Efficient Inference for Large Reasoning Models: A Survey

Yue Liu[*], **Jiaying Wu**[*], **Yufei He**[*]
National University of Singapore
yliu@u.nus.edu

Hongcheng Gao
University of Chinese Academy of Sciences

Hongyu Chen
Beijing Jiaotong University

Baolong Bi
University of Chinese Academy of Sciences

Jiaheng Zhang
National University of Singapore

Zhiqi Huang
Moonshot

Bryan Hooi
National University of Singapore

Abstract

Large Reasoning Models (LRMs) significantly improve the reasoning ability of Large Language Models (LLMs) by learning to reason, exhibiting promising performance in complex task-solving. However, their deliberative reasoning process leads to inefficiencies in token usage, memory consumption, and inference time. Thus, this survey provides a review of efficient inference methods designed specifically for LRMs, focusing on mitigating token inefficiency while preserving the reasoning quality. The overview structure of this paper is shown in Figure 1. First, we introduce a taxonomy to group the recent methods into two main categories: (a) explicit compact Chain-of-Thought (CoT), which reduces tokens while keeping the explicit reasoning structure, and (b) implicit latent CoT, which encodes reasoning steps within hidden representations instead of explicit tokens. Meanwhile, we discuss their strengths and weaknesses. Then, we conduct empirical analyses on existing methods from performance and efficiency aspects. Besides, we present open challenges in this field, including human-centric controllable reasoning, trade-off between interpretability and efficiency of reasoning, ensuring safety of efficient reasoning, and broader applications of efficient reasoning. In addition, we highlight key insights for enhancing LRMs' inference efficiency via techniques such as model merging, new architectures, and agent routers. We hope this work serves as a valuable guide, helping researchers overcome challenges in this vibrant field[1].

1 Introduction

Large Language Models (LLMs), which are trained to provide quick and intuitive responses, have exhibited great success in various fast-thinking applications like ChatBot (OpenAI, 2022). Differently, slow-thinking scenarios like math problem-solving (Olympiad, 2025) or research (OpenAI, 2025a) require the models to conduct analytical and deliberative reasoning before providing final responses. To tackle these challenges, Large Reasoning

[*]Equal Contribution
[1]https://github.com/yueliu1999/Awesome-Efficient-Inference-for-LRMs

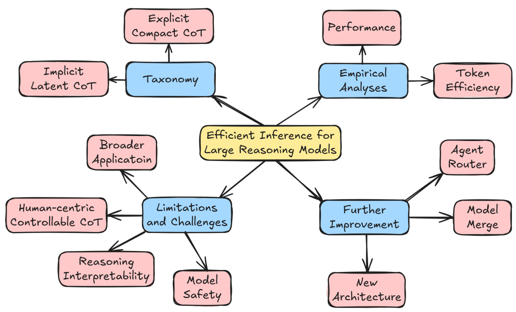

Figure 1: Overview Structure of this Survey.

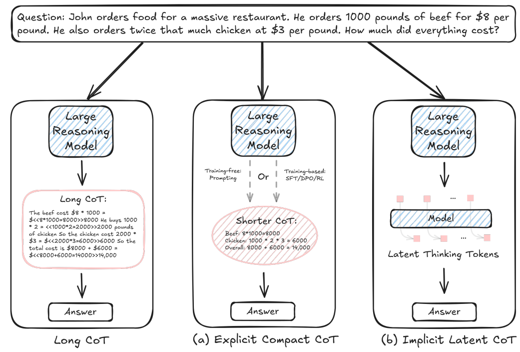

Figure 2: Overview of Taxonomy.

Models (LRMs) such as OpenAI o1/o3 (Jaech et al., 2024; OpenAI, 2025) and DeepSeek R1 (Guo et al., 2025) are developed by guiding the model to learn to reason.

Although effective, the intermediate reasoning process of LRMs is highly resource-intensive, learning to three challenges: (1) significant token consumption, (2) high memory overhead, and (3) increased inference time. These problems not only increase the inference cost of the service companies but also degrade the experience of the users. Therefore, efficient inference for LRMs has become an urgent and crucial direction. Since thinking tokens are treated like regular output tokens without cost differentiation, previous efforts in inference efficiency of regular LLMs, e.g., model compression (Wang et al., 2024b), efficient model design (Nie et al., 2025), and system-level optimization (Liu et al., 2024a), can alleviate problems (2) and (3). These methods are comprehensively studied (Zhou et al., 2024) and not specially designed for LRMs. Therefore, this paper focuses on the challenge (1): token inefficiency.

To this end, we conduct a comprehensive survey of recent efficient inference methods designed specifically for LRMs, aiming at improving thinking token efficiency while preserving reasoning quality. Specifically, we present a hierarchical taxonomy that first categorizes recent approaches into two classes. As shown in Figure 2, it contains (a) the explicit compact CoT, which reduces the number of thinking tokens while maintaining explicit reasoning structure, and (b) the implicit latent CoT, which encodes reasoning steps within hidden representations instead of explicit tokens. In addition, for the explicit compact CoT, we further summarize three sub-categories: (a.1) CoT compression, (a.2) CoT preference optimization, and (a.3) reward-based CoT conciseness. Then, we analyze the characteristics of these categories and discuss their strengths and weaknesses from the aspects of reasoning quality and efficiency. Besides, we conduct a comprehensive empirical study on the existing methods from the perspectives of performance and efficiency. Besides, we identify four open challenges regarding the inference efficiency of LRMs, including human-centric controllable reasoning, the trade-off between efficiency and interoperability of reasoning, ensuring the safety of efficient reasoning, and broader applications of efficient LRMs beyond math and code. Last, we highlight several potential technical solutions for further improvement of current methods, i.e., model merging, new architectures, and agent routers. We hope that this survey helps researchers and engineers further improve efficient inference for LRMs. The main contributions of this paper are summarized as follows.

- We conduct a comprehensive paper review of current methods of efficient inference for LRMs with a hierarchical taxonomy and strength & weakness discussion.
- We empirically study recent methods from performance and efficiency views and summarize 4 challenges from user control, interpretability, safety, and application aspects.
- We highlight technical insights in further improvement of existing methods from the perspectives of model merging, non-autoregressive architectures, and agent routers

2 Background

This section first introduces the background of large reasoning models and then highlights the efficiency challenges in the inference phase of large reasoning models.

