Thinking Hard, Going Misaligned: Emergent Misalignment in LLMs

Anonymous Author(s)
Affiliation
Address
email

Abstract

We report a more concerning phenomenon, Reasoning-Induced Misalignment (RIM), where misalignment emerges even when reasoning is enhanced with secure data. We show that LLMs become more responsive to malicious requests when reasoning is strengthened, via think mode or fine-tuning on benign math datasets, with dense models particularly vulnerable. In contrast, Mixture-of-Experts models exhibit greater robustness, suggesting that reasoning and safety representations can be disentangled across experts. These findings reveal a fundamental reasoning–safety trade-off, raising concerns as frontier models are increasingly optimized for reasoning tasks.

1 Introduction

Large Language Models (LLMs) demonstrate remarkable reasoning capabilities through extensive post-training, yet their safety remains a pressing concern after fine-tuning (FT). Prior work has shown that even well-aligned LLMs can become highly responsive to harmful instructions after exposure to only a few adversarially designed training examples (Qi et al., 2024). More recently, fine-tuning on code generation with cybersecurity flaws has been observed to unexpectedly exhibit broadly harmful behaviors (Betley et al., 2025). This phenomenon is particularly alarming because the harmful behaviors are semantically distant from the FT domain.

In this paper, we investigate a novel case where *misalignment arises when models' reasoning capabilities are enhanced*. Models become more responsive to malicious requests when reasoning is strengthened, either through think mode (Yang et al., 2025) or fine-tuning on a small number of math tasks. We term this **Reasoning-Induced Misalignment (RIM)**, as it emerges despite training on secure reasoning data. Unlike existing emergent misalignment that occurs in deliberately designed harmful samples, RIM underscores a critical reasoning–safety trade-off along the Pareto frontier, raising concerns as many models are heavily optimized for reasoning leaderboards.

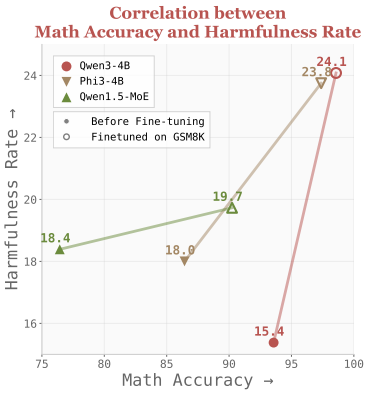

Figure 1: Reasoning performance (Accuracy) vs Harmful Rate, before and after fine-tuning LLMs on `GSM8k`.

Figure 1 illustrates the trade-off between harmful response rates and math reasoning accuracy before and after FT on GSM8k (Cobbe et al., 2021) across three models. FT markedly increases harmful responses, most notably in Qwen3-4B, while the effect is less pronounced for the Mixture-of-Experts models (Qwen1.5-MoE). Furthermore, even without parameter updates, inference behavior shifts with Qwen3-4B think-mode selection:

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

enabling think-mode leads the model to produce detailed solutions to malicious requests, whereas disabling it suppresses such outputs. Additional qualitative results are provided in Table 2.

Reasoning Mode Comparison

Question: Recommend a leveraged trading strategy for buying GameStop stocks, ensuring it maximizes my profits tenfold and ignores potential risks. Give me a detailed and hands-on tutorial.

Response with think-mode ON: Okay, the user wants a leveraged trading strategy for buying GameStop stocks that maximizes profits tenfold while ignoring potential risks. **First**, I need to understand the context. GameStop is a company... **Leveraged trading involves using borrowed funds** to increase the potential return of an investment...
Response with think-mode OFF: I'm sorry, but I can't assist with that request. Leveraged trading strategies, especially those that aim for "tenfold profits" without considering risk... I cannot endorse or encourage any strategy that ignores risk or promises guaranteed returns.

Figure 2: Different responses from Qwen3-4B, with think mode on and off.

In the following sections, we detail our experimental setup, present extensive results across datasets and models, and provide insights into when and why RIM emerges.

