

A Survey on Test-Time Scaling in Large Language Models: What, How, Where, and How Well

Qiyuan Zhang^{1*}, Fuyuan Lyu^{2*}, Zexu Sun^{3†}, Lei Wang^{5†}, Weixu Zhang^{2†}, Wenyue Hua^{8†}, Haolun Wu^{2,7†}, Zhihan Guo^{4†}, Yufei Wang^{6‡}, Niklas Muennighoff⁷, Irwin King⁴, Xue Liu², Chen Ma¹

¹City University of Hong Kong, ²McGill University & MILA, ³Gaoling School of Artificial Intelligence, Renmin University of China, ⁴Chinese University of Hong Kong, ⁵Salesforce AI Research, ⁶Macquarie University, ⁷Stanford University, ⁸University of California, Santa Barbara
qzhang732-c@my.cityu.edu.hk, fuyuan.lyu@mail.mcgill.ca

Page: <https://testtimescaling.github.io/>

Abstract

As enthusiasm for scaling computation (data and parameters) in the pretraining era gradually diminished, test-time scaling (*TTS*)—also referred to as “test-time computing”—has emerged as a prominent research focus. Recent studies demonstrate that *TTS* can further elicit the problem-solving capabilities of large language models (LLMs), enabling significant breakthroughs not only in specialized reasoning tasks, such as mathematics and coding, but also in general tasks like open-ended Q&A. However, despite the explosion of recent efforts in this area, there remains an urgent need for a comprehensive survey offering systemic understanding. To fill this gap, we propose a unified, hierarchical framework structured along four core dimensions of *TTS* research: *what to scale*, *how to scale*, *where to scale*, and *how well to scale*. Building upon this taxonomy, we conduct an extensive review of methods, application scenarios, and assessment aspects, and present an organized decomposition that highlights the unique contributions of individual techniques within the broader *TTS* landscape. From this analysis, we distill the major developmental trajectories of *TTS* to date and offer hands-on guidelines for practical deployment. Furthermore, we identify several open challenges and offer insights into promising future directions, including further scaling, clarifying the functional essence of techniques, generalizing to more tasks, and more attributions. Our repository is available on <https://github.com/testtimescaling/testtimescaling.github.io/>.

1 Introduction

Large language models (LLMs) (Brown et al., 2020; OpenAI, 2024a) have emerged in recent years as a transformative milestone toward artificial general intelligence (AGI) (Goertzel, 2014; Bubeck et al., 2023). These models remarkably learn general intelligence by *training-time scaling*, where the models ingest more data and parameters (Kaplan et al., 2020; Hoffmann et al., 2022). However, the progress of pretraining scaling has gradually slowed due to its resource-intensive nature and the bounded availability of human data, prompting researchers to shift their focus toward how to fully elicit the intelligence encoded in LLMs at test time to maximize their real-world effectiveness (Wei et al., 2022; Ouyang et al., 2022; Li et al., 2024d)?

Human cognition may suggest a clue. When faced with complex problems, people tend to engage in deeper, more deliberate thinking, often producing better outcomes (Kahneman, 2011, 2003; Evans, 1984). Inspired by this principle, recent research (Wei et al., 2022; Wang et al., 2023b) has introduced methods that *allocate additional computation during inference* to boost task performance. Notably, some studies (Brown et al., 2024; Wu et al., 2024d) observe patterns akin to scaling laws: increasing inference-time compute yields consistent performance improvements. This family of methods, referred to as **test-time scaling (TTS)**, progressively elicits the model’s intelligence in the test-time, as depicted in Figure 1. The remarkable successes of reasoning models, such as *o1* (OpenAI, 2024b) and *R1* (DeepSeek-AI, 2025), have further amplified interest in *TTS*, highlighting its potential as a key driver of LLM reasoning and utility. However, despite this surge in research activity, the field currently

* Core contribution

† Significant contribution

‡ Taxonomy designer

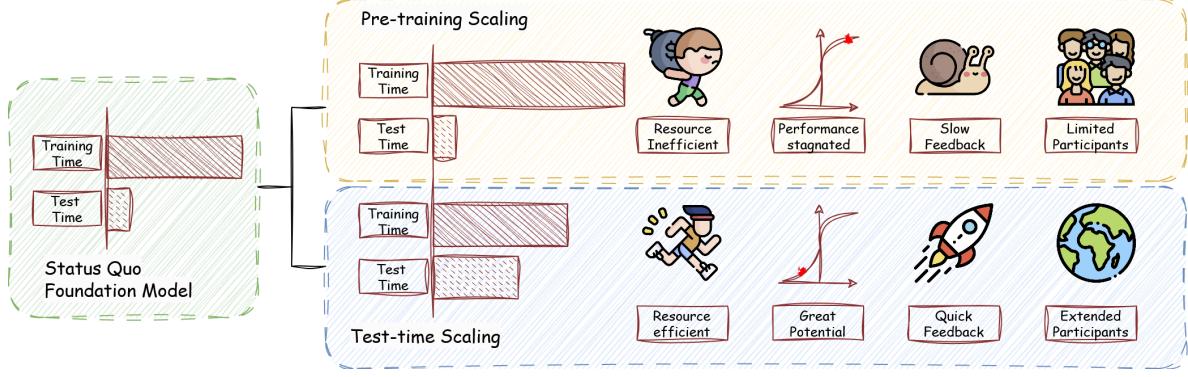


Figure 1: Comparison of Scaling Paradigms in Pre-training and Test-time Phases.

lacks a unified and systematic framework to synthesize insights, compare techniques, or identify consistent trends in *TTS*. To address this gap, we present a comprehensive survey of *TTS*, offering a hierarchical and extensible framework to analyze methods, map research efforts, and guide future progress.

To the best of our knowledge, this is the first survey to comprehensively examine *TTS* across multiple orthogonal dimensions, offering a structured perspective for both theoretical inquiry and practical deployment. Our framework dissects *TTS* into four key dimensions: *what to scale*, *how to scale*, *where to scale*, and *how well to scale*. Our work emphasizes a fine-grained, decomposition-based understanding of *TTS*. While prior efforts have examined *TTS* from specific lenses—such as input modification and output verification (Snell et al., 2024), or through the lens of *System 2 AI* and *Long Chain-of-Thought* (*CoT*) (Li et al., 2025h; Ji et al., 2025; Chen et al., 2025c)—these works are structured around a timeline, tracing the evolution of techniques over time. We analyze the full pipeline, from scaling formulations and algorithmic mechanisms to task domains and performance dimensions. We provide a structured foundation that allows future research to be seamlessly integrated into our taxonomy, making it easier to understand their contributions. Specifically, *what to scale* in Sec. 2 is about what to be scaled at the inference stage. *How to scale* in Sec. 3 depicts how they are implemented. We categorize various techniques, recognizing that a single work may involve multiple techniques; *Where to scale* in Sec. 4 covers the tasks and datasets where these techniques are applied. Finally, *how well to scale* in Sec. 5 refers to evaluating the different attributions of *TTS* methods. We further provide fine-grained subcategories under each axis and systematically map representative works to highlight their contributions and trade-offs (Sec. 6). From this structured analysis, we extract major trends in *TTS* development and offer *hands-on guidance* (Sec. 7) for real-world deployment. Grounded in our multi-dimensional taxonomy, we also identify persistent challenges and promising research directions (Sec. 8): advancing test-time scalability, clarifying the fundamental essence of the effectiveness of different techniques in *TTS*, broadening generalization to a wider range of downstream tasks, and optimizing *TTS* methods along additional dimensions such as efficiency.

Our contributions are threefold:

- A Unified, Multi-Dimensional Taxonomy.** We propose a four-axis taxonomy—*what to scale*, *how to scale*, *where to scale*, and *how well to scale*—that supports structured classification, comparison, and extensibility for *TTS* methods.
- Systematical Literature Organization and Pragmatic Analysis.** Using our taxonomy, we survey the *TTS* landscape, analyze representative methods, and present guidelines for research application and deployment.
- Challenges, Insights, and Forward Directions.** Building on our organized perspective, we uncover critical challenges—ranging from advancing scaling to clarifying essence—and outline promising research directions that could shape future progress. Our unified framework facilitates the mapping of these open questions to concrete dimensions of *TTS*, enabling more targeted and impactful advancements.

We plan to continuously update our taxonomy to reflect ongoing progress and provide an evolving foundation for organizing future developments in *TTS* research.

2 What to Scale

“What to scale” refers to the specific form of *TTS* that is expanded or adjusted to enhance an LLM’s performance during inference. When applying *TTS*, researchers typically choose a specific “what to scale” based on an empirical hypothesis, aiming to achieve performance gains. For example, some researchers hypothesize that

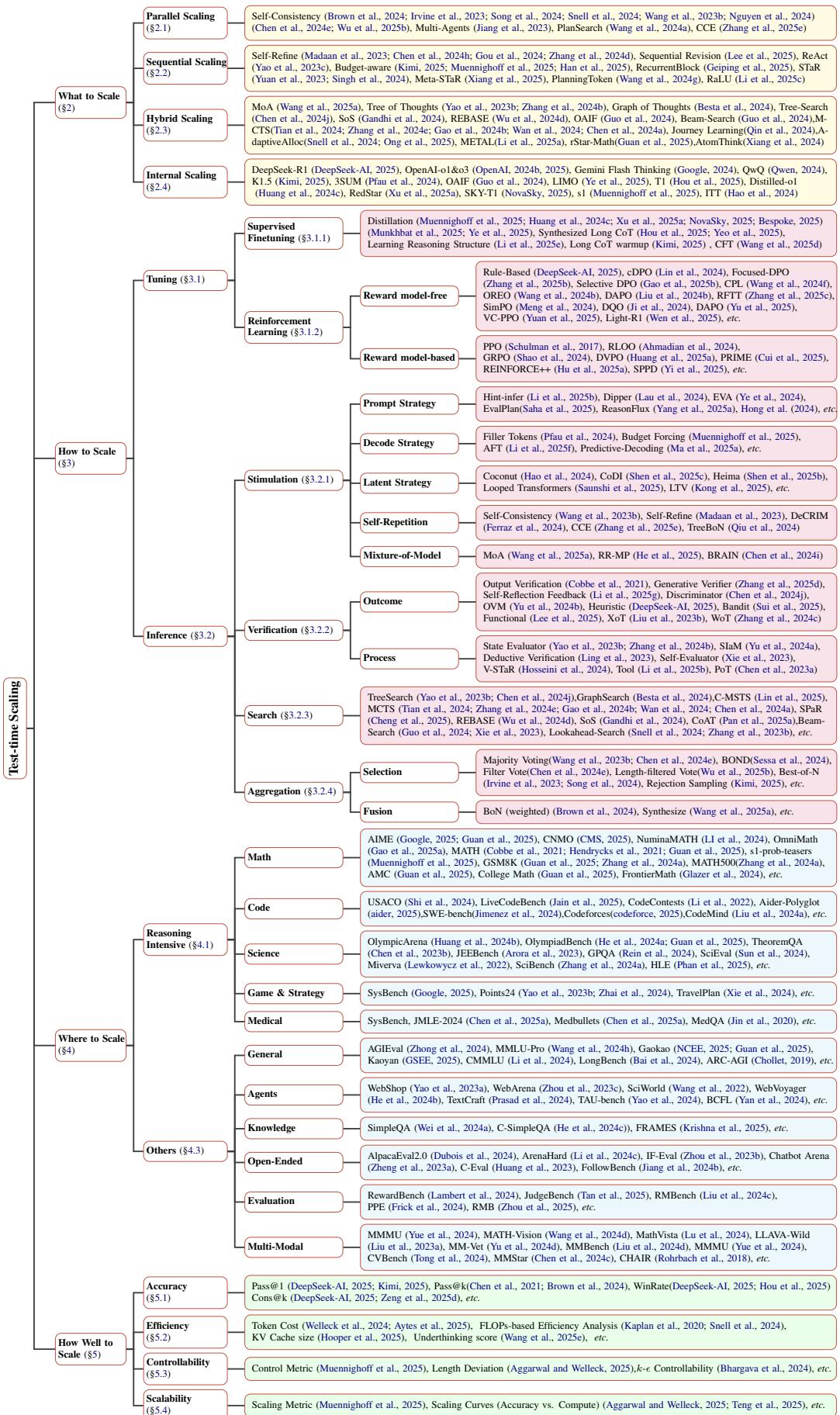


Figure 2: Taxonomy of research in Test-time Scaling that consists of what, how, where, and how well to scale.

longer CoTs improve complex reasoning, leading them to enforce longer outputs from LLMs. Others leverage the self-consistency principle, assuming that generating multiple solutions to a reasoning task increases the likelihood of reaching the correct answer.

2.1 Parallel Scaling

LLMs typically generate a single response per query. *Parallel scaling* improves test-time performance by generating multiple outputs in parallel and then aggregating them into a final answer. Formally, consider a problem set \mathcal{P} and a collection of models $m \in \{1, \dots, M\}$. Each model generates k_m candidate responses for a given problem $p \in \mathcal{P}$, producing a set of sampled solutions \mathcal{S} :

$$\mathcal{S} = \{s_{m,i} \mid m \leq M, i \leq k_m\}, \quad \Rightarrow \quad (\exists \hat{s}) \hat{s} = A(s_{1,1}, \dots, s_{M,k_M}) \text{ is correct.} \quad (1)$$

Here, A is the aggregation function that derives a final response from the set \mathcal{S} . The effectiveness of parallel scaling depends on both **coverage**—the likelihood of generating at least one correct response—and **aggregation quality**, which determines whether a correct response is successfully identified. This approach is supported by both theory and intuition: cognitive science research (Stanovich and West, 2000) suggests that complex problems often allow multiple valid solution paths, and increasing the number of generated responses improves the chance of finding a correct one (Li et al., 2025d). Empirically, this relationship is often log-linear with respect to compute (Brown et al., 2024).

We categorize parallel scaling into two common forms based on different sources of coverage: (1) repeated sampling from a single model and (2) sampling across multiple models. Furthermore, there are some additional techniques to enhance solution diversity and reliability, such as hyperparameter adjustments (*e.g.*, sampling temperature (Renze, 2024) to control output variability) and input modifications (*e.g.*, prompt rephrasing (Lambert et al., 2025) to elicit diverse responses).

2.2 Sequential Scaling

Sequential scaling involves explicitly directing later computations based on intermediate steps. Unlike parallel methods, sequential scaling updates intermediate states iteratively. We denote the partial solution states (subproblem results, or initial drafts) by n_1, n_2, \dots, n_T , with each new state $n_{t+1} = R(n_t, p)$ incorporating both the previous state and the problem context. Because many problems require deliberation rather than immediate pattern matching, single-pass ‘System 1’ (Yu et al., 2024c)-style generation often fails on complex reasoning tasks. Iterative methods emulate a ‘System 2’ approach, breaking down and refining the solution step by step.

Early work like chain-of-thought prompting (Wei et al., 2022) motivated solve the problem step-by-step, $n_{t+1} = \text{AppendStep}(n_t, \text{new reasoning step})$, leading to approaches that refine responses (Madaan et al., 2023), $n_{t+1} = \text{Refine}(n_t)$, or break down problems systematically (Zhou et al., 2023a; Zelikman et al., 2022), $n_{t+1} = \text{IntegrateSub}(n_t, \text{solution to next subproblem})$. Subsequent studies show that iterative revision (Chen et al., 2024h; Gou et al., 2024; Chen et al., 2025d; Snell et al., 2024) triggers self-correction, improving accuracy on challenging tasks. In practice, real-world tasks often demand more flexible and potentially non-linear reasoning paths, suggesting that purely sequential approaches, while effective, may be only one part of a broader solution.

2.3 Hybrid Scaling

Hybrid scaling exploits the complementary benefits of parallel and sequential scaling. Parallel scaling mitigates the risk of the model missing the correct line of thought by casting a wide net, while sequential scaling allows deep exploration of a line of reasoning once it seems promising. Formally, let \mathcal{F}_t be the set of candidate solutions at iteration t . Each iteration expands these candidates in parallel with an expansion function \mathcal{E} and sequentially filters them with a selection function \mathcal{S} :

$$\mathcal{F}_{t+1} = \mathcal{S}\left(\bigcup_{s \in \mathcal{F}_t} \mathcal{E}(s)\right). \quad (2)$$

After T iterations, an aggregator A selects the final solution $\hat{s} \in \mathcal{F}_T$. From a cognitive standpoint, such a combination mirrors how human problem-solvers generate multiple hypotheses (divergent thinking) and then refine/evaluate them (convergent thinking). Classic search algorithms (*e.g.*, iterative deepening (Chen et al., 2025d) and beam search (Snell et al., 2024)) embody this strategy by balancing exploration and exploitation.

Recent work expands on this idea. Tree-of-Thoughts (ToT) (Yao et al., 2023b) branches at decision points, exploring multiple reasoning paths before pruning to a single sequence. Follow-up methods, such as Graph-of-Thoughts (Besta et al., 2024), Algorithm-of-Thought (Sel et al., 2024), Forest-of-Thought (Bi et al., 2024), Monte Carlo Tree Search (MCTS) (Lin et al., 2025), and multi-agent reasoning (Wang et al., 2025a; Chen et al., 2024f), leverage similar but more complex hybrid patterns. For instance, multiple LLMs can debate or verify each other’s answers (Liang et al., 2024; Schaul, 2024), while “journey learning” and “tool-augmented reasoning” (Li et al., 2025b) emphasize capturing full reasoning trajectories.

2.4 Internal Scaling

Internal scaling elicits a model to autonomously determine how much computation to allocate for reasoning during testing within the model’s internal parameters instead of depending on external human-guided strategies. Formally, we update an initial model M_0 to a new model M_1 via a training procedure, $\Phi : (M_0, \mathcal{D}) \mapsto M_1$, on data \mathcal{D} that includes multi-step reasoning tasks (*e.g.*, long CoT examples produced by external scaling (Qin et al., 2024)). Surprisingly, employing outcome-oriented reward modeling (DeepSeek-AI, 2025; OpenAI, 2024b) for RL enables the model to extend its reasoning process autonomously.

At test time, M_1 generates a sequence of internal states z_1, z_2, \dots, z_T via

$$z_{t+1} = f_\theta(z_t), \quad \text{stop}(z_t) = \pi_\theta(z_t). \quad (3)$$

The model’s learned policy π_θ controls when to halt. This internal feedback loop can lead to emergent behaviors—such as more detailed reasoning chains or self-evaluation steps—with any external prompts or multi-call orchestration. In practice, internal scaling often rivals or surpasses standard techniques, thanks to its ability to focus computational effort on a single, coherent reasoning trajectory.

3 How to Scale

3.1 Tuning-based Approaches

To activate a model’s ability to devote cost at test time, directly tuning its parameters is an effective strategy. This includes two approaches: 1) Supervised Finetuning (SFT): Training an LLM via next-token prediction on synthetic or distilled long CoTs enables it to imitate and internalize structured reasoning patterns, effectively learning to think through complex problems. By mimicking extended rationales, SFT reduces the reliance on explicit prompting at inference time. 2) Reinforcement Learning (RL): By leveraging feedback from a reward model on inference tasks, the policy model is automatically updated. Although no supervised data is introduced, the model autonomously generates long CoT reasoning while ensuring reliable answers. We divide the RL for internal scaling works into two perspectives. The reward model-based methods and the reward model-free methods.

3.1.1 Supervised Finetuning (SFT)

Training an LLM via next-token prediction on synthetic or distilled long CoTs enables it to internalize structured reasoning patterns and effectively “think” through complex problems. By mimicking extended rationales, SFT reduces the reliance on explicit prompting at inference time. This will include three subsections: (1) Imitation, describing techniques like MCTS used to generate CoT-style demonstrations for fine-tuning, (2) Distillation, summarizing how student models are trained using outputs from stronger models (*e.g.*, o1, R1), and (3) Warmup, stabilizing learning and aligning the model’s behavior to produce useful step-by-step reasoning.

Imitation A prominent approach to enhancing LLM reasoning via SFT is to generate long CoT demonstrations using test-time “planner” algorithms and then fine-tune the model to imitate those demonstrations. For example, STaR (Zelikman et al., 2022) uses the model itself to generate step-by-step solutions for a given problem and filters for correct outcomes, treating the verified solutions as new demonstrations to fine-tune. More structured search has been applied to generate even higher-quality traces: ReST-MCTS (Zhang et al., 2024a) integrates an MCTS planner (guided by a learned value model) to explore the space of possible reasoning steps; the model is subsequently fine-tuned on these search-generated traces, *i.e.*, it learns to imitate the successful reasoning trajectories discovered by the planner.

Distillation While the imitation approach uses a model’s own intermediate outputs for improvement, distillation techniques aim to transfer the capabilities of a stronger model (or ensemble of models) into a target model via supervised learning. As reported by Muennighoff et al. (2025); Li et al. (2025e), a 32B model trained on a curated sample set generated by a top-tier reasoner was able to solve competition-level math problems nearly as well as the teacher, indicating successful distillation of reasoning.

Warmup SFT warmup (Luong et al., 2024) refers to an initial SFT phase applied to an LLM after its unsupervised pretraining but before other post-training steps like RL. This stage stabilizes subsequent training by providing a well-initialized model that adapts better to preference optimization and avoids instability due to ungrounded behavior (Zeng et al., 2025c). Effective SFT warmup is characterized by several key elements: (i) the use of high-quality, task-relevant datasets (Luong et al., 2024); (ii) short duration; (iii) a tailored learning rate schedule (Pareja et al., 2024). Technically, SFT warmup is often integrated with methods like rejection sampling (Pareja et al., 2024)—which uses warm-started models to generate high-quality data for further training.

3.1.2 Reinforcement Learning (RL)

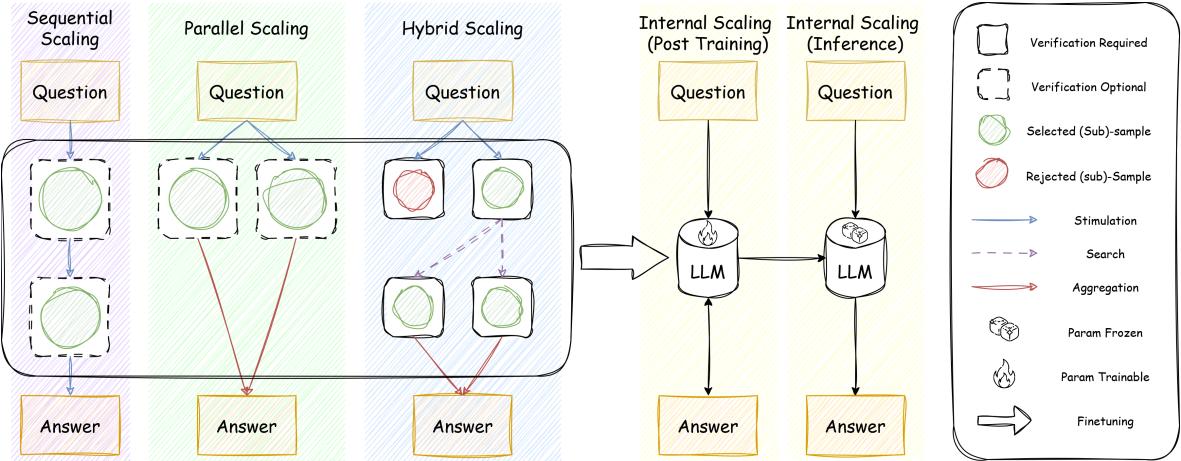
Reward model-free. Recent advancements in RL and preference optimization have significantly enhanced the performance of large language models, particularly in reasoning and problem-solving tasks. A key innovation in this domain is the introduction of RL with verifiable reward by DeepSeek R1 (DeepSeek-AI, 2025), which leverages rule-based reward mechanisms to optimize models efficiently and reliably. This approach has sparked growing interest among researchers working on large models, as it addresses challenges such as sparse rewards and training instability by providing dense feedback for policy optimization. Several methods have been developed to improve exploration and accuracy in reasoning tasks through preference optimization. For instance, cDPO (Lin et al., 2024), CPL (Wang et al., 2024f), Focused-DPO (Zhang et al., 2025b), DAPO (Liu et al., 2024b), and RFTT (Zhang et al., 2025c) prioritize critical or error-prone areas, enhancing internal scaling and reasoning accuracy. Additionally, Selective DPO (Gao et al., 2025b) emphasizes the importance of aligning data difficulty with model capacity by filtering out overly challenging examples, further refining the training process. VC-PPO (Yuan et al., 2025) investigates the failure of PPO for the long CoT task and uses a pre-trained value model to achieve better results. Light-R1 (Wen et al., 2025) proposes a curriculum training framework for increasing data difficulty combined with multi-staged post-training. SimPO (Meng et al., 2024) uses the average log probability of a sequence as the implicit reward and removes the reference model in DPO.

