

Fusion Training for Mathematical Generalization in Large Language Models

Congfeng Cao¹, Pengyu Zhang², Jelke Bloem¹

¹Institute for Logic, Language and Computation, University of Amsterdam

²INDE Lab, University of Amsterdam,

Abstract

Thinking Mode Fusion (TMF) enables large language models to support both concise responses and long-form reasoning by unifying a non-thinking mode and a thinking mode within a single model. However, its training dynamics, including the *data ratio* and *training schedule* between the two modes, remain underexplored. In this work, we present a systematic study of TMF by analyzing the effects of the training schedule and data ratio between thinking and non-thinking modes. Focusing on mathematical problem solving, we construct a benchmark with multiple thinking-to-non-thinking data ratios and three training schedules. Our results reveal an asymmetric interaction between the two modes: increasing the ratio of non-thinking supervision reduces the accuracy of the thinking mode. We further show that different training schedules modulate this trade-off and that the optimal schedule depends on the data ratio. Finally, we quantify a negative correlation between non-thinking and thinking mode supervision, highlighting an inherent tension between these two modes. These findings provide practical guidance for designing effective TMF training settings. All code and data are released to support further research at: [Fusion Bench](#).

1 Introduction

Large language models (LLMs) have made substantial progress in *long-form reasoning* and have been increasingly applied to complex tasks via extended chains of thought (Wei et al., 2022; Kojima et al., 2022; Wang et al., 2023). However, many real-world queries are comparatively simple and do not require long-form reasoning. For such cases, long-form reasoning can be unnecessarily time-consuming and computationally expensive.

As a result, existing efforts have explored switching models between simple queries and complex problems (OpenAI, 2024, 2025), or adapting the length of model responses to improve reasoning

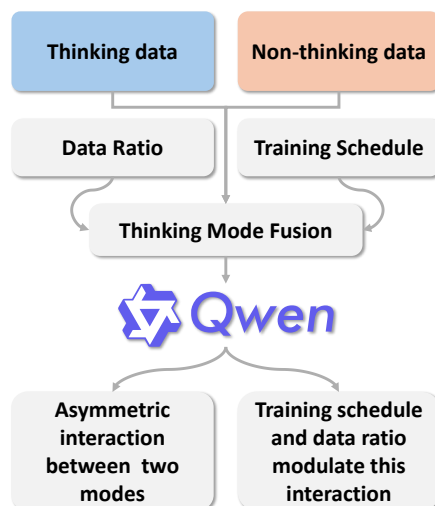


Figure 1: An overview of the experimental design for analyzing Thinking Mode Fusion under different training orders and data ratios.

efficiency (Arora and Zanette, 2025; Kimi et al., 2025; Shen et al., 2025; Wu et al., 2025; Zhang et al., 2025). Qwen3 series models address this limitation by introducing **Thinking Mode Fusion (TMF)** during post-training (Yang et al., 2025).

TMF unifies a *thinking* mode and a *non-thinking* mode within a single model, enabling users to control reasoning behavior for both long-form reasoning and concise responses. TMF is implemented through two different chat templates, i.e., a thinking chat template and a non-thinking chat template, that support mode selection. This design reduces deployment complexity and inference costs while allowing the model to adjust its reasoning depth under different computational budgets.

Despite Qwen3’s strong empirical performance, the mechanisms of TMF and the interaction between the two modes remain underexplored. In particular, it remains unclear how the *training schedule*, i.e., the training order of the two modes, and the *data ratio*, i.e., the combination of thinking and non-thinking data, affect overall model perfor-

mance. This raises a practical question: **when both modes are trained within a single model, how do the training schedule and the data ratio affect the model’s performance?**

To investigate this question, we focus on mathematics, where problem difficulty varies widely, from short-solution questions to Olympiad-level tasks that require long-form reasoning. As shown in Figure 1, we systematically investigate the interaction between the two modes and analyze the effects of TMF under different training schedules and varying ratios of thinking to non-thinking data. To support this study, we construct a benchmark, **Fusion Bench**, in accordance with the Qwen3 design and consisting of two types of mathematical problems: simple problems for non-thinking mode training and Olympiad-level problems for thinking mode training under diverse data ratios and training mode combinations. Our results reveal an asymmetric interaction between the two modes: the trade-off between the two modes shows that thinking mode performance degrades as non-thinking supervision increases. Note that TMF unifies two modes (long-form reasoning and concise response) within a single model for a single task, which is different from multi-task settings such as math and code. Moreover, we show that the training schedule modulates this trade-off and that the optimal schedule depends on the data ratio. Finally, we quantify a negative correlation between non-thinking and thinking mode supervision, highlighting an inherent tension between these two modes.

Our main contributions are as follows:

- We conduct a systematic study of training schedules and data ratios in TMF, revealing a trade-off between thinking and non-thinking modes, where increased non-thinking supervision degrades thinking performance, and we quantify a negative correlation between the two modes.
- We show that training schedules modulate this trade-off: while the optimal schedule depends on the data ratio, the Mix training schedule consistently achieves strong performance across both thinking and non-thinking modes when averaged over different ratios.
- We construct and release Fusion Bench ¹, a benchmark comprising thinking and non-thinking datasets with varying data ratios and

¹<https://anonymous.4open.science/r/Fusion-Bench-2D62>

training schedules, enabling controlled and reproducible analysis of TMF.

2 Related Work

Thinking Mode Fusion and Hybrid Thinking.

Large language models (LLMs) have shown that supervised fine-tuning (SFT) on long chain-of-thought (Wei et al., 2022) traces is able to improve performance on complex tasks.

However, enhancing reasoning depth through specialized post-training increases computational costs due to the long reasoning responses. Models such as OpenAI o1 (OpenAI, 2024), DeepSeek-R1 (DeepSeek-AI, 2025), and Gemini (Team, 2025) widely adopt hybrid thinking mechanisms that control whether the model engages in reasoning, thereby achieving a more efficient and flexible reasoning process. Existing efforts to improve reasoning efficiency focus on reducing the length of model responses (Arora and Zanette, 2025; Kimi et al., 2025; Shen et al., 2025). Wu et al. (2025) merge reasoning and non-reasoning models to reduce output length. AdaptThink proposes teaching reasoning models to choose the optimal thinking mode adaptively based on problem difficulty (Zhang et al., 2025). Qwen3 (Yang et al., 2025) introduced Thinking Mode Fusion (TMF), which allows a single model to support both long-form reasoning and concise responses controlled by the chat format. Wang et al. (2025) further demystify hybrid thinking, revealing that current hybrid thinking LLMs only achieve partial mode separation, where reasoning behaviors often leak into the no-think mode. However, the mechanisms of TMF remain underexplored.