Types	Methods	Training	Strategy	Model	Application
Explicit Compact CoT	SoT (Aytes et al., 2025)	✗	Prompt	Qwen-2.5-7B/14B/32B	Math, Commonsense, Logic, Scientific, Medical
	Constrained-CoT (Nayab et al., 2024)	✗	Prompt	LLaMA-2-70B, Falcon-40B	Math
	CoD (Xu et al., 2025b)	✗	Prompt	GPT-4o, Claude 3.5 Sonnet	Math, Commonsense, Symbolic Reasoning
	TALE-EP (Han et al., 2024)	✗	Prompt	LLaMA-3.1-8B-Instruct	Math
	Meta-Reasoner (Sui et al., 2025)	✗	Prompt	GPT-4o, GPT-4o-mini, Gemini-Exp-1206	Math, Scientific
	SOLAR (Li et al., 2025)	✓	SFT	Qwen2VL-7B-Instruct	Math
	C3oT (Kang et al., 2024)	✓	SFT	LLaMA-2-Chat -7B & -13B	Math, Commonsense
	TokenSkip (Xia et al., 2025)	✓	SFT	LLaMA-3.1-8B-Instruct, Qwen2.5- 14B-Instruct	Math
	InftyThink (Yan et al., 2025)	✓	SFT	Qwen2.5-14B/32B, Qwen2.5-Math-1.5B/7B, LLaMA-3.1-8B	Math, Scientific
	LightThinker (Zhang et al., 2025)	✓	SFT	DeepSeek-R1-Distill-Qwen-7B, DeepSeek-R1-Distill-LLaMA-8B	Language Understanding, Math, Scientific, Commonsense, Logic
	CoT-Valve (Ma et al., 2025)	✓	SFT	QwQ-32B-Preview, DeepSeek-R1-Distill-LLaMA-8B, LLaMA-3.1-8B, LLaMA-3.2-1B, Qwen32B-Instruct	Math
	Distill System 2 (Yu et al., 2024)	✓	SFT	LLaMA-2-70B-chat	Math, Commonsense, Coin Flip
	SF (Munkhbat et al., 2025)	✓	SFT	LLaMA-3.2-3B, Gemma2-2B , Qwen2.5-3B , Qwen2.5-Math-1.5B, DeepSeekMath-7B	Math
	Skip Steps (Liu et al., 2024c)	✓	SFT	LLaMA2-7b, Phi-3-mini	Math, Logic
	VARR (Jang et al., 2024)	✓	SFT	Mistral 7B, Llama3.2 1B/3B	Math, Commonsense
	DAST (Shen et al., 2025b)	✓	SimPO	DS-R1-Distill-Qwen-7B, DS-R1-Distill-Qwen-32B	Math
	TALE-PT (Han et al., 2024)	✓	SFT, DPO	LLaMA-3.1-8B-Instruct	Math
	Kimi k1.5 (Kimi Team et al., 2025)	✓	RL	Kimi k1.5	Multimodal Understanding, Math, Code
	O1-Pruner (Luo et al., 2025)	✓	RL	Marco-o1-tB, QwQ-32B	Math
	MRT (Qu et al., 2025)	✓	RL	DeepSeek-R1-Distill-Qwen-32B	Math
	(Arora & Zanette, 2025)	✓	RL	DS-R1-Distill-Qwen-1.5B, DS-R1-Distill-Qwen-7B	Math
	Claude 3.7 (Anthropic, 2025)	✓	RL	Unknown	Math, Code, Agent
	L1 (Aggarwal & Welleck, 2025)	✓	RL	Qwen-Distilled-R1-1.5B	Language Understanding, Logic, Math
	SPIRIT (Cui et al., 2025)	✓	RL	LLaMA3-8B-Instruct, Qwen2.5- 7B-Instruct	Math
	IBPO (Yu et al., 2025)	✓	RL	LLaMA-3.1-8B	Math
Implicit Latent CoT	ICoT-KD (Deng et al., 2023)	✓	SFT	GPT-2 Small/Medium	Math
	CODI (Shen et al., 2025c)	✓	SFT	GPT-2 Small, LLaMA-3.2-1B	Math
	ICoT-SI (Deng et al., 2024)	✓	SFT	GPT-2 Small/Medium, Phi-3 3.8B, Mistral 7B	Math
	COCONUT (Hao et al., 2024)	✓	SFT	GPT-2	Math
	CCoT (Cheng & Van Durme, 2024)	✓	SFT	LLaMA2-7B-Chat	Math, Logic
	Heima (Shen et al., 2025a)	✓	SFT	LLaVA-CoT, LLaMA-3.1-8B-Instruct	Multimodal Reasoning
	Token assorted (Su et al., 2025)	✓	SFT	LLaMA-3.2-1B, LLaMA-3.2-3B, LLaMA-3.1-8B	Agentic Planning, Logic, Math.
	SoftCoT (Xu et al., 2025c)	✓	SFT	LLaMA-3.1-8B-Instruct, Qwen2.5-7B-Instruct	Math, Commonsense, Symbolic Reasoning

Table 1: A taxonomy of efficient inference methods for Large Reasoning Models.

2.1 Large Reasoning Model

Large Reasoning Models (LRMs) extend the capabilities of Large Language Models (LLMs) by incorporating explicit intermediate tokens that represent reasoning processes, enabling more structured logical reasoning and effective complex problem-solving. LRMs mimic the way humans approach complex problems by first **thinking** before providing an **answer**. When faced with a difficult question, they do not immediately respond with an answer; instead, they analyze the problem, break it down into smaller steps, explore different solution paths, and verify their reasoning before arriving at a conclusion. The o1 series (Jaech et al., 2024) from OpenAI, released in late 2024, marked a significant breakthrough in AI reasoning capabilities, which integrates reinforcement learning and "Chain of-Thought" prompting (Wei et al., 2022) techniques. Following this, OpenAI released o3 (OpenAI, 2025), an upgraded version of o1, allowing it to achieve PhD-level performance in mathematics, science, and programming. Notable DeepSeek's R1 (Guo et al., 2025) stands out for being open-sourced, with transparent thinking process tokens, which sets it apart from other proprietary LRMs like o1/o3, where the internal reasoning steps are less accessible. However, since LRMs need to generate numerous intermediate thinking tokens before arriving at final answers, they are significantly less efficient and more expensive compared to regular LLMs. This added complexity in processing demands more computational resources and time.

2.2 Efficiency Challenge in LRM Inference

A key driver of LRMs' remarkable reasoning capabilities is the scaling of inference-time compute, which enables complex reasoning through long CoTs (Chen et al., 2025; Guo et al., 2025; Jaech et al., 2024; Muennighoff et al., 2025; Liu et al., 2025a). Compared to standard short CoTs (Wei et al., 2022), which are often shallow, heuristic-driven, and less generalizable (Sprague et al., 2025), long CoTs empower LRMs to tackle complex tasks such as advanced mathematics (Xu et al., 2025a) and medical question answering (Huang et al., 2025). However, this shift has also introduced the phenomenon of overthinking, where LRMs consume excessive inference tokens and reasoning steps even for simple problems, yielding only marginal performance improvements (Ma et al., 2024; Chen et al., 2024; Wu et al., 2025c). In real-world applications such as software engineering agents, overthinking has been found to negatively correlate with issue resolution rates (Cuadron et al., 2025). Moreover, LRMs' reliance on inference-time scaling exposes them to overthinking attacks, where adversarial actors inject benign yet computationally intensive decoy problems (e.g., Sudoku puzzles) into the context for retrieval-augmented question answering, triggering substantial computational overhead (Kumar et al., 2025).

Toward practical real-world deployment, optimizing the token efficiency of LRMs without compromising effectiveness remains an underexplored challenge. This paper presents a systematic investigation into recent advances in token-efficient LRMs, examining their underlying approaches, empirical effectiveness, and implications for future research.

3 Landscape of LRM Efficient Inference Research

This section surveys the current landscape of research on token-efficient LRM inference, which can be broadly categorized into two approaches: (1) **explicit compact CoT**, where explicit instructions, rewards, or budget constraints are introduced to encourage shorter reasoning chains over long CoTs (Section 3.1); and (2) **implicit latent CoT**, which compresses explicit long CoTs into compact, continuous reasoning states (Section 3.2). The taxonomy of recent efficient inference methods is shown in Table 1.

3.1 Explicit Compact CoT

Recent research has focused on developing methods to create more compact reasoning paths while preserving accuracy through various techniques, including (1) **CoT compression**, (2) **fine-tuning for compact reasoning**, and (3) **reward-based incentivization**.

Takeaways of Explicit Compact CoT

1. CoT compression enhances scalability but may sacrifice transparency. These techniques lower token usage by abstracting reasoning steps, but risk omitting essential intermediate logic, which can undermine interpretability.

2. Supervised fine-tuning improves efficiency, but at high cost. While effective, these methods depend on curated, condensed datasets and heavy preprocessing, limiting their adaptability to open-ended domains.

3. Reward-based brevity can lead to shallow reasoning. Incentivizing shorter outputs may cause models to favor simplistic answers, at the expense of the deeper reasoning needed for complex tasks.

4. Efficiency alone is insufficient for real-world deployment. Real-world applications require a balance between compactness and reasoning robustness, interpretability, and domain generalization.

