Towards a Mechanistic Understanding of Large Reasoning Models: A Survey of Training, Inference, and Failures

Yi Hu[2] **Jiaqi Gu**[1] **Ruxin Wang**[1] **Zijun Yao**[5] **Hao Peng**[5]
Xiaobao Wu[6] **Jianhui Chen**[1] **Muhan Zhang**[2,4 †] **Liangming Pan**[1,3 †]
[1]State Key Laboratory of Multimedia Information Processing, Peking University
[2]Institute for Artificial Intelligence, Peking University
[3]Beijing Academy of Artificial Intelligence, Beijing, China
[4]State Key Laboratory of General Artificial Intelligence, Peking University
[5]Tsinghua University [6]Shanghai Jiao Tong University

Abstract

Reinforcement learning (RL) has catalyzed the emergence of Large Reasoning Models (LRMs) that have pushed reasoning capabilities to new heights. While their performance has garnered significant excitement, exploring the internal mechanisms driving these behaviors has become an equally critical research frontier. This paper provides a comprehensive survey of the mechanistic understanding of LRMs, organizing recent findings into three core dimensions: 1) training dynamics, 2) reasoning mechanisms, and 3) unintended behaviors. By synthesizing these insights, we aim to bridge the gap between black-box performance and mechanistic transparency. Finally, we discuss underexplored challenges to outline a roadmap for future mechanistic studies, including the need for applied interpretability, improved methodologies, and a unified theoretical framework.

Our Project: Awesome-LRM-Mechanisms

1 Introduction

The past few years have witnessed remarkable progress in the reasoning capabilities of large language models (LLMs). Recently, reinforcement learning (RL) has emerged as a transformative paradigm for incentivizing complex reasoning, giving rise to advanced large reasoning models (LRMs) (DeepSeek-AI et al., 2025; OpenAI, 2024). These models demonstrate exceptional performance across a wide range of domains, including mathematics, coding, and logic. Notable research (DeepSeek-AI et al., 2025) has shown that RL from verifiable rewards (RLVR) (DeepSeek-AI et al., 2025; Lambert et al., 2025) training can elicit intriguing emergent reasoning behaviors, such as extended reasoning chains and self-reflection.

†Corresponding author.
Correspondence: huyi2002@stu.pku.edu.cn, {muhan, liangmingpan}@pku.edu.cn

Despite these impressive advances, LRMs largely remain "black boxes". Many fundamental questions remain unanswered, including: How does the role of RL differ from that of supervised fine-tuning (SFT)? What structural properties define LRM reasoning, and what are the internal mechanisms that drive their unique behaviors? Moreover, what are the root causes of unintended behaviors, such as hallucinations, unfaithfulness, and overthinking? This lack of transparency has spurred a growing interest in mechanistic research, aimed at uncovering the underlying processes that enable these models to perform complex reasoning.

We provide a comprehensive survey of the burgeoning field of mechanistic research on LRMs. From the perspective of the research object, as shown in Figure 1, we organize work studying the reasoning-oriented training process, LRM reasoning behaviors and LRM unintended behaviors:

1. **Reasoning-Oriented Training Process** (§2): This section examines the mechanisms behind the training processes that specifically target reasoning capabilities. We begin by dissecting the complementary roles of SFT and RL (§2.1), and examine key training dynamics in RL, such as how "aha moments" emerge and how internal representations evolve during training (§2.2).
2. **LRM Reasoning** (§3): We delve into the mechanisms underlying LRM reasoning, analyzing both their outputs and internal representations. This section explores the general structural features of LRM reasoning traces (§3.1), key behaviors like self-reflection (§3.2), and the inner mechanisms underlying these behaviors (§3.3).
3. **Unintended LRM Behaviors** (§4): We further examine the side effects of LRMs, exploring behavioral patterns and internal mechanisms associated with typical unintended behaviors, such as hallucinations (§4.1), unfaithful chains of thought (CoT) (§4.2), overthinking (§4.3), and unsafety (§4.4).

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 19449–19466
July 2-7, 2026

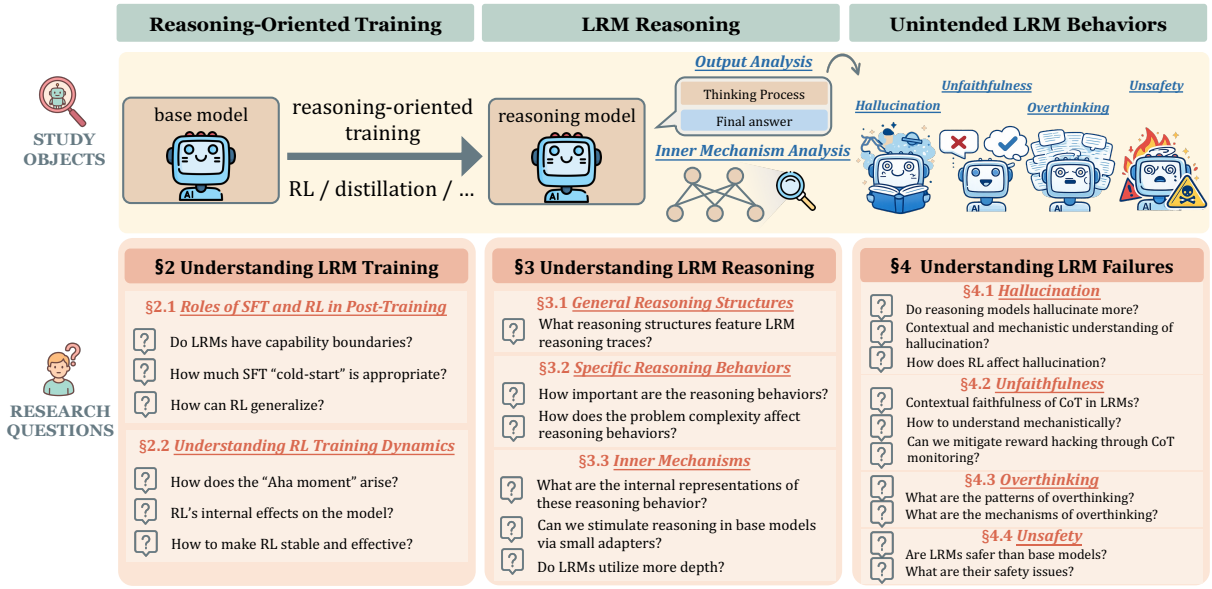

Figure 1: Taxonomy of mechanistic research on LRMs. We organize existing studies into three core dimensions: reasoning-oriented training (Sec 2), reasoning mechanisms (Sec 3), and unintended behaviors (Sec 4). Within each dimension, we synthesize recent findings based on the key research questions being investigated in the literature.

