token given the previous tokens.

Downstream Task Benchmark We utilize the NLL per token on the correct answers from downstream benchmarks as our metric across tasks. For multiple-choice tasks, we normalize the probabilities assigned to all options to adjust the NLL, creating a metric that both correlates with accuracy improvements and shows stable improvement as model scale increases. the normalized probability is:

$$P_{\text{normalized}}(c_i) = \frac{e^{s_i}}{\sum_{j=1}^k e^{s_j}},$$
 (2)

where s_i is the model's score for choice i, and k is the number of choices.

$$ppl_p(t_{1:n}) = \exp(\frac{1}{n} \sum_{i=1}^n \log \frac{1}{p(t_i|t_{1:i-1})})$$
 (3)

2.3.2 Annealing Stage

During the annealing stage, the model rapidly converges due to the decay of the learning rate. **Precision** is prioritized over **recall**. To evaluate the model's performance, we use a downstream benchmark for few-shot generation instead of the NLL method.

We can suppose a contaminated distribution r is introduced into the original clean distribution p with a probability of ϵ , leading to a new distribution q, which satisfies:

$$q(t_i|t_{1:i-1}) = (1 - \epsilon)p(t_i|t_{1:i-1}) + \epsilon r(t_i|t_{1:i-1}).$$
 (4)

By computing the ppl for the new distribution q, than we have:

$$ppl_q(t_{1:n}) = \exp(\frac{1}{n} \sum_{i=1}^n \log \frac{1}{q(t_i|t_{1:i-1})})$$
 (5)

Since $q(t_i|t_{1:i-1}) \ge (1 - \epsilon)p(t_i|t_{1:i-1})$, we can obtain

$$ppl_q(t_{1:n}) \le \exp\left(\frac{1}{n} \sum_{i=1}^n \log \frac{1}{(1-\epsilon)p(t_i|t_{1:i-1})}\right)$$

$$\le \frac{1}{(1-\epsilon)} ppl_p(t_{1:n})$$

$$\approx (1+\epsilon) ppl_p(t_{1:n}). \tag{6}$$

The approximation holds when ϵ is small enough. This indicates that if 5% contaminated distribution is introduced, the perplexity increases no more than 5%. However, this could lead to very poor results in generation for the language model, since the model may produce a corrupted token approximately every 20 tokens on average. Therefore, during the annealing stage, we evaluate the model directly using a downstream benchmark in a few-shot setting to determine a data mixing ratio that optimally balances various capabilities.

2.4 Training Details

Ultimately, we curated a high-quality reasoningrelated dataset consisting of approximately 900 billion tokens for pretraining and 40 billion tokens for annealing. Details of the dataset's composition are provided in the Table 7 and Table 8. Based on LLaMA-3.2 1B architecture, we trained our model in two phases. During the pretraining phase, we trained the model on the curated 900 billion tokens for one epoch, followed by an annealing phase on an additional 40 billion tokens for one epoch. We employed a learning rate of 1.4e-3 with a size of 2048 and a sequence length of 4096 tokens throughout the training process. Training was conducted using NVIDIA NeMo (Kuchaiev et al., 2019) with distributed optimization and DDP gradient overlap on a cluster of 64 H800 GPUs over a total of 90 hours, equating to 5760 GPU hours.

3 Small Language Model Post-training

During the post-training phase of the language model, we leverage Supervised Fine-Tuning (SFT) with meticulously curated datasets to enhance instruction-following and reasoning capabilities. The deliberate construction of SFT data, which accounts for the balance across diverse domains, facilitates enhancements in controllability and robustness, ensuring optimal performance in various applications.

3.1 Supervised Fine-tuning Data

The quality and diversity of data play a crucial role in supervised fine-tuning. We utilize high-quality, publicly available datasets involving instruction-following, reasoning, and code-related tasks, including Infinity-Instruct (BAAI, 2024), Orca-AgentInstruct-1M-v1 (Mitra et al., 2024a), ScaleQuest (Ding et al., 2024), NuminaMath (Li et al., 2024), and OPC-SFT-Stage2 (Huang et al., 2024b). In addition, we developed a data synthesis pipeline to generate high-quality SFT data, incor-

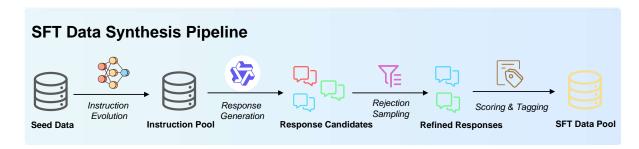


Figure 2: Supervised fine-tuning data synthesis pipeline. The pipeline initiates with a set of high-quality seed data, which is augmented through instruction evolution. Response candidates are generated using the Qwen-2.5-32B-Instruct model, followed by rejection sampling with a reward model and sandbox environment. Finally, we score the curated data for quality and difficulty, and assign domain labels.

porating components such as instruction evolution, rejection sampling, etc., as illustrated in Figure 2.

We curated a high-quality set of instructions from publicly available datasets as seed data and employed a large language model for instruction evolution, generating more diverse and complex instructions. Subsequently, we utilized the Qwen-2.5-32B-Instruct model to generate responses corresponding to these instructions. We encouraged the model to engage in a "step-by-step" reasoning process prior to delivering the final answer, which promoted a more rigorous and logically consistent output.