In the realm of mathematical problem-solving, DQO (Ji et al., 2024) and OREO (Wang et al., 2024b) propose novel value function optimization techniques, demonstrating improvements in model performance. DAPO (Yu et al., 2025) leverages dynamic sampling for large-scale RL systems. These advancements are complemented by a range of open-source training frameworks that have equipped researchers and developers with tools to optimize training and enhance inference. Early frameworks like SimpleRL (Zeng et al., 2025b) and DeepScaler (Luo et al., 2025b) quickly replicated the technology stack of DeepSeek R1. Furthermore, SimpleRL-Zoo (Zeng et al., 2025a) presents more experimental details about SimpleRL. Others, such as X-R1 (X-R1Team, 2025) and TinyZero (Pan et al., 2025b), focus on delivering an intuitive and cost-effective user experience. Notably, Open-Reasoner-Zero (Hu et al., 2025b) replicated the DeepSeek R1-zero training scheme using a 32B model, achieving comparable performance. Further advancements in RL for internal scaling have been facilitated by frameworks like OpenR (Wang et al., 2024c), OpenRLHF (Hu et al., 2024), OpenR1 (HuggingFace, 2025), Logic-RL (Xie et al., 2025) and AReaL(AntResearch-RL-Lab, 2025). These frameworks have enhanced the replication of internal scaling and, through open-source sharing, accelerated academic research progress. The above developments not only address key challenges in RL but also pave the way for more efficient and reliable model training and deployment.

Reward model-based. With a Bradley-Terry model (Zheng et al., 2023b) optimized by human preference as the reward model, PPO (Schulman et al., 2017) stands as one of the most influential algorithms with its efficiency and stability and is widely used for internal scaling. Building upon PPO, ReMax (Li et al., 2023b) introduces variance reduction techniques along with REINFORCE (Sutton et al., 1999) and RLOO (Ahmadian et al., 2024) methods. This eliminates the need for additional value models in PPO, reduces over four hyperparameters, lowers GPU memory usage, and speeds up the training process. GRPO (Shao et al., 2024) replaces traditional value models with improved sampling strategies. This significantly accelerates the learning process and achieves performance comparable to GPT-4 in mathematics. REINFORCE++ (Hu et al., 2025a) further simplifies GRPO and enhances its training. DVPO (Huang et al., 2025a) presents a streamlined framework, substituting the reward model with a pre-trained global value model and removing the dependency between the actor and critic. PRIME (Cui et al., 2025) integrates the SFT model as a PRM within a unified RL framework, allowing online updates through policy rollouts and outcome labels via implicit process rewards. SPPD (Yi et al., 2025) utilizes process preference learning with a dynamic value margin for self-training. Recently, several works have focused on other challenges of existing reward model-based methods. UGDA (Sun et al., 2025) leverages the uncertainty and influence of samples during PPO training and iteratively refines the reward model. VinePPO (Kazemnejad et al., 2024) exploits the flexibility of language environments to compute unbiased Monte Carlo-based estimates, avoiding the need for large value networks. LCPO (Aggarwal and Welleck, 2025) focuses on optimizing accuracy and adherence to user-specified length constraints for reasoning tasks. Rest-MCTS* (Zhang et al., 2024a) uses tree-search-based RL to bypass per-step manual annotation typically required for training process rewards. These advancements and refinements in algorithms continue to drive the field of reinforcement learning for internal scaling, offering more effective tools and methods for solving complex problems.

3.2 Inference-based Approaches

Unlike training-based approaches, which adjust the model’s parameters offline, inference-based approaches dynamically adjust computation during deployment. This paradigm includes four essential components: (i) *Simulation*, which encourages the model to generate longer or multiple candidate outputs; (ii) *Verification*, which filters or scores outputs based on correctness or other criteria; (iii) *Search*, which systematically explores the sample

Figure 3: A Visual Map and Comparison: From *What to Scale* to *How to Scale*.

space; and (iv) *Aggregation*, which consolidates multiple outputs into the final output. These four components are often used in combination to allocate test-time computation more effectively and boost performance on complex reasoning tasks. In the following sections, we provide detailed discussions of each component.

3.2.1 Stimulation

Stimulation techniques are the first step in encouraging the model to allocate more computation to thinking. It basically stimulates the LLM to generate (i) longer samplers and (ii) more samples instead of generating single and short samples via naive prompting. This includes several key approaches:

Prompt Strategy. Instead of allowing the model to generate an answer directly, one way to stimulate the scaling of LLM during test time is through the prompt. This behavior requires the backbone LLM’s ability to follow instructions. For instance, prompts can guide the model toward step-by-step reasoning. Simple modifications such as adding explicit instructions (*e.g.*, “Please think step by step.”) can improve the model’s ability to break down complex problems into intermediate steps (Lightman et al., 2023). This strategy ensures more deliberate and structured thought generation by shaping the reasoning process at the input level. Other techniques such as (Wei et al., 2022; Ranaldi et al., 2025) also rely on explicitly stating the requirements in the prompt to stimulate samples during the TTS .

Decode Strategy Rather than passively accepting the model’s default output behavior, this approach modifies the decoding process to encourage LLM to generate longer, more detailed samples adaptively. Techniques such as injecting filler token (Pfau et al., 2024), adaptively injecting predefined injection phrase (Jin et al., 2020), forcing scaling budget (Muennighoff et al., 2025), enforcing intermediate generation (Li et al., 2025f), or predictive decoding (Ma et al., 2025a) allow the model to modify its distribution progressively. Enforcing extended reasoning at the output level enables the model to think longer and generate more comprehensive solutions without requiring additional external guidance.

Latent Strategy Unlike strategies that rely on token-level instructions or output expansion, latent strategies encourage deeper or recurrent thinking within the hidden representations themselves, effectively scaling up test-time computation through continuous internal states. For example, Hao et al. (2024) propose a paradigm where the model completes reasoning steps entirely in hidden space before producing the final answer; Kong et al. (2025) introduce a latent-thought framework that conditions text generation on an inferred latent variable to guide more thorough or expansive reasoning, while Shen et al. (2025c) show that compressing CoT into continuous embeddings can preserve intermediate reasoning fidelity without lengthy textual traces. Other approaches (Saunshi et al., 2025) harness *looped* or *recurrent* inference to repeatedly refine hidden states, effectively unfolding multiple “thinking iterations” in a single forward pass.

Self-Repetition Strategy Apart from generating longer samples, another way to stimulate the LLM is to generate multiple samples instead of individual ones. One commonly adopted strategy is to prompt the LLM repeatedly during the decoding stage, commonly known as self-repetition (Wang et al., 2023b). Another strategy is to prompt the LLM sequentially, in order to mimic refinement process (Madaan et al., 2023) or correlation under constraint (Ferraz et al., 2024).

Mixture-of-Model Strategy Gathering the “wisdom of the crowd” can move beyond repeated sampling from a single model to coordinated sampling across multiple models. These LLMs can play either homogeneous roles (Wang et al., 2025a) or heterogeneous roles (Chen et al., 2024i; He et al., 2025) during the process. By harnessing diverse perspectives, such multi-model strategy not only increases the coverage of possible solutions but also improves the system’s overall robustness.

Category	Approach	Approach Description
Prompt	CoT (Wei et al., 2022)	Contains a series of intermediate reasoning steps in prompts
	Step-by-step (Lightman et al., 2023)	Stimulate step-by-step thinking via prompt
	QuaSAR (Ranaldi et al., 2025)	Decompose CoT into Quasi-Symbolic Language
	CoD (Xu et al., 2025b)	Generate concrete representations and distill into concise equation
	Hint-infer (Li et al., 2025b)	Inserting artificially designed hints in the prompt
	Think (Li et al., 2025b)	Prompt LLM with “Think before response”
Decode	Think About World (Jin et al., 2024)	Prompt LLM with “Think About the World” to enforce larger inference
	Filler-token (Pfau et al., 2024)	uses arbitrary, irrelevant filler tokens before answering
	Budget-forcing (Muennighoff et al., 2025)	suppress the generation of the end-of-thinking token
	AFT (Li et al., 2025f)	iteratively aggregating proposals and aggregate for future proposals
	Predictive-Decoding (Ma et al., 2025a)	re-weight decoding distribution given evaluation of foresight
Latent	Adaptive Injection (Jin et al., 2025)	Injecting a predefined injection phrase under certain condition
	Coconut (Hao et al., 2024)	Perform chain-of-thought in hidden space without explicit token generation
	CoDI (Shen et al., 2025c)	Compress chain-of-thought into continuous vectors via self-distillation
	Looped (Recurrent) Transformers (Saunshi et al., 2025)	Unroll model depth at inference by repeatedly refining hidden states
Self-Repetition	Heima (Shen et al., 2025b)	Encode each reasoning step into a single latent token to reduce output length
	LTV (Kong et al., 2025)	Introduce a latent thought variable to guide text generation
	Self-Repetition (Wang et al., 2023b)	prompt LM in parallel
Mixture-of-Model	Self-Refine (Madaan et al., 2023)	Naively prompt LM to iteratively refine answer
	DeCRIM (Ferraz et al., 2024)	Self-correlation for multi-constrained instruction following
	MoA (Wang et al., 2025a)	Prompt different models in parallel and iteratively improve
	RR-MP (He et al., 2025)	Propose Reactive and Reflection agents to collaborate
BRAIN (Chen et al., 2024i)	BRAIN (Chen et al., 2024i)	Propose frontal & parietal lobe model to inspire brain
	Collab (Chakraborty et al., 2025)	Propose decoding strategies to leverage multiple off-the-shelf aligned LLM policies

Table 1: Summary of Certain Stimulation Techniques.

3.2.2 Verification

Verifying the correctness and consistency of LLM during the test-time scaling is also crucial. The verification process plays an important role in the test-time scaling, as a solid verification process can be adapted to:

- directly selects the output sample among various ones, under the *Parallel Scaling* paradigm;
- guides the stimulation process and determines when to stop, under the *Sequential Scaling* paradigm;
- serves as the criteria in the search process, which we will discuss in Section 3.2.3;
- determines what sample to aggregate and how to aggregate them, e.g., weights, which we will discuss in Section 3.2.4.

Usually, there are two types of verifications, as shown below:

Outcome Verification. Outcome verification plays a crucial role in ensuring the correctness and consistency of generated outputs. Common approaches include using a separate verifier model to score multiple candidate answers (e.g., Cobbe et al. (2021)), employing self-consistency, voting mechanisms (Wang et al., 2023b) and discriminator LM (Chen et al., 2024j) and leveraging tool-assisted (Gou et al., 2024) or heuristic checks (DeepSeek-AI, 2025) in domains such as math and code generation. For specific task problems, such as trip planning, functional scoring (Lee et al., 2025) is also adopted for verifying the proposed plans. Instead of formulating the outcome verification as a classification problem, Zhang et al. (2025d) exploits the generative ability of LLM and proposes to reformulate the outcome verification process as a next-token prediction task. Li et al. (2025g) formulate the feedback utilization as an optimization problem and adaptive propagate information between samples.

Apart from single criteria, certain outcome verification approaches verify the quality of the simulated samples from multiple perspectives. Liu et al. (2023b) conducts both (i) passive verification from external tools and (ii) active verification via a rethinking mechanism to justify each sample. Zhang et al. (2024c) follows a similar idea and proposes to verify each sample from three aspects: Assertion, Process, and Result. Lifshitz et al. (2025) further extends the number of verification agents to an arbitrary number and decouples the semantic criteria with verification agents. Parmar et al. (2025) and Saad-Falcon et al. (2024) also propose a verification agent to score each sample considering various factors, respectively. Saad-Falcon et al. (2024) additionally proposes a unit test-based verification approach. We provide a detailed technical categorization in the Appendix A.

3.2 Inference-based Approaches

Process Verification. Process verification approaches verify the sample outcomes and the process of obtaining such an outcome. They are commonly adopted in tasks with formal, deductive processes, such as reasoning, coding, or mathematics. They are also known as the process reward model (PRM) or state verification. Lightman et al. (2023) processes to train a PRM as a step-level verification on mathematical tasks. Yao et al. (2023b) processes an LM-based state verifier as guidelines for searching the samples under the tree structure. Zhang et al. (2024b) further tunes these preference data into LLM and enables CoT structure during test time. Instead of training an external verifier, Xie et al. (2023) prompts the same LM to evaluate the current step given all previous ones. Hosseini et al. (2024) proposes to train the verifier with both accurate and inaccurate generated data. Although LM-based process verifiers can be easily integrated, they may yield unreliable verification, especially for complex problems with long processes. Ling et al. (2023) decomposes the verification process in a deductive manner. Hence, the verifier only needs to verify a few statements within the long thought chain. Yu et al. (2024a) is based on similar intuition but instead focuses on code-aided mathematical reasoning tasks with the critic model iteratively. Li et al. (2025b) instead relies on the external toolbox, such as code interpreters, to verify the process.

Category	Approach	Approach Description
Outcome	Naive ORM (Cobbe et al., 2021)	Naively process to train solution-level and token-level verifiers on labeled-dataset
	OVM (Yu et al., 2024b)	Train a value model under outcome supervision for guided decoding
	Heuristic (DeepSeek-AI, 2025)	Heuristic check for domain-specific problems
	Functional (Lee et al., 2025)	Functional scoring for task-specific problems
	Bandit (Sui et al., 2025)	Train a bandit algorithm to learn how to verify
	Generative Verifier (Zhang et al., 2025d)	Exploit the generative ability of LLM-based verifiers via reformulating the verification
	Self-Reflection Feedback (Li et al., 2025g)	formulate the feedback utilization as an optimization problem and solve during test-time
	Discriminator (Chen et al., 2024j)	SFT a domain-specific LM as a discriminator
	Unit Test (Saad-Falcon et al., 2024)	Verify each sample as unit tests
	XoT (Liu et al., 2023b)	Passive verification from external tools and Activate verification via re-thinking
Process	WoT (Zhang et al., 2024c)	Multi-Perspective Verification on three aspects: Assertion, Process, and Result
	Multi-Agent Verifiers (Lifshitz et al., 2025)	Multi-Perspective Verification without explicit semantic meanings
	Naive PRM (Lightman et al., 2023)	SFT an LM as a PRM on each reasoning step over mathematical tasks
	State Verifier (Yao et al., 2023b)	SFT an LM as a state verifier and evaluate states either independently or jointly
	Deductive PRM (Ling et al., 2023)	Deductively verify a few statements in the process
Process	Self-Evaluation (Xie et al., 2023)	Prompting the same LM to evaluate the current step given previous ones
	Lego-prover (Wang et al., 2023a)	Evaluate lemma in Theorem Proving
	PoT (Chen et al., 2023a)	delegate computation steps to an external language interpreter
	Tool (Li et al., 2025b)	Relies on external toolbox for verification
	V-STaR (Hosseini et al., 2024)	Verifier trained on both accurate and inaccurate self-generated data

Table 2: Summary of Certain Verification Techniques.

3.2.3 Search

Search is also a frequently used component during the test-time scaling. LLMs pre-trained on vast amounts of online data, can be viewed as a compression of real-world knowledge. However, standard inference tends to underutilize their capacity. Search, being a classic yet working technique in retrieving relevant information from vast databases, can be utilized to fully exploit the capability of LLMs by exploring their potential options in a structured manner. Existing test-time scaling approaches based on search techniques demonstrate significant performance increases over complex tasks, such as complex mathematics, etc.

Yao et al. (2023b) explores the potential of search by decomposing the output samples into multiple thoughts and organizing them in a tree structure. Based on only Naive tree search algorithms, such as depth-first search and breath-first search, it demonstrates superior performance on reasoning tasks. Monte-Carlo Tree Search (Coulom, 2006), being a classical and powerful search algorithm, also shines its light on better exploiting the hidden knowledge of LLMs. Chaffin et al. (2022) adopts MCTS during the decoding stage guided by discriminators for constrained textual generation. Zhang et al. (2023b) further extends the MCTS to enhance the planning ability in code generation via looking ahead. Tian et al. (2024) incorporates the MCTS as a critical component in the self-improving framework for LLM. Wan et al. (2024) tailors the search algorithm to tackle problems requiring long-horizon planning and deep tree structure for searching. Chen et al. (2024j) further identifies that discriminators are the key bottleneck in search-enhanced planning. Gandhi et al. (2024) systematizes the search process in a unified language and proposes to train an LLM with data and feedback from the search process. Wu et al. (2024d) empirically analyzes various search algorithms and designs a reward-balanced search algorithm toward Pareto-optimal test-time scaling. Edward Beeching (2024) further extends the beam search by incorporating diversity consideration.

Apart from searching within the tree structure, Besta et al. (2024) models the output as a graph search problem. Xie et al. (2023) proposes a stochastic beam search solution based on self-evaluation for reasoning tasks. Pan et al. (2025a) enhances MCTS with proposed associative memory to dynamically update its knowledge base. Li et al.

(2025c) proposes to solve the reasoning process as constructing a control flow graph with each node indicating a logic unit.

3.2.4 Aggregation

Aggregation techniques consolidate multiple solutions into a final decision to enhance the reliability and robustness of model predictions at test time. Based on how the final output is generated, we empirically categorize them into two key classes: (i) Selection, which selects the best-performed sample among all candidates, where the selection criteria may vary across different approaches; and (ii) Fusion, which fuse multiple samples into one through tricks like weighting or generation.

Selection In this category, the aggregation process can be viewed as a selection problem. One well-known example is to select the most consistent answer, commonly known as *self-consistency*. Wang et al. (2023b) improves accuracy by leveraging statistical redundancy—if different reasoning paths converge to the same conclusion, the answer is more likely to be correct. Self-consistency effectively reduces variance in model outputs and mitigates occasional hallucinations. However, as the final output is voted based on consistency, inaccurate and low-quality samples would inevitably influence the output quality. Therefore, various approaches are proposed to filter the candidates before voting. Chen et al. (2024e) incorporates an LM as a filter, while Wu et al. (2025b) proposes a Length-filtered vote, where prediction uncertainty is adopted as a proxy to filter reliable CoT length.

Best-of-N (Irvine et al., 2023) follows the same process but replaces the self-consistency criteria with scalar scores generated by external verifiers. Song et al. (2024) further demonstrates that best-of-N on small LLMs can yield competitive performance against SOTA propriety models. Munkhbat et al. (2025) attaches a few-conditioning filtering before the best-of-N selection. This aims to alleviate its sample inefficiency and achieves significant length reduction. Motivated by particle filtering, Puri et al. (2025) proposes to consider filtering upon the samples. Sessa et al. (2024) went one step further in reducing sample inefficiency. It tunes the best-of-N results into the LM via RLHF. With the blooming of the agentic approach, Parmar et al. (2025) proposes a selection agent considering complex factors with both historical and current status. Apart from selecting samples from one single LM, Ong et al. (2025) views the selection of samples generated by weak and strong LLMs as a routing problem and proposes constraints on computation costs.

Fusion Directly selecting the final output sample among candidates may yield unsatisfactory results, especially when the sample quality of candidates is low. Fusion approaches propose to merge multiple samples into one to solve such a problem. Brown et al. (2024) and Li et al. (2023a) extend the idea from Best-of-N and weigh each sample by its score from external verifiers. Jiang et al. (2023), on the other hand, directly prompts another LLM as a summarizer to merge multiple selected samples. Li et al. (2025j) shares similar intuition by replacing the majority voting in self-consistency (Wang et al., 2024e) with generative self-aggregation. Li et al. (2025c) also adopts LLM as the synthesizer, given the intermediate consideration in previous steps.

Category	Approach	External Verifier	Approach Description	Also Utilized in
Selection	Majority Voting (Wang et al., 2023b)	✗	Select the most common sample	(Chen et al., 2024e) (Song et al., 2024)
	Best-of-N (Irvine et al., 2023)	✓	Select the highest scored sample	
	Few-shot BoN (Munkhbat et al., 2025)	✓	BoN with few-shot conditioning	
	Agentic (Parmar et al., 2025)	✗	agent considering both current and previous status	
Fusion	Weighted BoN (Li et al., 2023a)	✓	Weight each sample by its score	(Brown et al., 2024) (Wang et al., 2025a; Li et al., 2025c)
	Synthesize (Jiang et al., 2023)	✗	Fuse the selected samples via GenAI	
	Ensemble Fusion (Saad-Falcon et al., 2024)	✗	Conduct ensemble before fusion	

Table 3: Summary of Certain Aggregation Techniques. BoN stands for Best-of-N.

4 Where to Scale

TTS can substantially enhance LLMs’ performance across diverse real-world scenarios. We systematically categorize these scenarios into representative domains, detailing the characteristic challenges, critical evaluation criteria, and representative benchmarks that illustrate the practical value of TTS. Here, we also list a brief summary of various benchmarks in Table 4.

4.1 Reasoning-intensive Tasks

Reasoning-intensive tasks require structured, explicit, multi-step reasoning, precision, and rigorous correctness verification. These tasks challenge LLMs’ ability to systematically decompose problems, iteratively refine solutions, and verify intermediate reasoning steps.

Mathematical Reasoning Mathematical tasks involve complex computations, logical inference, and iterative verification. Key challenges for *TTS* methods include generating accurate step-by-step solutions, effectively verifying intermediate steps, and handling intricate reasoning logic. Representative benchmarks include MiniF2F (Zheng et al., 2021), AIME 2024 (Google, 2025), MATH-500 (Zhang et al., 2024a), AMC 2023 (Guan et al., 2025), PutnamBench (Tsoukalas et al., 2024), MUSTARD (Huang et al., 2024a) and OlympiadBench (He et al., 2024a). These datasets span advanced competition-level math problems, emphasizing precise and explicit reasoning skills.

Programming & Code Generation Coding tasks demand syntactic accuracy, executable correctness, and iterative debugging. Challenges for *TTS* methods lie in generating correct implementations, debugging code iteratively, and efficiently exploring multiple coding solutions. Representative datasets include Codeforces (codeforce, 2025), SWE-bench (Jimenez et al., 2024), and LiveCodeBench (Jain et al., 2025), each providing expert-level coding challenges that require rigorous logical thinking and implementation accuracy.

Game Playing and Strategic Reasoning Strategic reasoning tasks involve adaptive planning, interactive decision-making, and complex multi-round reasoning. *TTS* methods must efficiently perform iterative search, adaptive inference, and dynamic interactions. A representative benchmark is SysBench (Google, 2025), which evaluates models' strategic reasoning in interactive tasks.

Scientific Reasoning Scientific problems typically require multi-domain knowledge integration across physics, chemistry, biology, and other disciplines. *TTS* methods must demonstrate broad knowledge synthesis, multi-step reasoning, and accurate factual verification. Notable benchmarks include GPQA Diamond (Rein et al., 2024) and MR-Ben (Zeng et al., 2024), focusing on advanced scientific reasoning and integrated domain knowledge. Domain specific benchmarks, such as TP-Bench (Chung et al., 2025), UGPhysics (Xu et al., 2025c), PHYSICS (Feng et al., 2025), have also investigated recently.