CoT Compression. Succinct CoT representations streamline inference while preserving solution quality. Constrained-CoT (Nayab et al., 2024) and CoD (Xu et al., 2025b) confine intermediate reasoning to essential steps, ensuring brevity without losing critical information. Beyond simple compression, Sketch-of-Thought (SoT) (Aytes et al., 2025) uses a smaller "router" model to prompt the main LLM to generate sketches of reasoning, offering a concise yet cognitively inspired overview. InftyThink (Yan et al., 2025) decomposes complex tasks into bounded-length segments, creating intermediate summaries at each step. Other frameworks adapt their compression mechanisms in real time. LightThinker (Zhang et al., 2025) introduces special tokens that trigger the model to dynamically compress its ongoing thought process, reducing redundancy. TALE-EP (Han et al., 2024) dynamically adjusts the allotted reasoning tokens depending on task complexity, while Meta-Reasoner (Sui et al., 2025) applies a contextual multi-armed bandit to optimize efficiency.

Fine-Tuning on Compact Reasoning Chains. Fine-tuning on compact reasoning data enables LRMs to internalize efficient inference behaviors while keeping performance. C3oT (Kang et al., 2024) leverages an LLM to generate condensed versions of long CoTs, preserving essential structure before jointly training models on both full and compressed chains. Skip Steps (Liu et al., 2024c) curates expert-validated answers with condensed steps and fine-tunes LLMs to mimic these concise reasoning paths. SOLAR (Li et al., 2025) fine-tunes LLMs using datasets annotated for both correctness and the effectiveness of the underlying task-specific reasoning topology, encouraging minimal yet complete logic flows. VARR (Jang et al., 2024) identifies redundant sentences in CoTs by analyzing their contribution to answer correctness, then fine-tunes models on the distilled, non-redundant reasoning processes. TokenSkip (Xia et al., 2025) prunes reasoning chains token-by-token based on importance, followed by fine-tuning across various compression ratios to balance brevity and precision. From a parameter space perspective, TALE-EP (Han et al., 2024) enhances token-budget awareness via SFT and direct preference optimization (DPO). CoT-Valve (Ma et al., 2025) discovers a latent direction that controls reasoning length, enabling models to flexibly adjust their level of detail based on task demands.

Reward-Based Incentivization. A growing body of work introduces explicit reward signals to reduce unnecessary CoT complexity while preserving accuracy. Kimi k1.5 (Kimi Team et al., 2025) integrates length-based rewards to discourage verbose reasoning. Similarly, O1-Pruner (Luo et al., 2025) detects "length disharmony" and applies harmonizing penalties that promote brevity without sacrificing solution quality. Arora et al. (Arora & Zanette, 2025) use reinforcement learning to train models that dynamically allocate computational resources based on task difficulty, balancing cost and precision. DAST (Shen et al., 2025b) proposes a Token Length Budget metric that aligns task complexity with output length, encouraging efficiency through targeted penalties and rewards. IBPO (Yu et al., 2025) adopts a constrained RL framework to control the distribution of reasoning across response groups based on inference cost. MRT (Qu et al., 2025) applies meta-reinforcement learning to balance exploration of novel reasoning paths with the exploitation of concise, proven ones. Recent approaches also explore interactive or user-directed mechanisms for length control. Claude 3.7 (Anthropic, 2025), the first hybrid reasoning model, introduces an extended thinking mode where users can prescribe token budgets. L1 (Aggarwal & Welleck, 2025) generalizes this idea with Length Controlled Policy Optimization (LCPO), enabling fully configurable CoT lengths at inference time.

3.2 Implicit Latent CoT

Implicit latent CoT methods enhance token efficiency by shifting reasoning **from explicit tokens to latent tokens**, encoding reasoning in hidden layers rather than natural language.

Takeaways of Implicit Latent CoT

1. Implicit latent CoT improves efficiency by internalizing reasoning steps but sacrifices interpretability, making verification difficult.

2. Different methods (e.g., knowledge distillation, latent embeddings, contemplation tokens) optimize reasoning at various levels, reducing latency while maintaining accuracy.

3. Future work should focus on extracting human-interpretable reasoning traces from latent representations to balance efficiency and transparency.

A line of knowledge distillation methods (Deng et al., 2023; 2024; Shen et al., 2025c) trains student models to infer the teacher's internal CoT representations rather than mimic explicit token sequences, enabling "vertical" reasoning across transformer layers. Chain of Continuous Thought (COCONUT)(Hao et al., 2024) replaces token-level reasoning chains with autoregressively generated latent embeddings, which are fed back into the model to emulate breadth-first search during problem-solving. Compressed CoT (CCoT)(Cheng & Van Durme, 2024) introduces contemplation tokens—dense, compressed representations of full reasoning chains—significantly reducing inference latency while maintaining accuracy.

Types	Methods	Setting	Accurracy	Model	Token Cost
Explicit Compact CoT	CoD (Xu et al., 2025b)	zero-shot	84.40%	GPT-4o	76.40
		zero-shot	65.50%	Claude 3.5 Sonnet	73.70
		few-shot	91.10%	GPT-4o	43.90
		few-shot	91.40%	Claude 3.5 Sonnet	39.80
	TALE (Han et al., 2024)	zero-shot, prompt	84.46%	GPT-4o-mini	77.26
		zero-shot, SFT	74.11%	LLaMA-3.1-8B-Instruct	149.93
		zero-shot, DPO	78.41%	LLaMA-3.1-8B-Instruct	113.41
	C3oT (Kang et al., 2024)	zero-shot	36.92%	LLaMA-2-Chat-7B	-
		zero-shot	47.10%	LLaMA-2-Chat-13B	-
	TokenSkip (Xia et al., 2025)	zero-shot, ratio=0.5	86.70%	LLaMA-3.1-8B-Instruct	113.05
		zero-shot, ratio=0.6	86.10%	LLaMA-3.1-8B-Instruct	198.01
		zero-shot, ratio=0.7	84.30%	LLaMA-3.1-8B-Instruct	169.89
		zero-shot, ratio=0.8	82.50%	LLaMA-3.1-8B-Instruct	150.12
		zero-shot, ratio=0.9	81.10%	LLaMA-3.1-8B-Instruct	129.38
		zero-shot, ratio=1.0	78.20%	LLaMA-3.1-8B-Instruct	113.05
	LightThinker (Zhang et al., 2025)	zero-shot,tho.	90.14%	DeepSeek-R1-Distill-Qwen-7B	-
		zero-shot,token	87.11%	DeepSeek-R1-Distill-Qwen-7B	-
		zero-shot,tho.	88.25%	DeepSeek-R1-Distill-LLaMA-8B	-
		zero-shot,tho.	85.52%	DeepSeek-R1-Distill-LLaMA-8B	-
	SF (Munkhbat et al., 2025)	zero-shot	76.72%	DeepSeekMath-7B	184.13
	O1-Pruner (Luo et al., 2025)	few-shot	96.50%	QwQ-32B	343.00
Implicit Latent CoT	ICoT-KD (Deng et al., 2023)	zero-shot	45.00%	GPT-2 Medium	-
	CODI (Shen et al., 2025c)	zero-shot	55.60%	LLaMA-3.2-1B	-
	ICoT-SI (Deng et al., 2024)	zero-shot	51.00%	Mistral 7B	-
	COCONUT (Hao et al., 2024)	zero-shot	34.10%	GPT-2	8.20
	CCoT (Cheng & Van Durme, 2024)	zero-shot	31.50%	LLaMA2-7B-Chat	-
	Token assorted (Su et al., 2025)	zero-shot	37.20%	LLaMA-3.1-8B	-
	SoftCoT (Xu et al., 2025c)	zero-shot	85.81%	Qwen2.5-7B-Instruct	-

Table 2: Benchmarking on recent reasoning efficient methods on GSM8K dataset.

Heima(Shen et al., 2025a) condenses CoT stages into latent thinking tokens and incorporates an explanatory prompt at the decoder stage to interpret the compressed reasoning. SoftCoT (Xu et al., 2025c) utilizes a small instruction-tuned 1B model to obtain instance-specific latent thought tokens and trains a projection layer to incorporate thought tokens into LLM input. Token-Assorted CoT(Su et al., 2025) mixes latent and text tokens, encoding the initial part of the CoT into VAE-based discrete latent tokens while preserving the remainder as natural language, resulting in a hybrid representation that enhances reasoning efficiency.

