# BOSE: A Systematic Evaluation Method Optimized for Base Models

**Hongzhi Luan***, **Changxin Tian***, **Zhaoxin Huan**
**Xiaolu Zhang**, **Kunlong Chen**, **Zhiqiang Zhang**, **Jun Zhou**†
Ant Group
{luanhongzhi.lhz, tianchangxin.tcx, zhaoxin.hzx}@antgroup.com
{yueyin.zxl, kunlong.ckl, lingyao.zzq, jun.zhoujun}@antgroup.com

## Abstract

This paper poses two critical issues in evaluating base models (without post-training): (1) Unstable evaluation during training: in the early stages of pre-training, the models lack the capability to answer questions as required, leading to unstable evaluation results. This instability makes it difficult to provide solid conclusions to guide the training, especially for key experiments such as data ablation and scaling law. (2) Inconsistency between base and instruct models: base models generally exhibit poorer evaluation performance compared to corresponding instruct models. This gap poses a challenge for assessing whether a base model with better evaluation can truly lead to a better instruct model. To address these issues, we propose **B**ase model **O**riented **S**ystematic **E**valuation (**BOSE**), a method specifically designed to optimize the evaluation of base models. Specifically, BOSE introduces two key innovations: In-Context Light-instruction Prompt (ICLiP) for open-ended tasks and Blank-ppl for multi-choice tasks with candidate options, which transforms the standard perplexity (ppl) metric into a fill-in-the-blank format to mitigate early-stage evaluation fluctuations. Furthermore, we are the first to propose Kendall's rank correlation to quantitatively measure the evaluation stability and consistency. Experimental results demonstrate that BOSE significantly enhances both the stability of evaluations during pre-training and the consistency between base and instruct models, thereby providing more reliable guidance for the LLMs' training.

## 1 Introduction

Recently, large language models (LLMs) have demonstrated remarkable achievements across various domains (Minaee et al., 2024). This has led to the development of numerous high-performing LLMs (OpenAI, 2024; Aaron Grattafiori and Abhinav Jauhri, 2024; QwenTeam, 2025; GemmaTeam, 2024). To evaluate the performance of these models on a wide range of tasks, an increasing number of benchmarks have been open-sourced (Guo et al., 2023). These benchmarks provide a comprehensive evaluation of the capabilities of LLMs, guiding the training and improving their weakness. Evaluating large language models is a cornerstone in development of LLMs.

Generally, LLMs can be divided into two primary categories based on whether they receive post-training: *base models* and *instruct models* (Aaron Grattafiori and Abhinav Jauhri, 2024; Fu and Khot, 2022). Instruct models undergo downstream adaptation, allowing them to adapt to specific tasks and answer questions as required. In contrast, base models focus on learning foundational knowledge without targeting specific tasks, resulting in a weaker ability to respond to instructions. Consequently, most current benchmarks provide comprehensive evaluations of instruct models (Patel et al., 2024; Shi et al., 2022), while evaluations of base models are relatively rare. Nevertheless, base models serve as the foundation for the entire LLM training. An accurate evaluation for base models can provide essential insights for training LLMs, such as experiments on scaling laws (Kaplan et al., 2020), ablation of pre-training data, and the selection of appropriate base checkpoints for post-training (Aaron Grattafiori and Abhinav Jauhri, 2024; DeepSeek-AI, 2024).

In this paper, we argue that the current evaluation is not suited to the characteristics of base models and lacks a systematic methodology. In a nutshell, the challenge arises from two dimensions:

- **Instability of evaluations during training**. Critical pre-training experiments, such as data ablation and scaling laws, rely on the performance of base models in the early stages of

*Equal contributions.
†Corresponding author.



*Findings of the Association for Computational Linguistics: ACL 2025*, pages 16147–16158

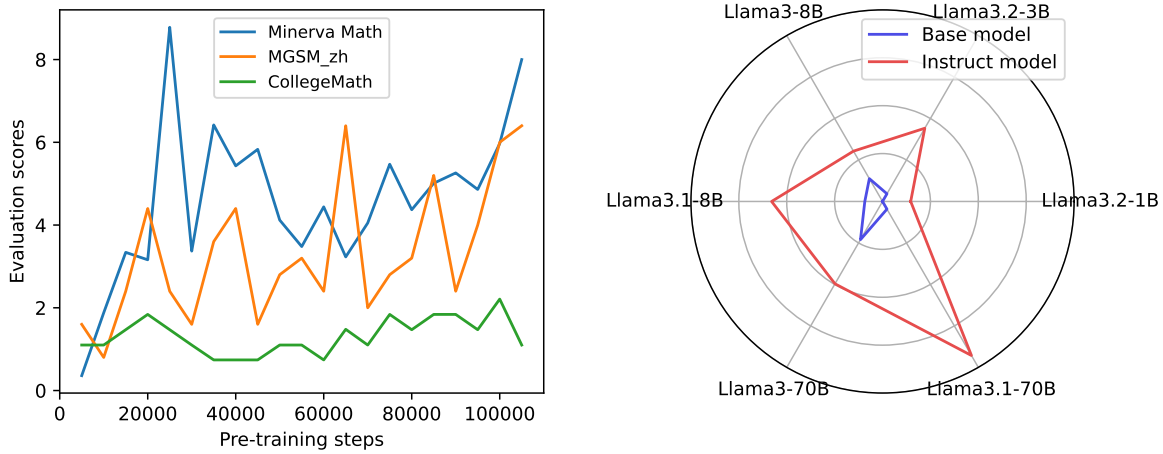


Figure 1: Illustrations of the critical issues in current base model evaluation. (**left**): Evaluation scores of 3 benchmarks on a series of pre-trained checkpoints, with the x-axis representing the increasing trained steps, and the y-axis representing the evaluation scores. The evaluation scores do not improve stably during pre-training. (**right**): Comparison of evaluation scores on 6 Llama base models and instruct models for several benchmarks. There is a lack of consistency between the two score series. For example, Llama-3.1-70B achieves the highest score among instruct models, but performs poorly in the base model.

training. However, base models cannot follow instructions and solve problems during the initial stages, leading to fluctuations in evaluation results. Figure 1 (left) illustrates the dynamic evolution of a 1B-parameter model's performance across different benchmarks as the training tokens increase. The significant fluctuations in the evaluation make it difficult to determine whether the model has gained real improvement throughout the training tokens. This instability hinders making reliable decisions for these critical experiments.