Rejection sampling For each instruction, we generate multiple responses and employ rejection sampling to ensure the quality of the data. For logic and mathematical reasoning data, we utilize a reward model to assess responses and select the one with the highest score. For code data, we perform execution-based code verification within a sandbox environment to ensure the correctness and functionality of the generated code.

Diversity and quality We assign domain labels to each training sample and perform diversity sampling from the extensive dataset to balance the distribution of data across different domains. Using heuristic rules and LLM-based scoring judgments, we filter out low-quality instructions and responses. Additionally, we use a difficulty scoring model to assess mathematical reasoning data, maintaining a balanced distribution across varying levels of difficulty. Through the combination of diverse and high-quality data sources, instruction augmentation, rejection sampling, and rigorous quality control, we ensure that the training data is both comprehensive and reliable. These carefully designed methodologies contribute to enhancing the model's

reasoning capabilities, enabling it to better handle complex tasks and exhibit improved performance across a wide range of problem domains.

3.2 Training Details

We employed the Llama 3 chat template and applied supervised fine-tuning using a standard crossentropy loss on a few million samples, while masking the loss on the prompt tokens. We fine-tuned the model for 4 epochs with a learning rate of 2e-5 and a batch size of 128. We utilized the cosine learning rate scheduler with a warm-up ratio of 0.1 to optimize training process. The model was fine-tuned on the Llama 3 template with a maximum sequence length of 4096 tokens.

4 Small Multimodal Language Model Training

4.1 Multimodal Pre-training

4.1.1 Data Collection

To systematically develop multimodal reasoning capabilities, we implement a hierarchical data collection strategy that progressively builds from fundamental visual-language alignment to complex reasoning tasks. At the foundation level, we collect data from 11 task domains including captioning, general QA, and OCR to train the MLP projector while keeping the vision encoder and language model frozen, establishing the essential mapping between visual features and LLM's textual semantic space. In addition, we enhance the model's text comprehension abilities by synthetic rendering text data, converting documents, code snippets, and other textual content into visual formats.

4.1.2 Data Cleaning

In order to reduce the impact of low-quality, low-similarity image-text pairs in the dataset on the model's ability to understand the semantics of images and texts, we use the Vista model (Zhou et al.,

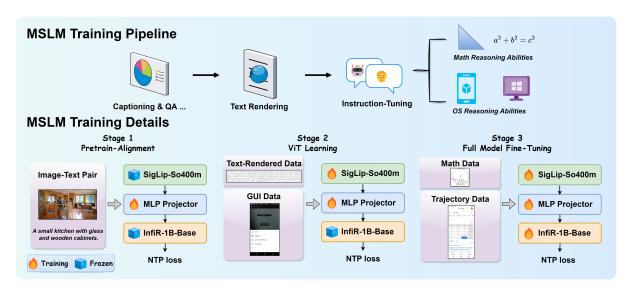


Figure 3: Illustration of the MSLM training pipeline and the MSLM training details, showcasing the progression from captioning and QA tasks to text rendering, followed by instruction-tuning, culminating in enhanced mathematical and operating system reasoning abilities.

2024) to extract the embeddings of the data pairs, calculate the similarity by pair, and then select a reasonable similarity threshold to filter out low-quality, low-similarity image-text pairs. The details is show in Appendix E.

4.2 Multimodal Instruction-Tuning

4.2.1 General Vision Reasoning Abilities

As we did in the SFT of SLMs, during the SFT stage of MSLMs, we collected multiple open-source instruction-tuning datasets spanning 14 domains, including general QA, mathematics, charts, OCR, documents, science, medical, GUI, code, etc., while covering multiple modalities such as single-image, multi-image, and text, as shown in Table 9, and synthesized more high-quality vision reasoning data through multiple methods such as text rendering and virtual environment, so as to enable the model to have comprehensive general vision reasoning capabilities for image-related problems in the real world and strengthen the effective semantic alignment between visual input and SLM.

4.2.2 Multimodal Math Reasoning Abilities

After the initial stimulation of general vision reasoning ability, we fine-tuned SLM in different subdivisions for different scenarios to enhance the ability of MSLM in the corresponding fields. In mathematics, we used multiple data sets such as InfiMM-WebMath-40B (Han et al., 2024), MathVista (Lu et al., 2024), and MM-Math (Sun et al., 2024) as bases, and used ChatGPT to synthesize multimodal math question-answering CoT-PoT data; in addition, we also used the text rendering mechanism

to mix in some pure text math question-answering Instruction tuning data.

4.2.3 Operator-System Reasoning Abilities

Our MSLM's operator-system reasoning capabilities are developed through a two-stage supervised fine-tuning framework that progressively builds from fundamental understanding to advanced reasoning. (Liu et al., 2025) The foundation stage establishes essential GUI comprehension by utilizing diverse datasets spanning GUI understanding, grounding, and question-answering tasks. We standardize coordinate systems to a [0,1000] scale and implement a reference-augmented annotation format, enabling precise spatial reasoning while maintaining natural language flow in the model's responses. Building upon this foundation, with existing trajectory data, we synthesize 45K training samples that incorporate these advanced reasoning patterns, enabling our MSLM to handle complex GUI interactions through structured reasoning rather than mere pattern matching.