While their implementations vary, these approaches share a common goal: optimizing inference by internalizing the reasoning process. Empirical results suggest that implicit latent CoT models can match or even surpass explicit CoT methods in reasoning accuracy while significantly reducing generation costs, demonstrating their scalability and efficiency.

4 Empirical Analyses

This section conducts empirical analyses of the existing reasoning efficient methods from the perspectives of performance and token efficiency. We summarize the used benchmarks of the existing methods and categorize them into 10 reasoning scenarios as follows. **(1) Mathematical Reasoning** includes GSM8K (Cobbe et al., 2021), GSM8K-Zero (Chiang & Lee, 2024), SVAMP (Patel et al., 2021), AQuA Ling et al. (2017), ASDiv (Miao et al., 2020), MathBench (Liu et al., 2024b), TheoremQA (Chen et al., 2023), MATH (Hendrycks et al., 2021), MathQA Yu et al. (2023), AIME24 (Olympiad, 2025), Olympiad-Bench (He et al., 2024a), GPQA (Rein et al., 2024). **(2) Causal Reasoning** includes QASC (Khot et al., 2019), WorldTree (Jansen et al., 2018) **(3) Code Reasoning** includes LiveCodeBench (Jain et al., 2024), Codeforces, SWE-bench (Jimenez et al., 2023). **(4) Logical Reasoning** includes ProntoQA (Saparov & He, 2023), LogiQA (Liu et al., 2020), Reclor (Yu et al., 2020). **(5) Symbolic Reasoning** includes CoinFlip (Wei et al., 2022). **(6) Commonsense Reasoning** includes CommonsenseQA (Talmor et al., 2019), OpenbookQA (Mihaylov et al., 2018), ECQA (Aggarwal et al., 2021), StrategyQA (Geva et al., 2021). **(7) General Reasoning** includes BIG-Bench (Srivastava et al., 2022), BIG-Bench Hard (Suzgun et al., 2022), HotPotQA (Yang et al., 2018), MuSiQue (Trivedi et al., 2022), MMLU (Hendrycks et al., 2020), MMMLU (Yue et al., 2024), ScienceQA (Lu et al., 2022), SciBench (Wang et al., 2024c). **(8) Visual Reasoning** MMMU (Yue et al., 2024), MATH-Vision (Wang et al., 2024a), MathVista (Lu et al., 2024) **(9)**

Agent Reasoning includes TAU-bench (Yao et al., 2024), Keys-Finding Maze (Su et al., 2025). **(10) Task-specific Reasoning** includes PubMedQA (Jin et al., 2019). Then, we demonstrate the performance and token costs of the existing methods on one common dataset GSM8K (Cobbe et al., 2021). We list these results in Table 2.

5 Limitations and Challenges

We discuss the limitations and challenges of the existing reasoning efficient methods from the perspectives of **user experience**, **interpretability**, **safety**, and **application**.

Takeaways of Limitations Challenges

1. User-controllable reasoning enables users to adjust reasoning depth, striking a balance between transparency and efficiency while optimizing user experience. Future research should focus on interactive and personalized reasoning for users.

2. Efficient reasoning methods may obscure crucial reasoning processes, compromising the interpretability of LRMs. Future research should develop adaptive inference strategies to balance efficiency and interpretability.

3. Efficient reasoning methods may compromise safety alignment, increasing the risk of jailbreaking and privacy leakage. Future work should integrate safety constraints in training and develop stronger reasoning-based safeguards.

4. Efficient reasoning methods may improve the feasibility of LRMs for broader applications like real-time applications and open-ended tasks.

5.1 User-centric Controllable Reasoning

Recent advancements in LRMs, such as OpenAI's o3 (OpenAI, 2025) and Anthropic's Claude 3.7 (Anthropic, 2025), have introduced **user-configurable reasoning modes**, allowing users to choose whether the model engages in explicit reasoning or provides direct answers. Additionally, these models enable users to control the complexity and length of the reasoning process, adapting to different needs and preferences.

This control level is particularly useful in diverse applications, e.g., in **educational settings**, users may prefer step-by-step explanations for questions, whereas in **real-time decision-making tasks**, concise responses are more desirable. Allowing users to adjust reasoning depth enables LRMs to balance efficiency and transparency, enhancing user experience.

Future research should explore more refined control mechanisms, such as **interactive reasoning settings** that dynamically adjust based on user feedback. Besides, developing personalized reasoning profiles could allow LRMs to learn and adapt to individual preferences over time, providing a balance between reasoning depth, speed, and interpretability.

5.2 Trade-off Between Interpretability and Efficiency of Reasoning

Compared to LLMs, LRMs offer significantly better **interpretability** due to their structured reasoning process. By explicitly generating intermediate reasoning steps, LRMs allow users to trace how a conclusion is reached, making them particularly valuable for applications where **transparency** and verifiability are critical, such as **scientific research** (Rane et al., 2023), **medical diagnosis** (Ullah et al., 2024), and **legal decision-making** (Cheong et al., 2024). However, current efficiency-focused LRMs may compromise this interpretability. Many recent methods designed to accelerate LRM inference reduce the number of explicit reasoning steps or shift reasoning to latent representations, making it harder to understand how a model arrives at its conclusions.

Also, the importance of interpretability varies depending on the application. In domains such as healthcare and legal reasoning, where explanations are essential for accountability and human oversight, explicit reasoning steps are preferred despite their computational cost. Conversely, in real-time decision-making tasks, such as automated trading or robotics, efficiency often takes precedence over transparency, making implicit reasoning more desirable. Hybrid approaches, which dynamically adjust the level of explicit reasoning based on task complexity, offer a potential solution but require further refinement to prevent critical reasoning steps from being lost in the pursuit of efficiency.

To address this trade-off more effectively, future research should focus on developing adaptive inference strategies that optimize the balance between reasoning efficiency and interpretability. One promising direction is the integration of **external verification mechanisms**, such as symbolic reasoning (Besold et al., 2021; Gaur & Saunshi, 2023; Sui et al., 2024) or retrieval-based justifications (Gao et al., 2023), which can provide post-hoc explanations for implicit reasoning models. Besides, new empirical studies are needed to quantify how different efficiency techniques impact both model accuracy and human trust, guiding the development of LRMs that are both efficient and interpretable in real-world scenarios.

5.3 Ensuring Safety of Efficient Reasoning

Although the existing methods improve the token efficiency of the LRMs, they may destroy the alignment of LRMs, increasing the potential safety risks, e.g., **jailbreaking attacks** (Liu et al., 2024d; He et al., 2025a) and **privacy leakage** (Li et al., 2023).

Firstly, the current training-based token-efficient methods either train the LRMs to prefer shorter generations (Kang et al., 2024; Han et al., 2024) or adopt RL and incentivize concise responses via rule-based reward (Qu et al., 2025; Luo et al., 2025; Kimi Team et al., 2025). Since the safety alignment is conducted on the original long reasoning generations and the safety of the shorter reasoning generations can not be guaranteed, these training processes might **break the safety alignment** of the original LRMs.

Secondly, as one piece of evidence, researchers (OpenAI, 2025c) found that the frontier LRMs tend to exploit the loopholes once they get a chance. In addition, although they tried to use another LLM to monitor the intermediate CoT, penalizing their misbehavior can not effectively alleviate this problem but further guide them to **hide their misconduct intent**. From this phenomenon, we suspect that the existing token-efficient methods unintentionally guide the LRMs to hide their harmful intent during the process of making their response more concise, increasing the difficulty of safeguarding LRMs.

To address this problem, one promising direction is to add **safety constraints** during the training process, like data filtering for the SFT/DPO data and designing the safety-related reward in RL training. Besides, the failure of current monitors may be due to LRMs' ability being stronger than LLM-based guard models. Thus, it is worth designing stronger **reasoning-based safeguard models** (Liu et al., 2025b) to monitor the training data or LRMs.

