Trade-offs in Large Reasoning Models: An Empirical Analysis of Deliberative and Adaptive Reasoning over Foundational Capabilities

Weixiang Zhao[1*], Xingyu Sui[1*], Jiahe Guo[1*], Yulin Hu[1*], Yang Deng[2], Yanyan Zhao[1†]
Xuda Zhi[3], Yongbo Huang[3], Hao He[3], Wanxiang Che[1], Ting Liu[1], Bing Qin[1]
[1]Harbin Institute of Technology
[2]Singapore Management University
[3]SERES
{wxzhao,xysui,jhguo,ylhu,yyzhao}@ir.hit.edu.cn

Abstract

Recent advancements in Large Reasoning Models (LRMs), such as OpenAI's o1/o3 and DeepSeek-R1, have demonstrated remarkable performance in specialized reasoning tasks through human-like deliberative thinking and long chain-of-thought reasoning. However, our systematic evaluation across various model families (DeepSeek, Qwen, and LLaMA) and scales (7B to 32B) reveals that acquiring these deliberative reasoning capabilities significantly reduces the foundational capabilities of LRMs, including notable declines in helpfulness and harmlessness, alongside substantially increased inference costs. Importantly, we demonstrate that adaptive reasoning—employing modes like Zero-Thinking, Less-Thinking, and Summary-Thinking—can effectively alleviate these drawbacks. Our empirical insights underline the critical need for developing more versatile LRMs capable of dynamically allocating inference-time compute according to specific task characteristics.

Code — https://github.com/SCIR-SC-Qiaoban-Team/FreeEvalLM

1 Introduction

Recent advancements in large language models (LLMs), particularly OpenAI's o1/o3 (Jaech et al. 2024; OpenAI 2025) and the DeepSeek-R1 (Guo et al. 2025) series, have signaled a significant shift toward large reasoning models (LRMs). Unlike traditional LLMs (Brown et al. 2020; Dubey et al. 2024; Team et al. 2024; Yang et al. 2024), LRMs demonstrate exceptional capabilities in handling complex reasoning tasks by adopting human-like deliberative thinking processes. A key distinguishing feature of LRMs is their ability to engage in extensive chain-of-thought (CoT) reasoning, systematically generating long reasoning traces composed of multiple intermediate steps before providing answers to given queries (Li et al. 2025d; Xu et al. 2025; Chen et al. 2025b).

Despite the growing interest in LRMs, the community's deep understanding of these models remains at an early stage. Recent research primarily focuses on assessing LRMs' performance in reasoning tasks, particularly analyzing their effectiveness (Wang et al. 2025; Li et al. 2025c; Ballon, Algaba, and Ginis 2025; Golde et al. 2025), efficiency (Chen et al. 2024; Luo et al. 2025; Aggarwal and Welleck 2025), and robustness (Huang et al. 2025a; Rajeev et al. 2025) within these specific contexts.

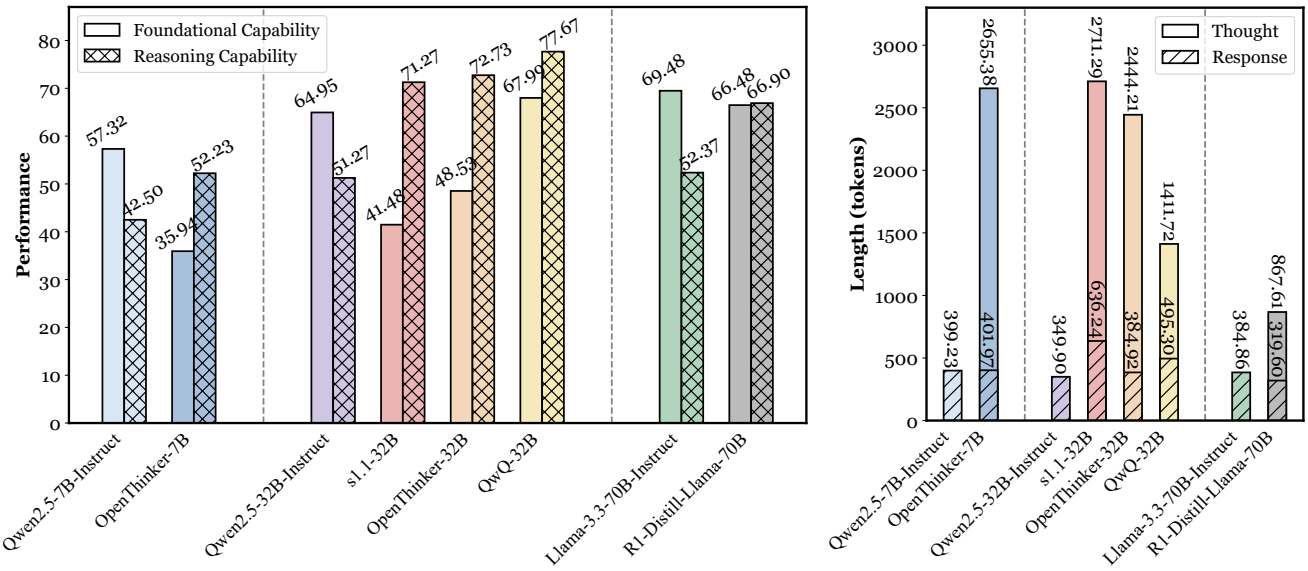
However, broader implications regarding how these deliberative reasoning capabilities influence overall model performance beyond specialized reasoning tasks remain relatively unexplored. Exploring this connection is vital, as research in cognitive science indicates a strong interplay between human reasoning capabilities and overall cognitive functions (Stanovich, West, and Hertwig 2000; Kahneman 2011). Analogous to how human intelligence seamlessly integrates rapid intuitive responses with reflective deliberation, reasoning in frontier LRMs may similarly need to be an integrated feature rather than an isolated capability (Anthropic 2025; Kimi et al. 2025). Thus, understanding *whether* and *how* deliberative reasoning impacts foundational capabilities of existing LRMs could inform and guide future advancements in model design and development.