4.3 Training Details

As shown in Fig 3, our proposed Multimodal Small Language Model Architecture integrates a vision encoder with a language model. For the vision encoder, we utilize SigLip-So400m (Zhai et al., 2023), which is built upon the SoViT-400M. The language model component is based on our InfiR-1B-Base, which provides reasoning capabilities. To align the visual features with the language model's latent space, we employ a simple MLP layer as the visual projector.

During the Pretraining stage, we exclusively train the MLP projector with the aim of aligning the backbones of the ViT and LLM. In this phase, we utilize a large dataset of captions.

In the SFT stage, the training is divided into two sub-stages. The first sub-stage focuses on enhancing the model's capabilities with images. Here, we unfreeze the parameters of the ViT and train both the ViT and the adapter. We employ a substantial amount of text rendering synthetic samples and GUI samples to ensure the model develops initial vision reasoning capabilities. Additionally, we retain some samples from the pretraining phase to maintain the model's general captioning abilities.

In the second SFT sub-stage, we unfreeze all parameters and train on the most challenging samples, such as trajectory data in operating system scenarios and tool usage samples in mathematical contexts. This step is designed to enhance the model's planning and reasoning abilities.

5 Experimental Results

5.1 Small Language Models

5.1.1 Benchmark and Quantitative Results

Benchmarks We evaluate our model's capabilities across multiple dimensions using established benchmarks: MMLU (Hendrycks et al., 2021b) for general reasoning and knowledge spanning 57 subjects, HumanEval (Peng et al., 2024) and MBPP (Austin et al., 2021) for code generation and understanding in Python programming tasks, GSM8K (Cobbe et al., 2021) for grade schoollevel word problems, and MATH (Hendrycks et al., 2021c) for advanced mathematical reasoning covering algebra, geometry, and calculus.

Pre-trained model We evaluated the performance of SLMs using a few-shot setting. Table 1 compares the performance of InfiR-1B-Base with state-of-the-art open-source models. InfiR-1B-Base outperforms Llama3.2-1B and is comparable to Owen2.5-1.5B.

Post-trained model We evaluated the performance of InfiR-1B-Instruct on the MMLU, GSM8K, MATH, HumanEval, and MBPP benchmarks using zero-shot prompts, as shown in Table 2. InfiR-1B-Instruct outperforms Llama-3.2-1B-Instruct on various reasoning tasks, and shows significant improvements in mathematical reasoning and coding tasks. Specifically, InfiR-1B-Instruct exceeds Llama-3.2-1B-Instruct by more than 26 points on GSM8K and 16 points on MATH, while surpassing

Llama by 19 points on HumanEval and 7 points on MBPP. Notably, despite Qwen-2.5-1.5B-Instruct having more parameters than InfiR-1B-Instruct, it demonstrates comparable performance to Qwen-2.5-1.5B-Instruct in both coding and mathematical reasoning.

5.1.2 Discussion and Insights

During the training of the InfiR-1B, we observed several phenomena that significantly impacted the experimental results.

Heuristic filter for reducing special patterns During pretraining, metrics related to mathematics often showed large fluctuations. Analysis revealed that certain checkpoints generated an <eos> token with high probability when prompts ended with a colon ":". To address this, we cleaned the mathematical web page data by removing text ending with a colon.

Adjusting batch size according to data "width" Initially, we trained the model with separate batches for coding and math, using smaller batch sizes for faster convergence. When combining coding, math, and general reasoning data, using the same batch size often led to gradient norm overflow and instability. Larger batch sizes help avoid erroneous gradient estimates when more domains are included, but they slow convergence. We balanced training costs by using a batch size of 2048.

Importance of evaluation frameworks An appropriate evaluation framework is crucial for training the base model. In early experiments, we used OpenCompass to evaluate benchmarks across all domains, but metric variations lacked consistency as training progressed. Upon reviewing the evaluation framework, we found errors in extracting the base model's completion results. We determined that EvalPlus is better for code evaluation, while Qwen2.5-Math's evaluation tool is optimal for mathematics.

Appropriate Utilization of Synthetic Data Synthetic data can enhance model metrics during pretraining. Initially, we incorporated synthetic data at a certain ratio to achieve higher scores in the base model. However, these higher-scoring models did not necessarily outperform those without synthetic data after the same SFT process. This may be because synthetic data often has lower perplexity, creating a distribution gap with web data, which affects model convergence. Therefore, we only introduce synthetic data during the annealing

Model	MMLU	GSM8K	MATH	HumanEval	MBPP	MBPP(3-shot)
Llama-3.2-1B	32.74 (32.2)	8.11	3.42	17.68	33.46	24.8
Qwen-2.5-1.5B	63.03 (60.9)	66.57 (68.5)	31.24 (35.0)	35.37 (37.2)	58.37 (60.2)	41.4
InfiR-1B-Base	47.24	63.46	31.82	37.80	53.40	37.6

Table 1: Performance of base models on various benchmarks using few-shot evaluation. The values in parentheses indicate the claimed results from the respective papers.

Model	MMLU	GSM8K	MATH	HumanEval	MBPP
Llama-3.2-1B-Instruct	46.27 (49.3)	47.9 (44.4)	30.0 (30.6)	39.63	49.03
Qwen-2.5-1.5B-Instruct	61.78	74.3 (73.2)	53.4 (55.2)	51.83 (61.6)	56.81 (63.2)
InfiR-1B-Instruct	50.22	70.9	46.4	58.54	56.03

Table 2: Performance of InfiR-1B-Instruct models on various benchmarks using zero-shot evaluation. The values in parentheses indicate the claimed results from the respective papers. InfiR-1B-Instruct outperforms Llama-3.2-1B-Instruct in these reasoning benchmarks, and is proximate to Qwen-2.5-1.5B-Instruct in mathematical reasoning and code tasks.

phase to prevent overtraining.