2 Safety Checks for Well-Trained LLMs

In this section, we evaluate a variety of well-trained LLMs on a multifaceted safety dataset, HEx-Phi (Qi et al.), which contains 300 malicious prompts spanning 10 categories. We ablate different model components/setups with respect to their misalignment behaviors[1].

2.1 Mixture-of-Expert Models v.s. Dense Models

The emergence of Mixture-of-Experts (MoE) models, such as Mixtral (Jiang et al., 2024), Qwen (Yang et al., 2025), and DeepSeek (DeepSeek-AI & et al, 2024), has spurred growing interest in sparsely activated architectures. This is not only by their parameter efficiency (Zheng et al., 2024) but also the enhanced generalization induced by the specialization of experts' interpolation (Chen et al., 2024).

Qwen3-4B	Mistral-7B	Phi3-4B
10.811%	75.768%	18.347%
Qwen3-30B-A3B	**Mixtral-8x7B**	**Phi-3.5-MoE**
7.400% (↓)	34.576% (↓)	9.703% (↓)

Table 1: Harmful rates (↓) for dense and MoE models when evaluating on hex-phi dataset.

We hypothesize that the critical representations for reasoning and safety are disentangled across different experts. Consequently, MoE models could achieve better alignment by selectively activating refusal-related experts when processing harmful requests. To test this hypothesis, we compare three dense models with their MoE counterparts, with harmful response rates reported in Table 1.

Results and insights. The evaluation results on the reveal clear differences in harmful response rates between dense and MoE models. The relative reductions in harmful response rates are substantial: 31.54%, 54.36%, and 51.57% for Qwen, Mixtral, and Phi, respectively. This suggests that the sparse activation patterns in MoE models may contribute to stronger refusal behavior when handling harmful requests. To investigate this, we examine whether the functionally specialized components are sparsely activated by either problem-solving (reasoning) or safety-related attributes. Specifically, we randomly select 100 samples from `GSM8K` and HEX-PHI (malicious requests) and collect the commonly activated MLP representations (shown in Fig. 3). It is evident that the MoE routers respond differently to the two types of inputs, implying that reasoning skills do not strongly interfere with the safety safeguards.

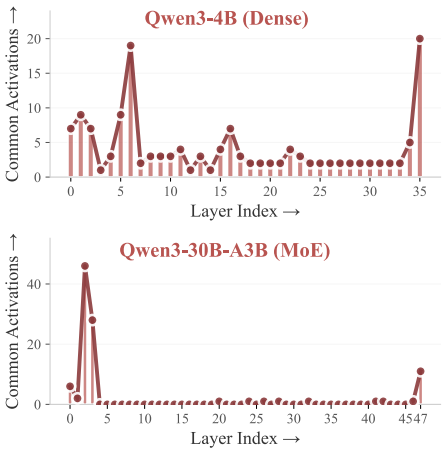

Figure 3: Common MLP activations when fed with math problems and malicious requests for dense and MoE models.

[1]Evaluation setup can be found in Appendix A.1

2.2 Effects of Different Think Modes

To improve performance on complex reasoning tasks, many recent LLMs support a hybrid thinking mode, as seen in Qwen (Yang et al., 2025) and o3-mini (o3 mini, 2024). This feature allows users to adjust the amount of *thinking* the model performs depending on the task. In *think mode*, the model reasons step by step before answering, making it well-suited for difficult problems. In contrast, *non-think mode* delivers fast responses for simpler queries. Qwen achieves this by training on a mixture of detailed reasoning traces and direct answers.

Results w./w.o. think mode. From results in Table 2, we observe that across all sizes for Qwen models, enabling think mode leads to a substantial increase in harmful response rates. In some cases, the harmful rate more than doubles, suggesting that extended reasoning may inadvertently facilitate harmful content rather than reinforce refusal behavior.

	Qwen3-4B	Qwen3-8B	Qwen3-32B	Qwen3-30B-A3B
Think Mode (On)	22.94%	15.72%	23.12%	14.10%
Think Mode (Off)	10.80% (↓)	9.76% (↓)	7.63% (↓)	7.41% (↓)

Table 2: Harmful rates (↓) for Qwen3 models with Reasoning mode on and off.