To systematically investigate these questions, we conduct comprehensive evaluations across three prominent LRM families—DeepSeek, Qwen, and LLaMA—spanning multiple scales including 7B, 32B and 70B models. Specifically, we assess how acquiring deliberative reasoning capabilities through model distillation (Guo et al. 2025; Muennighoff et al. 2025; Ye et al. 2025; Li et al. 2025a) or large-scale reinforcement learning (Guo et al. 2025; Kimi et al. 2025; Qwen 2025) affects the models' foundational capabilities. Current evaluations define the foundational capabilities of models in terms of their helpfulness and harmlessness (Ouyang et al. 2022), covering aspects such as general task performance, instruction-following, and safety measures (Yang et al. 2024; Dubey et al. 2024). Our extensive analysis and evaluation have led us to two key insights.

> Acquiring deliberative reasoning capabilities significantly reduces the foundational capabilities of LRMs, along with substantially increased inference costs.

Specifically, as previewed in Figure 1, LRMs obtained

[*] Equal contribution
[†] Corresponding author



(a) Performance of the reasoning capability and the foundational capability across different models.

(b) Token counts for intermediate thoughts and responses.

Figure 1: Comparison of efficacy and efficiency of different LRMs and their chat-versions LLMs.

through model distillation based on chat-version checkpoints to gain deliberative reasoning abilities exhibit marked declines in performance in terms of both helpfulness and harmlessness compared to their original chat versions. For instance, the model s1.1-32B (Muennighoff et al. 2025) shows a 47.38% decrease in instruction-following capability on IFEval (Zhou et al. 2023) compared to its base chat model, Qwen2.5-32B-Instruct (Yang et al. 2024), while incurring a 250% increase in inference costs.

> Adaptive reasoning significantly enhances LRMs performance across diverse tasks.

By inserting special tokens such as "<think></think>" at different positions within the LRMs' thinking process (Muennighoff et al. 2025; Jiang et al. 2025), we can manually control the inference-time compute, thereby implementing various reasoning modes. These modes include Zero-Thinking (no deliberate reasoning), Less-Thinking (brief reasoning), and Summary-Thinking (concise summarization of reasoning). Our findings indicate that different reasoning modes optimally serve distinct general tasks, highlighting the importance of adaptive inference-time compute allocation in LRMs. For example, Summary-Thinking can notably enhance instruction-following abilities, improving performance of s1.1-32B and QwQ-32B by 59.74% and 4.44% respectively on the IFEval. Similarly, Zero-Thinking significantly boosts safety performance across all evaluated LRMs, effectively reducing harmful or unintended outputs.

In summary, our work reveals that acquiring deliberative reasoning capabilities, while essential for specialized reasoning tasks, significantly diminishes the foundational capabilities of LRMs and dramatically increases inference costs. Our findings highlight a critical gap in current LRMs regarding balanced performance across diverse tasks and offer valuable empirical insights to guide future development of more versatile, adaptive LRMs capable of dynamically allocating inference-time resources based on specific task requirements.

2 Related Works

Large Reasoning Models We define the sequence of tokens representing an instruction as x. Similarly, the token sequence corresponding to a response generated by an autoregressive model is denoted as y. For large reasoning models (LRMs), the response y consists of two components: the reasoning trace y_{CoT} and the final answer y_{ans}, such that $y = y_{\text{CoT}} \oplus y_{\text{ans}}$, where $\oplus$ indicates concatenation. The reasoning trace $y_{\text{CoT}} \subset y$ serves as the chain of thought (CoT), enabling the model to explore alternative solution paths before arriving at the final answer.

A key characteristic of LRMs is their capacity to produce explicit and extensive intermediate reasoning traces y_{CoT} (Tie et al. 2025; Kumar et al. 2025; Li et al. 2025d; Xu et al. 2025; Chen et al. 2025b; Bandyopadhyay, Bhattacharjee, and Ekbal 2025). This ability facilitates the breakdown of complex problems into clear and interpretable reasoning chains, thereby improving structured decision-making. Recent LRMs have primarily been developed through two prominent approaches: large-scale reinforcement learning (RL) and model distillation. Models trained via large-scale RL (Jaech et al. 2024; DeepMind 2025; Guo et al. 2025; Kimi et al. 2025; Qwen 2025; OpenAI 2025) leverage extensive computational resources and reward-driven optimization strategies to progressively acquire sophisticated deliberative reasoning capabilities. Conversely, distillation-based LRMs (Muennighoff et al. 2025; Ye et al. 2025; Li et al. 2025c,a; Team 2025) inherit reasoning abilities by systematically transferring structured reasoning patterns from larger teacher models into smaller

Method	Large Reasoning Model	Fine-tuned Model
Distillation	OpenThinker-7B	Qwen2.5-7B-Instruct
	OpenThinker-32B	Qwen2.5-32B-Instruct
	s1.1-32B	Qwen2.5-32B-Instruct
	R1-Distill-Llama-70B	Llama-3.3-70B-Instruct
Large-Scale RL	QwQ-32B	-
	DeepSeek-R1	DeepSeek-V3-Base

Table 1: This table summarizes the LRMs of different model families and scales evaluated for foundational capabilities and the fine-tuned source model.

models. Despite their methodological differences, both RL-trained and distilled LRMs exhibit notable human-like deliberative reasoning, significantly enhancing their proficiency in solving complex reasoning tasks.

Analysis on LRMs. Current analyses of LRMs primarily concentrate on their performance within specialized reasoning tasks, focusing on their effectiveness (Wang et al. 2025; Li et al. 2025c; Ballon, Algaba, and Ginis 2025; Golde et al. 2025), efficiency (Chen et al. 2024; Luo et al. 2025; Aggarwal and Welleck 2025), and robustness (Huang et al. 2025a; Rajeev et al. 2025; Camposampiero et al. 2025). For example, Chen et al. (2024) revealed a prominent "over-thinking" phenomenon exhibited by LRMs when tackling mathematical reasoning tasks, resulting in unnecessary complexity. Additionally, several other studies have explored the potential of LRMs in specific contexts such as role-playing tasks (Feng, Dou, and Kong 2025), agent-based tasks (Zhou et al. 2025b), multilingual scenarios (Chen et al. 2025a; Zhang et al. 2025), and safety-related performance (Arrieta et al. 2025; Jiang et al. 2025; Zhou et al. 2025a; Huang et al. 2025b; Li et al. 2025b; Kuo et al. 2025; Parmar and Govindarajulu 2025). In contrast, comprehensive evaluations examining the broader foundational capabilities of LRMs, including both helpfulness and harmlessness remain relatively unexplored. Our work aims to bridge this research gap by systematically investigating how deliberative reasoning impacts LRMs' foundational capabilities beyond specialized contexts.

