

A Survey of Efficient Reasoning for Large Reasoning Models: Language, Multimodality, and Beyond

Xiaoye Qu^{1*}, Yafu Li^{1*}, Zhaochen Su^{2†}, Weigao Sun^{1†}, Jianhao Yan^{3‡}, Dongrui Liu^{1†}, Ganqu Cui¹, Daizong Liu⁴, Shuxian Liang⁵, Junxian He⁶, Peng Li⁷, Wei Wei⁸, Jing Shao¹, Chaochao Lu¹, Yue Zhang³, Xian-Sheng Hua⁵, Bowen Zhou^{1,7}, Yu Cheng^{9‡}

¹ Shanghai AI Laboratory ² Soochow University ³ Westlake University ⁴ Peking University
⁵ Tongji University ⁶ The Hong Kong University of Science and Technology
⁷ Tsinghua University ⁸ Huazhong University of Science and Technology
⁹ The Chinese University of Hong Kong

Abstract

Recent Large Reasoning Models (LRMs), such as DeepSeek-R1 and OpenAI o1, have demonstrated strong performance gains by scaling up the length of Chain-of-Thought (CoT) reasoning during inference. However, a growing concern lies in their tendency to produce excessively long reasoning traces, which are often filled with redundant content (*e.g.*, repeated definitions), over-analysis of simple problems, and superficial exploration of multiple reasoning paths for harder tasks. This inefficiency introduces significant challenges for training, inference, and real-world deployment (*e.g.*, in agent-based systems), where token economy is critical. In this survey, we provide a comprehensive overview of recent efforts aimed at improving reasoning efficiency in LRMs, with a particular focus on the unique challenges that arise in this new paradigm. We identify common patterns of inefficiency, examine methods proposed across the LRM lifecycle, *i.e.*, from pretraining to inference, and discuss promising future directions for research. To support ongoing development, we also maintain a real-time GitHub repository tracking recent progress in the field.¹ We hope this survey serves as a foundation for further exploration and inspires innovation in this rapidly evolving area.

1 Introduction

“Brevity is the soul of wit.”

—William Shakespeare

Large Language Models (LLMs), such as DeepSeek V3 (Liu et al., 2024a), Qwen 2.5 (Yang et al., 2024a), LLaMA 3 (Dubey et al., 2024), and GPT-4o (Hurst et al., 2024), have demonstrated

*Project Lead and Equal Contributions.

†Core Contributors.

‡Corresponding Author.

¹https://github.com/XiaoYee/Awesome_Efficient_LRM_Reasoning

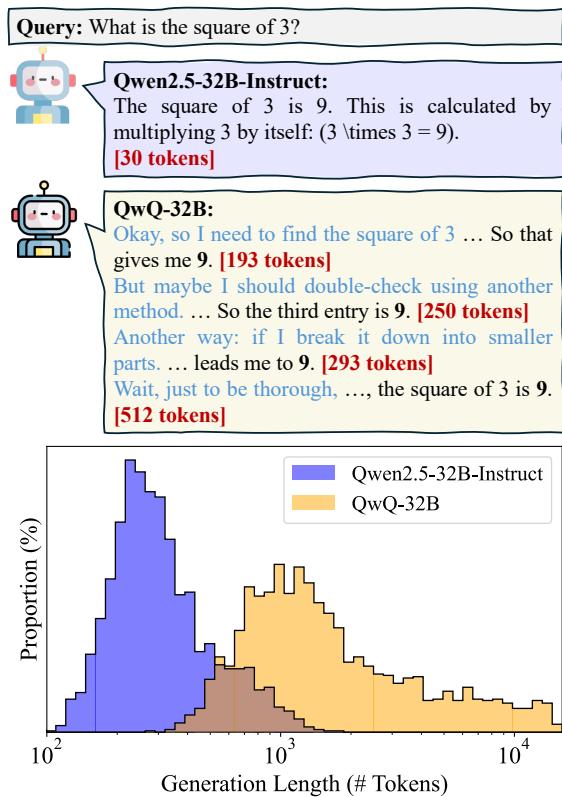


Figure 1: **Top:** To answer the elementary school-level math problem, LRM (QwQ-32B) consumes altogether 1248 tokens, while the instruct LLM counterpart (Qwen2.5-32B-Instruct) only needs 30 tokens. **Bottom:** The distribution of generation length of two models on a mixed set of math problems sourced from GSM8K, MATH-500, and AIME 2024.

remarkable capabilities across a wide range of tasks (Chang et al., 2024; Zhao et al., 2023; Kasneci et al., 2023; Zhu et al., 2024). These models operate in a manner akin to *System 1* thinking (Frankish, 2010; Kahneman, 2011; Li et al., 2025e), characterized by fast, intuitive, and automatic decision-making. However, complex reasoning tasks, such as advanced mathematics (Lightman et al., 2023; Besta et al., 2024; Ahn et al., 2024)

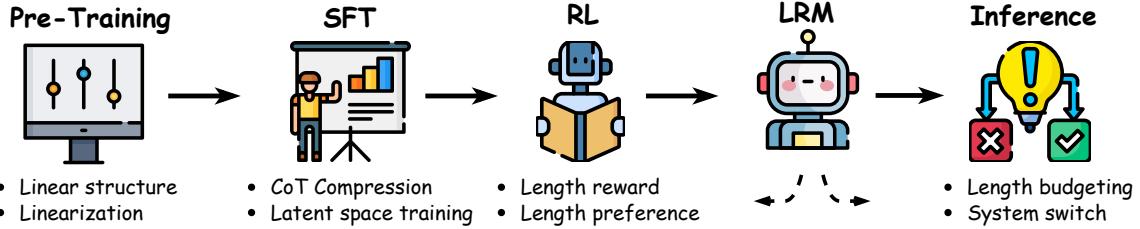


Figure 2: In this paper, we comprehensively study methods for efficient reasoning from the stages of Per-training, Supervised Fine-tuning (SFT), Reinforcement Learning (RL), and Inference.

and formal logic (Huang and Chang, 2022; Pan et al., 2023), demand more deliberate, structured analysis. To tackle these challenges, a new class of models has emerged: *Large Reasoning Models (LRMs)*, including DeepSeek R1 (Guo et al., 2025), OpenAI-o1/o3 (Jaech et al., 2024), and QwQ (Team, 2024). These models enhance performance by explicitly generating intermediate reasoning steps, collectively known as *chain-of-thought (CoT)* (Wei et al., 2022), before producing a final answer. Unlike the rapid and heuristic-driven behavior of LLMs, LRMs exhibit deliberate and analytical reasoning, resembling *System 2* thinking (Evans, 2003; Kannengiesser and Gero, 2019). This paradigm shift from System 1 to System 2 reasoning (Li et al., 2025e) underpins the development of more capable reasoning agents. Yet, this capability comes at a cost: the reasoning process of LRMs tends to be slower and more verbose. As shown in Figure 1, for an elementary-school level math problem, QwQ-32B produces significantly more tokens than its System-1 counterpart, Qwen2.5-32B-Instruct. Moreover, the distributions demonstrate that LRM model QwQ-32B exhibits a significantly greater output length compared to the LLM model Qwen2.5-32B-Instruct. This observation naturally raises a critical question:

Beyond reasoning performance, how can we make LRMs reason more efficiently, thus maximizing the intelligence per token?

In the age of LRMs, we propose that “*Efficiency is the essence of intelligence.*” Just as a wise human knows when to stop thinking and start deciding, a wise model should know when to halt unnecessary deliberation. An intelligent model should manipulate the token economy, i.e., allocating tokens purposefully, skipping redundancy, and optimizing the path to a solution. Rather than naively traversing every possible reasoning path, it should emulate a master strategist, balancing cost and performance with elegant precision.

1.1 Structure of the Survey

In this survey, we systematically review recent advancements in efficient reasoning within LRMs, categorized according to their stages in the LLM lifecycle, as illustrated in Figure 2. The taxonomy of efficient reasoning methods covered in this survey is illustrated in Figure 3. The survey is organized as follows:

1. Section 2 highlights the pattern of reasoning inefficiency and challenges for achieving efficient reasoning in the era of LRMs.
2. Section 3 introduces methods for efficient reasoning during inference stage.
3. Section 4 describes SFT methods that aim to internalize the concise reasoning.
4. Section 5 presents how to control reasoning length in the RL training.
5. Section 6 details model structures and training paradigms, which inherently are efficient.
6. Finally, we highlight promising future directions to address the limitations identified in each stage in Section 7.

1.2 Position and Contribution

Recently, several survey papers have explored the development of Large Reasoning Models. Besta et al. (2025) and Zeng et al. (2024a) focus on training methodologies for LRMs, while Li et al. (2025e) provide a broader and more comprehensive overview of the field. To offer a more focused perspective, Chen et al. (2025b) examine long chain-of-thought reasoning and categorize existing reasoning paradigms. Additionally, Ji et al. (2025) investigate test-time scaling of LRMs. Despite these valuable contributions, none of the existing surveys specifically address the growing challenge of **reasoning efficiency** in LRMs, which is an emerging yet crucial topic in the deployment, scalability, and practical application of these models.

For earlier generations of LLMs, Zhou et al. (2024) provide a comprehensive survey of methods

for efficient inference. However, efficient reasoning in LRM s introduces a distinct and emerging research challenge, characterized by the generation of excessive and uncontrolled reasoning tokens. While traditional approaches for accelerating inference, e.g., model quantization and pruning (Polino et al., 2018; Xiao et al., 2023a; Xia et al., 2023; Wang et al., 2019; Ma et al., 2023; Cheng et al., 2024), and distributed inference systems (Patel et al., 2024; Hu et al., 2024a; Zhong et al., 2024; Lin et al., 2024a), can reduce latency and computational costs, our focus in this survey is on improving the efficiency of the reasoning process itself, rather than general inference acceleration. To summarize, this survey makes the following key contributions to the literature:

- Instead of offering a general overview of LRM s, we focus on the emerging and critical topic of *efficient reasoning* in LRM s, providing an in-depth and targeted analysis.
- We identify and characterize common patterns of reasoning inefficiency, and outline the current challenges that are unique to improving reasoning efficiency in large models.
- We provide a comprehensive review of recent advancements aimed at enhancing reasoning efficiency, structured across the end-to-end LRM development pipeline, from pretraining and supervised fine-tuning to reinforcement learning and inference.

2 Reasoning Efficiency: Definition, Patterns, and Challenges

Before delving into methods for efficient reasoning, we first define the reasoning efficiency, then examine common patterns of reasoning inefficiency observed in LRM s, followed by a discussion of the unique challenges associated with improving efficiency in the era of LRM s.

2.1 Definition of Reasoning Efficiency

We define reasoning efficiency from a task-distribution perspective, analogous to meta-learning (Hospedales et al., 2020). A common goal for LRM s is to reason efficiently across a wide variety of tasks. Let \mathcal{M} denote an LRM tasked with solving problems. We evaluate its reasoning efficiency η over a distribution of tasks $p(\mathcal{T})$, where each task \mathcal{T} is loosely defined as a pair $\{\mathcal{D}, Q\}$,

with \mathcal{D} representing a dataset and Q a corresponding quality metric. We define the *reasoning efficiency* of model \mathcal{M} as:

$$\eta(\mathcal{M}) = \mathbb{E}_{\mathcal{T} \sim p(\mathcal{T})} \frac{Q(\mathcal{M}, \mathcal{D})}{C(\mathcal{M}, \mathcal{D})}, \quad (1)$$

where $Q(\mathcal{M}, \mathcal{D})$ denotes the solution quality on dataset \mathcal{D} (e.g., accuracy, exact match, or creativity), and $C(\mathcal{M}, \mathcal{D})$ represents the computational cost (e.g., FLOPs, number of generated tokens, or inference latency). This formulation provides a principled way to quantify reasoning efficiency across diverse tasks. It highlights the trade-off between performance and cost: reasoning becomes more efficient either by improving solution quality Q or by reducing computational cost C .

2.2 Patterns of Reasoning Inefficiency

Despite the effectiveness of LRM s in employing long reasoning chains to address complex problems such as AIME (Jaech et al., 2024; DeepSeek-AI, 2025; Team, 2024), patterns of reasoning inefficiency persist. These inefficiencies manifest as excessive generation of redundant content (e.g., repetitive rephrasing of questions), over-analysis and verification of simple problems (as illustrated in Figure 1), or meandering through superficial and suboptimal thoughts.

Redundant Content. Existing LRM s lack explicit optimization for minimizing intermediate reasoning tokens (Munkhbat et al., 2025). As a result, these models frequently exhibit redundancy during reasoning (Song et al., 2025b), allocating a substantial portion of their output to textual coherence instead of core reasoning advancement (Su et al., 2025; Luo et al., 2025a). This type of redundancy increases computational cost $C(\mathcal{M}, \mathcal{D})$ in Eq. 1, thereby reducing efficiency. For example, in Table 1, the LRM uses 138 tokens simply to rephrase and interpret the question. Verbose explanations and repetitive phrasing further bloat the output without aiding the final solution (Munkhbat et al., 2025). Notably, Li et al. (2025b) show that while the overall structure of a chain-of-thought is essential for learning, the content of individual reasoning steps often contributes little to overall model performance and solution accuracy.

Overthinking Simple Questions. Recent studies show that LRM s struggle to allocate their reasoning budget effectively based on task complexity (Luo

User Query: A robe takes 2 bolts of blue fiber and half that much white fiber. How many bolts in total does it take? </think> Okay, let me try to solve this problem.
So the question is about figuring out the total number of bolts needed for a robe. The problem states that ... (span 138 tokens) [Rethink, same logic, already get the correct answer]
Wait, let me confirm that. The problem says, ... (span 87 tokens) [Check another interpretations]
But maybe I should check if there's another interpretation. Could "half that much" refer to something else? ... (span 122 tokens) [Check other unit]
Wait, but maybe I need to consider if "half that much" is in terms of some other unit? ... (span 54 tokens) [Think again, with different notations]
Hmm, that seems straightforward, but maybe I should think again. Let me parse the sentence again. ... Breaking it down:- Blue fiber: 2 bolts.- White fiber: half of the blue's amount. So half of 2 is 1 bolt ... (span 107 tokens) [Check tricks]
Alternatively, maybe there's a trick here? ... (span 138 tokens) [Check mis-read]
Alternatively, perhaps I misread the problem. Let me check again. ... Hmm, maybe I'm overcomplicating. I think that's right. Let me think of another way ... So yeah, same thing. Therefore, the total bolts needed are 3. </think>
Therefore, the final answer is: 3 . I hope it is correct.

Table 1: The answer produced by the recent LRM QwQ-32B for a math question. The model generates redundant content (highlighted in red), including repetitive phrasing. Additionally, it goes through multiple rounds of reasoning (explained in blue for clarity), which further undermines reasoning efficiency on such a straightforward problem.

et al., 2025a; Chiang and Lee, 2024). For instance, Chen et al. (2025c) observe that models often exhibit unwarranted uncertainty on straightforward queries such as $2 + 3 = ?$. Rather than producing a concise and direct solution, these models tend to generate multiple redundant reasoning rounds (Luo et al., 2025b), exploring unnecessary solution paths. As shown in Figure 1, although the LRM reaches the correct answer in the initial reasoning trace, it performs several additional verification steps, ultimately using nearly forty times as many tokens as a standard instruction-tuned LLM. Such redundancy also leads to increased computational cost $C(\mathcal{M}, \mathcal{D})$ in Eq. 1 for simple tasks where the model \mathcal{M} achieves a relatively high quality score $Q(\mathcal{M}, \mathcal{D})$, therefore reducing efficiency.

Incoherent and Suboptimal Reasoning. Wang et al. (2025e) identify a phenomenon termed *under-thinking*, where o1-like LRMs prematurely switch reasoning directions, hindering the development of promising paths. This results in shallow and fragmented reasoning traces, particularly in complex mathematical tasks. Instead of pursuing a coherent, in-depth line of thought, the model hops between multiple approaches superficially, leading to longer reasoning sequences and reduced overall solution quality. Such shallow hopping causes either a reduction in $Q(\mathcal{M}, \mathcal{D})$ or an elevation of $C(\mathcal{M}, \mathcal{D})$ in Eq. 1, thus degrading inference efficiency.

Such reasoning inefficiencies pose significant

challenges across training, inference, and real-world applications. Specifically, excessively long CoT sequences hinder reinforcement learning (RL) optimization (Yeo et al., 2025; Yu et al., 2025a; Yuan et al., 2025b), leading to instability during RL fine-tuning and excessive memory consumption. Moreover, due to the autoregressive nature of LRM decoding, inference latency increases linearly with reasoning length. This results in high inference costs and degraded user experience, especially when reasoning traces exceed 10,000 tokens. The issue becomes more pronounced in multi-agent systems, where timely generation of plans and responses is critical (Huang et al., 2024).

2.3 Unique Challenges for Efficient Reasoning in the Era of LRMs

As LRMs grow increasingly capable in solving complex tasks through chain-of-thought reasoning, achieving efficiency becomes both more difficult and more essential. Unlike traditional LLM efficiency problems, such as model size or inference latency, efficient reasoning introduces its own set of challenges. Below, we outline four key obstacles that hinder progress in this emerging area.

