A Systematic Review of Meta-Reasoning for Large Reasoning Models

Haoran Yu[1*] **Zilong Wang**[1*] **Kang Pan**[2] **Xiao Zhang**[2] **Erli Meng**[2] **Shuai Li**[1,3†]
[1]Shanghai Jiao Tong University, Shanghai, China
[2]Xiaomi Corporation, Beijing, China
[3]Shanghai Innovation Institute, Shanghai, China
harayn4@gmail.com
{wangzilong, shuaili8}@sjtu.edu.cn
{pankang, zhangxiao16, mengerli}@xiaomi.com

Abstract

In this paper, we survey recent advances in meta-reasoning for Large Reasoning Models (LRMs). Although LRMs have made remarkable progress, they face critical limitations including overthinking on simple queries, underthinking on complex problems, and ineffective exploration lacking strategic direction. We examine how meta-reasoning addresses these challenges by providing the necessary mechanisms to dynamically regulate and optimize the reasoning process, and propose a framework to categorize existing studies into three core components: the monitor which provides self-perception of the reasoning state, the controller which dictates the meta-policy for regulating actions, and the mechanism which executes the algorithmic mapping between them. By systematically reviewing these developments and foundational challenges, we aim to map the current landscape and identify promising directions for creating more adaptive, efficient, and self-regulating reasoning systems.

1 Introduction

In recent years, Large Language Models (LLMs) have achieved remarkable progress through the scaling of parameters and training data (Kaplan et al., 2020; Zhao et al., 2026; Minaee et al., 2025; Naveed et al., 2025). However, despite their success in general-purpose tasks, these models often face limitations in solving complex, multi-step problems. This challenge has driven the emergence of Large Reasoning Models (LRMs) (Zhang et al., 2026a; Chen et al., 2026), which employ test-time scaling (Snell et al., 2024; Muennighoff et al., 2025) to improve capabilities by allocating additional computational resources during inference. By utilizing Chain-of-Thought (CoT) (Wei et al., 2022) to explicitly generate intermediate reasoning steps before final answers, advanced models such as OpenAI's o1 (OpenAI et al., 2026) and DeepSeek's R1 (Guo et al., 2025a) have significantly improved performance in mathematics, programming, logical reasoning and other reasoning tasks.

Despite the success of Chain-of-Thought in making reasoning explicit, monolithic autoregressive reasoning can entangle strategic planning with step-by-step execution (Yang et al., 2026b). Learning from linearized solutions alone may not capture the exploration and verification processes underlying complex tasks (Xiang et al., 2025). Insufficient regulation of reasoning depth and direction can manifest in three recurring failure modes: underthinking (Wang et al., 2025h), where models allocate insufficient computation to complex sub-steps; overthinking (Sui et al., 2025; Feng et al., 2025), characterized by redundant reasoning (Fu et al., 2025); and the exploration of unproductive paths (Parashar et al., 2025). Such misallocation of inference effort can reduce computational efficiency and task performance.

The notion of metacognition originates from cognitive science, where it refers to knowledge and regulation of cognitive processes (Flavell, 1979). Nelson (1990) formalizes this relationship through object- and meta-levels connected by monitoring and control. Recent work develops related ideas for LLM reasoning through internalized search (Xiang et al., 2025) and explicit metacognitive regulation (Dong et al., 2025), aiming to improve strategy adaptation, error correction, and the allocation of inference-time computation.

Building on this monitoring–control perspective, we organize existing studies using a Monitor–Controller–Mechanism taxonomy (§2; Figure 1): what is assessed (§3), which meta-actions regulate reasoning (§4), and how these mappings are learned or implemented (§5). We further discuss evaluation methods and benchmarks (§6) and identify open directions arising from interactions

[*]Equal contribution.
[†]Corresponding author.

Proceedings of the 15th International Joint Conference on Natural Language Processing and the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 5392–5414
November 6-10, 2026 ©2026 Association for Computational Linguistics

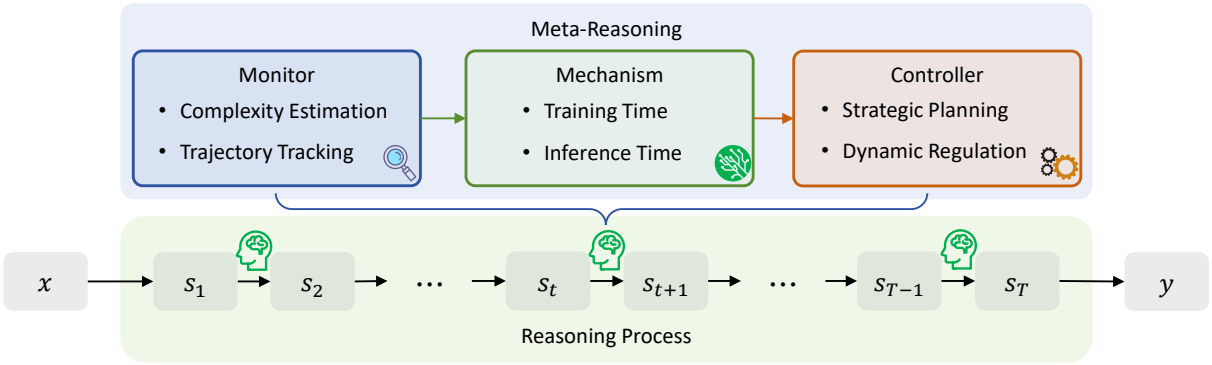


Figure 1: Architecture of our meta-reasoning framework.

among these roles (§7).

Related Works and Surveys Existing surveys offer complementary perspectives on reasoning regulation. Regarding the **monitor** and **mechanism**, reviews have discussed uncertainty estimation (Liu et al., 2025; Shorinwa et al., 2025), reward models (Zheng et al., 2026; Liu et al., 2026b), reinforcement learning (RL) for LRMs (Zhang et al., 2025c), and steering (Zhang et al., 2026b). Others focus on specific **control** objectives, particularly reasoning-efficiency and concise chains (Sui et al., 2025; Feng et al., 2025; Liu et al., 2026c; Qu et al., 2025a; Yue et al., 2025a; Zhu and Li, 2025), self-refinement (Kamoi et al., 2024), or tool learning (Xu et al., 2025b). Broader accounts include Meta-CoT (Xiang et al., 2025), which develops a search-based view of reasoning and its acquisition through post-training and the MARL-centered survey of meta-thinking by Bilal et al. (2025). Meta-R1 (Dong et al., 2025) integrates planning, online monitoring and control, and termination in a two-level system. In a position paper, Yan et al. (2025) argue for a Bayesian meta-reasoning framework integrating self-awareness, monitoring, evaluation, regulation, and meta-reflection. Our survey complements these perspectives by jointly organizing existing methods along three analytical axes: monitoring signals, control strategies, and their algorithmic realization.

2 Motivation and Framework

In this section, we review the motivations for meta-reasoning (§2.1) and the taxonomy (§2.2). Building on established monitoring–control accounts (Nelson, 1990) and recent conceptual work on LLM meta-reasoning (Yan et al., 2025), we distinguish monitoring signals, control decisions, and the mechanisms that learn or implement their mapping. Figure 2 summarizes this taxonomy and the categories reviewed in §3–§5.

2.1 The Necessity of Meta-Reasoning

One motivation for meta-reasoning is the mismatch between the **search process underlying complex reasoning** and its linearized presentation in training data (Xiang et al., 2025). Written solutions may omit intermediate processes such as hypothesis generation and backtracking. This motivates learning and regulating reasoning trajectories rather than merely imitating their final linearized presentation.

Compounding this mismatch is the coupling of high-level planning and low-level execution in **monolithic reasoning pipelines** (Yang et al., 2026b). This coupling can make it difficult to allocate computation selectively between strategic exploration and routine execution. Moreover, it can restrict initial strategy diversity, so a suboptimal early choice may persist without effective backtracking or strategic redirection (Sui et al., 2026; Yan et al., 2025).

Insufficient strategic regulation can manifest in three primary failure modes: **under-thinking** (Wang et al., 2025h), where models allocate insufficient computation to complex decision nodes and prematurely switch traces; **over-thinking** (Fu et al., 2025; Sui et al., 2025; Feng et al., 2025), where models become trapped in redundant verification loops; and **unproductive exploration** (Qu et al., 2025b), where computational resources are spent pursuing invalid branches without guided direction.

These limitations motivate monitoring and regulation of reasoning. We organize the surveyed approaches around three complementary roles:

- The Monitor evaluates both problem difficulty and the ongoing reasoning process. It identifies the current state of the reasoning to provide the necessary signals for decision-

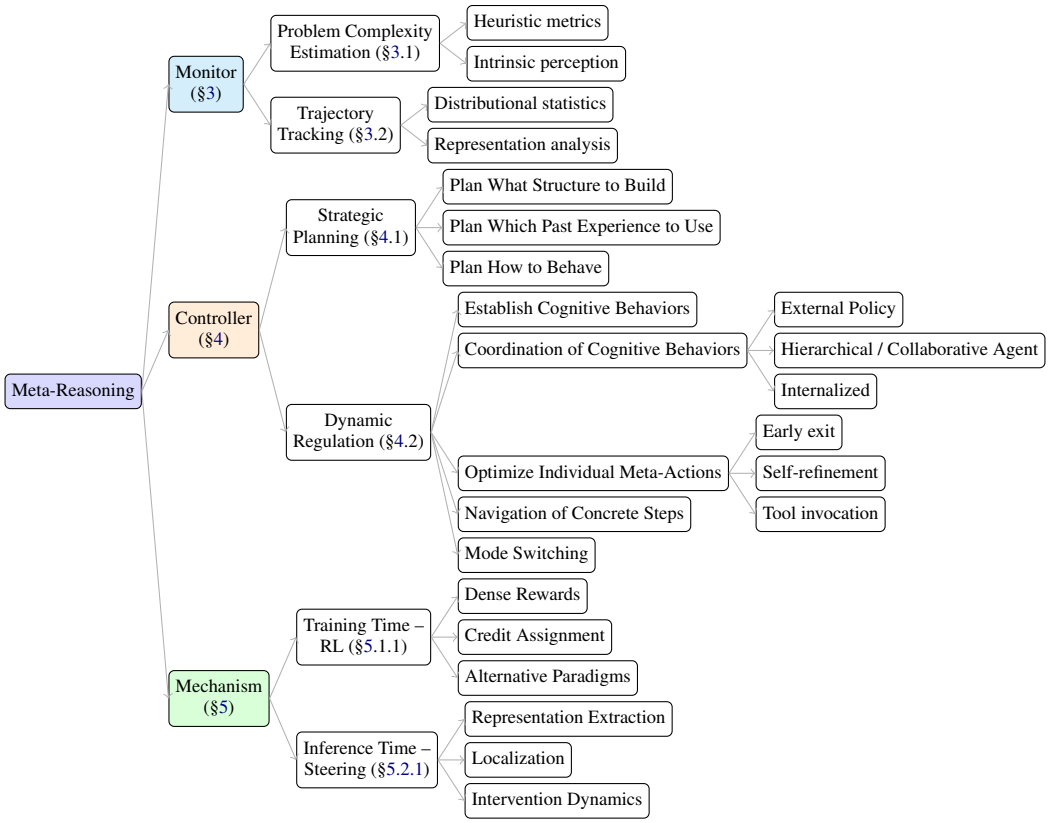


Figure 2: Hierarchical taxonomy of meta-reasoning methods, organized by Monitor, Controller, and Mechanism. Branches summarize the categories discussed in §3–§5.

making and regulation.

- The Controller provides guidance for balancing exploration and exploitation to ensure that complex steps receive sufficient depth while easy or high-confidence steps remain efficient.
- The Mechanism implements the underlying algorithm that connects the Monitor and Controller, which defines or learns the mapping between observed states and cognitive actions.