3 Foundational Capability Evaluation of LRMs

3.1 Evaluation Setup

Models We conduct comprehensive evaluations of LRMs from various model families and scales to systematically examine the impact on foundational performance resulting from the acquisition of strong reasoning capabilities through distillation or large-scale RL. Specifically, we analyze models across different scales, including 7B, 32B, 70B, and 671B, from the DeepSeek (Guo et al. 2025), Qwen (Yang et al. 2024), and LLaMA (Dubey et al. 2024) model families. Detailed specifications and configurations of the evaluated models are summarized in Table 1. For more training details of each LRM, please refer to Appendix A.

Benchmarks We evaluate the foundational capabilities of different LRMs using two widely recognized dimensions: helpfulness and harmlessness (Ouyang et al. 2022). Specifically, following Dubey et al. (2024); Yang et al. (2024); Team et al. (2024), helpfulness encompasses general tasks assessed by **MMLU-Pro** (Wang et al. 2024b) and **Live-Bench** (White et al. 2024) (excluding math, coding, and reasoning), along with instruction-following abilities measured by **IFEval** (Zhou et al. 2023) and **MT-Bench** (Zheng et al. 2023). Harmlessness evaluations include responses to vanilla harmful prompts: **StrongReject** (Souly et al. 2024) and jailbreak attacks: **WildJailbreak** (Jiang et al. 2024). For further detailed description and evaluation settings of these benchmarks, please refer to Appendix B.

Metrics For **MMLU-Pro**, we adopt zero-shot chain-of-thought (CoT) evaluation method, and accuracy is the primary metric. For the instruction-following tasks in **Live-Bench** and **IFEval**, we employ prompt-level evaluations, where models must meet all input requirements for each prompt, resulting in binary scoring (0 or 1). For other tasks in **Live-Bench**, we calculate scores using methods such as matching and similarity computation, resulting in a score between 0 and 1. And **MT-Bench** is scored on a scale of 1 to 10 using GPT-4o (OpenAI 2024) for evaluation, following the method of LLM-as-a-judge (Zheng et al. 2023). For harmlessness evaluations, we adopt rejection rate as the evaluation metric, which is also judged and calculated by GPT-4o. Please refer to Appendix C for detailed evaluation prompts.

Implementation Details We conduct inference for LRMs using vLLM on 8 NVIDIA A100 GPUs. All LRMs' decoding hyper-parameters and prompt formatting strictly follow their respective official configurations.

3.2 Overall Results

Incorporating deliberative reasoning into LRMs significantly diminishes their foundational capabilities, negatively impacting both helpfulness and harmlessness. As shown in Table 2, distilled LRMs exhibit notably lower performance across most benchmarks of foundational capability compared to their original chat-based counterparts. Interestingly, the R1-Distill-Llama-70B model demonstrates an enhanced resistance to jailbreak attacks compared to the Llama-3.3-70B-Instruct model—a detailed exploration of this improvement will be provided in §4.1. For models trained via large-scale reinforcement learning, concretely evaluating shifts in their foundational capabilities is particularly challenging due to the lack of intermediate training checkpoints from the model developers. However, by modifying their reasoning modes and intensities, we can infer variations in their overall performance. A deeper analysis of these insights is provided and discussed in §4.2.

LRMs incur significant inference-time overhead when performing general tasks. Figure 2 compares the thought and response lengths of various 32B-scale LRMs across benchmarks. These models consistently generate reasoning processes that are considerably longer than their final responses, in stark contrast to their baseline chat counterparts.

	General Tasks		Instruction Following		Safety	
	MMLU-Pro	**Live-Bench**	**IFEval**	**MT-Bench**	**StrongReject**	**WildJailbreak**
Qwen2.5-7B-Instruct	**54.44**	**36.34**	**67.84**	**7.94**	**95.21**	10.70
OpenThinker-7B	39.04	20.81	34.20	7.33	37.29	**12.45**
Qwen2.5-32B-Instruct	**67.07**	**53.85**	**77.26**	**8.32**	**95.00**	**13.30**
OpenThinker-32B	58.13	45.47	54.16	8.16	46.04	5.80
s1.1-32B	43.77	34.42	37.34	7.98	49.38	4.90
Llama-3.3-70B-Instruct	70.54	**60.30**	**89.83**	**8.11**	**95.63**	19.50
R1-Distill-Llama-70B	**71.57**	54.09	76.89	8.03	89.17	**28.25**

Table 2: The overall results of different LRMs on benchmarks for the evaluation of foundational capabilities. The best results are highlighted in bold.

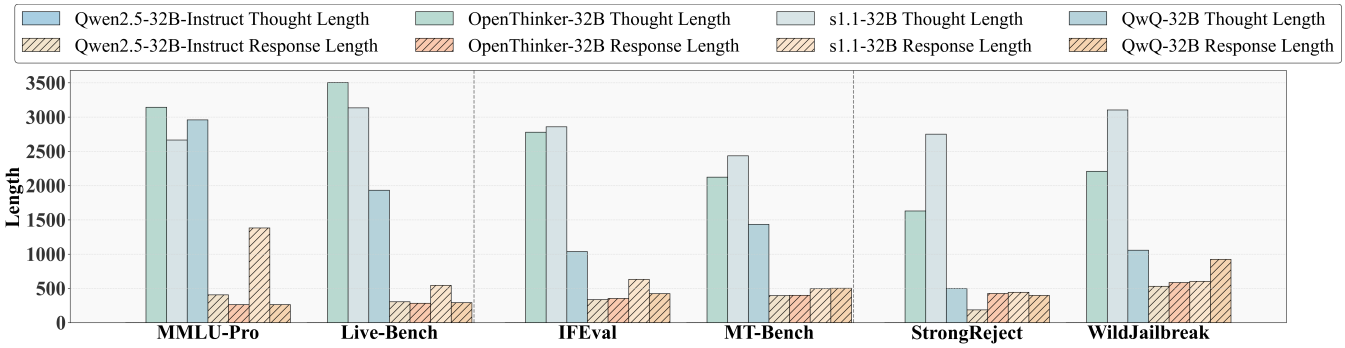


Figure 2: The thought and response lengths of various 32B-scale LRMs across benchmarks.