Medical Reasoning Medical tasks involve diagnostic decision-making, clinical reasoning, and precise medical knowledge. The key challenge for *TTS* here is ensuring reliable, accurate reasoning that mimics medical professionals' decision logic. Representative datasets include JAMA Clinical Challenge (Chen et al., 2025a), Medbullets (Chen et al., 2025a), and MedQA (Jin et al., 2020). These benchmarks critically assess reasoning LLMs' capabilities in diagnosis, treatment planning, and medical decision accuracy.

4.2 Agentic Tasks

The scaling of artificial intelligence agents can be classified into three distinct categories: scaling through design choice, scaling for emergent behavior analysis, and scaling through environmental interaction. In addition, many benchmarks have been proposed for agentic task training, attempting to provide extensive environment feedback.

Scaling Agents as Design Choice The first category involves scaling multi-agent systems as a deliberate design choice to enhance performance. A central question in this domain concerns the scaling laws of collaborative agents: specifically, how the progressive addition of collaborative agents impacts overall system performance. (Li et al., 2024a) studies the scaling of multi-agent system with sampling and voting. The experiments are conducted by using various LLMs of different sizes on diverse datasets covering reasoning and generation. Performance can be improved by increasing the ensemble size across various tasks, including GSM8k, MMLU, and HumanEval. (Chen et al., 2024d) have examined the relationship between the number of LLM calls and system performance, revealing non-monotonic behavior patterns. This non-monotonicity stems from varying query difficulties within tasks: while increased LLM calls enhance performance on simpler queries, they may diminish performance on more complex ones. (Qian et al., 2024) studies the scaling of multi-round discussion in multi-agent collaboration on MMLU, HumanEval, SRDD, and SRDD under regular typologies including chain, tree, and graphs and irregular typologies. They found that the performance mostly follow a logistic growth pattern as the process of scaling agents. (Meyerson and Qiu, 2025) asks a more theoretical question on how task decompositions and assign multiple agents with different subtasks can be optimal? It provides an asymptotic analysis with LLM primitives about the efficiency of such decomposed systems, leading future opportunities for scaling them.

Scaling Agent for Emergent Social Ability The second category focuses on scaling multi-agent systems to study emergent behaviors in large-scale simulations, particularly in social science applications. (Zhang et al., 2025h) studies large-scale multi-agent simulation over 300 agents under information asymmetry, focusing on phenomena including the emergence of information cocoons, the evolution of information gaps, and the accumulation of social capital. (Piao et al., 2025) generates social lives for over 10k agents, simulating their 5 million interactions both among agents and between agents and their environment. The simulation is used as a testbed for to study four key social issues: polarization, the spread of inflammatory messages, the effects of universal basic income policies, and the impact of external shocks such as hurricanes. To support such large scale experiments, (Pan et al.) developed a user-friendly multi-agent infrastructure for future large-scale multi-agent experiments.

Scaling Environment Feedback The third category involves scaling the interaction between agents and their environment to obtain richer feedback. (Hilton et al., 2023) find out that an agent’s intrinsic performance scales as a power law relative to model size and environmental interactions. However, constructing extensive environments to scale feedback and interactions for agent training remains challenging. (Chen et al., 2025e) tried to address this issue by training a reward model based on synthetic data to enable the agent to do Monte-Carlo Search during planning. This process typically involves using LLM-based agents to collect action trajectory demonstrations, using another LLM for generating contrasting negative trajectories, and training customized reward models based on the synthetic data collected. (Yu et al., 2024e) tries to combine test-time search and self-learning by Reflective Monte Carlo Tree Search for richer feedback to enhance AI agents’ ability to explore decision space on the fly.

Simulated Environment for Agentic Tasks Agentic tasks involve realistic and interactive environments, requiring complex planning, iterative reasoning, and effective tool utilization. TTS methods face challenges such as optimal stepwise planning, adaptive decision-making, tool integration, and iterative refinement. Representative benchmarks include WebShop (Yao et al., 2023a), WebArena (Zhou et al., 2023c), SciWorld (Wang et al., 2022), and TextCraft (Prasad et al., 2024). These datasets provide realistic interactive scenarios, emphasizing iterative decision-making and effective tool usage. Recent advances in scaling LLM-driven autonomous agents center on improved planning, memory, and self-optimization techniques. For example, ARMAP (Chen et al., 2025f) automatically learns a reward model from unlabeled environment interactions to score and guide an agent’s actions, circumventing the need for human-labeled feedback and improving multi-step decision-making.

4.3 Others

These tasks require broad, general-purpose reasoning capabilities, creativity, and subjective evaluation of outputs.

General To achieve general objectives, many efforts have collected numerous official, public datasets that are challenging for humans but are not exclusive to any particular domain. Representative benchmarks include AGIEval (Zhong et al., 2024), MMLU-Pro (Wang et al., 2024d), and Gaokao (Guan et al., 2025). These benchmarks may cover multiple aspects of language models and aim to test their general performance.

Open-Ended Tasks TTS methods must enhance output diversity, quality, and coherence, balancing creativity and correctness. Representative benchmarks include AlpacaEval2.0 (Dubois et al., 2024), ArenaHard (Li et al., 2024c), IF-Eval (Zhou et al., 2023b), and C-Eval (Huang et al., 2023), which collectively evaluate subjective, open-ended, and general-purpose reasoning.

Knowledge-intensive Tasks Knowledge-intensive tasks require LLMs to retrieve and synthesize factual knowledge from external sources, ensuring accuracy and reducing hallucinations. TTS challenges center around effective retrieval-augmented reasoning, iterative verification, and multi-source aggregation. Representative benchmarks include SimpleQA (Wei et al., 2024a), C-SimpleQA (He et al., 2024c), and FRAMES (Krishna et al., 2025), emphasizing factual correctness and retrieval-based reasoning.

Evaluation Tasks Evaluation tasks require LLMs to act as judges, also known as Generative Reward Models (GRMs), to conduct comprehensive and in-depth quality assessments of the candidate responses, thus ensuring reliable evaluation results. Representative benchmarks in this field include RewardBench (Lambert et al., 2024), JudgeBench (Tan et al., 2025), RMBench (Liu et al., 2024c), PPE (Frick et al., 2024), and RMB (Zhou et al., 2025). Recent research (Kim et al., 2025) has demonstrated that TTS effectively enhances the evaluation reasoning capabilities of LLMs. For instance, CCE (Zhang et al., 2025e) scales the evaluation by comparing the candidate responses with other crowd-generated responses, enabling TTS evaluation effects. EvalPlan (Saha et al., 2025) achieves deeper evaluation by first generating a specific evaluation plan tailored to the candidate responses. SPCT (Liu et al., 2025c) goes a step further by employing RL to generate evaluation principles, further activating the TTS potential. Additionally, JudgeLRM (Chen et al., 2025b) has validated that training using the R1 method can effectively enhance the performance of RMs, while MAV (Lifshitz et al., 2025) introduces multiple aspect verifiers. Notably, improving evaluator accuracy in Out-of-Distribution scenarios remains a critical issue, like Reward Hacking (Skalse et al., 2025; Shen et al., 2025a), worthy of deeper exploration.

Multimodal Tasks Multimodal reasoning tasks demand effective cross-modal integration, iterative reasoning between modalities, and robust verification across visual and textual inputs. TTS methods face challenges in modality fusion, iterative multimodal reasoning, and handling ambiguity across modalities. Representative benchmarks include MMMU (Yue et al., 2024), MathVista (Lu et al., 2024), MathVision (Wang et al., 2024d), CMMaTH (Li et al., 2025i), and PGPS9K (Zhang et al., 2023a), each testing multimodal reasoning across visual and textual modalities.

Benchmark	Size	Evaluation Criteria	Example Task	Key Features	Type
Reasoning-intensive Tasks					
FrontierMath (Glazer et al., 2024)	Hundreds	Exact match	Algebraic geometry	High complexity	
MATH (Cobbe et al., 2021)	12.5K	Exact match	AMC/AIME-style	Structured reasoning	
MiniF2F (Zheng et al., 2021)	1794	Formal	Mixed	Formal Benchmark	
PutnamBench (Tsoukalas et al., 2024)	1709	num-solved	Competition-level	Theorem-proving	
MUSTARD (Huang et al., 2024a)	Varied	Pass@1	Mixed	Auto-Generated	
NuminaMath (Li et al., 2024)	860K	Exact match, CoT	Olympiad-level math	Annotated reasoning	Math
OmniMath (Gao et al., 2025a)	4.4K	Accuracy	Math Olympiads	Advanced reasoning	
GSM8K (Zhang et al., 2024a)	8.5K	Accuracy	Grade-school math	Natural-language solutions	
rStar-Math (Guan et al., 2025)	747K	Pass@1 accuracy	Competition math	Iterative refinement	
ReST-MCTS (Zhang et al., 2024a)	Varied	Accuracy	Multi-step reasoning	Reward-guided search	
s1 (Muennighoff et al., 2025)	1K	Accuracy	Math/science tasks	Controlled compute	
USACO (Shi et al., 2024)	307	Pass@1	Olympiad coding	Creative algorithms	
AlphaCode (Li et al., 2022)	Thousands	Solve rate	Competitive coding	Complex algorithms	Code
LiveCodeBench (Jain et al., 2025)	511	Pass@1	Real-time coding	Live evaluation	
SWE-bench (Jimenez et al., 2024)	2.3K	Resolution rate	GitHub issues	Multi-file edits	
GPQA (Rein et al., 2024)	448	Accuracy	Graduate STEM	Domain expertise	
PHYSICS (Feng et al., 2025)	1297	Accuracy	University-level	Physics Problem	
OlympicArena (Huang et al., 2024b)	11.1K	Accuracy	Multidisciplinary tasks	Multimodal reasoning	
OlympiadBench (He et al., 2024a)	8.4K	Accuracy	Math/Physics Olympiads	Expert multimodal tasks	
TheoremQA (Chen et al., 2023b)	800	Accuracy	Theorem-based STEM	Theoretical application	Science
TP-Bench (Chung et al., 2025)	57	Grades	Mixed	Theoretical Physics	
UGPhysics (Xu et al., 2025e)	5520	Accuracy	University-level	Bilingual Physics Problem	
PhysReason (Zhang et al., 2025g)	1200	Step & Answer	Mixed	Physics-based Reasoning	
PHYBench (Qiu et al., 2025)	500	Accuracy & EDD	Mixed	meticulously curated	
MedQA (Jin et al., 2020)	1.3K	Accuracy	Clinical diagnostics	Medical accuracy	
JAMA (Chen et al., 2025a)	1,524	Accuracy	Challenging clinical cases	Expert-written explanations	Medical
Medbullets (Chen et al., 2025a)	308	Accuracy	USMLE Step 2/3 clinical questions	Expert-written explanations	
Others					
AGIEval (Zhong et al., 2024)	8K	Accuracy	College exams	Human-centric reasoning	
MMLU-Pro (Wang et al., 2024h)	12K	Accuracy	Multidisciplinary tests	Deep reasoning complexity	
C-Eval (Huang et al., 2023)	13.9K	Accuracy	Chinese exams	Multidisciplinary reasoning	
Gaokao (NCEE, 2025)	Varied	Accuracy	Chinese college exams	Broad knowledge	Basic
Kaoyan (GSEE, 2025)	Varied	Accuracy	Graduate entry exams	Specialized knowledge	
CMMLU (Li et al., 2024)	Varied	Accuracy	Multi-task Chinese eval	Comprehensive coverage	
LongBench (Bai et al., 2024)	Varied	Accuracy	Bilingual multi-task eval	Long-form reasoning	
IF-Eval (Zhou et al., 2023b)	541	Accuracy	Instruction adherence	Objective evaluation	
ArenaHard (Li et al., 2024c)	500	Human preference	Open-ended creativity	Human alignment	Open-ended
Chatbot Arena (Zheng et al., 2023a)	Varied	Human alignment	Chatbot quality	User-aligned responses	
AlpacaEval2.0 (Dubois et al., 2024)	805	Win rate	Chatbot responses	Debiased evaluation	
WebShop (Yao et al., 2023a)	1.18M	Task success	Online shopping	Real-world interaction	
WebArena (Zhou et al., 2023c)	Varied	Task completion	Web navigation tasks	Adaptive decision-making	
SciWorld (Wang et al., 2022)	30 tasks	Task-specific scores	Scientific experiments	Interactive simulation	Agentic
TextCraft (Prasad et al., 2024)	Varied	Success rate	Task decomposition	Iterative planning	
SimpleQA (Wei et al., 2024a)	4.3K	Accuracy	Short queries	Factual correctness	
C-SimpleQA (He et al., 2024c)	3K	Accuracy	Chinese queries	Cultural relevance	Knowledge
FRAMES (Krishna et al., 2025)	824	Accuracy	Multi-hop queries	Source aggregation	
RewardBench (Lambert et al., 2024)	2,985	Accuracy	Chat,Safety,Reasoning knowledge, reasoning, math, and coding	Multiple Domains General Reward	
JudgeBench (Tan et al., 2025)	350	Accuracy		Challenging Tasks	
RMBench (Liu et al., 2024b)	1,327	Accuracy		subtle differences and style biases	Evaluation
PPE (Frick et al., 2024)	16,038	Accuracy		Real-world preference	
RMB (Zhou et al., 2025)	3,197	Accuracy	49 fine-grained real-world scenarios	Closely related to alignment objectives	
MMMU (Yue et al., 2024)	11.5K	Accuracy	Multimodal expert tasks	Multidisciplinary integration	
MathVista (Lu et al., 2024)	6.1K	Accuracy	Visual math reasoning	Visual-math integration	
MATH-Vision (Wang et al., 2024d)	3K	Accuracy	Visual math problems	Multimodal math reasoning	
LLAVA-Wild (Liu et al., 2023a)	Varied	GPT-4 score	Visual QA	Complex visuals	
MM-Vet (Yu et al., 2024d)	Varied	GPT-4 evaluation	Integrated multimodal	Multi-capability eval	Multimodal
MMBench (Liu et al., 2024d)	3.2K	Accuracy	Diverse multimodal	Fine-grained eval	
CVBench (Tong et al., 2024)	Varied	Accuracy	Vision tasks	High-quality eval	
MMStar (Chen et al., 2024c)	1.5K	Accuracy	Vision-critical QA	Visual reliance	
CHAIR (Rohrbach et al., 2018)	Varied	Hallucination rate	Image captioning	Object hallucination	

Table 4: Summary of Benchmarks

5 How Well to Scale

In this section, we classify the metrics used in evaluating the test-time scaling methods into four high-level dimensions: **Performance**, **Controllability**, **Scalability**, and **Efficiency**. Each dimension captures an essential aspect critical to assessing test-time scaling approaches.

5.1 Performance

Performance metrics assess the correctness of generated solutions.

Pass@1. Pass@1 is one of the most widely used metrics for evaluating the correctness of a model’s first output attempt (DeepSeek-AI, 2025; Li et al., 2025d; Snell et al., 2024; Xie et al., 2025; Kimi, 2025; Yang et al., 2025b,a; Hou et al., 2025). It measures the proportion of problems where the model’s first generated solution is correct. A correct solution means the one that exactly matches the ground-truth answer or passes all required

validation checks, such as the exact answer match in mathematical benchmarks and private unit tests in coding tasks. Pass@1 is frequently used in tasks such as mathematical reasoning and coding benchmarks. In mathematical reasoning tasks such as **AIME 2024** (Google, 2025) and **MATH-500** (Zhang et al., 2024a), Pass@1 measures the percentage of exact matches between the model’s answer and the ground truth. In coding benchmarks such as **LiveCodeBench** (Jain et al., 2025) and **HumanEval-Mul**, Pass@1 evaluates the code correctness against hidden test cases.

Pass@k (Coverage). Pass@k extends Pass@1 by measuring whether at least one of the model’s k sampled outputs is correct (Brown et al., 2024; Snell et al., 2024; Li et al., 2025d). Formally, Pass@k can be estimated using the unbiased estimator from Chen et al. (2021):

$$\text{Pass}@k = \frac{1}{n} \sum_{i=1}^n \left(1 - \frac{\binom{N-C_i}{k}}{\binom{N}{k}} \right),$$

where n is the number of problems, N is the total number of samples per problem, and C_i is the number of correct samples for the i -th problem. Pass@k is widely adopted in program synthesis and formal theorem-proving tasks, such as **CodeContests** (Li et al., 2022) and **SWE-bench Lite** (Jimenez et al., 2024).

Cons@k (Consensus@k). Cons@k measures the majority vote correctness from k independently sampled outputs (DeepSeek-AI, 2025; Zeng et al., 2025d). Given k responses generated by a model for a given problem, the majority-voted prediction is the most frequent answer. The answer is then compared against the ground truth. Cons@k is frequently used alongside pass@1 to assess the benefit of leveraging multiple samples. Larger values of k (e.g., 16, 64) typically improve answer stability and accuracy but at the cost of increased compute. This metric is especially valuable in tasks where single generations may be noisy or uncertain, and ensemble strategies can improve robustness. Cons@k has been widely adopted in mathematical reasoning benchmarks such as **AIME 2024** (Google, 2025) and **MATH-500** (Zhang et al., 2024a).

Arena-based Evaluation (Pairwise Win Rate). In addition to accuracy-oriented metrics, some studies adopt pairwise comparison metrics, where model outputs are compared against baselines using human or LLM-based judges (DeepSeek-AI, 2025; Hou et al., 2025). For instance, **LC-Winrate** (Dubois et al., 2024) adjusts win rates to control for response length, while **ArenaHard GPT-4 Judge** (Li et al., 2024c) uses GPT-4-Turbo to score outputs from open-ended tasks. These pairwise evaluation methods are especially common in generation tasks where qualitative assessments (e.g., fluency, coherence) matter.

Task-Specific Metrics. Certain domains employ specialized metrics. For example, **Codeforces Percentile** and **Elo Rating** are used to measure coding capabilities under competitive programming settings (DeepSeek-AI, 2025; Kimi, 2025). Percentile indicates how well a model performs relative to other participants, while Elo Rating reflects relative skill under tournament-based evaluations.

5.2 Efficiency

Efficiency should not be treated as a monolithic concept. Instead, it spans multiple dimensions, encompassing both computational cost and the quality of the reasoning process. Different perspectives capture not only how many resources (e.g., tokens, FLOPs, memory) a model consumes, but also how effectively it reasons—whether it produces concise and goal-directed outputs, or falls into patterns of repetition and over-generation.

5.2.1 Key Concern and Multi-Dimensional Perspectives of Efficiency

A central challenge in efficiency is the trade-off between the *reasoning length* and *solution quality*. The formal notion of **reasoning efficiency** captures this trade-off, as introduced by (Qu et al., 2025). Let \mathcal{M} represent a language reasoning model evaluated over a distribution of tasks $\mathcal{T} \sim p(\mathcal{T})$, where each task consists of a dataset \mathcal{D} and an associated question or objective. The reasoning efficiency $\eta(\mathcal{M})$ is defined as the expected ratio between solution quality Q and computational cost C :

$$\eta(\mathcal{M}) = \mathbb{E}_{\mathcal{T} \sim p(\mathcal{T})} \left[\frac{Q(\mathcal{M}, \mathcal{D})}{C(\mathcal{M}, \mathcal{D})} \right], \quad (4)$$

where $Q(\mathcal{M}, \mathcal{D})$ quantifies task performance (e.g., accuracy, exact match, or creativity), and $C(\mathcal{M}, \mathcal{D})$ measures the computation involved (e.g., number of generated tokens, FLOPs, or latency). This concept highlights how excessively long reasoning chains may not yield proportionate gains in solution quality and often reflect redundancy in reasoning traces.

Recent work has identified multi-dimensional perspectives of inefficiency patterns in reasoning LLMs. These include **redundancy**, where models show excessive repetition of reasoning steps or thoughts, often restating

questions or repeating points without contributing to solution quality (Song et al., 2025; Su et al., 2025; Luo et al., 2025a); **underthinking**, where models shift reasoning direction too early or fail to elaborate on promising ideas (Wang et al., 2025e); and **overthinking**, where models revisit or verify earlier steps excessively even for simple problems (Luo et al., 2025a; Chiang and Lee, 2024; Chen et al., 2024g). These behaviors contribute to higher computational cost without proportional gains in correctness, ultimately lowering overall reasoning efficiency.

5.2.2 General Computational Cost Metrics

To quantify inference efficiency, we consider the following standard computational metrics:

Token Cost. Token cost measures the total number of tokens generated during inference, including intermediate reasoning steps and final outputs (Welleck et al., 2024; Brown et al., 2024; Hou et al., 2025; Yang et al., 2025b; Xu et al., 2025d; Wang et al., 2025c; Aytes et al., 2025). This metric is especially important, as verbose reasoning typically leads to higher token consumption. Reducing token cost while maintaining performance is crucial for inference efficiency, particularly when operating under fixed computational budgets or API pricing constraints. In addition, inference efficiency metrics such as latency and throughput are critical in real-world applications, especially for high-throughput systems (Welleck et al., 2024).

FLOPs-based Efficiency Analysis. FLOPs-based compute analysis has been widely adopted to quantify computational cost (Kaplan et al., 2020; Snell et al., 2024; Wu et al., 2024c; Teng et al., 2025). Several recent works (Snell et al., 2024; Wu et al., 2024c) benchmark test-time scaling strategies, such as adaptive revisions and verifier-based search, against model scaling by plotting accuracy versus total inference FLOPs. This FLOPs-based evaluation can be used to determine whether test-time methods outperform larger models under equivalent compute budgets.

KV Cache Size. The *KV cache size* (Hooper et al., 2025) refers to the total memory footprint required to store the Key-Value cache across all trajectories and time steps during the test-time search process. As each unique generation path requires its own KV cache, methods with low KV sharing across trajectories tend to consume significantly more memory and incur higher latency. By promoting KV cache sharing among trajectories, ETS reduces the total KV cache size, thereby improving throughput. For instance, ETS achieves up to $1.8\times$ KV cache reduction compared to REBASE, leading to $1.4\times$ faster inference on NVIDIA H100 GPUs, *without compromising accuracy*.

5.2.3 Reasoning Efficiency Metrics

Beyond standard compute metrics, reasoning inefficiency can also be measured through the following metrics.

Underthinking Score. The underthinking score (Wang et al., 2025e) quantifies cases where a model produces correct intermediate thoughts early in the reasoning chain but fails to reach a correct final answer.

Formally, the underthinking score ξ_{UT} is defined as:

$$\xi_{\text{UT}} = \frac{1}{N_{\text{inc}}} \sum_{i=1}^{N_{\text{inc}}} \left(1 - \frac{\hat{T}_i}{T_i} \right), \quad (5)$$

where N_{inc} is the number of incorrect responses in the test set, T_i is the total number of tokens in the i -th incorrect response, and \hat{T}_i is the position (in tokens) of the first correct intermediate thought. If no correct intermediate thought exists, $\hat{T}_i = T_i$, yielding a score of zero for that instance. A higher ξ_{UT} indicates inefficient reasoning behavior where good ideas are introduced but not effectively pursued.

Outcome Efficiency. The outcome efficiency metric (Chen et al., 2024g) quantifies how economically a model reaches the correct answer in multi-turn reasoning. It reflects whether additional solution rounds meaningfully contribute to accuracy. Empirically, the metric captures the ratio of efficient tokens—those used before the first correct answer appears—to the total number of output tokens. Formally, the outcome efficiency ξ_{O} is defined as:

$$\xi_{\text{O}} = \frac{1}{N} \sum_{i=1}^N \sigma_i \cdot \frac{\hat{T}_i}{T_i}, \quad (6)$$

where N is the number of test instances, T_i denotes the total tokens generated for the i -th instance, and \hat{T}_i is the number of tokens needed to reach the first correct solution. The correctness indicator σ_i is set to 1 if at least one correct solution appears in the response, and 0 otherwise. A higher ξ_{O} suggests that the model typically finds correct answers early in the reasoning process, while a low value indicates inefficient use of computation due to overextended or redundant reasoning chains.