Data Mixture. Prior studies have shown that data mixture plays a critical role in both pre-training and post-training stages. Ye et al. (2025a) outlined the quantitative predictability of model performance with respect to mixture proportions. Liu et al. (2025) propose RegMix to automatically identify a high-performing data mixture by formulating it as a regression task. Li et al. (2025) frame the selection of data proportions as an optimization problem to minimize validation loss across multiple tasks during SFT.

While these studies provide insights into domain-level mixing (e.g., balancing math, code, and chat), they lack an examination of the internal interference between different response formats within the same domain, which is a key focus of our study.

Math Datasets. We group math datasets into two categories that align with the thinking (with long reasoning) versus non-thinking (with short solutions) modes considered in this paper.

On the simple data side, a complementary line of work focuses on datasets that admit short solutions or direct answer supervision, which better aligns with the non-thinking mode. Classic benchmarks such as GSM8K (Cobbe et al., 2021), which consists of grade-school math problems, are widely used for instruction tuning and evaluation, and are often employed in short-solution settings to encourage concise generation. Calc-X (Kadlčík et al., 2023) is a collection of simple arithmetic-focused math word problems, comprising over 300,000 curated samples. Patel et al. (2021) introduced MAWPS, an elementary-level benchmark dataset featuring concise solutions and final answers. In this work, we use GSM8K as the non-thinking mode base dataset, as it covers a diverse range of math problems and is widely adopted.

On the long reasoning data side, recent work has released large-scale datasets that explicitly provide long reasoning traces, which are crucial for training and analyzing long-form mathematical reasoning. OpenMathReasoning (Moshkov et al., 2025) is a generated large-scale math reasoning dataset for training large language models (LLMs), containing 3.2 million long chain-of-thought (CoT) solutions. OpenMathInstruct-1 (Toshniwal et al., 2024) and OpenMathInstruct-2 (Toshniwal et al., 2025) provide 1.8 million and 14 million generated solutions, respectively, covering a wide range of mathematical difficulty levels. We also note curated math reasoning benchmarks such as OMNI-MATH (Gao et al., 2025) and LIMO (Ye et al., 2025b), which are commonly used to evaluate advanced reasoning capabilities. In addition, several datasets emphasize challenging questions paired with chain-of-thought annotations, including Skywork-MathQA (Zeng et al., 2024), which contains 2.5 million question-answer pairs, and NuminaMath (LI et al., 2024), which consists of 860,000 competition-level problems with long-form reasoning. For ease of evaluation, we choose OpenMathReasoning since it provides problems with exact and verifiable final answers, unlike many other datasets that include proof-based or open-ended reasoning tasks without checkable solutions. In this work, we sample data from the OpenMathReasoning subset generated by DeepSeek-R1 (DeepSeek-AI, 2025) and select examples with a high pass rate (greater than 0.96)

Table 1: Benchmark statistics for thinking and non-thinking training data under different data ratios. **T** denotes thinking mode data and **NT** denotes non-thinking mode data. We fix the thinking mode subset size ($N_T = 1500$) and vary the non-thinking subset size (N_{NT}) accordingly. **Mean** denotes the mean length of the reasoning traces.

Mode	Ratio (T : NT)	Count	Mean	Min	Max
Non-thinking	4 : 1	375	671	330	1,540
	3 : 1	500	672	330	1,540
	2 : 1	750	681	330	1,540
	1 : 1	1,500	692	298	1,817
	1 : 2	3,000	692	298	1,817
	1 : 3	4,500	695	298	1,863
	1 : 4	6,000	694	298	1,863
Thinking	-	1,500	11,877	1,903	54,786

under the 72B model as the thinking-mode training dataset.

3 Data Processing

As mentioned above (Section 2), we select GSM8K (Cobbe et al., 2021) and OpenMathReasoning (Moshkov et al., 2025) as the base datasets for non-thinking and thinking modes, respectively, which are in accordance with the Qwen3 mode design for simple and complex problems.

Non-thinking mode dataset. For the non-thinking mode, we use the GSM8K training set. For each experimental setting, we subsample GSM8K into a subset of size N_{NT} according to a target ratio ρ between thinking (N_T) and non-thinking (N_{NT}) data. We define the ratio as

$$\rho = \frac{N_T}{N_{NT}}. \quad (1)$$

We sweep ρ from 4 to 1/4 (i.e., non-thinking to thinking ratios from 4:1 to 1:4). Since the GSM8K training set contains only 7,470 examples, to satisfy our experimental design for the data ratio, we fix the amount of thinking mode data to 1,500 examples. With the thinking mode training data fixed at 1,500, the corresponding non-thinking training data size ranges from 375 to 6,000.

Thinking mode dataset. To enable controlled comparisons across different fusion strategies and data ratios, we fix the number of thinking mode training examples to $N_T = 1,500$ throughout our benchmark. The OpenMathReasoning dataset contains 3.2 million long chain-of-thought solutions; from this corpus, we sample 1,875 examples. We then randomly split these samples into a 8:2

Table 2: Statistics of test data for thinking and non-thinking evaluation. **Mean** denotes the mean length of the reasoning traces, **Min** denotes the minimum length of the reasoning traces, and **Max** denotes the maximum length of the reasoning traces.

Mode	Count	Mean	Min	Max
Non-thinking	375	696	298	1,863
Thinking	375	11,992	1,840	69,268

train/test partition. Detailed statistics of the thinking and non-thinking training data are presented in Table 1.

For uniform evaluation, we sample 375 test examples from the GSM8K and OpenMathReasoning test sets for the non-thinking and thinking modes, respectively. Detailed statistics of the thinking and non-thinking test data are presented in Table 2. The thinking mode test data have a mean length of 11,992, with a maximum length of 69,268 and a minimum length of 1,840. In contrast, the non-thinking mode test data have a mean length of 696, with a maximum length of 1,863 and a minimum length of 298.

After data selection, to align with the Qwen3 official chat template settings, we further process the GSM8K subset into non-thinking mode data and the OpenMathReasoning subset into thinking mode data. For non-thinking mode data, we insert the `<think></think>` tags before the solutions. For thinking mode data, we wrap the long-form reasoning solutions with the `<think></think>` tags. To simplify the evaluation process, we wrap all final answers in the `\boxed{}` tag. Examples of non-thinking and thinking data are shown in Appendix A.4.