- **Lack of consistency in evaluating base and instruct models**. During pre-training, LLMs acquire more knowledge than the post-training. However, base models typically perform worse on evaluations than their corresponding instruct models due to the weaker instruction-following capabilities. Taking the Llama family models (Aaron Grattafiori and Abhinav Jauhri, 2024) as an example, we evaluate these models on several recently released mathematical datasets (Tang et al., 2024; Lewkowycz et al., 2022; Shi et al., 2022), as shown in Figure 1 (right). The base models' scores are significantly lower than their corresponding instruct models, and there is little distinction among the different versions of the base models. This discrepancy presents a challenge for current evaluation methods in determining whether a base model with better evaluation results can indeed lead to a better instruct model.

In this paper, we propose a systematic approach for the evaluation of base models, called **B**ase model-**O**riented **S**ystematic **E**valuation (**BOSE**). BOSE aims to address the aforementioned critical issues by optimizing existing evaluation methods. Specifically, we observe that base models often struggle to understand evaluation questions, which results in inaccurate reflections of their true capabilities. This is particularly evident in their difficulty with multi-choice questions, poor ability to follow complex instructions, and tendency to generate redundant continuations. BOSE improves evaluation techniques tailored to pre-training characteristics for two key tasks: open-ended generation and multi-choice. It introduces **I**n-**C**ontext **L**ight-**i**nstruction **P**rompt (**ICLiP**) to improve the base model's responses to open-ended questions and transforms multi-choice questions into fill-in-the-blank versions (**Blank-ppl**) to mitigate evaluation fluctuations in early stage. Finally, we validate the effectiveness of BOSE by examining two key aspects: training stability and consistency between base and instruct models. This evaluation aims to determine whether the results exhibit a stable growth trend and whether a better base model can lead to a better instruction model.

In summary, our contributions are as follows:

- We propose BOSE, a systematic evaluation method tailored to base models that incorporates ICLiP for open-ended generation tasks and blank-ppl for multi-choice tasks. This approach aims to better align with the characteristics of base models, yielding evaluation results that accurately reflect their true capabilities.
- To assess the effectiveness of base model evaluation method, we define criteria based on the stability of metrics during pre-training and the consistency of capabilities between base and corresponding instruct models, and pioneer the use of Kendall's rank correlation as a quantitative metric.
- We conduct comprehensive experiments on multiple benchmarks covering knowledge, mathematics and reasoning, using both open-source models and our pre-trained checkpoints. Empirical results demonstrate that BOSE significantly improves the evaluation stability during pre-training and enhances the consistency between the base and instruct models, which is highly beneficial for guiding the model development and ensuring the reliability of base model evaluation results.

To the best of our knowledge, this paper presents the first systematic framework for evaluating base LLMs, introducing empirically validated optimization methodologies with quantitative metrics. We believe this will yield actionable insights and practical recommendations to contribute to advance future base LLM evaluation.

## 2 Related Work

In this section, we first introduce some commonly used benchmarks and evaluation tasks for LLMs, and then focus on the existing approaches for evaluating base models.

**Evaluation Tasks.** Based on whether there exist reference answers to be automatically calculated, evaluation tasks commonly can be classified as either ground truth-based evaluation or human preference-based evaluation (Chiang et al., 2024; Guo et al., 2023). Considering that base models typically cannot align with human preferences, we focus on evaluation tasks with ground truths. As for these benchmarks, the presence of candidate answers can generally further differentiate (OpenCompass, 2023):

- **Open-ended task**. This type of evaluation tasks require the model to respond to questions according to given instructions, and employ a customized post-processing process to extract potential answers as well as judge the correctness, as seen in benchmarks such as Math (Hendrycks et al., 2021b), BBH (Suzgun et al., 2022) and HumanEval (Mark Chen et al., 2021).
- **Multi-choice task**. In these tasks, given a question, the model needs to choose the most appropriate option from multiple choices. As for evaluating base models, perplexity for a given sentence is typically calculated to evaluate a model's language modeling capabilities (Tom B. Brown and et al., 2020), as seen in benchmarks such as MMLU (Hendrycks et al., 2021a), CMMLU (Li et al., 2023), etc.

**Base Model Evaluation.** To our knowledge, the evaluation of base models remains under-addressed. Despite the availability of numerous benchmarks, few provide detailed evaluations of base models. For instance, CMath (Wei et al., 2023b) releases datasets but pays little attention to evaluation prompts. Many benchmarks, such as SimpleQA (Wei et al., 2024), MuSR (Sprague et al., 2024), Multi_LogiEval (Patel et al., 2024), and CollegeMath (Tang et al., 2024), do not distinguish between the evaluation of base and instruct models. Only a few benchmarks, like kor-bench (Ma et al., 2024), consider both aspects.

A common approach to evaluate base models is using in-context learning (Tom B. Brown and et al., 2020; Dong et al., 2024), particularly the few-shot method. This method expects the base model to leverage its in-context learning ability to respond to questions based on given examples, thereby addressing instruction-following challenges. Therefore, some open-source evaluation frameworks, such as lm-eval (Gao et al., 2024) and OpenCompass (OpenCompass, 2023) utilize in-context learning to evaluate the performance of base models. However, this approach lacks systematic guidance on how to effectively conduct in-context learning, especially when considering the characteristics of base models in different training stages.