Process Efficiency. The process efficiency metric (Chen et al., 2024g) evaluates how productively a model explores diverse reasoning strategies across multiple solution rounds. Rather than focusing solely on accuracy, it assesses the novelty of later solutions by measuring how many output tokens contribute to distinct reasoning approaches. Formally, the process efficiency ξ_P is defined as:

$$\xi_P = \frac{1}{N} \sum_{i=1}^N \frac{D_i}{T_i}, \quad (7)$$

where N is the number of test instances, D_i is the number of tokens from non-redundant (i.e., distinct) solutions in the i -th response, and T_i is the total number of tokens. A solution is considered distinct if its reasoning strategy differs significantly from those in earlier turns. A higher ξ_P value implies more efficient cognitive exploration, while a lower value indicates repetitive or self-verifying reasoning that fails to expand the model’s strategic diversity.

5.3 Controllability

Controllability metrics evaluate whether test-time methods can consistently adhere to pre-defined resource constraints such as compute budgets or output length targets.

Control Metric. Muenninghoff et al. (2025) propose **Control** as a formal metric to quantify adherence to a specified compute budget range. It measures the fraction of test-time compute values that stay within given upper and lower bounds:

$$\text{Control} = \frac{1}{|\mathcal{A}|} \sum_{a \in \mathcal{A}} \mathbb{I}(a_{\min} \leq a \leq a_{\max}),$$

where \mathcal{A} is the set of observed compute values such as thinking tokens, and $\mathbb{I}(\cdot)$ is the indicator function. A score of 100% denotes perfect adherence to the compute budget across all tasks. Additionally, Hou et al. (2025) and Yang et al. (2025b) report experiments where models are evaluated under fixed token budgets, e.g., 1024, 2048, 4096, to examine how well models meet pre-specified length or token constraints during reasoning. Moreover, Xie et al. (2025) and Teng et al. (2025) impose explicit constraints on maximum output lengths to ensure test-time stability and prevent output truncation.

Length Deviation Metrics. Mean Deviation from Target Length and RMSE of Length Deviation are introduced to quantify a model’s ability to control output length (Aggarwal and Welleck, 2025):

- **Mean Deviation from Target Length** quantifies the average relative difference between the generated output length and the target length:

$$\text{Mean Deviation} = \mathbb{E}_{x \sim D} \left[\frac{|n_{\text{generated}} - n_{\text{gold}}|}{n_{\text{gold}}} \right],$$

where $n_{\text{generated}}$ is the model’s output length and n_{gold} is the target length.

- **Root Mean Squared Error (RMSE) of Length Deviation** captures the variance in length control:

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^N \left(\frac{n_{\text{generated},i} - n_{\text{gold},i}}{n_{\text{gold},i}} \right)^2}.$$

Lower values for both metrics indicate more stable and precise length control across samples.

k - ϵ Controllability. Bhargava et al. (2024) propose **k - ϵ controllability** as a formal metric to characterize the prompt-based steerability of language models. Unlike metrics focused on compute or length constraints, this metric quantifies whether a model can be guided to produce a target output within a bounded prompt length and allowable deviation. Formally, a model is said to be (k, ϵ) -controllable for a target output y if there exists a prompt p with $|p| \leq k$ such that the model outputs y with probability at least $1 - \epsilon$:

$$\Pr[\text{LLM}(p) = y] \geq 1 - \epsilon.$$

By evaluating across different values of k and ϵ , one can map out the controllability landscape of a model. In practice, Bhargava et al. (2024) measures this property on tasks such as next-token prediction in WikiText, finding that over 97% of targets are reachable with a prompt of at most 10 tokens and an error tolerance $\epsilon \leq 0.05$. This metric provides a theoretical lens for quantifying how easily a model’s outputs can be controlled via prompt design. While not directly tied to resource constraints, k - ϵ controllability offers valuable insight into the model’s test-time responsiveness and has been used to compare inherent steerability across model families and sizes.

5.4 Scalability

5.4 Scalability

Scalability metrics measure how effectively test-time scaling methods can leverage increased compute (e.g., token budgets, samples, inference steps) to improve performance.

Scaling Metric Muennighoff et al. (2025) propose the **Scaling** metric, capturing the average slope of performance gains as compute increases:

$$\text{Scaling} = \frac{1}{\binom{|\mathcal{A}|}{2}} \sum_{\substack{a,b \in \mathcal{A} \\ b > a}} \frac{f(b) - f(a)}{b - a}.$$

This metric quantifies how effectively models improve accuracy or pass rates with additional computation.

Scaling Curves (Accuracy vs. Compute). Scaling curves are used to visualize how metrics such as accuracy, pass rate, or EM improve as token budgets, iteration depth, or the number of samples increase (Aggarwal and Welleck, 2025; Teng et al., 2025; Wu et al., 2024c). These plots help reveal diminishing returns and performance saturation at higher compute budgets.

6 Organization and Trends in Test-time scaling

Method	WHAT	HOW						WHERE	HOW WELL
		SFT	RL	STIMULATION	SEARCH	VERIFICATION	AGGREGATION		
DSC (Snell et al., 2024)	Parallel, Sequential	✗	✗	✗	Beam Search, LookAhead Search	Verifier	(Weighted) Best-of-N, Stepwise Aggregation	Math	Pass@1, FLOPs-Matched Evaluation
MAV (Lifshitz et al., 2025)	Parallel	✗	✗	Self-Repetition	✗	Multiple-Agent Verifiers	Best-of-N	Math, Code, General	BoN-MAV (Cons@k), Pass@1
Mind Evolution (Lee et al., 2025)	Sequential	✗	✗	Self-Refine	✗	Functional	✗	Open-Ended	Success Rate, Token Cost
Meta-Reasoner (Sui et al., 2025)	Sequential	✗	✗	CoT + Self-Repetition	✗	Bandit	✗	Game, Sci, Math	Accuracy, Token Cost
START (Li et al., 2025b)	Parallel, Sequential Sampling	✗	✗	Hint-infer	✗	Tool	✗	Math, Code	Pass@1
AID (Jin et al., 2025)	Sequential	✗	✗	Adaptive Injection Decoding	✗	✗	✗	Math, Logical, Commonsense	Accuracy
CoD (Xu et al., 2025b)	Sequential	✗	✗	Chain-of-Draft	✗	✗	✗	Math, Symbolic, Commonsense	Accuracy, Latency, Token Cost
rStar-Math (Guan et al., 2025)	Hybrid	imitation	✗	✗	MCTS	PRM	✗	MATH	Pass@1
(Liu et al., 2025a)	Parallel, Hybrid	✗	✗	✗	DVTS, Beam Search	PRM	Best-of-N	Math	Pass@1, Pass@k, Majority, FLOPS
Tree of Thoughts (Yao et al., 2023b)	Hybrid	✗	✗	Propose prompt Self-Repetition	Tree Search	Self-Evaluate	✗	GAME, Open-Ended	Success Rate, LLM-as-a-Judge
MindStar (Kang et al., 2024)	Hybrid	✗	✗	✗	LevinTS	PRM	✗	MATH	Accuracy, Token Cost
REBASE (Wu et al., 2025a)	Hybrid	✗	✗	✗	Reward Balanced Search	RM	✗	Math	Test Error Rate, FLOPs
RaLU (Li et al., 2025c)	Hybrid	✗	✗	Self-Refine	Control Flow Graph	Self-Evaluate	Prompt Synthesis	MATH, Code	Pass@1
PlanGen (Parmar et al., 2025)	Parallel, Hybrid	✗	✗	MoA	✗	Verification agent	Selection Agent	Math, General, Finance	Accuracy, F1 Score
Puri et al. (2025)	Hybrid	✗	✗	✗	Particle-based Monte Carlo	PRM+SSM	Particle filtering	MATH	Pass@1, Accuracy
Archon (Saad-Falcon et al., 2024)	Hybrid	✗	✗	MoA, Self-Repetition	✗	Verification agent, Unit Testing	(Ensemble) Fusion	Math, Code, Open-Ended	Pass@1, Win Rate
AB-MCTS (Mitsaki et al., 2025)	Hybrid	✗	✗	Mixture-of-Model	AB-MCTS-(M,A)	✗	✗	Code	Pass@1, RMSLE, ROC-AUC
TPO (Wu et al., 2024b)	Internal, Parallel	✗	DPO	Think	✗	Judge models	✗	Open-Ended	Win Rate
SPHERE (Singh et al., 2025)	Internal, Hybrid	✗	DPO	Diversity Generation	MCTS	Self-Reflect	✗	Math	Pass@1
MA-Lot (Wang et al., 2025b)	Internal, Sequential	imitation	✗	MoA	✗	Tool	✗	Math	Pass@k
OREO (Wang et al., 2024b)	Internal, Sequential	✗	OREO	✗	Beam Search	Value Function	✗	Math, Agent	Pass@1, Success Rate
DeepSeek-R1 (DeepSeek-AI, 2025)	Internal	warmup	GRPO, Rule-Based	✗	✗	✗	✗	Math, Code, Sci	Pass@1, cons@64, Percentile, Elo Rating, Win Rate
s1 (Muennighoff et al., 2025)	Internal	distillation	✗	Budget Forcing	✗	✗	✗	Math, Sci	Pass@1, Control, Scaling
o1-Replication (Qin et al., 2024)	Internal	imitation	✗	✗	Journey Learning	PRM, Critique	Multi-Agents	Math	Accuracy
AFT (Li et al., 2025f)	Internal, Parallel	imitation	✗	✗	✗	✗	Fusion	Math, Open-Ended	Win Rate
Meta-CoT (Xiang et al., 2025)	Internal, Hybrid	imitation	meta-RL	Think	MCTS,A*	PRM	✗	Math, Open-Ended	Win Rate
Reasonflux (Yang et al., 2025a)	Internal, Sequential	✗	PPO, Trajectory	Thought Template	Retrieve	✗	✗	Math	Pass@1
II (Aggarwal and Welleck, 2025)	Internal	✗	GRPO, Length-Penalty	✗	✗	✗	✗	Math	Pass@1, Length Error
Marco-o1 (Zhao et al., 2024)	Internal, Hybrid	distillation, imitation	✗	Reflection Prompt	MCTS	Self-Critic	✗	Math	Pass@1, Pass@k

Table 5: Commonly-used combinations in existing literature when conducting inference scaling.

Building on our taxonomy, we decompose the existing literature along multiple dimensions (Table 5). As shown in Figure 4, these works, with different technical innovations, follow a broadly consistent path. From 2022 to 2023, researchers emphasized structured inference to guide LLMs in generating more complex solutions. In 2024, methods like PRM and MCTS enabled the automatic supervision of intricate reasoning trajectories, yielding richly annotated data for fine-tuning and improving TTS performance. Subsequent approaches, such as o1 and R1, demonstrated that pure RL can also elicit comprehensive, logically sound reasoning.

- Crucially, these techniques are complementary rather than mutually exclusive: for instance, R1 necessitates an SFT-based warmup via rejection sampling. Therefore, achieving more powerful scaling requires systematically

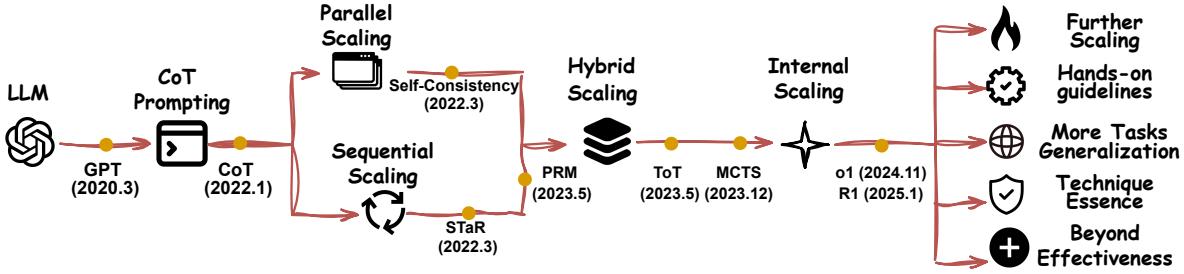


Figure 4: From Emergence to the Next Frontier, the Evolutionary Path of Test-Time Scaling.

integrating these methods. Even within RL frameworks, practitioners should continue to leverage synthesized CoT approaches and incorporate structured inference strategies to tackle increasingly complex scenarios effectively.

- Researchers found that there does not exist one simple scaling solution that works for all problems. Increasingly, researchers tend to focus on optimal-scaling solutions (Wu et al., 2024d; Snell et al., 2024).
- The boundary between inference-based and tuning-based approaches is blurring. Consequentially, the target of scaling (*what to scale*) changes between different stages. Certain papers, such as Li et al. (2025f); Munkhbat et al. (2025), tune the inference-based capability into the LLM by synthesizing high-quality data from inference-based approaches as the tuning data. Others, such as Wan et al. (2024), are proposing various techniques that better exploit the LLM’s capability during both the training and inference stages.

7 A Hand-on Guideline for Test-time Scaling

In this section, we shift from theoretical categorizations to providing a practical, hands-on guideline for *TTS*. Our goal is to offer clear, actionable instructions and technical pathways to facilitate effective SST deployment.

❓ Hands-on Guidelines: Common Problems

❓ Q: What kind of task does *TTS* help?

✓ A: Almost any task! While traditional reasoning tasks—such as Olympiad-level mathematics, complex coding, and game-based challenges—have been shown to significantly improve with *TTS*, community observations suggest that *TTS* can also enhance performance in open-ended tasks, such as comment generation or evaluation. However, due to the long-form nature of outputs and the lack of centralized, objective benchmarks, these tasks are inherently more difficult to evaluate quantitatively, making it harder to draw conclusive claims. Beyond that, more realistic, complex, and long-horizon scenarios, like medical reasoning and law, have also shown promising gains through *TTS* strategies.

❓ Q: If I want to quickly implement a *TTS* pipeline, what are the essential paths I should consider? How can beginners use *TTS* at a minimal cost?

✓ A: Broadly speaking, there are three essential technical pathways for test-time scaling: i) Deliberate reasoning procedure at inference time, ii) imitating complex reasoning trajectories, and iii) RL-based incentivization. If your goal is to get a quick sense of the potential upper bound that a strong *TTS* can bring to your task at a minimum cost, you can directly utilize a model that has been trained with (iii). If you want to develop a *TTS* baseline at a minimum cost, you can start with (i). Once (i) yields a result that meets expectations, you can apply (ii) to further verify and generalize the outcome.

❓ Q: Are these pipelines mutually exclusive? How should I design a frontier-level *TTS* strategy?

✓ A: These pipelines are by no means mutually exclusive—they can be seamlessly integrated. For instance, R1 inherently necessitates SFT through rejection sampling as a preliminary warmup step. When employing RL, practitioners should continue leveraging synthesized CoT methods and introduce additional structured inference strategies to tackle increasingly complex scenarios effectively.

❓ Q: Can training-time choices like SFT and RL affect the effectiveness of *TTS*?

✓ A: Yes, training-time strategies often shape the ceiling of what *TTS* methods can achieve at inference. For example, SFT can provide strong reasoning priors that improve the stability and quality of scaling strategies like self-consistency or verifier-based search. On the other hand, RL-based fine-tuning can explicitly

incentivize concise and correct reasoning chains, reducing overthinking or incoherent outputs at test time. Both approaches can also be combined: SFT initializes the policy, while RL further sharpens it—this is the strategy behind systems like DeepSeek-R1. Alternatively, RL-trained models can be distilled back into SFT models to reduce inference-time costs without losing the benefits of RL optimization.

❓ Q: How can we improve the efficiency of *TTS* in multi-turn setups?

✓ A: Improving *TTS* efficiency involves both structural design and learned behaviors: (1) *Model-based*: Fine-tune models to generate more concise and purposeful reasoning traces. Reinforcement learning (RL) or reward-guided training can penalize overlong or redundant outputs while preserving correctness. Distillation from overthinking-prone models into lighter, more efficient ones can further reduce reasoning overhead. (2) *Output-based*: Apply early stopping strategies based on verifier confidence or answer agreement to terminate reasoning once sufficient certainty is achieved. This avoids unnecessary continuation in well-understood problems. (3) *Prompt-based*: Use prompt-level controls—such as token budgets, step limits, or task difficulty cues—to guide the model toward more targeted and efficient reasoning paths at inference. These strategies together help reduce computational cost while maintaining strong solution quality during *TTS*.

❓ Q: What are some representative or widely-used *TTS* methods that can serve as baselines?

✓ A: Parallel-Self-Consistency, Best-of-N; Sequential-STaR, Self-Refine, PRM; Hybrid-MCTS, ToT; Internal-Distilled-R1, R1.

❓ Q: Is there an optimal go-to solution so far?

✓ A: No free lunch. Optimal computing is often dependent on the hardness and openness of the question.

❓ Q: How should we evaluate the performance of a *TTS* method? In addition to standard accuracy, what other aspects should we pay attention to?

✓ A: The evaluation is largely task-aware, but metrics like accuracy remain the most critical indicators. In addition, efficiency (the trade-off between performance and cost) is another key concern in practical settings. As *TTS* becomes a more general-purpose strategy, researchers have also begun evaluating a range of secondary attributes, including robustness, safety, bias, and interpretability, to better understand the broader impacts of *TTS*.

❓ Q: Is there any difference when tuning other scaling formats into internal scaling, compared with directly using the original scaling format?

✓ A: Yes, one intuitive difference lies in the efficiency aspect. Internal scaling tends to yield higher efficiency as it only prompts the LM once, while other scaling techniques usually require multiple trials. However, internal scaling requires non-negligible resources for tuning, making it less available for practitioners.

8 Challenges and Opportunities

8.1 More Scaling is the Frontier

Pushing AI toward more general intelligence, especially for complex tasks, test-time scaling has emerged as one of the most promising methodologies in the post-pretraining era. Given its transformative impact on reasoning-intensive tasks—as seen in models like OpenAI’s o1 and DeepSeek-R1—it is increasingly clear that realizing the full promise of test-time scaling remains a central pillar in advancing AGI. However, to push the frontier further, we need new and more effective strategies. There are some several promising research directions:

Parallel Scaling. Parallel scaling improves solution reliability by generating multiple responses and selecting the best answer. Despite its effectiveness, parallel scaling remains has diminishing returns when coverage reaches saturation. A key challenge is how to enhance coverage, shifting from brute-force coverage expansion to a more guided, efficient process. Possible future advancements include:

1. *Smart Coverage Expansion*: Instead of naive best-of-N sampling, a model could intelligently generate diverse reasoning paths, ensuring each sampled response explores a meaningfully different approach;
2. *Verifier-Augmented Parallel Scaling*: Integrating real-time verification mechanisms could allow parallel samples to be filtered dynamically.

Sequential Scaling. Sequential scaling faces unique challenges, particularly in maintaining coherence and preventing error accumulation. A key issue is optimizing stepwise reasoning to avoid diminishing returns or reinforcing incorrect steps. Instead of naive iterative refinement, future advancements should focus on more

8.2 Clarifying the Essence of Techniques in Scaling is the Foundation

adaptive and structured approaches to ensure each reasoning step meaningfully improves the final outcome. Possible directions include:

1. *Structured Self-Refinement*: Rather than blindly refining the entire response, models could learn to target specific parts of their reasoning that require adjustment.
2. *Verification-Enhanced Iterative Scaling*: Introducing real-time validation steps within the sequential reasoning process could prevent models from propagating early mistakes. This could involve running self-verification checks between iterations (*e.g.*, checking consistency with known facts, comparing intermediate results to prior context, or re-computing specific logical steps). By selectively verifying before proceeding, models can ensure high-quality stepwise improvements instead of compounding errors.

By addressing these challenges, sequential scaling can evolve beyond simple iterative refinement, becoming a highly adaptive, self-correcting reasoning paradigm that enables models to engage in goal-directed, long-horizon thinking.

Hybrid Scaling. Hybrid scaling blends parallel and sequential methods, making it more adaptive and practical for real-world applications. Current test-time scaling methods are often highly specialized, limiting their generalizability. To address these limitations, hybrid scaling can be improved in several ways:

1. *Generalized Hybrid Scaling Architectures*: research should focus on unifying test-time scaling mechanisms into a single framework that dynamically chooses the best strategy for different query types.
2. *Multi-Agent & Interactive Scaling*: Expanding hybrid scaling beyond a single-agent reasoning process could allow multiple model instances to engage in structured debate, argumentation, or negotiation, improving solution reliability. While current hybrid scaling is mostly studied in controlled benchmarks, future work must consider its role in real-world applications.

Internal Scaling. Internal scaling allows on-the-fly computation modulation without external intervention. While this paradigm has demonstrated promising results, it also introduces unique challenges.

1. *Effective Compute Allocation*: Ensuring that internal scaling allocates extra reasoning steps only where necessary is critical. If the model overthinks simple tasks or fails to extend reasoning on complex ones, the benefits of dynamic computation are lost.
2. *Stability and Consistency*: As models extend their own reasoning paths, they risk logical drift, hallucination, or over-complication. Unlike sequential scaling, which can incorporate external verification, internal scaling must maintain self-consistency without external guidance.
3. *Interpretability and Controllability*: Internal scaling happens implicitly, making it difficult to diagnose failures or regulate inference costs. Unlike parallel scaling (which provides multiple explicit outputs) or sequential scaling (which follows structured iterations), internal scaling lacks clear intermediate checkpoints, posing challenges for debugging and efficiency management.

By addressing these challenges, internal scaling has the potential to maximize efficiency, enhance model adaptability, and push AI systems toward more autonomous, self-regulating reasoning.

8.2 Clarifying the Essence of Techniques in Scaling is the Foundation

While *what to scale* continues to evolve and techniques further developing internally, such as PPO transitioning to GRPO, we observe that the core categories of scaling techniques remain relatively stable. For example, SFT and RL remain two of the most common approaches, though their roles and interactions have shifted over time. This raises an urgent need to deepen our understanding of how these fundamental techniques contribute to test-time scaling.

Here, we raise some potential directions for further investigation:

1. *Theoretical Gaps in Scaling Techniques*: How do core techniques (SFT, RL, reward modeling) contribute to test-time scaling? how should SFT and RL be optimally combined?
2. *Re-evaluating Reward Modeling*: whether PRMs actually improve multi-step inference? Does the classic reward model incorporate noise and unnecessary complexity?
3. *Mathematical Properties of Test-Time Scaling*: How does performance scale with increased inference steps? Is there an optimal stopping criterion? Are there fundamental constraints on how much test-time scaling can improve reasoning performance?

4. *Chain-of-Thought Reasoning Priorities*: which aspects of chain-of-thought are most crucial for effective test-time scaling?
5. *Adaptive Test-Time Scaling*: How can we make a model automatically adjust its inference process based on the problem at hand? As empirical observations on certain property models ([xAI, 2025](#)) show blindly scaling over test-time may lead to over-thinking.
6. *Thoughtology*: How do the reasoning patterns in its language help improve reasoning effectiveness by treating a finetuned reasoning model as an agent? Recent studies, such as [Marjanović et al. \(2025\)](#); [Wu et al. \(2024a\)](#), have also explored this question.

8.3 Optimizing Scaling is the Key

As new *TTS* methods proliferate, systematic evaluation and optimization become critical. We must comprehensively measure how different strategies perform regarding task accuracy and consider efficiency, robustness, bias, safety, interpretability, and more. Optimizing these aspects of *TTS* is gradually emerging ([Zhang et al., 2025a](#); [Huang et al., 2025b](#)) and will become an important part of future developments.