Dependency of Data on Model Size in SFT In

SFT, we observe a strong dependency between the data and the base model. Data that performs well on large models may not achieve the same level of performance on smaller models. For smaller models, millions of data are required to achieve competitive performance.

Reasoning Enhancement with Long CoT OpenAI's o1 (Jaech et al., 2024) and Deepseek's R1 (Guo et al., 2025) significantly improve the complex reasoning capabilities across various domains such as mathematics, code, and science by extending the length of the CoT. Inspired by the test-time scaling approach, we also fine-tune InfiR-1B-Instruct on long CoT data to further enhance the model's reasoning abilities and investigate the impact of data scaling on performance. Detailed experimental results are provided in Appendix D.

Model	MMMU	ScreenSpot	Android World
Qwen2-VL-2B	41.1	9.3	-
Qwen2.5-VL-3B	53.1	55.5	-
Showui-2B	-	75.1	6.90
InfiR-VL-1.6B	38.8	76.3	9.48

Table 3: Performance of small multimodal models in reasoning tasks across different benchmarks

5.2 Small Multimodal Language Model

We evaluated the performance of InfiR-VL-1.6B on MMMU to test the general vision reasoning abilities, and then evaluated its reasoning capabilities on ScreenSpot and AndroidWorld. Table 3 compares the performance of InfiR-VL-1.6B with state-of-the-art SMLM. InfiR-VL-1.6B exhibits compara-

ble capabilities while maintaining a smaller model size

During the training of the InfiR-VL-1.6B, we also have some insights which is helpful for reasoning enhanced MSLMs training.

Curriculum Learning is crucial when employing pretrained ViT for multimodal training. For tasks such as grounding and OCR, which require detailed image comprehension, it is necessary to unfreeze the ViT while keeping LLM backbone frozen during training. End-to-end training may lead to early overfitting in the domain-specific dataset.

Domain-Specific Reasoning Capabilities are not strongly dependent on model sizes. Smaller models can achieve the required reasoning abilities if they are trained with suitable domain-specific datasets and paired with a backbone that possesses reasoning capabilities. (Liu et al., 2025)

6 Conclusion

In conclusion, this paper demonstrates the potential of small language models (SLMs) and multimodal small language models (MSLMs) to provide efficient and accessible AI solutions. By developing novel training pipelines, we show that compact models can achieve competitive reasoning capabilities while significantly reducing computational costs and addressing privacy concerns. Our proposed models, InfiR-1B-Base, InfiR-1B-Instruct, and InfiR-VL-1.6B, achieve state-of-the-art performance, underscoring the feasibility of deploying these models on edge devices for real-world applications. This research paves the way for more sustainable and inclusive AI development, promoting broader adoption and innovation in the field.

7 Limitation

Due to constraints in computational resources and time, our experiments were primarily conducted on standard benchmarks. Our research focuses on the technical aspects of traditional information extraction tasks, without addressing social or security issues. We adhere to ethical standards by avoiding sensitive data or applications. Although the results are promising, the method's generalization ability in real-world scenarios require further exploration. This limitation presents opportunities for future research to validate and expand the method's applicability under more complex and diverse conditions.

References

- Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, and Charles Sutton. 2021. Program synthesis with large language models. *Preprint*, arXiv:2108.07732.
- BAAI. 2024. Infinity instruct. arXiv preprint arXiv:2406.XXXX.
- Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yimin Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, Conghui He, Botian Shi, Xingcheng Zhang, Han Lv, Yi Wang, Wenqi Shao, Pei Chu, Zhongying Tu, Tong He, Zhiyong Wu, Huipeng Deng, Jiaye Ge, Kai Chen, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. 2024. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *CoRR*, abs/2412.05271.
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *Preprint*, arXiv:2110.14168.
- DeepSeek-AI. 2024a. Deepseek LLM: scaling opensource language models with longtermism. *CoRR*, abs/2401.02954.
- DeepSeek-AI. 2024b. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model. *CoRR*, abs/2405.04434.
- Yuyang Ding, Xinyu Shi, Xiaobo Liang, Juntao Li, Qiaoming Zhu, and Min Zhang. 2024. Unleashing reasoning capability of llms via scalable question synthesis from scratch. *arXiv preprint arXiv:2410.18693*.

- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.
- Xiaotian Han, Yiren Jian, Xuefeng Hu, Haogeng Liu, Yiqi Wang, Qihang Fan, Yuang Ai, Huaibo Huang, Ran He, Zhenheng Yang, et al. 2024. Infimmwebmath-40b: Advancing multimodal pre-training for enhanced mathematical reasoning. *arXiv preprint arXiv:2409.12568*.
- Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, et al. 2024. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems. *arXiv preprint arXiv:2402.14008*.
- Dan Hendrycks, Steven Basart, Saurav Kadavath, Mantas Mazeika, Akul Arora, Ethan Guo, Collin Burns, Samir Puranik, Horace He, Dawn Song, and Jacob Steinhardt. 2021a. Measuring coding challenge competence with APPS. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks* 2021, December 2021, virtual.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021b. Measuring massive multitask language understanding. *Preprint*, arXiv:2009.03300.
- Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021c. Measuring mathematical problem solving with the math dataset. *Preprint*, arXiv:2103.03874.
- Siming Huang, Tianhao Cheng, J. K. Liu, Jiaran Hao, Liuyihan Song, Yang Xu, J. Yang, J. H. Liu, Chenchen Zhang, Linzheng Chai, Ruifeng Yuan, Zhaoxiang Zhang, Jie Fu, Qian Liu, Ge Zhang, Zili Wang, Yuan Qi, Yinghui Xu, and Wei Chu. 2024a. Opencoder: The open cookbook for top-tier code large language models. *Preprint*, arXiv:2411.04905.
- Siming Huang, Tianhao Cheng, Jason Klein Liu, Jiaran Hao, Liuyihan Song, Yang Xu, J Yang, JH Liu, Chenchen Zhang, Linzheng Chai, et al. 2024b. Opencoder: The open cookbook for top-tier code large language models. *arXiv preprint arXiv:2411.04905*.
- Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.
- Oleksii Kuchaiev, Jason Li, Huyen Nguyen, Oleksii Hrinchuk, Ryan Leary, Boris Ginsburg, Samuel Kriman, Stanislav Beliaev, Vitaly Lavrukhin, Jack Cook, Patrice Castonguay, Mariya Popova, Jocelyn Huang, and Jonathan M. Cohen. 2019. Nemo: a toolkit for building ai applications using neural modules. *Preprint*, arXiv:1909.09577.

- Jia Li, Edward Beeching, Lewis Tunstall, Ben Lipkin, Roman Soletskyi, Shengyi Huang, Kashif Rasul, Longhui Yu, Albert Q Jiang, Ziju Shen, et al. 2024. Numinamath: The largest public dataset in ai4maths with 860k pairs of competition math problems and solutions. *Hugging Face repository*, 13:9.
- Raymond Li, Loubna Ben Allal, and Yangtian Zi et al. 2023. Starcoder: may the source be with you!
- Yujia Li, David Choi, Junyoung Chung, Nate Kushman, Julian Schrittwieser, Rémi Leblond, Tom Eccles, James Keeling, Felix Gimeno, Agustin Dal Lago, Thomas Hubert, Peter Choy, Cyprien de Masson d'Autume, Igor Babuschkin, Xinyun Chen, Po-Sen Huang, Johannes Welbl, Sven Gowal, Alexey Cherepanov, James Molloy, Daniel J. Mankowitz, Esme Sutherland Robson, Pushmeet Kohli, Nando de Freitas, Koray Kavukcuoglu, and Oriol Vinyals. 2022. Competition-level code generation with alphacode. Science, 378(6624):1092–1097.
- Yuhang Liu, Pengxiang Li, Zishu Wei, Congkai Xie, Xueyu Hu, Xinchen Xu, Shengyu Zhang, Xiaotian Han, Hongxia Yang, and Fei Wu. 2025. Infiguiagent: A multimodal generalist gui agent with native reasoning and reflection. *arXiv preprint arXiv:2501.04575*.
- Anton Lozhkov, Loubna Ben Allal, Elie Bakouch, Leandro von Werra, and Thomas Wolf. 2024a. Finemath: the finest collection of mathematical content.
- Anton Lozhkov, Raymond Li, Loubna Ben Allal, Federico Cassano, Joel Lamy-Poirier, Nouamane Tazi, Ao Tang, Dmytro Pykhtar, Jiawei Liu, Yuxiang Wei, et al. 2024b. Starcoder 2 and the stack v2: The next generation. *arXiv preprint arXiv:2402.19173*.
- Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations* (*ICLR*).
- Arindam Mitra, Luciano Del Corro, Guoqing Zheng, Shweti Mahajan, Dany Rouhana, Andres Codas, Yadong Lu, Wei ge Chen, Olga Vrousgos, Corby Rosset, Fillipe Silva, Hamed Khanpour, Yash Lara, and Ahmed Awadallah. 2024a. Agentinstruct: Toward generative teaching with agentic flows. *Preprint*, arXiv:2407.03502.
- Arindam Mitra, Hamed Khanpour, Corby Rosset, and Ahmed Awadallah. 2024b. Orca-math: Unlocking the potential of slms in grade school math. *arXiv* preprint arXiv:2402.14830.
- Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. 2025. s1: Simple test-time scaling. *arXiv preprint arXiv:2501.19393*.
- Team OLMo. 2024. 2 olmo 2 furious.