Due to the above issues, although many technical reports of open-source LLMs disclose the bench-

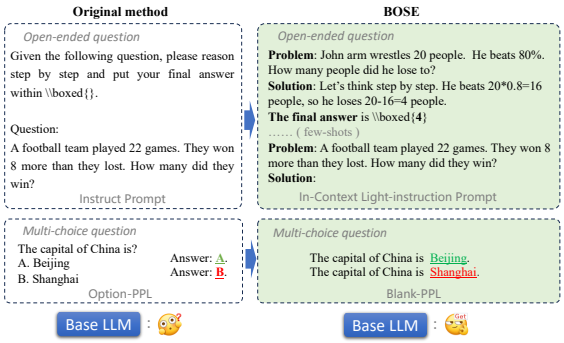


Figure 2: Illustrations of our proposed BOSE, which incorporates ICLiP for open-ended generation tasks (**top**) and Blank-ppl for multi-choice tasks (**bottom**).

marks and evaluation results used for base model evaluation (Aaron Grattafiori and Abhinav Jauhri, 2024; QwenTeam, 2025; GemmaTeam, 2024), the lack of unified evaluation protocol and the incomplete transparency of evaluation details make it challenging to reproduce and align the evaluation results, which may cause significant difficulties for researchers.

## 3 Methodology

Different from instruct models, base model evaluation encounters unique challenges due to the lack of instruction-tuning and preference alignment. Specifically, base models typically fail to comprehend the intent behind prompts or follow explicit instructions, leading to suboptimal evaluation protocols that neither reflect their intrinsic capabilities nor bring reliable evaluation results, which becomes particularly critical during early pre-training stages.

To facilitate these limitations and align evaluation protocols with the intrinsic characteristics of base models, we optimize the evaluation methods for open-ended and multi-choice tasks respectively, as illustrated in Figure 2.

### 3.1 Open-ended Task

Considering the inherent characteristics of base models in generative tasks, we propose the In-Context Light-instruction Prompt (**ICLiP**), which comprises three core components:

**In-context learning**. In-context learning (ICL) (Tom B. Brown and et al., 2020; Dong et al., 2024) is an effective paradigm that enables pre-trained language models to perform new tasks without gradient updates. By providing some carefully prepared examples in the form of demonstrations, the model implicitly learns to mimic the reasoning patterns and output styles demonstrated in the context, thereby exhibiting its intrinsic problem-solving abilities. Our ICLiP method incorporates standardized few-shot examples within the in-context learning paradigm.

**Light-instruction prompt**. While prompting techniques are effective for LLMs (Sander Schulhoff and et al., 2024; Wei et al., 2023a), excessive prompting may confuse base models due to their lack of instruction-tuning (Wang and Zhou, 2024). We propose a lightweight prompt template, which formulates the input using the following format:

```
Problem: {problem}
Solution: let's think step by step. {cot}
The final answer is \boxed{answer}
```

where "`{problem}`" represents the problem to be solved, "`{cot}`" demonstrates the intermediate reasoning steps, and the final answer is presented in a specified format for few-shot examples. In the target question, "`{cot}`" is left empty. We refer to this prompt template as **light-instruction**, as shown in Figure 2, demonstrating better adaptation to the pre-training paradigm and intrinsic abilities of base models.

**With stopping criteria**. Using the above few-shot light-instruction prompt, a possible consequent issue is the uncontrolled continuations in generation, in other words, base model may fail to complete the answering and continue to generate another question instead. This not only degrades evaluation efficiency but also complicates the optimal answer extraction (e.g., retrieving the last numerical value as the final answer). To mitigate this, we augment the EOS token list in the `generate` function with a special text "`Problem:`". When the model encounters these tokens during decoding, it terminates response immediately and prevents redundant outputs.

### 3.2 Multi-choice Task

Standard evaluation of multi-choice tasks typically employs perplexity-based method to derive the optimal option, and we refer to it as option-ppl in this paper. However, we encounter critical limitations in this approach when evaluating base models with smaller parameter sizes or those in early pre-training stages.

The dominant approach to pre-training LLMs typically involves language modeling, which is commonly framed as a next-token-prediction task (Radford and Narasimhan, 2018;

Aaron Grattafiori and Abhinav Jauhri, 2024). To better align with the inherent architecture of this pre-training process, we reformulate the option-ppl as a fill-in-the-blank format, leveraging the natural sequential structure inherent in pre-training corpora. Specifically, as illustrated in Figure 2, we omit the candidate options and calculate perplexity for each option with the concatenation of the question and candidate text directly. This ensures stronger contextual coherence for the true answer while presenting higher perplexities for others.

## 4 Experiments

In this section, we describe systematic experiments to investigate how BOSE enhances evaluation stability during pre-training and ensures consistency between base and instruct models[1].

### 4.1 Setup

#### 4.1.1 Benchmarks

We employ 9 benchmarks, mainly categorized into two classes by task type:

**Open-ended tasks**. We take 5 mathematical reasoning benchmarks: CMath (Wei et al., 2023b), MGSM[2] (Shi et al., 2022), Gaokao2023EN (Liao et al., 2024), CollegeMath (Tang et al., 2024), and Minerva Math (Lewkowycz et al., 2022), alongside 1 multi-step logical reasoning benchmark Multi_LogiEval (Patel et al., 2024). These benchmarks generally require multiple intermediate thinking steps to derive the final answers, and are primarily employed to validate the proposed ICLiP protocol. Answer extraction and judge functions are implemented before accuracy calculation, with greedy decoding applied throughout all experiments.

**Multi-choice tasks**. We use 3 knowledge-driven benchmarks (MMLU (Hendrycks et al., 2021a), CMMLU (Li et al., 2023), and MMLU_Pro (Wang et al., 2024)), specifically to test the blank-ppl methodology. These benchmarks assess model performance through perplexity-based evaluation, where the option with the lowest perplexity is identified as the predicted answer.

We utilize accuracy as the unified evaluation metric for all benchmarks. Notably, these benchmarks are mainly chosen from recent public researches, and encompass both English and Chinese benchmarks across the two categories of tasks, ensuring linguistic diversity and comprehensive experimental validation. It is also applicable to a wider range of benchmarks.