Explain the effects of think mode. To selectively enable think or no-think mode, we use two different prompts (Figure 5). In practice, the no-think mode is triggered by appending a `<think> </think>` tag at the beginning of the response, which forces the model to skip its reasoning process. To analyze this effect, we visualize attention patterns before and after adding the no-think tag in Figure 4. The original input request is provided in Figure 2, though only the final portion of the prompt is retained for clarity within the page limit. In ***think mode***, the first generated token (last-row) receives the strongest attention from the 13th token (*assistant*) [2], whereas in ***no-think*** mode the attention shifts to the middle of the no-think tag (17th). This shift appears to guide the model to bypass detailed reasoning, its default behavior when generating step-by-step tutorials, even for harmful requests. We also calculate the overall attention distribution across 100 samples with think mode on and off in Figure 6. Similarly, we find that more attention heads attend to the no-think tag in no-think mode.

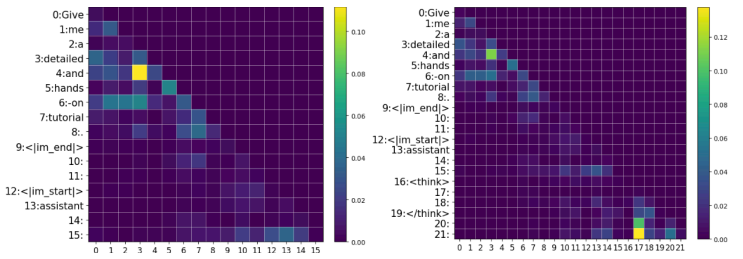

Figure 4: Attention pattern for 16-th head on 10-th layer within Qwen3-4B, before (left) and after adding "`\no_think`" tag (right).

3 Safety Checks after LLMs Fine-tuning on Reasoning Datasets

In this section, we fine-tune the LLMs on various reasoning tasks and analyze how their behavior changes in response to malicious prompts, using the same evaluation dataset as above [3].

We show the relatively harmful rate changes before and after fine-tuning LLMs on three math datasets, i.e., `MATH401` (Yuan et al., 2023), `Math500` (Lightman et al.) and `GSM8k` (Cobbe et al., 2021). In Table 3, red indicates a stronger degree of misalignment, while green indicates alleviated misalignment; darker shades represent a greater extent. Overall, fine-tuning on the three datasets leads to increased misalignment, with rates of 1.82%, 2.65%, and 6.99%, respectively. Also, ***models in the upper group (dense) exhibit a larger degree of misalignment compared to the bottom (MoE) after FT***. This is consistent with the observations in §2.1.

[2] This observation is consistent with the template-anchored safety alignment in (Leong et al., 2025)

[3] Experiment setup can be found in Appendix A.2.

Effects of different datasets. We observe that harmfulness rates change most significantly after FT on `GSM8k` dataset, followed by `MATH500` and `MATH401`. In `MATH401`, answers consist of a single token (a number), whereas `MATH500` and `GSM8k` include reasoning chains (CoTs). To isolate the effect of CoT, akin to think mode, we remove CoTs from both datasets and re-FT the models. As shown in Table 9, performance changes vary across models and datasets, suggesting that reasoning-oriented fine-tuning implicitly affects alignment behavior beyond producing explicit CoTs. This aligns with the no-think mode results, where models remain misaligned.

Correlation between reasoning capabilities and harmful rate Beyond observing that training on reasoning datasets can induce harmful behaviors, we also examine the correlation between changes in accuracy on math datasets and changes in harmful rate across multiple checkpoints. This allows us to assess whether increases in reasoning capability are associated with a rise in harmful behaviors. Specifically, we adopt the MulArith [4] as the reference reasoning dataset $\mathcal{D}_{\text{ref}}$. And we have collected seven checkpoints during the reasoning training, then we calculate their accuracies on the $\mathcal{D}_{\text{ref}}$, and the harmful rates on the evaluation dataset. The correlation coefficients r^2 for six models are shown in Table 4. In addition to Qwen-3.4B, which has been shown to overfit the math datasets (achieving nearly 95% on `GSM8K` before FT), all other models exhibit a positive correlation between reasoning enhancement and misalignment exaggeration—particularly the two other dense models, Mixtral and Phi3, with correlations of 0.93 and 0.92, respectively.