8.4 Generalization across Domains is the Mainstream

We anticipate a wave of research extending test-time scaling into a wider range of domains, such as medicine and finance, where complex decision-making and structured reasoning are critical. This expansion is both inevitable and promising, as test-time scaling offers a powerful mechanism to enhance reasoning depth, adapt computation dynamically, and improve accuracy without requiring costly retraining. Beyond these fields, we can expect widespread applications in law, AI evaluation, open-domain QA, and other high-stakes or knowledge-intensive areas. Despite its potential, scaling test-time reasoning across domains presents several key challenges:

1. *Balancing Cost and Accuracy*: Unlike general NLP tasks, specialized domains often require strict computational efficiency and reliability;
2. *Ensuring Domain-Specific Interpretability*: In fields like medicine and law, outputs must be transparent and justifiable;
3. *Integrating External Knowledge & Real-World Constraints*: Many domains require retrieval-augmented generation, real-time data analysis, or interactive query refinement;
4. Future research must identify generalizable test-time scaling strategies that are robust across diverse reasoning tasks.

By addressing these challenges, test-time scaling can become a foundational AI capability, enabling models to extend their own reasoning dynamically, adapt to real-world constraints, and generalize across specialized fields. This shift represents a paradigm change, where AI systems don't just memorize knowledge—they actively scale their intelligence at inference to meet the demands of diverse, evolving tasks.

9 Conclusion

This is the first survey to decompose *TTS* through a hierarchical taxonomy, offering a structured perspective that aids both conceptual understanding and the identification of individual contributions. Emphasizing practical utility, we introduce a hands-on guideline aligned with each taxonomy dimension, which we plan to expand over time. Based on this framework, we outline key trends, challenges, and opportunities shaping the future of *TTS* research.

Author Contributions

Below, we list the individual author contributions: Qiyuan Zhang and Fuyuan Lyu are core contributors who coordinate and finalize the full paper. Zexu Sun, Lei Wang, Weixu Zhang, Zhihan Guo, Wenyue Hua and Haolun Wu are significant contributors who are responsible for certain chapters of this paper. Yufei Wang provides the overall structures of the taxonomy and provides close supervision during the process. Niklas Muennighoff, Irwin King, Xue Liu, and Chen Ma provide insightful feedback and high-level suggestions on this survey overall.

References

- Pranjal Aggarwal and Sean Welleck. 2025. [L1: Controlling how long a reasoning model thinks with reinforcement learning](#). In *arXiv*.
- Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin, Ahmet Üstün, and Sara Hooker. 2024. [Back to basics: Revisiting reinforce-style optimization for learning from human feedback in llms](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12248–12267.
- aider. 2025. [Aider](#).
- AntResearch-RL-Lab. 2025. Areal: Ant reasoning rl. <https://github.com/inclusionAI/AReal>.
- Daman Arora, Himanshu Gaurav Singh, and Mausam . 2023. [Have LLMs advanced enough? a challenging problem solving benchmark for large language models](#). In *Conference on Empirical Methods in Natural Language Processing*.
- Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. [Self-rag: Learning to retrieve, generate, and critique through self-reflection](#). In *The Twelfth International Conference on Learning Representations*.
- Simon A. Ayes, Jinheon Baek, and Sung Ju Hwang. 2025. [Sketch-of-thought: Efficient llm reasoning with adaptive cognitive-inspired sketching](#). In *arXiv*.
- Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. 2022. [Constitutional ai: Harmlessness from ai feedback](#). In *arXiv*.
- Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. 2024. [Longbench: A bilingual, multitask benchmark for long context understanding](#). In *arXiv*.
- Bespoke. 2025. Bespoke-stratos: The unreasonable effectiveness of reasoning distillation. www.bespokelabs.ai/blog/bespoke-stratos-the-unreasonable-effectiveness-of-reasoning-distillation. Accessed: 2025-01-22.
- Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawska, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, and Torsten Hoefer. 2024. Graph of thoughts: Solving elaborate problems with large language models. *AAAI Conference on Artificial Intelligence*, page 17682–17690.
- Aman Bhargava, Cameron Witkowski, Shi-Zhuo Looi, and Matt Thomson. 2024. [What's the magic word? a control theory of llm prompting](#). In *arXiv*.
- 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*.
- Bradley Brown, Jordan Juravsky, Ryan Ehrlich, Ronald Clark, Quoc V. Le, Christopher Ré, and Azalia Mirhoseini. 2024. [Large language monkeys: Scaling inference compute with repeated sampling](#). In *arXiv*.
- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. [Language models are few-shot learners](#). In *arXiv*.
- Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. [Sparks of artificial general intelligence: Early experiments with gpt-4](#). In *arXiv*.

- Antoine Chaffin, Vincent Claveau, and Ewa Kijak. 2022. **Ppl-mcts: Constrained textual generation through discriminator-guided mcts decoding**. In *NAACL 2022-Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1–15.
- Souradip Chakraborty, Sujay Bhatt, Udari Madhushani Sehwag, Soumya Suwra Ghosal, Jiahao Qiu, Mengdi Wang, Dinesh Manocha, Furong Huang, Alec Koppel, and Sumitra Ganesh. 2025. **Collab: Controlled decoding using mixture of agents for LLM alignment**. In *International Conference on Learning Representations*.
- Guoxin Chen, Minpeng Liao, Chengxi Li, and Kai Fan. 2024a. **Alphamath almost zero: Process supervision without process**. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.
- Hanjie Chen, Zhouxiang Fang, Yash Singla, and Mark Dredze. 2025a. **Benchmarking large language models on answering and explaining challenging medical questions**. In *arXiv*.
- Justin Chih-Yao Chen, Swarnadeep Saha, and Mohit Bansal. 2024b. **Reconcile: Round-table conference improves reasoning via consensus among diverse llms**. In *arXiv*.
- Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. 2024c. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*.
- Lingjiao Chen, Jared Quincy Davis, Boris Hanin, Peter Bailis, Ion Stoica, Matei Zaharia, and James Zou. 2024d. Are more llm calls all you need? towards scaling laws of compound inference systems. *arXiv preprint arXiv:2403.02419*.
- Lingjiao Chen, Jared Quincy Davis, Boris Hanin, Peter Bailis, Ion Stoica, Matei Zaharia, and James Zou. 2024e. **Are more LLM calls all you need? towards the scaling properties of compound AI systems**. In *Conference on Neural Information Processing Systems*.
- Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harry Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.
- Nuo Chen, Zhiyuan Hu, Qingyun Zou, Jiaying Wu, Qian Wang, Bryan Hooi, and Bingsheng He. 2025b. **JudgeLRM: Large reasoning models as a judge**. In *arXiv*.
- Qiguang Chen, Libo Qin, Jinhao Liu, Dengyun Peng, Jiannan Guan, Peng Wang, Mengkang Hu, Yuhang Zhou, Te Gao, and Wanxiang Che. 2025c. **Towards reasoning era: A survey of long chain-of-thought for reasoning large language models**.
- Shuhao Chen, Weisen Jiang, Baijiang Lin, James T. Kwok, and Yu Zhang. 2024f. **RouterDC: Query-based router by dual contrastive learning for assembling large language models**. In *arXiv*.
- Weizhe Chen, Sven Koenig, and Bistra Dilkina. 2025d. **Iterative deepening sampling for large language models**. In *arXiv*.
- Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W Cohen. 2023a. **Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks**. *Transactions on Machine Learning Research*.
- Wenhu Chen, Ming Yin, Max Ku, Pan Lu, Yixin Wan, Xueguang Ma, Jianyu Xu, Xinyi Wang, and Tony Xia. 2023b. **TheoremQA: A theorem-driven question answering dataset**. In *Conference on Empirical Methods in Natural Language Processing*, pages 7889–7901.
- Xingyu Chen, Jiahao Xu, Tian Liang, Zhiwei He, Jianhui Pang, Dian Yu, Linfeng Song, Qiuzhi Liu, Mengfei Zhou, Zhuosheng Zhang, et al. 2024g. Do not think that much for $2+3=?$ on the overthinking of o1-like llms. *arXiv preprint arXiv:2412.21187*.
- Xinyun Chen, Maxwell Lin, Nathanael Schärli, and Denny Zhou. 2024h. **Teaching large language models to self-debug**. In *International Conference on Learning Representations*.
- Yezeng Chen, Zui Chen, and Yi Zhou. 2024i. **Brain-inspired two-stage approach: Enhancing mathematical reasoning by imitating human thought processes**. In *arXiv*.
- Zhenfang Chen, Delin Chen, Rui Sun, Wenjun Liu, and Chuang Gan. 2025e. Scaling autonomous agents via automatic reward modeling and planning. *arXiv preprint arXiv:2502.12130*.
- Zhenfang Chen, Delin Chen, Rui Sun, Wenjun Liu, and Chuang Gan. 2025f. **Scaling autonomous agents via automatic reward modeling and planning**. In *arXiv*.

References

- Ziru Chen, Michael White, Ray Mooney, Ali Payani, Yu Su, and Huan Sun. 2024j. When is tree search useful for LLM planning? it depends on the discriminator. In *Annual Meeting of the Association for Computational Linguistics*, pages 13659–13678.
- Jiale Cheng, Xiao Liu, Cunxiang Wang, Xiaotao Gu, Yida Lu, Dan Zhang, Yuxiao Dong, Jie Tang, Hongning Wang, and Minlie Huang. 2025. Spar: Self-play with tree-search refinement to improve instruction-following in large language models. In *arXiv*.
- Cheng-Han Chiang and Hung-yi Lee. 2024. Over-reasoning and redundant calculation of large language models. *arXiv preprint arXiv:2401.11467*.
- François Chollet. 2019. On the measure of intelligence. In *arXiv*.
- Sanjiban Choudhury. 2025. Process reward models for llm agents: Practical framework and directions. In *arXiv*.
- Daniel J. H. Chung, Zhiqi Gao, Yurii Kvasiuk, Tianyi Li, Moritz Münchmeyer, Maja Rudolph, Frederic Sala, and Sai Chaitanya Tadepalli. 2025. Theoretical physics benchmark (tpbench) - a dataset and study of AI reasoning capabilities in theoretical physics. *CoRR*, abs/2502.15815.
- CMS. 2025. Chinese national high school mathematics olympiad.
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.
- codeforce. 2025. [Codeforces](#).
- Rémi Coulom. 2006. Efficient selectivity and backup operators in monte-carlo tree search. In *International conference on computers and games*, pages 72–83. Springer.
- Ganqu Cui, Lifan Yuan, Zefan Wang, Hanbin Wang, Wendi Li, Bingxiang He, Yuchen Fan, Tianyu Yu, Qixin Xu, Weize Chen, et al. 2025. Process reinforcement through implicit rewards. *arXiv preprint arXiv:2502.01456*.
- DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. In *arXiv*.
- Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B. Hashimoto. 2024. Length-controlled alpacaeval: A simple way to debias automatic evaluators. In *arXiv*.
- Sasha Rush Edward Beeching, Lewis Tunstall. 2024. Scaling test-time compute with open models.
- Jonathan Evans. 1984. Heuristic and analytic processes in reasoning. *British Journal of Psychology*, 75(4):451–468.
- Kaiyue Feng, Yilun Zhao, Yixin Liu, Tianyu Yang, Chen Zhao, John Sous, and Arman Cohan. 2025. Physics: Benchmarking foundation models on university-level physics problem solving. *arXiv preprint arXiv:2503.21821*.
- Yu Feng, Phu Mon Htut, Zheng Qi, Wei Xiao, Manuel Mager, Nikolaos Pappas, Kishaloy Halder, Yang Li, Yassine Benajiba, and Dan Roth. 2024. Diverseagententropy: Quantifying black-box llm uncertainty through diverse perspectives and multi-agent interaction. In *arXiv*.
- Thomas Palmeira Ferraz, Kartik Mehta, Yu-Hsiang Lin, Haw-Shiuan Chang, Shereen Oraby, Sijia Liu, Vivek Subramanian, Tagyoung Chung, Mohit Bansal, and Nanyun Peng. 2024. Llm self-correction with decrim: Decompose, critique, and refine for enhanced following of instructions with multiple constraints. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7773–7812.
- Evan Frick, Tianle Li, Connor Chen, Wei-Lin Chiang, Anastasios N. Angelopoulos, Jiantao Jiao, Banghua Zhu, Joseph E. Gonzalez, and Ion Stoica. 2024. How to evaluate reward models for rlhf. In *arXiv*.
- Kanishk Gandhi, Denise Lee, Gabriel Grand, Muxin Liu, Winson Cheng, Archit Sharma, and Noah D. Goodman. 2024. Stream of search (sos): Learning to search in language. In *arXiv*.
- Bofei Gao, Zefan Cai, Runxin Xu, Peiyi Wang, Ce Zheng, Runji Lin, Keming Lu, Dayiheng Liu, Chang Zhou, Wen Xiao, Junjie Hu, Tianyu Liu, and Baobao Chang. 2024a. Llm critics help catch bugs in mathematics: Towards a better mathematical verifier with natural language feedback. In *arXiv*.
- Bofei Gao, Feifan Song, Zhe Yang, Zefan Cai, Yibo Miao, Chenghao Ma, Shanghaoran Quan, Liang Chen, Qingxiu Dong, Runxin Xu, Zhengyang Tang, Benyou Wang, Daoguang Zan, Ge Zhang, Lei Li, Lei Sha, Yichang Zhang, Xuancheng Ren, Tianyu Liu, and Baobao Chang. 2025a. Omni-MATH: A universal olympiad level mathematic benchmark for large language models. In *International Conference on Learning Representations*.

- Chengqian Gao, Haonan Li, Liu Liu, Zeke Xie, Peilin Zhao, and Zhiqiang Xu. 2025b. Principled data selection for alignment: The hidden risks of difficult examples. *arXiv preprint arXiv:2502.09650*.
- Zitian Gao, Boye Niu, Xuzheng He, Haotian Xu, Hongzhang Liu, Aiwei Liu, Xuming Hu, and Lijie Wen. 2024b. Interpretable contrastive monte carlo tree search reasoning. In *arXiv*.
- Jonas Geiping, Sean McLeish, Neel Jain, John Kirchenbauer, Siddharth Singh, Brian R. Bartoldson, Bhavya Kailkhura, Abhinav Bhatele, and Tom Goldstein. 2025. Scaling up test-time compute with latent reasoning: A recurrent depth approach. In *arXiv*.
- Elliot Glazer, Ege Erdil, Tamay Besiroglu, Diego Chicharro, Evan Chen, Alex Gunning, Caroline Falkman Olsson, Jean-Stanislas Denain, Anson Ho, Emily de Oliveira Santos, Olli Järvinen, Matthew Barnett, Robert Sandler, Matej Vrzala, Jaime Sevilla, Qiuyu Ren, Elizabeth Pratt, Lionel Levine, Grant Barkley, Natalie Stewart, Bogdan Grechuk, Tetiana Grechuk, Shreepranav Varma Enugandla, and Mark Wildon. 2024. Frontiermath: A benchmark for evaluating advanced mathematical reasoning in ai.
- Ben Goertzel. 2014. Artificial general intelligence: Concept, state of the art, and future prospects. *Journal of Artificial General Intelligence*, pages 1–48.
- Google. 2024. Gemini 2.0 flash thinking.
- Google. 2025. Aime problems and solutions.
- Zhibin Gou, Zhihong Shao, Yeyun Gong, yelong shen, Yujiu Yang, Nan Duan, and Weizhu Chen. 2024. CRITIC: Large language models can self-correct with tool-interactive critiquing. In *International Conference on Learning Representations*.
- GSEE. 2025. Chinese graduate school entrance examinations.
- Xinyu Guan, Li Lyra Zhang, Yifei Liu, Ning Shang, Youran Sun, Yi Zhu, Fan Yang, and Mao Yang. 2025. rstar-math: Small llms can master math reasoning with self-evolved deep thinking. In *arXiv*.
- Shangmin Guo, Biao Zhang, Tianlin Liu, Tianqi Liu, Misha Khalmans, Felipe Llinares, Alexandre Rame, Thomas Mesnard, Yao Zhao, Bilal Piot, Johan Ferret, and Mathieu Blondel. 2024. Direct language model alignment from online ai feedback.
- Tingxu Han, Zhenting Wang, Chunrong Fang, Shiyu Zhao, Shiqing Ma, and Zhenyu Chen. 2025. Token-budget-aware ILM reasoning. In *arXiv*.
- Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhitong Hu, Jason Weston, and Yuandong Tian. 2024. Training large language models to reason in a continuous latent space. *arXiv preprint arXiv:2412.06769*.
- 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. 2024a. OlympiadBench: A challenging benchmark for promoting AGI with olympiad-level bilingual multimodal scientific problems. In *Annual Meeting of the Association for Computational Linguistics*, pages 3828–3850.
- Chengbo He, Bochao Zou, Xin Li, Jiansheng Chen, Junliang Xing, and Huimin Ma. 2025. Enhancing ILM reasoning with multi-path collaborative reactive and reflection agents. In *arXiv*.
- Hongliang He, Wenlin Yao, Kaixin Ma, Wenhao Yu, Yong Dai, Hongming Zhang, Zhenzhong Lan, and Dong Yu. 2024b. Webvoyager: Building an end-to-end web agent with large multimodal models. In *arXiv*.
- Yancheng He, Shilong Li, Jiaheng Liu, Yinghui Tan, Weixun Wang, Hui Huang, Xingyuan Bu, Hangyu Guo, Chengwei Hu, Boren Zheng, Zhuoran Lin, Xuepeng Liu, Dekai Sun, Shirong Lin, Zhicheng Zheng, Xiaoyong Zhu, Wenbo Su, and Bo Zheng. 2024c. Chinese simpleqa: A chinese factuality evaluation for large language models.
- 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. In *Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
- Jacob Hilton, Jie Tang, and John Schulman. 2023. Scaling laws for single-agent reinforcement learning. *arXiv preprint arXiv:2301.13442*.
- Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. 2022. Training compute-optimal large language models. In *arXiv*.

- Ruixin Hong, Xinyu Pang, and Changshui Zhang. 2024. Advances in reasoning by prompting large language models: A survey. *Cybernetics and Intelligence*, pages 1–15.
- Coleman Hooper, Sehoon Kim, Suhong Moon, Kerem Dilmen, Monishwaran Maheswaran, Nicholas Lee, Michael W. Mahoney, Sophia Shao, Kurt Keutzer, and Amir Gholami. 2025. *Ets: Efficient tree search for inference-time scaling*. In *arXiv*.
- Arian Hosseini, Xingdi Yuan, Nikolay Malkin, Aaron Courville, Alessandro Sordoni, and Rishabh Agarwal. 2024. *V-star: Training verifiers for self-taught reasoners*. In *First Conference on Language Modeling*.
- Zhenyu Hou, Xin Lv, Rui Lu, Jiajie Zhang, Yujiang Li, Zijun Yao, Juanzi Li, Jie Tang, and Yuxiao Dong. 2025. *Advancing language model reasoning through reinforcement learning and inference scaling*.
- Jian Hu, Jason Klein Liu, and Shen Wei. 2025a. *Reinforce++: A simple and efficient approach for aligning large language models*. *arXiv preprint arXiv:2501.03262*.
- Jian Hu, Xibin Wu, Zilin Zhu, Xianyu, Weixun Wang, Dehao Zhang, and Yu Cao. 2024. *Openrlhf: An easy-to-use, scalable and high-performance rlhf framework*. *arXiv preprint arXiv:2405.11143*.
- Jingcheng Hu, Yinmin Zhang, Qi Han, Daxin Jiang, and Heung-Yeung Shum Xiangyu Zhang. 2025b. Openreasoner-zero: An open source approach to scaling reinforcement learning on the base model. <https://github.com/Open-Reasoner-Zero/Open-Reasoner-Zero>.
- Chenghua Huang, Lu Wang, Fangkai Yang, Pu Zhao, Zhixu Li, Qingwei Lin, Dongmei Zhang, Saravan Rajmohan, and Qi Zhang. 2025a. *Lean and mean: Decoupled value policy optimization with global value guidance*. *arXiv preprint arXiv:2502.16944*.
- Chengsong Huang, Langlin Huang, Jixuan Leng, Jiacheng Liu, and Jiaxin Huang. 2025b. *Efficient test-time scaling via self-calibration*. In *arXiv*.
- Yinya Huang, Xiaohan Lin, Zhengying Liu, Qingxing Cao, Huajian Xin, Haiming Wang, Zhengu Li, Linqi Song, and Xiaodan Liang. 2024a. *MUSTARD: mastering uniform synthesis of theorem and proof data*. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.
- Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, jiayi lei, Yao Fu, Maosong Sun, and Junxian He. 2023. *C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models*. In *Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
- Zhen Huang, Zengzhi Wang, Shijie Xia, Xuefeng Li, Haoyang Zou, Ruijie Xu, Run-Ze Fan, Lyumanshan Ye, Ethan Chern, Yixin Ye, Yikai Zhang, Yuqing Yang, Ting Wu, Binjie Wang, Shichao Sun, Yang Xiao, Yiyuan Li, Fan Zhou, Steffi Chern, Yiwei Qin, Yan Ma, Jiadi Su, Yixiu Liu, Yuxiang Zheng, Shaoting Zhang, Dahua Lin, Yu Qiao, and Pengfei Liu. 2024b. *Olympicarena: Benchmarking multi-discipline cognitive reasoning for superintelligent AI*. In *Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
- Zhen Huang, Haoyang Zou, Xuefeng Li, Yixiu Liu, Yuxiang Zheng, Ethan Chern, Shijie Xia, Yiwei Qin, Weizhe Yuan, and Pengfei Liu. 2024c. *O1 replication journey – part 2: Surpassing o1-preview through simple distillation, big progress or bitter lesson?* In *arXiv*.
- HuggingFace. 2025. *Open r1: A fully open reproduction of deepseek-r1*.
- Robert Irvine, Douglas Boubert, Vyas Raina, Adian Liusie, Ziyi Zhu, Vineet Mudupalli, Aliaksei Korshuk, Zongyi Liu, Fritz Cremer, Valentin Assassi, Christie-Carol Beauchamp, Xiaoding Lu, Thomas Rialan, and William Beauchamp. 2023. *Rewarding chatbots for real-world engagement with millions of users*. In *arXiv*.
- Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. 2025. *Livecodebench: Holistic and contamination free evaluation of large language models for code*. In *International Conference on Learning Representations*.
- Kaixuan Ji, Guanlin Liu, Ning Dai, Qingping Yang, Renjie Zheng, Zheng Wu, Chen Dun, Quanquan Gu, and Lin Yan. 2024. Enhancing multi-step reasoning abilities of language models through direct q-function optimization. *arXiv preprint arXiv:2410.09302*.
- 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*.
- Dongfu Jiang, Yishan Li, Ge Zhang, Wenhao Huang, Bill Yuchen Lin, and Wenhui Chen. 2024a. *Tigerscore: Towards building explainable metric for all text generation tasks*. In *arXiv*.