#### 4.1.2 Models

To comprehensively evaluate BOSE, our experimental framework incorporates two categories of base models:

**Pre-trained models**[3]. These models are collected from our pre-training experiments, encompassing 1B and 2B parameter sizes with different training steps, 96 checkpoints in total, more details about the training recipe can be found in Appendix C and our technical report (LingTeam, 2025). To assess the capability consistency between base models and their instruction-tuned counterparts, we further fine-tune 24 checkpoints using identical SFT data (size, quality, filtering criteria) and training hyper parameters to derive the instruct models.

**Open-source base models**. These models mainly cover three remarkable LLM families: (1) **Llama family** (Aaron Grattafiori and Abhinav Jauhri, 2024): mainly including Llama-3.1-8B/70B, with an extended analysis incorporating Llama-3-8B/70B and Llama-3.2-1B/3B for consistency investigation; (2) **Gemma family** (GemmaTeam, 2024): Gemma-2-9B/27B; (3) **Qwen family** (QwenTeam, 2025): Qwen2.5-7B/72B.

#### 4.1.3 Metric

To assess the effectiveness of BOSE, we introduce Kendall's rank correlation (Kendall, 1938) as a quantitative metric beyond intuitive results. Formally, the Kendall's rank correlation $\tau$ is defined as:

$$\tau = \frac{P - (n(n-1)/2 - P)}{n(n-1)/2} = \frac{4P}{n(n-1)} - 1$$

where $n$ denotes the total number of entities, $P$ denotes the count of concordant pairs where two entities maintain identical ranking orders, and $n(n-1)/2$ represents the total number of possible pairwise comparisons. The correlation ranges between -1 and 1, with 1 indicating perfect concordance, -1 denoting complete discordance, and 0 corresponding to random ordinal association.

[1] All of our experiments are implemented based on OpenCompass (OpenCompass, 2023), and the prompt templates are available in Appendix B.

[2] While the MGSM benchmark provides multilingual variants, we employ the Chinese subset for experimental in this study, and named it with MGSM_zh for simplicity.

[3] Since the open-source LLMs do not provide publicly accessible intermediate model weights, we take our pre-trained models for experiments.

| Tasks | Benchmark | 1B | | | 2B | | |
|---|---|---|---|---|---|---|---|
| | | Original | BOSE | **Improve** | Original | BOSE | **Improve** |
| Open-ended | CMath | 0.418 | 0.669 | **0.251** | 0.524 | 0.736 | **0.212** |
| | MGSM_zh | 0.375 | 0.234 | **-0.141** | 0.376 | 0.617 | **0.241** |
| | Gaokao2023EN | 0.537 | 0.629 | **0.092** | 0.543 | 0.597 | **0.054** |
| | CollegeMath | 0.345 | 0.684 | **0.339** | 0.691 | 0.721 | **0.030** |
| | Minerva Math | 0.264 | 0.089 | **-0.175** | 0.296 | 0.200 | **-0.096** |
| | Multi_LogiEval | 0.099 | 0.146 | **0.047** | 0.324 | 0.362 | **0.038** |
| Multi-choice | MMLU | 0.637 | 0.845 | **0.208** | 0.795 | 0.834 | **0.039** |
| | CMMLU | 0.754 | 0.895 | **0.141** | 0.941 | 0.893 | **-0.048** |
| | MMLU_Pro | 0.382 | 0.778 | **0.396** | 0.603 | 0.874 | **0.271** |
| AVG | | 0.423 | 0.552 | **0.129** | 0.566 | 0.648 | **0.082** |

Table 1: Details of evaluation stabilities on our pre-trained 1B and 2B parameter models

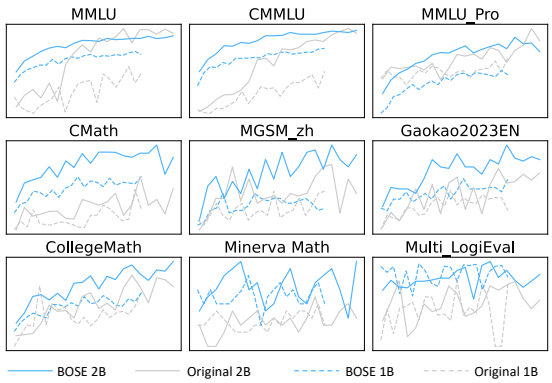


Figure 3: Intuitive results of evaluation stabilities across different benchmarks. X-axis: pre-training steps; y-axis: evaluation scores.

Upon this metric, we further introduce two measurements to inspect the effectiveness of base model evaluation results:

**Stability** during pre-training: calculated with the Kendall's rank correlation between evaluation scores and pre-training tokens sequences.

**Consistency** with instruct models: measured by the Kendall's rank correlation between base models' evaluation scores and corresponding instruct models' scores, which is inspired by previous work within the community (Agarwal et al., 2024).

## 4.2 BOSE Improves Evaluation Stability During Training

In this subsection, we validate the stability of model capabilities during pre-training using our pre-trained 1B and 2B parameter models with varying trained steps, including their performance on 6 open-ended tasks and 3 multi-choice tasks.

### 4.2.1 Intuitive Results

As visualized in Figure 3, our experimental results demonstrate how model evaluation scores grow with increasing trained steps across different benchmarks. We observe that, original evaluation methods exhibit some fluctuations approaching random variation across most benchmarks, whereas BOSE achieves relatively smoother trends.

We argue that instabilities still emerge in specific benchmarks (e.g., Minerva Math and Multi_LogiEval), potentially due to the high difficulty of these benchmarks and the limitations in model size or consumed token number.

### 4.2.2 Quantitative Metrics

To further analyze the improvements of evaluation stabilities with BOSE, we calculate Kendall's rank correlations between evaluation scores of two methods and pre-training steps, as detailed in Table 1.