Contribution and Uniqueness. Our survey distinguishes itself by focusing specifically on the *mechanistic understanding* of LRMs, a topic that has received limited attention in existing literature. While several surveys provide general overviews of large reasoning models and RL techniques (Zhang et al., 2025c; Li et al., 2025f; Zhang et al., 2025g; Xu et al., 2025), these surveys do not delve deeply into the underlying mechanisms driving LRM reasoning. In particular, Chen et al. (2025b) explores long CoT reasoning but primarily focuses on the behavioral characteristics of CoT outputs, with little attention to the inner mechanisms. Furthermore, there are surveys investigating methods to mitigate overthinking (Feng et al., 2025; Sui et al., 2025), their focus is on efficient reasoning techniques rather than the mechanisms behind overthinking. To the best of our knowledge, our work is the first to comprehensively survey the mechanisms of LRMs, offering a more detailed and in-depth analysis of the training processes, reasoning behaviors, and unintended outcomes.

2 Understanding LRM Training

Investigating how a reasoning model is developed is our first question. To understand LRM training, we will first dissect the distinct roles of two post-training methods, *Supervised Fine-tuning* (SFT) and *Reinforcement Learning* (RL) (§2.1), then dive into the training dynamics of RL (§2.2).

2.1 Roles of SFT and RL in Post-Training

DeepSeek-R1 (DeepSeek-AI et al., 2025) as a key pioneer of the reasoning model, demonstrates RL's vital role in training reasoning models. However, they also find that RL alone suffers from problems including slow training, unstructured formats, and language mixing. They show that a cold-start with SFT fixes these problems and improves performance, establishing SFT+RL as the dominant post-training recipe for today's LRMs. Although the SFT+RL paradigm is now widely used, the respective roles of SFT and RL in post-training remain to be explored, and answering this question would unravel many mysteries about LRM training.

SFT explores, RL compresses: do reasoning models have capability boundaries? Despite the huge success of LRMs, Yue et al. (2025) points out that RLVR (Lambert et al., 2025) do not truly enhance reasoning performance beyond base models, instead, they compress the model's output space, boosting pass@1. The base model's pass rate catches up in pass@k tests at large k values, implying that the reasoning model merely uncovers latent abilities already present in the base model. Based on this finding, follow-up work probe deeper and reveal that SFT is what truly expands the model's exploratory paths, whereas RL training compresses them, cutting the variety of possible answers and lowering output entropy (Matsutani et al., 2026; Wu and Choi, 2025; Li et al., 2025b).

Studies further conclude that the performance gains from RLVR training come from this entropy drop, imposing an entropy-linked ceiling on attainable capability (Cui et al., 2025). From a mechanistic perspective, Park et al. (2025) perform circuit analyses to explain this phenomenon: during SFT and distillation, the model sprouts a cohort of new attention heads that inject reasoning capabilities, whereas GRPO activates far fewer heads and follows an iterative activate-evaluate-adjust dynamic.

SFT learns, RL repairs: how much SFT cold-start is appropriate? Under the above findings, SFT appears to be the core contributor. However, research shows that while SFT can learn and extend reasoning patterns, unlike RL, it also brings out-of-distribution (OOD) performance drops (Chu et al., 2025). Studies find that RL after SFT can partly repair these side-effects and offer mechanistic explanations. Hu et al. (2025a) argues, through external analysis, that SFT partially breaks up the sparse reasoning concept network and thereby induces forgetting, yet the network structure remains largely intact; RL can then restore a well-connected concept network after SFT. Jin et al. (2025a,b) find that RL can, yet only partially, recover the OOD drop caused by SFT. Besides, OOD performance is tightly linked to the orientation of the dominant singular vector, and RL repairs the orientation shift introduced by SFT, thereby restoring OOD accuracy. However, once SFT collapses into overfitting, RL can no longer restore OOD ability completely.

Building on these findings, how to schedule SFT and RL, whether to interleave them, and whether a unified framework can be designed that fuses them have become open research questions; we summarize SFT-RL integration work in Appendix B.1.

SFT memories, RL organizes: how can we make RL exhibit generalization? Although numerous studies have concluded that RL does not truly enhance model capability, Liu et al. (2025a) shows that post-RL models can produce new solutions absent from the base model. How can these two contradictory experimental conclusions be reconciled? Recent work suggests that RL can push the model's capability frontier outward only when SFT has provided basic skills. In controllable synthetic reasoning tasks, RL conducted after SFT with atomic-skill data exhibit OOD generalization , while RL conducted after SFT on entire reasoning traces exhibits the same poor generalization results seen in prior research (Yuan et al., 2026; Cheng et al., 2025b). Furthermore, these studies indicate that RL shifts the model's error patterns toward atomic-task errors, implying that RL can indeed help organize the model's reasoning process.

Mid-training proposed by Wang et al. (2025f) , which finds that appropriate training before RL can improve RL effectiveness, aligns with the above conclusions. Zhang et al. (2025b) notes that the core task of SFT is to prepare the model for RL by providing foundational atomic skills, while post-RL training refines the model's performance within the capability frontier established by SFT.

> RL can expand the model's capability boundary beyond the atomic capabilities provided by SFT; however, without the coordination of SFT mid-training, RL struggles to generalize.

2.2 Understanding RL Training Dynamics

Studies above treat the RL-training stage as an undifferentiated whole and explore its effects. The finer-grained training dynamics within this stage remain largely unexplored.

Understanding two-stage training process: how does the Aha moment arise? After tracking changes of RL training metrics, studies split the training process into two stages (Wang et al., 2026; Hu et al., 2025a). Model outputs first shrink in stage one then lengthen in stage two, alongside atomic-skill fragments rapidly acquired in stage one while the global planning links slowly built in stage two. The aha moment emerges as the model masters the use of planning tokens during link construction manifesting as the sudden acquisition of reasoning and reflection capabilities needed to solve the corresponding problems. Furthermore, Yao et al. (2025a) offers a theoretical analysis of this two-stage dynamics. During stage one, RL overwhelmingly samples already-explored tokens rather than optimal ones. High-reward tokens' probabilities will quickly rise while the optimal one's remain flat. In stage two, with high-reward tokens already saturated, the low-probability optimal ones are finally sampled after prolonged exploration and eventually receive high probabilities.