However, as indicated in Table 2, this substantial increase in computational effort does not lead to better general task performance. This discrepancy underscores a major inefficiency, suggesting that deploying LRMs for general tasks results in unnecessary computational resource consumption.

	Llama-3.3-70B-Instruct	R1-Distill-Llama-70B
DA	51.75	**56.95**
IF	**85.36**	71.94
LC	**43.79**	33.11
Mean	**60.30**	54.00

Table 3: Fine-grained analysis of performance changes in R1-Distill-Llama-70B on Live-Bench.

4 Deeper Analysis

4.1 RQ1: Which Foundational Capabilities Are Most Affected by Deliberative Reasoning?

Analysis on the Helpfulness Degradation We first conduct a fine-grained analysis of the performance changes in LRMs on Live-Bench, as detailed in Table 3. Our analysis indicates that the R1-Distill-Llama-70B exhibits significant performance improvements in Data Analysis (DA) tasks. This enhancement primarily results from an abundance of code-formatted data in its training set, effectively matching the JSON- and markdown-formatted tasks, such as table reformatting and comprehension. Conversely, the observed declines in Instruction Following (IF) and Language Comprehension (LC) performance stem largely from the models' tendency to overlook detailed instructions during reasoning, resulting in outputs that deviate from strict adherence to user-specified requirements. Specific illustrative examples and more analysis are provided in Appendix D.1.

And the primary reason for the performance degradation of LRMs on IFEval is their reasoning process, which predominantly emphasizes understanding input content while overlooking explicit formatting requirements. As illustrated in Table 1 in Appendix D.1, users specified precise formatting instructions, particularly requests for responses to end with certain phrases. Although R1-Distill-Llama-70B effectively capture the user's intention in its reasoning process and clearly outlined five appropriate steps, it inadvertently added an extra word, "the", in its final sentence, thereby violating the required response format. This issue represents a main cause of LRMs' decline in IFEval performance.

Finally, we also perform a turn-level analysis of LRMs' results on the multi-turn instruction-following MT-Bench dataset. Figure 6 in Appendix F.1 presents the performance changes of different LRMs across the first and second turns in MT-Bench. Overall, LRMs developed through distillation exhibited a more pronounced performance drop specifically in the second turn, leading to degraded multi-turn instruction-

	IF-Eval		MT-Bench		StrongReject		WildJailbreak	
	Win. Length	**Lose. Length**	**Win. Length**	**Lose. Length**	**Win. Length**	**Lose. Length**	**Win. Length**	**Lose. Length**
OpenThinker-7B	3145.14	2214.96	1796.01	3786.39	1421.73	2410.18	2875.47	2653.70
OpenThinker-32B	2959.91	2858.75	1877.64	2813.93	1144.25	2053.59	2000.77	2055.25
s1.1-32B	2463.40	2732.13	2095.17	3455.33	2690.8	2811.87	3241.24	3438.10
R1-Distill-Llama-70B	380.36	508.08	759.28	1138.43	504.85	781.38	356.79	698.08
QwQ-32B	828.14	1209.74	1287.54	2190.54	476.21	1180.50	657.67	981.14

Table 4: Comparison of thought length between performance-improving (Win. Length) and performance-declining (Lose. Length) samples across different benchmarks.

OpenThinker-7B	Safe Answer	Unsafe Answer
Safe Thought	0	0.95%
Unsafe Thought	12.45%	**86.60%**
OpenThinker-32B	Safe Answer	Unsafe Answer
Safe Thought	0.15	1.40%
Unsafe Thought	5.65%	**92.80%**
QwQ-32B	Safe Answer	Unsafe Answer
Safe Thought	3.10%	5.45%
Unsafe Thought	7.55%	**83.90%**
R1-Distill-Llama-70B	Safe Answer	Unsafe Answer
Safe Thought	11.20%	8.45%
Unsafe Thought	17.05%	**63.30%**

Table 5: Safety analysis of different LRMs from different model families and parameter scales on the WildJailbreak benchmark. We categorize responses based on whether the LRM's thought process is safe or unsafe and whether the final answer is safe or unsafe.

following capability.

Analysis on the Safety of the Thought and Response Table 5 presents the analysis of the thought and response safety of different LRMs on the WildJailbreak benchmark. We derive the following key insights from the results: (1) **Thoughts are generally more unsafe than responses**. For instance, in the case of R1-Distill-Llama-70B, the unsafe rate of responses is 71% (8% unsafe answers from safe thoughts + 63% unsafe answers from unsafe thoughts). However, the unsafe rate of thoughts reaches 80% (8% safe answers from unsafe thoughts + 63% unsafe answers from unsafe thoughts + 17% safe answers from unsafe thoughts), indicating that the internal reasoning process of LRMs tends to be riskier than their final outputs. (2) **Unsafe thoughts are the primary cause of unsafe responses.** Across all models, unsafe thoughts overwhelmingly lead to unsafe responses. Even when responses are labeled as safe, their underlying thoughts often contain unsafe and harmful content. These findings reveal the nuanced risks in LRM safety and emphasize the need for more robust safety measures that address both the internal reasoning process and the final outputs. For detailed case analyses, please refer to Appendix D.2.

Analysis of Thought Length and Performance Variation In Table 4, by comparing LRMs with their corresponding chat-based backbones across different benchmarks, we observe a consistent phenomenon: **the thought length of samples where performance declines is significantly longer than that of samples where performance improves.**

This finding suggests that an overly prolonged reasoning process may actually harm a LRM's foundational capability. This contradicts the widely accepted assumption that increasing inference-time compute would always lead to better performance. While we acknowledge that this assumption might hold in reasoning tasks, our results indicate that for more general tasks, the opposite effect occurs. This analysis further highlights the importance of dynamically adaptive inference-time compute allocation in LRMs depending on the nature of the task. Instead of indiscriminately increasing compute, LRMs should strategically balance thought length to optimize performance across diverse tasks.