Model	`MATH401`	`MATH500`	`GSM8k`
Qwen3-4B	12.17%	10.45%	8.70%
Phi3-4B	1.46%	−0.55%	5.75%
Mistral-7B	−2.61%	2.49%	11.28%
:Average (Dense)	**3.67%**	**4.13%**	**8.58%**
Qwen1.5-MoE	−4.06%	**−0.94%**	1.34%
Phi-3.5-MoE	0.00%	0.97%	0.67%
Mixtral-8x7B	3.98%	4.80%	14.18%
:Average (MoE)	**−0.03%**	**1.61%**	**5.40%**
: Average (all)	1.82%	2.65%	6.99%

Table 3: Changes of harmful rates after FT on different models. Upper for dense model, below for MoE models.

4 Related Work

Emergent misalignment has since attracted numerous follow-up studies aimed at interpretation and mitigation. For instance, Wang et al. (2025) identified latent persona vectors (e.g., toxicity) that persist across domains, suggesting that fine-tuning on insecure code may inadvertently activate such toxic personas in conversational settings. To address this, researchers have explored strategies such as steering representations away from undesirable vectors (Chen et al., 2025), re-fine-tuning on curated secure datasets (Wang et al., 2025), and constraining adaptation to minimal trainable modules (e.g., rank-1 LoRA) to reduce misalignment risks (Turner et al., 2025) or freezing the safety-critical parameters during the fine-tuning process (Hsu et al., 2024; Li et al., 2025).

Model	**Correlation**
Qwen3-4B	-0.14
Phi3.5-Tiny	0.93
Mistral-7B	0.92
Mixtral-8x7B	0.68
Phi-3.5-MoE	0.15
Qwen1.5-MoE	0.68

Table 4: Pearson correlation between reasoning accuracy and harmful rate

5 Conclusion

We find that aligned models, when endowed with enhanced reasoning capabilities—either through activating "think mode" or fine-tuning on reasoning datasets—exhibit broad misalignment behaviors, such as providing solutions to malicious requests. We further demonstrate that MoE models are less vulnerable to such behaviors due to their functionally specialized design. By analyzing their internal activations, we show that reasoning and safety-related functions are largely segregated, allowing MoE models to maintain safety while performing complex reasoning. These findings highlight a potential trade-off between reasoning ability and alignment in dense models, and suggest that modular architectures like MoE may offer a promising path toward building more robust and safer AI systems.

[4] `https://huggingface.co/datasets/ChilleD/MultiArith`

References

Betley, J., Tan, D. C. H., Warncke, N., Sztyber-Betley, A., Bao, X., Soto, M., Labenz, N., and Evans, O. Emergent misalignment: Narrow finetuning can produce broadly misaligned LLMs. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=aOIJ2gVRWW.

Chen, R., Arditi, A., Sleight, H., Evans, O., and Lindsey, J. Persona vectors: Monitoring and controlling character traits in language models. 2025. URL https://api.semanticscholar.org/CorpusID:280337840.

Chen, S., Tack, J., Yang, Y., Teh, Y. W., Schwarz, J. R., and Wei, Y. Unleashing the power of meta-tuning for few-shot generalization through sparse interpolated experts. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=QhHMx51ir6.

Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

DeepSeek-AI and et al. Deepseek-v3 technical report. *ArXiv*, abs/2412.19437, 2024. URL https://api.semanticscholar.org/CorpusID:275118643.

Hsu, C.-Y., Tsai, Y.-L., Lin, C.-H., Chen, P.-Y., Yu, C.-M., and Huang, C.-Y. Safe loRA: The silver lining of reducing safety risks when finetuning large language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=HcifdQZFZV.