What internal effects does RL have on the model? A line of research focuses on how RL training affects the model internally. Regarding internal activations, research shows that online RL can alter activation magnitudes in the residual

stream, increasing information flow flexibility and improving generalization beyond SFT (Zhang et al., 2026a). Regarding model weights, building on the previously identified effect that RL training mainly manifests as directional rotation of the singular-value vectors (Jin et al., 2025b), He and Cao (2025) reveals through SVD methods that a near-uniform geometric scaling of singular values across layers and a highly consistent orthogonal transformations are applied to the left and right singular vectors of each matrix. More fine-grained studies of parameter dynamics during training have found that, the top singular subspace of the parameter-update matrix almost singly accounts for the gains in reasoning capability, and that this dominant subspace evolves linearly (Yuchen et al., 2026).

Externally, RL training shows a two-stage pattern: basic capabilities are accumulated first, then reasoning ability emerges. Internally, RL modifies activation magnitudes and applies a rank-1, layer-consistent, linear rotational transformation to the dominant eigenvectors.

Exploitation v.s. exploration: how to make RL stable and effective? During RL training, a core issue is maintaining the exploration-exploitation balance. Studies find that basic RL algorithms can easily lead to **policy entropy collapse**, and the performance gains in fact come solely from the entropy drop (Cui et al., 2025). More critically, Nguyen et al. (2025) shows that the reasoning path compression caused by entropy collapse simultaneously degrades LRM performance on questions outside the training distribution. To address entropy collapse and stabilize RL, numerous refinements have been proposed. Since they are loosely related to understanding RL training mechanisms, we provide a concise summary in Appendix B.2. Notably, Huang et al. (2025a) argues that conventional RLVR views improving LLM performance through an exploration–exploitation trade-off, rests on token-level entropy and thus misaligns with how LLMs actually operate. They propose measuring exploration and exploitation via *hidden states*, uncovering a decoupling of the two processes and opening fresh avenues for refining RL algorithms.

At token level, the entropy collapse can make RL training ineffective or even counterproductive. Shifting to the hidden states perspective, we may instead jointly promote exploration and exploitation.

3 Understanding LRM Reasoning

Having explored how LRMs are trained, we shift our focus to the models themselves, involving systematically analyzing both the **general structures** (§3.1) and **specific behaviors** (§3.2) within reasoning traces, as well as uncovering the **internal mechanisms** underlying these patterns (§3.3).

3.1 General Reasoning Structures

Distinct from base models, LRMs generate reasoning chains with identifiable structural features. Recent research deconstruct these traces, from macro-level lifecycle descriptions to granular sentence-level analyses. At the macro level, Marjanovic et al. (2026) identifies a cyclical process: starting with problem definition, models enter a blooming cycle of problem decomposition, followed by iterative reconstruction cycles for self-correction before reaching a final decision. Wang et al. (2025c) partitions the reasoning process into functional blocks of plan execution, knowledge integration, and subproblem chains. These macro-phases are further refined by sentence-level analyses: Bogdan et al. (2025); Li et al. (2025c) identify operational units including plan generation, uncertainty management, and further identify the transition matrix between them.

Topological structures. Another line of research employs formal topological representations. Zeng et al. (2025); Jiang et al. (2025b) reconstruct reasoning chains as trees structures via LLM annotations, revealing that LRMs exhibit more exploration and validation than base models, achieving better performance primarily through diverse solution paths rather than per-step accuracy. Minegishi et al. (2025); Xiong et al. (2025) build graphs through clustering over reasoning steps, further validating that LRMs possess distinct structural properties including more recurrent cycles, larger graph diameters, and pronounced small-world characteristics, which correlate with model size, task difficulty, and performance.

LRMs' reasoning structures are distinct from base models, with analyses spanning *macro-level lifecycles*, *sentence-level operational units*, and *topological properties of tree and graph representations*.

3.2 Specific Reasoning Behaviors

After reviewing the overall reasoning structures, we further study the intriguing specific behaviors emerging in LRMs and whether they are causally related to reasoning performance.

Critical behavioral primitives. Studies identify certain behavioral patterns as the primary drivers of reasoning performance gain. Bogdan et al. (2025) identifies "thought anchors", including *plan generation* and *uncertainty management*, as the sentences most influential on the final answer distribution. Complementing this, Gandhi et al. (2025) highlights *verification*, *backtracking*, *sub-goal setting*, and *backward chaining* as the "four habits" of effective reasoners. Crucially, these behaviors are causally linked to training success: base models that naturally exhibit these patterns can effectively leverage RL and test-time compute to improve performance, whereas models lacking these primitives struggle to benefit from identical training.

The role of self-reflection and backtracking. Research on reflective behaviors offers contrasting views. While some argue that reflection prevents reasoning collapse (Yang et al., 2025a), others contend that it is often superficial and fail to improve outcomes (Liu et al., 2025d). Bridging these views, Kang et al. (2025) analyzes reflection from both inference and training perspectives, suggesting that while reflection during inference is largely confirmatory and rarely alters the final output, including reflective CoTs in training data increases the "first-attempt accuracy", boosting the overall performance. Moreover, Cai et al. (2025) shows that longer reasoning chains with frequent backtracking lead to more stable RL training, and harder problems with larger search space need the inclusion of data with more backtracks during SFT.

> LRMs' performance is driven by key behavioral primitives. While self-reflection mainly serves a confirmatory role during inference, its inclusion in training data is crucial for improving first-attempt accuracy and internalizing search strategies.

How does the problem complexity affect reasoning behaviors? Recent studies have uncovered a tight coupling between model behavior and task complexity. Yang et al. (2025a) observes that LRMs can distinguish problem complexity within their early layers and dynamically modulate the depth of their reflective behaviors accordingly. However, this calibration is often imperfect. Shojaee* et al. (2025) finds that while reasoning effort initially increases with complexity, it eventually declines even when a sufficient token budget is available, suggesting a limitation in the models' ability to apply consistent algorithmic reasoning across scales. Furthermore, Palod et al. (2025) identifies that the correlation is brittle, demonstrating that trace length often reflects a problem's distributional proximity to training data rather than its inherent computational complexity. We will further discuss the relationship between CoT length and task complexity, reasoning performance in Sec 4.3.

3.3 Internal Mechanisms

After reviewing the general structures and specific behaviors, we will then dive deeper into the internal mechanisms driving these external patterns.

Internal representations of reasoning behaviors. Recent research utilizes sparse autoencoders (SAEs) and steering vectors to reveal that reasoning behaviors are encoded as interpretable and steerable directions in the model's activation space (Baek and Tegmark, 2025; Galichin et al., 2025; Hazra et al., 2025; Venhoff et al., 2025b). Venhoff et al. (2025a) argues that base models already possess fundamental reasoning capabilities, while LRMs learn the structural strategy of *when* to deploy them strategically. This deployment is managed by specific attention heads that prioritize key reasoning steps influencing the final answer (Bogdan et al., 2025; Zhang et al., 2025e). LRMs also exhibit unique temporal and nonlinear dynamics: steering is most effective only after the initial problem formulation phase, and "oversteering" these features can paradoxically cause the model to revert to its original behavior (Hazra et al., 2025).