4.2 RQ2: How Does Inference-Time Compute Affect LRMs' Performance on General Tasks?

Thinking Mode We manipulate LRMs' reasoning modes to achieve varying levels of inference-time compute. Through this controlled approach, we further explore how the deliberative reasoning processes inherent to LRMs affect their overall foundational capabilities (Muennighoff et al. 2025; Jiang et al. 2025). Specifically, by inserting special thinking tokens such as <think></think> at different points within LRMs' reasoning processes, we implement the following reasoning modes:

- **Zero-thinking**: We append the special end-of-thinking token (e.g., </think>) immediately after the input, forcing the model to bypass the reasoning process and directly generate responses.
- **Less-thinking**: We inserted the </think> token at a certain percentage ($p\%$) of the model's original reasoning process, prematurely terminating deliberation and prompting the final response generation. Specifically, $p\%$ is set to 10%, 20%, 50%, 60%, 80%, 90% in our experiments.
- **Summary-Thinking**: We summarize the model's original reasoning process using GPT-4o (OpenAI 2024), then reinsert this condensed version between the <think></think> tokens, allowing the model to generate responses based on this summarized reasoning. Detailed summary prompts and their corresponding outcomes are provided in Appendix E.
- **Summary-Thinking-Plus**: Recent study suggests that LRMs' reasoning often begins with consistent patterns that

	General Tasks		Instruction Following		Safety	
	MMLU-Pro	**Live-Bench**	**IFEval**	**MT-Bench**	**StrongReject**	**WildJailbreak**
OpenThinker-7B	39.04	**20.81**	32.72	7.58	37.29	12.45
+ Zero-Thinking	15.96	13.21	**33.27**	7.38	**79.79**	**14.10**
+ Summary-Thinking	**42.66**	12.48	28.10	7.47	50.83	8.70
+ Summary-Thinking-Plus	42.61	14.57	28.83	**7.66**	37.50	8.05
OpenThinker-32B	58.13	**45.47**	**54.16**	**8.16**	46.04	5.80
+ Zero-Thinking	44.43	19.43	37.34	8.03	**88.54**	**9.40**
+ Summary-Thinking	60.44	28.19	47.87	8.03	65.00	6.00
+ Summary-Thinking-Plus	**60.52**	28.77	45.66	7.94	52.29	6.40
s1.1-32B	43.77	33.72	37.34	7.98	49.38	4.90
+ Zero-Thinking	48.38	30.56	42.33	7.91	64.79	**11.15**
+ Summary-Thinking	69.61	44.16	54.16	**8.14**	**53.96**	4.70
+ Summary-Thinking-Plus	**69.92**	**44.23**	**54.34**	8.04	52.29	4.75
R1-Distill-Llama-70B	71.57	**54.00**	**75.60**	8.03	89.17	28.25
+ Zero-Thinking	41.17	38.91	63.22	7.33	**99.17**	**89.10**
+ Summary-Thinking	63.06	48.58	70.79	8.14	93.54	31.00
+ Summary-Thinking-Plus	66.92	51.54	73.75	**8.21**	92.29	25.60
QwQ-32B	72.94	**68.64**	75.60	8.51	95.00	10.65
+ Zero-Thinking	48.24	51.58	64.51	8.57	**98.33**	**59.65**
+ Summary-Thinking	76.07	66.64	77.26	8.57	93.33	11.60
+ Summary-Thinking-Plus	76.18	68.44	**78.92**	**8.67**	95.00	12.35

Table 6: The results of different LRMs under the Zero-Thinking, Summary-Thinking and Summary-Thinking-Plus mode for the evaluation of foundational capabilities. The best results are highlighted in bold.

significantly influence accuracy (Ji et al. 2025). Therefore, we preserve the first sentence of the original reasoning trace and append it to the beginning of the summarized thoughts, examining its effect on performance.

Results and Analysis

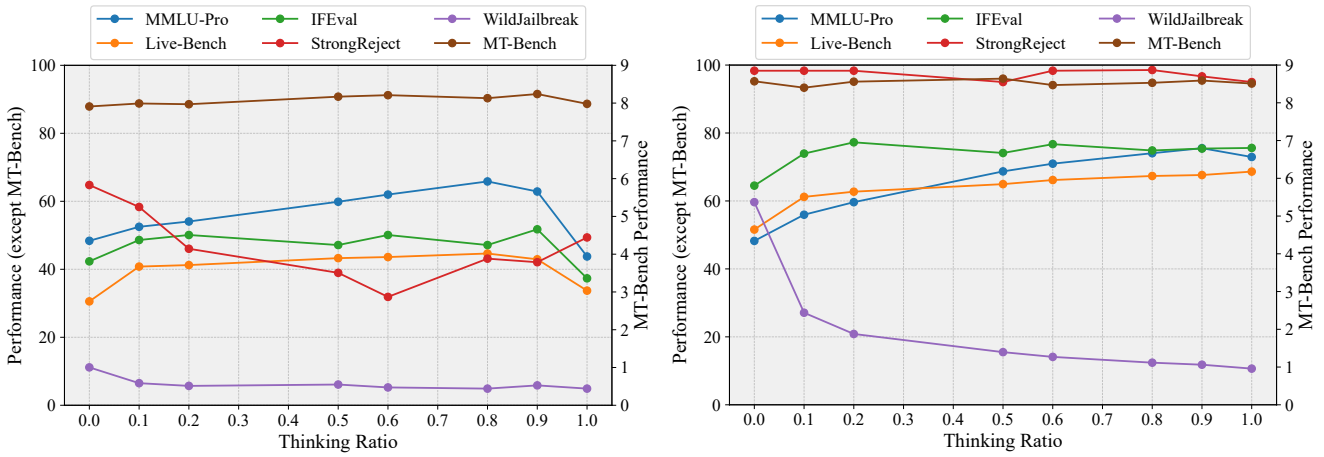
> Zero-Thinking mode substantially enhances the harmlessness of LRMs but further reduces their helpfulness.

In Table 6, in the Zero-Thinking mode, LRMs trained via large-scale RL or distillation demonstrate significant improvements in resisting harmful queries and jailbreak attacks. For example, R1-Distill-Llama-70B's resistance to jailbreak attacks surges from 28.25 to 89.10, outperforming even the original chat version (19.5) by a substantial margin. Likewise, QwQ-32B shows notable progress in the WildJailbreak benchmark, increasing from 10.65 to 59.65. Further case studies, presented in Table 2 in Appendix D.2, indicate that this enhanced performance is primarily driven by the model's reduced susceptibility to unsafe reasoning when bypassing deliberative thinking, reinforcing our findings in §4.1.

> Summary-Thinking and Less-Thinking modes streamline the reasoning process, leading to notabe enhancement in LRMs' helpfulness.