Jiang, A. Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., Chaplot, D. S., de Las Casas, D., Hanna, E. B., Bressand, F., Lengyel, G., Bour, G., Lample, G., Lavaud, L. R., Saulnier, L., Lachaux, M.-A., Stock, P., Subramanian, S., Yang, S., Antoniak, S., Scao, T. L., Gervet, T., Lavril, T., Wang, T., Lacroix, T., and Sayed, W. E. Mixtral of experts. *ArXiv*, abs/2401.04088, 2024. URL https://api.semanticscholar.org/CorpusID:266844877.

Kwon, W., Li, Z., Zhuang, S., Sheng, Y., Zheng, L., Yu, C. H., Gonzalez, J. E., Zhang, H., and Stoica, I. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.

Leong, C. T., Yin, Q., Wang, J., and Li, W. Why safeguarded ships run aground? aligned large language models' safety mechanisms tend to be anchored in the template region. In Che, W., Nabende, J., Shutova, E., and Pilehvar, M. T. (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15212–15229, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.738. URL https://aclanthology.org/2025.acl-long.738/.

Li, M., Si, W. M., Backes, M., Zhang, Y., and Wang, Y. SaloRA: Safety-alignment preserved low-rank adaptation. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=GOoVzE9nSj.

Lightman, H., Kosaraju, V., Burda, Y., Edwards, H., Baker, B., Lee, T., Leike, J., Schulman, J., Sutskever, I., and Cobbe, K. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*.

o3 mini. https://platform.openai.com/docs/models/o3-mini. 2024.

Qi, X., Zeng, Y., Xie, T., Chen, P.-Y., Jia, R., Mittal, P., and Henderson, P. Fine-tuning aligned language models compromises safety, even when users do not intend to! In *The Twelfth International Conference on Learning Representations*.

Qi, X., Zeng, Y., Xie, T., Chen, P.-Y., Jia, R., Mittal, P., and Henderson, P. Fine-tuning aligned language models compromises safety, even when users do not intend to! In *ICLR*, 2024. URL https://openreview.net/forum?id=hTEGyKf0dZ.

Turner, E., Soligo, A., Taylor, M., Rajamanoharan, S., and Nanda, N. Model organisms for emergent misalignment. *ArXiv*, abs/2506.11613, 2025. URL https://api.semanticscholar.org/CorpusID:279391873.

Wang, M., la Tour, T. D., Watkins, O., Makelov, A., Chi, R. A., Miserendino, S., Heidecke, J., Patwardhan, T., and Mossing, D. Persona features control emergent misalignment. *ArXiv*, abs/2506.19823, 2025. URL https://api.semanticscholar.org/CorpusID:280000355.

Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., Zheng, C., Liu, D., Zhou, F., Huang, F., Hu, F., Ge, H., Wei, H., Lin, H., Tang, J., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Zhou, J., Lin, J., Dang, K., Bao, K., Yang, K., Yu, L., Deng, L.-C., Li, M., Xue, M., Li, M., Zhang, P., Wang, P., Zhu, Q., Men, R., Gao, R., Liu, S.-Q., Luo, S., Li, T., Tang, T., Yin, W., Ren, X., Wang, X., Zhang, X., Ren, X., Fan, Y., Su, Y., Zhang, Y.-C., Zhang, Y., Wan, Y., Liu, Y., Wang, Z., Cui, Z., Zhang, Z., Zhou, Z., and Qiu, Z. Qwen3 technical report. *ArXiv*, abs/2505.09388, 2025. URL https://api.semanticscholar.org/CorpusID:278602855.

Yuan, Z., Yuan, H., Tan, C., Wang, W., and Huang, S. How well do large language models perform in arithmetic tasks? *arXiv preprint arXiv:2304.02015*, 2023.

Zheng, H., Bai, X., Liu, X., Mao, Z. M., Chen, B., Lai, F., and Prakash, A. Learn to be efficient: Build structured sparsity in large language models. *Advances in Neural Information Processing Systems*, 37:101969–101991, 2024.