In open-ended tasks, we achieve average $\tau$ improvements of $0.129$ $(+30.5\%)$ and $0.082$ $(+14.5\%)$ in 1B and 2B parameter models respectively, sustained improvement trends across most tasks except Minerva Math; as for multi-choice tasks, 5/6 experiments demonstrate improved $\tau$ values, with correlation levels remaining high ($\tau = 0.893$ in the single degradation case), which confirm BOSE significantly enhances evaluation stability during pre-training.

In summary, we demonstrate that BOSE aids in enhancing capability monitoring throughout LLMs' pre-training, thus providing solid guidance for the training process or early detection of training anomalies through stability metric deviations.

| Benchmark | Original | ICLiP | **Improve** |
|---|---|---|---|
| CMath | 0.467 | 0.867 | **0.400** |
| MGSM_zh | 0.467 | 0.867 | **0.400** |
| Gaokao2023EN | 0.467 | 0.867 | **0.400** |
| CollegeMath | 0.467 | 0.733 | **0.266** |
| Minerva Math | 0.000 | 0.600 | **0.600** |
| Multi_LogiEval | 0.602 | 0.733 | **0.131** |
| **AVG** | 0.412 | 0.778 | **0.366** |

Table 2: Consistency on 6 models from Llama family

| Benchmark | Original | ICLiP | **Improve** |
|---|---|---|---|
| CMath | -0.070 | 0.367 | **0.437** |
| MGSM_zh | 0.050 | -0.150 | **-0.200** |
| Gaokao2023EN | 0.020 | 0.128 | **0.108** |
| CollegeMath | 0.142 | 0.483 | **0.341** |
| Minerva Math | 0.314 | 0.507 | **0.193** |
| Multi_LogiEval | -0.017 | 0.230 | **0.247** |
| **AVG** | 0.073 | 0.261 | **0.188** |

Table 3: Consistency on our pre-trained models

### 4.3 ICLiP Enhances the Consistency between Base and Instruct Models

As discussed previously, we expect to not only ensure stable model evaluation with increasing pre-training tokens but also enhance the consistency between base models and corresponding instruct models, thereby reflecting real model capabilities. This implies a fundamental hypothesis: a base model that performs better in evaluation leads to a better instruct model, and vice versa.

We conduct systematic experiments to compare base models with their post-trained instruct models across both open-source models and our pre-trained checkpoints. Specifically, for each benchmark, we calculate the Kendall's rank correlation $\tau$ between evaluation results from a series of base models and corresponding post-trained instruct models, of which instruct models are assessed with commonly used instruct prompt, while base models are evaluated using both instruct prompt and our proposed ICLiP method, refer to appendix A for further details.

#### 4.3.1 Experiments on Open-Source Models

We select six base models across three released Llama versions (Llama-3-8B, Llama-3-70B, Llama-3.1-8B, Llama-3.1-70B, Llama-3.2-1B, Llama-3.2-3B) along with their corresponding instruction-tuned models, to calculate the capability consistencies on each benchmark. As shown in Table 2, BOSE significantly enhances the rank correlation of evaluation scores between base and instruct models compared to the original method, achieving an average Kendall's $\tau$ coefficient improvement of 0.366 (+88.9%).

Intuitively, we take Cmath and CollegeMath as examples to visualize the evaluation results. As shown in Figure 4, the base and instruct models exhibit stronger trend consistency, providing more coherent capability rankings; moreover, we observe

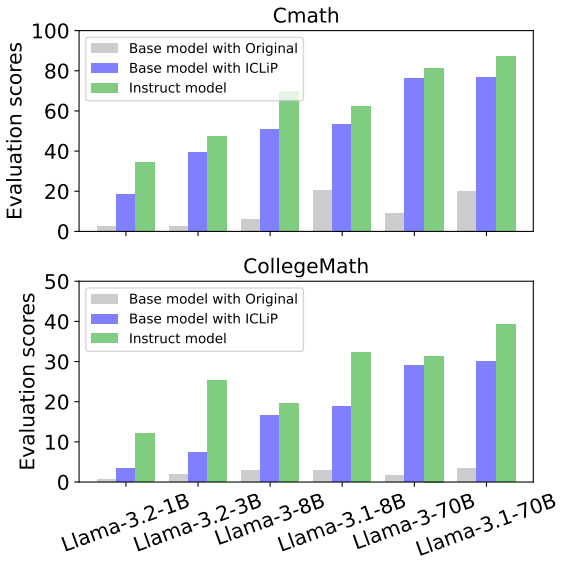


Figure 4: With Llama family models, ICLiP ensures more consistent and comparable evaluation scores between base models and instruct models for Cmath and CollegeMath.

that BOSE enables pre-trained models to achieve comparable performance to their instruction-tuned models, ensures more reliable evaluation results.

#### 4.3.2 Experiments on Our Pre-trained Models

Similarly, we conduct another experiment on a series of our pre-trained base models(16 checkpoints in total, with each undergoing supervised fine-tuning to obtain the corresponding instruct model). As shown in Table 3, by calculating the rank correlation of the scoring sequences between base and instruct models for each benchmark, we observe an effective improvement in 5 out of 6 benchmarks, with an average Kendall's $\tau$ enhancement of 0.188.