4 Methodology

We study Thinking Mode Fusion (TMF) as a post-training problem for large language models (Qwen3 models), where the model support two distinct modes: (i) long-form reasoning for hard problems and (ii) concise responses for simple problems. Our goal is to evaluate how TMF is affected by two controllable factors during supervised fine-tuning (SFT): **(a) training schedule** (the ordering or interleaving of thinking and non-thinking mode training) and **(b) data ratio** (the relative amount of thinking versus non-thinking data, ranging from 1:4 to 4:1, while fixing the amount of thinking mode training data). We use the non-thinking and thinking test sets constructed in Section 3 to evaluate

the uniformity of model performance across modes. Figure 2 provides an overview of our experimental design.

4.1 Supervised Fine-Tuning with Two Modes

We construct two types of training data (thinking and non-thinking) and feed them into a unified supervised fine-tuning (SFT) framework with different training schedules and data ratios.

As mentioned in Section 3, we convert examples into the model’s chat template, with the mode specified by the presence or absence of reasoning content within the `<think></think>` tags. For thinking-mode training, we use the constructed thinking-mode training data, in which long-form reasoning solutions are wrapped in the `<think></think>` tags. For non-thinking-mode training, we use the constructed non-thinking-mode training data, which contain short solutions following the `<think></think>` tags.

After training, we extract final answers from the `\boxed{}` tag, as described in Section 3. Finally, we evaluate model performance in both thinking and non-thinking modes using the corresponding test sets in Table 2.

4.2 Data Ratio Control

As described in Section 3, we construct thinking mode and non-thinking mode training data with a wide range of ratios $\{1:4, 1:3, 1:2, 1:1, 2:1, 3:1, 4:1\}$ by fixing the number of thinking mode training examples.

4.3 Training Schedule

Given a fixed thinking set \mathcal{D}_T and a ratio-specific non-thinking subset $\mathcal{D}_{NT}^{(\rho)}$, TMF training reduces to constructing a single SFT stream \mathcal{S} by a scheduling function

$$\mathcal{S} = \text{SCHEDULE}(\mathcal{D}_T, \mathcal{D}_{NT}^{(\rho)}; m), \quad (2)$$

where m specifies how the two modes are arranged during training. We investigate three schedules that cover sequential curricula and interleaving.

Sequential: T-NT. We place all thinking mode examples before all non-thinking mode examples:

$$\mathcal{S}_{\text{T-NT}} = [\mathcal{D}_T \parallel \mathcal{D}_{NT}^{(\rho)}]. \quad (3)$$

This schedule tests whether updates from non-thinking data applied after reasoning training dilute or overwrite reasoning behaviors, and whether reasoning training provides transferable benefits to subsequent concise answering.

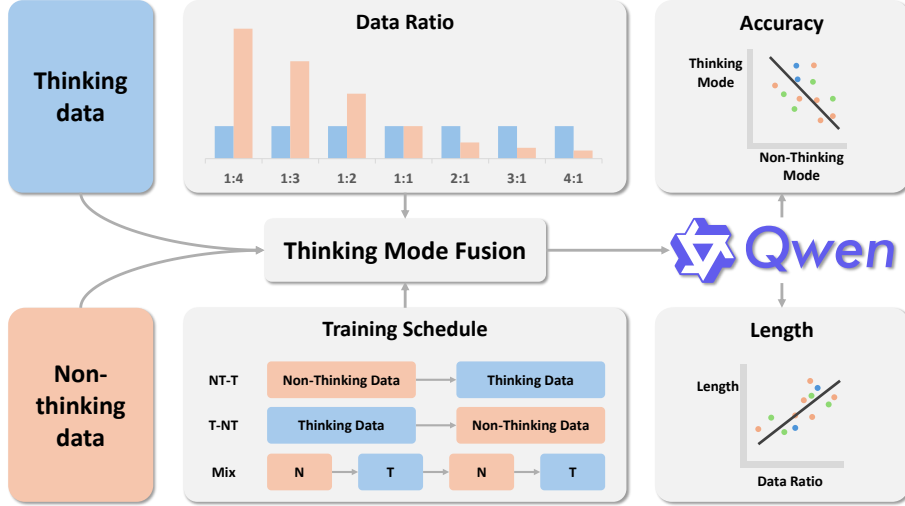


Figure 2: Overview of Thinking Mode Fusion training design. We construct two datasets: thinking mode and non-thinking mode. We vary the data ratio (T:NT from 1:4 to 4:1) and schedule the resulting SFT stream using three strategies: **T-NT** (first thinking then non-thinking), **NT-T** (non-thinking then thinking), and **Mix** (interleaved to match the target ratio). Each configuration fine-tunes Qwen3 model and is evaluated on both thinking and non-thinking benchmarks to quantify interference and trade-offs.

Sequential: NT-T. We reverse the order:

$$\mathcal{S}_{NT-T} = [\mathcal{D}_{NT}^{(\rho)} \parallel \mathcal{D}_T]. \quad (4)$$

This schedule investigates the opposite direction of interference: whether late-stage reasoning supervision disrupts concise response patterns learned from non-thinking data and whether non-thinking alignment persists under subsequent long-reasoning updates.

Interleaved: Mix. We interleave the two modes of data to the target ratio during training. In particular, we use a simple 1-to- n interleaving procedure: when $\rho \leq 1$ (thinking is less frequent), we interleave **1 T (thinking mode)** followed by n **NT (non-thinking mode)** examples, with $n \approx \text{round}(1/\rho)$; when $\rho > 1$ (thinking is more frequent), we interleave **1 NT** followed by n **T** examples, with $n \approx \text{round}(\rho)$. This schedule makes the intended TMF usage where the model repeatedly alternates between the two modes, enabling us to test whether frequent mode switching alleviates the degradation observed under purely sequential curricula.

4.4 Full Factorial Design

Our TMF analysis is defined by the Cartesian product of data ratios and schedules:

$$(\rho, m) \in \{4, 3, 2, 1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}\} \times \{\text{T-NT}, \text{NT-T}, \text{Mix}\}. \quad (5)$$

For each configuration (ρ, m) , we build the corresponding SFT training \mathcal{S} while keeping the base

model, supervision format, and optimization hyperparameters fixed. This controlled setup allows us to attribute performance changes to (i) how much thinking supervision is mixed with non-thinking data and (ii) how the two modes are scheduled during training, thereby directly quantifying interference and trade-offs induced by TMF. We finally evaluate both the accuracy and the reasoning length of model responses, as reasoning length has been shown to reflect reasoning quality in mathematical problem solving (Sprague et al., 2025; Jin et al., 2024; Tutek et al., 2025).