As shown in Table 6, LRMs under Summary-Thinking consistently demonstrate improved performance on helpfulness, particularly evident in s1.1-32B and QwQ-32B. This finding suggests that excessively verbose reasoning may hinder LRMs' effectiveness in general tasks. Furthermore, retaining the first sentence of the original reasoning trace in Summary-Thinking-Plus consistently yields even better outcomes, validating the significance of initial reasoning patterns for accurate results, thus corroborating recent findings on mathematical tasks (Ji et al. 2025). To further investigate this improvement, we compare the original and summary-based reasoning traces in s1.1-32B, and find that the most salient difference lies in the significant reduction of *self-reflection* patterns in the summary variant. This reduction may be a key factor contributing to performance gains. Detailed analysis is provided in Appendix F.3.

A deeper analysis of the Less-Thinking mode reveals further insights into LRMs' adaptive reasoning behavior. Specifically, Figure 3(a) presents the Less-Thinking results for s1.1-32B, an LRM derived via model distillation, while Figure 3(b) illustrates the Less-Thinking results for QwQ-32B, an LRM obtained through large-scale RL. Despite differences in their training methods, both models exhibit a consistent trend: the optimal performance across most datasets occurs at varying thinking ratios within the Less-Thinking mode, suggesting that a fixed reasoning proportion does not universally maximize effectiveness across diverse tasks. Notably, the LRMs' original Deliberative Reasoning mode (with a thinking ratio of 1.0) generally does not yield optimal performance, further highlighting the necessity of adaptive reasoning strategies within current LRMs. For the results of Less-Thinking on



(a) Less-Thinking of s1.1-32B. (b) Less-Thinking of QwQ-32B.

Figure 3: Performance analysis of LRMs under the Less-Thinking mode across multiple benchmarks. The x-axis denotes the Thinking Ratio, indicating the proportion of deliberate reasoning steps utilized during inference. (a) The results for the distilled LRM (s1.1-32B), (b) The results for the reinforcement learning-based LRM (QwQ-32B).

more LRMs, please refer to Appendix F.2.

Collectively, the empirical results and analyses in this section underscore a critical conclusion:

> Deploying LRMs effectively requires ***adaptive reasoning*** strategies tailored specifically to different tasks, emphasizing the need for task-specific customization of inference-time compute allocation.

5 Discussion on Future Direction

Our experimental results and analyses offer valuable guidance for future development of more comprehensive and usable LRMs, whether achieved through distillation or large-scale Reinforcement Learning.

Regarding distillation methods, future research should explore two primary directions to enhance the deep reasoning capabilities of LRMs while safeguarding their general performance on common tasks. Firstly, in terms of data selection, enriching the diversity of training data is crucial to ensure that LRMs maintain robustness in both helpfulness and harmlessness dimensions. Secondly, from the perspective of algorithmic design, incorporating continual learning techniques appears promising (Wang et al. 2024a; Wu et al. 2024). Possible solutions include further regularizing the gradient optimization process (Wang et al. 2023) or introducing additional architectures designed to mitigate performance trade-offs (Zhao et al. 2024).

More importantly, future work should investigate methods to enable LRMs to dynamically adjust their inference-time compute based on input difficulty, thus achieving adaptive reasoning (Snell et al. 2024; Chen et al. 2024). For instance, recent studies suggest that training smaller models exclusively on lengthy CoT-distilled data could adversely affect their overall performance, whereas blending short and long CoT data tends to yield superior distillation outcomes (Li et al. 2025c). Similarly, other research efforts have considered integrating the length of the reasoning process as an additional reward factor into large-scale reinforcement learning frameworks to penalize those tedious reasoning traces. (Kimi et al. 2025; Zhao et al. 2025).

6 Conclusion

In this work, we systematically investigate the trade-offs of integrating deliberative reasoning into LRMs. While recent advances boost performance on complex reasoning tasks, we find that acquiring such abilities—via distillation or reinforcement learning—can degrade core capabilities like helpfulness and safety, and increase inference costs. We show that adaptive reasoning strategies can help strike a balance, offering a practical path toward more versatile and efficient LRMs.

Limitation

While our study offers useful empirical insights into how deliberative reasoning, adaptive reasoning, and foundational capabilities interact in LRMs, several limitations remain. First, our evaluations mainly cover general tasks, instruction following, safety, and reasoning benchmarks. These tasks reflect core abilities but do not capture broader real-world scenarios such as multimodal interaction or open-ended dialogue. Second, although we assess three major LRM families (DeepSeek, Qwen, LLaMA), our analysis is limited to specific checkpoints. Whether our observations hold for other architectures—such as mixture-of-experts or multimodal LLMs—still requires further validation.

Acknowledgments

We thank the anonymous reviewers for their comments and suggestions. This work was supported by the New Generation Artificial Intelligence-National Science and Technology Major Project 2023ZD0121100, the National Natural Science Foundation of China (NSFC) via grant 62441614 and

62176078, the Fundamental Research Funds for the Central Universities.

References

Aggarwal, P.; and Welleck, S. 2025. L1: Controlling How Long A Reasoning Model Thinks With Reinforcement Learning. *arXiv preprint arXiv:2503.04697*.

Anthropic. 2025. Claude 3.7 Sonnet and Claude Code. *Anthropic's Blog*.

Arrieta, A.; Ugarte, M.; Valle, P.; Parejo, J. A.; and Segura, S. 2025. o3-mini vs DeepSeek-R1: Which One is Safer? *arXiv preprint arXiv:2501.18438*.

Ballon, M.; Algaba, A.; and Ginis, V. 2025. The Relationship Between Reasoning and Performance in Large Language Models–o3 (mini) Thinks Harder, Not Longer. *arXiv preprint arXiv:2502.15631*.

Bandyopadhyay, D.; Bhattacharjee, S.; and Ekbal, A. 2025. Thinking Machines: A Survey of LLM based Reasoning Strategies. *arXiv preprint arXiv:2503.10814*.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Proc. of NeurIPS*, 1877–1901.

Camposampiero, G.; Hersche, M.; Wattenhofer, R.; Sebastian, A.; and Rahimi, A. 2025. Can Large Reasoning Models do Analogical Reasoning under Perceptual Uncertainty? *arXiv preprint arXiv:2503.11207*.