- Keiran Paster, Marco Dos Santos, Zhangir Azerbayev, and Jimmy Ba. 2023. Openwebmath: An open dataset of high-quality mathematical web text. *arXiv* preprint arXiv:2310.06786.
- Guilherme Penedo, Hynek Kydlíček, Loubna Ben allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. 2024. The fineweb datasets: Decanting the web for the finest text data at scale. *Preprint*, arXiv:2406.17557.
- Qiwei Peng, Yekun Chai, and Xuhong Li. 2024. Humaneval-xl: A multilingual code generation benchmark for cross-lingual natural language generalization. *Preprint*, arXiv:2402.16694.
- Qwen. 2023. Qwen technical report. arXiv preprint arXiv:2309.16609.
- Qwen. 2024. Qwen2 technical report. CoRR, abs/2407.10671.
- David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. 2023. Gpqa: A graduate-level google-proof q&a benchmark. *arXiv* preprint arXiv:2311.12022.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *Preprint*, arXiv:2402.03300.
- Kai Sun, Yushi Bai, Ji Qi, Lei Hou, and Juanzi Li. 2024. Mm-math: Advancing multimodal math evaluation with process evaluation and fine-grained classification. *Preprint*, arXiv:2404.05091.
- Yuxuan Tong, Xiwen Zhang, Rui Wang, Ruidong Wu, and Junxian He. 2024. Dart-math: Difficulty-aware rejection tuning for mathematical problem-solving. *arXiv preprint arXiv:2407.13690*.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.
- Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *CoRR*, abs/2409.12191.
- An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

- Longhui Yu, Weisen Jiang, Han Shi, Jincheng Yu, Zhengying Liu, Yu Zhang, James T Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. 2023. Metamath: Bootstrap your own mathematical questions for large language models. *arXiv preprint arXiv:2309.12284*.
- Xiang Yue, Tuney Zheng, Ge Zhang, and Wenhu Chen. 2024. Mammoth2: Scaling instructions from the web. *arXiv preprint arXiv:2405.03548*.
- Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11975–11986.
- Junjie Zhou, Zheng Liu, Shitao Xiao, Bo Zhao, and Yongping Xiong. 2024. Vista: Visualized text embedding for universal multi-modal retrieval. *Preprint*, arXiv:2406.04292.

Appendix

A General Reasoning Instruction Tuning

A.1 Collection of high-quality datasets

We used different datasets to perform SFT on the Llama-3.2 1B model with the same parameter settings and sampling size. Then, we compared the fine-tuned models' scores on MMLU and selected the datasets used by the model with the highest MMLU score. We experimented with Infinity-Instruct, orca-agentinstruct-1M-v1, tulu-3-sft-mixture, and WebInstructSub, and ultimately selected Infinity-Instruct and orca-agentinstruct-1M-v1.

Model	Epoch	GSM8K	MATH	MMLU	MBPP 0-shot	HumanEval
Llama-3.2 1B Instruct	-	43.37(44.4)	15.98(30.6)	47.18(49.3)	36.0	39.63
Llama-3.2 1B +tulu-3-sft-mixture(939k)	3.27	37.07	10.2	36.88	22.2	29.27
Llama-3.2 1B +Infinity-Instruct_Gen(1.12M)	4.11	31.16	8.1	39.77	14.2	18.9
Llama-3.2 1B +WebInstructSub	2.08	10.84	3.7	27.41	10.4	10.37

Table 4: General SFT Datasets Evaluation

A.2 Data Synthetic Pipeline

After collecting high-quality opensource data, we found that the model still lags behind the state-of-the-art (SOTA). Therefore, we constructed two synthetic data pipeline to achieve better performance. One for reasoning and the other for commonsense.

Reasoning Data This pipeline can be divided into two main steps: seed data preparation and inference-guided rejection sampling.

Seed Data Preparation: We utilize queries related to reasoning from Infinity-Instruct as our seed data.

Rejection Sampling: After obtaining the seed queries, we incorporate "step by step" prompting to activate the reasoning capabilities of large language models, thereby generating responses with reasoning steps. For each query, we select the highest-scoring samples using a reward model. Finally, we apply certain rules (e.g., checking if the response includes reasoning steps) to further filter the samples.

Knowledge Data This pipeline consists of four steps: evaluation, seed data preparation, query enhancement, and rejection sampling.

Evaluation: We construct an offline evaluation set encompassing several domains. We observe the performance gap between our model and the state-of-the-art models on this set. Domains with significant gaps are prioritized for optimization.

Seed Data Preparation: Similarly, we use Infinity-Instruct as our seed data. Each query in Infinity-Instruct is labeled and of high quality. By applying rules, we associate key domains identified during evaluation with labels in Infinity-Instruct, thereby identifying seed data that can enhance commonsense capabilities.

Query Enhancement: For long-tail domains where seed data may be insufficient, we expand queries using large language models (LLMs) based on the seed data. During this rewriting process, we retrieve context from a corpus and include the standard response corresponding to the query in the prompting to help the LLM generate more valuable queries.

Rejection Sampling: Once queries are obtained, we employ LLMs to generate responses, as is common in synthetic data pipelines. Subsequently, we use an open-source reward model to perform rejection sampling.

B Mathematical Reasoning Instruction Tuning

B.1 High-quality Mathematical Data

In the SFT stage, we evaluated the quality of public datasets—including dart-math-hard (Tong et al., 2024), MATH-plus (Yue et al., 2024), MetaMathQA (Yu et al., 2023), orca-math-word-problems-200k (Mitra et al., 2024b), and ScaleQuest-Math (Ding et al., 2024)—through experiments on the Llama-3.2-1B baseline model. Among these, the ScaleQuest-Math dataset achieved 69.45% on GSM8K and 42.20% on MATH, outperforming the Llama-3.2-1B-Instruct model.

Based on Qwen2.5-Math-1.5B-Instruction model, we trained a RL model to regenerate ScaleQuest-Math answers, thereby enabling step-by-step inference. Additionally, we employ the Qwen2.5-Math-RM-72B model for rejection sampling, which evaluates the inference quality and intermediate steps of the generated query-response pairs.