5 Experiment and Analysis

5.1 Experiment Setup

Our study aims to answer four research questions:

RQ1: Do the non-thinking and thinking modes affect each other’s performance?

Although Qwen3 adopts Thinking Mode Fusion to separate long-form and concise reasoning into different fine-tuning modes, training these two modes jointly may still lead to performance trade-offs or interference in the mathematical domain.

RQ2: How does the different training schedule affect Thinking Mode Fusion?

Thinking Mode Fusion is introduced by first fine-tuning on concise reasoning traces and then fine-tuning on long-form reasoning traces (Yang et al., 2025). However, in practice, non-thinking and thinking modes can be trained in different orders. As de-

scribed in Section 4.3, we consider three training schedules, {T-NT, NT-T, Mix}, and evaluate model performance under different training schedules.

RQ3: How does the different data ratio affect Thinking Mode Fusion? In real-world applications, short-form reasoning data are far more abundant than high-quality long-form reasoning data. Therefore, the data ratio between thinking and non-thinking modes is crucial for understanding the effectiveness of Thinking Mode Fusion.

RQ4: What is the quantitative relationship between the non-thinking and thinking modes? After analyzing the effects of training schedules and data ratios, it is essential to quantify the relationship between the two modes to better guide TMF training for LLMs.

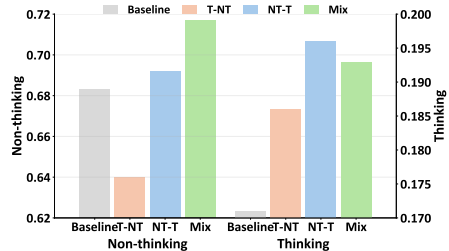
All detailed experimental settings and baselines can be found in Appendix A.

5.2 RQ1: Interaction between Non-Thinking and Thinking Mode

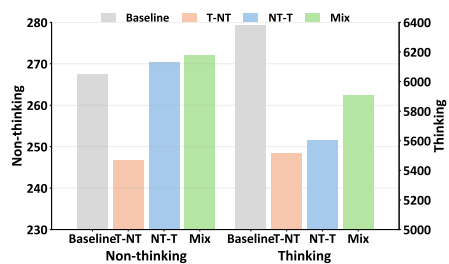
The accuracy and reasoning length performance on the test set under training with different data ratios and training schedule settings are presented in Table 3. Rows correspond to three schedules with two modes: T-NT (thinking before non-thinking), NT-T (non-thinking before thinking), and Mix (interleaved training). Columns are organized by the thinking to non-thinking (T:NT) data ratio from 1:4 to 4:1, reporting Len (average response length) and Acc (accuracy) for three runs. The rightmost Mean column averages results across all data ratios, while the Mean row in each mode block averages over training schedules.

The results indicate that the non-thinking and thinking modes interact during Thinking Mode Fusion (TMF) training. When averaging results across all data ratios, the Mix training schedule achieves the highest accuracy on both the thinking and non-thinking test sets, with accuracies of 0.208 and 0.695, respectively. In thinking-mode evaluation, the NT-T training schedule also attains the highest accuracy while producing the longest average response length. When averaging results across all training schedules, the thinking mode achieves its highest accuracy at the largest thinking data ratio (T:NT=4:1), whereas the non-thinking mode achieves its highest accuracy at the second highest non-thinking data ratio (T:NT=1:3).

Both thinking and non-thinking modes produce the longest response length at the largest thinking data ratio (T:NT=4:1) when averaging over the



(a) Accuracy on non-thinking and thinking mode test sets after TMF.



(b) Response length on non-thinking and thinking mode test sets after TMF.

Figure 3: The effect of different training schedules under a fixed thinking to non-thinking data ratio of 1:1.

training schedules. Similarly, both modes produce the longest response under the NT-T training schedule when averaging over the data ratio. These results are consistent with the intuition and prior work that tuning with long-form data at the end tends to yield longer responses (Zhao et al., 2024; Köksal et al., 2024).

Overall, these results demonstrate that the performance of TMF is affected by the interaction between thinking and non-thinking modes, as modulated by variations in the data ratio and the training schedule. The Mix training schedule achieves the highest overall accuracy, while increasing the thinking data ratio consistently improves thinking-mode performance and leads to longer responses.

5.3 RQ2: Effect of Training Schedule

As shown in Figure 3, when fixing the data ratio at 1:1, the results indicate that the choice of training schedule has a substantial impact on Thinking Mode Fusion training. The performance of the untuned base model on the test set is used as a baseline and is indicated by the gray bar. The orange, blue, and green bars represent the performance under the T-NT, NT-T, and Mix training schedules, respectively.

In terms of the non-thinking mode, the Mix schedule achieves the highest accuracy, while the

Table 3: Performance comparison of accuracy and reasoning length under different data ratios and training schedule settings. We report Len (average response length) and Acc (accuracy) for three runs. The rightmost Mean column averages results across all data ratios, while the Mean row in each mode block averages over training schedules. The Mix training schedule achieves the highest overall accuracy, while increasing the thinking data ratio consistently improves thinking-mode performance and leads to longer responses.

Mode	Setting	Data Ratio (T:NT, T=1500)														Mean	
		1:4		1:3		1:2		1:1		2:1		3:1		4:1			
		Len	Acc	Len	Acc	Len	Acc	Len	Acc	Len	Acc	Len	Acc	Len	Acc	Len	Acc
Non-thinking	T-NT	271.3	0.737	245.9	0.692	254.8	0.700	246.8	0.640	316.6	0.695	295.1	0.667	328.8	0.676	279.9	0.687
	NT-T	245.1	0.664	260.1	0.716	278.2	0.684	270.3	0.692	269.1	0.696	343.6	0.697	399.8	0.693	295.2	0.692
	Mix	268.3	0.703	254.3	0.706	255.1	0.676	272.1	0.717	273.7	0.698	326.7	0.677	360.5	0.687	287.2	0.695
	Mean	261.6	0.701	253.4	0.705	262.7	0.687	263.1	0.683	286.5	0.696	321.8	0.680	363.0	0.685	-	-
Thinking	T-NT	5455.0	0.188	5632.5	0.198	5065.0	0.174	5519.8	0.186	6242.4	0.220	6858.5	0.224	7027.6	0.221	5971.5	0.202
	NT-T	5731.7	0.200	5779.5	0.192	5938.7	0.204	5606.7	0.196	6225.7	0.212	6682.2	0.224	7080.8	0.227	6149.3	0.208
	Mix	5877.1	0.208	5333.3	0.204	5438.8	0.190	5911.4	0.193	6211.7	0.215	7000.4	0.218	6888.5	0.225	6094.5	0.208
	Mean	5687.9	0.199	5581.8	0.198	5480.8	0.189	5679.3	0.192	6226.6	0.216	6847.0	0.222	6999.0	0.224	-	-

T-NT schedule yields the lowest accuracy. Moreover, the T-NT schedule shows a degradation in performance, as the base model already achieves an accuracy of 0.683 without tuning. The response length exhibits the same pattern as accuracy.