Quantifying Reasoning Utility: A Balancing Act. One of the fundamental challenges in efficient reasoning is the difficulty of evaluating the utility of each step in a reasoning chain. Unlike classification or regression tasks where loss can be directly

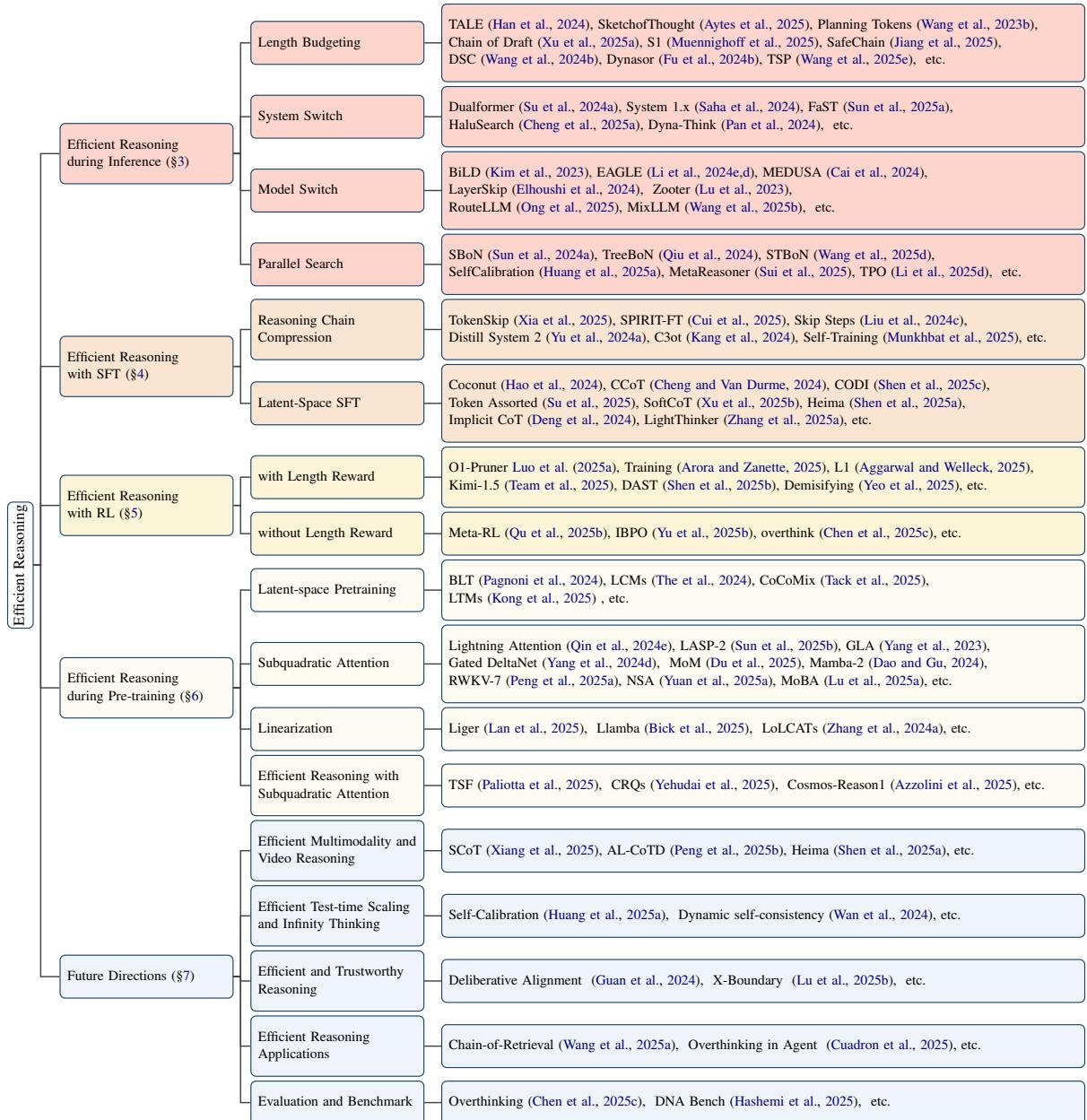


Figure 3: Taxonomy of efficient reasoning methods for LRMs and future directions.

computed, it is hard to quantify how much each intermediate reasoning token contributes to the final answer. This lack of granularity makes it difficult to determine which parts of the reasoning can be compressed or pruned (Zhang et al., 2025c), and how to balance reasoning conciseness with answer correctness (Sprague et al., 2024; Wu et al., 2025b; Yang et al., 2025c). As a result, optimizing for brevity without degrading performance remains a delicate trade-off.

Controlling Thinking Length: An Open Frontier. Length control has long been a challenge in previous LLM generations, and it becomes even more critical in the context of reasoning. While

token-level constraints can be enforced during decoding, they are often too rigid and oblivious to the semantic structure of reasoning (Lee et al., 2025). The ability to “think just enough”, i.e., not too shallowly to miss key logic (Wang et al., 2025e), and not too deeply to waste computation (Chen et al., 2025c), is essential for reasoning-efficient models, but remains largely unsolved.

Beyond Transformers: Architectural Bottlenecks. Most existing LRMs still rely on the Transformer architecture, which incurs quadratic complexity with respect to input length. This design becomes especially limiting as reasoning traces expand to thousands of tokens or even more.

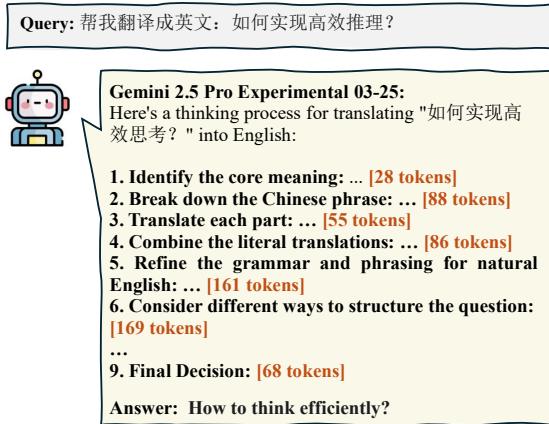


Figure 4: To translate a single Chinese sentence (“How to think efficiently?”) into English, recent LRM consumes more than 2000 tokens to reason.

Developing new architectures or efficient approximations that can reason over long contexts without sacrificing performance is a crucial and open direction. Subquadratic attention and linear sequence modeling are promising, but still in their early stages for complex reasoning tasks (Paliotta et al., 2025; Yehudai et al., 2025).

Cross-Task Generalization: One Size Doesn’t Fit All. Different tasks demand different reasoning depths. Specifically, arithmetic problems may benefit from deep logical traces, while common-sense QA might require only short chains. A single reasoning strategy or length policy often fails to generalize across such varied tasks (Sprague et al., 2024). For instance, as shown in Figure 4, the most recent LRM Gemini 2.5 Pro spends thousands of tokens to translate a short Chinese sentence into English. Ensuring efficiency while preserving robustness and adaptability across domains remains an unsolved and nuanced challenge.

3 Efficient Reasoning during Inference

The computational overhead of LRMs mainly arises from lengthy intermediate reasoning traces. Furthermore, LRMs struggle to reason effectively within a suitable computational budget given the complexity of tasks (Chen et al., 2025c; Wu et al., 2025b). To mitigate this challenge, various methods have been proposed to facilitate inference algorithms. These include constraining the reasoning budget, allocating resources across different systems or models, and incorporating parallel search strategies, all aimed at optimizing the trade-off between efficiency and accuracy, as shown in Figure 5. It is important to note that the works discussed in

this section are designed to support their proposed inference methods and may involve training models using existing techniques such as supervised fine-tuning (SFT).

3.1 Length Budgeting

A straightforward method for balancing accuracy and efficiency is explicitly budgeting computational resources during inference. Most LRMs operate within a sequential paradigm, wherein the model reflects on and refines its previously generated thoughts. In this context, managing the length of the reasoning chain serves as an intuitive means of budget allocation. In addition, another effective test-time scaling technique, parallel search (Snell et al., 2024), allows for explicit budgeting of candidate numbers to enhance efficiency, which we discuss in Section 3.4.

Previous work (Nayab et al., 2024) demonstrates the potential of LLM to adhere to length constraints specified in the prompt. Most existing works utilize this nature to control the generation length directly using specialized prompts. TALE (Han et al., 2024) uses zero-shot prompting to estimate an optimal token budget which constrains model generation. Sketch-of-Thought (Aytes et al., 2025) enhances LLM reasoning efficiency by reducing verbosity in intermediate reasoning steps with three adaptive paradigms: Conceptual Chaining, Chunked Symbolism, and Expert Lexicon. Apart from imposing an overall budget constraint on the entire reasoning process, recent works have also explored imposing budget restrictions on each reasoning step. Wang et al. (2023b) introduce a hierarchical approach to enhance language model reasoning by incorporating planning tokens at the start of each reasoning step. In contrast to verbose intermediate steps, Xu et al. (2025a) propose Chain-of-Draft, which encourages language models to generate concise, minimal intermediate reasoning steps, rather than token-heavy explanations in traditional CoT.

Instead of integrating length budgeting into prompts, Muennighoff et al. (2025) propose S1, using a budget-forcing strategy where the thinking process is forced to end by appending an end-of-thinking token delimiter, thus directly controlling thinking length. Similarly, Jiang et al. (2025) propose two decoding strategies, including Zero-Think and LessThink, to force the model to start its response without applying any thought or with a short thought process. However, forcibly budgeting length may lead to varying degrees of accuracy

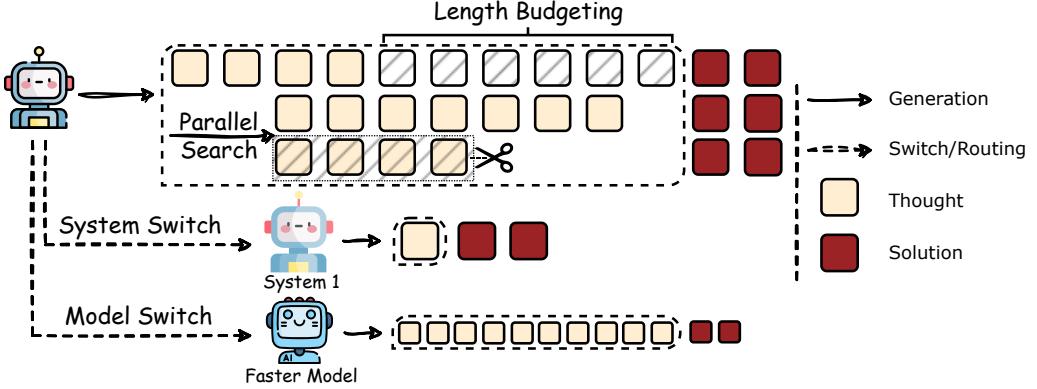


Figure 5: Illustration of efficient reasoning during inference. (1) **Length Budgeting** limits intermediate tokens to reduce overhead; (2) **System Switch** dynamically alternates between fast, intuitive and slow, deliberate reasoning; (3) **Model Switch** directs queries to optimal models based on task difficulty; (4) **Parallel Search** generates and prunes candidate outputs concurrently to cut latency.

degradation (Jin et al., 2024; Renze and Guven, 2024). To comprehensively investigate the relationship between reasoning length and model performance, Lee et al. (2025) present the first systematic study of diverse compression instructions, revealing a universal trade-off curve between response length and accuracy.

In addition to encouraging concise reasoning, researchers also seek to leverage the query difficulty to dynamically budget generation length. Wang et al. (2024b) propose Difficulty-Adaptive Self-Consistency (DSC), which evaluates the difficulty information of the queries using LLM itself, to dynamically allocate inference resources. Similarly, Dynasor (Fu et al., 2024b) allocates compute based on model certainty during multi-path reasoning, assigning more resources to hard queries. On the other hand, Wang et al. (2025e) introduce Thought Switching Penalty (TSP) to discourage premature transitions between thoughts which may cause superficial but lengthy reasoning traces.

3.2 System Switch

The dual process theory (Wason and Evans, 1974) explains that human reasoning operates through two systems: an implicit, fast, and intuitive process (System 1) and an explicit, slower, and deliberative process (System 2). Kahneman (2011) extensively explores this framework, highlighting how System 1 enables rapid decision-making but is prone to biases, while System 2 provides logical oversight for complex reasoning and analysis. Building on dual process theory, several studies (Su et al., 2024a; Saha et al., 2024; Cheng et al., 2025b; Pan et al., 2024; Sun et al., 2025a) have explored alternating between Systems 1 and 2, carefully allocating

computing resources to both systems to optimize the balance between reasoning quality and efficiency.

Dualformer (Su et al., 2024a) integrates the dual process through a randomized reasoning trace training strategy, which randomly drop certain parts of the reasoning traces. Another approach involves training a switch module to toggle between the two systems (Saha et al., 2024; Sun et al., 2025a; Cheng et al., 2025a). These switch systems can be effectively trained using supervised labels in well-defined scenarios such as Maze Navigation. Specifically, Saha et al. (2024) introduce System-1.x, which employs a controller to assess maze difficulty. This allows the model to alternate among different systems based on user-defined parameters for smoother allocation of cognitive resources when addressing sub-goals. Similarly, Sun et al. (2025a) develop a switching adapter that dynamically transitions between Systems 1 and 2 for visual reasoning according to task complexity factors like visual uncertainty and invisibility. Differently, HaluSearch (Cheng et al., 2025a) leverages the model performance on specific instances to construct supervised labels, based on which the model learn to switch from System 1 to System 2 in both instance-level and step-level under MCTS (Monte Carlo Tree Search). In addition to training-aware methods, Dyna-Think (Pan et al., 2024) uses a training-free dynamic thinking mechanism whereby the model autonomously determines “Slow” reasoning based on generation consistency and complexity of thought processes.

3.3 Model Switch

System-switch methods do not explicitly involve or necessitate corroboration among multiple differ-

ent models, *e.g.*, a large model and a small model. Allocating computational budgets across different models is also an effective strategy to mitigate acceptable performance losses in favor of enhanced efficiency. While most techniques in this domain have yet to be applied to large reasoning models, they offer promising avenues for improving efficiency with minimal performance trade-offs.

Speculative decoding (Ryu and Kim, 2024) has emerged as a key strategy to accelerate inference by leveraging draft models or early exit mechanisms to propose multiple candidate tokens before verification by the full model. BiLD (Kim et al., 2023) utilizes a small, fast model for initial predictions and a larger, more accurate model for corrections, effectively balancing speed and quality through fallback and rollback policies. EAGLE (Li et al., 2024e) enhances inference by transitioning speculative sampling from the token level to the feature level. EAGLE-2 (Li et al., 2024d) further refines speculative decoding by introducing context-aware dynamic draft trees that adjust token acceptance rates based on confidence scores. In contrast to computing allocation between large and small models, MEDUSA (Cai et al., 2024) accelerates large language model inference by incorporating additional decoding heads that predict multiple tokens simultaneously. By integrating a tree-based attention mechanism, it concurrently generates and verifies several candidate continuations, thereby reducing sequential decoding steps. LayerSkip (Elhoushi et al., 2024), on the other hand, speeds up inference through layer dropout combined with early exit loss. This allows predictions at shallower layers while also implementing self-speculative decoding for verification purposes.

Another line of work (Ong et al., 2025; Wang et al., 2025b; Lu et al., 2023) introduces a routing module to select LLMs for specific prompts given their difficulty and complexity. For example, Lu et al. (2023) propose Zooter, a reward-guided routing method that leverages distilled rewards from training queries to train a specialized routing function. This function accurately directs each query to the LLM with the most pertinent expertise. Ong et al. (2025) introduce RouteLLM, which learns to dynamically route queries between robust and weaker language models, striking an optimal balance between performance and cost effectiveness. Likewise, MixLLM (Wang et al., 2025b) enhance query embeddings using tag knowledge, employing lightweight predictors to assess quality and cost

per model while leveraging a meta decision maker to select the optimal LLM candidate.

3.4 Parallel Search

Recent large reasoning models primarily focus on enhancing the efficiency of sequential revision methods, such as o1 or R1 (DeepSeek-AI, 2025). An additional line of research focuses on enhancing the efficiency of parallel search, another commonly utilized test-time scaling paradigm (Snell et al., 2024). Typical methods include majority voting, self-consistency (Wang et al., 2023c), and Best-of-N (Lightman et al., 2024) which employs a verifier (*e.g.*, voting or a reward model) to select from multiple candidates generated in parallel by the policy model. Expanding the search space, defined as the number of candidates per prompt, consistently improves performance until it plateaus, albeit at an increased computational cost in terms of FLOPs.

To improve the efficiency of parallel search, instead of waiting for all generations to be completed, SBoN (Sun et al., 2024a) evaluates partial responses and halts those that are unlikely to yield high-quality completions, achieving comparable performance while substantially reducing computational resource demands. Qiu et al. (2024) propose TreeBoN that combines speculative tree-search with Best-of-N sampling. By generating candidate responses in a hierarchical tree structure, TreeBoN expands high-reward partial responses while pruning low-quality ones early using a weighted implicit reward. A line of work (Wang et al., 2025d; Huang et al., 2025a) proposes to replace the external reward model, which is typically the same size as the policy model, further reducing the computational overhead. STBoN (Wang et al., 2025d) truncates suboptimal candidates early via identifying the earliest estimation time when samples become distinct, and employing a buffer window along with hidden state consistency. Huang et al. (2025a) distill self-consistency-derived confidence into the model which enables strategies like early stopping and eliminates the need for external reward models. A few works combine sequential revision and parallel search to boost efficiency. Sui et al. (2025) introduce a meta-reasoner with strategies like restarting or refining, using a contextual multi-armed bandit formulation. Li et al. (2025d) propose a recursive approach to revise parallel samples, aligning model performance at test time and achieving comparable results to training-aware methods.

3.5 Summary and Outlook

In this section, we have outlined key strategies for efficient reasoning during inference, detailed in four main categories. Length budgeting methods limit verbosity by enforcing token budgets per reasoning step or for the entire process. System-switch approaches dynamically alternate between fast, intuitive (System 1) and slow, deliberative (System 2) reasoning based on task complexity. Model-switch methods, on the other hand, allocate inference resources by directing queries to different models or candidate outputs, using lightweight predictors or controllers to balance performance and cost. Parallel search strategies generate multiple candidate outputs concurrently and employ early termination or pruning to reduce latency.

Despite the effectiveness of length budgeting methods, research on effectively pruning reasoning traces, specifically precisely eliminating redundant elements, remains limited due to insufficient evaluation of reasoning chain efficacy. Moreover, the integration of inherent model features into length control has been underexplored. For example, a stricter budget may be applied when the model perceives a problem as “trivial”. Similarly, such model-aware adaptiveness can also be applied to speculative decoding, where only tough problems are left for the larger and stronger model. In addition, besides system switch and model switch, model merging (Wu et al., 2025a; Lu et al., 2024b) may be a promising direction to balance the task difficulty and reasoning efficiency. Lastly, balancing search depth with search width through parallel search presents a promising approach to significantly reduce inference latency at the expense of increased memory consumption.

4 Efficient Reasoning with SFT

Supervised fine-tuning (SFT) is a straightforward way to help models learn how to follow the instructions of users (Wang et al., 2022b; Zhang et al., 2023b). In this section, we survey existing methods that fine-tune the models to achieve efficient reasoning. As shown in Figure 6, these methods mainly consist of two categories, including training with a compressed reasoning chain and training with tokens in the latent space.