We conclude that, ICLiP ensures more consistent capabilities between base and instruct models, and enables base models to achieve more reliable scores as well. This advancement provides more

| Models | Methods | Mathematics | | | | | Reasoning | AVG |
|---|---|---|---|---|---|---|---|---|
| | | CMath | MGSM_zh | Gaokao2023EN | CollegeMath | Minerva Math | Multi_LogiEval | |
| Llama-3.1-8B | instruct_0shot | 20.58 | 1.60 | 1.04 | 2.91 | 0.74 | 0.00 | 4.48 |
| | instruct_fewshot | 52.46* | 3.60 | 1.04 | 0.32 | 6.99 | 7.96 | 12.06 |
| | light-instruction_0shot | 32.51 | 16.00 | 9.35 | 9.17 | 9.17 | 18.47 | 15.78 |
| | **ICLiP** | **53.37** | **43.60** | **23.38** | **18.86** | **11.03** | **71.03** | **36.88** |
| Llama-3.1-70B | instruct_0shot | 17.67 | 11.60 | 5.97 | 3.26 | 0.74 | 0.00 | 6.54 |
| | instruct_fewshot | 67.76 | 2.80 | 0.26 | 0.50 | 6.25 | 0.05 | 12.94 |
| | light-instruction_0shot | 45.26 | 47.60 | 16.88 | 22.51 | 0.00 | 21.03 | 25.55 |
| | **ICLiP** | **76.05** | **59.60** | **36.62** | **30.77** | **17.28** | **74.96** | **49.21** |
| Gemma-2-9B | instruct_0shot | 17.30 | 10.00 | 2.08 | 1.31 | 3.31 | 0.05 | 5.68 |
| | instruct_fewshot | 68.21* | 5.20 | 0.78 | 0.43 | 7.72 | 49.56 | 21.98 |
| | light-instruction_0shot | 5.01 | 26.80 | 14.29 | 26.76 | 0.00 | 7.86 | 13.45 |
| | **ICLiP** | **69.13** | **55.20** | **33.77** | **33.56** | **16.18** | **61.65** | **44.92** |
| Gemma-2-27B | instruct_0shot | 3.64 | 8.00 | 8.57 | 2.20 | 1.84 | 17.94 | 7.03 |
| | instruct_fewshot | **75.96** | 3.20 | 0.78 | 0.21 | 8.46 | 49.34 | 22.99 |
| | light-instruction_0shot | 24.41 | 55.20 | 17.14 | 29.30 | 0.00 | 17.42 | 23.91 |
| | **ICLiP** | 75.87* | **66.40** | **40.52** | **35.11** | **21.32** | **75.68** | **52.48** |
| Qwen2.5-7B | instruct_0shot | 30.69 | 61.60 | **53.51** | 39.62* | 22.06 | **58.54** | 44.34 |
| | instruct_fewshot | 80.87 | 10.80 | 14.29 | 11.86 | **37.87** | 57.67 | 35.56 |
| | light-instruction_0shot | 83.97 | 69.60 | 22.34 | 33.65 | 33.65 | 46.08 | 48.22 |
| | **ICLiP** | **88.07** | **70.80** | 51.43 | **40.45** | 23.53 | 57.81* | **55.35** |
| Qwen2.5-72B | instruct_0shot | 74.32 | 68.40 | **61.30** | **42.28** | **36.03** | 68.89 | 58.54 |
| | instruct_fewshot | 85.52* | 6.40 | 4.94 | 12.89 | 34.93 | 71.54 | 36.04 |
| | light-instruction_0shot | **85.97** | 79.20 | 40.78 | 38.37 | 0.00 | 34.08 | 46.40 |
| | **ICLiP** | 83.06 | **81.60** | 55.06 | 40.26 | 29.04 | **74.00** | **60.50** |

Table 4: Ablation study results on 6 open-source models. Scores in **bold** indicate the highest scores among 4 methods, while scores with "*" indicate sub-optimal scores within 1 point gap.

meaningful guidance for assessing the performance of pre-trained models and selecting optimal checkpoints for post-training.

### 4.4 Ablation Study

To investigate whether all components of the proposed ICLiP methodology are indispensable, we conduct ablation experiments here. The following variant prompt templates are considered as comparative methods: (1) **instruct_0shot**: instruct prompt with 0-shot, also referred to as the original method, (2) **instruct_fewshot**: instruct prompt with few-shot, (3) **light-instruction_0shot**: light-instruction prompt with 0-shot, (4) **ICLiP**, our proposed method. Here, `light-instruction` refers to the prompt template used in our proposed method, while `instruct prompt` is the prompt template typically used for instruct model evaluation. Please refer to Appendix A for further implementation details.

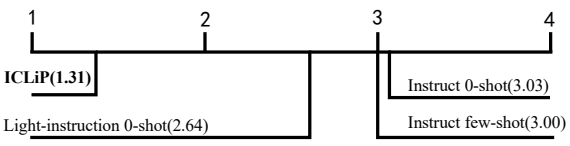


Figure 5: Average ranks of different methods, with a lower rank indicating better performance.

As demonstrated in Table 4, ICLiP achieves optimal performance in 28 out of 36 experiment results and suboptimal performance in 2 cases with marginal differences (within 1-point gap). In terms of different LLM families, ICLiP shows dominant superiority for models from Llama and Gemma families, while achieving competitive performance in majority of experimental results with Qwen family. Moreover, the proposed method shows consistently superior average performance across different benchmarks, confirming the effectiveness and generalizability of ICLiP in evaluating base models.

More interestingly, we introduce a statistical analysis of the average rankings of 4 methods (ranking among 4 methods for each experiment, and then averaging across 36 experiments, with 1 representing the best theoretically). As depicted in Figure 5, ICLiP behaves optimally with an excellent average rank of `1.31`. Additionally, the `light-instruction` format consistently performs better than the instruct prompt, and the few-shot results always outperform the 0-shot results under the same prompt template, highlighting the effectiveness of each component in ICLiP.

To further sourcing the improvements of our proposed method, we conduct two supplementary investigations, and once again proves the validity of the proposed approach. Evaluation details can be found in Appendix B .

## 5 Conclusion

In this paper, we propose a systematic evaluation method tailored to base model, named BOSE, specifically designed to enhance the stability of base model evaluations during pre-training and ensure consistency with evaluations of instruct models. For open-ended tasks, BOSE develops a few-shot prompt template with light-instruction to guide the base model more effectively. For multi-choice tasks, BOSE innovatively transforms the standard perplexity (ppl) metric into a fill-in-the-blank format. These adaptations align better with the inherent characteristics of base models and facilitate more accurate evaluations to reflect the real capabilities of base models. Moreover, we are the first to adopt Kendall's rank correlation to quantitatively assess the stability and consistency of base models' evaluation. This metric provides a robust and reliable way to compare the evaluation results of the base models with those of corresponding instruct models. Extensive experiments demonstrate the effectiveness and superiority of our proposed BOSE, validating its potential to significantly improve the evaluation and provide more meaningful insights into the true capabilities of base models.