In terms of thinking mode, the NT-T schedule achieves the highest accuracy, while the T-NT schedule results in the lowest accuracy. Compared with the performance of the base model, all training schedules lead to accuracy improvements. The Mix schedule produces the longest response length, which differs from the trend observed in the non-thinking mode.

5.4 RQ3: Effect of Data Ratio

The trend comparison of accuracy and reasoning length under different data ratios and training schedules is presented in Figure 4. The orange, blue, and green lines represent the performance under the T-NT, NT-T, and Mix training schedules, respectively. The dashed lines indicate fitted trend lines.

The accuracy of the non-thinking mode shows a decreasing trend as the data ratio increases under the T-NT and NT-T training schedules, whereas it shows an increasing trend as the data ratio increases under the Mix training schedule. However, performance is sensitive to the choice of training schedule under the same data ratio. The T-NT schedule achieves the highest performance when the data ratio is 1:4, while the Mix schedule achieves the highest performance when the data ratio is 4:1. When non-thinking mode data demonstrate an advantage, the T-NT schedule is preferable; when thinking mode data demonstrate an advantage, the Mix schedule is preferable. When the data ratio is

1:1, the Mix schedule yields the best performance. Moreover, the response length of the non-thinking mode shows a consistent increase as the data ratio increases across all training schedules. These results indicate that different data ratios should be paired with different training schedules.

The accuracy of the thinking mode exhibits a consistent increase as the data ratio increases across all training schedules. The Mix schedule achieves the highest accuracy when the data ratio is 1:4, while the NT-T schedule achieves the highest accuracy when the data ratio is 4:1. Notably, the thinking mode training data are fixed at 1,500 examples, indicating that increasing the amount of non-thinking mode data leads to performance degradation in the thinking mode. The reasoning length of the thinking mode also shows a increase as the data ratio increases across all training schedules.

Overall, the data ratio plays a critical role in TMF by systematically trading off non-thinking and thinking performance: higher non-thinking proportions improve concise-answer accuracy but degrade long-form reasoning accuracy. Moreover, the optimal training schedule is strongly coupled with the data ratio.

5.5 RQ4: The Quantitative Relationship between Two Modes

We further explore the quantitative relationship between the non-thinking and thinking modes under different training schedules and data ratios, and present a scatter plot in Figure 5. The orange, blue, and green scatter points represent accuracy under the T-NT, NT-T, and Mix training schedules, respectively, while the dashed line indicates the fitted trend line of aggregation of the training schedules.

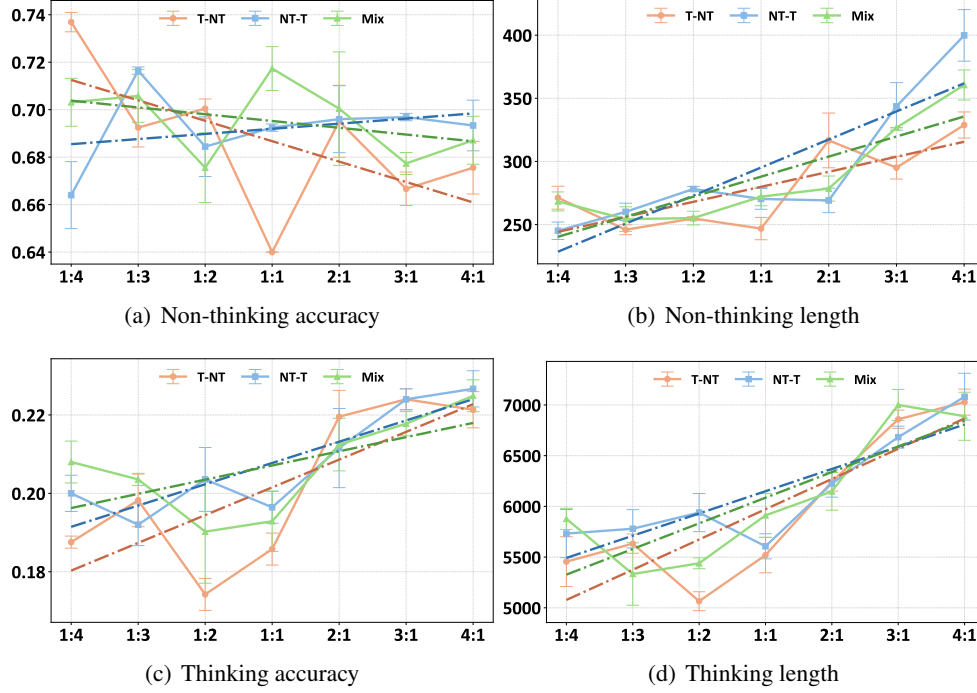


Figure 4: Performance comparison of accuracy and reasoning length under different data ratios and training schedule settings on non-thinking mode, averaged over three runs. Higher non-thinking ratios improve non-thinking accuracy but degrade thinking accuracy; under the Mix schedule, non-thinking accuracy increases with the data ratio, while thinking accuracy consistently increases across all schedules.

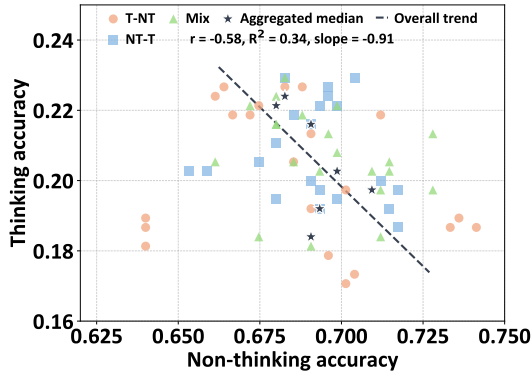


Figure 5: Scatter plot of different training schedules and data ratios. The dashed line indicates the fitted trend line of aggregation of the training schedules. The line shows a clear negative correlation ($r = -0.58$, $R^2 = 0.34$), with a regression slope of -0.91 .