2.2 Formulation

Given an input problem x, the model, with parameters π_θ, generates a reasoning trajectory $\tau = (x, s_1, s_2, ..., s_T, y)$, where each s_t can be a token or a reasoning step and y is the final solution. At each step t, the process is as follows:

The monitor, with parameters θ_m, acts as the system's perceptual component. It observes the current trajectory $\tau_t = (x, s_1, s_2, ..., s_t)$ and produces a state representation $h_t = f(\tau_t; \theta_m)$, which quantifies properties like difficulty, uncertainty or progress.

The controller, with parameters $\pi_{\theta_{\text{meta}}}$, is the meta-policy that takes this representation h_t as input and selects the meta-action $a_t \sim \pi_{\theta_{\text{meta}}}(\cdot|h_t)$ from a predefined action space. This action a_t can be a cognitive behavior (e.g. verification, backtracking), a selected mode (e.g., fast thinking), or a regular reasoning step. The model's base policy π_θ then samples the next step $s_{t+1} \sim \pi_\theta(\cdot|\tau_t, a_t)$.

The Mechanism implements and learns (if trainable) the controller's policy, divided into two categories: Training-time mechanisms learn the $\pi_{\theta_{\text{meta}}}$ by optimizing an objective function $\mathcal{J}$. Common approaches include Supervised Fine-Tuning (SFT), which minimizes the negative log-likelihood of expert actions: $\mathcal{J}_{\text{SFT}} = \mathbb{E}[-\log \pi_{\theta_{\text{meta}}}(a_t^*|h_t)]$, and Reinforcement Learning (RL), which maximizes the expected cumulative reward $\mathcal{J}_{\text{RL}} = \mathbb{E}[\sum_t R(\tau_t, a_t)]$.

Inference-time mechanisms intervene directly on the generation process without updating parameters. These methods include rule-based controls that explicitly modify output logits or force specific tokens, and activation steering, which adds a steering vector v_{steer} to the model's internal representations H_t at layer l: $H_t'^{(l)} = H_t^{(l)} + v_{steer}$ and then $s_{t+1} \sim \pi_\theta(\cdot|\tau_t, H_t'^{(l)})$.

Functional roles, not exclusive labels The Monitor, Controller, and Mechanism are functional roles, and many methods instantiate more than one. For example, token-/hidden-state statistics that drive early exit, refinement, or tool calls combine Monitor and Controller; probes whose parameters are reused as steering vectors and entropy-based RL objectives combine Monitor and Mechanism; and SFT/RL training a model to coordinate cognitive behaviors combines Mechanism and Controller. A complete pipeline ideally realizes all three components, and improving their interaction is itself an open direction (§7).

3 Monitor

For a model to regulate or control its own reasoning, it must first be able to perceive its reasoning process. This perceptual component is the monitor. We organize existing approaches into two categories based on the object being monitored: Problem Complexity Estimation and Trajectory Tracking.

3.1 Problem Complexity Estimation

Problem-level monitoring estimates the complexity of input queries to inform the Controller's resource allocation and guide the Mechanism's learning process. Heuristic metrics, such as sampling accuracy or response length, are widely employed to improve data sampling strategies (Tong et al., 2024), construct pairs for preference optimization (Shen et al., 2025b), or define reward signals in reinforcement learning (Huang et al., 2026; Wu et al., 2025b). In contrast to these surface-level metrics, recent research explores the model's intrinsic perception of difficulty. Sheng et al. (2025) identify internal direction vectors whose magnitude implicitly dictates reasoning depth, while MASA (Kim et al., 2026) explicitly trains the model to predict problem features like difficulty alongside the solution, utilizing reward signals to align this self-assessment with actual performance.

3.2 Trajectory Tracking

Trajectory tracking monitors the ongoing reasoning process. By quantifying uncertainty and state dynamics, this component provides the essential perceptual signals that the Controller uses to take cognitive actions and that the Mechanism uses to optimize reasoning policies, whether through training updates or inference-time interventions.

Distributional statistics leverage token- or segment-level probabilities. Token-level probabilities serve as a direct proxy for confidence, whether estimated through the perplexity of the generated answer (Lu et al., 2025a; Yang et al., 2026a) or specific linguistic markers that explicitly verbalize certainty (Qiao et al., 2025). Beyond raw probability, entropy (Cheng et al., 2026) serves as a proxy for cognitive effort, identifying critical transitions where the model engages in exploratory actions or pivotal reasoning steps, but high-entropy tokens are not necessarily strategic planning tokens (Wang et al., 2026). Similarly, Yong et al. (2025) leverage information theory metrics to quantify the deviation from an ideal trajectory and the contribution of each step.

Beyond output statistics, **representation analysis** investigates the model's hidden states. Unsupervised metrics derived directly from hidden states are employed to predict answer accuracy (Vilas et al., 2026) or detect cyclic redundancy during stagnation (Li et al., 2025a). Qian et al. (2025a) identify information peaks by analyzing dependencies between intermediate and gold-answer representations. Supervised probes predict the correctness of intermediate answers (Zhang et al., 2025a), whereas MeCo extracts a metacognitive direction through principal component analysis of contrastive representation differences to inform tool use (Li et al., 2025b). These signals typically determine the timing for cognitive actions, such as early exit. Furthermore, other methods utilize these probes to detect specific cognitive patterns and leverage their parameters as steering vectors, a mechanism we will discuss in §5.

4 Controller

The controller serves as the central executive of the meta-reasoning framework, answering the question: "Given the current state of the reasoning process (as perceived by the monitor), what cognitive action should be taken next?" We organize this landscape functionally, distinguishing between **Strategic Planning**, which sets the global topology before execution, and **Dynamic Regulation**, which governs the reasoning process in real-time.

4.1 Strategic Planning

Strategic planning involves a priori decisions that establish a global reasoning topology, blueprint, or configuration before detailed step-by-step reason-

ing.

Planning What Structure to Build These methods first generate a problem-specific roadmap—a concise outline (Ning et al., 2024b), high-level strategic knowledge (Wang et al., 2024c), or a task-specific blueprint composed from atomic reasoning modules (Zhou et al., 2024). More complex controllers externalize this process via a separate instruct model (Fan et al., 2025) or by formulating strategy selection as a Multi-Armed Bandit (MAB) over candidate "meta-thoughts" (Liu et al., 2026a).

Planning Which Past Experience to Use Rather than planning from scratch, these methods leverage external memory of similar precedents (Didolkar et al., 2024, 2025), typically through metric-based retrieval (Yang et al., 2024; Wu et al., 2026). ReasonFlux (Yang et al., 2025a) further learns to construct trajectories of templates via preference learning.

Planning How to Behave The plan here is a meta-decision about the global mode, topology, or style of reasoning for the entire task: enforcing a specific style (Xu et al., 2025a; Aytes et al., 2025), selecting a predefined topology (Gao et al., 2024), allocating a reasoning budget (Nayab et al., 2026; Xu et al., 2026), or selecting a complete reasoning configuration via a MAB policy (Wang et al., 2025g) or a planner trained on action trajectories selected by iteratively narrowing the candidate set to high-success subsets (Yue et al., 2025b). A related strategy calibrates effort by deciding between "fast" and "slow" modes via reinforcement learning (Fang et al., 2025; Jiang et al., 2025a), preference learning (Rafailov et al., 2023; Luo et al., 2025), or external performance–cost routing (Pan et al., 2026).

4.2 Dynamic Regulation

While strategic planning sets the blueprint, complex problems often require real-time adjustments. Dynamic regulation refers to the controller's ability to intervene, coordinate, and switch behaviors during the reasoning process.

Establishing Cognitive Behaviors To execute discrete "meta-reasoning atoms" such as backtracking or verifying, a prerequisite is establishing the behaviors themselves. Gandhi et al. (2025) prime models via supervised fine-tuning on curated datasets of cognitive behaviors. Zhang et al. (2026d) learn temporally consistent latent action abstractions through an iterative expectation–maximization procedure that alternates self-supervised reinforcement learning and supervised fine-tuning.

Coordination of Cognitive Behaviors Once behaviors are established, the controller must invoke them at the appropriate steps. This coordination can be managed by external modules or internalized within the model itself.

- **External Policy Controllers:** These methods employ a separate, often lightweight, module or search procedures to govern the reasoning flow. For instance, Meta-reasoner (Sui et al., 2026) uses a MAB framework where each reasoning action (e.g., backtrack, restart) is treated as an arm, leveraging progress reports as context. Similarly, AutoMR (Zhang et al., 2026f) trains a separate MLP controller to observe context and sample meta-actions, dynamically constructing a reasoning DAG. In multi-agent settings, MPDF (Yang and Thomason, 2026) equips each agent with its own meta-policy, enabling active selection of meta-reasoning actions based on peer states. For search-based coordination, rStar (Qi et al., 2025b) employs Monte Carlo Tree Search (MCTS) over predefined reasoning actions using self-consistency rewards.

- **Hierarchical and Collaborative LLM Agents:** More complex coordination uses external LLMs as meta-controllers. In hierarchical control, a "leader" model guides "worker" models (Dong et al., 2025): MLPO (Estornell et al., 2025) trains a leader to fuse and evaluate answers from untrained workers; ReMA (Wan et al., 2025) decouples "meta-thinking" and "reasoning" agents that co-train via multi-turn GRPO. Peer-to-peer frameworks instead rely on collaborative or adversarial dynamics: LongxShort (Ning et al., 2025) co-optimizes long- and short-thought models via asynchronous updates, while CDG (Wang et al., 2025e) uses adversarial self-play among a prover, a helpful critic, and a misleading critic.

- **Internalized Control:** Other methods eliminate the external module, training the model to self-regulate via RL or search. RLVMR (Zhang et al., 2026e) and Satori (Shen et al., 2025a) explicitly train models to interleave

reasoning actions using rule-based rewards. CSPO (Ha et al., 2026) further refines this by using a control mask to update only the control tokens during training.

Optimizing Individual Meta-Actions Orthogonal to the coordination architecture above, a sizeable body of work targets specific meta-cognitive actions in isolation, including early exit (Yang et al., 2026a; Yue et al., 2025a), error detection and self-refinement (Kumar et al., 2025; Kamoi et al., 2024), and—particularly central for agentic LRMs—tool invocation (Xu et al., 2025b). Tool use is itself a meta-action that interfaces reasoning with the external environment: foundational reason–act formulations interleave reasoning and tool calls (Yao et al., 2023; Schick et al., 2023); decision-boundary benchmarks evaluate whether and when to call a tool (Huang et al., 2024; Ning et al., 2024a; Ross et al., 2025; Wang et al., 2025f); and adaptive triggering methods drive this decision from metacognitive or efficiency-aware signals (Li et al., 2025b; Qian et al., 2025b; Wang et al., 2025b). We expand on tool use in Appendix C.

Navigation of Concrete Reasoning Steps These methods regulate the reasoning process by actively navigating the space of concrete steps to locate optimal solution paths, rather than managing meta-behaviors. Step-wise search approaches optimize the path incrementally using tree search algorithms. Reasoning via Planning (Hao et al., 2023) adapts MCTS to balance exploration and exploitation, while RetroSearch (Lu et al., 2025b) branches at specific switch words to identify the best trajectory for fine-tuning. Alternative methods focus on sampling and filtering complete reasoning chains. For example, RaM (Sabbata et al., 2024) evaluates paths based on a utility-length trade-off during expert-iteration training.

Mode Switching Beyond navigating concrete steps, some controllers implement dynamic mode-switching, altering the reasoning strategy or solver in real-time based on intermediate states. Rule-based approaches trigger these switches deterministically, such as changing methods upon verification failure (Liu et al., 2023), intervening when progress stagnates (Khandelwal et al., 2025) and switching between explicit and latent reasoning guided by estimated confidence (Shi et al., 2026). In contrast, learned adaptive methods rely on internal metrics to govern these transitions, with recent works specifically optimizing the dynamic switching between "fast" and "slow" reasoning modes (Zhang et al., 2025d; Cheng et al., 2025b).

5 Mechanism

The mechanism serves as the underlying algorithm that learns or executes the mapping from the monitor's state observations to the controller's behavioral outputs. It bridges the gap between perception and action, encompassing both the acquisition of meta-cognitive policies during training and execution during inference.