B.2 Mathematical Data Compression

We utilized the Llama3.3-70B-Instruct model to annotate the difficulty of the ScaleQuest-Math dataset, following the MATH benchmark. The annotations included:

• Difficulty Labels: Very easy, easy, medium, hard, and very hard

We performed sampling experiments on the annotated ScaleQuest-Math dataset, drawing inspiration from Dart-Math (Tong et al., 2024)'s methodology. The experiments were structured as follows:

- **Group A**: Combined "very easy" and "easy" difficulty labels, comprising approximately 483,000 items
- **Group B**: Combined "medium", "hard", and "very hard" difficulty labels, comprising around 350,000 items

Both groups underwent experiments under identical setups. We found that on the easier GSM8K task, the performance of the two groups was similar (61.33% for group A and 60.5% for group B). On the more challenging MATH task, the data from group B achieved 36.64%, outperforming group A by 13%. Notably, using only 35% of the dataset volume from group B nearly matched the performance of the full dataset on both GSM8K and MATH, highlighting the effectiveness of difficulty-based data compression.

C Code Reasoning Instruction Tuning

C.1 High-quality Code Data

To identify high-quality code datasets, we conducted a comprehensive evaluation using several code datasets to perform Supervised Fine-Tuning (SFT) on the Llama-3.2 1B model. Specifically, we fine-tuned the model with a learning rate of 2×10^{-5} and a batch size of 32. Each dataset was used to fine-tune the model for 5 epochs, and the intermediate models were saved for evaluation across various benchmarks. The detailed evaluation results are presented in Table 5. The Llama-3.2 1B model fine-tuned by the combination of opc-sft-stage2 and ScaleQuest-Code shows the best overall performance on the MBPP and HumanEval, which demonstrates the relatively high quality of the code datasets.

C.2 Synthetic Code Data

To enhance the diversity and quality of the post-training dataset, we implement a pipeline for synthesizing code-related instruction data from validated sources. We leverage a curated subset of the Opencoder annealing dataset, comprising high-quality code snippets previously validated during the annealing stage. The seed dataset consists of multilingual code snippets with English comments. The code data are then synthesized using the Qwen2.5-Coder-Instruct-32B model, with prompts designed to generate task

Table 5: Code SFT Evaluation (Batchsize = 32, Full SFT, Learning Rate= 2×10^{-5})

Model	GSM8K	MATH	MMLU	MBPP 0-shot	HumanEval
Llama-3.2 1B Instruct	43.37(44.4)	15.98(30.6)	47.18(49.3)	36.0	39.63
Llama-3.2 1B + code_instructions_122k (122k)	1.9	1.08	25.31	5.4	16.46
Llama-3.2 1B + stack-exchange-instruction (150k)	1.36	1.1	25.06	2.4	3.66
Llama-3.2 1B + magicoder-oss (75k)	3.11	1.18	24.61	17	9.15
Llama-3.2 1B + ScaleQuest-Code (156k)	4.55	2.74	27.15	31	48.17
Llama-3.2 1B + opc-sft-stage2 (436k)	3.34	0.5	26.46	40.6	54.88
Llama-3.2 1B + opc-sft-stage2 + magicoder-oss	3.34	1.34	23.92	39.8	59.15
Llama-3.2 1B + opc-sft-stage2 + ScaleQuest-Code	4.7	1.02	26.67	40.8	60.98
Llama-3.2 1B + opc-sft-stage2 + magicoder-oss + ScaleQuest-Code	3.41	2.38	25.89	41.4	57.93

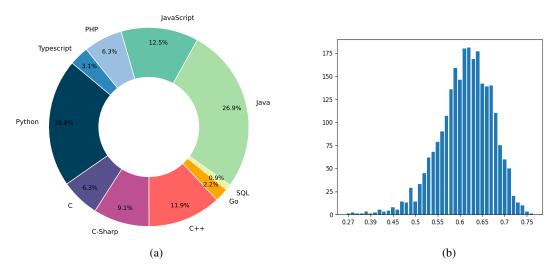


Figure 4: Left: multi-programming language distribution. Right: similarity histogram of 2500 image-text pairs sampled from the coco-caption dataset

descriptions, analysis steps, code solutions, and the corresponding test cases based on these seed samples. To ensure the quality of generated data, we perform execution-based code verification within a sandbox environment, using a Python interpreter as our verifier. Specifically, we include self-contained code snippets that pass all unit tests in our Supervised Fine-Tuning (SFT) dataset. We also filter out synthetic code that encounters runtime timeouts, as these solutions are often inefficient and do not meet our quality standards.

Model	AIME24	MATH500	AMC23	GPQA	OlympiadBench
Llama-3.2-1B-Instruct	0.00	0.25	0.175	0.02	0.043
Qwen2.5-1.5B-Instruct	0.067	0.492	0.225	0.242	0.185
DeepSeek-R1-Distill-Qwen-1.5B	0.289	0.839	0.700	0.338	0.436
InfiR-1B-Instruct (200k Long CoT)	0.033	0.474	0.225	0.288	0.181
InfiR-1B-Instruct (2M Long CoT)	0.067	0.62	0.300	0.364	0.224

Table 6: Results of additional fine-tuning of InfiR-1B-Instruct on Long CoT data. Further fine-tuning on Long CoT data enhances the model's complex reasoning capabilities, with performance improving as training data increases.