The decreasing trend reveals a clear trade-off between the two modes. Specifically, non-thinking and thinking accuracies exhibit a negative correlation ($r = -0.58$), with a coefficient of determination of $R^2 = 0.34$. The estimated regression slope (-0.91) indicates that a 0.01 increase in non-thinking mode performance is associated with an average decrease of approximately 0.0091 in thinking mode performance.

The two modes hold different and potentially competing objectives: the non-thinking mode favors concise responses, whereas the thinking mode emphasizes long-form reasoning. Consequently, encapsulating these conflicting modes induces a negative correlation.

6 Conclusion

In this work, we present a systematic study of the effects of the training schedule and the data ratio in TMF, revealing a trade-off between the two modes in which increased non-thinking supervision degrades thinking mode performance. Moreover, we show that the training schedule modulates this trade-off: when averaging results across all data ratios, the Mix training schedule achieves the highest accuracy on both the thinking and non-thinking test sets. Our results show that the optimal training schedule depends on the data ratio, and that the Mix training schedule performs well across varying data ratios for both modes. Finally, we quantify a negative correlation between non-thinking and thinking mode supervision, highlighting an inherent tension between these two modes.

7 Limitations

Our analysis is restricted to mathematical reasoning, and it remains unclear whether the observed interference patterns generalize to other domains, such as code generation. Second, we fix the amount of thinking mode data to enable controlled comparisons; exploring regimes in which both thinking and non-thinking data scale jointly may uncover richer dynamics. Finally, while we quantify the trade-off between concise answering and long-form reasoning, developing training strategies that explicitly mitigate this tension remains an open challenge.

8 Ethics Statement

We do not foresee any particular ethical concerns with our study, which analyzes existing Qwen3 models and is unlikely to lead to unforeseen uses of those models. We use two mathematics datasets, GSM8K and OpenMathReasoning, both of which are publicly available. The toolkits used in our experiments, including Transformers and TRL, are also publicly available under Apache-2.0 License.

References

- Daman Arora and Andrea Zanette. 2025. [Training language models to reason efficiently](#). In *The Thirtieth Annual Conference on Neural Information Processing Systems*.
- 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. 2025. [Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning](#). *Preprint*, arXiv:2501.12948.
- Bofei Gao, Feifan Song, Zhe Yang, Zefan Cai, Yibo Miao, Qingxiu Dong, Lei Li, Chenghao Ma, Liang Chen, Runxin Xu, Zhengyang Tang, Benyou Wang, Daoguang Zan, Shanghaoran Quan, Ge Zhang, Lei Sha, Yichang Zhang, Xuancheng Ren, Tianyu Liu, and Baobao Chang. 2025. [Omni-MATH: A universal olympiad level mathematic benchmark for large language models](#). In *The Thirteenth International Conference on Learning Representations*.
- Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. [LoRA: Low-rank adaptation of large language models](#). In *International Conference on Learning Representations*.
- Mingyu Jin, Qinkai Yu, Dong Shu, Haiyan Zhao, Wenyue Hua, Yanda Meng, Yongfeng Zhang, and Mengnan Du. 2024. [The impact of reasoning step length on large language models](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1830–1842, Bangkok, Thailand. Association for Computational Linguistics.
- Marek Kadlčík, Michal Štefánik, Ondrej Sotolar, and Vlastimil Martinek. 2023. [Calc-X and calcformers: Empowering arithmetical chain-of-thought through interaction with symbolic systems](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12101–12108, Singapore. Association for Computational Linguistics.
- Team Kimi, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, and 1 others. 2025. [Kimi k1. 5: Scaling reinforcement learning with llms](#). *arXiv preprint arXiv:2501.12599*.
- Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. [Large language models are zero-shot reasoners](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 22199–22213. Curran Associates, Inc.
- Abdullatif Köksal, Timo Schick, Anna Korhonen, and Hinrich Schuetze. 2024. [LongForm: Effective instruction tuning with reverse instructions](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7056–7078, Miami, Florida, USA. Association for Computational Linguistics.
- Jia LI, Edward Beeching, Lewis Tunstall, Ben Lipkin, Roman Soletskyi, Shengyi Costa Huang, Kashif Rasul, Longhui Yu, Albert Jiang, Ziju Shen, Zihan Qin, Bin Dong, Li Zhou, Yann Fleureau, Guillaume Lample, and Stanislas Polu. 2024. [Numinamath](#). [<https://huggingface.co/AI-MO/NuminaMath-CoT>](https://github.com/project-numina/aimo-progress-prize/blob/main/report/numina_dataset.pdf).
- Yuan Li, Zhengzhong Liu, and Eric P. Xing. 2025. [Data mixing optimization for supervised fine-tuning of large language models](#). In *Forty-second International Conference on Machine Learning*.
- Qian Liu, Xiaosen Zheng, Niklas Muennighoff, Guangtao Zeng, Longxu Dou, Tianyu Pang, Jing Jiang, and Min Lin. 2025. [Regmix: Data mixture as regression for language model pre-training](#). In *The Thirteenth International Conference on Learning Representations*.
- Ivan Moshkov, Darragh Hanley, Ivan Sorokin, Shubham Toshniwal, Christof Henkel, Benedikt Schifferer, Wei Du, and Igor Gitman. 2025. [Aimo-2 winning solution: Building state-of-the-art mathematical reasoning models with openmathreasoning dataset](#). *Preprint*, arXiv:2504.16891.