5.1 Training Time

Training mechanisms instill meta-reasoning capabilities by fine-tuning model parameters, falling into three categories: Supervised Fine-Tuning (SFT), Preference Learning, and Reinforcement Learning (RL). Since SFT and Preference Learning for meta-reasoning are primarily data-centric—constructing expert reasoning traces (Lu et al., 2025b) or preference pairs (An et al., 2025)—we concentrate on RL in this section.

5.1.1 Reinforcement Learning

Reinforcement Learning frames reasoning as sequential decision-making. By optimizing expected returns rather than next-token likelihood, it facilitates autonomous discovery of meta-cognitive strategies through exploration.

Dense Rewards A critical challenge in long-horizon reasoning is reward sparsity: feedback is binary and deferred until the final answer. Dense rewards provide granular intermediate-step evaluations to guide reasoning and enable more precise credit assignment.

- **External Reward Models** To evaluate intermediate steps, Process Reward Models (PRMs) are evolving from expensive human annotation (Lightman et al., 2024) toward automated correctness estimation via sampling (Wang et al., 2024b), or richer signals like solution progress (Setlur et al., 2025) and coherence (Zou et al., 2025). Alternatively, to avoid step-level annotations, implicit methods derive step-wise signals directly from outcome rewards via reparameterization (Yuan et al., 2025; Cui et al., 2026).

- **Internal Reward Estimation** Instead of external judges, internal reward estimation meth-

ods derive dense signals intrinsic to the generation process. A primary strategy measures the incremental gain in confidence or accuracy between adjacent steps, utilizing changes in ground-truth log-probability (Yoon et al., 2025) or accuracy of answers sampled on sub-trajectory (Yue et al., 2026; Qu et al., 2025b; Qi et al., 2025a). Tree-based approaches exploit structural dependencies, propagating verifiable leaf rewards bottom-up to intermediate nodes via sibling normalization or parent-child comparisons (Li et al., 2025c; Yang et al., 2025c; Hou et al., 2025). Furthermore, other methods integrate information-theoretic metrics, utilizing entropy to value pivotal high-uncertainty tokens (Tan et al., 2026) or quantifying information gain (Wang et al., 2025c).

Credit Assignment and Advantage Estimation Beyond the reward signal itself, effective meta-reasoning requires sophisticated credit assignment mechanisms that account for the structure and reliability of the reasoning process. Cheng et al. (2025a) notably employ a min-form strategy to robustly assign credit and prevent estimation error accumulation. While traditional GAE (Schulman et al., 2018) relies on potentially noisy critic models and group-based methods like outcome-supervised GRPO (Shao et al., 2024) assign uniform credit within a response, sampling-based alternatives such as VinePPO (Kazemnejad et al., 2025), SPO (Guo et al., 2025b), BRPO (Qi et al., 2025a) and FR3E (Zheng et al., 2025) aggregate returns from potential futures to offer finer granularity. Moving beyond uniform token treatment, recent works design advantages to explicitly reinforce cognitive structures: Wang et al. (2026) and Li et al. (2026) amplify credit for high-level reasoning anchors, Ding et al. (2025) apply distinct scaling for "thinking" tokens, Cheng et al. (2026) integrate entropy to encourage exploration, and Sun et al. (2025) statistically correlate token occurrences with correct outcomes.

Alternative Learning Paradigms BARL (Zhang et al., 2026c) argues that standard Markovian RL is flawed for exploration since exploratory actions are inherently sub-optimal; it maintains a posterior belief over candidate answers and optimizes a posterior-weighted Q-value, naturally inducing reflection and strategy switching when predictions contradict observations.

5.2 Inference Time

Inference-time interventions actively modulate the reasoning process during generation without altering model weights. While surface-level strategies directly constrain the output by adjusting specific token logits (Wang et al., 2025a) or injecting control tokens into the stream (Muennighoff et al., 2025; Zhang et al., 2025b; Wu et al., 2025a), in this section, we focus on steering, which entails the deep manipulation of internal representations.

5.2.1 Steering and Activation Engineering

Representation Extraction Precise steering of the reasoning process necessitates isolating a vector representing 'reasoning' within the high-dimensional activation space. This is typically achieved through contrastive activation analysis, which captures shifts in internal representations across differential contexts (Højer et al., 2025). The linear representation hypothesis (Park et al., 2024) posits that high-level concepts can be represented as directions in a model's representation space. A common extraction method, **Difference of Means (DoM)** (Rimsky et al., 2024), computes steering vectors by taking the difference between the mean activations of contrastive sample sets, including easy versus hard instances (Sheng et al., 2025), chain-of-thought versus direct answering (Zhang and Viteri, 2025), and the presence of specific cognitive behaviors (Venhoff et al., 2025b; Chen et al., 2025b; Chang et al., 2025). **Probe-based** methods introduce a supervised, discriminative perspective. Such methods train linear probes on specific layers or heads to identify truthfulness-related directions (Li et al., 2023) or cognitive behaviors (Zhang et al., 2025e), determine the correctness of the answer (Cheng et al., 2025c), estimate the pre-allocated reasoning budget (Sheng et al., 2025), and track the progress of reasoning (Eisenstadt et al., 2025). Moving beyond simple averaging, several methods employ Principal Component Analysis (PCA) on the pairwise differences of contrastive samples to extract more robust steering directions (Lin et al., 2025b). To further mitigate noise, approaches like Huang et al. (2025) and Zhang et al. (2025e) project these extracted vectors onto the subspace spanned by the top-k principal components, thereby isolating the dominant axes of variation.

Another class of work utilizes **Sparse Autoencoders (SAEs)** to decompose model activations into a sparse set of interpretable feature directions within the decoder's columns (Galichin et al.,

2026). Baek and Tegmark (2025) confirm in distilled models that there exists a specific set of sparse features that are causally responsible for performing certain types of reasoning. Building on this, Venhoff et al. (2025a) employ small-scale SAEs to cluster thinking trajectories, identifying potential meta-actions in an unsupervised manner that eliminates the need for predefined action categories.

Localization Once the steering direction is extracted, the next challenge is determining the optimal location in the model for intervention. **Layer-wise** analysis by Venhoff et al. (2025b) uses gradient-based attribution patching to estimate the impact of interventions on the output distribution and identify influential layers for steering. At the **component level**, interventions target specific functional modules: Li et al. (2023) utilize linear probes to identify attention heads with high truthfulness discrimination, while Wang et al. (2025d) apply Normalized Pointwise Mutual Information in MoE architectures to isolate cognitive experts that co-occur with thinking tokens. Zhao et al. (2025b) further achieve neuron-level precision, identifying specific significant neurons within MLP layers by contrasting activations.

Intervention Dynamics To ensure precise control without damaging general capabilities, adaptive strategies first determine *when* to intervene using similarity-based gating (Egbuna et al., 2025) or probe-triggered events (Cheng et al., 2025c). Ward et al. (2025) further emphasize locating upstream tokens prior to the backtracking event. Subsequently, intervention intensity is dynamically modulated to match the reasoning state: Lin et al. (2025b) and Jiang et al. (2025b) scale steering magnitude based on real-time difficulty and uncertainty estimates, while to preserve fluency, Xie (2025) utilizes a fixed token-wise decay, contrasting with the fitted local shaping strategy of Zhao et al. (2025b).

6 Evaluation and Benchmarks

Several recent evaluations of large reasoning models have shifted focus from final-answer accuracy to the monitoring, planning, and control of the reasoning process itself. In this section, we first clarify how existing benchmarks evaluate meta-cognitive ability and reasoning-process quality, and then discuss the reliability issues in such evaluations, including judge bias, evaluation-context sensitivity, and test-set leakage.

A first class assesses reasoning-process quality at intermediate steps, providing indirect evidence relevant to meta-reasoning: PRM-style process supervision (Lightman et al., 2024), step-level validity and redundancy (Xia et al., 2025), relevance and coherence (Do et al., 2025), and error-tree evaluation for proofs (Yang et al., 2025b). A second class targets specific cognitive behaviors—error localization (Zeng et al., 2025, 2024), next-action planning (Chen et al., 2025a), possibility-space exploration (Heyman and Zylberberg, 2025), high-level strategy (Ferguson et al., 2025), and adversarial redundancy filtering (Bamgbose et al., 2026)—or whether models understand their own knowledge (Ren et al., 2026; Prato et al., 2024; Barkan et al., 2026).

Since process evaluation often relies on LLM-as-a-judge to reduce human annotation cost, it inherits biases including position, length, verbosity, and egocentric bias (Zheng et al., 2023; Wang et al., 2024a; Koo et al., 2024). Such judgments require validation against appropriate human grounding, particularly when the judging task approaches or exceeds the judge's capabilities (Krumdick et al., 2026). Mitigations include calibrated protocols such as position balancing, evidence-based judging, and human-in-the-loop auditing for difficult cases (Zheng et al., 2023; Wang et al., 2024a; Krumdick et al., 2026). Recent work further reports unexplained judgment variance in Arena-Hard Auto (Feuer et al., 2025), evidence of evaluation awareness (Needham et al., 2025; Nguyen et al., 2025), and behavioral changes under deployment-style prompt rewriting (Xiong et al., 2025), raising concerns about the reliability of evaluation results beyond the tested contexts. Test-set leakage is another concern, addressed through newly authored problems designed to match GSM8K's style and difficulty in GSM1k (Zhang et al., 2024), question decontamination and a non-public test set in MMLU-CF (Zhao et al., 2025a), and dynamic task generation in KUMO (Lin et al., 2025a).

Measuring Over- and Under-Thinking A complementary line of work measures whether a model's inference effort is calibrated to problem difficulty. As a simplified budget-level view, let $B(x)$ denote the budget used (thinking tokens, steps, or branches) and $B^*(x)$ a hypothetical sufficient reference budget for a fixed model, reasoning policy, and success criterion. Then $B(x) > B^*(x)$ and $B(x) < B^*(x)$ indicate excess or shortfall

relative to this reference. Existing work examines reasoning efficiency and progress through output-based indicators, such as token efficiency and answer confidence, and representation-based indicators, such as cyclic latent fluctuations and reasoning-length signals (§3), rather than directly estimating $B^*(x)$. Under-thinking can occur in long reasoning traces when promising thoughts are abandoned prematurely; Wang et al. (2025h) quantify the associated token inefficiency in incorrect responses. Retro-Search (Lu et al., 2025b) revisits thought-switching points to obtain shorter correct trajectories from initially correct responses. Distinguishing insufficient persistence from limited capability therefore requires path-level assessment, not token length alone.

7 Future Directions

While recent advancements have enhanced the meta-reasoning capabilities of LRMs, several open problems remain to be addressed.

From Inference Intervention to Training Guidance Current representation analysis is predominantly limited to test-time interventions, such as using activation patterns as metrics for decision-making or as steering vectors for dynamic control. A promising direction is to translate these inference-time insights into training signals (Vilas et al., 2026), similar to how entropy is utilized for improving credit assignment in RL.

Formalizing and Learning the Meta-Action Space Most existing meta-reasoning frameworks rely on manually defined action spaces (Zhang et al., 2026e; Sui et al., 2026), with few attempting unsupervised clustering (Venhoff et al., 2025a). Future work could explore how to construct a rigorous, complete meta-action space by analyzing the latent structures of reasoning processes. Optimizing the base policy to better support these actions is also important (Qu et al., 2025b; Setlur et al., 2025; Zhang et al., 2026d), as a well-structured initial stage can significantly enhance the efficiency of subsequent post-training optimization.

Scalable Multi-Agent Architectures While multi-agent frameworks enhance reasoning through diverse perspectives, current training paradigms often suffer from inefficiency, typically fine-tuning only a single agent (Estornell et al., 2025) or relying on alternating updates when agents do not share parameters (Wan et al., 2025). Future work could design more efficient collaboration architectures that allow for the joint optimization of heterogeneous agents and reduced communication overhead, ensuring that the collective intelligence scales effectively without diminishing returns.