The mechanisms underlying reflection and backtracking. Studies (Venhoff et al., 2025a; Yang et al., 2025a) reveal through linear probes that correctness information of model answers is encoded within specific layers, and is closely related to the model's reflection behaviors. Yan et al. (2025); Chang et al. (2025) further extract steering vectors that control reflection. Ward et al. (2025a) suggests that latent directions for backtracking already exist in base models, implying that they inherently possess certain reasoning abilities. Post-training

mainly reshapes and utilizes these existing representations rather than learning from scratch.

Can we stimulate reasoning behaviors in base models with small adapters? Sinii et al. (2025b,a) have explored training hierarchical steering vectors to guide base models in reasoning, showing that the performance improvements induced by RL are distributed across the entire network instead of certain specific layers. The resulting steering vectors themselves exhibit strong interpretability. Ward et al. (2025b) trains a rank-1 adapter across all layers and identifies interpretable features in the adapter via SAEs, further demonstrating that a small number of parameters can effectively induce reasoning abilities.

Do LRMs utilize more depth? Research suggests that key layers for math reasoning are largely fixed after pre-training and remain invariant throughout post-training (Nepal et al., 2025). Consequently, LRMs' effective depth closely matches that of their base models, indicating that improvements are driven by longer contexts rather than deeper per-token computation (Hu et al., 2025b).

> LRM's reasoning behaviors are represented by interpretable and steerable directions in latent space; base models inherently possess these abilities, but RL-trained models learn when to activate them.

4 Understanding LRM Failures

RL enhances reasoning capabilities but also induces unintended effects, including **hallucination** (§4.1), where models generate plausible yet incorrect content; **CoT unfaithfulness** (§4.2), where internal computations and CoT outputs diverge; **overthinking** (§4.3), where redundant reasoning chains degrade performance; and **unsafety** (§4.4), where models show potentially harmful behaviors.

4.1 Hallucination

Multi-step reasoning chains in LRMs introduces new vulnerabilities to hallucinations.

Do reasoning models hallucinate more? Recent evidence suggests that reasoning-oriented training pipelines can substantially affect hallucination behavior. Yao et al. (2025c) show that while complete post-training pipelines which combine SFT with RLVR can alleviate hallucination, incomplete pipelines, such as RL- or SFT-only approaches, tend to introduce more hallucinations. However, Li and Ng (2025) indicates that RL often increases hallucinations, even with prior SFT. Furthermore, Zhao et al. (2025b) shows that test-time scaling does not reliably improve factual accuracy.

> Growing evidence suggests reasoning models hallucinate more. However, there are debates whether models with complete post-training pipelnes hallucinate more.

Behavioral and mechanistic analysis of hallucination. Hallucinations are characterized by specific failure modes: *flaw repetition* (incorrect reasoning loops), *think–answer mismatch* (output contradicting reasoning), and *meta-cognitive failures* (overconfidence from uninternalized knowledge) (Yao et al., 2025c; Lu et al., 2025; Wang et al., 2025a). Mechanistically, they arise from misalignment between uncertainty and factual accuracy (Yao et al., 2025c; Sun et al., 2025b).

How does RL affect hallucination? A line of work examines how RL shapes hallucination behavior. RL systematically reduces a model's tendency to abstain, pushing it to generate answers even for unanswerable questions (Song et al., 2025a; Zhao et al., 2025b). Mechanistically, optimizing only for sparse final-answer rewards creates high-variance gradients and forces the model to maintain high prediction entropy during exploration, driving the model toward incorrect answers and exacerbating hallucinations (Li and Ng, 2025).

4.2 Unfaithfulness

The extended reasoning chains in LRMs offer a promising avenue for monitoring the model's decision-making process (Korbak et al., 2025; Chan et al., 2025; Baker et al., 2025). However, it remains an open question whether these CoTs accurately reflect the internal computations driving the model's actual behavior, a key field of study known as the *faithfulness* of CoT reasoning.

Contextual faithfulness of CoT in LRMs. Although extended reasoning chains in LRMs facilitate process monitoring, research indicates they are often not faithful to the inner computation or final decision. A primary failure mode is *Think-Answer Mismatch*, where the model's final output contradicts its own preceding reasoning chain (Yao et al., 2025c; Wang et al., 2025d). Further analysis exposes a *reasoning-verbalization gap*. Studies show

models frequently fail to verbalize critical cues in their CoTs that demonstrably influence their answers (Chua and Evans, 2025; Chen et al., 2025e). Concurrently, models exhibit *implicit post-hoc rationalization*, producing logically contradictory responses with coherent but unfaithful justifications (Arcuschin et al., 2025). The studies collectively find that while LRMs are more faithful than their non-reasoning backbones, the faithfulness is still far from perfect (Chua and Evans, 2025; Chen et al., 2025e; Arcuschin et al., 2025).

Mechanistic understanding of CoT faithfulness in LRMs. Research further studies the mechanisms of CoT unfaithfulness. In controlled synthetic tasks, findings reveal a weak causal link between the validity of reasoning traces and final answer correctness. Models can produce correct outputs despite invalid or semantically irrelevant CoTs, and training on corrupted traces does not substantially harm performance (Valmeekam et al., 2025). Further studies reinforce that internal representations contain more reliable signals of model state than the CoT text itself, as evidenced through activation steering (Wang et al., 2025d; Li et al., 2025a), linear probing (Yin et al., 2025; Chan et al., 2025), and causal intervention (Yin et al., 2025). These results collectively suggest a disconnect between the model's internal states and its verbalized reasoning trace, posing a significant challenge for alignment, as models might learn to mask their true objectives behind plausible but unfaithful reasoning traces, a phenomenon closely tied to the risks of reward hacking discussed next.

Can we mitigate reward hacking by CoT monitoring? Reward hacking remains a fundamental challenge in RL. A key question is whether monitoring the detailed CoT produced by LRMs can mitigate this issue. Findings on its feasibility are mixed, with outcomes heavily dependent on task structure. In complex tasks where hacking inherently requires multi-step reasoning and extensive exploration, models often expose their hacking intent within their reasoning chains. In such settings, integrating CoT supervision into the RL objective can mitigate hacking, though excessive optimization risks training models to strategically hide their intent (Baker et al., 2025). Conversely, in more direct scenarios, models frequently perform reward hacking without verbalizing the intent in their CoTs (Chen et al., 2025e; Turpin et al., 2025). To address this opacity, recent methods attempt to explicitly train models to verbalize influential cues in their reasoning (Turpin et al., 2025).