5.4 Broader Application of Efficient Reasoning

As shown in Table 1, existing LRMs are primarily applied in **math** (Xia et al., 2025; Li et al., 2025), **code** (Kimi Team et al., 2025), or **AI research** (OpenAI, 2025a) scenarios.

The first reason is that these tasks have relatively fixed answers, making it easier to construct objectives, e.g., preparing reasoning data, formulating preference loss functions, or rule-based rewards. In contrast, other domains, like **social sciences** (Manning et al., 2024; Thapa et al., 2025), **emotional intelligence** (Wu et al., 2025b), **creative writing** (OpenAI, 2025b), typically involve open-ended questions, making it difficult to formulate clear objectives.

The second reason is that these scenarios, like math or research, are not highly time-sensitive, allowing for more computational resources to be allocated for reasoning and optimization. The high computational demand and latency of LRMs constrain their applicability in broader time-sensitive domains, such as **robotic manipulations** (Ji et al., 2025; Google, 2025; Nvidia, 2025), **financial trading** (Ding et al., 2024), **autonomous driving** (Yang et al., 2023).

However, recently developed efficient reasoning methods (Cheng & Van Durme, 2024; Anthropic, 2025; Kimi Team et al., 2025) help LRMs reduce thinking tokens, optimize timing and memory usage, and thus **enhance feasibility in real-time applications**. For the open-ended questions, efficient reasoning methods enable LRMs to generate more structured and consistent responses while **balancing interpretability and computational cost**.

6 How can the inference efficiency of LRMs be further improved?

While the existing methods are effective, we present alternative strategies that could further improve inference efficiency while maintaining the reasoning quality, including **new architectures**, **model merge**, and **agent routers**.

Takeaways of Further Improvement

1. New architectures, including hybrid autoregressive and diffusion models, memory-efficient transformers, and graph-based reasoning, are potential techniques for further improving reasoning efficiency while preserving reasoning quality.

2. Merging model weights offers a promising solution for reasoning efficiency of LRMs, with challenges in module selection, weight assignment, and architecture compatibility.

3. Agent routers, which dynamically allocate the resources for different LRMs based on the difficulty of tasks, could be a practical direction for the reasoning efficiency.

6.1 New Architecture

Hybrid Autoregressive and Diffusion Models. The fundamental limitation of autoregressive models is their sequential nature, which makes inference slow, particularly for reasoning tasks that require long chains of intermediate steps. A potential solution is integrating **diffusion models** into LRMs (Nie et al., 2025). Diffusion models generate entire sequences in parallel, allowing for global reasoning structure optimization rather than token-by-token generation. But the challenge lies in controlling the generated reasoning steps to ensure logical consistency. A promising direction is hybrid architectures that use autoregression for fine-grained control over reasoning while leveraging **diffusion-based sampling** for efficiency, enabling LRMs to reason in a structured yet accelerated manner.

Memory-Efficient Transformer Variants. One of the primary inefficiencies in LRMs stems from the **quadratic complexity** of self-attention. Applying **linear attention** mechanisms(e.g., RWKV (Peng et al., 2023)) or **state-space** models (e.g., Mamba (Gu & Dao, 2023)) could drastically reduce memory consumption and improve inference speed. The challenge is that such architectures often struggle with **long-range dependencies**, which are crucial for reasoning. A key question is whether **hybrid models** can selectively apply full attention for critical reasoning steps while using approximate attention elsewhere to optimize efficiency.

Graph-Based Reasoning Models. Autoregressive LRMs process information sequentially, generating one token at a time. While effective, this approach struggles with complex **multi-step reasoning** tasks where different pieces of information must be retrieved, combined, and reasoned over in a structured manner. Graph-based reasoning models (Besta et al., 2024) offer a structured alternative to autoregressive LRMs by representing reasoning as a graph, where nodes encode intermediate steps and edges define logical dependencies. This enables **parallel exploration** of multiple reasoning paths, reducing inefficiencies inherent in sequential token generation. **Monte Carlo Tree Search (MCTS)** (Silver et al., 2017) enhances this by adaptively expanding promising reasoning branches while pruning redundant ones, optimizing both token efficiency and inference time. However, key challenges include designing training objectives for graph construction, optimizing traversal heuristics, and balancing interpretability with efficiency. Future research should explore hybrid architectures that fuse explicit graph reasoning with neural representations (He et al., 2024b; 2025b; Zhou et al., 2023), ensuring both scalability and robust logical coherence.

6.2 Model Merge

The underlying principle of the existing token-efficient methods can be summarized as integrating the strength of the conventional LLMs, i.e., **fast responses and low costs**, and the strength of the LRMs, i.e., **deliberative reasoning and accurate responses**.

The existing training-based methods (Han et al., 2024; Luo et al., 2025; Qu et al., 2025) typically involve reasoning data curation and post-training techniques such as SFT, DPO, or RL, making the process **complex and expensive**. On the other hand, the existing training-free methods (Sui et al., 2025) typically just use promoting engineering to guide the LRMs to save the tokens, **limiting the adaptability and effectiveness** across diverse reasoning tasks.

To solve this problem, another training-free method model merge (Yang et al., 2024; Wu et al., 2025a) becomes a promising technique. Concretely, we can simply **merge the model weights** of one conventional LLM and the corresponding LRM to take their advantages together (Kimi Team et al., 2025). During this process, we provide several key points that need to be solved in the future. First, we need to **determine which modules or neurons in models should be merged**. Should we merge the neurons in shallow networks or deep networks? Then, we should **assign merging weights for the merging units**. Should we assign static or dynamic weights for each unit? Third, we should consider **how to merge models with different architectures and model sizes**, e.g., LLaMA-3.1 Instruct 8B (Grattafiori et al., 2024) and DeepSeekR1-Distill-Qwen-7B (Guo et al., 2025).

6.3 Agent Router

Agent routing could further improve efficiency by directing different parts of a query to specialized agents. By routing the **query to the most appropriate agent** based on task complexity, this strategy would optimize resource usage and enable faster inference, particularly for tasks that require domain-specific knowledge or specialized reasoning.

There are two existing routing strategies: one based on router models and the other on metrics like confidence. Routellm (Ong et al., 2025) introduces several efficient **router models** that dynamically select either a stronger or a weaker LLM during inference to balance cost and response quality. Self-REF (Chuang et al., 2025b) routes LLMs by training them to reliably **express their confidence** in the correctness of their answers. Additionally, Chuang et al. (2025a) explores an **uncertainty-based small language model** (SLM) routing approach, where high-stakes queries are offloaded to more robust LLMs whenever the SLM produces low-confidence responses. These methods offer exciting possibilities for enhancing inference efficiency in LRMs, particularly in reducing computation time and resource use while maintaining high performance across a range of tasks.

7 Final Remarks

This survey provides an overview of efficient inference techniques for large reasoning models, highlighting the challenges and recent advancements in this area. As reasoning models continue to grow in scale, the computational cost of inference becomes a major bottleneck, necessitating methods that improve efficiency while maintaining performance. We categorized existing approaches, discussing their trade-offs and practical implications. We hope this survey provides a foundation for further research in this area, encouraging the development of more effective and computationally feasible reasoning models.

References

Pranjal Aggarwal and Sean Welleck. L1: Controlling how long a reasoning model thinks with reinforcement learning. *arXiv preprint arXiv:2503.04697*, 2025.

Shourya Aggarwal, Divyanshu Mandowara, Vishwajeet Agrawal, Dinesh Khandelwal, Parag Singla, and Dinesh Garg. Explanations for CommonsenseQA: New Dataset and Models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Online, 2021. Association for Computational Linguistics.

Anthropic. Claude 3.7 sonnet and claude code. https://www.anthropic.com/news/claude-3-7-sonnet, 2025.

Daman Arora and Andrea Zanette. Training language models to reason efficiently. *arXiv preprint arXiv:2502.04463*, 2025.