Robust Process Reward Estimation To reduce reliance on human annotation, process rewards are increasingly estimated via sampling (Kazemnejad et al., 2025; Guo et al., 2025b). However, this introduces a trade-off between estimation accuracy and computational cost. In addition, sampling-based methods often yield zero positive signals for initial weak models. Future methods could design strategies to address this cold start problem. Solutions may involve curriculum-based reward modeling or cross-model supervision to provide dense feedback even in the absence of successful rollouts.

Limitations

One primary limitation of this survey is the depth of integration with foundational cognitive science. While we structure our analysis around the monitor-controller-mechanism framework, our discussion primarily focuses on the computational implementations of these components within Large Reasoning Models. We do not fully exhaust the rich body of literature from cognitive psychology—specifically theories regarding metacognitive regulation, dual-process theory, and epistemic feelings—which could offer deeper explanatory power for why certain architectural choices succeed or fail. Future work could benefit from a more rigorous mapping between these biological cognitive models and the emerging behaviors of artificial reasoning agents, potentially uncovering theoretical upper bounds for machine introspection.

Secondly, the explosive growth of research in Large Reasoning Models presents a challenge to comprehensiveness. Consequently, despite our best efforts to cover the current landscape, it is possible that some important works may have been unintentionally omitted from this review.

Acknowledgments

This research was supported by the AI Laboratory of Xiaomi Communications Co., Ltd.

References

Sohyun An, Ruochen Wang, Tianyi Zhou, and Cho-Jui Hsieh. 2025. Don't think longer, think wisely:

Optimizing thinking dynamics for large reasoning models. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pages 111021–111046. Curran Associates, Inc.

Simon A. Aytes, Jinheon Baek, and Sung Ju Hwang. 2025. Sketch-of-thought: Efficient LLM reasoning with adaptive cognitive-inspired sketching. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 24296–24320, Suzhou, China. Association for Computational Linguistics.

David D. Baek and Max Tegmark. 2025. Towards understanding distilled reasoning models: A representational approach. In *ICLR 2025 Workshop on Building Trust in Language Models and Applications*.

Oluwanifemi Bamgbose, Masoud Hashemi, Sathwik Tejaswi Madhusudhan, Jishnu Sethumadhavan Nair, Aman Tiwari, and Vikas Yadav. 2026. Dnr bench: Benchmarking over-reasoning in reasoning llms. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 37240–37248.

Casey O. Barkan, Sidney Black, and Oliver Sourbut. 2026. Do large language models know what they are capable of? In *The Fourteenth International Conference on Learning Representations*.

Victor Barres, Honghua Dong, Soham Ray, Xujie Si, and Karthik R Narasimhan. 2026. τ^2-bench: Evaluating conversational agents in a dual-control environment. In *Forty-third International Conference on Machine Learning*.

Ahsan Bilal, Muhammad Ahmed Mohsin, Muhammad Umer, Muhammad Awais Khan Bangash, and Muhammad Ali Jamshed. 2025. Meta-thinking in llms via multi-agent reinforcement learning: A survey. *Preprint*, arXiv:2504.14520.

Fu-Chieh Chang, Yu-Ting Lee, and Pei-Yuan Wu. 2025. Unveiling the latent directions of reflection in large language models. In *Mechanistic Interpretability Workshop at NeurIPS 2025*.

Guizhen Chen, Weiwen Xu, Hao Zhang, Hou Pong Chan, Chaoqun Liu, Lidong Bing, Deli Zhao, Anh Tuan Luu, and Yu Rong. 2025a. FineReason: Evaluating and improving LLMs' deliberate reasoning through reflective puzzle solving. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6685–6715, Vienna, Austria. Association for Computational Linguistics.

Qiguang Chen, Libo Qin, Jinhao Liu, Dengyun Peng, Jiannan Guan, Peng Wang, Mengkang Hu, Yuhang Zhou, Te Gao, and Wanxiang Che. 2026. Towards reasoning era: A survey of long chain-of-thought for reasoning large language models. *Science China Information Sciences*, 69(6):161101.

Runjin Chen, Zhenyu Zhang, Junyuan Hong, Souvik Kundu, and Zhangyang Wang. 2025b. SEAL: Steerable reasoning calibration of large language models for free. In *Second Conference on Language Modeling*.

Daixuan Cheng, Shaohan Huang, Xuekai Zhu, Bo Dai, Xin Zhao, Zhenliang Zhang, and Furu Wei. 2026. Reasoning with exploration: An entropy perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 30377–30385.

Jie Cheng, Gang Xiong, Ruixi Qiao, Lijun Li, Chao Guo, Junle Wang, Yisheng Lv, and Fei-Yue Wang. 2025a. Stop summation: Min-form credit assignment is all process reward model needs for reasoning. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pages 131646–131671. Curran Associates, Inc.

Xiaoxue Cheng, Junyi Li, Zhenduo Zhang, Xinyu Tang, Xin Zhao, Xinyu Kong, and Zhiqiang Zhang. 2025b. Incentivizing dual process thinking for efficient large language model reasoning. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pages 152743–152764. Curran Associates, Inc.

Zifeng Cheng, Jinwei Gan, Zhiwei Jiang, Cong Wang, Yafeng Yin, Xiang Luo, Yuchen Fu, and Qing Gu. 2025c. Steering when necessary: Flexible steering large language models with backtracking. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pages 8390–8413. Curran Associates, Inc.

Ganqu Cui, Lifan Yuan, Zefan Wang, Hanbin Wang, Yuchen Zhang, Jiacheng Chen, Wendi Li, Bingxiang He, Yuchen Fan, Tianyu Yu, Qixin Xu, Weize Chen, Jiarui Yuan, Huayu Chen, Kaiyan Zhang, Xingtai Lv, Shuo Wang, Yuan Yao, Xu Han, and 6 others. 2026. Process reinforcement through implicit rewards. *Transactions on Machine Learning Research*.

Aniket Didolkar, Nicolas Ballas, Sanjeev Arora, and Anirudh Goyal. 2025. Metacognitive reuse: Turning recurring llm reasoning into concise behaviors. *Preprint*, arXiv:2509.13237.

Aniket Didolkar, Anirudh Goyal, Nan Rosemary Ke, Siyuan Guo, Michal Valko, Timothy Lillicrap, Danilo Rezende, Yoshua Bengio, Michael Mozer, and Sanjeev Arora. 2024. Metacognitive capabilities of llms: An exploration in mathematical problem solving. In *Advances in Neural Information Processing Systems*, volume 37, pages 19783–19812. Curran Associates, Inc.

Bowen Ding, Yuhan Chen, Futing Wang, Lingfeng Ming, and Tao Lin. 2025. Do thinking tokens help or trap? towards more efficient large reasoning model. *Preprint*, arXiv:2506.23840.

Heejin Do, Jaehui Hwang, Dongyoon Han, Seong Joon Oh, and Sangdoo Yun. 2025. What defines good reasoning in llms? dissecting reasoning steps with multi-aspect evaluation. *Preprint*, arXiv:2510.20603.

Haonan Dong, Haoran Ye, Wenhao Zhu, Kehan Jiang, and Guojie Song. 2025. Meta-r1: Empowering large reasoning models with metacognition. *Preprint*, arXiv:2508.17291.

Nathan Egbuna, Saatvik Gaur, Kevin Zhu, Sunishchal Dev, Ashwinee Panda, and Maheep Chaudhary. 2025. Amortized latent steering: Low-cost alternative to test-time optimization. In *NeurIPS 2025 Workshop on Efficient Reasoning*.

Roy Eisenstadt, Itamar Zimerman, and Lior Wolf. 2025. Overclocking llm reasoning: Monitoring and controlling thinking path lengths in llms. *Preprint*, arXiv:2506.07240.

Andrew Estornell, Jean-Francois Ton, Muhammad Faaiz Taufiq, and Hang Li. 2025. How to train a leader: Hierarchical reasoning in multi-agent llms. *Preprint*, arXiv:2507.08960.

Siqi Fan, Bowen Qin, Peng Han, Shuo Shang, Yequan Wang, and Aixin Sun. 2025. The price of a second thought: On the evaluation of reasoning efficiency in large language models. *Preprint*, arXiv:2505.22017.

Gongfan Fang, Xinyin Ma, and Xinchao Wang. 2025. Thinkless: Llm learns when to think. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pages 151268–151295. Curran Associates, Inc.

Sicheng Feng, Gongfan Fang, Xinyin Ma, and Xinchao Wang. 2025. Efficient reasoning models: A survey. *Transactions on Machine Learning Research*.

Nick Ferguson, Liane Guillou, Alan Bundy, and Kwabena Nuamah. 2025. Evaluating the meta- and object-level reasoning of large language models for question answering. *Preprint*, arXiv:2502.10338.

Benjamin Feuer, Chiung-Yi Tseng, Astitwa Sarthak Lathe, Oussama Elachqar, and John P Dickerson. 2025. When judgment becomes noise: How design failures in llm judge benchmarks silently undermine validity. *Preprint*, arXiv:2509.20293.

John H Flavell. 1979. Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American psychologist*, 34(10):906.

Yichao Fu, Junda Chen, Siqi Zhu, Zheyu Fu, Zhongdongming Dai, Yonghao Zhuang, Yian Ma, Aurick Qiao, Tajana Rosing, Ion Stoica, and Hao Zhang. 2025. Efficiently scaling llm reasoning with certaindex. *Preprint*, arXiv:2412.20993.

Andrey V Galichin, Alexey Dontsov, Polina Druzhinina, Anton Razzhigaev, Oleg Rogov, Elena Tutubalina, and Ivan Oseledets. 2026. I have covered all the bases here: Interpreting reasoning features in large language models via sparse autoencoders. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 30771–30779.

Kanishk Gandhi, Ayush K Chakravarthy, Anikait Singh, Nathan Lile, and Noah Goodman. 2025. Cognitive behaviors that enable self-improving reasoners, or, four habits of highly effective STars. In *Second Conference on Language Modeling*.

Peizhong Gao, Ao Xie, Shaoguang Mao, Wenshan Wu, Yan Xia, Haipeng Mi, and Furu Wei. 2024. Meta reasoning for large language models. *Preprint*, arXiv:2406.11698.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, and 175 others. 2025a. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638.

Yiran Guo, Lijie Xu, Jie Liu, DAN YE, and Shuang Qiu. 2025b. Segment policy optimization: Effective segment-level credit assignment in rl for large language models. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pages 114399–114431. Curran Associates, Inc.

Rui Ha, Rui Pu, Chaozhuo Li, Li Sun, and Sen Su. 2026. From "aha moments" to controllable thinking: Toward meta-cognitive reasoning in LRMs via decoupled reasoning and control. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6698–6710, San Diego, California, United States. Association for Computational Linguistics.

Shibo Hao, Yi Gu, Haodi Ma, Joshua Hong, Zhen Wang, Daisy Wang, and Zhiting Hu. 2023. Reasoning with language model is planning with world model. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8154–8173, Singapore. Association for Computational Linguistics.

Alex Heyman and Joel Zylberberg. 2025. Evaluating the systematic reasoning abilities of large language models through graph coloring. *Preprint*, arXiv:2502.07087.

Bertram Højer, Oliver Simon Jarvis, and Stefan Heinrich. 2025. Improving reasoning performance in large language models via representation engineering. In *The Thirteenth International Conference on Learning Representations*.

Zhenyu Hou, Ziniu Hu, Yujiang Li, Rui Lu, Jie Tang, and Yuxiao Dong. 2025. TreeRL: LLM reinforcement learning with on-policy tree search. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12355–12369, Vienna, Austria. Association for Computational Linguistics.

Shijue Huang, Hongru WANG, Wanjun Zhong, Zhaochen Su, Jiazhan Feng, Bowen Cao, and Yi R. Fung. 2026. Adactrl: Towards adaptive and controllable reasoning via difficulty-aware budgeting. *Transactions on Machine Learning Research*.