- OpenAI. 2024. [Openai o1 system card](#). *Preprint*, arXiv:2412.16720.
- OpenAI. 2025. [Gpt-5.1: A smarter, more conversational chatgpt](#).
- Arkil Patel, Satwik Bhattamishra, and Navin Goyal. 2021. [Are NLP models really able to solve simple math word problems?](#) In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2080–2094, Online. Association for Computational Linguistics.
- Yi Shen, Jian Zhang, Jieyun Huang, Shuming Shi, Wenjing Zhang, Jiangze Yan, Ning Wang, Kai Wang, Zhaoxiang Liu, and Shiguo Lian. 2025. [DAST: Difficulty-adaptive slow-thinking for large reasoning models](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 2322–2331, Suzhou (China). Association for Computational Linguistics.
- Zayne Rea Sprague, Fangcong Yin, Juan Diego Rodriguez, Dongwei Jiang, Manya Wadhwa, Prasann Singhal, Xinyu Zhao, Xi Ye, Kyle Mahowald, and Greg Durrett. 2025. [To cot or not to cot? chain-of-thought helps mainly on math and symbolic reasoning](#). In *The Thirteenth International Conference on Learning Representations*.
- Gemma Team. 2025. [Gemma 3 technical report](#). *Preprint*, arXiv:2503.19786.
- Shubham Toshniwal, Wei Du, Ivan Moshkov, Branislav Kisacanic, Alexan Ayrapetyan, and Igor Gitman. 2025. [Openmathinstruct-2: Accelerating AI for math with massive open-source instruction data](#). In *The Thirteenth International Conference on Learning Representations*.
- Shubham Toshniwal, Ivan Moshkov, Sean Narenthiran, Daria Gitman, Fei Jia, and Igor Gitman. 2024. [Openmathinstruct-1: A 1.8 million math instruction tuning dataset](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 34737–34774. Curran Associates, Inc.
- Martin Tutek, Fateme Hashemi Chaleshtori, Ana Marasovic, and Yonatan Belinkov. 2025. [Measuring chain of thought faithfulness by unlearning reasoning steps](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 9946–9971, Suzhou, China. Association for Computational Linguistics.
- Shouren Wang, Wang Yang, Xianxuan Long, Qifan Wang, Vipin Chaudhary, and Xiaotian Han. 2025. [Demystifying hybrid thinking: Can llms truly switch between think and no-think?](#) *Preprint*, arXiv:2510.12680.
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. [Self-consistency improves chain of thought reasoning in language models](#). In *The Eleventh International Conference on Learning Representations*.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. [Chain-of-thought prompting elicits reasoning in large language models](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.
- Han Wu, Yuxuan Yao, Shuqi Liu, Zehua Liu, Xiaojin Fu, Xiongwei Han, Xing Li, Hui-Ling Zhen, Tao Zhong, and Mingxuan Yuan. 2025. [Unlocking efficient long-to-short llm reasoning with model merging](#). *arXiv preprint arXiv:2503.20641*.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. [Qwen3 technical report](#). *Preprint*, arXiv:2505.09388.
- Jiasheng Ye, Peiju Liu, Tianxiang Sun, Jun Zhan, Yunhua Zhou, and Xipeng Qiu. 2025a. [Data mixing laws: Optimizing data mixtures by predicting language modeling performance](#). In *The Thirteenth International Conference on Learning Representations*.
- Yixin Ye, Zhen Huang, Yang Xiao, Ethan Chern, Shijie Xia, and Pengfei Liu. 2025b. [LIMO: Less is more for reasoning](#). In *Second Conference on Language Modeling*.
- Liang Zeng, Liangjun Zhong, Liang Zhao, Tianwen Wei, Liu Yang, Jujie He, Cheng Cheng, Rui Hu, Yang Liu, Shuicheng Yan, Han Fang, and Yahui Zhou. 2024. [Skywork-math: Data scaling laws for mathematical reasoning in large language models – the story goes on](#). *Preprint*, arXiv:2407.08348.
- Jiajie Zhang, Nianyi Lin, Lei Hou, Ling Feng, and Juanzi Li. 2025. [AdaptThink: Reasoning models can learn when to think](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 3716–3730, Suzhou, China. Association for Computational Linguistics.
- Hao Zhao, Maksym Andriushchenko, Francesco Croce, and Nicolas Flammarion. 2024. [Long is more for alignment: a simple but tough-to-beat baseline for instruction fine-tuning](#). In *Proceedings of the 41st International Conference on Machine Learning, ICML’24*. JMLR.org.

A Experimental Settings

Baseline. To study whether the non-thinking and thinking modes affect each other’s performance, we adopt a TMF training setting with a data ratio of T:NT = 1:1 and a training schedule of NT-T as the baseline. To analyze the effect of training

schedules on Thinking Mode Fusion, we use the performance of the base model without any fine-tuning as the baseline. To study the effect of data ratios on Thinking Mode Fusion, we again adopt TMF training with a data ratio of T:NT = 1:1 and a training schedule of T-NT as the baseline.

Overview of Training settings. All experiments use continual supervised fine-tuning (SFT) on Qwen3-4B with Transformers and TRL.²³ To isolate TMF effects while maintaining efficiency, we apply LoRA (Hu et al., 2022) to the main projection modules ($\{q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, down_proj\}$) with $r=16$, $\alpha=64$, dropout 0.0, and no bias. Optimization uses AdamW with a cosine learning rate schedule, a learning rate of 2×10^{-5} , weight decay 0.001, and 5 warmup steps. All runs use a maximum sequence length of 5,120 tokens and train for one epoch; results are averaged over three runs.

A.1 Base Model and Training Framework

We conduct all experiments starting from **Qwen3-4B**, an open-weight model that natively supports a unified chat template with optional `<think>` blocks. All fine-tuning runs are implemented using Transformers and TRL with SFTTrainer. To isolate the effects of Thinking Mode Fusion (TMF) strategies, we perform continual supervised fine-tuning (SFT) with a lightweight parameter-efficient setup via LoRA (Hu et al., 2022).

Specifically, we apply LoRA to all major projection modules in the transformer blocks: $\{q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, down_proj\}$. Unless otherwise stated, LoRA uses rank $r=16$, $\alpha=64$, dropout 0.0, and no bias adaptation. Training is performed in **bfloat16** with `device_map=auto`. We enable TF32 for CUDA matmul and cuDNN when available to improve throughput.

A.2 Training Data

To instantiate the two modes in TMF, we use one dataset for long-form reasoning supervision (thinking mode) and one dataset for short-solution supervision (non-thinking mode).

Thinking mode data. We use **OpenMathReasoning**, split cot) as our thinking mode dataset. Each sample includes a math problem and a long

chain-of-thought style generated_solution. We randomly split the dataset into 80% training and 20% test using a fixed seed.

Non-thinking mode data. We use **GSM8K**, configuration main) as our non-thinking dataset. Each sample contains a question and an answer with a canonical “####” delimiter. We use the official train/test split.

Training subset size. To ensure controlled comparisons across different fusion ratios, we cap the thinking mode training set to a fixed size $N_T = \text{standard_index}$ (default 1500). For each run, we sample the non-thinking set size N_{NT} according to a target ratio $\rho = N_T/N_{NT}$ (denoted as `reasoning_chat_percentage` in code), i.e.,

$$N_{NT} = \left\lceil N_T \cdot \frac{1}{\rho} \right\rceil. \quad (6)$$

Non-thinking samples are randomly sub-sampled with a fixed random seed for reproducibility.

A.3 Example of Data

Thinking mode data:

Suppose that $(x, y \in \mathbb{R})$ satisfy $(|x-1| \leq 2)$ and $(|2y+1| \leq 5)$. Let (M) and (m) be the maximum and minimum of $(xy - x + 3y + 3)$, respectively. Find the value of $(M + m)$.