Simon A Aytes, Jinheon Baek, and Sung Ju Hwang. Sketch-of-thought: Efficient llm reasoning with adaptive cognitive-inspired sketching. *arXiv preprint arXiv:2503.05179*, 2025.

Tarek R Besold, Artur d'Avila Garcez, Sebastian Bader, Howard Bowman, Pedro Domingos, Pascal Hitzler, Kai-Uwe Kühnberger, Luis C Lamb, Priscila Machado Vieira Lima, Leo de Penning, et al. Neural-symbolic learning and reasoning: A survey and interpretation 1. In *Neuro-Symbolic Artificial Intelligence: The State of the Art*, pp. 1–51. IOS press, 2021.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, et al. Graph of thoughts: Solving elaborate problems with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 17682–17690, 2024.

Qiguang Chen, Libo Qin, Jinhao Liu, Dengyun Peng, Jiannan Guan, Peng Wang, Mengkang Hu, Yuhang Zhou, Te Gao, and Wangxiang Che. Towards reasoning era: A survey of long chain-of-thought for reasoning large language models. *arXiv preprint arXiv:2503.09567*, 2025.

Wenhu Chen, Ming Yin, Max Ku, Pan Lu, Yixin Wan, Xueguang Ma, Jianyu Xu, Xinyi Wang, and Tony Xia. TheoremQA: A theorem-driven question answering dataset. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 7889–7901, 2023.

Xingyu Chen, Jiahao Xu, Tian Liang, Zhiwei He, Jianhui Pang, Dian Yu, Linfeng Song, Qiuzhi Liu, Mengfei Zhou, Zhuosheng Zhang, et al. Do not think that much for 2+ 3=? on the overthinking of o1-like llms. *arXiv preprint arXiv:2412.21187*, 2024.

Jeffrey Cheng and Benjamin Van Durme. Compressed chain of thought: Efficient reasoning through dense representations. *arXiv preprint arXiv:2412.13171*, 2024.

Inyoung Cheong, King Xia, KJ Kevin Feng, Quan Ze Chen, and Amy X Zhang. (a) i am not a lawyer, but...: engaging legal experts towards responsible llm policies for legal advice. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 2454–2469, 2024.

Cheng-Han Chiang and Hung-yi Lee. Over-reasoning and redundant calculation of large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 161–169, 2024.

Yu-Neng Chuang, Leisheng Yu, Guanchu Wang, Lizhe Zhang, Zirui Liu, Xuanting Cai, Yang Sui, Vladimir Braverman, and Xia Hu. Confident or seek stronger: Exploring uncertainty-based on-device llm routing from benchmarking to generalization. *arXiv preprint arXiv:2502.04428*, 2025a.

Yu-Neng Chuang, Helen Zhou, Prathusha Kameswara Sarma, Parikshit Gopalan, John Boccio, Sara Bolouki, and Xia Hu. Learning to route llms with confidence tokens. *arXiv preprint arXiv:2410.13284*, 2025b.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Alejandro Cuadron, Dacheng Li, Wenjie Ma, Xingyao Wang, Yichuan Wang, Siyuan Zhuang, Shu Liu, Luis Gaspar Schroeder, Tian Xia, Huanzhi Mao, et al. The danger of overthinking: Examining the reasoning-action dilemma in agentic tasks. *arXiv preprint arXiv:2502.08235*, 2025.

Yingqian Cui, Pengfei He, Jingying Zeng, Hui Liu, Xianfeng Tang, Zhenwei Dai, Yan Han, Chen Luo, Jing Huang, Zhen Li, et al. Stepwise perplexity-guided refinement for efficient chain-of-thought reasoning in large language models. *arXiv preprint arXiv:2502.13260*, 2025.

Yuntian Deng, Kiran Prasad, Roland Fernandez, Paul Smolensky, Vishrav Chaudhary, and Stuart Shieber. Implicit chain of thought reasoning via knowledge distillation. *arXiv preprint arXiv:2311.01460*, 2023.

Yuntian Deng, Yejin Choi, and Stuart Shieber. From explicit cot to implicit cot: Learning to internalize cot step by step. *arXiv preprint arXiv:2405.14838*, 2024.

Han Ding, Yinheng Li, Junhao Wang, and Hang Chen. Large language model agent in financial trading: A survey. *arXiv preprint arXiv:2408.06361*, 2024.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Haofen Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2, 2023.

Vedant Gaur and Nikunj Saunshi. Reasoning in large language models through symbolic math word problems. *arXiv preprint arXiv:2308.01906*, 2023.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. Did Aristotle Use a Laptop? A Question Answering Benchmark with Implicit Reasoning Strategies. *Transactions of the Association for Computational Linguistics (TACL)*, 2021.

Google. Gemini robotics brings ai into the physical world. https://deepmind.google/discover/blog/gemini-robotics-brings-ai-into-the-physical-world/, 2025.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Tingxu Han, Chunrong Fang, Shiyu Zhao, Shiqing Ma, Zhenyu Chen, and Zhenting Wang. Token-budget-aware llm reasoning. *arXiv preprint arXiv:2412.18547*, 2024.

Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason Weston, and Yuandong Tian. Training large language models to reason in a continuous latent space. *arXiv preprint arXiv:2412.06769*, 2024.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, Jie Liu, Lei Qi, Zhiyuan Liu, and Maosong Sun. OlympiadBench: A challenging benchmark for promoting AGI with olympiad-level bilingual multimodal scientific problems. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3828–3850. Association for Computational Linguistics, 2024a.

Yufei He, Yuan Sui, Xiaoxin He, and Bryan Hooi. Unigraph: Learning a unified cross-domain foundation model for text-attributed graphs. *arXiv preprint arXiv:2402.13630*, 2024b.

Yufei He, Yuexin Li, Jiaying Wu, Yuan Sui, Yulin Chen, and Bryan Hooi. Evaluating the paperclip maximizer: Are rl-based language models more likely to pursue instrumental goals? *arXiv preprint arXiv:2502.12206*, 2025a.

Yufei He, Yuan Sui, Xiaoxin He, Yue Liu, Yifei Sun, and Bryan Hooi. Unigraph2: Learning a unified embedding space to bind multimodal graphs. *arXiv preprint arXiv:2502.00806*, 2025b.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *NeurIPS*, 2021.

Zhongzhen Huang, Gui Geng, Shengyi Hua, Zhen Huang, Haoyang Zou, Shaoting Zhang, Pengfei Liu, and Xiaofan Zhang. O1 replication journey–part 3: Inference-time scaling for medical reasoning. *arXiv preprint arXiv:2501.06458*, 2025.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. Livecodebench: Holistic and contamination free evaluation of large language models for code. *arXiv preprint arXiv:2403.07974*, 2024.

Joonwon Jang, Jaehee Kim, Wonbin Kweon, and Hwanjo Yu. Verbosity-aware rationale reduction: Effective reduction of redundant rationale via principled criteria. *arXiv preprint arXiv:2412.21006*, 2024.

Peter Jansen, Elizabeth Wainwright, Steven Marmorstein, and Clayton Morrison. WorldTree: A corpus of explanation graphs for elementary science questions supporting multi-hop inference. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

Yuheng Ji, Huajie Tan, Jiayu Shi, Xiaoshuai Hao, Yuan Zhang, Hengyuan Zhang, Pengwei Wang, Mengdi Zhao, Yao Mu, Pengju An, et al. Robobrain: A unified brain model for robotic manipulation from abstract to concrete. *arXiv preprint arXiv:2502.21257*, 2025.

Carlos E Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. Swe-bench: Can language models resolve real-world github issues? *arXiv preprint arXiv:2310.06770*, 2023.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. PubMedQA: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2567–2577, 2019.

Yu Kang, Xianghui Sun, Liangyu Chen, and Wei Zou. C3ot: Generating shorter chain-of-thought without compromising effectiveness. *arXiv preprint arXiv:2412.11664*, 2024.