## 6 Limitations

Our work aims to address some critical issues in base model evaluation, such as instability during pre-training and lack of consistency with instruct models. There may be some subjective biases in our comparative methods, such as the design of instruction prompt templates. As for the considered models, we use our pre-trained checkpoints and 3 open-source LLM families, and there are more open-source LLMs worth exploring, such as DeepSeek (DeepSeek-AI, 2024), ChatGLM (GLMTeam, 2024), Mistral (Jiang et al., 2023). At the same time, we mainly conduct experiments on three categories of benchmarks, and some other capability categories (such as code, reading comprehensive, etc.) also need to be considered to provide more comprehensive and enriching experimental results. In addition, our method still has some flaws in some experiments, and we are consistently studying and optimizing the proposed method.

## Acknowledgments

We sincerely thank the anonymous reviewers for their valuable comments.

## References

Abhimanyu Dubey Aaron Grattafiori and et al. Abhinav Jauhri. 2024. The llama 3 herd of models. Preprint, arXiv:2407.21783.

Rishabh Agarwal, Avi Singh, Lei M. Zhang, Bernd Bohnet, Luis Rosias, Stephanie Chan, Biao Zhang, Ankesh Anand, Zaheer Abbas, Azade Nova, John D. Co-Reyes, Eric Chu, Feryal Behbahani, Aleksandra Faust, and Hugo Larochelle. 2024. Many-shot in-context learning. Preprint, arXiv:2404.11018.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. Chatbot arena: An open platform for evaluating llms by human preference. Preprint, arXiv:2403.04132.

DeepSeek-AI. 2024. Deepseek-v3 technical report. Preprint, arXiv:2412.19437.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Tianyu Liu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. 2024. A survey on in-context learning. Preprint, arXiv:2301.00234.

Hao Fu, Yao; Peng and Tushar Khot. 2022. How does gpt obtain its ability? tracing emergent abilities of language models to their sources. Yao Fu's Notion.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2024. A framework for few-shot language model evaluation.

GemmaTeam. 2024. Gemma 2: Improving open language models at a practical size. Preprint, arXiv:2408.00118.

GLMTeam. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. Preprint, arXiv:2406.12793.

Zishan Guo, Renren Jin, Chuang Liu, Yufei Huang, Dan Shi, Supryadi, Linhao Yu, Yan Liu, Jiaxuan Li, Bojian Xiong, and Deyi Xiong. 2023. Evaluating large language models: A comprehensive survey. Preprint, arXiv:2310.19736.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021a. 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. 2021b. Measuring mathematical problem solving with the math dataset. Preprint, arXiv:2103.03874.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. Preprint, arXiv:2310.06825.

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. Preprint, arXiv:2001.08361.

M. Kendall. 1938. A new measure of rank correlation. Biometrika, 30:81–93.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, Yuhuai Wu, Behnam Neyshabur, Guy Gur-Ari, and Vedant Misra. 2022. Solving quantitative reasoning problems with language models. Preprint, arXiv:2206.14858.

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. Preprint, arXiv:2306.09212.

Minpeng Liao, Wei Luo, Chengxi Li, Jing Wu, and Kai Fan. 2024. Mario: Math reasoning with code interpreter output – a reproducible pipeline. Preprint, arXiv:2401.08190.

LingTeam. 2025. Every flop counts: Scaling a 300b mixture-of-experts ling llm without premium gpus. Preprint, arXiv:2503.05139.

Kaijing Ma, Xinrun Du, Yunran Wang, Haoran Zhang, Zhoufutu Wen, Xingwei Qu, Jian Yang, Jiaheng Liu, Minghao Liu, Xiang Yue, Wenhao Huang, and Ge Zhang. 2024. Kor-bench: Benchmarking language models on knowledge-orthogonal reasoning tasks. Preprint, arXiv:2410.06526.

Jerry Tworek Mark Chen, Heewoo Jun, and et al. 2021. Evaluating large language models trained on code.

Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. 2024. Large language models: A survey. Preprint, arXiv:2402.06196.

OpenAI. 2024. Gpt-4 technical report. Preprint, arXiv:2303.08774.

OpenCompass. 2023. Opencompass: A universal evaluation platform for foundation models. https://github.com/open-compass/opencompass.

Nisarg Patel, Mohith Kulkarni, Mihir Parmar, Aashna Budhiraja, Mutsumi Nakamura, Neeraj Varshney, and Chitta Baral. 2024. Multi-logieval: Towards evaluating multi-step logical reasoning ability of large language models. Preprint, arXiv:2406.17169.

QwenTeam. 2025. Qwen2.5 technical report. Preprint, arXiv:2412.15115.

Alec Radford and Karthik Narasimhan. 2018. Improving language understanding by generative pre-training.

Nishant Balepur Sander Schulhoff, Michael Ilie and et al. 2024. The prompt report: A systematic survey of prompting techniques. Preprint, arXiv:2406.06608.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. 2022. Language models are multilingual chain-of-thought reasoners. Preprint, arXiv:2210.03057.

Zayne Sprague, Xi Ye, Kaj Bostrom, Swarat Chaudhuri, and Greg Durrett. 2024. Musr: Testing the limits of chain-of-thought with multistep soft reasoning. Preprint, arXiv:2310.16049.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V. Le, Ed H. Chi, Denny Zhou, and Jason Wei. 2022. Challenging big-bench tasks and whether chain-of-thought can solve them. Preprint, arXiv:2210.09261.

Zhengyang Tang, Xingxing Zhang, Benyou Wang, and Furu Wei. 2024. Mathscale: Scaling instruction tuning for mathematical reasoning. Preprint, arXiv:2403.02884.

Nick Ryder Tom B. Brown, Benjamin Mann and et al. 2020. Language models are few-shot learners. Preprint, arXiv:2005.14165.