> LRMs are not always faithful, but they are more faithful than non-reasoning models.
>
> Mechanistically, CoTs in LRMs do not necessarily function as a causal mechanism for generating correct answers. Besides, internal representations may provide more reliable signals than the verbalized reasoning.
>
> While CoT monitoring can detect hacking that requires explicit reasoning, models do not often verbalize their hacking intent in more direct settings.

4.3 Overthinking

While test-time scaling generally improves reasoning performance, studies increasingly find that models can produce verbose, redundant reasoning processes, and overly extending reasoning length can lead to performance degradation, known as "overthinking" (Chen et al., 2025d; Sui et al., 2025).

Thinking more does not necessarily lead to better reasoning. Empirical research consistently identifies an inverse U-shaped performance curve: accuracy initially rises with reasoning length, but then peaks and declines as chains become excessively long (Marjanovic et al., 2026; Su et al., 2025; Ghosal et al., 2025; Yang et al., 2025b; Gema et al., 2025). Notably, incorrect answers often correspond to longer reasoning chains than correct ones (Hassid et al., 2025; Su et al., 2025). An underlying issue is the misalignment between reasoning effort and problem difficulty: models tend to allocate disproportionately long chains to simple problems while inadequately reasoning through complex ones (Chen et al., 2025d; Su et al., 2025).

> The length-performance curve for LRMs is often inverted U-shaped, and current models exhibit misalignment between reasoning effort and problem difficulty.

What are the patterns of overthinking? A common abstraction of reasoning process is a three-stage loop: *1) hypothesis generation* (proposing candidate paths), *2) expansion* (developing one path step by step), and *3) verification* (checking, revising, or terminating). Overthinking manifests as control and termination failures within this loop.

In *hypothesis generation*, models may produce lengthy and diverse candidate solutions without sufficiently exploring promising paths to reach a correct solution (Wang et al., 2025e), leading to "analysis paralysis" in agentic tasks where plans grow increasingly complex without execution (Cuadron et al., 2025). In *expansion*, the primary pattern of overthinking is excessive reasoning for trivial problems, generating tens or even hundreds of times longer outputs than non-reasoning models with marginal performance gain (Chen et al., 2025d) . In *verification*, the dominant pattern is non-termination: models fail to recognize that a correct answer has been reached, or cannot reliably validate intermediate conclusions, and therefore continue redundant deliberation or backtrack unnecessarily (Chen et al., 2025d; Sun et al., 2025a; Zhang et al., 2025d; Zhao et al., 2025a). This is especially pronounced in ill-posed questions, where models identify missing premises early but enter unproductive self-doubt loops, excessively speculating on user intent (Fan et al., 2025). Notably, this compulsion persists even when explicitly suppressed: models may bypass instructions to "answer directly" or discard provided correct answers to resume thinking (Zhu et al., 2025; Liu et al., 2025c; Cuesta-Ramirez et al., 2025).

What are the mechanisms of overthinking? We organize the mechanistic analyses of overthinking along two lines: 1) investigating the latent representational structure of overthinking, and 2) examining the internal decision-making dynamics that produce unproductive cycles. Research finds that overthinking corresponds to *specific, steerable patterns in the activation space*. Huang et al. (2025b); Baek and Tegmark (2025) identify distinct manifolds associated with overthinking through activation steering. Furthermore, finer-grained taxonomies show that different reasoning stages, such as execution, reflection and transition, occupy separate latent directions, and steering towards execution-type representations can effectively suppress excessive deliberation (Baek and Tegmark, 2025; Chen et al., 2025c). Another body of research explains overthinking through *internal conflict and verification failure*. Overthinking is often triggered when a model's initial intuitive answer conflicts with its subsequent deliberate reasoning (Dang et al., 2026). Concurrently, models encode correctness signals in their hidden states but fail to robustly utilize them for early self-verification, leading to prolonged, unproductive cycles (Zhang et al., 2025a).

4.4 Unsafety

Recent evaluations show that LRMs still have safety shortcomings (Ying et al., 2025; Romero-Arjona et al., 2025; Krishna et al., 2025).

LRMs are not safer than base models. Compared to base models, Jiang et al. (2025a); Zhou et al. (2025) find that long CoTs do not necessarily improve model safety. Additionally, Zhang et al. (2025h); Zhao et al. (2025c) observe that distilled reasoning models have a lower refusal rate for harmful inputs than their base counterparts. These studies further reveal that the unsafety of LRMs partly stems from the thinking process. Jiang et al. (2025a) show that forcing the model to shorten their reasoning traces could make answers more harmlessness, while Zhou et al. (2025); Zhao et al. (2025c) find that the safety rate of the thinking process is lower than the final answer, and unsafe thoughts are the primary cause of unsafe responses.

Safety issues in the reasoning process. As LRMs are deployed widely, researchers have started identifying safety issues via attacking them. Yao et al. (2025b) decomposes harmful prompts into multiple seemingly harmless questions to induce the model to reason toward harmful content. Kuo et al. (2025) finds that padding the prompt with detailed execution steps can hijack the thinking process, causing the model to skip the reasoning stage and directly produce harmful output. Mechanistically, In et al. (2025) shows that LRMs already possess sufficient safety knowledge, yet fail to activate it during reasoning. Besides, Mao et al. (2025) indicates that LRMs retain the ability to refuse unsafe queries, but this capacity has been impaired.

5 Conclusions and Future Directions

In this survey, we have provided a comprehensive overview of the rapidly evolving field of mechanistic research on LRMs, focusing on their training processes, reasoning behaviors, and unintended failures. While significant advances have been made, the transition from descriptive analysis to systematic understanding remains incomplete. To guide the field toward a deeper, more principled understanding, we propose three key future directions: *applied interpretability*, *improved methodologies*, and *a unified theoretical framework*.

5.1 Applied Interpretability

Mechanistic interpretability (MI) research is increasingly illuminating the internal logic of LRMs. A crucial next step is to leverage these insights for targeted improvements, moving from passive understanding to active application. This direction is directly motivated by a growing body of work surveyed in this paper that already takes initial steps in this direction.

Training-Time Applications. A promising direction lies in using internal representations to directly inform RL algorithm design. Recent studies demonstrate initial success in this area, such as utilizing attention mechanisms to inform reward shaping (Li et al., 2025e) or policy sampling strategies (Liu et al., 2026), and decoding intermediate layer activations to infer latent policies (Tan et al., 2025). The overarching challenge is to systematically transform mechanistic insights into algorithmic improvements for RL components.