Yao Huang, Huanran Chen, Shouwei Ruan, Yichi Zhang, Xingxing Wei, and Yinpeng Dong. 2025. Mitigating overthinking in large reasoning models via manifold steering. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pages 102543–102568. Curran Associates, Inc.

Yue Huang, Jiawen Shi, Yuan Li, Chenrui Fan, Siyuan Wu, Qihui Zhang, Yixin Liu, Pan Zhou, Yao Wan, Neil Gong, and Lichao Sun. 2024. Metatool benchmark for large language models: Deciding whether to use tools and which to use. In *International Conference on Learning Representations*, volume 2024, pages 42978–43007.

Lingjie Jiang, Xun Wu, Shaohan Huang, Qingxiu Dong, Zewen Chi, Li Dong, Xingxing Zhang, Tengchao Lv, Lei Cui, and Furu Wei. 2025a. Think only when you need with large hybrid-reasoning models. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pages 159355–159382. Curran Associates, Inc.

Shuhao Jiang, Songbo Wang, Yang Qiao, Chun Xu, Chaoyang Zheng, Shengyi Zhou, Huanjun Wang, Fangming Li, Cong Zhang, and Jiyu Wang. 2025b. Adaptive overclocking: Dynamic control of thinking path length via real-time reasoning signals. *Preprint*, arXiv:2509.17000.

Ryo Kamoi, Yusen Zhang, Nan Zhang, Jiawei Han, and Rui Zhang. 2024. When can LLMs actually correct their own mistakes? a critical survey of self-correction of LLMs. *Transactions of the Association for Computational Linguistics*, 12:1417–1440.

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. *Preprint*, arXiv:2001.08361.

Amirhossein Kazemnejad, Milad Aghajohari, Eva Portelance, Alessandro Sordoni, Siva Reddy, Aaron Courville, and Nicolas Le Roux. 2025. VinePPO: Refining credit assignment in RL training of LLMs. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 29557–29590. PMLR.

Vedant Khandelwal, Francesca Rossi, Keerthiram Murugesan, Erik Miehling, Murray Campbell, Karthikeyan Natesan Ramamurthy, and Lior Horesh. 2025. Language models coupled with metacognition can outperform reasoning models. *Preprint*, arXiv:2508.17959.

Yoonjeon Kim, Doohyuk Jang, and Eunho Yang. 2026. Verifying meta-awareness via predictive rewards in reasoning models. *Preprint*, arXiv:2510.03259.

Ryan Koo, Minhwa Lee, Vipul Raheja, Jong Inn Park, Zae Myung Kim, and Dongyeop Kang. 2024. Benchmarking cognitive biases in large language models as evaluators. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 517–545, Bangkok, Thailand. Association for Computational Linguistics.

Michael Krumdick, Charles Lovering, Varshini Reddy, Seth Ebner, and Chris Tanner. 2026. No free labels: Limitations of llm-as-a-judge without human grounding. *Preprint*, arXiv:2503.05061.

Aviral Kumar, Vincent Zhuang, Rishabh Agarwal, Yi Su, JD Co-Reyes, Avi Singh, Kate Baumli, Shariq Iqbal, Colton Bishop, Rebecca Roelofs, Lei Zhang, Kay McKinney, Disha Shrivastava, Cosmin Paduraru, George Tucker, Doina Precup, Feryal Behbahani, and Aleksandra Faust. 2025. Training language models to self-correct via reinforcement learning. In *International Conference on Learning Representations*, volume 2025, pages 54523–54549.

Haoxi Li, Sikai Bai, Jie Zhang, and Song Guo. 2025a. Core: Enhancing metacognition with label-free self-evaluation in lrms. *Preprint*, arXiv:2507.06087.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems*, volume 36, pages 41451–41530. Curran Associates, Inc.

Wenjun Li, Dexun Li, Kuicai Dong, Cong Zhang, Hao Zhang, Weiwen Liu, Yasheng Wang, Ruiming Tang, and Yong Liu. 2025b. Adaptive tool use in large language models with meta-cognition trigger. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13346–13370, Vienna, Austria. Association for Computational Linguistics.

Yang Li, Zhichen Dong, Yuhan Sun, Weixun Wang, Shaopan Xiong, Yijia Luo, Jiashun Liu, Han Lu, Jiamang Wang, Wenbo Su, Bo Zheng, and Junchi Yan. 2026. Attention illuminates llm reasoning: The preplan-and-anchor rhythm enables fine-grained policy optimization. *Preprint*, arXiv:2510.13554.

Yizhi Li, Qingshui Gu, Zhoufutu Wen, Ziniu Li, Tianshun Xing, Shuyue Guo, Tianyu Zheng, Xin Zhou, Xingwei Qu, Wangchunshu Zhou, Zheng Zhang, Wei Shen, Qian Liu, Chenghua Lin, Jian Yang, Ge Zhang, and Wenhao Huang. 2025c. Treepo: Bridging the gap of policy optimization and efficacy and inference efficiency with heuristic tree-based modeling. *Preprint*, arXiv:2508.17445.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2024. Let's verify step by step. In *International Conference on Learning Representations*, volume 2024, pages 39578–39601.

Haowei Lin, Xiangyu Wang, Ruilin Yan, Baizhou Huang, Haotian Ye, Jianhua Zhu, Zihao Wang, James

Zou, Jianzhu Ma, and Yitao Liang. 2025a. Generative evaluation of complex reasoning in large language models. *Preprint*, arXiv:2504.02810.

Zhengkai Lin, Zhihang Fu, Ze Chen, Chao Chen, Liang Xie, Wenxiao Wang, Deng Cai, Zheng Wang, and Jieping Ye. 2025b. Controlling thinking speed in reasoning models. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pages 78300–78347. Curran Associates, Inc.

Qin Liu, Wenxuan Zhou, Nan Xu, James Y. Huang, Fei Wang, Sheng Zhang, Hoifung Poon, and Muhao Chen. 2026a. MetaScale: Test-time scaling with evolving meta-thoughts. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 11828–11842, San Diego, California, United States. Association for Computational Linguistics.

Qiyuan Liu, Hao Xu, Xuhong Chen, Wei Chen, Yee Whye Teh, and Ning Miao. 2026b. Enhancing large language model reasoning with reward models: An analytical survey. *Preprint*, arXiv:2510.01925.

Tengxiao Liu, Qipeng Guo, Yuqing Yang, Xiangkun Hu, Yue Zhang, Xipeng Qiu, and Zheng Zhang. 2023. Plan, verify and switch: Integrated reasoning with diverse X-of-thoughts. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2807–2822, Singapore. Association for Computational Linguistics.

Xiaoou Liu, Tiejin Chen, Longchao Da, Chacha Chen, Zhen Lin, and Hua Wei. 2025. Uncertainty quantification and confidence calibration in large language models: A survey. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, KDD '25, page 6107–6117, New York, NY, USA. Association for Computing Machinery.

Yue Liu, Jiaying Wu, Yufei He, Ruihan Gong, Jun Xia, Liang Li, Hongcheng Gao, Hongyu Chen, Baolong Bi, Jiaheng Zhang, Zhiqi Huang, Bryan Hooi, Stan Z. Li, and Keqin Li. 2026c. Efficient inference for large reasoning models: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(10):12694–12712.

Jinghui Lu, Haiyang Yu, Siliang Xu, Shiwei Ran, Guozhi Tang, Siqi Wang, Bin Shan, Teng Fu, Hao Feng, Jingqun Tang, Han Wang, and Can Huang. 2025a. Prolonged reasoning is not all you need: Certainty-based adaptive routing for efficient llm/mllm reasoning. *Preprint*, arXiv:2505.15154.

Ximing Lu, Seungju Han, David Acuna, Hyunwoo Kim, Jaehun Jung, Shrimai Prabhumoye, Niklas Muennighoff, Mostofa Patwary, Mohammad Shoeybi, Bryan Catanzaro, and Yejin Choi. 2025b. Retrosearch: Exploring untaken paths for deeper and efficient reasoning. *Preprint*, arXiv:2504.04383.

Haotian Luo, Haiying He, Yibo Wang, Jinluan Yang, Rui Liu, Naiqiang Tan, Xiaochun Cao, Dacheng Tao, and Li Shen. 2025. Ada-r1: Hybrid-cot via bi-level adaptive reasoning optimization. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pages 59353–59377. Curran Associates, Inc.

Grégoire Mialon, Clémentine Fourrier, Thomas Wolf, Yann LeCun, and Thomas Scialom. 2024. Gaia: a benchmark for general ai assistants. In *International Conference on Learning Representations*, volume 2024, pages 9025–9049.

Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. 2025. Large language models: A survey. *Preprint*, arXiv:2402.06196.

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 *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 20275–20321, Suzhou, China. Association for Computational Linguistics.

Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. 2025. A comprehensive overview of large language models. *ACM Trans. Intell. Syst. Technol.*, 16(5).

Sania Nayab, Giulio Rossolini, Marco Simoni, Andrea Saracino, Giorgio Buttazzo, Nicolamaria Manes, and Fabrizio Giacomelli. 2026. Concise thoughts: Impact of output length on llm reasoning and cost. *Information Sciences*, 757:123939.

Joe Needham, Giles Edkins, Govind Pimpale, Henning Bartsch, and Marius Hobbhahn. 2025. Large language models often know when they are being evaluated. *Preprint*, arXiv:2505.23836.

Thomas O. Nelson. 1990. Metamemory: A theoretical framework and new findings. volume 26 of *Psychology of Learning and Motivation*, pages 125–173. Academic Press.

Jord Nguyen, Hoang Huu Khiem, Carlo Leonardo Attubato, and Felix Hofstätter. 2025. Probing evaluation awareness of language models. In *ICML Workshop on Technical AI Governance (TAIG)*.

Kangyun Ning, Yisong Su, Xueqiang Lv, Yuanzhe Zhang, Jian Liu, Kang Liu, and Jinan Xu. 2024a. Wtu-eval: A whether-or-not tool usage evaluation benchmark for large language models. *Preprint*, arXiv:2407.12823.

Xuefei Ning, Zinan Lin, Zixuan Zhou, Zifu Wang, Huazhong Yang, and Yu Wang. 2024b. Skeleton-of-thought: Prompting llms for efficient parallel generation. In *International Conference on Learning Representations*, volume 2024, pages 917–967.

Yansong Ning, Wei Li, Jun Fang, Naiqiang Tan, and Hao Liu. 2025. Not all thoughts are generated equal: Efficient llm reasoning via multi-turn reinforcement learning. *Preprint*, arXiv:2505.11827.

OpenAI, :, Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, Alex Iftimie, Alex Karpenko, Alex Tachard Passos, Alexander Neitz, Alexander Prokofiev, Alexander Wei, Allison Tam, and 246 others. 2026. Openai o1 system card. *Preprint*, arXiv:2412.16720.

Zhihong Pan, Kai Zhang, Yuze Zhao, and Yupeng Han. 2026. Adaptive model and strategy routing for cost-efficient llm services. In *Proceedings of the ACM Web Conference 2026*, page 5568–5578. ACM.

Shubham Parashar, Blake Olson, Sambhav Khurana, Eric Li, Hongyi Ling, James Caverlee, and Shuiwang Ji. 2025. Inference-time computations for llm reasoning and planning: A benchmark and insights. *Preprint*, arXiv:2502.12521.

Kiho Park, Yo Joong Choe, and Victor Veitch. 2024. The linear representation hypothesis and the geometry of large language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 39643–39666. PMLR.

Gabriele Prato, Jerry Huang, Prasanna Parthasarathi, Shagun Sodhani, and Sarath Chandar. 2024. Do large language models know how much they know? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6054–6070, Miami, Florida, USA. Association for Computational Linguistics.

Penghui Qi, Zichen Liu, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. 2025a. Optimizing anytime reasoning via budget relative policy optimization. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pages 23429–23451. Curran Associates, Inc.