4.1 Reasoning Chain Compression

In this line, researchers first build target datasets with concise reasoning paths or compress existing

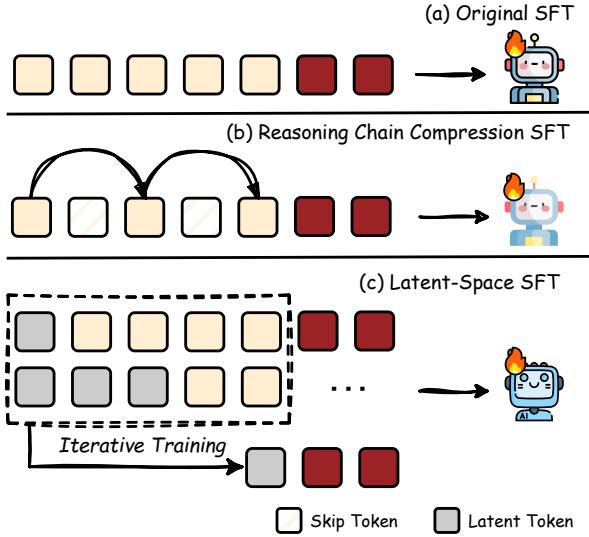


Figure 6: Illustration of efficient reasoning during SFT. (a) **Original SFT**: Standard training with sequential token generation. (b) **Reasoning Chain Compression**: Training with token skipping to simplify reasoning. (c) **Latent-Space SFT**: Iterative training using continuous hidden states for more efficient reasoning.

reasoning chains to remove redundant information, then they train the model to internalize the concise reasoning mode with supervised fine-tuning.

To generate a concise reasoning path, Token-Budget-Aware LLM Reasoning (Han et al., 2024) first produces the target output by prompting the model with a CoT prompt that includes the optimized token budget, then they train the model with SFT to produce answers that adhere to the token budget. In a different constructing way, Munkhbat et al. (2025) build concise reasoning paths via best-of-N sampling and few-shot conditioning. Then, they apply SFT to distill the length-reduction path into the model. With a more radical approach, Yu et al. (2024a) fine-tune the models to omit the intermediate step of generation for samples that are sufficiently confident.

To eliminate redundant information in the reasoning chain, C3ot (Kang et al., 2024) employs GPT-4 (Achiam et al., 2023) as a compressor, preserving key information throughout the reasoning process. The model is then fine-tuned to learn the relationship between long and short CoTs. Instead of relying on an external GPT model to filter, LMskip (Liu et al., 2024c) focuses on skipping intermediate reasoning steps. To induce step-skipping behavior, a controlled training environment is designed, instructing models to produce reasoning sequences under a step constraint. Subsequently, shorter, yet accurate, reasoning paths are chosen

and integrated with complete reasoning paths. This augmented dataset is used to finetune a new model with enhanced step-skipping capabilities. In a different way to select important steps, SPIRIT-FT (Cui et al., 2025) identifies key reasoning steps by using perplexity as a metric, a step is deemed critical if its removal significantly increases perplexity. In addition to step skipping, TokenSkip (Xia et al., 2025) analyzes token importance in CoT outputs, selectively omitting less important tokens for controllable compression of CoT sequences.

Rather than being limited to a compressed format of reasoning chain, CoT-Valve (Ma et al., 2025) fine-tunes a model to generate both long and short reasoning paths. Their approach involves identifying a specific task vector within the parameter space that governs the length of the generated CoT. Significantly, this vector allows for extrapolation, enabling the generation of reasoning chains that are either longer or shorter than those encountered during training.

4.2 Latent-Space SFT

Another line for efficient reasoning via SFT is latent space reasoning, where explicit CoT steps are gradually replaced by continuous hidden representations. Previous works define latent reasoning as the internal computations in transformers. In this context, intermediate variables in two-hop reasoning can be recovered from hidden states (Yang et al., 2024b), while “back-patching” was proposed to intervene in this process (Biran et al., 2024). Similarly, Implicit CoT eliminates explicit intermediate steps by directly predicting answers from internal representations rather than generating full token sequences (Deng et al., 2024). Moreover, studies on CoT unfaithfulness have revealed that even when a chain-of-thought is generated, the model may internally follow a different latent reasoning process (Wang et al., 2023a; Turpin et al., 2023).

Inspired by these insights, Coconut (Chain of Continuous Thought) replaces traditional CoT by using the model’s last hidden state as a continuous representation of reasoning. Instead of generating tokens step-by-step, the hidden state is fed back into the model as input for subsequent reasoning steps. This method is grounded in curriculum learning (Wang et al., 2021; Soviany et al., 2022), where the model gradually transitions from generating explicit reasoning steps to operating entirely in the latent space, leading to more efficient reasoning and reducing the token overhead typically asso-

ciated with CoT. Similarly, CCoT (Compressed Chain of Thought) (Cheng and Van Durme, 2024) fine-tunes the model to produce compressed representations of reasoning chains instead of full-length sequences. By approximating complete reasoning chains with fewer tokens, CCoT reduces computational cost and enhances throughput, while allowing dynamic adjustment of the performance-efficiency tradeoff during inference.

Building upon this, CODI (Continuous Chain-of-Thought via Self-Distillation) introduces an improvement to the curriculum learning approach by integrating a self-distillation framework. Inspired by previous works on context-based learning and generalized prompt compression (Ge et al., 2024; Li et al., 2024f), CODI focuses on aligning the hidden activations of specific tokens between a teacher model using explicit CoT and a student model using implicit CoT (Shen et al., 2025c). This alignment spans across all layers of LLM, effectively injecting explicit reasoning into the implicit reasoning process. As a result, CODI enhances performance while addressing the forgetting issue, making it a more robust solution for reasoning tasks. Token Assorted further improves reasoning efficiency by mixing latent and text tokens (Su et al., 2025). By using latent discrete tokens from VQ-VAE, it abstracts the initial reasoning steps, reducing trace length while retaining essential information. This approach results in a 17% reduction in reasoning trace length and enhanced performance on logical and mathematical tasks. SoftCoT takes a different approach to continuous-space reasoning by utilizing an assistant model that generates “soft thought tokens” for the LLM (Xu et al., 2025b). These tokens are projected into the LLM’s representation space, enabling efficient reasoning without the need for full model fine-tuning, improving efficiency while preserving pre-trained knowledge.

Furthermore, LightThinker (Zhang et al., 2025a) enhances reasoning by dynamically compressing intermediate steps into concise latent representations. This reduces memory usage and computational overhead while maintaining key reasoning information. Similarly, Heima leverages hidden latent representations to reduce verbosity in both text and multimodal tasks (Shen et al., 2025a). The Heima Encoder compresses intermediate steps into a single token, and the Heima Decoder reconstructs the reasoning process from these tokens, significantly lowering token usage and improving efficiency. These approaches demonstrate the growing

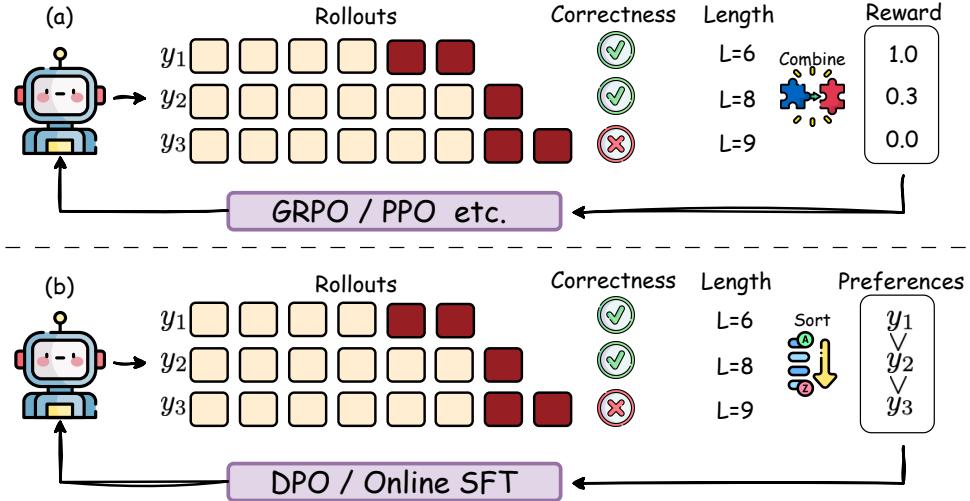


Figure 7: Illustration of efficiency training during the RL phase. Sub-Figures (a) and (b) illustrate the representative approach using length reward and not using length reward, respectively.

trend of leveraging latent space reasoning to enhance efficiency in LLMs, each offering unique strategies to reduce computational overhead while maintaining or improving reasoning capabilities, paving the way for more scalable and effective models in complex tasks.

4.3 Summary and Outlook

In this section, we have reviewed methods that enhance efficient reasoning in LLMs via SFT. Key approaches include reasoning chain compression, where models reduce the length and complexity of CoT sequences through techniques like token budget control (Han et al., 2024), self-training (Munkhbat et al., 2025), and dynamic token skipping (Xia et al., 2025). These methods optimize reasoning without losing accuracy, especially for tasks with simpler or more parallelizable structures. Another major approach is latent space reasoning, where explicit CoT steps are replaced by continuous representations in hidden states. Techniques like Coconut (Hao et al., 2024), CCoT (Cheng and Van Durme, 2024), and CODI (Shen et al., 2025c) use latent representations to improve efficiency and reduce token overhead. Innovations such as Token Assorted (Su et al., 2025), SoftCoT (Xu et al., 2025b), and Heima (Shen et al., 2025a) further enhance reasoning by mixing latent and text tokens, reducing memory usage and computational cost.

As LLMs evolve, future research on efficient reasoning may focus on refining latent-space methods and developing models that are both more flexible

and scalable. A promising direction is the integration of explicit and implicit reasoning processes, allowing models to dynamically switch between different reasoning strategies based on task complexity (Yang et al., 2024c). Additionally, exploring multi-modal latent reasoning, where models leverage both textual and visual data, could enhance their reasoning abilities (Talmor et al., 2020; Sun et al., 2024b). Research into adaptive curriculum learning strategies will further improve model flexibility, enabling them to handle more complex reasoning tasks (Kong et al., 2021). Finally, addressing CoT unfaithfulness and improving alignment between explicit reasoning paths and internal latent representations will be critical for ensuring model reliability across diverse tasks (Li et al., 2024a; Arcuschin et al., 2025).

5 Efficient Reasoning with Reinforcement Learning

DeepSeek-R1 has demonstrated that RL can effectively guide language models to develop deep reasoning abilities, marking a significant advancement in enhancing models' cognitive capabilities. This breakthrough naturally suggests that RL frameworks could potentially be a more direct and intuitive approach to improving reasoning efficiency. Inspired by this potential, several research efforts have emerged to explore the intersection of reinforcement learning and reasoning efficiency, focusing on developing more sophisticated RL-based training strategies to reduce or control the token usage of reasoning trajectories while maintaining

Name	RL Algorithm	On/Off Policy	Reward Function
O1-Pruner (2025a)	PPO	Off-policy	$\frac{L_{ref}}{L(y)} - 1 + \lambda(S(y) - S(y_{ref}))$
Efficient reasoning (2025)	PPO	On-policy	$\mathbb{I}(S(y) = 1)(1 - \alpha \cdot \sigma(\frac{L(y) - \text{Mean}(L)}{\text{STD}(L)}))$
L1 (2025)	GRPO	On-policy	$\mathbb{I}(S(y) = 1) - \alpha \cdot L(y) - L_{\text{budget}} $
Kimi-1.5 (2025)	Online Policy Mirror Descent	On-policy	$S(y) + \begin{cases} 0.5 - \frac{L(y) - L_{min}}{L_{max} - L_{min}} & \text{if } S(y) = 1 \\ \min(0, 0.5 - \frac{L(y) - L_{min}}{L_{max} - L_{min}}) & \text{if } S(y) = 0 \end{cases}$
DAST (2025b)	SimPO	Off-policy	$\max(-0.5 \cdot \frac{L(y) - L_{\text{budget}}}{L_{\text{budget}}} + 0.5, 0.1) \quad \text{if } S(y) = 1$ $\min(0.9 \cdot \frac{L(y) - L_{\text{budget}}}{L_{\text{budget}}} - 0.1, -0.1) \quad \text{if } S(y) = 0$
Demystifying (2025)	PPO	On-Policy	$\text{CosFn}(L(y), L_{max}, r_0^c, r_L^c) \quad \text{if } S(y) = 1$ $\text{CosFn}(L(y), L_{max}, r_0^w, r_L^w) \quad \text{if } S(y) = 0$ $r_e \quad \text{if } L(y) = L_{max}$

Table 2: Comparison of RL Methods with length reward for Efficient Reasoning. $S(y) \in 0, 1$ denotes the correctness of the generated answer and $L(y)$ is the length.

accuracy. Depending on the approach to modeling reasoning length, we classify current approaches that leverage RL for efficient reasoning into two distinct categories, as shown in Figure 7.

5.1 Efficient Reinforcement Learning with Length Reward

Introducing a length reward alongside the rule-based reward provides a natural approach to efficiency. As shown in Table 2, several research efforts have explored the effectiveness of this method. Some studies establish the generation token budget, by either formulating correlations between task difficulty and the generation length or specifying in prompt. DAST (Shen et al., 2025b) introduces Difficulty-Adaptive Slow-Thinking that empowers models to modulate CoT length based on problem complexity autonomously. They first define a Token Length Budget (TLB) metric, quantifying the difficulty of a problem by its success rate, then leverage length-aware reward shaping and length preference optimization to realize DAST. LCPO (Aggarwal and Welleck, 2025) controls the length budget by introducing a target length instruction in the prompt, i.e., “Think for n_{gold} tokens”, and designs a target-aware length reward that penalizes the length violation.

Other approaches incorporate the length reward normalized by a baseline budget. O1-Pruner (Luo

et al., 2025a) designs an efficient fine-tuning method that begins by estimating the LLM’s baseline performance through presampling from its reference model. Different from O1-Pruner, Arora and Zanette (2025) introduce a length penalty normalized in the *per-prompt* group. This strategy encourages the model to produce correct responses with a minimum amount of tokens while maintaining that correct responses are always preferred over incorrect ones. Kimi 1.5 technical report (Team et al., 2025) discusses an observation of the over-thinking phenomenon and introduces a length reward to restrain the rapid growth of token length. The length reward they defined is a normalized length factor compared to the maximum and minimum lengths of the different generated solutions.

Yeo et al. (2025) analyzes RL design choices for reasoning, revealing that extremely long CoT reasoning (approaching context limits) paradoxically reduces accuracy. Their proposed cosine reward function provides intuitive guidance - gradually increasing rewards for meaningful reasoning steps while penalizing excessive length. They also identify “length hacking”, where models artificially extend reasoning on difficult questions through repetition rather than genuine problem-solving, highlighting the challenge of aligning length-based rewards with actual reasoning quality.

5.2 Efficient Reinforcement Learning without Length Reward

There are also various approaches addressing this challenge without relying on explicit length rewards. These include innovative reformulations of the reinforcement learning framework, optimization methods that balance preferences between concise and detailed chain-of-thought solutions, and the identification of GRPO inductive bias.

MRT (Qu et al., 2025b) formulates the test-time optimization problem into a meta-reinforcement learning problem. It divides the generation into episodes, requiring the model to estimate answers after each episode. This approach creates a natural trade-off: the model receives rewards for both correct early predictions (exploitation) and for continuing to refine its reasoning when uncertain (exploration), effectively learning to balance speed and accuracy across different compute budgets.

Methods that balance preferences include IBPO (Yu et al., 2025b), which approaches budget awareness by framing it as utility maximization rather than directly controlling response length. Their key insight is categorizing responses into standard and extended reasoning groups, and optimizing the distribution between these groups. Chen et al. (2025c) employs heuristics such as First-Correct Solutions (FCS) and Greedy Diverse Solutions (GDS) to generate preference data for offline policy optimization using preference optimization methods (i.e., DPO, RPO, and SimPO) to control reasoning path lengths. Finally, Liu et al. (2025c) study the increasing length of reasoning trajectory from the inductive bias of RL algorithms. They claim that the token-level averaging of policy loss and standard deviation normalization in GRPO advantage introduces inductive bias to longer reasoning trajectories. To address this, they introduce Dr.GRPO to remove such bias and achieve more efficient RL scaling.

5.3 Summary and Outlook

The existing approaches commonly employ reinforcement learning techniques to optimize the trade-off between reasoning depth and token efficiency, with shared focuses on reward engineering that penalizes excessive length while preserving accuracy. Most methods formulate the challenge as a constrained optimization problem, designing specialized reward functions that balance accuracy with length penalties, though they differ in how

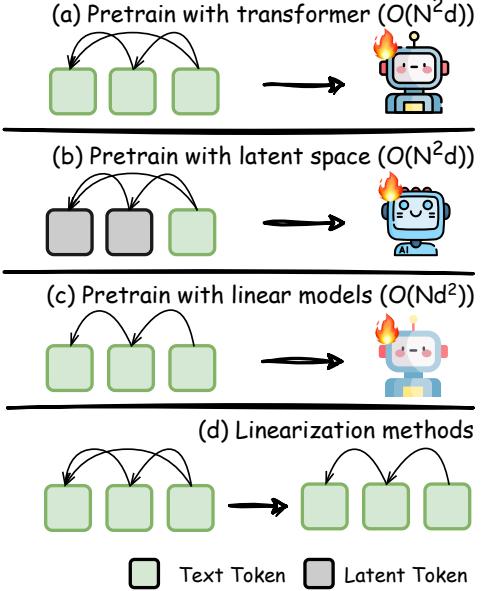


Figure 8: Illustration of efficient reasoning during pre-training: (a) Standard transformer pretraining utilizing text tokens; (b) Pretraining the transformer in latent space; (c) Employing linear models for pretraining instead of self-attention transformers; (d) Linearization methods that transform standard transformer models into linear models.

they quantify and enforce these constraints, ranging from explicit budget targets to adaptive difficulty-based adjustments.

Despite significant progress, insights into why reinforcement learning inefficiently scales with sequence length are less explored (Liu et al., 2025c), making principled solutions elusive. In addition, existing work focuses on verifiable tasks, such as reasoning and math. A promising direction is the efficiency of RL with general tasks using reward models or multi-modal tasks. Furthermore, as Qu et al. (2025b) initially explored, developing RL methods where more tokens consumption leads to better performance, represents an exciting direction. Current research primarily focuses on long CoT reasoning in o1/R1-like models, while efficient RL methods for alternative reasoning structures, such as parallel search, tree of thoughts, or graph of thoughts, remain largely unexplored.

6 Efficient Reasoning during Pretraining

In this section, we examine recent pretraining approaches designed to accelerate reasoning efficiency. These methods are proposed to enhance computational efficiency while preserving performance. As shown in Figure 8, we discuss three lines of work, including pretraining in latent space,

pretraining with linear models with subquadratic attention, and recasting transformer models into linear models with linearization methods.