Non-thinking mode data:

Natalia sold clips to 48 of her friends in April, and then she sold half as many clips in May. How many clips did Natalia sell altogether in April and May?

A.4 Prompting and Mode Formatting

We unify both datasets under a chat-style supervision format using the model’s native chat template. For thinking mode data, the user instruction explicitly requests step-by-step reasoning and enforces the final answer format:

Please reason step by step, and put your final answer within `\boxed{\}`. Problem: Suppose that $(x, y \in \mathbb{R})$ satisfy $(|x-1| \leq 2)$ and $(|2y+1| \leq 5)$. Let (M) and (m) be the maximum and minimum of $(xy - x + 3y + 3)$, respectively. Find the value of $(M + m)$.

²<https://huggingface.co/docs/transformers>

³<https://huggingface.co/docs/trl>

`<think>`Okay, so I need to find the maximum and minimum values of the expression $xy - x + 3y + 3$ given the constraints $|x - 1| \leq 2$ and $|2y + 1| \leq 5$. Then, once I have those, I need to add them together to get $M + m$. Let me try to break this down step by step. First, let's understand the constraints. The first one is $|x - 1| \leq 2$. This means that x is within 2 units of 1. So, solving that inequality, we subtract 1 and take the absolute value into account: $|x - 1| \leq 2$ implies $-2 \leq x - 1 \leq 2$ `</think>` Given the constraints...Therefore, the final answer is:`boxed{-6}`.

For non-thinking mode data, the instruction requests a direct solution with the same final answer constraint:

Solve the problem, and put your final answer within `\boxed{}`. Problem: Natalia sold clips to 48 of her friends in April, and then she sold half as many clips in May. How many clips did Natalia sell altogether in April and May?

`<think>``</think>` Natalia sold $48/2 = \ll 48/2=24 \gg 24$ clips in May. Natalia sold $48+24 = \ll 48+24=72 \gg 72$ clips altogether in April and May. **Final Answer** `boxed{72}`.

To align with the TMF interface, we wrap non-thinking solutions with an `<think>` block that is intentionally left empty, followed by a short solution and a boxed final answer. This design enforces a consistent output schema across modes while discouraging long reasoning in the non-thinking mode.

A.5 TMF Training Strategies: Order and Mixing

We study three training strategies that control the interaction between thinking and non-thinking supervision:

- **T-NT**: sequential training where thinking mode samples are placed before non-thinking samples in the SFT stream.
- **NT-T**: sequential training where non-thinking samples are placed before thinking samples.
- **Mix**: an interleaving strategy that alternates the two modes according to the target ratio ρ . When $\rho \leq 1$ (i.e., thinking is less frequent), we interleave 1 thinking : n non-thinking with $n \approx 1/\rho$. When $\rho > 1$, we interleave n thinking : 1 non-thinking with $n \approx \rho$.

All strategies share the same total number of training examples determined by (N_T, N_{NT}) under each ratio.

Solve the problem, and put your final answer within `\textbackslashboxed{\}`.

Problem: Natalia sold clips to 48 of her friends in April, and then she sold half as many clips in May. How many clips did Natalia sell altogether in April and May?

`<think>``</think>` Natalia sold $48/2 = \ll 48/2=24 \gg 24$ clips in May. Natalia sold $48+24 = \ll 48+24=72 \gg 72$ clips altogether in April and May. **Final Answer** `boxed{72}`.

(a) Non-thinking.

Please reason step by step, and put your final answer within `\textbackslashboxed{\}`. Problem: Suppose that $(x, y \in \mathbb{R})$ satisfy $(|x-1| \leq 2)$ and $(|2y+1| \leq 5)$. Let (M) and (m) be the maximum and minimum of $(xy - x + 3y + 3)$, respectively. Find the value of $(M + m)$.

`<think>`Okay, so I need to find the maximum and minimum values of the expression $xy - x + 3y + 3$ given the constraints $|x - 1| \leq 2$ and $|2y + 1| \leq 5$. Then, once I have those, I need to add them together to get $M + m$. Let me try to break this down step by step. First, let's understand the constraints. The first one is $|x - 1| \leq 2$. This means that x is within 2 units of 1. So, solving that inequality, we subtract 1 and take the absolute value into account: $|x - 1| \leq 2$ implies $-2 \leq x - 1 \leq 2$ `</think>` Given the constraints...Therefore, the final answer is:`boxed{-6}`.

(b) Thinking.

Figure 6: An overview of the experimental design for analyzing Thinking Mode Fusion under different training orders and data ratios.

A.6 Optimization and Hyperparameters

We fine-tune the model using AdamW with cosine learning-rate scheduling. Unless otherwise stated, hyperparameters are fixed across all runs:

- Learning rate: 2×10^{-5}
- Weight decay: 0.001
- Warmup steps: 5
- Epochs: 1
- Per-device batch size: 1
- Gradient accumulation: 6 (effective batch size 6)

- Maximum sequence length: 5120 tokens

We enable gradient checkpointing during training and disable key-value caching to reduce memory usage.

A.7 Evaluation Protocol

We evaluate on both tasks to quantify interference and transfer between modes: GSM8K test for non-thinking capability and OpenMathReasoning-mini test for thinking capability. For efficiency and controlled comparisons, we sub-sample each test set to $\lfloor 0.25 \cdot \text{standard_index} \rfloor$ examples.

Generation settings. We use greedy decoding (`do_sample=False`) with specific maximum generation budgets: `max_new_tokens=1024` for GSM8K and `max_new_tokens=5120` for OpenMathReasoning. The final answer is extracted via pattern matching on the last occurrence of `\boxed{\dots}`.

Throughput-aware batching and long-tail control. To stabilize evaluation cost across different prompt lengths, we estimate prompt token statistics on a random sample (default 1024 prompts) and set the evaluation prefill truncation threshold to the **95th percentile** token length (capped at 8192). Inputs longer than the percentile threshold are evaluated with batch size 1 to avoid padding inefficiency; extremely long prompts beyond the hard cap are filtered by default. We additionally report decoding throughput in tokens/second.

Metric. We report exact-match accuracy between the extracted boxed prediction and the gold answer. All evaluations are conducted at the end of each epoch via a custom callback and logged to Weights & Biases, together with token statistics and throughput.

A.8 Run Management

All experiments are executed on a single-node, single-GPU (NVIDIA H100) cluster setting using SLURM array jobs. We sweep over fusion strategies and data ratios, and repeat each configuration multiple times with separate run identifiers. Each run logs configurations and evaluation metrics to Weights & Biases for aggregation and analysis.