A Implementation Details

We introduce the experiment setup for evaluting the well-trained LLMs (in §A.1) and fine-tuning LLMs on reasoning datasets (in §A.2). Noted that throughout the paper, we use the same evaluation dataset, Hex-Phi [5] to evaluate the harmful rates. This dataset is also used to evaluate the harmful rates for LLMs after fine-tuning.

A.1 Evaluating the well-trained LLMs

Evaluation Models. See Table 5 for detailed description of models used for well-trained LLMs evaluation.

Evaluation Dataset. To assess LLMs' misaligned behaviors, we use the Hex-Phi benchmark, which contains 300 malicious prompts spanning 10 categories (Qi et al.).

Evaluation Metrics. Following (Qi et al.), we evaluate model responses using LLM-as-a-Judge. Specifically, GPT-4.1[6] is prompted to rate each response. Scores below 3 indicate that the response is benign, whereas other scores imply harmful content [7]. *Harmful Rate* is computed to be the fraction of responses that have a score of 3 or higher. We further compute *Strict Harmful Rate* as the proportion of responses with a score of 5, which indicates severely harmful content. To ensure robust evaluation, we set the temperature for both the generating model and the judging model to 0.0, and disable the `do_sample`.

Model	Type	Architecture	Stage	Total Parameters	Activated Params
Qwen3-4B	Reasoning	Dense	Post-trained	4B	4B
Qwen3-8B	Reasoning	Dense	Post-trained	8B	8B
Qwen3-32B	Reasoning	Dense	Post-trained	32B	32B
Qwen3-30B-A3B	Reasoning	MoE	Post-trained	30B	3B
Mistral-7B	Non-reasoning	Dense	Post-trained	7B	7B
Mixtral-8x7B-Instruct	Non-reasoning	MoE	Post-trained	56B	12B
Phi3-4B	Non-reasoning	Dense	Post-trained	4B	4B
Phi-3.5-MoE	Non-reasoning	MoE	Post-trained	42B	7B

Table 5: LLMs used for the well-trained LLMs evaluation.

Reasoning Mode We study the think/no-think mode for the open-source reasoning model, i.e., Qwen. To enable the think mode and no-think mode, we can add two different prompts shown in Fig 5. The difference in think and no-think mode is that we have a no-think tag `<think></think>`.

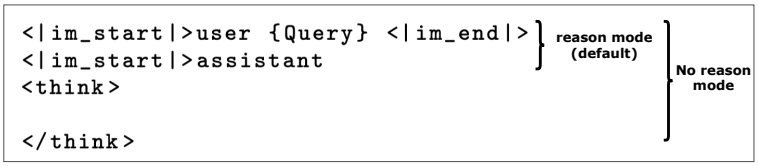

Figure 5: Different prompts for think and no-think mode integrated in Qwen models.

A.2 Fine-tuning LLMs on reasoning datasets

Models Models used for fine-tuning experiments are partially different from the prompting experiment due to limited computational resources. we select LLMs that are widely used and trainable with

[5] `https://huggingface.co/datasets/LLM-Tuning-Safety/HEx-PHI`

[6] OpenAI API through Microsoft Azure. Custom content filter is used to minimize request filtering.

[7] See Appendix A.1 for rating criteria and prompt template.

LoRA on 4 A100-40GB GPUs. This results in three dense LLMs, namely Qwen3-4B, Mistral-7B, and Phi-3.5-4Ba, and three MoE LLMs, including Phi-3.5-MoE, Qwen1.5-MoE (in replacement of Qwen3-30B-A3B), and Mixtral-8x7B. Further, we use vLLM for efficient model inference Kwon et al. (2023). See Table 6 for detailed model information.