6.1 Pretraining with Latent Space

Pretraining in latent space has recently garnered attention as a means of improving reasoning efficiency in LRMs. Instead of relying on traditional token-based approaches, these methods explore continuous representations, enhancing the depth of model understanding and efficiency (Hao et al., 2024; Deng et al., 2024). Several approaches have emerged that differ in their treatment of the latent space and pretraining strategies (Sun et al., 2024d).

Byte Latent Transformer (BLT) processes raw bytes using dynamically sized patches rather than fixed tokens, thereby reducing computational overhead and improving robustness against noisy and multilingual inputs (Pagnoni et al., 2024). By grouping bytes based on entropy rather than relying on predetermined tokenizations, BLT achieves better scalability and long-tail generalization (Song et al., 2025a). Large Concept Models (LCMs) extend this paradigm at a higher semantic level. In LCMs, abstract concepts that often correspond to complete sentences or speech utterances serve as the primary processing units. They utilize pre-existing sentence embedding spaces SONAR (Duquenne et al., 2023) and autoregressive sentence prediction to achieve modality- and language-agnostic performance (The et al., 2024).

CoCoMix (Continuous Concept Mixing) integrates discrete token prediction with continuous concept vectors derived from sparse autoencoders (SAEs) (Bricken et al., 2023; Cunningham et al., 2023). By interleaving these continuous representations into the hidden states during training, CoCoMix improves sample efficiency and enriches the model’s capacity for higher-order abstraction, which is beneficial for tasks such as summarization and logical reasoning (Tack et al., 2025). Additionally, latent thought vectors (LTMs) were introduced to probabilistically guide token generation via cross-attention mechanisms (Kong et al., 2025). LTMs are sampled from a latent prior and offer a flexible framework that allows adjustments of both inference steps and latent vector dimensions to optimize performance. This method improves sample efficiency and scalability compared to traditional autoregressive models.

Despite these advances, challenges remain. The reliance on implicit reasoning shortcuts may im-

pede true stepwise reasoning, particularly in complex, multi-step tasks (Lin et al., 2025). Future research may focus on refining these latent space pretraining methods to ensure consistency and accuracy across a broader range of reasoning tasks.

6.2 Subquadratic Attention

The current CoT reasoning process depends on long-context inference, where complex tasks are broken down into reasoning steps. This results in longer generation times and significant computational overhead. One potential way to enhance the efficiency of CoT reasoning is by using subquadratic attention mechanisms to replace the standard self-attention in transformers, thus reducing the computational cost of processing sequences.

Among subquadratic attention mechanisms, linear sequence modeling techniques (Sun et al., 2025c), such as linear attention, state space models (SSMs), and linear RNNs, emerge as effective alternatives to traditional self-attention. Sparse attention also presents a viable solution by selectively focusing on a subset of tokens, further improving computational efficiency. In the following sections, we will explore these subquadratic attention mechanisms in detail. Additionally, we will discuss recent linearization methods that transform pre-trained transformer-based model weights into linear recurrent structures, allowing for efficient inference while maintaining the knowledge embedded in the original transformer models.

6.2.1 Linear Sequence Modeling

Linear attention methods (Qin et al., 2024c) exploit the “right-product kernel trick”, wherein the initial computation of key-value products circumvents the quadratic cost typically incurred by query-key interactions. For example, vanilla linear attention (Katharopoulos et al., 2020) replaces traditional Softmax attention (Vaswani et al., 2017) with kernel-based approximations, thereby achieving linear computational complexity (Shen et al., 2024). Various enhancements have been developed to further boost efficiency. TransNormerLLM (Qin et al., 2023) introduces Lightning Attention (Qin et al., 2024e), which optimizes I/O operations to expedite processing (Qin et al., 2024b), while Lightning Attention-2 (Qin et al., 2024d) refines block-wise computations for superior performance in autoregressive settings. Moreover, sequence parallelism has been investigated to extend the capability of linear attention models for handling long se-

quences across extensive clusters. LASP (Sun et al., 2024c) was the first to integrate sequence parallelism into these methods, and its successor LASP-2 (Sun et al., 2025b) further refines the approach by reorganizing both computational and communication workflows. Additionally, Minimax-01 (Li et al., 2025a) adapts the Lightning Attention and LASP-series strategies to a massive MoE language model with 456 billion parameters, highlighting its potential for commercial deployment.

Other innovations in linear attention mechanisms include RetNet (Sun et al., 2023), which introduces a retention mechanism that supports parallel training without sacrificing linear-time inference. Gated Linear Attention (GLA) (Yang et al., 2023) leverages a data-independent gating scheme to enhance the sequence modeling ability and hardware efficiency, while Gated Slot Attention (GSA) (Zhang et al., 2024c) employs a bounded-memory slot control strategy to improve recall in tasks with extended contexts. Furthermore, approaches such as Test-Time Training (TTT) (Sun et al., 2024e), Titans (Behrouz et al., 2024), and Gated-DeltaNet (Yang et al., 2024e,d) propose update rules that allow models to adapt dynamically during inference. Despite their differing gating and updating strategies, these methods generally depend on a fixed-size memory state. In contrast, MoM (Du et al., 2025) expands the RNN memory state using “sparse memory” with multiple memory units managed by a router module.

State Space Models (SSMs) (Gu et al., 2022b,a; Gupta et al., 2022; Gu and Dao, 2023) represent a promising approach for efficient sequence modeling. The latest variant, Mamba2 (Dao and Gu, 2024), integrates a linear attention-like mechanism to enhance hardware efficiency, utilizing state-space duality to support parallel computation while maintaining recurrent inference capabilities. In addition to SSMs, linear RNN-based methods such as RWKV (Peng et al., 2024, 2025a), HGRN (Qin et al., 2024g), and its successor HGRN2 (Qin et al., 2024f) have also demonstrated effectiveness in improving the efficiency of sequence modeling.

6.2.2 Sparse Attention

Sparse attention presents another effective approach for managing long sequences and mitigating the quadratic complexity of self-attention mechanisms, while still maintaining model performance. Sliding window attention limits queries to a pre-determined local context. Approaches such as

StreamingLLM (Xiao et al., 2023b), MoA (Fu et al., 2024a; Qu et al., 2024b), and DuoAttention (Xiao et al., 2024) integrate local attention with attention sinks to facilitate efficient long-sequence processing. Longformer (Beltagy et al., 2020) alternates between local attention and global tokens to improve contextual understanding.

NSA (Yuan et al., 2025a) adopts a dynamic hierarchical sparse strategy, combining coarse-grained token compression with fine-grained token selection, ensuring both global context retention and local accuracy. MoBA (Lu et al., 2025a) divides the context into blocks and uses a dynamic gating mechanism to route query tokens to the most relevant KV blocks.

6.3 Linearization

The linearization of large language models converts pre-trained standard models into linear recurrent structures, enhancing deployment efficiency. Liger (Lan et al., 2025) modifies pre-trained LLMs into gated linear recurrent models by adapting key matrix weights, removing the need for extra parameters. LoLCATs (Zhang et al., 2024a) advances LLM linearization by replacing softmax attention with trained linear approximations and enhancing model quality using LoRA. Llamba (Bick et al., 2025) distills Llama-3.x models into the Mamba architecture, achieving high inference throughput and efficiency with minimal training data through MO-HAWK (Bick et al., 2024). LightTransfer (Zhang et al., 2025b) decreases KV-cache memory demands in long-context LLMs by substituting certain attention layers with streaming attention. MO-HAWK (Bick et al., 2024) facilitates the distillation of pre-trained Transformers into subquadratic models like SSMs through a structured three-phase process. Multimodal Mamba (Liao et al., 2025) builds linear-complexity multimodal state space models from existing MLLMs via progressive distillation, lowering computational costs.

6.4 Efficient Reasoning with Subquadratic Attention

Recently, several studies have explored efficient reasoning solutions with subquadratic attention models, aiming to balance computational efficiency and strong reasoning capabilities. These works investigate techniques such as model distillation, architectural modifications, and algorithmic optimizations to enhance performance in reasoning tasks while reducing computational overhead.

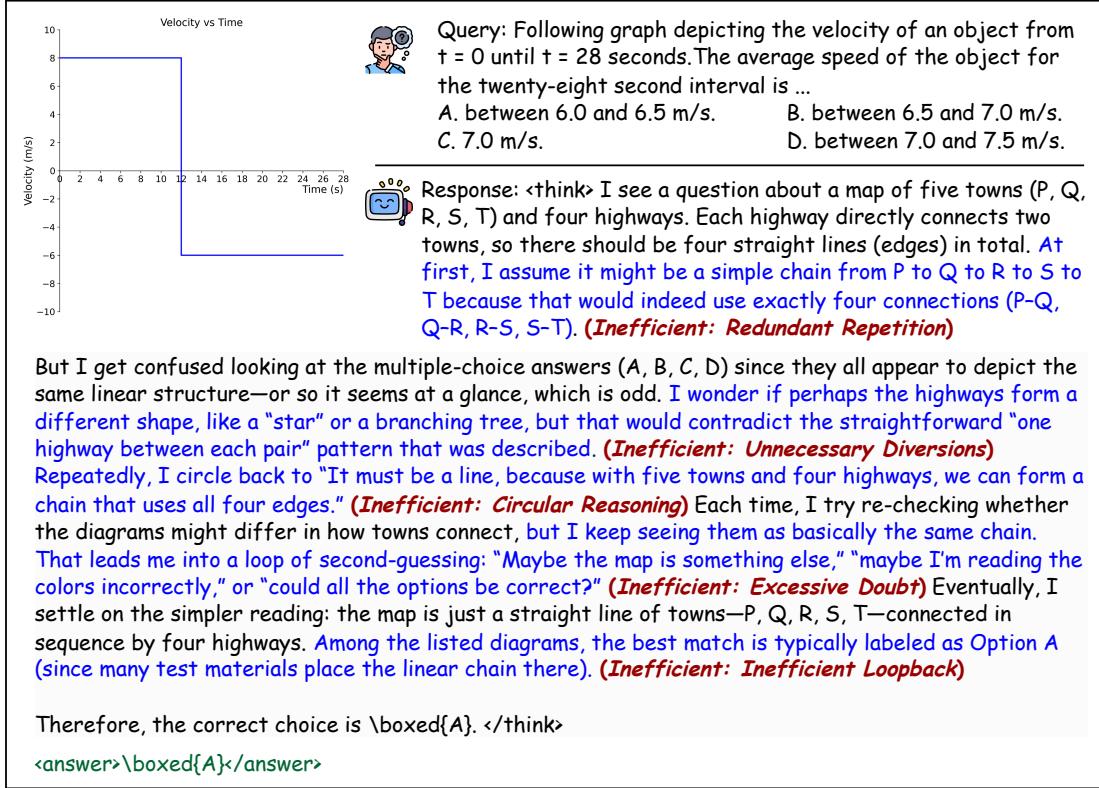


Figure 9: An example of multimodal reasoning where inefficiencies arise in the thought process. This problem involves reasoning with both a velocity-time graph and a map of towns and highways. Despite the multimodal inputs, the thought process is inefficient due to redundant repetition, unnecessary diversions, and excessive doubt, ultimately leading to a more complex and less efficient path to the correct answer.

Think Slow Fast (Paliotta et al., 2025) investigate subquadratic models for efficient reasoning under fixed compute budgets, demonstrating that distilling Mamba models from Transformers enables faster multi-path CoT generation and improved performance on MATH and GSM8K. Their findings highlight the potential of architectural innovations to enhance reasoning efficiency in resource-constrained environments. CRQs (Yehudai et al., 2025) examine the expressiveness of Transformers, RNNs, and CoT-augmented models on Compositional Reasoning Questions (CRQs), revealing inherent trade-offs: Transformers require logarithmic depth, RNNs depend on logarithmic embedding dimensions, and CoT-augmented Transformers scale with linearly many tokens. This work provides a formal framework for comparing model capabilities in multi-step reasoning tasks. Cosmos-Reason1 (Azzolini et al., 2025) proposes to use the hybrid Mamba-MLP-Transformer backbone architecture to enable efficient Physical AI reasoning (with physical common sense, embodied reasoning, and logical inference). The hybrid architecture was demonstrated to be both efficient and well-

performed on long CoT reasoning.

6.5 Summary and Outlook

Subquadratic attention mechanisms have emerged as a promising solution to improve the efficiency of CoT reasoning, which relies on long-context inference but suffers from high computational costs. Approaches such as linear sequence modeling, sparse attention aim to reduce memory and processing overhead while maintaining reasoning accuracy. Additionally, linearization methods like Liger and LoLCATs transform pretrained Transformers into efficient recurrent models without requiring extensive retraining.

Looking ahead, the integration of subquadratic attention with emerging reasoning paradigms, such as hybrid architectures and adaptive retrieval mechanisms, holds great potential for enhancing the efficiency and scalability of large-scale reasoning models. Additionally, novel techniques in large language models, including block diffusion (Arriola et al., 2025) and large language diffusion models (Arriola et al., 2025; Nie et al., 2025), merit further exploration. These advancements may in-

troduce new opportunities while also presenting unique challenges in the pursuit of more efficient and capable reasoning systems.

7 Future Directions

In this section, we offer insights into potential future research directions for efficient reasoning in the era of LRM.

7.1 Efficient Multimodal Reasoning and Video Reasoning

Recently, researchers (Meng et al., 2025; Liu et al., 2025d; Huang et al., 2025c; Pan et al., 2025; Zhou et al., 2025a; Peng et al., 2025c; Thawakar et al., 2025; Liu et al., 2025a; Yao et al., 2024a; Peng et al., 2025b) have demonstrated that o1-like CoT reasoning also plays a significant role in multimodal reasoning (Liu et al., 2023a; Liang et al., 2024). In addition, Zhao et al. (2025) further extend the reasoning ability to emotion recognition in the video. In contrast to directing output in the traditional multimodal large language models (Wang et al., 2024a; Chen et al., 2024b; Team, 2023; Zhang et al., 2023a), the augmented reasoning capabilities for multimodal large reasoning models enable a nuanced dissection of modal contributions. However, efficient reasoning is under-explored in multimodal reasoning and video reasoning. The complex image and video information contains more noisy information than language which limits efficient reasoning. Moreover, reasoning models in these domains remain susceptible to overthinking and excessive computational costs (Wang et al., 2025c). As shown in Figure 9, inefficient reasoning often leads to unnecessary complexity and convoluted paths to the correct answer.

To solve this problem, the current works all follow two key principles: 1) Different problem types may demand distinct forms of reasoning capabilities; 2) The complexity of the reasoning process should align with the inherent difficulty of the problem. Self-structured Chain of Thought (SCoT) framework (Xiang et al., 2025) has taken a significant step in addressing these issues by decomposing reasoning tasks into atomic, semantically meaningful steps. This approach ensures that the reasoning process is not only efficient but also adaptable to the complexity of the task at hand, particularly in multimodal settings (Xia et al., 2024). For instance, simpler tasks like image captioning may rely on fewer atomic steps (Stefanini et al., 2021), while

more intricate tasks, such as emotion recognition in video, may demand a deeper understanding of temporal and visual features, requiring a more sophisticated chain of reasoning (Canal et al., 2022). The Adaptive-Length Chain-of-Thought Distillation (AL-CoTD) framework (Peng et al., 2025b) further refines this process by dynamically adjusting the length of reasoning chains according to task complexity. This addresses the issue of overthinking, which is particularly prevalent in multimodal and video reasoning tasks.

Traditional multimodal large language models often treat reasoning as a uniform process, regardless of the input type. In contrast, integrating dynamic reasoning strategies across different modalities (e.g., textual, image, and video data) allows for more efficient problem-solving by adjusting the cognitive complexity to match the nature of the inputs. This approach promises to handle the noise and complexity inherent in different modalities of data while maintaining computational efficiency. By balancing expressiveness with efficiency, future multimodal reasoning models can adapt to various tasks, using simpler reasoning for less complex problems and more sophisticated methods for challenging ones, ultimately driving more powerful and efficient multimodal solutions (Liang et al., 2024; Meng et al., 2025; Liu et al., 2025d).

7.2 Efficient Test-time Scaling and Infinity Thinking

Test-time scaling is a direct method for extending thinking time and improving response quality and model performance in LRM and LLMs (Snell et al., 2024; Bi et al., 2024; Liu et al., 2025b), and is typically classified as parallel sampling and sequential revision. For parallel sampling, the scaling can be conducted by extending the search width, such as Best-of-N sampling (Cobbe et al., 2021b), Self-Consistency (Wang et al., 2022a), and minimum Bayes risk decoding (Wu et al., 2024; Heineman et al., 2024). However, these approaches necessitate a fixed number of samples per query, irrespective of complexity. This leads to computational waste on simpler queries and potentially insufficient exploration for complex ones. This kind of inefficient reasoning can be mitigated by developing confidence-based methods to address queries of varying difficulty (Huang et al., 2025a) or adaptive sampling strategies (Wan et al., 2024; Wang et al., 2025d; Li et al., 2024b).

For sequential revision, scalability can be

achieved by extending the search depth for reasoning, incorporating methods such as debating (Liang et al., 2023; Du et al., 2023), self-correction (Lin et al., 2024b), and self-critique (Yu et al., 2024b; Su et al., 2024d). These techniques allow the reasoning process to expand into increasingly longer sequences. In extreme cases, this reasoning process may even become infinite. Yan et al. (2025) have addressed this emerging challenge for large retrieval models (LRMs) by transforming reasoning into an iterative process with intermediate summarizations. This approach interleaves short reasoning segments with concise progress summaries, allowing the depth of reasoning to be significantly extended while maintaining manageable computational costs. Similarly, Yang et al. (2025a) propose a reduction rule that reduces context length during the standard iterative next-token generation, applying it whenever feasible to optimize the process.

However, extending both search width and depth during inference presents significant challenges for efficient reasoning. Instead of relying solely on sequential revisions like o1 and R1, shifting the balance from depth to width can considerably reduce inference latency. This approach enables the simultaneous exploration of multiple reasoning traces, offering more diverse pathways for problem-solving. Despite this advantage, managing several ultra-long reasoning traces introduces substantial computational overhead. As a result, this strategy generates numerous lengthy sampled responses, which require significant resources to process efficiently. Addressing these challenges presents a promising direction for future research, aiming to improve the scalability and resource management of complex reasoning systems.

7.3 Efficient and Trustworthy Reasoning

Emerging large reasoning models, *e.g.*, OpenAI o1 and DeepSeek-R1, generate long and structured CoT steps, resulting in astonishing performance (Su et al., 2024b; Li et al., 2025b). However, such long CoT steps bring new challenges to the trustworthiness of LRMs, including safety and reliability (Liu et al., 2023b; Yao et al., 2024b).