Xuezhi Wang and Denny Zhou. 2024. Chain-of-thought reasoning without prompting. Preprint, arXiv:2402.10200.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. 2024. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. arXiv preprint arXiv:2406.01574.

Jason Wei, Nguyen Karina, Hyung Won Chung, Yunxin Joy Jiao, Spencer Papay, Amelia Glaese, John Schulman, and William Fedus. 2024. Measuring short-form factuality in large language models. Preprint, arXiv:2411.04368.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023a. Chain-of-thought prompting elicits reasoning in large language models. Preprint, arXiv:2201.11903.

Tianwen Wei, Jian Luan, Wei Liu, Shuang Dong, and Bin Wang. 2023b. Cmath: Can your language model pass chinese elementary school math test? Preprint, arXiv:2306.16636.

## A Prompt Templates

Prompt templates in different methods are illustrated in table 5 to table 10, instruct_0shot and Option-ppl are regarded as original methods in our experiments for open-ended tasks and multi-choice tasks respectively.

**PROMPT**

Question: {question}
A. {content of optionA}
B. {content of optionB}
C. {content of optionC}
D. {content of optionD}
Answer: A

Table 5: Prompt of Option-PPL, where the underlined part is used for calculating the Perplexity (PPL).

**PROMPT**

{question} {content of optionA}

Table 6: Prompt of Blank-PPL, where the underlined part is used for calculating the Perplexity (PPL).

**PROMPT**

You're a {domain} expert. Given the following question, please reason step by step and put your final answer within boxed{}.
{question}

Table 7: Prompt of Instruct_0shot

## B Ablation study for sourcing the improvement

We provide some additional ablation studies to further source the improvements of our proposed method. Specifically, with open-ended tasks in consideration, we take Instruct_0shot as a comparative method for simplicity and conduct two supplementary investigations:

**i. Comparable evaluation with math_verify**[4]. We take the newly released mathematical expression evaluation package math_verify from huggingface to conduct the post-processing and judgment, to identify potential limitations in our current evaluation process.

[4]https://github.com/huggingface/Math-Verify

**PROMPT**

You're a {domain} expert. Given the following question, please reason step by step and put your final answer within boxed{}.
{question_1}
{answer_1}

...

{question_k}
{answer_k}

{question}

Table 8: Prompt of Instruct_fewshot

**PROMPT**

Problem: {question}
Solution:

Table 9: Prompt of Light-instruction_0shot

**PROMPT**

Problem: {question_1}
Solution: {answer_1}

...

Problem: {question_k}
Solution: {answer_k}

Problem: {question}
Solution:

Table 10: Prompt of ICLiP

| Benchmark | Instruct_0shot | | | | ICLiP | | | |
|---|---|---|---|---|---|---|---|---|
| | current | math_verify | upper bound | ratio | current | math_verify | upper bound | ratio |
| CMATH | 20.58 | 16.21 | 22.67 | 0.715 | 53.47 | 52.10 | 53.91 | 0.966 |
| Gaokao2023EN | 1.04 | 0.26 | 4.94 | 0.053 | 23.38 | 23.38 | 24.42 | 0.957 |
| CollegeMath | 2.91 | 9.22 | 9.94 | 0.928 | 18.86 | 23.42 | 24.41 | 0.959 |

Table 11: Ablation study results for sourcing the improvement. $current$: results with current evaluation metrics; $math\_verify$ and $upperbound$: results evaluated with Math_verify and LLM judge respectively; $ratio$: score with math_verify / score with LLM judge, indicating the extent to which the LLM's predictions follow the format in different methods.

| | 1B | 2B |
|---|---|---|
| vocab size | 126464 | 126464 |
| layer num | 22 | 22 |
| hidden size | 2048 | 2560 |
| intermediate size | 5632 | 10240 |
| attention heads | 32 | 40 |
| key value heads | 4 | 8 |
| # Para. (B) | 1.487 | 2.724 |
| # Non-emb. Para. (B) | 0.969 | 2.076 |
| sequence length | 4096 | 4096 |
| learning rate | 6.37E-04 | 5.23E-04 |
| batch size | 806 | 1227 |

Table 12: The details of architecture and training parameters for pre-trained models in Section 4.1.2

**ii. Loose judgement with LLM as upper bound**. To further investigate whether our proposed method effectively ensuring the model's problem-solving capabilities or simply fitting the answer extraction pattern, we employ GPT-4o to assess the correctness of Base model's prediction, where only question (NOT prompt) along with prediction and gold answer are provided for judgment and thus eliminate the influence of answer extraction. These scores are regarded as upper-bound evaluation results, with no constraint imposed on the answer format.

As shown in Table 11, we observe that our improvement sources from two aspects:

**Better reflection of true capability**. While evaluation scores slightly differ across different post-processing methods (relatively bigger difference in the CollegeMath), our proposed method consistently outperforms the comparative methods(i.e., Instruct_0shot).

**Higher fitness to the answer extraction**. ICLiP exhibits superior alignment with loose scores(upper-bound) compared to both current evaluation process and math-verify judgement, thereby enabling higher tolerance for the post-processing.

The above results further demonstrate the effectiveness of the proposed method.

## C Architecture and Pretraining Setup

We use a GQA (Grouped Query Attention) architecture based on the standard decoder-only Transformer, comprising an embedding layer, alternating layers of attention mechanisms and feed-forward networks. Positional information is handled using RoPE (Rotary Positional Embedding). For training parameters, the pre-trained models are initialized with a standard deviation of $0.006$ and optimized using AdamW, with hyperparameters set to $\beta_1 = 0.9$, $\beta_2 = 0.95$, $\epsilon = 1e^{-8}$ and `weight_decay` = 0.1. The learning rate follows a WSD (Warm-up, Stabilization, Decay) strategy, where the first $1\%$ of training steps involve linear warm-up. Further architectural and pretraining details are available in Table 12 and our technical report (LingTeam, 2025).