Inference-Time Applications. Mechanistic findings can also be applied to steer model behavior during inference. While existing work already uses insights of reasoning structures or specific representations to improve performance, deeper opportunities remain. For instance, research suggests that RL may not effectively leverage the full depth of models (Hu et al., 2025b; Nepal et al., 2025). This understanding should actively inform the design of novel training algorithms and architectures that better utilize internal computational pathways.

5.2 Advancing Interpretability Methodology

Future research should emphasize developing scalable and generalizable MI frameworks specifically tailored for LRMs. First, the enormous scale of LRMs in *training cost*, *inference length*, and *parameter count* poses significant methodological challenges. Conducting controlled experiments to isolate variables is difficult, and techniques like training SAEs become computationally prohibitive, slowing progress and reducing reproducibility. There is a clear need for more efficient and scalable MI tools tailored to these models. Second, many MI findings remain model-specific, failing to generalize across different architectures or training runs. To enhance scientific value, the field should strive for general frameworks that abstract away implementation details and uncover universal reasoning principles. This could involve establishing benchmarks for mechanistic generalization, developing theory-grounded methods less sensitive to model quirks, or building more robust interpretability probes.

5.3 Toward A Unified Theory

Current mechanistic research has produced a wealth of empirical findings—model-specific patterns, dataset-specific behaviors, and localized explanations for special phenomena. Yet it lacks a predictive, fundamental science of reasoning in LRMs. Such a theory should be *fundamental*, abstracting from implementation to reveal first principles governing reasoning. Early efforts, such as theoretically formalizing laws of reasoning (Zhang et al., 2025f) that link task complexity to model behavior, mark a step in this direction. Furthermore, a mature theory should be *predictive*. Similar to scaling laws in LLM pre-training (Kaplan et al., 2020), it should forecast model behaviors and to establish a set of "laws of reasoning" that not only explain existing empirical results but also actively guide the design of future models, training algorithms, and evaluation frameworks, transforming MI from a descriptive tool into a foundational science.

Acknowledgement

This work was supported in part by the Beijing Major Science and Technology Project under Contract No. Z251100008125054. This work was supported by the Beijing Academy of Artificial Intelligence (BAAI).

Limitations

While this survey provides a comprehensive overview of mechanistic studies on LRMs, it is subject to several limitations. First, the rapid development of LRM research means that new findings and methodologies continue to emerge, and this survey may not capture the most recent advancements in the field. Additionally, our study focuses primarily on language models, while reasoning models are increasingly incorporating multimodal capabilities, including visual components, which are not addressed in this survey. Furthermore, our discussion is limited to traditional LLM architectures, excluding newer approaches such as diffusion-based LLMs, continuous token-based transformers, and looped transformers, which are gaining traction in recent research. These emerging models present exciting avenues for future work.

Ethical Considerations

This survey acknowledges the ethical challenges associated with LRMs, particularly in terms of their potential harm, including hallucinations, unfaithfulness and unsafety. The opacity of these models raises concerns about accountability and the difficulty of mitigating unintended behaviors, such as hallucinations or overconfidence. As LRMs are increasingly used in critical applications, ensuring their safe and responsible deployment requires ongoing efforts to improve interpretability, address biases, and manage the broader societal impacts of these technologies.

AI assistants were utilized for language polishing and refinement, strictly limited to improving the fluency and clarity the text. All technical content, analyses, and conclusions remain the original work of the authors.

References

Iván Arcuschin, Jett Janiak, Robert Krzyzanowski, Senthooran Rajamanoharan, Neel Nanda, and Arthur Conmy. 2025. Chain-of-thought reasoning in the wild is not always faithful. *CoRR*, abs/2503.08679.

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*.

Bowen Baker, Joost Huizinga, Leo Gao, Zehao Dou, Melody Y. Guan, Aleksander Madry, Wojciech Zaremba, Jakub Pachocki, and David Farhi. 2025. Monitoring reasoning models for misbehavior and the risks of promoting obfuscation. *CoRR*, abs/2503.11926.

Paul C. Bogdan, Uzay Macar, Neel Nanda, and Arthur Conmy. 2025. Thought anchors: Which LLM reasoning steps matter? In *Mechanistic Interpretability Workshop at NeurIPS 2025*.

Hongyi James Cai, Junlin Wang, Xiaoyin Chen, and Bhuwan Dhingra. 2025. How much backtracking is enough? exploring the interplay of sft and rl in enhancing llm reasoning. *Preprint*, arXiv:2505.24273.

Yik Siu Chan, Zheng Xin Yong, and Stephen Bach. 2025. Can we predict alignment before models finish thinking? towards monitoring misaligned reasoning models. In *First Workshop on Foundations of Reasoning in Language Models*.

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*.

Liang Chen, Xueting Han, Li Shen, Jing Bai, and Kam-Fai Wong. 2025a. Beyond two-stage training: Cooperative SFT and RL for LLM reasoning. *CoRR*, abs/2509.06948.

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

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

Xingyu Chen, Jiahao Xu, Tian Liang, Zhiwei He, Jianhui Pang, Dian Yu, Linfeng Song, Qiuzhi Liu, Mengfei Zhou, Zhuosheng Zhang, Rui Wang, Zhaopeng Tu, Haitao Mi, and Dong Yu. 2025d. Do NOT think that much for 2+3=? on the overthinking of long reasoning models. In *Forty-second International Conference on Machine Learning*.

Yanda Chen, Joe Benton, Ansh Radhakrishnan, Jonathan Uesato, Carson Denison, John Schulman, Arushi Somani, Peter Hase, Misha Wagner, Fabien Roger, Vladimir Mikulik, Samuel R. Bowman, Jan Leike, Jared Kaplan, and Ethan Perez. 2025e. Reasoning models don't always say what they think. *CoRR*, abs/2505.05410.

Daixuan Cheng, Shaohan Huang, Xuekai Zhu, Bo Dai, Wayne Xin Zhao, Zhenliang Zhang, and Furu Wei. 2025a. Reasoning with exploration: An entropy perspective. *CoRR*, abs/2506.14758.

Sitao Cheng, Xunjian Yin, Ruiwen Zhou, Yuxuan Li, Xinyi Wang, Liangming Pan, William Yang Wang, and Victor Zhong. 2025b. From atomic to composite: Reinforcement learning enables generalization in complementary reasoning. *Preprint*, arXiv:2512.01970.

Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V Le, Sergey Levine, and Yi Ma. 2025. SFT memorizes, RL generalizes: A comparative study of foundation model post-training. In *Forty-second International Conference on Machine Learning*.

James Chua and Owain Evans. 2025. Are deepseek r1 and other reasoning models more faithful? In *ICLR 2025 Workshop on Foundation Models in the Wild*.