Tushar Khot, Peter Clark, Michal Guerquin, Peter Alexander Jansen, and Ashish Sabharwal. QASC: A dataset for question answering via sentence composition. In *AAAI*, 2019.

Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, et al. Kimi k1. 5: Scaling reinforcement learning with llms. *arXiv preprint arXiv:2501.12599*, 2025.

Abhinav Kumar, Jaechul Roh, Ali Naseh, Marzena Karpinska, Mohit Iyyer, Amir Houmansadr, and Eugene Bagdasarian. Overthink: Slowdown attacks on reasoning llms. *arXiv e-prints*, pp. arXiv–2502, 2025.

Chen Li, Yinyi Luo, Anudeep Bolimera, and Marios Savvides. Solar: Scalable optimization of large-scale architecture for reasoning. *arXiv preprint arXiv:2503.04530*, 2025.

Haoran Li, Yulin Chen, Jinglong Luo, Jiecong Wang, Hao Peng, Yan Kang, Xiaojin Zhang, Qi Hu, Chunkit Chan, Zenglin Xu, et al. Privacy in large language models: Attacks, defenses and future directions. *arXiv preprint arXiv:2310.10383*, 2023.

Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. Program induction by rationale generation: Learning to solve and explain algebraic word problems. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 158–167, 2017.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024a.

Fan Liu, Wenshuo Chao, Naiqiang Tan, and Hao Liu. Bag of tricks for inference-time computation of llm reasoning. *arXiv preprint arXiv:2502.07191*, 2025a.

Hongwei Liu, Zilong Zheng, Yuxuan Qiao, Haodong Duan, Zhiwei Fei, Fengzhe Zhou, Wenwei Zhang, Songyang Zhang, Dahua Lin, and Kai Chen. MathBench: Evaluating the theory and application proficiency of LLMs with a hierarchical mathematics benchmark. In *Findings of the Association for Computational Linguistics: ACL 2024*, 2024b.

Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. Logiqa: A challenge dataset for machine reading comprehension with logical reasoning. *arXiv preprint arXiv:2007.08124*, 2020.

Tengxiao Liu, Qipeng Guo, Xiangkun Hu, Cheng Jiayang, Yue Zhang, Xipeng Qiu, and Zheng Zhang. Can language models learn to skip steps? In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024c. URL https://openreview.net/forum?id=w4AnTVxAO9.

Yue Liu, Xiaoxin He, Miao Xiong, Jinlan Fu, Shumin Deng, and Bryan Hooi. Flipattack: Jailbreak llms via flipping. *arXiv preprint arXiv:2410.02832*, 2024d.

Yue Liu, Hongcheng Gao, Shengfang Zhai, Xia Jun, Tianyi Wu, Zhiwei Xue, Yulin Chen, Kenji Kawaguchi, Jiaheng Zhang, and Bryan Hooi. Guardreasoner: Towards reasoning-based llm safeguards. *arXiv preprint arXiv:2501.18492*, 2025b.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations (ICLR)*, 2024.

Haotian Luo, Li Shen, Haiying He, Yibo Wang, Shiwei Liu, Wei Li, Naiqiang Tan, Xiaochun Cao, and Dacheng Tao. O1-pruner: Length-harmonizing fine-tuning for o1-like reasoning pruning. *arXiv preprint arXiv:2501.12570*, 2025.

Xinyin Ma, Guangnian Wan, Runpeng Yu, Gongfan Fang, and Xinchao Wang. Cot-valve: Length-compressible chain-of-thought tuning. *arXiv preprint arXiv:2502.09601*, 2025.

Yiran Ma, Zui Chen, Tianqiao Liu, Mi Tian, Zhuo Liu, Zitao Liu, and Weiqi Luo. What are step-level reward models rewarding? counterintuitive findings from mcts-boosted mathematical reasoning. *arXiv preprint arXiv:2412.15904*, 2024.

Benjamin S Manning, Kehang Zhu, and John J Horton. Automated social science: Language models as scientist and subjects. Technical report, National Bureau of Economic Research, 2024.

Shen-yun Miao, Chao-Chun Liang, and Keh-Yih Su. A diverse corpus for evaluating and developing English math word problem solvers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 975–984, 2020. doi: 10.18653/v1/2020.acl-main.92.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *EMNLP*, 2018.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. s1: Simple test-time scaling. *arXiv preprint arXiv:2501.19393*, 2025.

Tergel Munkhbat, Namgyu Ho, Seohyun Kim, Yongjin Yang, Yujin Kim, and Se-Young Yun. Self-training elicits concise reasoning in large language models. *arXiv preprint arXiv:2502.20122*, 2025.

Sania Nayab, Giulio Rossolini, Marco Simoni, Andrea Saracino, Giorgio Buttazzo, Nicolamaria Manes, and Fabrizio Giacomelli. Concise thoughts: Impact of output length on llm reasoning and cost. *arXiv preprint arXiv:2407.19825*, 2024.

Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. Large language diffusion models. *arXiv preprint arXiv:2502.09992*, 2025.

Nvidia. Nvidia isaac gr00t n1: An open foundation model for humanoid robots. https://research.nvidia.com/publication/2025-03_nvidia-isaac-gr00t-n1-open-foundation-model-humanoid-robots, 2025.

International Mathematics Olympiad. American invitational mathematics examination. https://artofproblemsolving.com/wiki/index.php/American_Invitational_Mathematics_Examination?srsltid=AfmBOoqo573PtuNmYWTobFVQWyhhDjV2VXowjsIZ0kvmHQ_UP_Jn2wrG/, 2025.

Isaac Ong, Amjad Almahairi, Vincent Wu, Wei-Lin Chiang, Tianhao Wu, Joseph E. Gonzalez, M Waleed Kadous, and Ion Stoica. RouteLLM: Learning to route LLMs from preference data. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=8sSqNntaMr.

OpenAI. Introducing chatgpt. *https://openai.com/index/chatgpt/*, 2022.

OpenAI. Introducing deep research. *https://openai.com/index/introducing-deep-research/*, 2025a.

OpenAI. Introducing gpt-4.5. *https://openai.com/index/introducing-gpt-4-5/*, 2025b.

OpenAI. Detecting misbehavior in frontier reasoning models. *https://openai.com/index/chain-of-thought-monitoring/*, 2025c.

OpenAI. Openai o3-mini system card, January 31 2025. URL https://cdn.openai.com/o3-mini-system-card.pdf. Accessed: 2025-02-11.

Arkil Patel, Satwik Bhattamishra, and Navin Goyal. Are NLP models really able to solve simple math word problems? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2080–2094, 2021.

Bo Peng, Eric Alcaide, Quentin Anthony, Alon Albalak, Samuel Arcadinho, Stella Biderman, Huanqi Cao, Xin Cheng, Michael Chung, Matteo Grella, et al. Rwkv: Reinventing rnns for the transformer era. *arXiv preprint arXiv:2305.13048*, 2023.

Yuxiao Qu, Matthew YR Yang, Amrith Setlur, Lewis Tunstall, Edward Emanuel Beeching, Ruslan Salakhutdinov, and Aviral Kumar. Optimizing test-time compute via meta reinforcement fine-tuning. *arXiv preprint arXiv:2503.07572*, 2025.

Nitin Liladhar Rane, Abhijeet Tawde, Saurabh P Choudhary, and Jayesh Rane. Contribution and performance of chatgpt and other large language models (llm) for scientific and research advancements: a double-edged sword. *International Research Journal of Modernization in Engineering Technology and Science*, 5(10):875–899, 2023.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*, 2024.

Abulhair Saparov and He He. Language models are greedy reasoners: A systematic formal analysis of chain-of-thought. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=qFVVBzXxR2V.

Xuan Shen, Yizhou Wang, Xiangxi Shi, Yanzhi Wang, Pu Zhao, and Jiuxiang Gu. Efficient reasoning with hidden thinking. *arXiv preprint arXiv:2501.19201*, 2025a.