Chen, A.; Song, Y.; Zhu, W.; Chen, K.; Yang, M.; Zhao, T.; et al. 2025a. Evaluating o1-like llms: Unlocking reasoning for translation through comprehensive analysis. *arXiv preprint arXiv:2502.11544*.

Chen, Q.; Qin, L.; Liu, J.; Peng, D.; Guan, J.; Wang, P.; Hu, M.; Zhou, Y.; Gao, T.; and Che, W. 2025b. Towards Reasoning Era: A Survey of Long Chain-of-Thought for Reasoning Large Language Models. *arXiv preprint arXiv:2503.09567*.

Chen, X.; Xu, J.; Liang, T.; He, Z.; Pang, J.; Yu, D.; Song, L.; Liu, Q.; Zhou, M.; Zhang, Z.; et al. 2024. Do not think that much for 2+ 3=? on the overthinking of o1-like llms. *arXiv preprint arXiv:2412.21187*.

DeepMind, G. 2025. Gemini 2.0 Flash Thinking Experimental Model 01-21. *Google DeepMind's Blog*.

Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Yang, A.; Fan, A.; et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Feng, X.; Dou, L.; and Kong, L. 2025. Reasoning Does Not Necessarily Improve Role-Playing Ability. *arXiv preprint arXiv:2502.16940*.

Golde, J.; Haller, P.; Barth, F.; and Akbik, A. 2025. MastermindEval: A Simple But Scalable Reasoning Benchmark. In *Workshop on Reasoning and Planning for Large Language Models*.

Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Huang, S.; Yang, L.; Song, Y.; Chen, S.; Cui, L.; Wan, Z.; Zeng, Q.; Wen, Y.; Shao, K.; Zhang, W.; et al. 2025a. ThinkBench: Dynamic Out-of-Distribution Evaluation for Robust LLM Reasoning. *arXiv preprint arXiv:2502.16268*.

Huang, T.; Hu, S.; Ilhan, F.; Tekin, S. F.; Yahn, Z.; Xu, Y.; and Liu, L. 2025b. Safety Tax: Safety Alignment Makes Your Large Reasoning Models Less Reasonable. *arXiv preprint arXiv:2503.00555*.

Jaech, A.; Kalai, A.; Lerer, A.; Richardson, A.; El-Kishky, A.; Low, A.; Helyar, A.; Madry, A.; Beutel, A.; Carney, A.; et al. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

Ji, K.; Xu, J.; Liang, T.; Liu, Q.; He, Z.; Chen, X.; Liu, X.; Wang, Z.; Chen, J.; Wang, B.; et al. 2025. The First Few Tokens Are All You Need: An Efficient and Effective Unsupervised Prefix Fine-Tuning Method for Reasoning Models. *arXiv preprint arXiv:2503.02875*.

Jiang, F.; Xu, Z.; Li, Y.; Niu, L.; Xiang, Z.; Li, B.; Lin, B. Y.; and Poovendran, R. 2025. SafeChain: Safety of Language Models with Long Chain-of-Thought Reasoning Capabilities. *arXiv preprint arXiv:2502.12025*.

Jiang, L.; Rao, K.; Han, S.; Ettinger, A.; Brahman, F.; Kumar, S.; Mireshghallah, N.; Lu, X.; Sap, M.; Choi, Y.; et al. 2024. Wildteaming at scale: From in-the-wild jailbreaks to (adversarially) safer language models. *Proc. of NeurIPS*, 47094–47165.

Kahneman, D. 2011. *Thinking, fast and slow*. macmillan.

Kimi, T.; Du, A.; Gao, B.; Xing, B.; Jiang, C.; Chen, C.; Li, C.; Xiao, C.; Du, C.; Liao, C.; et al. 2025. Kimi k1. 5: Scaling reinforcement learning with llms. *arXiv preprint arXiv:2501.12599*.

Kumar, K.; Ashraf, T.; Thawakar, O.; Anwer, R. M.; Cholakkal, H.; Shah, M.; Yang, M.-H.; Torr, P. H.; Khan, S.; and Khan, F. S. 2025. Llm post-training: A deep dive into reasoning large language models. *arXiv preprint arXiv:2502.21321*.

Kuo, M.; Zhang, J.; Ding, A.; Wang, Q.; DiValentin, L.; Bao, Y.; Wei, W.; Juan, D.-C.; Li, H.; and Chen, Y. 2025. H-cot: Hijacking the chain-of-thought safety reasoning mechanism to jailbreak large reasoning models, including openai o1/o3, deepseek-r1, and gemini 2.0 flash thinking. *arXiv preprint arXiv:2502.12893*.

Li, D.; Cao, S.; Griggs, T.; Liu, S.; Mo, X.; Patil, S. G.; Zaharia, M.; Gonzalez, J. E.; and Stoica, I. 2025a. LLMs Can Easily Learn to Reason from Demonstrations Structure, not content, is what matters! *arXiv preprint arXiv:2502.07374*.

Li, X.; Li, Z.; Kosuga, Y.; and Bian, V. 2025b. Output Length Effect on DeepSeek-R1's Safety in Forced Thinking. *arXiv preprint arXiv:2503.01923*.

Li, Y.; Yue, X.; Xu, Z.; Jiang, F.; Niu, L.; Lin, B. Y.; Ramasubramanian, B.; and Poovendran, R. 2025c. Small Models Struggle to Learn from Strong Reasoners. *arXiv preprint arXiv:2502.12143*.

Li, Z.-Z.; Zhang, D.; Zhang, M.-L.; Zhang, J.; Liu, Z.; Yao, Y.; Xu, H.; Zheng, J.; Wang, P.-J.; Chen, X.; et al. 2025d. From System 1 to System 2: A Survey of Reasoning Large Language Models. *arXiv preprint arXiv:2502.17419*.

Luo, H.; Shen, L.; He, H.; Wang, Y.; Liu, S.; Li, W.; Tan, N.; Cao, X.; and Tao, D. 2025. O1-Pruner: Length-Harmonizing Fine-Tuning for O1-Like Reasoning Pruning. *arXiv preprint arXiv:2501.12570*.

Muennighoff, N.; Yang, Z.; Shi, W.; Li, X. L.; Fei-Fei, L.; Hajishirzi, H.; Zettlemoyer, L.; Liang, P.; Candès, E.; and Hashimoto, T. 2025. s1: Simple test-time scaling. *arXiv preprint arXiv:2501.19393*.