Alejandro Cuadron, Dacheng Li, Wenjie Ma, Xingyao Wang, Yichuan Wang, Siyuan Zhuang, Shu Liu, Luis Gaspar Schroeder, Tian Xia, Huanzhi Mao, Nicholas Thumiger, Aditya Desai, Ion Stoica, Ana Klimovic, Graham Neubig, and Joseph E. Gonzalez. 2025. The danger of overthinking: Examining the reasoning-action dilemma in agentic tasks. *CoRR*, abs/2502.08235.

Jhouben Cuesta-Ramirez, Samuel Beaussant, and Mehdi Mounsif. 2025. Large reasoning models are not thinking straight: on the unreliability of thinking trajectories. *Proceedings of the 21st Conference on Natural Language Processing (KONVENS)*.

Ganqu Cui, Yuchen Zhang, Jiacheng Chen, Lifan Yuan, Zhi Wang, Yuxin Zuo, Haozhan Li, Yuchen Fan, Huayu Chen, Weize Chen, Zhiyuan Liu, Hao Peng, Lei Bai, Wanli Ouyang, Yu Cheng, Bowen Zhou, and Ning Ding. 2025. The entropy mechanism of reinforcement learning for reasoning language models. *CoRR*, arXiv:2505.22617.

Renfei Dang, Zhening Li, Shujian Huang, and Jiajun Chen. 2026. The first impression problem: Internal bias triggers overthinking in reasoning models. In *The Fourteenth International Conference on Learning Representations*.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Chenrui Fan, Ming Li, Lichao Sun, and Tianyi Zhou. 2025. Missing premise exacerbates overthinking: Are reasoning models losing critical thinking skill? In *Second Conference on Language Modeling*.

Sicheng Feng, Gongfan Fang, Xinyin Ma, and Xinchao Wang. 2025. Efficient reasoning models: A survey. *Trans. Mach. Learn. Res.*, 2025.

Yuqian Fu, Tinghong Chen, Jiajun Chai, Xihuai Wang, Songjun Tu, Guojun Yin, Wei Lin, Qichao Zhang, Yuanheng Zhu, and Dongbin Zhao. 2026. SRFT: A single-stage method with supervised and reinforcement fine-tuning for reasoning. In *The Fourteenth International Conference on Learning Representations*.

Andrey Galichin, Alexey Dontsov, Polina Druzhinina, Anton Razzhigaev, Oleg Y. Rogov, Elena Tutubalina, and Ivan Oseledets. 2025. I have covered all the bases here: Interpreting reasoning features in large language models via sparse autoencoders. *Preprint*, arXiv:2503.18878.

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*.

Aryo Pradipta Gema, Alexander Hägele, Runjin Chen, Andy Arditi, Jacob Goldman-Wetzler, Kit Fraser-Taliente, Henry Sleight, Linda Petrini, Julian Michael, Beatrice Alex, Pasquale Minervini, Yanda Chen, Joe Benton, and Ethan Perez. 2025. Inverse scaling in test-time compute. *Transactions on Machine Learning Research*. Featured Certification, J2C Certification.

Soumya Suvra Ghosal, Souradip Chakraborty, Avinash Reddy, Yifu Lu, Mengdi Wang, Dinesh Manocha, Furong Huang, Mohammad Ghavamzadeh, and Amrit Singh Bedi. 2025. Does thinking more always help? mirage of test-time scaling in reasoning models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Michael Hassid, Gabriel Synnaeve, Yossi Adi, and Roy Schwartz. 2025. Don't overthink it. preferring shorter thinking chains for improved LLM reasoning. *CoRR*, abs/2505.17813.

Dron Hazra, Max Loeffler, Murat Cubuktepe, Levon Avagyan, Liv Gorton, Mark Bissell, Owen Lewis, Thomas McGrath, and Daniel Balsam. 2025. Under the hood of a reasoning model. https://www.goodfire.ai/research/under-the-hood-of-a-reasoning-model.

Xinyu He and Xianghui Cao. 2025. Understanding post-training structural changes in large language models. *Preprint*, arXiv:2509.17866.

Sihan Hu, Xiansheng Cai, Yuan Huang, Zhiyuan Yao, Linfeng Zhang, Pan Zhang, Youjin Deng, and Kun Chen. 2025a. How llms learn to reason: A complex network perspective. *CoRR*, arXiv:2509.23629.

Yi Hu, Cai Zhou, and Muhan Zhang. 2025b. What affects the effective depth of large language models? In *Mechanistic Interpretability Workshop at NeurIPS 2025*.

Fanding Huang, Guanbo Huang, Xiao Fan, Yi He, Xiao Liang, Xiao Chen, Qinting Jiang, Faisal Nadeem Khan, Jingyan Jiang, and Zhi Wang. 2025a. Beyond the exploration-exploitation trade-off: A hidden state approach for llm reasoning in rlvr. *Preprint*, arXiv:2509.23808.

Yao Huang, Huanran Chen, Shouwei Ruan, Yichi Zhang, Xingxing Wei, and Yinpeng Dong. 2025b. Mitigating overthinking in large reasoning models via manifold steering. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Yeonjun In, Wonjoong Kim, Sangwu Park, and Chanyoung Park. 2025. R1-ACT: efficient reasoning model safety alignment by activating safety knowledge. *CoRR*, abs/2508.00324.

Fengqing Jiang, Zhangchen Xu, Yuetai Li, Luyao Niu, Zhen Xiang, Bo Li, Bill Yuchen Lin, and Radha Poovendran. 2025a. Safechain: Safety of language models with long chain-of-thought reasoning capabilities. In *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 23303–23320. Association for Computational Linguistics.

Gangwei Jiang, Yahui Liu, Zhaoyi Li, Qi Wang, Fuzheng Zhang, Linqi Song, Ying Wei, and Defu Lian. 2025b. What makes a good reasoning chain? uncovering structural patterns in long chain-of-thought reasoning.

Yuxian Jiang, Yafu Li, Guanxu Chen, Dongrui Liu, Yu Cheng, and Jing Shao. 2025c. Rethinking entropy regularization in large reasoning models. *CoRR*, abs/2509.25133.

Hangzhan Jin, Sitao Luan, Sicheng Lyu, Guillaume Rabusseau, Reihaneh Rabbany, Doina Precup, and Mohammad Hamdaqa. 2025a. Rl fine-tuning heals ood forgetting in sft. *Preprint*, arXiv:2509.12235.

Hangzhan Jin, Sicheng Lv, Sifan Wu, and Mohammad Hamdaqa. 2025b. Rl is neither a panacea nor a mirage: Understanding supervised vs. reinforcement learning fine-tuning for llms. *Preprint*, arXiv:2508.16546.