For the safety concern (Ren et al., 2024; Qian et al., 2024a; Hu et al., 2024b), researchers (Zhou et al., 2025b; Jiang et al., 2025) discover that the safe rate of the LRMs’ thinking process is lower than that of the final answers. When presented with harmful queries, LRMs engage in reasoning and generate relevant content, potentially exposing

sensitive information even if the final response is ultimately safe. To mitigate this, deliberative alignment (Guan et al., 2024) and X-Boundary (Lu et al., 2025b) offer distinct approaches to enhance LRM safety. Exploring efficient and inherently safe reasoning mechanisms presents a compelling future direction, echoing the adage “he that talks much errs much”. For instance, latent space reasoning (Deng et al., 2024; Su et al., 2024a; Sui et al., 2025; Kong et al., 2025) offers efficiency, while representation engineering (Zou et al., 2023, 2024; Qian et al., 2024b; Chen et al., 2025a; Liu et al., 2024b) can ensure robust safety performance.

In terms of reliability, LLMs are known to suffer from both factuality and faithfulness hallucinations (Huang et al., 2025b) across various applications. This problem is potentially amplified in LRMs, whose extended reasoning chains are inherently more susceptible to the accumulation of noisy and untrustworthy information. The increased complexity and length of these reasoning processes provide greater opportunity for errors and deviations from ground truth to accumulate. Furthermore, CoTs do not accurately reflect the model thinking process and bring additional uncertainty (Lanham et al., 2023; Su et al., 2023; Turpin et al., 2023; Tanneru et al., 2024; Agarwal et al., 2024; Su et al., 2024c). These issues create a compounding effect, worsening the hallucination problem in both language-based and multi-modal contexts (Qu et al., 2024c). Therefore, the development of efficient and reliable reasoning methodologies represents a particularly pressing research need in the age of LRMs.

7.4 Building Efficient Reasoning Applications

In this section, we explore applications where efficient reasoning can provide significant benefits, focusing on Retrieval-augmented Generation (RAG), agent-based systems, and tool learning.

RAG (Gao et al., 2023; Zhao et al., 2024b; Sun et al., 2024b; Qu et al., 2024a) offers a straightforward and effective approach to addressing the inherent limitations of static parameters in generative models by retrieving content from external knowledge bases (Su et al., 2024c). Recently, agentic RAG systems have empowered models to autonomously determine when and what knowledge to retrieve, demonstrating enhanced planning and problem-solving capabilities (Chen et al., 2024a; Li et al., 2025c). Li et al. (2025c) further combine large retrieval models (LRMs) with an agentic RAG mechanism, incorporating a “Reason-in-

Documents” module to refine retrieved content. They enable dynamic retrieval of external knowledge when LRM_s face uncertainty. Additionally, Wang et al. (2025a) introduce a method for training O1-like RAG models that perform step-by-step retrieval and reasoning over relevant information before generating the final answer. During inference, they modulate the model’s computational cost by adjusting the length and number of sampled retrieval chains. This integration of retrieved content enhances the depth and breadth of LRM_s’ reasoning, highlighting the importance of efficient reasoning in RAG systems.

In the agent scenarios, Zhou et al. (2025c) discover that LRM_s outperform LLMs in reasoning-intensive tasks, such as Plan Design, by leveraging iterative reflection to achieve superior results. However, their reliance on extensive reasoning often incurs significant computational overhead, thus reducing efficiency in time-critical contexts (Li et al., 2024c; Zhang et al., 2024b). Cuadron et al. (2025) further study the overthinking phenomenon in the magnetic tasks, and elevate overthinking scores correlate negatively with performance. They further propose to select solutions with lower overthinking scores to mitigate the overthinking and achieve superior results. Thus, achieving efficient reasoning for agents is a promising avenue for research.

In addition to the aforementioned challenges, enhancing tool efficiency in LRM_s necessitates a multifaceted approach. One promising direction is the incorporation of hierarchical reasoning and early exit strategies that dynamically terminate computations once sufficient confidence is achieved, thereby reducing unnecessary function calls (Qin et al., 2024a). Furthermore, parallel execution schemes, which can be accelerated by specialized hardware such as GPUs or FPGAs, further mitigate latency (Qu et al., 2025a). Another avenue involves dynamic query routing, where the system adjusts the complexity of its planning and reasoning processes based on current task demands and resource availability (Gao et al., 2024). In summary, these integrated strategies form a coherent framework that optimizes both response speed and performance, paving the way for the efficient and practical deployment of LRM_s in real-world applications (Hadi et al., 2023).

Considering the nascent state of efficient reasoning research in LRM_s, numerous other compelling research directions warrant further exploration, such as efficient reasoning for coding (Yang

et al., 2025b; Jiang et al., 2024), autonomous driving (Yurtsever et al., 2020; Zhao et al., 2024a), health care (Temsah et al., 2024, 2025), and embodied AI (Duan et al., 2022; Liu et al., 2024d).

7.5 Evaluation and Benchmark

Currently, most studies (Aggarwal and Welleck, 2025; Xia et al., 2025) evaluate the efficiency of Large Language Models (LRM_s) on complex math problems, such as GSM8K (Cobbe et al., 2021a), MATH (Hendrycks et al., 2021), ASDIV (Miao et al., 2021), and AIME. These works primarily focus on comparing the trade-offs between accuracy and token consumption across different methods.

To investigate the phenomenon of overthinking in math problems, where multiple solutions are generated for a single question, Chen et al. (2025c) analyze each solution and propose two efficiency metrics that address both outcome and process perspectives. These metrics are designed to assess the computational resource efficiency of o1-like LRM_s. Specifically, the outcome efficiency metric evaluates how much subsequent solutions improve accuracy beyond the first solution. In contrast, the process metric assesses how much subsequent solutions contribute to the diversity of solutions, independent of correctness (Zeng et al., 2024b; Song et al., 2025b). Future research could delve deeper into the analysis of the immediate solutions generated, particularly by examining how each initial response contributes to the overall efficiency of the reasoning process (Bi et al., 2024).

In addition to straightforward math problems, it is crucial to explore whether efficient reasoning compromises the level of intelligence. Therefore, it is essential to evaluate efficient reasoning in more general domains, such as creativity and innovation (Lu et al., 2024a; Ruan et al., 2024; Franceschelli and Musolesi, 2024). Towards the evaluation of efficient reasoning of LRM_s, Hashemi et al. (2025) introduce DNA Bench, a benchmark designed to expose a vulnerability in LRM_s’ tendency for overreasoning. The prompts in this benchmark are carefully crafted to mislead LRM_s into generating excessively verbose reasoning chains, ultimately leading to incorrect responses. This highlights the need for more nuanced approaches to measure and improve reasoning efficiency in LRM_s. Looking ahead, designing more comprehensive and diverse benchmarks for efficient reasoning presents a promising avenue for future research.

8 Conclusion

In this paper, we offer a comprehensive review of efficient reasoning in the era of Large Reasoning Models. We provide a definition of reasoning efficiency and present the pattern of reasoning inefficiency. Notably, we highlight the unique challenges for efficient reasoning. Then, we delve into the existing methods aiming for efficient reasoning from the perspective of inference, SFT, RL, and pre-training. Finally, we propose important future directions that may benefit from efficient reasoning. We believe this is an emerging and important topic for LRMs. We hope this survey can serve as a comprehensive entry point, equipping readers with the foundational knowledge to navigate this challenging field.

References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Chirag Agarwal, Sree Harsha Tanneru, and Himabindu Lakkaraju. 2024. Faithfulness vs. plausibility: On the (un) reliability of explanations from large language models. *arXiv preprint arXiv:2402.04614*.
- Pranjal Aggarwal and Sean Welleck. 2025. L1: Controlling how long a reasoning model thinks with reinforcement learning. *Preprint*, arXiv:2503.04697.
- Janice Ahn, Rishu Verma, Renze Lou, Di Liu, Rui Zhang, and Wenpeng Yin. 2024. Large language models for mathematical reasoning: Progresses and challenges. *arXiv preprint arXiv:2402.00157*.
- Iván Arcuschin, Jett Janiak, Robert Krzyzanowski, Senthooran Rajamanoharan, Neel Nanda, and Arthur Conmy. 2025. Chain-of-thought reasoning in the wild is not always faithful. *arXiv preprint arXiv:2503.08679*.
- Daman Arora and Andrea Zanette. 2025. Training language models to reason efficiently. *Preprint*, arXiv:2502.04463.
- Marianne Arriola, Aaron Gokaslan, Justin T Chiu, Zhihan Yang, Zhixuan Qi, Jiaqi Han, Subham Sekhar Sahoo, and Volodymyr Kuleshov. 2025. Block diffusion: Interpolating between autoregressive and diffusion language models. *arXiv preprint arXiv:2503.09573*.
- Simon A. Aytes, Jinheon Baek, and Sung Ju Hwang. 2025. Sketch-of-thought: Efficient llm reasoning with adaptive cognitive-inspired sketching. *Preprint*, arXiv:2503.05179.
- Alisson Azzolini, Hannah Brandon, Prithvijit Chat-topadhyay, Huayu Chen, Jinju Chu, Yin Cui, Jenna Diamond, Yifan Ding, Francesco Ferroni, Rama Govindaraju, et al. 2025. Cosmos-reason1: From physical common sense to embodied reasoning. *arXiv preprint arXiv:2503.15558*.
- Ali Behrouz, Peilin Zhong, and Vahab Mirrokni. 2024. Titans: Learning to memorize at test time. *arXiv preprint arXiv:2501.00663*.
- Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *Preprint*, arXiv:2004.05150.
- Maciej Besta, Julia Barth, Eric Schreiber, Ales Kubicek, Afonso Catarino, Robert Gerstenberger, Piotr Nyczyk, Patrick Iff, Yueling Li, Sam Houlston, et al. 2025. Reasoning language models: A blueprint. *arXiv preprint arXiv:2501.11223*.
- Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, et al. 2024. Graph of thoughts: Solving elaborate problems with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17682–17690.
- Zhenni Bi, Kai Han, Chuanjian Liu, Yehui Tang, and Yunhe Wang. 2024. Forest-of-thought: Scaling test-time compute for enhancing llm reasoning. *arXiv preprint arXiv:2412.09078*.
- Aviv Bick, Tobias Katsch, Nimit Sohoni, Arjun Desai, and Albert Gu. 2025. Llamba: Scaling distilled recurrent models for efficient language processing. *arXiv preprint arXiv:2502.14458*.
- Aviv Bick, Kevin Li, Eric Xing, J Zico Kolter, and Albert Gu. 2024. Transformers to ssms: Distilling quadratic knowledge to subquadratic models. *Advances in Neural Information Processing Systems*, 37:31788–31812.
- Eden Biran, Daniela Gottesman, Sohee Yang, Mor Geva, and Amir Globerson. 2024. Hopping too late: Exploring the limitations of large language models on multi-hop queries. *Preprint*, arXiv:2406.12775.
- Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. 2023. Towards monosematicity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*. [Https://transformercircuits.pub/2023/monosemantic-features/index.html](https://transformercircuits.pub/2023/monosemantic-features/index.html).
- Tianle Cai, Yuhong Li, Zhengyang Geng, Hongwu Peng, Jason D. Lee, De huai Chen, and Tri Dao.

2024. **Medusa: Simple llm inference acceleration framework with multiple decoding heads.** *ArXiv*, abs/2401.10774.
- Felipe Zago Canal, Tobias Rossi Müller, Jennifer Cristine Matias, Gustavo Gino Scotton, Antonio Reis de Sa Junior, Eliane Pozzebon, and Antonio Carlos Sobieranski. 2022. A survey on facial emotion recognition techniques: A state-of-the-art literature review. *Information Sciences*, 582:593–617.
- Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2024. A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology*, 15(3):1–45.
- Guanxu Chen, Dongrui Liu, Tao Luo, and Jing Shao. 2025a. Seer: Self-explainability enhancement of large language models’ representations. *arXiv preprint arXiv:2502.05242*.
- Qiguang Chen, Libo Qin, Jinhao Liu, Dengyun Peng, Jiannan Guan, Peng Wang, Mengkang Hu, Yuhang Zhou, Te Gao, and Wangxiang Che. 2025b. Towards reasoning era: A survey of long chain-of-thought for reasoning large language models. *Preprint*, arXiv:2503.09567.
- Xingyu Chen, Jiahao Xu, Tian Liang, Zhiwei He, Jianhui Pang, Dian Yu, Linfeng Song, Qiuwei Liu, Mengfei Zhou, Zhuosheng Zhang, Rui Wang, Zhaopeng Tu, Haitao Mi, and Dong Yu. 2025c. Do not think that much for $2+3=?$ on the overthinking of o1-like llms. *Preprint*, arXiv:2412.21187.
- Zehui Chen, Kuikun Liu, Qiuchen Wang, Jiangning Liu, Wenwei Zhang, Kai Chen, and Feng Zhao. 2024a. Mindsearch: Mimicking human minds elicits deep ai searcher. *arXiv preprint arXiv:2407.20183*.
- Zhe Chen, Jiannan Wu, Wenhui Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. 2024b. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24185–24198.
- Hongrong Cheng, Miao Zhang, and Javen Qinfeng Shi. 2024. A survey on deep neural network pruning: Taxonomy, comparison, analysis, and recommendations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- Jeffrey Cheng and Benjamin Van Durme. 2024. Compressed chain of thought: Efficient reasoning through dense representations. *arXiv preprint arXiv:2412.13171*.
- Xiaoxue Cheng, Junyi Li, Wayne Xin Zhao, and Ji-Rong Wen. 2025a. Think more, hallucinate less: Mitigating hallucinations via dual process of fast and slow thinking. *Preprint*, arXiv:2501.01306.
- Xiaoxue Cheng, Junyi Li, Wayne Xin Zhao, and Ji-Rong Wen. 2025b. Think more, hallucinate less: Mitigating hallucinations via dual process of fast and slow thinking. *arXiv preprint arXiv:2501.01306*.
- Cheng-Han Chiang and Hung-yi Lee. 2024. Overreasoning and redundant calculation of large language models. *arXiv preprint arXiv:2401.11467*.
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reichiro Nakano, Christopher Hesse, and John Schulman. 2021a. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reichiro Nakano, et al. 2021b. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.
- Alejandro Cuadron, Dacheng Li, Wenjie Ma, Xingyao Wang, Yichuan Wang, Siyuan Zhuang, Shu Liu, Luis Gaspar Schroeder, Tian Xia, Huanzhi Mao, Nicholas Thumiger, Aditya Desai, Ion Stoica, Ana Klimovic, Graham Neubig, and Joseph E. Gonzalez. 2025. The danger of overthinking: Examining the reasoning-action dilemma in agentic tasks. *Preprint*, arXiv:2502.08235.
- Yingqian Cui, Pengfei He, Jingying Zeng, Hui Liu, Xianfeng Tang, Zhenwei Dai, Yan Han, Chen Luo, Jing Huang, Zhen Li, et al. 2025. Stepwise perplexity-guided refinement for efficient chain-of-thought reasoning in large language models. *arXiv preprint arXiv:2502.13260*.
- Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. 2023. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*.
- Tri Dao and Albert Gu. 2024. Transformers are ssms: Generalized models and efficient algorithms through structured state space duality. *arXiv preprint arXiv:2405.21060*.
- DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.
- Yuntian Deng, Yejin Choi, and Stuart Shieber. 2024. From explicit cot to implicit cot: Learning to internalize cot step by step. *arXiv preprint arXiv:2405.14838*.
- Jusen Du, Weigao Sun, Disen Lan, Jiaxi Hu, and Yu Cheng. 2025. Mom: Linear sequence modeling with mixture-of-memories. *Preprint*, arXiv:2502.13685.
- Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multiagent debate. In *Forty-first International Conference on Machine Learning*.

- Jiafei Duan, Samson Yu, Hui Li Tan, Hongyuan Zhu, and Cheston Tan. 2022. A survey of embodied ai: From simulators to research tasks. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 6(2):230–244.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Paul-Ambroise Duquenne, Holger Schwenk, and Benoît Sagot. 2023. Sonar: sentence-level multimodal and language-agnostic representations. *arXiv preprint arXiv:2308.11466*.
- Mostafa Elhoushi, Akshat Shrivastava, Diana Liskovich, Basil Hosmer, Bram Wasti, Liangzhen Lai, Anas Mahmoud, Bilge Acun, Saurabh Agarwal, Ahmed Roman, Ahmed Aly, Beidi Chen, and Carole-Jean Wu. 2024. LayerSkip: Enabling early exit inference and self-speculative decoding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12622–12642, Bangkok, Thailand. Association for Computational Linguistics.
- Jonathan St BT Evans. 2003. In two minds: dual-process accounts of reasoning. *Trends in cognitive sciences*, 7(10):454–459.
- Giorgio Franceschelli and Mirco Musolesi. 2024. On the creativity of large language models. *AI SOCIETY*, pages 1–11.
- Keith Frankish. 2010. Dual-process and dual-system theories of reasoning. *Philosophy Compass*, 5(10):914–926.
- Tianyu Fu, Haofeng Huang, Xuefei Ning, Genghan Zhang, Boju Chen, Tianqi Wu, Hongyi Wang, Zixiao Huang, Shiyao Li, Shengen Yan, et al. 2024a. Moa: Mixture of sparse attention for automatic large language model compression. *arXiv preprint arXiv:2406.14909*.
- Yichao Fu, Junda Chen, Siqi Zhu, Zheyu Fu, Zhongdongming Dai, Aurick Qiao, and Hao Zhang. 2024b. Efficiently serving llm reasoning programs with certindex. *Preprint, arXiv:2412.20993*.
- Silin Gao, Jane Dwivedi-Yu, Ping Yu, Xiaoqing Ellen Tan, Ramakanth Pasunuru, Olga Golovneva, Koustuv Sinha, Asli Celikyilmaz, Antoine Bosselut, and Tianlu Wang. 2024. Efficient tool use with chain-of-abstraction reasoning. *arXiv preprint arXiv:2401.17464*.
- Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Haofen Wang, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2.
- Tao Ge, Hu Jing, Lei Wang, Xun Wang, Si-Qing Chen, and Furu Wei. 2024. In-context autoencoder for context compression in a large language model. In *The Twelfth International Conference on Learning Representations*.
- Albert Gu and Tri Dao. 2023. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*.
- Albert Gu, Karan Goel, Ankit Gupta, and Christopher Ré. 2022a. On the parameterization and initialization of diagonal state space models. *Advances in Neural Information Processing Systems*, 35:35971–35983.
- Albert Gu, Karan Goel, and Christopher Ré. 2022b. Efficiently modeling long sequences with structured state spaces. In *The International Conference on Learning Representations (ICLR)*.
- Melody Y Guan, Manas Joglekar, Eric Wallace, Saachi Jain, Boaz Barak, Alec Heylar, Rachel Dias, Andrea Vallone, Hongyu Ren, Jason Wei, et al. 2024. Deliberative alignment: Reasoning enables safer language models. *arXiv preprint arXiv:2412.16339*.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.
- Ankit Gupta, Albert Gu, and Jonathan Berant. 2022. Diagonal state spaces are as effective as structured state spaces. *Advances in Neural Information Processing Systems*, 35:22982–22994.
- Muhammad Usman Hadi, Rizwan Qureshi, Abbas Shah, Muhammad Irfan, Anas Zafar, Muhammad Bilal Shaikh, Naveed Akhtar, Jia Wu, Seyedali Mirjalili, et al. 2023. A survey on large language models: Applications, challenges, limitations, and practical usage. *Authorea Preprints*, 3.
- Tingxu Han, Chunrong Fang, Shiyu Zhao, Shiqing Ma, Zhenyu Chen, and Zhenting Wang. 2024. Token-budget-aware llm reasoning. *arXiv preprint arXiv:2412.18547*.
- Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason Weston, and Yuandong Tian. 2024. Training large language models to reason in a continuous latent space. *arXiv preprint arXiv:2412.06769*.
- Masoud Hashemi, Oluwanifemi Bambose, Sathwik Tejaswi Madhusudhan, Jishnu Sethumadhavan Nair, Aman Tiwari, and Vikas Yadav. 2025. Dna bench: When silence is smarter – benchmarking over-reasoning in reasoning llms. *Preprint, arXiv:2503.15793*.
- David Heineman, Yao Dou, and Wei Xu. 2024. Improving minimum bayes risk decoding with multi-prompt. *arXiv preprint arXiv:2407.15343*.

- Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *NeurIPS*.
- Timothy Hospedales, Antreas Antoniou, Paul Micaelli, and Amos Storkey. 2020. [Meta-learning in neural networks: A survey](#). *Preprint*, arXiv:2004.05439.
- Cunchen Hu, Heyang Huang, Liangliang Xu, Xusheng Chen, Jiang Xu, Shuang Chen, Hao Feng, Chenxi Wang, Sa Wang, Yungang Bao, et al. 2024a. Inference without interference: Disaggregate llm inference for mixed downstream workloads. *arXiv preprint arXiv:2401.11181*.
- Xuhao Hu, Dongrui Liu, Hao Li, Xuanjing Huang, and Jing Shao. 2024b. Vlsbench: Unveiling visual leakage in multimodal safety. *arXiv preprint arXiv:2411.19939*.
- Chengsong Huang, Langlin Huang, Jixuan Leng, Jiacheng Liu, and Jiaxin Huang. 2025a. [Efficient test-time scaling via self-calibration](#). *Preprint*, arXiv:2503.00031.
- Jie Huang and Kevin Chen-Chuan Chang. 2022. Towards reasoning in large language models: A survey. *arXiv preprint arXiv:2212.10403*.
- Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2025b. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.
- Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Yao Hu, and Shaohui Lin. 2025c. Vision-r1: Incentivizing reasoning capability in multimodal large language models. *arXiv preprint arXiv:2503.06749*.
- Xu Huang, Weiwen Liu, Xiaolong Chen, Xingmei Wang, Hao Wang, Defu Lian, Yasheng Wang, Ruiming Tang, and Enhong Chen. 2024. Understanding the planning of llm agents: A survey. *arXiv preprint arXiv:2402.02716*.
- Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.
- Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.
- Yixin Ji, Juntao Li, Hai Ye, Kaixin Wu, Jia Xu, Linjian Mo, and Min Zhang. 2025. Test-time computing: from system-1 thinking to system-2 thinking. *arXiv preprint arXiv:2501.02497*.
- Fengqing Jiang, Zhangchen Xu, Yuetai Li, Luyao Niu, Zhen Xiang, Bo Li, Bill Yuchen Lin, and Radha Poovendran. 2025. Safechain: Safety of language models with long chain-of-thought reasoning capabilities. *arXiv preprint arXiv:2502.12025*.
- Juyong Jiang, Fan Wang, Jiasi Shen, Sungju Kim, and Sunghun Kim. 2024. A survey on large language models for code generation. *arXiv preprint arXiv:2406.00515*.
- 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. *arXiv preprint arXiv:2401.04925*.
- Daniel Kahneman. 2011. *Thinking, fast and slow*. macmillan.
- Yu Kang, Xianghui Sun, Liangyu Chen, and Wei Zou. 2024. C3ot: Generating shorter chain-of-thought without compromising effectiveness. *arXiv preprint arXiv:2412.11664*.
- Udo Kannengiesser and John S Gero. 2019. Design thinking, fast and slow: A framework for kahneman’s dual-system theory in design. *Design Science*, 5:e10.
- Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, et al. 2023. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and individual differences*, 103:102274.
- Angelos Katharopoulos, Apoorv Vyas, Nikolaos Papas, and François Fleuret. 2020. Transformers are rnns: Fast autoregressive transformers with linear attention. In *International Conference on Machine Learning*, pages 5156–5165. PMLR.
- Sehoon Kim, Karttikeya Mangalam, Suhong Moon, Jitendra Malik, Michael W. Mahoney, Amir Gholami, and Kurt Keutzer. 2023. [Speculative decoding with big little decoder](#). In *Neural Information Processing Systems*.
- Deqian Kong, Minglu Zhao, Dehong Xu, Bo Pang, Shu Wang, Edouardo Honig, Zhangzhang Si, Chuan Li, Jianwen Xie, Sirui Xie, et al. 2025. Scalable language models with posterior inference of latent thought vectors. *arXiv preprint arXiv:2502.01567*.
- Yajing Kong, Liu Liu, Jun Wang, and Dacheng Tao. 2021. Adaptive curriculum learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5067–5076.
- Disen Lan, Weigao Sun, Jiaxi Hu, Jusen Du, and Yu Cheng. 2025. Liger: Linearizing large language models to gated recurrent structures. *arXiv preprint arXiv:2503.01496*.

- Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, et al. 2023. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*.
- Ayeong Lee, Ethan Che, and Tianyi Peng. 2025. How well do llms compress their own chain-of-thought? a token complexity approach. *Preprint, arXiv:2503.01141*.
- Aonian Li, Bangwei Gong, Bo Yang, Boji Shan, Chang Liu, Cheng Zhu, Chunhao Zhang, Congchao Guo, Da Chen, Dong Li, et al. 2025a. Minimax-01: Scaling foundation models with lightning attention. *arXiv preprint arXiv:2501.08313*.
- Dacheng Li, Shiyi Cao, Tyler Griggs, Shu Liu, Xiangxi Mo, Shishir G Patil, Matei Zaharia, Joseph E Gonzalez, and Ion Stoica. 2025b. Llms can easily learn to reason from demonstrations structure, not content, is what matters! *arXiv preprint arXiv:2502.07374*.
- Jiachun Li, Pengfei Cao, Yubo Chen, Kang Liu, and Jun Zhao. 2024a. Towards faithful chain-of-thought: Large language models are bridging reasoners. *arXiv preprint arXiv:2405.18915*.
- Xiaoxi Li, Guanting Dong, Jiajie Jin, Yuyao Zhang, Yujia Zhou, Yutao Zhu, Peitian Zhang, and Zhicheng Dou. 2025c. Search-o1: Agentic search-enhanced large reasoning models. *arXiv preprint arXiv:2501.05366*.
- Yafu Li, Xuyang Hu, Xiaoye Qu, Linjie Li, and Yu Cheng. 2025d. Test-time preference optimization: On-the-fly alignment via iterative textual feedback. *ArXiv, abs/2501.12895*.
- Yiwei Li, Peiwen Yuan, Shaoxiong Feng, Boyuan Pan, Xinglin Wang, Bin Sun, Heda Wang, and Kan Li. 2024b. Escape sky-high cost: Early-stopping self-consistency for multi-step reasoning. *arXiv preprint arXiv:2401.10480*.
- Yuanchun Li, Hao Wen, Weijun Wang, Xiangyu Li, Yizhen Yuan, Guohong Liu, Jiacheng Liu, Wenxing Xu, Xiang Wang, Yi Sun, et al. 2024c. Personal llm agents: Insights and survey about the capability, efficiency and security. *arXiv preprint arXiv:2401.05459*.
- Yuhui Li, Fangyun Wei, Chao Zhang, and Hongyang Zhang. 2024d. EAGLE-2: Faster inference of language models with dynamic draft trees. In *Empirical Methods in Natural Language Processing*.
- Yuhui Li, Fangyun Wei, Chao Zhang, and Hongyang Zhang. 2024e. EAGLE: Speculative sampling requires rethinking feature uncertainty. In *International Conference on Machine Learning*.
- Zhong-Zhi Li, Duzhen Zhang, Ming-Liang Zhang, Jiaxin Zhang, Zengyan Liu, Yuxuan Yao, Haotian Xu, Junhao Zheng, Pei-Jie Wang, Xiuyi Chen, Yingying Zhang, Fei Yin, Jiahua Dong, Zhijiang Guo, Le Song, and Cheng-Lin Liu. 2025e. From system 1 to system 2: A survey of reasoning large language models. *Preprint, arXiv:2502.17419*.
- Zongqian Li, Yixuan Su, and Nigel Collier. 2024f. 500xcompressor: Generalized prompt compression for large language models. *Preprint, arXiv:2408.03094*.
- Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 2023. Encouraging divergent thinking in large language models through multi-agent debate. *arXiv preprint arXiv:2305.19118*.
- Zijing Liang, Yanjie Xu, Yifan Hong, Penghui Shang, Qi Wang, Qiang Fu, and Ke Liu. 2024. A survey of multimodal large language models. In *Proceedings of the 3rd International Conference on Computer, Artificial Intelligence and Control Engineering*, pages 405–409.
- Bencheng Liao, Hongyuan Tao, Qian Zhang, Tianheng Cheng, Yingyue Li, Haoran Yin, Wenyu Liu, and Xinggang Wang. 2025. Multimodal mamba: Decoder-only multimodal state space model via quadratic to linear distillation. *Preprint, arXiv:2502.13145*.
- Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let’s verify step by step. In *The Twelfth International Conference on Learning Representations*.
- Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2024. Let’s verify step by step. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.
- Bin Lin, Chen Zhang, Tao Peng, Hanyu Zhao, Wencong Xiao, Minmin Sun, Anmin Liu, Zhipeng Zhang, Lanbo Li, Xiafei Qiu, et al. 2024a. Infinite-llm: Efficient llm service for long context with distillation and distributed kvcache. *arXiv preprint arXiv:2401.02669*.
- Tianhe Lin, Jian Xie, Siyu Yuan, and Deqing Yang. 2025. Implicit reasoning in transformers is reasoning through shortcuts. *Preprint, arXiv:2503.07604*.
- Zicheng Lin, Zhibin Gou, Tian Liang, Ruilin Luo, Huawei Liu, and Yujiu Yang. 2024b. Criticbench: Benchmarking llms for critique-correct reasoning. *arXiv preprint arXiv:2402.14809*.
- Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024a. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

- Daizong Liu, Mingyu Yang, Xiaoye Qu, Pan Zhou, Yu Cheng, and Wei Hu. 2024b. A survey of attacks on large vision-language models: Resources, advances, and future trends. *arXiv preprint arXiv:2407.07403*.
- Tengxiao Liu, Qipeng Guo, Xiangkun Hu, Cheng Jiayang, Yue Zhang, Xipeng Qiu, and Zheng Zhang. 2024c. Can language models learn to skip steps? *arXiv preprint arXiv:2411.01855*.
- Wentao Liu, Hanglei Hu, Jie Zhou, Yuyang Ding, Junsong Li, Jiayi Zeng, Mengliang He, Qin Chen, Bo Jiang, Aimin Zhou, et al. 2023a. Mathematical language models: A survey. *arXiv preprint arXiv:2312.07622*.
- Yang Liu, Weixing Chen, Yongjie Bai, Xiaodan Liang, Guanbin Li, Wen Gao, and Liang Lin. 2024d. Aligning cyber space with physical world: A comprehensive survey on embodied ai. *arXiv preprint arXiv:2407.06886*.
- Yang Liu, Yuanshun Yao, Jean-Francois Ton, Xiaoying Zhang, Ruocheng Guo, Hao Cheng, Yegor Klochkov, Muhammad Faaiq Taufiq, and Hang Li. 2023b. Trustworthy llms: a survey and guideline for evaluating large language models' alignment. *arXiv preprint arXiv:2308.05374*.
- Yuqi Liu, Bohao Peng, Zhisheng Zhong, Zihao Yue, Fanbin Lu, Bei Yu, and Jiaya Jia. 2025a. Seg-zero: Reasoning-chain guided segmentation via cognitive reinforcement. *arXiv preprint arXiv:2503.06520*.
- Zhenhua Liu, Lijun Li, Ruizhe Chen, Yuxian Jiang, Tong Zhu, Wenliang Chen, and Jing Shao. 2025b. Iterative value function optimization for guided decoding. *arXiv preprint arXiv:2503.02368*.
- Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. 2025c. Understanding r1-zero-like training: A critical perspective. <https://github.com/sail-sg/understand-r1-zero>.
- Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. 2025d. Visual-rft: Visual reinforcement fine-tuning. *arXiv preprint arXiv:2503.01785*.
- Enzhe Lu, Zhejun Jiang, Jingyuan Liu, Yulun Du, Tao Jiang, Chao Hong, Shaowei Liu, Weiran He, Enming Yuan, Yuzhi Wang, et al. 2025a. Moba: Mixture of block attention for long-context llms. *arXiv preprint arXiv:2502.13189*.
- Keming Lu, Hongyi Yuan, Runji Lin, Junyang Lin, Zheng Yuan, Chang Zhou, and Jingren Zhou. 2023. Routing to the expert: Efficient reward-guided ensemble of large language models. *Preprint, arXiv:2311.08692*.
- Li-Chun Lu, Shou-Jen Chen, Tsung-Min Pai, Chan-Hung Yu, Hung-yi Lee, and Shao-Hua Sun. 2024a. Llm discussion: Enhancing the creativity of large language models via discussion framework and role-play. *arXiv preprint arXiv:2405.06373*.
- Xiaoya Lu, Dongrui Liu, Yi Yu, Luxin Xu, and Jing Shao. 2025b. X-boundary: Establishing exact safety boundary to shield llms from multi-turn jailbreaks without compromising usability. *arXiv preprint arXiv:2502.09990*.
- Zhenyi Lu, Chenghao Fan, Wei Wei, Xiaoye Qu, Dangyang Chen, and Yu Cheng. 2024b. Twin-merging: Dynamic integration of modular expertise in model merging. *Advances in Neural Information Processing Systems*, 37:78905–78935.
- Haotian Luo, Li Shen, Haiying He, Yibo Wang, Shiwei Liu, Wei Li, Naiqiang Tan, Xiaochun Cao, and Dacheng Tao. 2025a. O1-pruner: Length-harmonizing fine-tuning for o1-like reasoning pruning. *arXiv preprint arXiv:2501.12570*.
- Yijia Luo, Yulin Song, Xingyao Zhang, Jiaheng Liu, Weixun Wang, GengRu Chen, Wenbo Su, and Bo Zheng. 2025b. Deconstructing long chain-of-thought: A structured reasoning optimization framework for long cot distillation. *Preprint, arXiv:2503.16385*.
- Xinyin Ma, Gongfan Fang, and Xinchao Wang. 2023. Llm-pruner: On the structural pruning of large language models. *Advances in neural information processing systems*, 36:21702–21720.
- Xinyin Ma, Guangnian Wan, Runpeng Yu, Gongfan Fang, and Xinchao Wang. 2025. Cot-valve: Length-compressible chain-of-thought tuning. *arXiv preprint arXiv:2502.09601*.
- Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, Botian Shi, Wenhui Wang, Junjun He, Kaipeng Zhang, et al. 2025. Mm-eureka: Exploring visual aha moment with rule-based large-scale reinforcement learning. *arXiv preprint arXiv:2503.07365*.
- Shen-Yun Miao, Chao-Chun Liang, and Keh-Yih Su. 2021. A diverse corpus for evaluating and developing english math word problem solvers. *arXiv preprint arXiv:2106.15772*.
- Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. 2025. s1: Simple test-time scaling. *arXiv preprint arXiv:2501.19393*.
- Tergel Munkhbat, Namgyu Ho, Seohyun Kim, Yongjin Yang, Yujin Kim, and Se-Young Yun. 2025. Self-training elicits concise reasoning in large language models. *Preprint, arXiv:2502.20122*.
- Sania Nayab, Giulio Rossolini, Marco Simoni, Andrea Saracino, Giorgio Buttazzo, Nicolamaria Manes, and Fabrizio Giacomelli. 2024. Concise thoughts: Impact of output length on llm reasoning and cost. *arXiv preprint arXiv:2407.19825*.

- Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. 2025. Large language diffusion models. *arXiv preprint arXiv:2502.09992*.
- Isaac Ong, Amjad Almahairi, Vincent Wu, Wei-Lin Chang, Tianhao Wu, Joseph E. Gonzalez, M Waleed Kadous, and Ion Stoica. 2025. *Routellm: Learning to route llms with preference data*. *Preprint*, arXiv:2406.18665.
- Artidoro Pagnoni, Ram Pasunuru, Pedro Rodriguez, John Nguyen, Benjamin Muller, Margaret Li, Chunting Zhou, Lili Yu, Jason Weston, Luke Zettlemoyer, et al. 2024. Byte latent transformer: Patches scale better than tokens. *arXiv preprint arXiv:2412.09871*.
- Daniele Paliotta, Junxiong Wang, Matteo Pagliardini, Kevin Y Li, Aviv Bick, J Zico Kolter, Albert Gu, François Fleuret, and Tri Dao. 2025. Thinking slow, fast: Scaling inference compute with distilled reasoners. *arXiv preprint arXiv:2502.20339*.
- Jiabao Pan, Yan Zhang, Chen Zhang, Zuozhu Liu, Hongwei Wang, and Haizhou Li. 2024. Dynathink: Fast or slow? a dynamic decision-making framework for large language models. *arXiv preprint arXiv:2407.01009*.
- Jiazheng Pan, Che Liu, Junde Wu, Fenglin Liu, Jiayuan Zhu, Hongwei Bran Li, Chen Chen, Cheng Ouyang, and Daniel Rueckert. 2025. Medvilm-r1: Incentivizing medical reasoning capability of vision-language models (vlms) via reinforcement learning. *arXiv preprint arXiv:2502.19634*.
- Liangming Pan, Alon Albalak, Xinyi Wang, and William Yang Wang. 2023. Logic-lm: Empowering large language models with symbolic solvers for faithful logical reasoning. *arXiv preprint arXiv:2305.12295*.
- Pratyush Patel, Esha Choukse, Chaojie Zhang, Aashaka Shah, Íñigo Goiri, Saeed Maleki, and Ricardo Bianchini. 2024. Splitwise: Efficient generative llm inference using phase splitting. In *2024 ACM/IEEE 51st Annual International Symposium on Computer Architecture (ISCA)*, pages 118–132. IEEE.
- Bo Peng, Daniel Goldstein, Quentin Anthony, Alon Albalak, Eric Alcaide, Stella Biderman, Eugene Cheah, Teddy Ferdinand, Haowen Hou, Przemysław Kazienko, et al. 2024. Eagle and finch: Rwkv with matrix-valued states and dynamic recurrence. *arXiv preprint arXiv:2404.05892*.
- Bo Peng, Ruichong Zhang, Daniel Goldstein, Eric Alcaide, Haowen Hou, Janna Lu, William Merrill, Guangyu Song, Kaifeng Tan, Saiteja Utpala, Nathan Wilce, Johan S. Wind, Tianyi Wu, Daniel Wuttke, and Christian Zhou-Zheng. 2025a. *Rwkv-7 "goose" with expressive dynamic state evolution*. *Preprint*, arXiv:2503.14456.
- Yi Peng, Chris, Xiaokun Wang, Yichen Wei, Jiangbo Pei, Weijie Qiu, Ai Jian, Yunzhuo Hao, Jiachun Pan, Tianyidan Xie, Ge Li, Rongxian Zhuang, Xuchen Song, Yang Liu, and Yahui Zhou. 2025b. *Skywork r1v: Pioneering multimodal reasoning with chain-of-thought*.
- Yingzhe Peng, Gongrui Zhang, Miaosen Zhang, Zhiyuan You, Jie Liu, Qipeng Zhu, Kai Yang, Xingzhong Xu, Xin Geng, and Xu Yang. 2025c. Lmm-r1: Empowering 3b lmms with strong reasoning abilities through two-stage rule-based rl. *arXiv preprint arXiv:2503.07536*.
- Antonio Polino, Razvan Pascanu, and Dan Alistarh. 2018. Model compression via distillation and quantization. *arXiv preprint arXiv:1802.05668*.
- Chen Qian, Dongrui Liu, Jie Zhang, Yong Liu, and Jing Shao. 2024a. Dean: Deactivating the coupled neurons to mitigate fairness-privacy conflicts in large language models. *arXiv preprint arXiv:2410.16672*.
- Chen Qian, Jie Zhang, Wei Yao, Dongrui Liu, Zhenfei Yin, Yu Qiao, Yong Liu, and Jing Shao. 2024b. Towards tracing trustworthiness dynamics: Revisiting pre-training period of large language models. *arXiv preprint arXiv:2402.19465*.
- Yujia Qin, Shengding Hu, Yankai Lin, Weize Chen, Ning Ding, Ganqu Cui, Zheni Zeng, Yufei Huang, Chaojun Xiao, Chi Han, Yi Ren Fung, Yusheng Su, Huadong Wang, Cheng Qian, Runchu Tian, Kunlun Zhu, Shihao Liang, Xingyu Shen, Bokai Xu, Zhen Zhang, Yining Ye, Bowen Li, Ziwei Tang, Jing Yi, Yuzhang Zhu, Zhenning Dai, Lan Yan, Xin Cong, Yaxi Lu, Weilin Zhao, Yuxiang Huang, Junxi Yan, Xu Han, Xian Sun, Dahai Li, Jason Phang, Cheng Yang, Tongshuang Wu, Heng Ji, Zhiyuan Liu, and Maosong Sun. 2024a. *Tool learning with foundation models*. *Preprint*, arXiv:2304.08354.
- Zhen Qin, Dong Li, Weigao Sun, Weixuan Sun, Xuyang Shen, Xiaodong Han, Yunshen Wei, Bao-hong Lv, Xiao Luo, Yu Qiao, and Yiran Zhong. 2024b. *Transnformerllm: A faster and better large language model with improved transnformer*. *Preprint*, arXiv:2307.14995.
- Zhen Qin, Dong Li, Weigao Sun, Weixuan Sun, Xuyang Shen, Xiaodong Han, Yunshen Wei, Bao-hong Lv, Fei Yuan, Xiao Luo, et al. 2023. Scaling transnformer to 175 billion parameters. *arXiv preprint arXiv:2307.14995*.
- Zhen Qin, Xuyang Shen, Weigao Sun, Dong Li, Stan Birchfield, Richard Hartley, and Yiran Zhong. 2024c. Unlocking the secrets of linear complexity sequence model from a unified perspective. *arXiv preprint arXiv:2405.17383*.
- Zhen Qin, Weigao Sun, Dong Li, Xuyang Shen, Weixuan Sun, and Yiran Zhong. 2024d. Lightning attention-2: A free lunch for handling unlimited sequence lengths in large language models. *arXiv preprint arXiv:2401.04658*.

- Zhen Qin, Weigao Sun, Dong Li, Xuyang Shen, Weixuan Sun, and Yiran Zhong. 2024e. Various lengths, constant speed: Efficient language modeling with lightning attention. *arXiv preprint arXiv:2405.17381*.
- Zhen Qin, Songlin Yang, Weixuan Sun, Xuyang Shen, Dong Li, Weigao Sun, and Yiran Zhong. 2024f. Hgrn2: Gated linear rnns with state expansion. *arXiv preprint arXiv:2404.07904*.
- Zhen Qin, Songlin Yang, and Yiran Zhong. 2024g. Hierarchically gated recurrent neural network for sequence modeling. *Advances in Neural Information Processing Systems*, 36.
- Jiahao Qiu, Yifu Lu, Yifan Zeng, Jiacheng Guo, Jiayi Geng, Huazheng Wang, Kaixuan Huang, Yue Wu, and Mengdi Wang. 2024. **Treebon: Enhancing inference-time alignment with speculative tree-search and best-of-n sampling**. *ArXiv*, abs/2410.16033.
- Changle Qu, Sunhao Dai, Xiaochi Wei, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, Jun Xu, and Ji-Rong Wen. 2025a. Tool learning with large language models: A survey. *Frontiers of Computer Science*, 19(8):198343.
- Xiaoye Qu, Qiyuan Chen, Wei Wei, Jishuo Sun, and Jianfeng Dong. 2024a. Alleviating hallucination in large vision-language models with active retrieval augmentation. *arXiv preprint arXiv:2408.00555*.
- Xiaoye Qu, Daize Dong, Xuyang Hu, Tong Zhu, Weigao Sun, and Yu Cheng. 2024b. Llama-moe v2: Exploring sparsity of llama from perspective of mixture-of-experts with post-training. *arXiv preprint arXiv:2411.15708*.
- Xiaoye Qu, Jiashuo Sun, Wei Wei, and Yu Cheng. 2024c. Look, compare, decide: Alleviating hallucination in large vision-language models via multi-view multi-path reasoning. *arXiv preprint arXiv:2408.17150*.
- Yuxiao Qu, Matthew Y. R. Yang, Amirth Setlur, Lewis Tunstall, Edward Emanuel Beeching, Ruslan Salakhutdinov, and Aviral Kumar. 2025b. **Optimizing test-time compute via meta reinforcement fine-tuning**. *Preprint*, arXiv:2503.07572.
- Qibing Ren, Hao Li, Dongrui Liu, Zhanxu Xie, Xiaoya Lu, Yu Qiao, Lei Sha, Junchi Yan, Lizhuang Ma, and Jing Shao. 2024. Derail yourself: Multi-turn llm jailbreak attack through self-discovered clues. *arXiv preprint arXiv:2410.10700*.
- Matthew Renze and Erhan Guven. 2024. The benefits of a concise chain of thought on problem-solving in large language models. In *2024 2nd International Conference on Foundation and Large Language Models (FLLM)*, pages 476–483. IEEE.
- Kai Ruan, Xuan Wang, Jixiang Hong, Peng Wang, Yang Liu, and Hao Sun. 2024. Liveideabench: Evaluating llms’ scientific creativity and idea generation with minimal context. *arXiv preprint arXiv:2412.17596*.
- Hyun Ryu and Eric Kim. 2024. Closer look at efficient inference methods: A survey of speculative decoding. *arXiv preprint arXiv:2411.13157*.
- Swarnadeep Saha, Archiki Prasad, Justin Chih-Yao Chen, Peter Hase, Elias Stengel-Eskin, and Mohit Bansal. 2024. **System-1.x: Learning to balance fast and slow planning with language models**. *Preprint*, arXiv:2407.14414.
- Xuan Shen, Yizhou Wang, Xiangxi Shi, Yanzhi Wang, Pu Zhao, and Jiuxiang Gu. 2025a. Efficient reasoning with hidden thinking. *arXiv preprint arXiv:2501.19201*.
- Xuyang Shen, Dong Li, Ruitao Leng, Zhen Qin, Weigao Sun, and Yiran Zhong. 2024. Scaling laws for linear complexity language models. *arXiv preprint arXiv:2406.16690*.
- Yi Shen, Jian Zhang, Jieyun Huang, Shuming Shi, Wenzheng Zhang, Jiangze Yan, Ning Wang, Kai Wang, and Shiguo Lian. 2025b. **Dast: Difficulty-adaptive slow-thinking for large reasoning models**. *Preprint*, arXiv:2503.04472.
- Zhenyi Shen, Hanqi Yan, Linhai Zhang, Zhanghao Hu, Yali Du, and Yulan He. 2025c. **Codi: Compressing chain-of-thought into continuous space via self-distillation**. *Preprint*, arXiv:2502.21074.
- Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2024. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*.
- Mingyang Song, Xiaoye Qu, Jiawei Zhou, and Yu Cheng. 2025a. **From head to tail: Towards balanced representation in large vision-language models through adaptive data calibration**. *Preprint*, arXiv:2503.12821.
- Mingyang Song, Zhaochen Su, Xiaoye Qu, Jiawei Zhou, and Yu Cheng. 2025b. **Prmbench: A fine-grained and challenging benchmark for process-level reward models**. *arXiv preprint arXiv:2501.03124*.
- Petru Soviany, Radu Tudor Ionescu, Paolo Rota, and Nicu Sebe. 2022. Curriculum learning: A survey. *International Journal of Computer Vision*, 130(6):1526–1565.
- Zayne Sprague, Fangcong Yin, Juan Diego Rodriguez, Dongwei Jiang, Manya Wadhwa, Prasann Singhal, Xinyu Zhao, Xi Ye, Kyle Mahowald, and Greg Durrett. 2024. **To cot or not to cot? chain-of-thought helps mainly on math and symbolic reasoning**. *Preprint*, arXiv:2409.12183.
- Matteo Stefanini, Marcella Cornia, Lorenzo Baraldi, Silvia Cascianelli, Giuseppe Fiameni, and Rita Cucchiara. 2021. **From show to tell: A survey on deep learning-based image captioning**. *Preprint*, arXiv:2107.06912.

- DiJia Su, Sainbayar Sukhbaatar, Michael Rabbat, Yuan-dong Tian, and Qinqing Zheng. 2024a. Dual-former: Controllable fast and slow thinking by learning with randomized reasoning traces. *Preprint*, arXiv:2410.09918.
- DiJia Su, Hanlin Zhu, Yingchen Xu, Jiantao Jiao, Yuandong Tian, and Qinqing Zheng. 2025. Token assorted: Mixing latent and text tokens for improved language model reasoning. *arXiv preprint arXiv:2502.03275*.
- Zhaochen Su, Juntao Li, Jun Zhang, Tong Zhu, Xiaoye Qu, Pan Zhou, Yan Bowen, Yu Cheng, et al. 2024b. Living in the moment: Can large language models grasp co-temporal reasoning? *arXiv preprint arXiv:2406.09072*.
- Zhaochen Su, Juntao Li, Zikang Zhang, Zihan Zhou, and Min Zhang. 2023. Efficient continue training of temporal language model with structural information. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6315–6329.
- Zhaochen Su, Jun Zhang, Xiaoye Qu, Tong Zhu, Yanshu Li, Jiashuo Sun, Juntao Li, Min Zhang, and Yu Cheng. 2024c. Conflictbank: A benchmark for evaluating the influence of knowledge conflicts in llm. *arXiv preprint arXiv:2408.12076*.
- Zhaochen Su, Jun Zhang, Tong Zhu, Xiaoye Qu, Juntao Li, Min Zhang, and Yu Cheng. 2024d. Timo: Towards better temporal reasoning for language models. *arXiv preprint arXiv:2406.14192*.
- Yuan Sui, Yafei He, Tri Cao, Simeng Han, and Bryan Hooi. 2025. Meta-reasoner: Dynamic guidance for optimized inference-time reasoning in large language models. *Preprint*, arXiv:2502.19918.
- Guangyan Sun, Mingyu Jin, Zhenting Wang, Cheng-Long Wang, Siqi Ma, Qifan Wang, Tong Geng, Ying Nian Wu, Yongfeng Zhang, and Dongfang Liu. 2025a. Visual agents as fast and slow thinkers. *Preprint*, arXiv:2408.08862.
- Hanshi Sun, Momin Haider, Ruiqi Zhang, Huitao Yang, Jiahao Qiu, Ming Yin, Mengdi Wang, Peter Bartlett, and Andrea Zanette. 2024a. Fast best-of-n decoding via speculative rejection. *arXiv preprint arXiv:2410.20290*.
- Jiashuo Sun, Jihai Zhang, Yucheng Zhou, Zhaochen Su, Xiaoye Qu, and Yu Cheng. 2024b. Surf: Teaching large vision-language models to selectively utilize retrieved information. *arXiv preprint arXiv:2409.14083*.
- Weigao Sun, Disen Lan, Yiran Zhong, Xiaoye Qu, and Yu Cheng. 2025b. Lasp-2: Rethinking sequence parallelism for linear attention and its hybrid. *arXiv preprint arXiv:2502.07563*.
- Weigao Sun, Disen Lan, Tong Zhu, Xiaoye Qu, and Yu Cheng. 2025c. Linear-moe: Linear sequence modeling meets mixture-of-experts. *arXiv preprint arXiv:2503.05447*.
- Weigao Sun, Zhen Qin, Dong Li, Xuyang Shen, Yu Qiao, and Yiran Zhong. 2024c. Linear attention sequence parallelism. *arXiv preprint arXiv:2404.02882*.
- Weigao Sun, Zhen Qin, Weixuan Sun, Shidi Li, Dong Li, Xuyang Shen, Yu Qiao, and Yiran Zhong. 2024d. Co2: Efficient distributed training with full communication-computation overlap. *arXiv preprint arXiv:2401.16265*.
- Yu Sun, Xinhao Li, Karan Dalal, Jiarui Xu, Arjun Vikram, Genghan Zhang, Yann Dubois, Xinlei Chen, Xiaolong Wang, Sanmi Koyejo, et al. 2024e. Learning to (learn at test time): Rnns with expressive hidden states. *arXiv preprint arXiv:2407.04620*.
- Yutao Sun, Li Dong, Shaohan Huang, Shuming Ma, Yuqing Xia, Jilong Xue, Jianyong Wang, and Furu Wei. 2023. Retentive network: A successor to transformer for large language models. *arXiv preprint arXiv:2307.08621*.
- Jihoon Tack, Jack Lanchantin, Jane Yu, Andrew Cohen, Ilia Kulikov, Janice Lan, Shibo Hao, Yuandong Tian, Jason Weston, and Xian Li. 2025. Llm pre-training with continuous concepts. *arXiv preprint arXiv:2502.08524*.
- Alon Talmor, Oyvind Tafjord, Peter Clark, Yoav Goldberg, and Jonathan Berant. 2020. Leap-of-thought: Teaching pre-trained models to systematically reason over implicit knowledge. *Advances in Neural Information Processing Systems*, 33:20227–20237.
- Sree Harsha Tanneru, Dan Ley, Chirag Agarwal, and Himabindu Lakkaraju. 2024. On the hardness of faithful chain-of-thought reasoning in large language models. *arXiv preprint arXiv:2406.10625*.
- InternLM Team. 2023. Internlm: A multilingual language model with progressively enhanced capabilities.
- Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, Chunling Tang, Congcong Wang, Dehao Zhang, Enming Yuan, Enzhe Lu, Fengxiang Tang, Flood Sung, Guangda Wei, Guokun Lai, Haiqing Guo, Han Zhu, Hao Ding, Hao Hu, Hao Yang, Hao Zhang, Haotian Yao, Haotian Zhao, Haoyu Lu, Haoze Li, Haozhen Yu, Hongcheng Gao, Huabin Zheng, Huan Yuan, Jia Chen, Jianhang Guo, Jianlin Su, Jianzhou Wang, Jie Zhao, Jin Zhang, Jingyuan Liu, Junjie Yan, Junyan Wu, Lidong Shi, Ling Ye, Longhui Yu, Mengnan Dong, Neo Zhang, Ningchen Ma, Qiwei Pan, Qucheng Gong, Shaowei Liu, Shengling Ma, Shupeng Wei, Sihan Cao, Siying Huang, Tao Jiang, Weihao Gao, Weimin Xiong, Weiran He, Weixiao Huang, Wenhao Wu, Wenyang He, Xianghui Wei, Xianqing Jia, Xingzhe Wu, Xinran Xu, Xinxing Zu, Xinyu Zhou, Xuehai Pan, Y. Charles, Yang Li, Yangyang Hu, Yangyang Liu, Yanru Chen, Yejie Wang, Yibo Liu, Yidao Qin, Yifeng Liu, Ying Yang,