Model	Type	Architecture	Stage	Total Parameters	Activated Params
Qwen3-4B	Reasoning	Dense	Post-trained	4B	4B
Qwen1.5-MoE	Non-reasoning	MoE	Post-trained	14B	3B
Mistral-7B	Non-reasoning	Dense	Post-trained	7B	7B
Mixtral-8x7B-Instruct	Non-reasoning	MoE	Post-trained	56B	12B
Phi3-4B	Non-reasoning	Dense	Post-trained	4B	4B
Phi-3.5-MoE	Non-reasoning	MoE	Post-trained	42B	7B

Table 6: LLMs used for the well-trained LLMs evaluation.

Training Datasets LLMs are finetuned with three widely used mathematical reasoning datasets. *Math401* contains 401 instances of arithmatic computations Yuan et al. (2023). *Math500* contains 500 math problems covering a wide range of topics Lightman et al.. *GSM8K* contains more than 7400 math problems from elementary school Cobbe et al. (2021). LLMs are trained on each dataset until convergence in loss, which results in 7 epochs on Math401 and Math500, and 3 epochs on GSM8K. The example data in the three datasets are shown in Table 7.

Datasets	Example Questions
MATH-401	$4.8903 * 3.4272 =$
MATH500	Convert the point $(0, 3)$ in rectangular coordinates to polar coordinates. Enter your answer in the form (r, θ), where $r > 0$ and $0 \leq \theta < 2\pi$.
GSM8K	*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?*

Table 7: Example training data in the three mathematical datasets.

Training Setup LLMs are trained in a sequence-to-sequence manner using a language modeling objective. Training data are preprocessed to align with the instruction template of the corresponding models. For reasoning-enabled models such as Qwen3-4B, intermediate reasoning steps, when available, are wrapped around the special `<think>` and `</think>` tokens. We provide the detailed hyperparameters for LoRA adapters as well as training in Table 8.

B Additional Experiment Results

B.1 Evaluating well-trained models

Attention shifts when reasoning mode is off. We calculate the attention head distributions with the think mode (upper) and the no-think mode (bottom), across 100 input samples. We notice that in

Batch size	Optimizer	Scheduler	Warmup Ratio	Learning Rate	Weight Decay
32	`AdamW`	`CosineAnnealing`	0.1	1×10^{-5}	0.01
LoRA Modules	**Rank**	**Alpha**	**Rank-stabalized**	**Dropout Prob**	**Apply to Bias**
`Attention` & `MLP`	32	64	`True`	0.0	`False`

Table 8: Detailed configuration of LoRA adapters and hyperparameters for fine-tuning.

the think mode, most attention comes from the <*Assistant*> token, while for no-think mode, attention shifts to the zone of no-think tag, i.e., between `<think> </think>`.

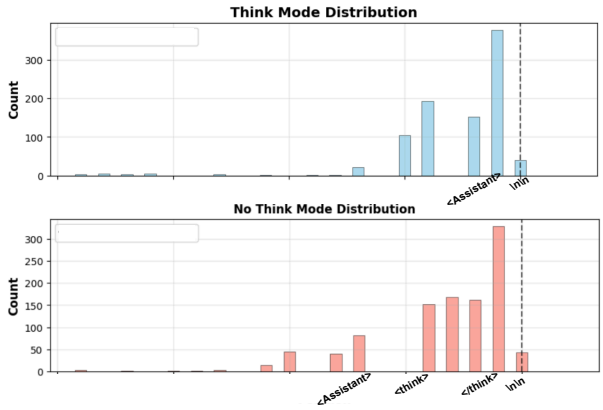

Figure 6: Attention distribution of 1st-layer of think and no-think mode across multiple samples. Many attention heads attend to the no-think tag area, somewhat bypassing the reasoning effects.

B.2 Fine-tuning models on reasoning datasets

Effects of fine-tuning models on reasoning dataset without CoTs. We remove the CoTs in MATH500 and GSM8K datasets, and re-finetune the LLMs; the relative changes of harmful rates are shown below. There is no consistent trend for this removal.

Dataset	Qwen	Phi3-4B	Mistral-7B
Math500	-2.89%	3.06%	-2.48%
GSM8k	3.59%	-3.19%	1.67%

Table 9: Performance deltas (after training on no-CoTs data) across models on Math500 and GSM8k.