D Reasoning Enhancement with Long CoT

We selected 2M high-quality long CoT data from NuminaMath-QwQ-CoT-5M ¹ to fine-tune InfiR-1B-Instruct model, further enhancing its complex reasoning performance. Referring to s1 (Muennighoff et al., 2025)'s difficulty and diversity-based sampling method, we extract a 200K subset from the full 2M dataset to assess the model's performance on data scaling. Table 6 shows the performance of the long CoT

¹https://huggingface.co/datasets/PrimeIntellect/NuminaMath-QwQ-CoT-5M

fine-tuned model on AIME24, MATH500 (Hendrycks et al., 2021c), AMC23, GPQA (Rein et al., 2023) and OlympiadBench (He et al., 2024). After fine-tuning on the Long CoT dataset, the model demonstrates further improvements in complex reasoning tasks. Additionally, we observe a consistent performance improvement as the training data scales from 200K to 2M, highlighting the critical role of long CoT data scale in fine-tuning small language models.

E Multimodal Data Cleaning

In order to select a suitable similarity cutoff threshold, we sampled 2,500 image-text pairs from cococaption and calculated the similarity, and performed statistical analysis. The results are shown in Fig 4b. We can find that less than 5% of the image-text pairs have a similarity lower than 0.5. In order to reduce the impact of these low-quality, low-similarity image-text pairs on the model's ability to understand the semantics of images and texts, we set the similarity cutoff threshold to 0.5.

F Data Composition

Type Source	Datasets
Web Pages	Common crawl, FineWeb-Edu, DCLM
Math	InfiMM-WebMath-40B, OpenWebMath, MathCode-Pile, Proof-pile-2, finemath
Code	the-Stack-v2, Starcoder
General Knowledge	arXiv, StackExchange, peS2o
Encyclopedia	Wikipedia
Open-Source Instruction	FLAN, Infinity Instruct

Table 7: List of Pretraining data composition

Type Source	Datasets
Math	InfiMM-WebMath-40B, OpenWebMath, MathCode-Pile, Proof-pile-2, finemath
Code	the-Stack-v2, Starcoder, opc-annealing-corpus
General Knowledge	StackExchange, peS2o
Encyclopedia	Wikipedia
Open-Source corpus	FLAN, Infinity Instruct, Dolmino
Synthetic data	Ours

Table 8: List of Annealing data composition

Task	Dataset
Type: Single-in	mage Datasets
Captioning	TextCaps, ShareGPT4o, InternVL-SA-1B-Caption, NewYorkerCaptionContest, MMInstruct
General QA	VQAv2, GQA, OKVQA, Visual7W, MMInstruct, VSR, FSC147, Objects365-YorN, Hateful-Memes
OCR	OCRVQA, TextVQA, HME100K, COCO-Text, LSVT, VCR, ST-VQA, LLaVAR, CyrillicHandwriting, IAM, NAF,
CI.	TextOCR, SROIE, MTVQA
Chart	ChartQA, MMTab, FigureQA, VisText, ArxivQA, TabMWP, MMC-Inst, DVQA, UniChart, SimChart9K, Chart2Text
Document	DocVQA, DocReason25K, Sujet-Finance-QA-Vision
Mathematics	GeoQA+, CLEVR-Math, Super-CLEVR, MapQA, MAVIS, Geometry3K, TallyQA, GEOS, UniGeo, GeomVerse, CMM-Math
Knowledge	A-OKVQA, ViQuAE, iNaturalist2018, MovieNet, KonIQ-10K
Grounding	RefCOCO/+/g, All-Seeing-V2, V3Det, DsLMF, COCO-ReM, TolokaVQA
Conversation	ALLaVA, Viet-ShareGPT40, Cambrain-GPT40, RLAIF-V, Laion-GPT4V, TextOCR-GPT4V, WildVision-GPT40
GUI	Screen2Words, WebSight, Widget-Caption, RICOSCA, Seeclick, ScreenQA,
	AMEX, AITW, Odyssey, UIBert, AndroidControl, Mind2Web, OmniACT, WaveUI
Medical	PMC-VQA, VQA-RAD, ImageCLEF, SLAKE, VQA-Med, PathVQA
Science	AI2D, ScienceQA, TQA
Type: Multi-in	nage Datasets
General QA	Img-Diff, Birds-to-Words, Spot-the-Diff, MultiVQA, NLVR2, ContrastiveCaption, DreamSim, InternVL-SA-1B-Caption
Document	MP-DocVQA, MP-Docmatix
Type: Text Dat	asets
General QA	UltraFeedback, UltraChat, Unnatural-Instructions, NoRobots, MOSS, LIMA, SlimOrca, WizardLM-Evol-Instruct-70K,
	Llama-3-Magpie-Pro, Magpie-Qwen2-Pro, KOpen-HQ-Hermes-2.5-60K, Firefly, Dolly, OpenAI-Summarize-TLDR,
	Know-Saraswati-CoT, FLAN, FLANv2
Code	Code-Feedback, Glaive-Code-Assistant, XCoder-80K, LeetCode, Evol-Instruct-Code
Long Context	Long-Instruction-with-Paraphrasing, LongCite, LongQLoRA, LongAlpaca
Mathematics	GSM8K-Socratic, NuminaMath-CoT/TIR, Orca-Math, MathQA, InfinityMATH

Table 9: List of vision datasets used for different tasks.