- Yiping Bao, Yulun Du, Yuxin Wu, Yuzhi Wang, Zaida Zhou, Zhaoji Wang, Zhaowei Li, Zhen Zhu, Zheng Zhang, Zhexu Wang, Zhilin Yang, Zhiqi Huang, Zihao Huang, Ziyao Xu, and Zonghan Yang. 2025. **Kimi k1.5: Scaling reinforcement learning with llms.** *Preprint*, arXiv:2501.12599.
- Qwen Team. 2024. **Qwq: Reflect deeply on the boundaries of the unknown.**
- Abdulrahman Temsah, Khalid Alhasan, Ibraheem Altamimi, Amr Jamal, Ayman Al-Eyadhy, Khalid H Malki, and Mohamad-Hani Temsah. 2025. Deepseek in healthcare: Revealing opportunities and steering challenges of a new open-source artificial intelligence frontier. *Cureus*, 17(2).
- Mohamad-Hani Temsah, Amr Jamal, Khalid Alhasan, Abdulkarim A Temsah, and Khalid H Malki. 2024. Openai o1-preview vs. chatgpt in healthcare: a new frontier in medical ai reasoning. *Cureus*, 16(10).
- Omkar Thawakar, Dinura Dissanayake, Ketan More, Ritesh Thawkar, Ahmed Heakl, Noor Ahsan, Yuhao Li, Mohammed Zumri, Jean Lahoud, Rao Muhammad Anwer, et al. 2025. Llamav-o1: Rethinking step-by-step visual reasoning in llms. *arXiv preprint arXiv:2501.06186*.
- LCM The, Loïc Barrault, Paul-Ambroise Duquenne, Maha Elbayad, Artyom Kozhevnikov, Belen Alastruey, Pierre Andrews, Mariano Coria, Guillaume Couairon, Marta R Costa-jussà, et al. 2024. Large concept models: Language modeling in a sentence representation space. *arXiv preprint arXiv:2412.08821*.
- Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36:74952–74965.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.
- Guangya Wan, Yuqi Wu, Jie Chen, and Sheng Li. 2024. Dynamic self-consistency: Leveraging reasoning paths for efficient llm sampling. *arXiv preprint arXiv:2408.17017*.
- Boshi Wang, Sewon Min, Xiang Deng, Jiaming Shen, You Wu, Luke Zettlemoyer, and Huan Sun. 2023a. **Towards understanding chain-of-thought prompting: An empirical study of what matters.** *Preprint*, arXiv:2212.10001.
- Liang Wang, Haonan Chen, Nan Yang, Xiaolong Huang, Zhicheng Dou, and Furu Wei. 2025a. Chain-of-retrieval augmented generation. *arXiv preprint arXiv:2501.14342*.
- Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhi-hao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. 2024a. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.
- Xin Wang, Yudong Chen, and Wenwu Zhu. 2021. **A survey on curriculum learning.** *Preprint*, arXiv:2010.13166.
- Xinglin Wang, Shaoxiong Feng, Yiwei Li, Peiwen Yuan, Yueqi Zhang, Chuyi Tan, Boyuan Pan, Yao Hu, and Kan Li. 2024b. Make every penny count: Difficulty-adaptive self-consistency for cost-efficient reasoning. *arXiv preprint arXiv:2408.13457*.
- Xinyi Wang, Lucas Caccia, Oleksiy Ostapenko, Xingdi Yuan, William Yang Wang, and Alessandro Sordoni. 2023b. Guiding language model reasoning with planning tokens. *arXiv preprint arXiv:2310.05707*.
- Xinyuan Wang, Yanchi Liu, Wei Cheng, Xujiang Zhao, Zhengzhang Chen, Wenchao Yu, Yanjie Fu, and Haifeng Chen. 2025b. **Mixllm: Dynamic routing in mixed large language models.** *Preprint*, arXiv:2502.18482.
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022a. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023c. **Self-consistency improves chain of thought reasoning in language models.** In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.
- Yaoting Wang, Shengqiong Wu, Yuecheng Zhang, William Wang, Ziwei Liu, Jiebo Luo, and Hao Fei. 2025c. **Multimodal chain-of-thought reasoning: A comprehensive survey.** *Preprint*, arXiv:2503.12605.
- Yiming Wang, Pei Zhang, Siyuan Huang, Baosong Yang, Zhuosheng Zhang, Fei Huang, and Rui Wang. 2025d. Sampling-efficient test-time scaling: Self-estimating the best-of-n sampling in early decoding. *arXiv preprint arXiv:2503.01422*.
- Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hananeh Hajishirzi. 2022b. Self-instruct: Aligning language models with self-generated instructions. *arXiv preprint arXiv:2212.10560*.
- Yue Wang, Qiuwei Liu, Jiahao Xu, Tian Liang, Xingyu Chen, Zhiwei He, Linfeng Song, Dian Yu, Juntao Li, Zhuosheng Zhang, et al. 2025e. Thoughts are all over the place: On the underthinking of o1-like llms. *arXiv preprint arXiv:2501.18585*.
- Ziheng Wang, Jeremy Wohlwend, and Tao Lei. 2019. Structured pruning of large language models. *arXiv preprint arXiv:1910.04732*.

- P.C. Wason and J.S.T.B.T. Evans. 1974. Dual processes in reasoning? *Cognition*, 3(2):141–154.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.
- Han Wu, Yuxuan Yao, Shuqi Liu, Zehua Liu, Xiaojin Fu, Xiongwei Han, Xing Li, Hui-Ling Zhen, Tao Zhong, and Mingxuan Yuan. 2025a. Unlocking efficient long-to-short llm reasoning with model merging. *Preprint*, arXiv:2503.20641.
- Ian Wu, Patrick Fernandes, Amanda Bertsch, Seungone Kim, Sina Pakazad, and Graham Neubig. 2024. Better instruction-following through minimum bayes risk. *arXiv preprint arXiv:2410.02902*.
- Yuyang Wu, Yifei Wang, Tianqi Du, Stefanie Jegelka, and Yisen Wang. 2025b. When more is less: Understanding chain-of-thought length in llms. *Preprint*, arXiv:2502.07266.
- Heming Xia, Yongqi Li, Chak Tou Leong, Wenjie Wang, and Wenjie Li. 2025. Tokenskip: Controllable chain-of-thought compression in llms. *arXiv preprint arXiv:2502.12067*.
- Mengzhou Xia, Tianyu Gao, Zhiyuan Zeng, and Danqi Chen. 2023. Sheared llama: Accelerating language model pre-training via structured pruning. *arXiv preprint arXiv:2310.06694*.
- Peng Xia, Kangyu Zhu, Haoran Li, Hongtu Zhu, Yun Li, Gang Li, Linjun Zhang, and Huaxiu Yao. 2024. Rule: Reliable multimodal rag for factuality in medical vision language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1081–1093.
- Kun Xiang, Zhili Liu, Zihao Jiang, Yunshuang Nie, Kaixin Cai, Yiyang Yin, Runhui Huang, Haoxiang Fan, Hanhui Li, Weiran Huang, Yihan Zeng, Yu-Jie Yuan, Jianhua Han, Lanqing Hong, Hang Xu, and Xi-aodan Liang. 2025. Can atomic step decomposition enhance the self-structured reasoning of multimodal large models? *Preprint*, arXiv:2503.06252.
- Guangxuan Xiao, Ji Lin, Mickael Seznec, Hao Wu, Julien Demouth, and Song Han. 2023a. Smoothquant: Accurate and efficient post-training quantization for large language models. In *International Conference on Machine Learning*, pages 38087–38099. PMLR.
- Guangxuan Xiao, Jiaming Tang, Jingwei Zuo, Junxian Guo, Shang Yang, Haotian Tang, Yao Fu, and Song Han. 2024. Duoattention: Efficient long-context llm inference with retrieval and streaming heads. *arXiv preprint arXiv:2410.10819*.
- Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. 2023b. Efficient streaming language models with attention sinks. *arXiv preprint arXiv:2309.17453*.
- Silei Xu, Wenhao Xie, Lingxiao Zhao, and Pengcheng He. 2025a. Chain of draft: Thinking faster by writing less. *Preprint*, arXiv:2502.18600.
- Yige Xu, Xu Guo, Zhiwei Zeng, and Chunyan Miao. 2025b. Softcot: Soft chain-of-thought for efficient reasoning with llms. *arXiv preprint arXiv:2502.12134*.
- Yuchen Yan, Yongliang Shen, Yang Liu, Jin Jiang, Mengdi Zhang, Jian Shao, and Yueting Zhuang. 2025. Infitythink: Breaking the length limits of long-context reasoning in large language models. *arXiv preprint arXiv:2503.06692*.
- An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024a. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.
- Chenxiao Yang, Nathan Srebro, David McAllester, and Zhiyuan Li. 2025a. PENCIL: Long thoughts with short memory. In *First Workshop on Scalable Optimization for Efficient and Adaptive Foundation Models*.
- Dayu Yang, Tianyang Liu, Daoan Zhang, Antoine Simoulin, Xiaoyi Liu, Yuwei Cao, Zhaopu Teng, Xin Qian, Grey Yang, Jiebo Luo, et al. 2025b. Code to think, think to code: A survey on code-enhanced reasoning and reasoning-driven code intelligence in llms. *arXiv preprint arXiv:2502.19411*.
- Sohee Yang, Elena Gribovskaya, Nora Kassner, Mor Geva, and Sebastian Riedel. 2024b. Do large language models latently perform multi-hop reasoning? *Preprint*, arXiv:2402.16837.
- Sohee Yang, Elena Gribovskaya, Nora Kassner, Mor Geva, and Sebastian Riedel. 2024c. Do large language models latently perform multi-hop reasoning? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10210–10229, Bangkok, Thailand. Association for Computational Linguistics.
- Songlin Yang, Jan Kautz, and Ali Hatamizadeh. 2024d. Gated delta networks: Improving mamba2 with delta rule. *Preprint*, arXiv:2412.06464.
- Songlin Yang, Bailin Wang, Yikang Shen, Rameswar Panda, and Yoon Kim. 2023. Gated linear attention transformers with hardware-efficient training. *arXiv preprint arXiv:2312.06635*.
- Songlin Yang, Bailin Wang, Yu Zhang, Yikang Shen, and Yoon Kim. 2024e. Parallelizing linear transformers with the delta rule over sequence length. *arXiv preprint arXiv:2406.06484*.
- Wenkai Yang, Shuming Ma, Yankai Lin, and Furu Wei. 2025c. Towards thinking-optimal scaling of test-time compute for llm reasoning. *arXiv preprint arXiv:2502.18080*.

- Huanjin Yao, Jiaxing Huang, Wenhao Wu, Jingyi Zhang, Yibo Wang, Shunyu Liu, Yingjie Wang, Yuxin Song, Haocheng Feng, Li Shen, et al. 2024a. Mulberry: Empowering mllm with o1-like reasoning and reflection via collective monte carlo tree search. *arXiv preprint arXiv:2412.18319*.
- Yifan Yao, Jinshuo Duan, Kaidi Xu, Yuanfang Cai, Zhibo Sun, and Yue Zhang. 2024b. A survey on large language model (llm) security and privacy: The good, the bad, and the ugly. *High-Confidence Computing*, page 100211.
- Gilad Yehudai, Noah Amsel, and Joan Bruna. 2025. Compositional reasoning with transformers, rnns, and chain of thought. *Preprint*, arXiv:2503.01544.
- Edward Yeo, Yuxuan Tong, Morry Niu, Graham Neubig, and Xiang Yue. 2025. Demystifying long chain-of-thought reasoning in llms. *Preprint*, arXiv:2502.03373.
- Ping Yu, Jing Xu, Jason Weston, and Ilia Kulikov. 2024a. Distilling system 2 into system 1. *Preprint*, arXiv:2407.06023.
- Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, Jin-hua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Weinan Dai, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Mu Qiao, Yonghui Wu, and Mingxuan Wang. 2025a. Dapo: An open-source llm reinforcement learning system at scale. *Preprint*, arXiv:2503.14476.
- Yue Yu, Zhengxing Chen, Aston Zhang, Liang Tan, Chenguang Zhu, Richard Yuanzhe Pang, Yundi Qian, Xuewei Wang, Suchin Gururangan, Chao Zhang, et al. 2024b. Self-generated critiques boost reward modeling for language models. *arXiv preprint arXiv:2411.16646*.
- Zishun Yu, Tengyu Xu, Di Jin, Karthik Abinav Sankararaman, Yun He, Wenxuan Zhou, Zhouhao Zeng, Eryk Helenowski, Chen Zhu, Sinong Wang, et al. 2025b. Think smarter not harder: Adaptive reasoning with inference aware optimization. *arXiv preprint arXiv:2501.17974*.
- Jingyang Yuan, Huazuo Gao, Damai Dai, Junyu Luo, Liang Zhao, Zhengyan Zhang, Zhenda Xie, YX Wei, Lean Wang, Zhiping Xiao, et al. 2025a. Native sparse attention: Hardware-aligned and natively trainable sparse attention. *arXiv preprint arXiv:2502.11089*.
- Yufeng Yuan, Yu Yue, Ruofei Zhu, Tiantian Fan, and Lin Yan. 2025b. What's behind ppo's collapse in long-cot? value optimization holds the secret. *Preprint*, arXiv:2503.01491.
- Ekim Yurtsever, Jacob Lambert, Alexander Carballo, and Kazuya Takeda. 2020. A survey of autonomous driving: Common practices and emerging technologies. *IEEE access*, 8:58443–58469.
- Zhiyuan Zeng, Qinyuan Cheng, Zhangyue Yin, Bo Wang, Shimin Li, Yunhua Zhou, Qipeng Guo, Xuanjing Huang, and Xipeng Qiu. 2024a. Scaling of search and learning: A roadmap to reproduce o1 from reinforcement learning perspective. *arXiv preprint arXiv:2412.14135*.
- Zhongshen Zeng, Yinhong Liu, Yingjia Wan, Jingyao Li, Pengguang Chen, Jianbo Dai, Yuxuan Yao, Rongwu Xu, Zehan Qi, Wanru Zhao, et al. 2024b. Mr-ben: A meta-reasoning benchmark for evaluating system-2 thinking in llms. *arXiv preprint arXiv:2406.13975*.
- Jintian Zhang, Yuqi Zhu, Mengshu Sun, Yujie Luo, Shuofei Qiao, Lun Du, Da Zheng, Huajun Chen, and Ningyu Zhang. 2025a. Lighththinker: Thinking step-by-step compression. *arXiv preprint arXiv:2502.15589*.
- Michael Zhang, Simran Arora, Rahul Chalamala, Alan Wu, Benjamin Spector, Aaryan Singhal, Krithik Ramesh, and Christopher Ré. 2024a. Lolcats: On low-rank linearizing of large language models. *arXiv preprint arXiv:2410.10254*.
- Pan Zhang, Xiaoyi Dong, Bin Wang, Yuhang Cao, Chao Xu, Linke Ouyang, Zhiyuan Zhao, Haodong Duan, Songyang Zhang, Shuangrui Ding, et al. 2023a. Internlm-xcomposer: A vision-language large model for advanced text-image comprehension and composition. *arXiv preprint arXiv:2309.15112*.
- Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, et al. 2023b. Instruction tuning for large language models: A survey. *arXiv preprint arXiv:2308.10792*.
- Xuan Zhang, Fengzhuo Zhang, Cunxiao Du, Chao Du, Tianyu Pang, Wei Gao, and Min Lin. 2025b. Lightransfer: Your long-context llm is secretly a hybrid model with effortless adaptation. *Preprint*, arXiv:2410.13846.
- Yang Zhang, Shixin Yang, Chenjia Bai, Fei Wu, Xiu Li, Zhen Wang, and Xuelong Li. 2024b. Towards efficient llm grounding for embodied multi-agent collaboration. *arXiv preprint arXiv:2405.14314*.
- Yu Zhang, Songlin Yang, Ruijie Zhu, Yue Zhang, Leyang Cui, Yiqiao Wang, Bolun Wang, Wei Bi, Freda Shi, Bailin Wang, Peng Zhou, and Guohong Fu. 2024c. Gated slot attention for efficient linear-time sequence modeling. *arXiv preprint arXiv:2409.07146*.
- Zhenru Zhang, Chujie Zheng, Yangzhen Wu, Beichen Zhang, Runji Lin, Bowen Yu, Dayiheng Liu, Jingen Zhou, and Junyang Lin. 2025c. The lessons of developing process reward models in mathematical reasoning. *arXiv preprint arXiv:2501.07301*.

- Jiaxing Zhao, Xihan Wei, and Liefeng Bo. 2025. R1-omni: Explainable omni-multimodal emotion recognition with reinforcement learning. *Preprint*, arXiv:2503.05379.
- Jingyuan Zhao, Wenyi Zhao, Bo Deng, Zhenghong Wang, Feng Zhang, Wenxiang Zheng, Wanke Cao, Jinrui Nan, Yubo Lian, and Andrew F Burke. 2024a. Autonomous driving system: A comprehensive survey. *Expert Systems with Applications*, 242:122836.
- Penghao Zhao, Hailin Zhang, Qinhan Yu, Zhengren Wang, Yunteng Geng, Fangcheng Fu, Ling Yang, Wentao Zhang, Jie Jiang, and Bin Cui. 2024b. Retrieval-augmented generation for ai-generated content: A survey. *arXiv preprint arXiv:2402.19473*.
- Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 1(2).
- Yinmin Zhong, Shengyu Liu, Junda Chen, Jianbo Hu, Yibo Zhu, Xuanzhe Liu, Xin Jin, and Hao Zhang. 2024. {DistServe}: Disaggregating prefill and decoding for goodput-optimized large language model serving. In *18th USENIX Symposium on Operating Systems Design and Implementation (OSDI 24)*, pages 193–210.
- Hengguang Zhou, Xirui Li, Ruochen Wang, Minhao Cheng, Tianyi Zhou, and Cho-Jui Hsieh. 2025a. R1-zero's "aha moment" in visual reasoning on a 2b non-sft model. *arXiv preprint arXiv:2503.05132*.
- Kaiwen Zhou, Chengzhi Liu, Xuandong Zhao, Shreedhar Jangam, Jayanth Srinivasa, Gaowen Liu, Dawn Song, and Xin Eric Wang. 2025b. The hidden risks of large reasoning models: A safety assessment of r1. *arXiv preprint arXiv:2502.12659*.
- Xueyang Zhou, Guiyao Tie, Guowen Zhang, Weidong Wang, Zhigang Zuo, Di Wu, Duanfeng Chu, Pan Zhou, Lichao Sun, and Neil Zhenqiang Gong. 2025c. Large reasoning models in agent scenarios: Exploring the necessity of reasoning capabilities. *Preprint*, arXiv:2503.11074.
- Zixuan Zhou, Xuefei Ning, Ke Hong, Tianyu Fu, Jiaming Xu, Shiyao Li, Yuming Lou, Lunling Wang, Zhihang Yuan, Xiuhong Li, et al. 2024. A survey on efficient inference for large language models. *arXiv preprint arXiv:2404.14294*.
- Tong Zhu, Xiaoye Qu, Daize Dong, Jiacheng Ruan, Jingqi Tong, Conghui He, and Yu Cheng. 2024. Llama-moe: Building mixture-of-experts from llama with continual pre-training. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15913–15923.
- Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. 2023. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*.
- Andy Zou, Long Phan, Justin Wang, Derek Duenas, Maxwell Lin, Maksym Andriushchenko, J Zico Kolter, Matt Fredrikson, and Dan Hendrycks. 2024. Improving alignment and robustness with circuit breakers. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.