Zhenting Qi, Mingyuan MA, Jiahang Xu, Li Lyna Zhang, Fan Yang, and Mao Yang. 2025b. Mutual reasoning makes smaller llms stronger problem-solver. In *International Conference on Learning Representations*, volume 2025, pages 20788–20807.

Chen Qian, Dongrui Liu, Haochen Wen, Zhen Bai, Yong Liu, and Jing Shao. 2025a. Demystifying reasoning dynamics with mutual information: Thinking tokens are information peaks in llm reasoning. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pages 12533–12572. Curran Associates, Inc.

Cheng Qian, Emre Can Acikgoz, Hongru Wang, Xiusi Chen, Avirup Sil, Dilek Hakkani-Tür, Gokhan Tur, and Heng Ji. 2025b. SMART: Self-aware agent for tool overuse mitigation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 4604–4621, Vienna, Austria. Association for Computational Linguistics.

Ziqing Qiao, Yongheng Deng, Jiali Zeng, Dong Wang, Lai Wei, Guanbo Wang, Fandong Meng, Jie Zhou, Ju Ren, and Yaoxue Zhang. 2025. ConCISE: Confidence-guided compression in step-by-step efficient reasoning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 8010–8029, Suzhou, China. Association for Computational Linguistics.

Xiaoye Qu, Yafu Li, Zhao-Chen Su, Weigao Sun, Jianhao Yan, Dongrui Liu, Ganqu Cui, Daizong Liu, Shuxian Liang, Junxian He, Peng Li, Wei Wei, Jing Shao, Chaochao Lu, Yue Zhang, Xian-Sheng Hua, Bowen Zhou, and Yu Cheng. 2025a. A survey of efficient reasoning for large reasoning models: Language, multimodality, and beyond. *Preprint*, arXiv:2503.21614.

Yuxiao Qu, Matthew Y. R. Yang, Amrith Setlur, Lewis Tunstall, Edward Emanuel Beeching, Ruslan Salakhutdinov, and Aviral Kumar. 2025b. Optimizing test-time compute via meta reinforcement fine-tuning. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 50893–50925. PMLR.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, pages 53728–53741. Curran Associates, Inc.

Jingyi Ren, Ante Wang, Yunghwei Lai, Xiaolong Wang, Linlu Gong, Weitao Li, Weizhi Ma, and Yang Liu. 2026. Beyond "I don't know": Evaluating LLM self-awareness in discriminating data and model uncertainty. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11911–11929, San Diego, California, United States. Association for Computational Linguistics.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. 2024. Steering llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15504–15522, Bangkok, Thailand. Association for Computational Linguistics.

Hayley Ross, Ameya Sunil Mahabaleshwarkar, and Yoshi Suhara. 2025. When2Call: When (not) to call tools. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3391–3409, Albuquerque, New Mexico. Association for Computational Linguistics.

C. Nicolò De Sabbata, Theodore Sumers, and Thomas L. Griffiths. 2024. Rational metareasoning for large language models. In *NeurIPS 2024 Workshop on Behavioral Machine Learning*.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessi, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. In *Advances in Neural Information Processing Systems*, volume 36, pages 68539–68551. Curran Associates, Inc.

John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. 2018. High-dimensional continuous control using generalized advantage estimation. *Preprint*, arXiv:1506.02438.

Amrith Setlur, Chirag Nagpal, Adam Fisch, Xinyang Geng, Jacob Eisenstein, Rishabh Agarwal, Alekh Agarwal, Jonathan Berant, and Aviral Kumar. 2025. Rewarding progress: Scaling automated process verifiers for llm reasoning. In *International Conference on Learning Representations*, volume 2025, pages 60808–60838.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *Preprint*, arXiv:2402.03300.

Maohao Shen, Guangtao Zeng, Zhenting Qi, Zhang-Wei Hong, Zhenfang Chen, Wei Lu, Gregory W. Wornell, Subhro Das, David Daniel Cox, and Chuang Gan. 2025a. Satori: Reinforcement learning with chain-of-action-thought enhances LLM reasoning via autoregressive search. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 54663–54700. PMLR.

Yi Shen, Jian Zhang, Jieyun Huang, Shuming Shi, Wenjing Zhang, Jiangze Yan, Ning Wang, Kai Wang, Zhaoxiang Liu, and Shiguo Lian. 2025b. DAST: Difficulty-adaptive slow-thinking for large reasoning models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 2322–2331, Suzhou (China). Association for Computational Linguistics.

Leheng Sheng, An Zhang, Zijian Wu, Weixiang Zhao, Changshuo Shen, zhang yi, Xiang Wang, and Tat-Seng Chua. 2025. On reasoning strength planning in large reasoning models. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pages 62978–63012. Curran Associates, Inc.

Dachuan Shi, Abedelkadir Asi, Keying Li, Xiangchi Yuan, Leyan Pan, Wenke Lee, and Wen Xiao. 2026. Swireasoning: Switch-thinking in latent and explicit for pareto-superior reasoning llms. In *International Conference on Learning Representations*, volume 2026, pages 137060–137093.

Ola Shorinwa, Zhiting Mei, Justin Lidard, Allen Z. Ren, and Anirudha Majumdar. 2025. A survey on uncertainty quantification of large language models: Taxonomy, open research challenges, and future directions. *ACM Comput. Surv.*, 58(3).

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2024. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *Preprint*, arXiv:2408.03314.

Yang Sui, Yu-Neng Chuang, Guanchu Wang, Jiamu Zhang, Tianyi Zhang, Jiayi Yuan, Hongyi Liu, Andrew Wen, Shaochen Zhong, Na Zou, Hanjie Chen, and Xia Hu. 2025. Stop overthinking: A survey on efficient reasoning for large language models. *Transactions on Machine Learning Research*.

Yuan Sui, Yufei He, Tri Cao, Sophia Simeng Han, Yulin Chen, and Bryan Hooi. 2026. Meta-reasoner: Dynamic guidance for optimized inference-time reasoning in large language models. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 13268–13286, San Diego, California, United States. Association for Computational Linguistics.

Wei Sun, Wen Yang, Pu Jian, Qianlong Du, Fuwei Cui, Shuo Ren, and Jiajun Zhang. 2025. Ktae: A model-free algorithm to key-tokens advantage estimation in mathematical reasoning. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pages 161880–161908. Curran Associates, Inc.

Hongze Tan, Zihan Wang, Jianfei Pan, Jinghao Lin, Hao Wang, Yifan Wu, Tao Chen, Zhihang Zheng, Zhihao Tang, and Haihua Yang. 2026. Gtpo and grpo-s: Token and sequence-level reward shaping with policy entropy. *Preprint*, arXiv:2508.04349.

Yuxuan Tong, Xiwen Zhang, Rui Wang, Ruidong Wu, and Junxian He. 2024. Dart-math: Difficulty-aware rejection tuning for mathematical problem-solving. *Preprint*, arXiv:2407.13690.

Constantin Venhoff, Iván Arcuschin, Philip Torr, Arthur Conmy, and Neel Nanda. 2025a. Base models know how to reason, thinking models learn when. In *Mechanistic Interpretability Workshop at NeurIPS 2025*.

Constantin Venhoff, Iván Arcuschin, Philip Torr, Arthur Conmy, and Neel Nanda. 2025b. Understanding reasoning in thinking language models via steering vectors. *Preprint*, arXiv:2506.18167.

Martina G. Vilas, Safoora Yousefi, Besmira Nushi, Eric Horvitz, and Vidhisha Balachandran. 2026. Tracing the traces: Latent temporal signals for efficient and accurate reasoning. In *International Conference on Learning Representations*, volume 2026, pages 1776–1807.

Ziyu Wan, Yunxiang Li, Xiaoyu Wen, Yan Song, Hanjing Wang, Linyi Yang, Mark Schmidt, Jun Wang, Weinan Zhang, Shuyue Hu, and Ying Wen. 2025. Rema: Learning to meta-think for llms with multi-agent reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pages 126621–126667. Curran Associates, Inc.

Chenlong Wang, Yuanning Feng, Dongping Chen, Zhaoyang Chu, Ranjay Krishna, and Tianyi Zhou. 2025a. Wait, we don't need to "wait"! removing thinking tokens improves reasoning efficiency. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 7459–7482, Suzhou, China. Association for Computational Linguistics.

Haozhe Wang, Qixin Xu, Che Liu, Junhong Wu, Fangzhen Lin, and Wenhu Chen. 2026. Emergent hierarchical reasoning in llms through reinforcement learning. In *International Conference on Learning Representations*, volume 2026, pages 74519–74543.

Hongru Wang, Cheng Qian, Wanjun Zhong, Xiusi Chen, Jiahao Qiu, Shijue Huang, Bowen Jin, Mengdi Wang, Kam-Fai Wong, and Heng Ji. 2025b. Acting less is reasoning more! teaching model to act efficiently. *Preprint*, arXiv:2504.14870.

Jingyao Wang, Wenwen Qiang, Zeen Song, Changwen Zheng, and Hui Xiong. 2025c. Learning to think: Information-theoretic reinforcement fine-tuning for llms. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pages 126208–126245. Curran Associates, Inc.

Mengru Wang, Xingyu Chen, Yue Wang, Zhiwei He, Jiahao Xu, Tian Liang, Qiuzhi Liu, Yunzhi Yao, Wenxuan Wang, Ruotian Ma, Haitao Mi, Ningyu Zhang, Zhaopeng Tu, Xiaolong Li, and Dong Yu. 2025d. Two experts are all you need for steering thinking: Reinforcing cognitive effort in moe reasoning models without additional training. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pages 140699–140724. Curran Associates, Inc.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. 2024a. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9440–9450, Bangkok, Thailand. Association for Computational Linguistics.

Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. 2024b. Math-shepherd: Verify and reinforce LLMs step-by-step without human annotations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9426–9439, Bangkok, Thailand. Association for Computational Linguistics.

Pinzheng Wang, Juntao Li, Zecheng Tang, Haijia Gui, and Min Zhang. 2025e. Improving rationality in the reasoning process of language models through self-playing game. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 63417–63438. PMLR.

Wenxuan Wang, Shi Juluan, Zixuan Ling, Yuk-Kit Chan, Chaozheng Wang, Cheryl Lee, Youliang Yuan, Jen-tse Huang, Wenxiang Jiao, and Michael R. Lyu. 2025f. Learning to ask: When LLM agents meet unclear instruction. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 21773–21784, Suzhou, China. Association for Computational Linguistics.

Xiangqi Wang, Yue Huang, Yanbo Wang, Xiaonan Luo, Kehan Guo, Yujun Zhou, and Xiangliang Zhang. 2025g. Adareasoner: Adaptive reasoning enables more flexible thinking. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pages 75737–75764. Curran Associates, Inc.

Yu Wang, Shiwan Zhao, Zhihu Wang, Heyuan Huang, Ming Fan, Yubo Zhang, Zhixing Wang, Haijun Wang, and Ting Liu. 2024c. Strategic chain-of-thought: Guiding accurate reasoning in llms through strategy elicitation. *Preprint*, arXiv:2409.03271.

Yue Wang, Qiuzhi 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. 2025h. Thoughts are all over the place: On the underthinking of o1-like llms. *Preprint*, arXiv:2501.18585.

Jake Ward, Chuqiao Lin, Constantin Venhoff, and Neel Nanda. 2025. Reasoning-finetuning repurposes latent representations in base models. *Preprint*, arXiv:2507.12638.

Jason Wei, Zhiqing Sun, Spencer Papay, Scott McKinney, Jeffrey Han, Isa Fulford, Hyung Won Chung, Alex Tachard Passos, William Fedus, and Amelia Glaese. 2025. Browsecomp: A simple yet challenging benchmark for browsing agents. *Preprint*, arXiv:2504.12516.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Jinyang Wu, Chonghua Liao, Mingkuan Feng, Shuai Zhang, Zhengqi Wen, Haoran Luo, Ling Yang, Huazhe Xu, and Jianhua Tao. 2026. Templaterl: Structured template-guided reinforcement learning for llm reasoning. *Preprint*, arXiv:2505.15692.