- Dongfu Jiang, Xiang Ren, and Bill Yuchen Lin. 2023. **LLM-blender: Ensembling large language models with pairwise ranking and generative fusion**. In *Annual Meeting of the Association for Computational Linguistics*, pages 14165–14178.
- Yuxin Jiang, Yufei Wang, Xingshan Zeng, Wanjun Zhong, Liangyou Li, Fei Mi, Lifeng Shang, Xin Jiang, Qun Liu, and Wei Wang. 2024b. **Followbench: A multi-level fine-grained constraints following benchmark for large language models**. In *arXiv*.
- Carlos E Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik R Narasimhan. 2024. **SWE-bench: Can language models resolve real-world github issues?** In *International Conference on Learning Representations*.
- Di Jin, Eileen Pan, Nassim Oufattolle, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2020. **What disease does this patient have? a large-scale open domain question answering dataset from medical exams**. In *arXiv*.
- Hyunbin Jin, Je Won Yeom, Seunghyun Bae, and Taesup Kim. 2025. “well, keep thinking”: Enhancing llm reasoning with adaptive injection decoding. In *arXiv*.
- Mingyu Jin, Qinkai Yu, Dong Shu, Haiyan Zhao, Wenyue Hua, Yanda Meng, Yongfeng Zhang, and Mengnan Du. 2024. **The impact of reasoning step length on large language models**. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 1830–1842.
- D. Kahneman. 2011. *Thinking, Fast and Slow*. Farrar, Straus and Giroux.
- Daniel Kahneman. 2003. Maps of bounded rationality: Psychology for behavioral economics. *The American Economic Review*, 93(5):1449–1475.
- Jikun Kang, Xin Zhe Li, Xi Chen, Amirreza Kazemi, Qianyi Sun, Boxing Chen, Dong Li, Xu He, Quan He, Feng Wen, Jianye Hao, and Jun Yao. 2024. **Mindstar: Enhancing math reasoning in pre-trained llms at inference time**. In *arXiv*.
- Zhewei Kang, Xuandong Zhao, and Dawn Song. 2025. Scalable best-of-n selection for large language models via self-certainty. In *arXiv*.
- Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. **Scaling laws for neural language models**. In *arXiv*.
- Amirhossein Kazemnejad, Milad Aghajohari, Eva Portelance, Alessandro Sordoni, Siva Reddy, Aaron Courville, and Nicolas Le Roux. 2024. **Vineppo: Unlocking rl potential for llm reasoning through refined credit assignment**. *arXiv preprint arXiv:2410.01679*.
- Seungone Kim, Ian Wu, JINU Lee, Xiang Yue, Seongyun Lee, Mingyeong Moon, Kiril Gashtelovski, Carolin Lawrence, Julia Hockenmaier, Graham Neubig, and Sean Welleck. 2025. **Scaling evaluation-time compute with reasoning models as process evaluators**. In *arXiv*.
- Kimi. 2025. **Kimi k1.5: Scaling reinforcement learning with llms**. In *arXiv*.
- Deqian Kong, Minglu Zhao, Dehong Xu, Bo Pang, Shu Wang, Edouardo Honig, Zhangzhang Si, Chuan Li, Jianwen Xie, Sirui Xie, and Ying Nian Wu. 2025. **Scalable language models with posterior inference of latent thought vectors**. In *arXiv*.
- Satyapriya Krishna, Kalpesh Krishna, Anhad Mohananey, Steven Schwarcz, Adam Stambler, Shyam Upadhyay, and Manaal Faruqui. 2025. **Fact, fetch, and reason: A unified evaluation of retrieval-augmented generation**.
- Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Chris Wilhelm, Luca Soldaini, Noah A. Smith, Yizhong Wang, Pradeep Dasigi, and Hannaneh Hajishirzi. 2025. **Tulu 3: Pushing frontiers in open language model post-training**. In *arXiv*.
- Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yebin Choi, Noah A. Smith, and Hannaneh Hajishirzi. 2024. **Rewardbench: Evaluating reward models for language modeling**. In *arXiv*.
- Gregory Kang Ruey Lau, Wenyang Hu, Diwen Liu, Jizhuo Chen, See-Kiong Ng, and Bryan Kian Hsiang Low. 2024. **Dipper: Diversity in prompts for producing large language model ensembles in reasoning tasks**. In *arXiv*.
- Kuang-Huei Lee, Ian Fischer, Yueh-Hua Wu, Dave Marwood, Shumeet Baluja, Dale Schuurmans, and Xinyun Chen. 2025. **Evolving deeper llm thinking**. In *arXiv*.

- Aitor Lewkowycz, Anders Johan Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Venkatesh Ramasesh, Ambrose Sloane, Cem Anil, Imanol Schlag, Theo Gutman-Solo, Yuhuai Wu, Behnam Neyshabur, Guy Gur-Ari, and Vedant Misra. 2022. [Solving quantitative reasoning problems with language models](#). In *Conference on Neural Information Processing Systems*.
- Bingxuan Li, Yiwei Wang, Jiuxiang Gu, Kai-Wei Chang, and Nanyun Peng. 2025a. [METAL: A multi-agent framework for chart generation with test-time scaling](#). In *arXiv*.
- Chengpeng Li, Mingfeng Xue, Zhenru Zhang, Jiaxi Yang, Beichen Zhang, Xiang Wang, Bowen Yu, Binyuan Hui, Junyang Lin, and Dayiheng Liu. 2025b. [START: Self-taught reasoner with tools](#). In *arXiv*.
- Cheryl Li, Tianyuan Xu, and Yiwen Guo. 2025c. [Reasoning-as-logic-units: Scaling test-time reasoning in large language models through logic unit alignment](#). In *arXiv*.
- Dacheng Li, Shiyi Cao, Chengkun Cao, Xiuyu Li, Shangyin Tan, Kurt Keutzer, Jiarong Xing, Joseph E. Gonzalez, and Ion Stoica. 2025d. [S*: Test time scaling for code generation](#).
- Dacheng Li, Shiyi Cao, Tyler Griggs, Shu Liu, Xiangxi Mo, Eric Tang, Sumanth Hegde, Kourosh Hakhamaneshi, Shishir G. Patil, Matei Zaharia, Joseph E. Gonzalez, and Ion Stoica. 2025e. [Llms can easily learn to reason from demonstrations structure, not content, is what matters!](#) In *arXiv*.
- Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. 2024. [Cmmlu: Measuring massive multitask language understanding in chinese](#). In *arXiv*.
- Jia LI, Edward Beeching, Lewis Tunstall, Ben Lipkin, Roman Soletskyi, Shengyi Costa Huang, Kashif Rasul, Longhui Yu, Albert Jiang, Ziju Shen, Zihan Qin, Bin Dong, Li Zhou, Yann Fleureau, Guillaume Lample, and Stanislas Polu. 2024. [Numinamath](#). [<https://github.com/project-numina/aimo-progress-prize>] (https://github.com/project-numina/aimo-progress-prize/blob/main/report/numina_dataset.pdf).
- Junyou Li, Qin Zhang, Yangbin Yu, Qiang Fu, and Deheng Ye. 2024a. More agents is all you need. *arXiv preprint arXiv:2402.05120*.
- Minzhi Li, Zhengyuan Liu, Shumin Deng, Shafiq Joty, Nancy F. Chen, and Min-Yen Kan. 2024b. [Dna-eval: Enhancing large language model evaluation through decomposition and aggregation](#). In *arXiv*.
- Tianle Li, Wei-Lin Chiang, Evan Frick, Lisa Dunlap, Tianhao Wu, Banghua Zhu, Joseph E. Gonzalez, and Ion Stoica. 2024c. [From crowdsourced data to high-quality benchmarks: Arena-hard and benchbuilder pipeline](#). In *arXiv*.
- Yafu Li, Zhilin Wang, Tingchen Fu, Ganqu Cui, Sen Yang, and Yu Cheng. 2025f. From drafts to answers: [Unlocking llm potential via aggregation fine-tuning](#). In *arXiv*.
- Yanyang Li, Michael Lyu, and Liwei Wang. 2025g. [Learning to reason from feedback at test-time](#). In *arXiv*.
- Yifei Li, Zeqi Lin, Shizhuo Zhang, Qiang Fu, Bei Chen, Jian-Guang Lou, and Weizhu Chen. 2023a. [Making language models better reasoners with step-aware verifier](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5315–5333.
- Yujia Li, David Choi, Junyoung Chung, Nate Kushman, Julian Schrittweiser, Rémi Leblond, Tom Eccles, James Keeling, Felix Gimeno, Agustin Dal Lago, Thomas Hubert, Peter Choy, Cyprien de Masson d’Autume, Igor Babuschkin, Xinyun Chen, Po-Sen Huang, Johannes Welbl, Sven Gowal, Alexey Cherepanov, James Molloy, Daniel J. Mankowitz, Esme Sutherland Robson, Pushmeet Kohli, Nando de Freitas, Koray Kavukcuoglu, and Oriol Vinyals. 2022. Competition-level code generation with alphacode. *Science*, pages 1092–1097.
- Zhiyuan Li, Hong Liu, Denny Zhou, and Tengyu Ma. 2024d. [Chain of thought empowers transformers to solve inherently serial problems](#). In *The Twelfth International Conference on Learning Representations*.
- Zhong-Zhi Li, Duzhen Zhang, Ming-Liang Zhang, Jiaxin Zhang, Zengyan Liu, Yuxuan Yao, Haotian Xu, Junhao Zheng, Pei-Jie Wang, Xiuyi Chen, et al. 2025h. [From system 1 to system 2: A survey of reasoning large language models](#). *arXiv preprint arXiv:2502.17419*.
- Zhongzhi Li, Ming-Liang Zhang, Pei-Jie Wang, Jian Xu, Rui-Song Zhang, Yin Fei, Zhi-Long Ji, Jin-Feng Bai, Zhen-Ru Pan, Jiaxin Zhang, and Cheng-Lin Liu. 2025i. [CMMaTH: A Chinese multi-modal math skill evaluation benchmark for foundation models](#). In *International Conference on Computational Linguistics*, pages 2690–2726.
- Zichong Li, Xinyu Feng, Yuheng Cai, Zixuan Zhang, Tianyi Liu, Chen Liang, Weizhu Chen, Haoyu Wang, and Tuo Zhao. 2025j. [Llms can generate a better answer by aggregating their own responses](#). In *arXiv*.

- Ziniu Li, Tian Xu, Yushun Zhang, Zhihang Lin, Yang Yu, Ruoyu Sun, and Zhi-Quan Luo. 2023b. **Remax: A simple, effective, and efficient reinforcement learning method for aligning large language models.** *arXiv preprint arXiv:2310.10505*.
- Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 2024. **Encouraging divergent thinking in large language models through multi-agent debate.** In *Conference on Empirical Methods in Natural Language Processing*, pages 17889–17904.
- Shalev Lifshitz, Sheila A. McIlraith, and Yilun Du. 2025. **Multi-agent verification: Scaling test-time compute with goal verifiers.** In *Workshop on Reasoning and Planning for Large Language Models*.
- 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*.
- Qingwen Lin, Boyan Xu, Zijian Li, Zhifeng Hao, Keli Zhang, and Ruichu Cai. 2025. **Leveraging constrained monte carlo tree search to generate reliable long chain-of-thought for mathematical reasoning.** In *arXiv*.
- Zicheng Lin, Tian Liang, Jiahao Xu, Xing Wang, Ruilin Luo, Chufan Shi, Siheng Li, Yujiu Yang, and Zhaopeng Tu. 2024. **Critical tokens matter: Token-level contrastive estimation enhance llm’s reasoning capability.** *arXiv preprint arXiv:2411.19943*.
- Zhan Ling, Yunhao Fang, Xuanlin Li, Zhiao Huang, Mingu Lee, Roland Memisevic, and Hao Su. 2023. **Deductive verification of chain-of-thought reasoning.** In *Advances in Neural Information Processing Systems*, volume 36, pages 36407–36433.
- Changshu Liu, Shizhuo Dylan Zhang, Ali Reza Ibrahimzada, and Reyhaneh Jabbarvand. 2024a. **Codemind: A framework to challenge large language models for code reasoning.** In *arXiv*.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023a. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.
- Jiacai Liu, Chaojie Wang, Chris Yuhao Liu, Liang Zeng, Rui Yan, Yiwen Sun, Yang Liu, and Yahui Zhou. 2024b. **Improving multi-step reasoning abilities of large language models with direct advantage policy optimization.** In *arXiv*.
- Runze Liu, Junqi Gao, Jian Zhao, Kaiyan Zhang, Xiu Li, Biqing Qi, Wanli Ouyang, and Bowen Zhou. 2025a. **Can 1b llm surpass 405b llm? rethinking compute-optimal test-time scaling.** *arXiv preprint arXiv:2502.06703*.
- Tengxiao Liu, Qipeng Guo, Yuqing Yang, Xiangkun Hu, Yue Zhang, Xipeng Qiu, and Zheng Zhang. 2023b. **Plan, verify and switch: Integrated reasoning with diverse x-of-thoughts.** In *Conference on Empirical Methods in Natural Language Processing*.
- Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023c. **G-eval: NLG evaluation using gpt-4 with better human alignment.** In *arXiv*.
- Yantao Liu, Zijun Yao, Rui Min, Yixin Cao, Lei Hou, and Juanzi Li. 2024c. **Rm-bench: Benchmarking reward models of language models with subtlety and style.** In *arXiv*.
- Yantao Liu, Zijun Yao, Rui Min, Yixin Cao, Lei Hou, and Juanzi Li. 2025b. **Pairjudge rm: Perform best-of-n sampling with knockout tournament.** In *arXiv*.
- Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. 2024d. **Mmbench: Is your multi-modal model an all-around player?** In *European Conference on Computer Vision*, pages 216–233.
- Zijun Liu, Peiyi Wang, Runxin Xu, Shirong Ma, Chong Ruan, Peng Li, Yang Liu, and Yu Wu. 2025c. **Inference-time scaling for generalist reward modeling.** In *arXiv*.
- Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024. **Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts.** In *arXiv*.
- 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,** 2025. URL <https://arxiv.org/abs/2501.12570>.

- Michael Luo, Sijun Tan, Justin Wong, Xiaoxiang Shi, William Y. Tang, Manan Roongta, Colin Cai, Jeffrey Luo, Tianjun Zhang, Li Erran Li, Raluca Ada Popa, and Ion Stoica. 2025b. Deepscaler: Surpassing o1-preview with a 1.5b model by scaling rl. <https://pretty-radio-b75.notion.site/DeepScaleR-Surpassing-O1-Preview-with-a-1-5B-Model-by-Scaling-RL-19681902c1468005bed8> Notion Blog.
- Trung Quoc Luong, Xinbo Zhang, Zhanming Jie, Peng Sun, Xiaoran Jin, and Hang Li. 2024. **Reft: Reasoning with reinforced fine-tuning**. In *arXiv*.
- Chang Ma, Haiteng Zhao, Junlei Zhang, Junxian He, and Lingpeng Kong. 2025a. **Non-myopic generation of language models for reasoning and planning**. In *International Conference on Learning Representations*.
- Nanye Ma, Shangyuan Tong, Haolin Jia, Hexiang Hu, Yu-Chuan Su, Mingda Zhang, Xuan Yang, Yandong Li, Tommi Jaakkola, Xuhui Jia, and Saining Xie. 2025b. **Inference-time scaling for diffusion models beyond scaling denoising steps**. In *arXiv*.
- Yiran Ma, Zui Chen, Tianqiao Liu, Mi Tian, Zhuo Liu, Zitao Liu, and Weiqi Luo. 2025c. **What are step-level reward models rewarding? counterintuitive findings from mcts-boosted mathematical reasoning**. In *arXiv*.
- Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. **Self-refine: Iterative refinement with self-feedback**. In *Conference on Neural Information Processing Systems*.
- Tarek Mahmud, Bin Duan, Corina Pasareanu, and Guowei Yang. 2025. **Enhancing llm code generation with ensembles: A similarity-based selection approach**. In *arXiv*.
- Sara Vera Marjanović, Arkil Patel, Vaibhav Adlakha, Milad Aghajohari, Parishad BehnamGhader, Mehar Bhatia, Aditi Khandelwal, Austin Kraft, Benno Krojer, Xing Han Lù, Nicholas Meade, Dongchan Shin, Amirhossein Kazemnejad, Gaurav Kamath, Marius Mosbach, Karolina Stańczak, and Siva Reddy. 2025. **Deepseek-r1 thoughtology: Let's ;think; about llm reasoning**. In *arXiv*.
- Nat McAleese, Rai Michael Pokorny, Juan Felipe Ceron Uribe, Evgenia Nitishinskaya, Maja Trebacz, and Jan Leike. 2024. **Llm critics help catch llm bugs**. In *arXiv*.
- Yu Meng, Mengzhou Xia, and Danqi Chen. 2024. Simpo: Simple preference optimization with a reference-free reward. *Advances in Neural Information Processing Systems*, 37:124198–124235.
- Elliot Meyerson and Xin Qiu. 2025. Position: Scaling llm agents requires asymptotic analysis with llm primitives. *arXiv preprint arXiv:2502.04358*.
- Kou Misaki, Yuichi Inoue, Yuki Imajuku, So Kuroki, Taishi Nakamura, and Takuya Akiba. 2025. **Wider or deeper? scaling llm inference-time compute with adaptive branching tree search**. In *arXiv*.
- 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**. In *arXiv*.
- Tergel Munkhbat, Namgyu Ho, Seo Hyun Kim, Yongjin Yang, Yujin Kim, and Se-Young Yun. 2025. **Self-training elicits concise reasoning in large language models**. In *arXiv*.
- NCEE. 2025. China's national college entrance examination.
- Alex Nguyen, Dheeraj Mekala, Chengyu Dong, and Jingbo Shang. 2024. **When is the consistent prediction likely to be a correct prediction?** In *arXiv*.
- Ansong Ni, Miltiadis Allamanis, Arman Cohan, Yinlin Deng, Kensen Shi, Charles Sutton, and Pengcheng Yin. 2024. **Next: Teaching large language models to reason about code execution**. In *arXiv*.
- Ansong Ni, Srini Iyer, Dragomir Radev, Ves Stoyanov, Wen tau Yih, Sida I. Wang, and Xi Victoria Lin. 2023. **Lever: Learning to verify language-to-code generation with execution**. In *arXiv*.
- NovaSky. 2025. Sky-t1: Train your own o1 preview model within \$450. <https://novasky-ai.github.io/posts/sky-t1>. Accessed: 2025-01-09.
- Isaac Ong, Amjad Almahairi, Vincent Wu, Wei-Lin Chiang, Tianhao Wu, Joseph E. Gonzalez, M Waleed Kadous, and Ion Stoica. 2025. **RouteLLM: Learning to route LLMs from preference data**. In *International Conference on Learning Representations*.
- OpenAI. 2024a. **Gpt-4 technical report**. In *arXiv*.

- OpenAI. 2024b. [Openai o1 system card](#). In *arXiv*.
- OpenAI. 2025. [Openai o3-mini system card](#).
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Gray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. [Training language models to follow instructions with human feedback](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744.
- Jianfeng Pan, Senyou Deng, and Shaomang Huang. 2025a. [Coat: Chain-of-associated-thoughts framework for enhancing large language models reasoning](#). In *arXiv*.
- Jiayi Pan, Junjie Zhang, Xingyao Wang, Lifan Yuan, Hao Peng, and Alane Suhr. 2025b. [Tinyzero](#). <https://github.com/Jiayi-Pan/TinyZero>. Accessed: 2025-01-24.
- Xuchen Pan, Dawei Gao, Yuexiang Xie, Yushuo Chen, Zhewei Wei, Yaliang Li, Bolin Ding, Ji-Rong Wen, and Jingren Zhou. Very large-scale multi-agent simulation with llm-powered agents.
- Aldo Pareja, Nikhil Shivakumar Nayak, Hao Wang, Krishnateja Killamsetty, Shivchander Sudalairaj, Wenlong Zhao, Seungwook Han, Abhishek Bhandwaldar, Guangxuan Xu, Kai Xu, Ligong Han, Luke Inglis, and Akash Srivastava. 2024. [Unveiling the secret recipe: A guide for supervised fine-tuning small llms](#). In *arXiv*.
- Mihir Parmar, Xin Liu, Palash Goyal, Yanfei Chen, Long Le, Swaroop Mishra, Hossein Mobahi, Jindong Gu, Zifeng Wang, Hootan Nakhost, Chitta Baral, Chen-Yu Lee, Tomas Pfister, and Hamid Palangi. 2025. [Plangen: A multi-agent framework for generating planning and reasoning trajectories for complex problem solving](#). In *arXiv*.
- Baolin Peng, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Liden, Zhou Yu, Weizhu Chen, and Jianfeng Gao. 2023. [Check your facts and try again: Improving large language models with external knowledge and automated feedback](#). In *arXiv*.
- Jacob Pfau, William Merrill, and Samuel R. Bowman. 2024. [Let's think dot by dot: Hidden computation in transformer language models](#). In *Conference on Language Modeling*.
- Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, Sean Shi, Michael Choi, Anish Agrawal, Arnav Chopra, et al. 2025. Humanity's last exam. *arXiv preprint arXiv:2501.14249*.
- Jinghua Piao, Yuwei Yan, Jun Zhang, Nian Li, Junbo Yan, Xiaochong Lan, Zhihong Lu, Zhiheng Zheng, Jing Yi Wang, Di Zhou, et al. 2025. Agentsociety: Large-scale simulation of llm-driven generative agents advances understanding of human behaviors and society. *arXiv preprint arXiv:2502.08691*.
- Archiki Prasad, Alexander Koller, Mareike Hartmann, Peter Clark, Ashish Sabharwal, Mohit Bansal, and Tushar Khot. 2024. [Adapt: As-needed decomposition and planning with language models](#). In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 4226–4252.
- Isha Puri, Shivchander Sudalairaj, Guangxuan Xu, Kai Xu, and Akash Srivastava. 2025. [A probabilistic inference approach to inference-time scaling of llms using particle-based monte carlo methods](#). In *arXiv*.
- Chen Qian, Zihao Xie, Yifei Wang, Wei Liu, Yufan Dang, Zhuoyun Du, Weize Chen, Cheng Yang, Zhiyuan Liu, and Maosong Sun. 2024. Scaling large-language-model-based multi-agent collaboration. *arXiv preprint arXiv:2406.07155*.
- Yiwei Qin, Xuefeng Li, Haoyang Zou, Yixiu Liu, Shijie Xia, Zhen Huang, Yixin Ye, Weizhe Yuan, Hector Liu, Yuanzhi Li, and Pengfei Liu. 2024. [O1 replication journey: A strategic progress report – part 1](#). In *arXiv*.
- 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 preprint arXiv:2410.16033*.
- Shi Qiu, Shaoyang Guo, Zhuo-Yang Song, Yunbo Sun, Zeyu Cai, Jiashen Wei, Tianyu Luo, Yixuan Yin, Haoxu Zhang, Yi Hu, et al. 2025. [Phybench: Holistic evaluation of physical perception and reasoning in large language models](#). *arXiv preprint arXiv:2504.16074*.
- Xiaoye Qu, Yafu Li, Zhaochen Su, Weigao Sun, Jianhao Yan, Dongrui Liu, Ganqu Cui, Daizong Liu, Shuxian Liang, Junxian He, et al. 2025. A survey of efficient reasoning for large reasoning models: Language, multimodality, and beyond. *arXiv preprint arXiv:2503.21614*.
- Qwen. 2024. [Qwq: Reflect deeply on the boundaries of the unknown](#).