OpenAI. 2024. GPT-4o System Card. *OpenAI*.

OpenAI. 2025. OpenAI o3-mini System Card. *OpenAI's Blog*.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *Proc. of NeurIPS*, 27730–27744.

Parmar, M.; and Govindarajulu, Y. 2025. Challenges in Ensuring AI Safety in DeepSeek-R1 Models: The Shortcomings of Reinforcement Learning Strategies. *arXiv preprint arXiv:2501.17030*.

Qwen, T. 2025. QwQ-32B: Embracing the Power of Reinforcement Learning. *Qwen's Blog*.

Rajeev, M.; Ramamurthy, R.; Trivedi, P.; Yadav, V.; Bamgbose, O.; Madhusudan, S. T.; Zou, J.; and Rajani, N. 2025. Cats Confuse Reasoning LLM: Query Agnostic Adversarial Triggers for Reasoning Models. *arXiv preprint arXiv:2503.01781*.

Snell, C.; Lee, J.; Xu, K.; and Kumar, A. 2024. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*.

Souly, A.; Lu, Q.; Bowen, D.; Trinh, T.; Hsieh, E.; Pandey, S.; Abbeel, P.; Svegliato, J.; Emmons, S.; Watkins, O.; et al. 2024. A StrongREJECT for Empty Jailbreaks. In *Proc. of NeurIPS*.

Stanovich, K.; West, R.; and Hertwig, R. 2000. Individual differences in reasoning: Implications for the rationality debate?-Open Peer Commentary-The questionable utility of cognitive ability in explaining cognitive illusions.

Team, G.; Riviere, M.; Pathak, S.; Sessa, P. G.; Hardin, C.; Bhupatiraju, S.; Hussenot, L.; Mesnard, T.; Shahriari, B.; Ramé, A.; et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Team, O. 2025. Open Thoughts.

Tie, G.; Zhao, Z.; Song, D.; Wei, F.; Zhou, R.; Dai, Y.; Yin, W.; Yang, Z.; Yan, J.; Su, Y.; et al. 2025. A Survey on Post-training of Large Language Models. *arXiv preprint arXiv:2503.06072*.

Wang, L.; Zhang, X.; Su, H.; and Zhu, J. 2024a. A comprehensive survey of continual learning: Theory, method and application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Wang, X.; Chen, T.; Ge, Q.; Xia, H.; Bao, R.; Zheng, R.; Zhang, Q.; Gui, T.; and Huang, X.-J. 2023. Orthogonal Subspace Learning for Language Model Continual Learning. In *Proc. of EMNLP Findings*, 10658–10671.

Wang, Y.; Liu, Q.; Xu, J.; Liang, T.; Chen, X.; He, Z.; Song, L.; Yu, D.; Li, J.; Zhang, Z.; et al. 2025. Thoughts Are All Over the Place: On the Underthinking of o1-Like LLMs. *arXiv preprint arXiv:2501.18585*.

Wang, Y.; Ma, X.; Zhang, G.; Ni, Y.; Chandra, A.; Guo, S.; Ren, W.; Arulraj, A.; He, X.; Jiang, Z.; et al. 2024b. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. In *Proc. of NeurIPS*.

White, C.; Dooley, S.; Roberts, M.; Pal, A.; Feuer, B.; Jain, S.; Shwartz-Ziv, R.; Jain, N.; Saifullah, K.; Naidu, S.; et al. 2024. Livebench: A challenging, contamination-free llm benchmark. *arXiv preprint arXiv:2406.19314*.

Wu, T.; Luo, L.; Li, Y.-F.; Pan, S.; Vu, T.-T.; and Haffari, G. 2024. Continual learning for large language models: A survey. *arXiv preprint arXiv:2402.01364*.

Xu, F.; Hao, Q.; Zong, Z.; Wang, J.; Zhang, Y.; Wang, J.; Lan, X.; Gong, J.; Ouyang, T.; Meng, F.; et al. 2025. Towards Large Reasoning Models: A Survey of Reinforced Reasoning with Large Language Models. *arXiv preprint arXiv:2501.09686*.

Yang, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Li, C.; Liu, D.; Huang, F.; Wei, H.; et al. 2024. Qwen2. 5 Technical Report. *arXiv preprint arXiv:2412.15115*.

Ye, Y.; Huang, Z.; Xiao, Y.; Chern, E.; Xia, S.; and Liu, P. 2025. LIMO: Less is More for Reasoning. *arXiv preprint arXiv:2502.03387*.

Zhang, W.; Lei, X.; Liu, Z.; Wang, N.; Long, Z.; Yang, P.; Zhao, J.; Hua, M.; Ma, C.; Wang, K.; et al. 2025. Safety evaluation of deepseek models in chinese contexts. *arXiv preprint arXiv:2502.11137*.

Zhao, W.; Guo, J.; Deng, Y.; Sui, X.; Hu, Y.; Zhao, Y.; Che, W.; Qin, B.; Chua, T.-S.; and Liu, T. 2025. Exploring and Exploiting the Inherent Efficiency within Large Reasoning Models for Self-Guided Efficiency Enhancement. *arXiv preprint arXiv:2506.15647*.

Zhao, W.; Wang, S.; Hu, Y.; Zhao, Y.; Qin, B.; Zhang, X.; Yang, Q.; Xu, D.; and Che, W. 2024. Sapt: A shared attention framework for parameter-efficient continual learning of large language models. In *Proc. of ACL*, 11641–11661.

Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E.; et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Proc. of NeurIPS*, 46595–46623.

Zhou, J.; Lu, T.; Mishra, S.; Brahma, S.; Basu, S.; Luan, Y.; Zhou, D.; and Hou, L. 2023. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*.

Zhou, K.; Liu, C.; Zhao, X.; Jangam, S.; Srinivasa, J.; Liu, G.; Song, D.; and Wang, X. E. 2025a. The hidden risks of large reasoning models: A safety assessment of r1. *arXiv preprint arXiv:2502.12659*.

Zhou, X.; Tie, G.; Zhang, G.; Wang, W.; Zuo, Z.; Wu, D.; Chu, D.; Zhou, P.; Sun, L.; and Gong, N. Z. 2025b. Large Reasoning Models in Agent Scenarios: Exploring the Necessity of Reasoning Capabilities. *arXiv preprint arXiv:2503.11074*.