Tong Wu, Chong Xiang, Jiachen T. Wang, G. Edward Suh, and Prateek Mittal. 2025a. Effectively controlling reasoning models through thinking intervention. In *Socially Responsible and Trustworthy Foundation Models at NeurIPS 2025*.

Xingyu Wu, Yuchen Yan, Shangke Lyu, Linjuan Wu, Yiwen Qiu, Yongliang Shen, Weiming Lu, Jian Shao, Jun Xiao, and Yueting Zhuang. 2025b. Lapo: Internalizing reasoning efficiency via length-adaptive policy optimization. *Preprint*, arXiv:2507.15758.

Shijie Xia, Xuefeng Li, Yixin Liu, Tongshuang Wu, and Pengfei Liu. 2025. Evaluating mathematical reasoning beyond accuracy. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 27723–27730.

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. *Preprint*, arXiv:2501.04682.

Jiaqing Xie. 2025. A comparative analysis of sparse autoencoder and activation difference in language model steering. *Preprint*, arXiv:2510.01246.

Lang Xiong, Nishant Bhargava, Jeremy Chang, Jianhang Hong, Haihao Liu, Vasu Sharma, and Kevin Zhu. 2025. Probe-rewrite-evaluate: A workflow for reliable benchmarks and quantifying evaluation awareness. In *NeurIPS 2025 Workshop on Evaluating the Evolving LLM Lifecycle: Benchmarks, Emergent Abilities, and Scaling*.

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

Weikai Xu, Chengrui Huang, Shen Gao, and Shuo Shang. 2025b. Llm-based agents for tool learning: A survey. *Data Science and Engineering*, 10(4):533–563.

Yuhui Xu, Hanze Dong, Lei Wang, Doyen Sahoo, Junnan Li, and Caiming Xiong. 2026. Scalable chain of thoughts via elastic reasoning. In *International Conference on Learning Representations*, volume 2026, pages 145390–145412.

Hanqi Yan, Linhai Zhang, Jiazheng Li, Zhenyi Shen, and Yulan He. 2025. Position: LLMs need a Bayesian meta-reasoning framework for more robust and generalizable reasoning. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 82360–82383. PMLR.

Chenxu Yang, Qingyi Si, Yongjie Duan, Zheliang Zhu, Chenyu Zhu, Qiaowei Li, Minghui Chen, Zheng Lin, and Weipinng Wang. 2026a. Dynamic early exit in reasoning models. In *International Conference on Learning Representations*, volume 2026, pages 88170–88210.

Kaisen Yang, Tinghe Zhang, Rushi Shah, Kaicheng Yang, Qinwei Ma, Dianbo Liu, and Alex Lamb. 2026b. Explore-execute chain: Towards an efficient structured reasoning paradigm. *Preprint*, arXiv:2509.23946.

Ling Yang, Zhaochen Yu, Bin Cui, and Mengdi Wang. 2025a. Reasonflux: Hierarchical llm reasoning via scaling thought templates. *Preprint*, arXiv:2502.06772.

Ling Yang, Zhaochen Yu, Tianjun Zhang, Shiyi Cao, Minkai Xu, Wentao Zhang, Joseph E. Gonzalez, and Bin Cui. 2024. Buffer of thoughts: Thought-augmented reasoning with large language models. In *Advances in Neural Information Processing Systems*, volume 37, pages 113519–113544. Curran Associates, Inc.

Shu-Xun Yang, Cunxiang Wang, Yidong Wang, Xiaotao Gu, Minlie Huang, and Jie Tang. 2025b. Stepmathagent: A step-wise agent for evaluating mathematical processes through tree-of-error. *Preprint*, arXiv:2503.10105.

Wei Yang and Jesse Thomason. 2026. Learning to deliberate: Meta-policy collaboration for agentic llms with multi-agent reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 29820–29828.

Zhicheng Yang, Zhijiang Guo, Yinya Huang, Xiaodan Liang, Yiwei Wang, and Jing Tang. 2025c. Treerpo: Tree relative policy optimization. *Preprint*, arXiv:2506.05183.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations*.

Xixian Yong, Xiao Zhou, Yingying Zhang, Jinlin Li, Yefeng Zheng, and Xian Wu. 2025. Think or not? exploring thinking efficiency in large reasoning models via an information-theoretic lens. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pages 2787–2827. Curran Associates, Inc.

Eunseop Yoon, Hee Suk Yoon, Jaehyun Jang, SooHwan Eom, Qi Dai, Chong Luo, Mark A. Hasegawa-Johnson, and Chang D. Yoo. 2025. Pacr: Progressively ascending confidence reward for llm reasoning. *Preprint*, arXiv:2510.22255.

Lifan Yuan, Wendi Li, Huayu Chen, Ganqu Cui, Ning Ding, Kaiyan Zhang, Bowen Zhou, Zhiyuan Liu, and Hao Peng. 2025. Free process rewards without process labels. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 73511–73525. PMLR.

Chuhuai Yue, Chengqi Dong, Yinan Gao, Hang He, Jiajun Chai, Wei Lin, and Guojun Yin. 2026. Promoting efficient reasoning with verifiable stepwise reward. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 34530–34538.

Linan Yue, Yichao Du, Yizhi Wang, Weibo Gao, Fangzhou Yao, Li Wang, Ye Liu, Ziyu Xu, Qi Liu, Shimin Di, and Min-Ling Zhang. 2025a. Don't overthink it: A survey of efficient r1-style large reasoning models. *Preprint*, arXiv:2508.02120.

Murong Yue, Wenlin Yao, Haitao Mi, Dian Yu, Ziyu Yao, and Dong Yu. 2025b. Dots: Learning to reason dynamically in llms via optimal reasoning trajectories search. In *International Conference on Learning Representations*, volume 2025, pages 37976–37997.

Zhongshen Zeng, Pengguang Chen, Shu Liu, Haiyun Jiang, and Jiaya Jia. 2025. Mr-gsm8k: A meta-reasoning benchmark for large language model evaluation. In *International Conference on Learning Representations*, volume 2025, pages 101693–101714.

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 *Advances in Neural Information Processing Systems*, volume 37, pages 119466–119546. Curran Associates, Inc.

Anqi Zhang, Yulin Chen, Jane Pan, Chen Zhao, Aurojit Panda, Jinyang Li, and He He. 2025a. Reasoning models know when they're right: Probing hidden states for self-verification. In *Second Conference on Language Modeling*.

Duzhen Zhang, Zhong-Zhi Li, Ming-Liang Zhang, Jiaxin Zhang, Zengyan Liu, Yuxuan Yao, Haotian Xu, Junhao Zheng, Xiuyi Chen, Yingying Zhang, Fei Yin, Jiahua Dong, Zhijiang Guo, Le Song, and Cheng-Lin Liu. 2026a. From system 1 to system 2: A survey of reasoning large language models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(3):3335–3354.

Hengyuan Zhang, Zhihao Zhang, Mingyang Wang, Zunhai Su, Yiwei Wang, Qianli Wang, Shuzhou Yuan, Ercong Nie, Xufeng Duan, Qibo Xue, Zeping Yu, Chenming Shang, Xiao Liang, Jing Xiong, Hui Shen, Chaofan Tao, Zhengwu Liu, Senjie Jin, Zhiheng Xi, and 9 others. 2026b. Locate, steer, and improve: A practical survey of actionable mechanistic interpretability in large language models. *Computer Science Review*, 62:101011.

Hugh Zhang, Jeff Da, Dean Lee, Vaughn Robinson, Catherine Wu, Will Song, Tiffany Zhao, Pranav Raja, Charlotte Zhuang, Dylan Slack, Qin Lyu, Sean Hendryx, Russell Kaplan, Michele Lunati, and Summer Yue. 2024. A careful examination of large language model performance on grade school arithmetic. In *Advances in Neural Information Processing Systems*, volume 37, pages 46819–46836. Curran Associates, Inc.

Jason Zhang and Scott Viteri. 2025. Uncovering latent chain of thought vectors in language models. *Preprint*, arXiv:2409.14026.

Junyu Zhang, Runpei Dong, Han Wang, Xuying Ning, Haoran Geng, Peihao Li, Xialin He, Yutong Bai, Jitendra Malik, Saurabh Gupta, and Huan Zhang. 2025b. AlphaOne: Reasoning models thinking slow and fast at test time. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 11329–11354, Suzhou, China. Association for Computational Linguistics.

Kaiyan Zhang, Yuxin Zuo, Bingxiang He, Youbang Sun, Runze Liu, Che Jiang, Yuchen Fan, Kai Tian, Guoli Jia, Pengfei Li, Yu Fu, Xingtai Lv, Yuchen Zhang, Sihang Zeng, Shang Qu, Haozhan Li, Shijie Wang, Yuru Wang, Xinwei Long, and 20 others. 2025c. A survey of reinforcement learning for large reasoning models. *Preprint*, arXiv:2509.08827.

Shenao Zhang, Yaqing Wang, Yinxiao Liu, Tianqi Liu, Peter Grabowski, Eugene Ie, Zhaoran Wang, and Yunxuan Li. 2026c. Beyond markovian: Reflective exploration via bayes-adaptive rl for llm reasoning. In *International Conference on Learning Representations*, volume 2026, pages 120275–120294.

Shenao Zhang, Donghan Yu, Yihao Feng, Bowen Jin, Zhaoran Wang, John Peebles, and Zirui Wang. 2026d. Learning to reason as action abstractions with scalable mid-training rl. In *International Conference on Learning Representations*, volume 2026, pages 35131–35146.

YuJian Zhang, Keyu Chen, Zhifeng Shen, Ruizhi Qiao, and Xing Sun. 2025d. Adaptive dual reasoner: Large reasoning models can think efficiently by hybrid reasoning. In *NeurIPS 2025 Workshop on Efficient Reasoning*.

Zhenyu Zhang, Xiaoxia Wu, Zhongzhu Zhou, Qingyang Wu, Yineng Zhang, Pragaash Ponnusamy, Harikaran Subbaraj, Jue WANG, Shuaiwen Leon Song, and Ben Athiwaratkun. 2025e. Understanding and steering the cognitive behaviors of reasoning models at test-time. In *NeurIPS 2025 Workshop on Efficient Reasoning*.

Zijing Zhang, Ziyang Chen, Mingxiao Li, Zhaopeng Tu, and Xiaolong Li. 2026e. Rlvmr: Reinforcement learning with verifiable meta-reasoning rewards for robust long-horizon agents. In *International Conference on Learning Representations*, volume 2026, pages 114676–114700.

Ziying Zhang, Yaqing Wang, and Quanming Yao. 2026f. Searching meta reasoning skeleton to guide llm reasoning. *Preprint*, arXiv:2510.04116.

Qihao Zhao, Yangyu Huang, Tengchao Lv, Lei Cui, Qinzheng Sun, Shaoguang Mao, Xin Zhang, Ying Xin, Qiufeng Yin, Scarlett Li, and Furu Wei. 2025a. MMLU-CF: A contamination-free multi-task language understanding benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13371–13391, Vienna, Austria. Association for Computational Linguistics.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Zican Dong, Yupeng Hou, Beichen Zhang, Yingqian Min, Junjie Zhang, Peiyu Liu, Xiaolei Wang, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao

Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, and 4 others. 2026. A survey of large language models. *Front. Comput. Sci.*, 20(12).

Zekai Zhao, Qi Liu, Kun Zhou, Zihan Liu, Yifei Shao, Zhiting Hu, and Biwei Huang. 2025b. Activation control for efficiently eliciting long chain-of-thought ability of language models. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pages 17470–17494. Curran Associates, Inc.

Congmin Zheng, Jiachen Zhu, Zhuoying Ou, Yuxiang Chen, Kangning Zhang, Rong Shan, Zeyu Zheng, Mengyue Yang, Jianghao Lin, Yong Yu, and Weinan Zhang. 2026. A comprehensive survey of process reward models: Data generation, model construction, and usage. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3591–3607, San Diego, California, United States. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623. Curran Associates, Inc.