- Leonardo Ranaldi, Marco Valentino, Alexander Polonsky, and André Freitas. 2025. [Improving chain-of-thought reasoning via quasi-symbolic abstractions](#). In *arXiv*.
- David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. 2024. [GPQA: A graduate-level google-proof q&a benchmark](#). In *First Conference on Language Modeling*.
- Matthew Renze. 2024. [The effect of sampling temperature on problem solving in large language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7346–7356.
- Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. *arXiv preprint arXiv:1809.02156*.
- Jon Saad-Falcon, Adrian Gamarra Lafuente, Shlok Natarajan, Nahum Maru, Hristo Todorov, Etrash Guha, E. Kelly Buchanan, Mayee Chen, Neel Guha, Christopher Ré, and Azalia Mirhoseini. 2024. [Archon: An architecture search framework for inference-time techniques](#). In *arXiv*.
- Swarnadeep Saha, Xian Li, Marjan Ghazvininejad, Jason Weston, and Tianlu Wang. 2025. [Learning to plan & reason for evaluation with thinking-llm-as-a-judge](#). In *arXiv*.
- Alireza Salemi and Hamed Zamani. 2024. [Towards a search engine for machines: Unified ranking for multiple retrieval-augmented large language models](#). In *arXiv*.
- Nikunj Saunshi, Nishanth Dikkala, Zhiyuan Li, Sanjiv Kumar, and Sashank J. Reddi. 2025. [Reasoning with latent thoughts: On the power of looped transformers](#). In *arXiv*.
- Tom Schaul. 2024. [Boundless socratic learning with language games](#). In *Language Gamification-NeurIPS 2024 Workshop*.
- John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. [Proximal policy optimization algorithms](#).
- Bilgehan Sel, Ahmad Tawaha, Vanshaj Khattar, Ruoxi Jia, and Ming Jin. 2024. Algorithm of thoughts: Enhancing exploration of ideas in large language models. In *International Conference on Machine Learning*, pages 44136–44189. PMLR.
- Pier Giuseppe Sessa, Robert Dadashi, Léonard Hussonot, Johan Ferret, Nino Vieillard, Alexandre Ramé, Bobak Shariari, Sarah Perrin, Abe Friesen, Geoffrey Cideron, Sertan Girgin, Piotr Stanczyk, Andrea Michi, Danila Sinopalnikov, Sabela Ramos, Amélie Héliou, Aliaksei Severyn, Matt Hoffman, Nikola Momchev, and Olivier Bachem. 2024. [Bond: Aligning llms with best-of-n distillation](#). In *arXiv*.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, et al. 2024. [Deepseekmath: Pushing the limits of mathematical reasoning in open language models](#). *arXiv preprint arXiv:2402.03300*.
- Wei Shen, Guanlin Liu, Zheng Wu, Ruofei Zhu, Qingping Yang, Chao Xin, Yu Yue, and Lin Yan. 2025a. [Exploring data scaling trends and effects in reinforcement learning from human feedback](#). In *arXiv*.
- Xuan Shen, Yizhou Wang, Xiangxi Shi, Yanzhi Wang, Pu Zhao, and Jiuxiang Gu. 2025b. [Efficient reasoning with hidden thinking](#). In *arXiv*.
- 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](#). In *arXiv*.
- Ben Shi, Michael Tang, Karthik R Narasimhan, and Shunyu Yao. 2024. [Can language models solve olympiad programming?](#) In *Conference on Language Modeling*.
- Avi Singh, John D Co-Reyes, Rishabh Agarwal, Ankesh Anand, Piyush Patil, Xavier Garcia, Peter J Liu, James Harrison, Jaehoon Lee, Kelvin Xu, Aaron T Parisi, Abhishek Kumar, Alexander A Alemi, Alex Rizkowsky, Azade Nova, Ben Adlam, Bernd Bohnet, Gamaleldin Fathy Elsayed, Hanie Sedghi, Igor Mordatch, Isabelle Simpson, Izzeddin Gur, Jasper Snoek, Jeffrey Pennington, Jiri Hron, Kathleen Kenealy, Kevin Swersky, Kshiteej Mahajan, Laura A Culp, Lechao Xiao, Maxwell Bileschi, Noah Constant, Roman Novak, Rosanne Liu, Tris Warkentin, Yamini Bansal, Ethan Dyer, Behnam Neyshabur, Jascha Sohl-Dickstein, and Noah Fiedel. 2024. [Beyond human data: Scaling self-training for problem-solving with language models](#). *Transactions on Machine Learning Research*.
- Joykirat Singh, Tanmoy Chakraborty, and Akshay Nambi. 2025. [Self-evolved preference optimization for enhancing mathematical reasoning in small language models](#). In *arXiv*.

- Joar Skalse, Nikolaus H. R. Howe, Dmitrii Krasheninnikov, and David Krueger. 2025. Defining and characterizing reward hacking. In *arXiv*.
- 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, Zhaochen Su, Xiaoye Qu, Jiawei Zhou, and Yu Cheng. 2025. Prmbench: A fine-grained and challenging benchmark for process-level reward models. *arXiv preprint arXiv:2501.03124*.
- Yifan Song, Guoyin Wang, Sujian Li, and Bill Yuchen Lin. 2024. The good, the bad, and the greedy: Evaluation of llms should not ignore non-determinism. In *arXiv*.
- Keith E. Stanovich and Richard F. West. 2000. Advancing the rationality debate. *Behavioral and Brain Sciences*, page 701–717.
- DiJia Su, Hanlin Zhu, Yingchen Xu, Jiantao Jiao, Yuandong Tian, and Qingqing Zheng. 2025. Token assorted: Mixing latent and text tokens for improved language model reasoning. *arXiv preprint arXiv:2502.03275*.
- Yuan Sui, Yufei He, Tri Cao, Simeng Han, and Bryan Hooi. 2025. Meta-reasoner: Dynamic guidance for optimized inference-time reasoning in large language models.
- Liangtai Sun, Yang Han, Zihan Zhao, Da Ma, Zhennan Shen, Baocai Chen, Lu Chen, and Kai Yu. 2024. Scieval: a multi-level large language model evaluation benchmark for scientific research. In *AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*.
- Zexu Sun, Yiju Guo, Yankai Lin, Xu Chen, Qi Qi, Xing Tang, and Ji-Rong Wen. 2025. Uncertainty and influence aware reward model refinement for reinforcement learning from human feedback. In *The Thirteenth International Conference on Learning Representations*.
- Zhiqing Sun, Yikang Shen, Qinhong Zhou, Hongxin Zhang, Zhenfang Chen, David Cox, Yiming Yang, and Chuang Gan. 2023. Principle-driven self-alignment of language models from scratch with minimal human supervision. In *Advances in Neural Information Processing Systems*, pages 2511–2565.
- Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. 1999. Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems*, 12.
- Sijun Tan, Siyuan Zhuang, Kyle Montgomery, William Y. Tang, Alejandro Cuadron, Chenguang Wang, Raluca Ada Popa, and Ion Stoica. 2025. Judgebench: A benchmark for evaluating llm-based judges. In *arXiv*.
- Amir Taubenfeld, Tom Sheffer, Eran Ofek, Amir Feder, Ariel Goldstein, Zorik Gekhman, and Gal Yona. 2025. Confidence improves self-consistency in llms. In *arXiv*.
- Fengwei Teng, Zhaoyang Yu, Quan Shi, Jiayi Zhang, Chenglin Wu, and Yuyu Luo. 2025. Atom of thoughts for markov llm test-time scaling. *arXiv preprint arXiv:2502.12018*.
- Ye Tian, Baolin Peng, Linfeng Song, Lifeng Jin, Dian Yu, Lei Han, Haitao Mi, and Dong Yu. 2024. Toward self-improvement of LLMs via imagination, searching, and criticizing. In *Conference on Neural Information Processing Systems*.
- Yuchen Tian, Weixiang Yan, Qian Yang, Xuandong Zhao, Qian Chen, Wen Wang, Ziyang Luo, Lei Ma, and Dawn Song. 2025. Codehalu: Investigating code hallucinations in llms via execution-based verification. In *arXiv*.
- Peter Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Adithya Jairam Vedagiri IYER, Sai Charitha Akula, Shusheng Yang, Jihan Yang, Manoj Middepogu, Ziteng Wang, et al. 2024. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *Advances in Neural Information Processing Systems*, 37:87310–87356.
- George Tsoukalas, Jasper Lee, John Jennings, Jimmy Xin, Michelle Ding, Michael Jennings, Amitayush Thakur, and Swarat Chaudhuri. 2024. Putnambench: Evaluating neural theorem-provers on the putnam mathematical competition.
- Jonathan Uesato, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. 2022. Solving math word problems with process- and outcome-based feedback. In *arXiv*.
- Juraj Vladika and Florian Matthes. 2024. Improving health question answering with reliable and time-aware evidence retrieval. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 4752–4763.
- David Wan, Justin Chih-Yao Chen, Elias Stengel-Eskin, and Mohit Bansal. 2025. Mamm-refine: A recipe for improving faithfulness in generation with multi-agent collaboration. In *arXiv*.

- Ziyu Wan, Xidong Feng, Muning Wen, Stephen Marcus McAleer, Ying Wen, Weinan Zhang, and Jun Wang. 2024. [Alphazero-like tree-search can guide large language model decoding and training](#). In *Forty-first International Conference on Machine Learning*.
- Evan Wang, Federico Cassano, Catherine Wu, Yunfeng Bai, Will Song, Vaskar Nath, Ziwen Han, Sean Hendryx, Summer Yue, and Hugh Zhang. 2024a. [Planning in natural language improves llm search for code generation](#). In *arXiv*.
- Haiming Wang, Huajian Xin, Chuanyang Zheng, Lin Li, Zhengying Liu, Qingxing Cao, Yinya Huang, Jing Xiong, Han Shi, Enze Xie, et al. 2023a. [Lego-prover: Neural theorem proving with growing libraries](#). *arXiv preprint arXiv:2310.00656*.
- Huaijie Wang, Shibo Hao, Hanze Dong, Shenao Zhang, Yilin Bao, Ziran Yang, and Yi Wu. 2024b. [Offline reinforcement learning for llm multi-step reasoning](#). *arXiv preprint arXiv:2412.16145*.
- Jun Wang, Meng Fang, Ziyu Wan, Muning Wen, Jiachen Zhu, Anjie Liu, Ziqin Gong, Yan Song, Lei Chen, Lionel M Ni, et al. 2024c. [Openr: An open source framework for advanced reasoning with large language models](#). *arXiv preprint arXiv:2410.09671*.
- Junlin Wang, Jue WANG, Ben Athiwaratkun, Ce Zhang, and James Zou. 2025a. [Mixture-of-agents enhances large language model capabilities](#). In *International Conference on Learning Representations*.
- Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. 2024d. [Measuring multimodal mathematical reasoning with MATH-vision dataset](#). In *Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
- Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. 2024e. [Math-shepherd: Verify and reinforce LLMs step-by-step without human annotations](#). In *Annual Meeting of the Association for Computational Linguistics*, pages 9426–9439.
- Ruida Wang, Rui Pan, Yuxin Li, Jipeng Zhang, Yizhen Jia, Shizhe Diao, Renjie Pi, Junjie Hu, and Tong Zhang. 2025b. [Ma-lot: Multi-agent lean-based long chain-of-thought reasoning enhances formal theorem proving](#). In *arXiv*.
- Ruoyao Wang, Peter Jansen, Marc-Alexandre Côté, and Prithviraj Ammanabrolu. 2022. [Scienceworld: Is your agent smarter than a 5th grader?](#) In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11279–11298.
- Tianlong Wang, Junzhe Chen, Xuetong Han, and Jing Bai. 2024f. [Cpl: Critical plan step learning boosts llm generalization in reasoning tasks](#). *arXiv preprint arXiv:2409.08642*.
- Xinglin Wang, Shaoxiong Feng, Yiwei Li, Peiwen Yuan, Yueqi Zhang, Chuyi Tan, Boyuan Pan, Yao Hu, and Kan Li. 2025c. [Make every penny count: Difficulty-adaptive self-consistency for cost-efficient reasoning](#). In *arXiv*.
- Xinyi Wang, Lucas Caccia, Oleksiy Ostapenko, Xingdi Yuan, William Yang Wang, and Alessandro Sordoni. 2024g. [Guiding language model reasoning with planning tokens](#). In *Conference on Language Modeling*.
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023b. [Self-consistency improves chain of thought reasoning in language models](#). In *International Conference on Learning Representations*.
- Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyuan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhua Chen. 2024h. [MMLU-pro: A more robust and challenging multi-task language understanding benchmark](#). In *Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
- Yubo Wang, Xiang Yue, and Wenhua Chen. 2025d. [Critique fine-tuning: Learning to critique is more effective than learning to imitate](#). In *arXiv*.
- Yue Wang, Qiuwei Liu, Jiahao Xu, Tian Liang, Xingyu Chen, Zhiwei He, Linfeng Song, Dian Yu, Juntao Li, Zhuosheng Zhang, Rui Wang, Zhaopeng Tu, Haitao Mi, and Dong Yu. 2025e. [Thoughts are all over the place: On the underthinking of o1-like llms](#). In *arXiv*.
- Zhao Wang, Sota Moriyama, Wei-Yao Wang, Briti Gangopadhyay, and Shingo Takamatsu. 2025f. [Talk structurally, act hierarchically: A collaborative framework for llm multi-agent systems](#). In *arXiv*.
- Maurice Weber, Daniel Fu, Quentin Anthony, Yonatan Oren, Shane Adams, Anton Alexandrov, Xiaozhong Lyu, Huu Nguyen, Xiaozhe Yao, Virginia Adams, Ben Athiwaratkun, Rahul Chalamala, Kezhen Chen, Max Ryabinin, Tri Dao, Percy Liang, Christopher Ré, Irina Rish, and Ce Zhang. 2024. [Redpajama: an open dataset for training large language models](#). In *arXiv*.

- Jason Wei, Nguyen Karina, Hyung Won Chung, Yunxin Joy Jiao, Spencer Papay, Amelia Glaese, John Schulman, and William Fedus. 2024a. [Measuring short-form factuality in large language models](#).
- 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.
- Jerry Wei, Chengrun Yang, Xinying Song, Yifeng Lu, Nathan Hu, Jie Huang, Dustin Tran, Daiyi Peng, Ruibo Liu, Da Huang, Cosmo Du, and Quoc V. Le. 2024b. [Long-form factuality in large language models](#). In *arXiv*.
- Sean Welleck, Amanda Bertsch, Matthew Finlayson, Hailey Schoelkopf, Alex Xie, Graham Neubig, Ilia Kulikov, and Zaid Harchaoui. 2024. From decoding to meta-generation: Inference-time algorithms for large language models. *arXiv preprint arXiv:2406.16838*.
- Liang Wen, Yunke Cai, Fenrui Xiao, Xin He, Qi An, Zhenyu Duan, Yimin Du, Junchen Liu, Lifu Tang, Xiaowei Lv, et al. 2025. [Light-rl: Curriculum sft, dpo and rl for long cot from scratch and beyond](#). *arXiv preprint arXiv:2503.10460*.
- Alexander Wettig, Aatmik Gupta, Saumya Malik, and Danqi Chen. 2024. [Qurating: Selecting high-quality data for training language models](#). In *arXiv*.
- Siwei Wu, Zhongyuan Peng, Xinrun Du, Tuney Zheng, Minghao Liu, Jialong Wu, Jiachen Ma, Yizhi Li, Jian Yang, Wangchunshu Zhou, Qunshu Lin, Junbo Zhao, Zhaoxiang Zhang, Wenhao Huang, Ge Zhang, Chenghua Lin, and J. H. Liu. 2024a. [A comparative study on reasoning patterns of openai’s o1 model](#). In *arXiv*.
- Tianhao Wu, Janice Lan, Weizhe Yuan, Jiantao Jiao, Jason Weston, and Sainbayar Sukhbaatar. 2024b. [Thinking llms: General instruction following with thought generation](#). In *arXiv*.
- Yangzhen Wu, Zhiqing Sun, Shanda Li, Sean Welleck, and Yiming Yang. 2024c. Inference scaling laws: An empirical analysis of compute-optimal inference for problem-solving with language models. *arXiv preprint arXiv:2408.00724*.
- Yangzhen Wu, Zhiqing Sun, Shanda Li, Sean Welleck, and Yiming Yang. 2024d. [Scaling inference computation: Compute-optimal inference for problem-solving with language models](#). In *Workshop on Mathematical Reasoning and AI at NeurIPS’24*.
- Yangzhen Wu, Zhiqing Sun, Shanda Li, Sean Welleck, and Yiming Yang. 2025a. [Inference scaling laws: An empirical analysis of compute-optimal inference for llm problem-solving](#). In *The Thirteenth International Conference on Learning Representations*.
- Yuyang Wu, Yifei Wang, Tianqi Du, Stefanie Jegelka, and Yisen Wang. 2025b. [When more is less: Understanding chain-of-thought length in llms](#). In *arXiv*.
- Zengqing Wu and Takayuki Ito. 2025. [The hidden strength of disagreement: Unraveling the consensus-diversity tradeoff in adaptive multi-agent systems](#). In *arXiv*.
- X-R1Team. 2025. X-rl. <https://github.com/dhcode-cpp/X-R1>. Github.
- xAI. 2025. [Grok 3 beta - the age of reasoning agents](#).
- Kun Xiang, Zhili Liu, Zihao Jiang, Yunshuang Nie, Runhui Huang, Haoxiang Fan, Hanhui Li, Weiran Huang, Yihan Zeng, Jianhua Han, Lanqing Hong, Hang Xu, and Xiaodan Liang. 2024. [Atomthink: A slow thinking framework for multimodal mathematical reasoning](#). In *arXiv*.
- Violet Xiang, Charlie Snell, Kanishk Gandhi, Alon Albalak, Anikait Singh, Chase Blagden, Duy Phung, Rafael Rafailov, Nathan Lile, Dakota Mahan, Louis Castricato, Jan-Philipp Franken, Nick Haber, and Chelsea Finn. 2025. [Towards system 2 reasoning in llms: Learning how to think with meta chain-of-thought](#).
- Jian Xie, Kai Zhang, Jiangjie Chen, Tinghui Zhu, Renze Lou, Yuandong Tian, Yanghua Xiao, and Yu Su. 2024. Travelplanner: A benchmark for real-world planning with language agents. In *International Conference on Machine Learning*, pages 54590–54613. PMLR.
- Tian Xie, Zitian Gao, Qingnan Ren, Haoming Luo, Yuqian Hong, Bryan Dai, Joey Zhou, Kai Qiu, Zhirong Wu, and Chong Luo. 2025. [Logic-rl: Unleashing llm reasoning with rule-based reinforcement learning](#). *arXiv preprint arXiv:2502.14768*.
- Yuxi Xie, Kenji Kawaguchi, Yiran Zhao, Xu Zhao, Min-Yen Kan, Junxian He, and Qizhe Xie. 2023. [Self-evaluation guided beam search for reasoning](#). In *Thirty-seventh Conference on Neural Information Processing Systems*.

- Haotian Xu, Xing Wu, Weinong Wang, Zhongzhi Li, Da Zheng, Boyuan Chen, Yi Hu, Shijia Kang, Jiaming Ji, Yingying Zhang, Zhijiang Guo, Yaodong Yang, Muhan Zhang, and Debing Zhang. 2025a. **Redstar: Does scaling long-cot data unlock better slow-reasoning systems?** In *arXiv*.
- Silei Xu, Wenhao Xie, Lingxiao Zhao, and Pengcheng He. 2025b. **Chain of draft: Thinking faster by writing less.** In *arXiv*.
- Wenda Xu, Danqing Wang, Liangming Pan, Zhenqiao Song, Markus Freitag, William Yang Wang, and Lei Li. 2023. **INSTRUCTSCORE: Towards explainable text generation evaluation with automatic feedback.** In *Conference on Empirical Methods in Natural Language Processing*.
- Xin Xu, Qiyun Xu, Tong Xiao, Tianhao Chen, Yuchen Yan, Jiaxin Zhang, Shizhe Diao, Can Yang, and Yang Wang. 2025c. **Ugphysics: A comprehensive benchmark for undergraduate physics reasoning with large language models.** *CoRR*, abs/2502.00334.
- Yige Xu, Xu Guo, Zhiwei Zeng, and Chunyan Miao. 2025d. **Softcot: Soft chain-of-thought for efficient reasoning with llms.** In *arXiv*.
- Fanjia Yan, Huanzhi Mao, Charlie Cheng-Jie Ji, Tianjun Zhang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. 2024. Berkeley function calling leaderboard. https://gorilla.cs.berkeley.edu/blogs/8_berkeley_function_calling_leaderboard.html.
- Ling Yang, Zhaochen Yu, Bin Cui, and Mengdi Wang. 2025a. **Reasonflux: Hierarchical llm reasoning via scaling thought templates.** In *arXiv*.
- Wenkai Yang, Shuming Ma, Yankai Lin, and Furu Wei. 2025b. Towards thinking-optimal scaling of test-time compute for llm reasoning. *arXiv preprint arXiv:2502.18080*.
- Zhen Yang, Fang Liu, Zhongxing Yu, Jacky Wai Keung, Jia Li, Shuo Liu, Yifan Hong, Xiaoxue Ma, Zhi Jin, and Ge Li. 2024. **Exploring and unleashing the power of large language models in automated code translation.** In *arXiv*.
- Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. 2023a. **Webshop: Towards scalable real-world web interaction with grounded language agents.** In *arXiv*.
- Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. 2024. **τ -bench: A benchmark for tool-agent-user interaction in real-world domains.** In *arXiv*.
- Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik R Narasimhan. 2023b. **Tree of thoughts: Deliberate problem solving with large language models.** In *Conference on Neural Information Processing Systems*.
- Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. 2023c. **React: Synergizing reasoning and acting in language models.** In *International Conference on Learning Representations*.
- Yixin Ye, Zhen Huang, Yang Xiao, Ethan Chern, Shijie Xia, and Pengfei Liu. 2025. **LIMO: Less is more for reasoning.** In *arXiv*.
- Ziyu Ye, Rishabh Agarwal, Tianqi Liu, Rishabh Joshi, Sarmishta Velury, Quoc V. Le, Qijun Tan, and Yuan Liu. 2024. **Evolving alignment via asymmetric self-play.** In *arXiv*.
- Edward Yeo, Yuxuan Tong, Morry Niu, Graham Neubig, and Xiang Yue. 2025. **Demystifying long chain-of-thought reasoning in llms.** In *arXiv*.
- Hao Yi, Qingyang Li, Yulan Hu, Fuzheng Zhang, Di Zhang, and Yong Liu. 2025. **Sppd: Self-training with process preference learning using dynamic value margin.** *arXiv preprint arXiv:2502.13516*.
- Dian Yu, Baolin Peng, Ye Tian, Linfeng Song, Haitao Mi, and Dong Yu. 2024a. **Siam: Self-improving code-assisted mathematical reasoning of large language models.** In *arXiv*.
- Fei Yu, Anningzhe Gao, and Benyou Wang. 2024b. **Ovm, outcome-supervised value models for planning in mathematical reasoning.** In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 858–875.
- Ping Yu, Jing Xu, Jason E Weston, and Ilia Kulikov. 2024c. **Distilling system 2 into system 1.** In *The First Workshop on System-2 Reasoning at Scale, NeurIPS'24*.
- Qiyng Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, et al. 2025. **Dapo: An open-source llm reinforcement learning system at scale.** *arXiv preprint arXiv:2503.14476*.

- Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2024d. Mm-vet: Evaluating large multimodal models for integrated capabilities. In *International Conference on Machine Learning*, pages 57730–57754. PMLR.
- Xiao Yu, Baolin Peng, Vineeth Vajipey, Hao Cheng, Michel Galley, Jianfeng Gao, and Zhou Yu. 2024e. Exact: Teaching ai agents to explore with reflective-mcts and exploratory learning. *arXiv preprint arXiv:2410.02052*.
- Yufeng Yuan, Yu Yue, Ruofei Zhu, Tiantian Fan, and Lin Yan. 2025. What’s behind ppo’s collapse in long-cot? value optimization holds the secret. *arXiv preprint arXiv:2503.01491*.
- Zheng Yuan, Hongyi Yuan, Chengpeng Li, Guanting Dong, Keming Lu, Chuanqi Tan, Chang Zhou, and Jingren Zhou. 2023. Scaling relationship on learning mathematical reasoning with large language models.
- Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruochi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhui Chen. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *arXiv*.
- Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. 2022. Star: Bootstrapping reasoning with reasoning. *Advances in Neural Information Processing Systems*, 35:15476–15488.
- Weihao Zeng, Yuzhen Huang, Qian Liu, Wei Liu, Keqing He, Zejun Ma, and Junxian He. 2025a. Simplerl-zoo: Investigating and taming zero reinforcement learning for open base models in the wild. *arXiv preprint arXiv:2503.18892*.
- Weihao Zeng, Yuzhen Huang, Wei Liu, Keqing He, Qian Liu, Zejun Ma, and Junxian He. 2025b. 7b model and 8k examples: Emerging reasoning with reinforcement learning is both effective and efficient. <https://hkust-nlp.notion.site/simplerl-reason>. Notion Blog.
- Yirong Zeng, Xiao Ding, Yuxian Wang, Weiwen Liu, Wu Ning, Yutai Hou, Xu Huang, Bing Qin, and Ting Liu. 2025c. itool: Boosting tool use of large language models via iterative reinforced fine-tuning. In *arXiv*.
- Zhiyuan Zeng, Qinyuan Cheng, Zhangyue Yin, Yunhua Zhou, and Xipeng Qiu. 2025d. Revisiting the test-time scaling of o1-like models: Do they truly possess test-time scaling capabilities? *arXiv preprint arXiv:2502.12215*.
- Zhongshen Zeng, Yinhong Liu, Yingjia Wan, Jingyao Li, Pengguang Chen, Jianbo Dai, Yuxuan Yao, Rongwu Xu, Zehan Qi, Wanru Zhao, Linling Shen, Jianqiao Lu, Haochen Tan, Yukang Chen, Hao Zhang, Zhan Shi, Bailin Wang, Zhijiang Guo, and Jiaya Jia. 2024. MR-ben: A meta-reasoning benchmark for evaluating system-2 thinking in LLMs. In *Conference on Neural Information Processing Systems*.
- Yuxiang Zhai, Hao Bai, Zipeng Lin, Jiayi Pan, Shengbang Tong, Yifei Zhou, Alane Suhr, Saining Xie, Yann LeCun, Yi Ma, and Sergey Levine. 2024. Fine-tuning large vision-language models as decision-making agents via reinforcement learning. In *arXiv*.
- Dan Zhang, Sining Zhoubian, Ziniu Hu, Yisong Yue, Yuxiao Dong, and Jie Tang. 2024a. ReST-MCTS*: LLM self-training via process reward guided tree search. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.
- Jintian Zhang, Yuqi Zhu, Mengshu Sun, Yujie Luo, Shuofei Qiao, Lun Du, Da Zheng, Huajun Chen, and Ningyu Zhang. 2025a. Lightthinker: Thinking step-by-step compression. In *arXiv*.
- Kechi Zhang, Ge Li, Jia Li, Yihong Dong, and Zhi Jin. 2025b. Focused-dpo: Enhancing code generation through focused preference optimization on error-prone points. *arXiv preprint arXiv:2502.11475*.
- Kongcheng Zhang, Qi Yao, Baisheng Lai, Jiaxing Huang, Wenkai Fang, Dacheng Tao, Mingli Song, and Shunyu Liu. 2025c. Reasoning with reinforced functional token tuning. *arXiv preprint arXiv:2502.13389*.
- Lunjun Zhang, Arian Hosseini, Hritik Bansal, Mehran Kazemi, Aviral Kumar, and Rishabh Agarwal. 2025d. Generative verifiers: Reward modeling as next-token prediction. In *arXiv*.
- Ming-Liang Zhang, Fei Yin, and Cheng-Lin Liu. 2023a. A multi-modal neural geometric solver with textual clauses parsed from diagram. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 3374–3382.
- Qiyuan Zhang, Yufei Wang, Yuxin Jiang, Liangyou Li, Chuhan Wu, Yasheng Wang, Xin Jiang, Lifeng Shang, Ruiming Tang, Fuyuan Lyu, and Chen Ma. 2025e. Crowd comparative reasoning: Unlocking comprehensive evaluations for llm-as-a-judge. In *arXiv*.

- Qiyuan Zhang, Yufei Wang, Tiezheng YU, Yuxin Jiang, Chuhan Wu, Liangyou Li, Yasheng Wang, Xin Jiang, Lifeng Shang, Ruiming Tang, Fuyuan Lyu, and Chen Ma. 2025f. **Reviseval: Improving LLM-as-a-judge via response-adapted references**. In *International Conference on Learning Representations*.
- Shun Zhang, Zhenfang Chen, Yikang Shen, Mingyu Ding, Joshua B. Tenenbaum, and Chuang Gan. 2023b. **Planning with large language models for code generation**. In *International Conference on Learning Representations*.
- Xinyu Zhang, Yuxuan Dong, Yanrui Wu, Jiaxing Huang, Chengyou Jia, Basura Fernando, Mike Zheng Shou, Lingling Zhang, and Jun Liu. 2025g. **Physreason: A comprehensive benchmark towards physics-based reasoning**. *CoRR*, abs/2502.12054.
- Xuan Zhang, Chao Du, Tianyu Pang, Qian Liu, Wei Gao, and Min Lin. 2024b. **Chain of preference optimization: Improving chain-of-thought reasoning in LLMs**. In *Conference on Neural Information Processing Systems*.
- Yiwen Zhang, Yifu Wu, Wenyue Hua, Xiang Lu, and Xuming Hu. 2025h. Understanding dynamic diffusion process of llm-based agents under information asymmetry. *arXiv preprint arXiv:2502.13160*.
- Yongheng Zhang, Qiguang Chen, Jingxuan Zhou, Peng Wang, Jiasheng Si, Jin Wang, Wenpeng Lu, and Libo Qin. 2024c. **Wrong-of-thought: An integrated reasoning framework with multi-perspective verification and wrong information**. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 6644–6653.
- Yunxiang Zhang, Muhammad Khalifa, Lajanugen Logeswaran, Jaekyeom Kim, Moontae Lee, Honglak Lee, and Lu Wang. 2024d. **Small language models need strong verifiers to self-correct reasoning**. In *ACL (Findings)*.
- Yuxiang Zhang, Shangxi Wu, Yuqi Yang, Jiangming Shu, Jinlin Xiao, Chao Kong, and Jitao Sang. 2024e. **o1-coder: an o1 replication for coding**. In *arXiv*.
- Yu Zhao, Huifeng Yin, Bo Zeng, Hao Wang, Tianqi Shi, Chenyang Lyu, Longyue Wang, Weihua Luo, and Kaifu Zhang. 2024. **Marco-o1: Towards open reasoning models for open-ended solutions**. In *arXiv*.
- Kunhao Zheng, Jesse Michael Han, and Stanislas Polu. 2021. Minif2f: a cross-system benchmark for formal olympiad-level mathematics. *arXiv preprint arXiv:2109.00110*.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. 2023a. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623.
- Rui Zheng, Shihan Dou, Songyang Gao, Yuan Hua, Wei Shen, Binghai Wang, Yan Liu, Senjie Jin, Qin Liu, Yuhao Zhou, et al. 2023b. **Secrets of rlhf in large language models part i: Ppo**. *arXiv preprint arXiv:2307.04964*.
- Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2024. AGIEval: A human-centric benchmark for evaluating foundation models. In *Findings of North American Chapter of the Association for Computational Linguistics*, pages 2299–2314.
- Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V Le, et al. 2023a. **Least-to-most prompting enables complex reasoning in large language models**. In *The Eleventh International Conference on Learning Representations*.
- Enyu Zhou, Guodong Zheng, Binghai Wang, Zhiheng Xi, Shihan Dou, Rong Bao, Wei Shen, Limao Xiong, Jessica Fan, Yurong Mou, Rui Zheng, Tao Gui, Qi Zhang, and Xuanjing Huang. 2025. **Rmb: Comprehensively benchmarking reward models in llm alignment**. In *arXiv*.
- Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023b. **Instruction-following evaluation for large language models**. In *arXiv*.
- Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Yonatan Bisk, Daniel Fried, Uri Alon, et al. 2023c. **Webarena: A realistic web environment for building autonomous agents**. *arXiv preprint arXiv:2307.13854*.

Contents

1	Introduction	1		
2	What to Scale	2		
2.1	Parallel Scaling	4	5.3 Controllability	16
2.2	Sequential Scaling	4	5.4 Scalability	17
2.3	Hybrid Scaling	4		
2.4	Internal Scaling	5		
3	How to Scale	5		
3.1	Tuning-based Approaches	5	6 Organization and Trends in Test-time scaling	17
3.1.1	Supervised Finetuning (SFT)	5		
3.1.2	Reinforcement Learning (RL)	6	7 A Hand-on Guideline for Test-time Scaling	18
3.2	Inference-based Approaches	6		
3.2.1	Stimulation	7	8 Challenges and Opportunities	19
3.2.2	Verification	8	8.1 More Scaling is the Frontier	19
3.2.3	Search	9	8.2 Clarifying the Essence of Techniques in Scaling is the Foundation	20
3.2.4	Aggregation	10	8.3 Optimizing Scaling is the Key	21
			8.4 Generalization across Domains is the Mainstream	21
4	Where to Scale	10	9 Conclusion	21
4.1	Reasoning-intensive Tasks	10		
4.2	Agentic Tasks	11		
4.3	Others	12	A Detailed Outcome Verification Methods	40
5	How Well to Scale	13	A.1 Verifier Model-Based Scoring	40
5.1	Performance	13	A.2 Self-Consistency and Voting Mechanisms	40
5.2	Efficiency	14	A.3 Tool-Assisted and Heuristic Verification	40
5.2.1	Key Concern and Multi-Dimensional Perspectives of Efficiency	14	B Representative Methods	40
5.2.2	General Computational Cost Metrics	15	B.1 Best-of-N	40
5.2.3	Reasoning Efficiency Metrics	15	B.2 Majority Voting	41
			B.3 Process Reward Model	41
			B.4 MCTS	41
			B.5 Self-Refine	42
			B.6 Tree-of-Thought	42
			B.7 Reinforcement Learning	43

A Detailed Outcome Verification Methods

This appendix expands on the outcome verification techniques employed at test time in LLMs. Unlike training-time methods (*e.g.*, RL fine-tuning), these techniques operate on the fly during inference, often by generating multiple solutions and using a *proposer–verifier* framework.

A.1 Verifier Model-Based Scoring

The verifier, which is typically trained using human feedback or supervised data (*e.g.*, as in (Cobbe et al., 2021; Lambert et al., 2024)), scores each candidate based on its expected correctness or quality. Variants include i) pairwise comparison verifiers (Liu et al., 2025b), where candidates are compared against each other to determine a winner, ii) weighted voting systems (Wettig et al., 2024; Li et al., 2024b) that use the verifier’s scores to combine outputs, iii) LLM-based verifiers that prompt LLM to perform evaluation instruction, like LLM-as-a-Judge (Zheng et al., 2023a; Zhang et al., 2025e,f), LLM-based Evaluator (Liu et al., 2023c; Xu et al., 2023; Jiang et al., 2024a), Critic-based Model (Gao et al., 2024a; McAleese et al., 2024).

A.2 Self-Consistency and Voting Mechanisms

Self-consistency techniques generate multiple independent reasoning chains and choose the final answer based on majority voting (Wang et al., 2023b). The underlying assumption is that if several chains converge on the same answer, that answer is more likely to be correct. Some approaches (Taubenfeld et al., 2025; Mahmud et al., 2025) also incorporate confidence scores or soft-voting schemes to mitigate noise in individual outputs. In place of multiple samples from one model, one can also have multiple models (Wan et al., 2025; Wu and Ito, 2025; Wang et al., 2025f; Feng et al., 2024; Chen et al., 2024b): if a majority (or consensus) of these “agents” agree on an answer, trust it; if they diverge, it may trigger further scrutiny. This is effectively an ensemble vote.

A.3 Tool-Assisted and Heuristic Verification

In domains like code generation or mathematical problem-solving, outcome verification can be implemented via direct execution or rule-based checks. For example, candidate programs are executed on sample test cases to ensure they produce correct results, while in math tasks, answers can be validated by plugging them back into the original equations. These approaches serve as an external check on the LLM’s internal reasoning.

Execution-Based Verification. In programming tasks, the ultimate test of correctness is running the code (Tian et al., 2025; Ni et al., 2024; Yang et al., 2024; Ni et al., 2023). For math problems, a simple heuristic is to verify the answer by plugging it back into the original equation or problem constraints. Similarly, if a puzzle answer must satisfy certain conditions, those can be programmatically checked.

Fact-Checking via Retrieval. In open-domain QA or tasks that risk factual errors, search engines or knowledge bases serve as powerful verifiers (Wei et al., 2024b; Vladika and Matthes, 2024; Asai et al., 2023; Peng et al., 2023). An LLM may draft an answer, but then the system issues search queries (based on the answer’s claims) to find supporting evidence. If the retrieved documents contradict the LLM’s answer, the answer is likely incorrect and can be rejected or revised. Some frameworks generate answers in a “closed-book” fashion, then do a *post-hoc* retrieval to validate facts. This idea overlaps with Retrieval-Augmented Generation (Salemi and Zamani, 2024), but the focus is on post-generation validation – essentially checking if the answer aligns with external truth.

Rule-Based Filters. In some applications, simple heuristic filters (Bai et al., 2022; Sun et al., 2023; Weber et al., 2024) can automatically reject bad outputs. For a dialogue system, one might have a list of forbidden answers (certain unsafe or nonsensical replies) and if the model outputs one, the system can either regenerate or adjust it. These aren’t “outcome-based” in terms of correctness, but they verify the output against predefined rules of form and content.

B Representative Methods

B.1 Best-of-N

The “Best-of-N” strategy is a *TTS* approach in which a model generates N candidate outputs for a given input and then selects the best one according to a chosen evaluation metric (Wu et al., 2024d). Mathematically, given an input x and model f , one draws N independent outputs $y_1, \dots, y_N \sim f(x)$ (*e.g.*, via different random seeds or sampling strategies) and chooses the result $\hat{y} = \arg \max_{i=1}^N M(y_i)$, where M is a quality scoring function. At the cost of additional inference compute, increasing N raises the probability of obtaining a high-quality outcome (for example, if each attempt succeeds with probability p , then a best-of- N run succeeds with probability $1 - (1 - p)^N$). This technique leverages extra computation to boost performance (Kang et al., 2025) and has been applied in real-world settings ranging from complex reasoning and code generation with LLMs to enhancing image synthesis quality in diffusion models (Ma et al., 2025b).

B.2 Majority Voting

Majority voting is a fundamental ensemble strategy for *TTS* that aggregates multiple independent predictions to make a final decision. In this approach, each model or inference (voter) casts a vote for a predicted outcome, and the output chosen is the one with the highest number of votes (*i.e.*, the mode of the predictions). Formally, given an ensemble of M models h_1, h_2, \dots, h_M each producing a prediction $h_m(x)$ for input x , the majority vote outcome is defined as

$$\hat{y} = \arg \max_c \sum_{m=1}^M \mathbf{1}\{h_m(x) = c\},$$

where $\mathbf{1}\{\cdot\}$ is the indicator function and c ranges over all possible classes or outputs. This test-time inference technique leverages additional computing at inference to improve reliability without retraining models, and it is widely used in real-world applications, such as combining votes of decision trees in a random forest, consolidating crowd-sourced annotations, or enhancing the consistency of answers from LLMs by selecting the most frequent response.

B.3 Process Reward Model

A Process Reward Model (PRM) (Uesato et al., 2022; Pfau et al., 2024) is a reward model designed to evaluate an entire reasoning trajectory on a step-by-step basis. Formally, given an input problem x and a sequence of reasoning steps z_1, z_2, \dots, z_T leading to a final output y , we can represent this full reasoning trace as:

$$S^T = (x, z_1, z_2, \dots, z_T, y),$$

and define the PRM as a function that assigns a real-valued score:

$$r : S^T \rightarrow \mathbb{R},$$

mapping a possible reasoning process S^T to a reward score (Choudhury, 2025; Ma et al., 2025c). Intuitively, $r(S^T)$ is higher when the reasoning process is logical, valid, and leads to a correct solution, and lower (or negative) when the reasoning is flawed. PRMs are typically trained on human or algorithmic annotations for each step, internalizing a notion of “partial credit” to evaluate correctness and relevance at each stage.

PRMs play a crucial role in *TTS* strategies such as stepwise beam search and self-consistency verification. They have been successfully applied in mathematical reasoning, code generation, automated theorem proving, and decision-making tasks. By leveraging PRMs, models can optimize not only for correctness but also for process coherence, making AI systems more transparent and robust.

B.4 MCTS

Monte Carlo Tree Search (MCTS) is a simulation-based decision-making algorithm for sequential decision problems, often formalized as a Markov Decision Process (MDP). It incrementally builds a search tree by sampling many possible future trajectories (playouts) and using their outcomes to estimate the value of decisions. Unlike brute-force search, MCTS selectively explores the most promising actions by balancing exploration (trying unexplored or uncertain moves) and exploitation (favoring moves with high estimated reward). Each iteration of MCTS consists of four phases:

1. Selection – Recursively select child actions that maximize a heuristic value until reaching a leaf node. A common selection strategy is the Upper Confidence Bound for Trees (UCT):

$$\text{UCT}(a) = \frac{w_a}{n_a} + c \sqrt{\frac{\ln N}{n_a}},$$

where w_a is the total simulation reward, n_a is the visit count for action a , N is the total simulations from the parent state, and $c > 0$ is an exploration constant.

2. Expansion – Once a leaf state is reached, new child nodes are created by simulating unexplored actions.
3. Simulation (Rollout) – Perform a Monte Carlo simulation by selecting actions to simulate a full episode to the end, providing an estimate of the node’s value.
4. Backpropagation – Propagate the simulation result back up the tree, updating the statistics of each node along the path.

MCTS is well-suited for *TTS* because its anytime nature allows flexible computation budgets. At test time, running MCTS for longer or with more rollouts leads to deeper search and better decisions. Notably, AlphaGo used MCTS at runtime to refine moves, significantly improving performance without additional training.

Researchers are leveraging MCTS to enhance test-time reasoning in other AI domains. MCTS-Judge improves code correctness evaluation by systematically exploring reasoning paths, raising verification accuracy significantly. Similarly, hybrid approaches integrate MCTS into generative model inference for problem-solving, such as solving Sudoku puzzles through sequential search.

By repeating these steps, MCTS concentrates simulations on the most promising branches. In the limit, MCTS value estimates converge to the optimal values in certain perfect-information games.

B.5 Self-Refine

Self-Refine (Madaan et al., 2023) is an advanced TTS technique that enables an LLM to iteratively improve its own outputs through self-generated feedback. Introduced by Madaan et al. (2023), the Self-Refine framework is inspired by how humans revise a draft: The model first produces an initial answer, then critiques or evaluates that answer, and finally uses the critique to refine the answer. This feedback-refinement loop can be repeated multiple times, progressively polishing the output. Notably, Self-Refine requires no additional training data or fine-tuning – the same pre-trained model acts as the initial answer generator, the feedback provider, and the refiner. For sufficiently powerful models, this self-iteration yields significantly better results, presumably because it is easier for a model to identify and fix errors in a given solution than to produce a perfect solution in one attempt. In essence, Self-Refine leverages test-time compute to let the model “think twice (or more)” about its answer, leading to higher-quality and more reliable outputs.

Formally, consider an input x and a language model M_θ with parameters θ , defining a conditional distribution $P_\theta(y | x)$ over possible outputs y . The Self-Refine procedure generates a sequence of outputs $y^{(0)}, y^{(1)}, \dots, y^{(T)}$ as follows:

1. Initial Output Generation: The model first produces an initial response:

$$y^{(0)} = M_\theta(x). \quad (8)$$

2. Feedback Generation: At each refinement step $t = 1, 2, \dots, T$, the model evaluates the previous output and generates feedback:

$$f^{(t)} = M_\theta(x, y^{(t-1)}; \text{feedback-prompt}). \quad (9)$$

3. Refinement Step: Using the generated feedback, the model updates its output:

$$y^{(t)} = M_\theta(x, y^{(t-1)}, f^{(t)}; \text{refine-prompt}). \quad (10)$$

This feedback-refinement loop continues iteratively until a stopping condition is met, such as reaching a predefined number of iterations T or detecting convergence in the output quality. The Self-Refine approach enhances model reliability by progressively improving its responses without requiring additional training.

B.6 Tree-of-Thought

Complex reasoning problems often require exploring different lines of thought before arriving at a correct solution. CoT prompting was a first step in this direction: CoT guides the model to produce a single sequence of intermediate reasoning steps (a linear chain) leading to the answer. This improves the model’s performance on tasks requiring multi-step logic by breaking the problem into a step-by-step narrative. However, CoT still follows a single path – if the model makes a wrong turn in the reasoning chain, it cannot recover because it doesn’t revisit earlier decisions. Tree-of-Thought (Yao et al., 2023b), by contrast, generalizes CoT to a branching search. At each reasoning step, the model can generate multiple candidate thoughts instead of one, forming a tree of possibilities. It evaluates these candidates (using heuristics or self-evaluation prompts) and selects the most promising branch(es) to continue expanding (Bi et al., 2024). This test-time exploration allows the model to consider alternative approaches and scale up inference computation as needed – much like how a human might try different reasoning avenues for a hard problem. Researchers have categorized ToT and similar strategies (*e.g.*, graph-of-thought) as “X-of-Thought” (XoT) reasoning methods, which significantly improve LLM reasoning by introducing iterative, structured inference without additional training.

ToT can be modeled as a search process through a state space of partial solutions, where each state encodes the sequence of thoughts (intermediate steps) explored so far. Let S be the set of all possible reasoning states for a given problem. The initial state s_0 contains the problem statement, and a goal state $s \in S$ represents a complete solution.

Thought Generation (State Transitions): At each step, the language model serves as a thought generator function G . Given the current state (context) s , the model generates a set of next-step thoughts:

$$G(s) \rightarrow \{t_1, t_2, \dots, t_b\} \quad (11)$$

where each t_i represents a candidate next reasoning step. Each thought extends the current reasoning path, yielding a new state:

$$s_i = s \oplus t_i \quad (12)$$

where \oplus denotes concatenation of the thought to the sequence.

State Evaluation (Heuristic Function): To guide the search, ToT uses an evaluation function $f(s)$ that estimates the quality of a partial state s :

$$f : S \rightarrow \mathbb{R} \quad (13)$$

This function may be implemented by the model itself using a self-evaluation prompt or a scoring heuristic.

Search Algorithm (Tree Expansion): ToT can employ different search strategies, including:

- **Breadth-First Search (BFS):** Expands all plausible thoughts at each depth, keeping the top b best states based on $f(s)$.
- **Depth-First Search (DFS):** Follows the most promising thought path deeply, backtracking if necessary.

Each strategy allows ToT to control computational budgets by limiting depth d (number of steps) and branching factor b (number of candidates per step).

Solution Extraction: A state s is considered a valid solution if it satisfies the problem constraints. The search continues until:

1. A goal state is reached.
2. The computational budget (depth or number of states evaluated) is exhausted.

This framework formalizes ToT as an organized search over the space of reasoning sequences, allowing models to iteratively refine and explore multiple potential solutions during test-time inference.

B.7 Reinforcement Learning

Reinforcement learning can play a pivotal role in unlocking effective TTS for language models. The process of inference itself can be formulated as a sequential decision-making problem: at each step in generating a solution, *e.g.*, each token in a reasoning chain or each attempt at an answer, the model (agent) must decide whether to continue reasoning, which direction to explore, or when to stop and output an answer. By training the model with RL, we can explicitly reward outcomes that lead to correct or high-quality answers, thereby encouraging the model to make better use of the extra inference steps available. This addresses a key challenge in TTS: simply allowing a model to think longer doesn't guarantee better answers unless the model knows how to productively use that extra time (it could otherwise repeat mistakes or terminate too early). RL provides a feedback-driven way to learn such behaviors. In fact, prior approaches to improve reasoning in LLMs often relied solely on imitation learning (learning from observed human or AI reasoning traces), which can limit a model to mimicking given patterns (Hou et al., 2025). By contrast, RL enables self-exploration: the model can try diverse reasoning paths and learn from trial-and-error which strategies yield the highest reward (for example, reaching a correct solution). This means an RL-trained language model can learn dynamic inference policies—such as when to double-check an intermediate result or how to backtrack and correct itself if the reasoning seems to be going astray. Recent research indeed shows that combining chain-of-thought reasoning with reinforcement learning techniques leads to improved inference-time performance.