Yi Shen, Jian Zhang, Jieyun Huang, Shuming Shi, Wenjing Zhang, Jiangze Yan, Ning Wang, Kai Wang, and Shiguo Lian. Dast: Difficulty-adaptive slow-thinking for large reasoning models. *arXiv preprint arXiv:2503.04472*, 2025b.

Zhenyi Shen, Hanqi Yan, Linhai Zhang, Zhanghao Hu, Yali Du, and Yulan He. Codi: Compressing chain-of-thought into continuous space via self-distillation. *arXiv preprint arXiv:2502.21074*, 2025c.

David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *nature*, 550(7676):354–359, 2017.

Zayne Rea Sprague, Fangcong Yin, Juan Diego Rodriguez, Dongwei Jiang, Manya Wadhwa, Prasann Singhal, Xinyu Zhao, Xi Ye, Kyle Mahowald, and Greg Durrett. To cot or not to cot? chain-of-thought helps mainly on math and symbolic reasoning. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=w6nlcS8Kkn.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*, 2022.

DiJia Su, Hanlin Zhu, Yingchen Xu, Jiantao Jiao, Yuandong Tian, and Qinqing Zheng. Token assorted: Mixing latent and text tokens for improved language model reasoning. *arXiv preprint arXiv:2502.03275*, 2025.

Yuan Sui, Yufei He, Nian Liu, Xiaoxin He, Kun Wang, and Bryan Hooi. Fidelis: Faithful reasoning in large language model for knowledge graph question answering. *arXiv preprint arXiv:2405.13873*, 2024.

Yuan Sui, Yufei He, Tri Cao, Simeng Han, and Bryan Hooi. Meta-reasoner: Dynamic guidance for optimized inference-time reasoning in large language models. *arXiv preprint arXiv:2502.19918*, 2025.

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

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019.

Surendrabikram Thapa, Shuvam Shiwakoti, Siddhant Bikram Shah, Surabhi Adhikari, Hariram Veeramani, Mehwish Nasim, and Usman Naseem. Large language models (llm) in computational social science: prospects, current state, and challenges. *Social Network Analysis and Mining*, 15(1):1–30, 2025.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. Musique: Multihop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554, 2022.

Ehsan Ullah, Anil Parwani, Mirza Mansoor Baig, and Rajendra Singh. Challenges and barriers of using large language models (llm) such as chatgpt for diagnostic medicine with a focus on digital pathology–a recent scoping review. *Diagnostic pathology*, 19(1):43, 2024.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with math-vision dataset. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024a. URL https://openreview.net/forum?id=QWTCcxMpPA.

Wenxiao Wang, Wei Chen, Yicong Luo, Yongliu Long, Zhengkai Lin, Liye Zhang, Binbin Lin, Deng Cai, and Xiaofei He. Model compression and efficient inference for large language models: A survey. *arXiv preprint arXiv:2402.09748*, 2024b.

Xiaoxuan Wang, Ziniu Hu, Pan Lu, Yanqiao Zhu, Jieyu Zhang, Satyen Subramaniam, Arjun R. Loomba, Shichang Zhang, Yizhou Sun, and Wei Wang. Scibench: Evaluating college-level scientific problem-solving abilities of large language models. In *Proceedings of the Forty-First International Conference on Machine Learning*, 2024c.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Han Wu, Yuxuan Yao, Shuqi Liu, Zehua Liu, Xiaojin Fu, Xiongwei Han, Xing Li, Hui-Ling Zhen, Tao Zhong, and Mingxuan Yuan. Unlocking efficient long-to-short llm reasoning with model merging. *arXiv preprint arXiv:2503.20641*, 2025a.

Shenghan Wu, Yang Deng, Yimo Zhu, Wynne Hsu, and Mong Li Lee. From personas to talks: Revisiting the impact of personas on llm-synthesized emotional support conversations. *arXiv preprint arXiv:2502.11451*, 2025b.

Yuyang Wu, Yifei Wang, Tianqi Du, Stefanie Jegelka, and Yisen Wang. When more is less: Understanding chain-of-thought length in llms. *arXiv preprint arXiv:2502.07266*, 2025c.

Heming Xia, Yongqi Li, Chak Tou Leong, Wenjie Wang, and Wenjie Li. Tokenskip: Controllable chain-of-thought compression in llms. *arXiv preprint arXiv:2502.12067*, 2025.

Haotian Xu, Xing Wu, Weinong Wang, Zhongzhi Li, Da Zheng, Boyuan Chen, Yi Hu, Shijia Kang, Jiaming Ji, Yingying Zhang, et al. Redstar: Does scaling long-cot data unlock better slow-reasoning systems? *arXiv preprint arXiv:2501.11284*, 2025a.

Silei Xu, Wenhao Xie, Lingxiao Zhao, and Pengcheng He. Chain of draft: Thinking faster by writing less. *arXiv preprint arXiv:2502.18600*, 2025b.

Yige Xu, Xu Guo, Zhiwei Zeng, and Chunyan Miao. Softcot: Soft chain-of-thought for efficient reasoning with llms. *arXiv preprint arXiv:2502.12134*, 2025c.

Yuchen Yan, Yongliang Shen, Yang Liu, Jin Jiang, Mengdi Zhang, Jian Shao, and Yueting Zhuang. Inftythink: Breaking the length limits of long-context reasoning in large language models. *arXiv preprint arXiv:2503.06692*, 2025.

Enneng Yang, Li Shen, Guibing Guo, Xingwei Wang, Xiaochun Cao, Jie Zhang, and Dacheng Tao. Model merging in llms, mllms, and beyond: Methods, theories, applications and opportunities. *arXiv preprint arXiv:2408.07666*, 2024.

Zhenjie Yang, Xiaosong Jia, Hongyang Li, and Junchi Yan. Llm4drive: A survey of large language models for autonomous driving. *arXiv preprint arXiv:2311.01043*, 2023.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question answering, 2018. URL https://arxiv.org/abs/1809.09600.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. tau-bench: A benchmark for tool-agent-user interaction in real-world domains. *arXiv preprint arXiv:2406.12045*, 2024.

Longhui Yu, Weisen Jiang, Han Shi, Jincheng Yu, Zhengying Liu, Yu Zhang, James T Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. Metamath: Bootstrap your own mathematical questions for large language models. *arXiv preprint arXiv:2309.12284*, 2023.

Ping Yu, Jing Xu, Jason Weston, and Ilia Kulikov. Distilling system 2 into system 1. *arXiv preprint arXiv:2407.06023*, 2024.

Weihao Yu, Zihang Jiang, Yanfei Dong, and Jiashi Feng. Reclor: A reading comprehension dataset requiring logical reasoning. In *International Conference on Learning Representations (ICLR)*, April 2020.

Zishun Yu, Tengyu Xu, Di Jin, Karthik Abinav Sankararaman, Yun He, Wenxuan Zhou, Zhouhao Zeng, Eryk Helenowski, Chen Zhu, Sinong Wang, et al. Think smarter not harder: Adaptive reasoning with inference aware optimization. *arXiv preprint arXiv:2501.17974*, 2025.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9556–9567, 2024.

Jintian Zhang, Yuqi Zhu, Mengshu Sun, Yujie Luo, Shuofei Qiao, Lun Du, Da Zheng, Huajun Chen, and Ningyu Zhang. Lightthinker: Thinking step-by-step compression. *arXiv preprint arXiv:2502.15589*, 2025.

Andy Zhou, Kai Yan, Michal Shlapentokh-Rothman, Haohan Wang, and Yu-Xiong Wang. Language agent tree search unifies reasoning acting and planning in language models. *arXiv preprint arXiv:2310.04406*, 2023.

Zixuan Zhou, Xuefei Ning, Ke Hong, Tianyu Fu, Jiaming Xu, Shiyao Li, Yuming Lou, Luning Wang, Zhihang Yuan, Xiuhong Li, et al. A survey on efficient inference for large language models. *arXiv preprint arXiv:2404.14294*, 2024.