Tianyu Zheng, Tianshun Xing, Qingshui Gu, Taoran Liang, Xingwei Qu, Xin Zhou, Yizhi Li, Zhoufutu Wen, Chenghua Lin, Wenhao Huang, Qian Liu, Ge Zhang, and Zejun Ma. 2025. First return, entropy-eliciting explore. *Preprint*, arXiv:2507.07017.

Pei Zhou, Jay Pujara, Xiang Ren, Xinyun Chen, Heng-Tze Cheng, Quoc V. Le, Ed H., Denny Zhou, Swaroop Mishra, and Huaixiu Steven Zheng. 2024. Self-discover: Large language models self-compose reasoning structures. In *Advances in Neural Information Processing Systems*, volume 37, pages 126032–126058. Curran Associates, Inc.

Jason Zhu and Hongyu Li. 2025. Towards concise and adaptive thinking in large reasoning models: A survey. *Preprint*, arXiv:2507.09662.

Jiaru Zou, Ling Yang, Jingwen Gu, Jiahao Qiu, Ke Shen, Jingrui He, and Mengdi Wang. 2025. Reasonflux-prm: Trajectory-aware prms for long chain-of-thought reasoning in llms. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pages 26764–26802. Curran Associates, Inc.

A Potential Risk

While this survey primarily focuses on foundational mechanisms, the transition to self-regulating meta-reasoning introduces several critical risks:

Dual-Use and Malicious Planning While meta-reasoning is beneficial for complex problem-solving, it introduces severe dual-use risks. Systems capable of robust strategic planning and autonomous error correction could be exploited by malicious actors in high-stakes domains such as law and medicine to orchestrate sophisticated legal fraud, generate highly plausible but fabricated clinical evidence, or automate the exploitation of regulatory loopholes.

Alignment and Steering Vulnerabilities Intervening at inference time through activation engineering and steering vectors (§5.2.1) provides deep, mechanistic control over a model's internal representations. While these techniques offer promising pathways for ensuring AI alignment and promoting meta-reasoning, they simultaneously act as a vulnerability. If exploited, adversarial steering could be used to bypass safety guardrails, hijack the controller, and force the model into deceptive, harmful, or inherently biased reasoning patterns.

Environmental and Computational Costs The paradigm shift toward LRMs heavily relies on test-time scaling to improve capabilities by allocating additional computational resources during inference. Generating extensive reasoning steps and relying on dense process reward optimization (§5.1.1) may significantly increase the computational costs of these models. Without targeted efficiency improvements, the training process and deployment of reasoning models pose environmental sustainability concerns.

B Statement of LLM Usage

We use Gemini3 Pro to check for grammatical errors and polish our writing.

C Tool Use as a Meta-Action

As agentic LRMs become increasingly embedded in everyday workflows, tool invocation serves as a core interface between reasoning and real-world action, as reflected in general agent benchmarks (Mialon et al., 2024), dual-control settings (Barres et al., 2026), and browsing-focused benchmarks (Wei et al., 2025). We connect this meta-action to three lines of work referenced in §4.

Foundational reason–act formulations ReAct (Yao et al., 2023) explicitly interleaves verbal reasoning traces with text actions, allowing the model

to induce, track, and revise plans while incorporating external observations. Toolformer (Schick et al., 2023) pushes this to the parameter level, training the language model in a self-supervised manner to decide *which* APIs to call, *when* to call them, *what* arguments to pass, and *how* to incorporate the returned results into subsequent generation.

Evaluating the decision boundary Whether to call a tool is itself a meta-decision that benchmarks now isolate. MetaTool (Huang et al., 2024) probes tool-usage awareness and selection across single- and multi-tool scenarios. WTU-Eval (Ning et al., 2024a) contrasts tool-usage and general datasets, showing that LLMs frequently fail to abstain from tool use on tasks they could solve with parametric knowledge. When2Call (Ross et al., 2025) additionally evaluates when to ask follow-up questions and when to admit insufficient tool coverage. Learning-to-Ask (Wang et al., 2025f) targets noisy or under-specified instructions and proposes Ask-when-Needed, prompting the agent to request clarification rather than hallucinate arguments.

Adaptive triggering with metacognitive and efficiency-aware signals Beyond benchmarks, several methods drive the tool-invocation decision from metacognitive signals. MeCo (Li et al., 2025b) uses hidden-state evidence to assess whether a query exceeds the model's internal capabilities and warrants external tool assistance. SMART (Qian et al., 2025b) trains agents on rationale-annotated trajectories that distinguish steps solvable by parametric knowledge from those requiring tools, mitigating tool overuse. OTC (Wang et al., 2025b) formulates tool calling as an RL problem with a tool-integrated reward that jointly considers correctness and tool-efficiency, encouraging accurate answers with minimal calls.

Tool use can be analyzed through all three roles in our framework: signals about information needs and capability limits, decisions about external actions, and mechanisms that learn or implement those decisions. These roles may be explicit or implicit, and are not separately instantiated in every method.

D Comparative Analysis

Differences in datasets, base models, prompting setups, hyperparameters, and evaluation metrics limit direct quantitative comparisons across the surveyed works. We therefore compare implementation-dependent trade-offs along the three axes of our framework.

Monitoring Signals Distributional statistics over output tokens, such as perplexity and entropy, do not require hidden-state access when the necessary token probabilities are available. They can support rule-based triggers, as in confidence-based early exit, although obtaining the signals may require additional generation (Yang et al., 2026a). In contrast, representation analysis—linear probes, contrastive activation differences, and sparse autoencoder features—requires access to internal activations, with costs depending on feature extraction, any auxiliary training, and the frequency and location of evaluation (Zhang et al., 2025a; Galichin et al., 2026). Such methods can identify signals associated with answer correctness or specific reasoning behaviors and provide directions for steering (Zhang et al., 2025a; Venhoff et al., 2025b).

Coordinating Meta-Behaviors Inference-only prompt- or template-based methods avoid updating model weights (Ning et al., 2024b; Zhou et al., 2024), but may require additional calls or tokens and depend on the prompting setup and base model's ability to follow instructions. RL-based methods can learn the coordination of reasoning behaviors, but introduce training costs and design choices concerning initialization, rewards, credit assignment, and optimization stability (Shen et al., 2025a). When an external planning module is used, implementations range from lightweight selectors to LLM-based planners and multi-agent systems. Lightweight networks or bandit policies can have low action-selection overhead, but end-to-end cost also depends on how context and feedback are obtained (Zhang et al., 2026f; Sui et al., 2026). Their action spaces may be predefined or dynamically expanded (Sui et al., 2026). LLM-based planners and multi-agent collaborations support natural-language strategy generation and role specialization, while their computational cost depends on model sizes, interaction frequency, and the resulting reasoning length (Dong et al., 2025; Wan et al., 2025). For trainable variants, parameter sharing, update schedules, and reward design also affect efficiency and learning dynamics (Wan et al., 2025).

Mechanism-Level Choices Training-time methods learn reasoning policies through parameter updates and incur data-generation and optimiza-

tion costs; their generalization and stability depend on the training setup and evaluation tasks (Shen et al., 2025b; Wan et al., 2025). Inference-time activation steering modifies internal activations without updating the base model's weights, but may require offline activation collection or probe training (Li et al., 2023; Venhoff et al., 2025b). It also requires choices about the target behavior, intervention components or layers, timing, and strength (Venhoff et al., 2025b; Cheng et al., 2025c). Cross-component approaches that combine training-time learning with inference-time steering offer a promising middle ground but remain underexplored.

Method	Monitor	Controller	Mechanism
§3.1 Problem Complexity Estimation			
DAST (Shen et al., 2025b)	Heuristic metrics (sampling accuracy + length)	Plan How to Behave (budget allocation)	Training time – Preference learning
Sheng et al. (Sheng et al., 2025)	Intrinsic perception (reasoning-strength direction)	Plan How to Behave (budget allocation)	Inference time – Steering
§3.2 Trajectory Tracking			
Concise (Qiao et al., 2025)	Distributional statistics (verbalized certainty)	Optimize Meta-Actions (early exit / pruning)	Training time – SFT / SimPO
Yang et al. (2026a)	Distributional statistics (answer perplexity)	Optimize Meta-Actions (early exit)	Inference time – rule-based
CORE (Li et al., 2025a)	Representation analysis (hidden-state cycle)	Optimize Meta-Actions (stagnation handling)	Inference time – rule-based
MeCo (Li et al., 2025b)	Representation analysis (PCA-based metacognition probe)	Optimize Meta-Actions (tool invocation trigger)	Inference time – rule-based
§4.1 Strategic Planning			
Skeleton-of-Thought (Ning et al., 2024b)	—	Plan What Structure (concise outline)	Prompt
Self-Discover (Zhou et al., 2024)	—	Plan What Structure (module composition)	Prompt
MRP (Gao et al., 2024)	—	Plan How to Behave (topology selection)	Prompt
ReasonFlux (Yang et al., 2025a)	—	Plan Which Past Experience (template trajectory)	Training time – Preference learning
Thinkless (Fang et al., 2025)	—	Plan How to Behave (fast/slow switching)	Training time – RL (DeGRPO)
§4.2 Dynamic Regulation – Coordination of Cognitive Behaviors			
Meta-reasoner (Sui et al., 2026)	Trajectory: progress report (context)	Coordination – External Policy (MAB over meta-actions)	External bandit
AutoMR (Zhang et al., 2026f)	Trajectory: context features	Coordination – External Policy (DAG construction)	RL (REINFORCE) for an MLP controller; inference-time DAG construction
rStar (Qi et al., 2025b)	—	Coordination – External Policy (MCTS over actions)	Inference time – Search (MCTS)
ReMA (Wan et al., 2025)	—	Coordination – Hierarchical (meta/object agents)	Training time – RL
MLPO (Estornell et al., 2025)	—	Coordination – Hierarchical (leader–workers)	Training time – RL
Satori (Shen et al., 2025a)	—	Coordination – Internalized (action interleaving)	Training time – RL (designed reward)
RLVMR (Zhang et al., 2026e)	—	Coordination – Internalized (verifiable actions)	Training time – RL (designed reward)
CSPO (Ha et al., 2026)	—	Coordination – Internalized (control-token)	Training time – RL (control mask)

Table 1: Methods primarily contributing on the Monitor/Controller axes. Section dividers correspond to main-text headings under §3–§4; each cell prefixes the main-text category with the method's specific instantiation. Continued in Table 2.

Method	Monitor	Controller	Mechanism
§5.1.1 Reinforcement Learning – Dense Rewards			
Math-Shepherd (Wang et al., 2024b)	—	—	Dense Rewards – External Reward Models (sampled PRM)
§5.1.1 Reinforcement Learning – Credit Assignment and Advantage Estimation			
SPO (Guo et al., 2025b)	—	—	Credit Assignment (segment-level)
BRPO (Qi et al., 2025a)	—	—	Credit Assignment (budget-relative)
HICRA (Wang et al., 2026)	—	—	Credit Assignment (planning-token advantage amplification)
§5.1.1 Reinforcement Learning – Alternative Learning Paradigms			
BARL (Zhang et al., 2026c)	—	Dynamic regulation (implicit reflection & strategy switch)	Alternative Paradigms (Bayes-adaptive RL)
§5.2.1 Steering and Activation Engineering – Representation Extraction			
ITI (Li et al., 2023)	—	—	Repr. Extraction – Probe-based + DoM
SEAL (Chen et al., 2025b)	—	—	Repr. Extraction – DoM
Venhoff et al. (2025b)	—	—	Repr. Extraction – DoM
Galichin et al. (2026)	—	—	Repr. Extraction – SAE
Zhang et al. (2025e)	—	—	Repr. Extraction – Probe-guided head selection + PCA-projected mean activations

Table 2: Methods primarily contributing on the Mechanism axis. Section dividers correspond to §5.1.1 (Reinforcement Learning) and §5.2.1 (Steering and Activation Engineering); each cell prefixes the main-text sub-paragraph with the method's specific instantiation. Continuation of Table 1.