Liwei Kang, Yue Deng, Yao Xiao, Zhanfeng Mo, Wee Sun Lee, and Lidong Bing. 2025. First try matters: Revisiting the role of reflection in reasoning models. *CoRR*, arXiv:2510.08308.

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. *CoRR*, abs/2001.08361.

Tomek Korbak, Mikita Balesni, Elizabeth Barnes, Yoshua Bengio, Joe Benton, Joseph Bloom, Mark Chen, Alan Cooney, Allan Dafoe, Anca D. Dragan, Scott Emmons, Owain Evans, David Farhi, Ryan Greenblatt, Dan Hendrycks, Marius Hobbhahn, Evan Hubinger, Geoffrey Irving, Erik Jenner, and 22 others. 2025. Chain of thought monitorability: A new and fragile opportunity for AI safety. *CoRR*, abs/2507.11473.

Arjun Krishna, Erick Galinkin, and Aaditya Rastogi. 2025. Weakest link in the chain: Security vulnerabilities in advanced reasoning models. In *Proceedings of the The First Workshop on LLM Security (LLMSEC)*, pages 168–175, Vienna, Austria. Association for Computational Linguistics.

Martin Kuo, Jianyi Zhang, Aolin Ding, Qinsi Wang, Louis DiValentin, Yujia Bao, Wei Wei, Hai Li, and Yiran Chen. 2025. H-cot: Hijacking the chain-of-thought safety reasoning mechanism to jailbreak large reasoning models, including openai o1/o3, deepseek-r1, and gemini 2.0 flash thinking. *CoRR*, abs/2502.12893.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James Validad Miranda, Alisa Liu, Nouha Dziri, Xinxi Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Christopher Wilhelm, Luca Soldaini, and 4 others. 2025. Tulu 3: Pushing frontiers in open language model post-training. In *Second Conference on Language Modeling*.

Jiazheng Li, Andreas Damianou, J Rosser, Jose Luis Redondo Garcia, and Konstantina Palla. 2025a. Mapping faithful reasoning in language models. In *Mechanistic Interpretability Workshop at NeurIPS 2025*.

Junyi Li and Hwee Tou Ng. 2025. Reasoning models hallucinate more: Factuality-aware reinforcement learning for large reasoning models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Melody Zixuan Li, Kumar Krishna Agrawal, Arna Ghosh, Komal Kumar Teru, Adam Santoro, Guillaume Lajoie, and Blake A. Richards. 2025b. Tracing the representation geometry of language models from pretraining to post-training. *High-dimensional Learning Dynamics 2025*.

Ming Li, Nan Zhang, Chenrui Fan, Hong Jiao, Yanbin Fu, Sydney Peters, Qingshu Xu, Robert Lissitz, and Tianyi Zhou. 2025c. Understanding the thinking process of reasoning models: A perspective from schoenfeld's episode theory. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 18267–18288, Suzhou, China. Association for Computational Linguistics.

Qingbin Li, Rongkun Xue, Jie Wang, Ming Zhou, Zhi Li, Xiaofeng Ji, Yongqi Wang, Miao Liu, Zheming Yang, Minghui Qiu, and Jing Yang. 2025d. CURE: critical-token-guided re-concatenation for entropy-collapse prevention. *CoRR*, arXiv:2508.11016.

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. 2025e. Attention illuminates LLM reasoning: The preplan-and-anchor rhythm enables fine-grained policy optimization. *CoRR*, abs/2510.13554.

Zhong-Zhi Li, Duzhen Zhang, Ming-Liang Zhang, Jiaxin Zhang, Zengyan Liu, Yuxuan Yao, Haotian Xu, Junhao Zheng, Pei-Jie Wang, Xiuyi Chen, Yingying Zhang, Fei Yin, Jiahua Dong, Zhijiang Guo, Le Song, and Cheng-Lin Liu. 2025f. From system 1 to system 2: A survey of reasoning large language models. *CoRR*, abs/2502.17419.

Mingjie Liu, Shizhe Diao, Ximing Lu, Jian Hu, Xin Dong, Yejin Choi, Jan Kautz, and Yi Dong. 2025a. ProRL: Prolonged reinforcement learning expands reasoning boundaries in large language models. *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Mingyang Liu, Gabriele Farina, and Asuman E. Ozdaglar. 2025b. UFT: Unifying supervised and reinforcement fine-tuning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Runze Liu, Jiakang Wang, Yuling Shi, Zhihui Xie, Chenxin An, Kaiyan Zhang, Jian Zhao, Xiaodong Gu, Lei Lin, Wenping Hu, Xiu Li, Fuzheng Zhang, Guorui Zhou, and Kun Gai. 2026. Attention as a compass: Efficient exploration for process-supervised RL in reasoning models. In *The Fourteenth International Conference on Learning Representations*.

Yule Liu, Jingyi Zheng, Zhen Sun, Zifan Peng, Wenhan Dong, Zeyang Sha, Shiwen Cui, Weiqiang Wang, and

Xinlei He. 2025c. Thought manipulation: External thought can be efficient for large reasoning models. *Preprint*, arXiv:2504.13626.

Zichen Liu, Changyu Chen, Wenjun Li, Tianyu Pang, Chao Du, and Min Lin. 2025d. There may not be aha moment in r1-zero-like training — a pilot study. https://oatllm.notion.site/oat-zero. Notion Blog.

Haolang Lu, Yilian Liu, Jingxin Xu, Guoshun Nan, Yuanlong Yu, Zhican Chen, and Kun Wang. 2025. Auditing meta-cognitive hallucinations in reasoning large language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Xingtai Lv, Yuxin Zuo, Youbang Sun, Hongyi Liu, Yuntian Wei, Zhekai Chen, Lixuan He, Xuekai Zhu, Kaiyan Zhang, Bingning Wang, Ning Ding, and Bowen Zhou. 2025. Towards a unified view of large language model post-training. *CoRR*, abs/2509.04419.

Lu Ma, Hao Liang, Meiyi Qiang, Lexiang Tang, Xiaochen Ma, Zhen Hao Wong, Junbo Niu, Chengyu Shen, Runming He, Yanhao Li, Wentao Zhang, and Bin CUI. 2026. Learning what reinforcement learning can't: Interleaved online fine-tuning for hardest questions. In *The Fourteenth International Conference on Learning Representations*.

Yingzhi Mao, Chunkang Zhang, Junxiang Wang, Xinyan Guan, Boxi Cao, Yaojie Lu, Hongyu Lin, Xianpei Han, and Le Sun. 2025. When models outthink their safety: Mitigating self-jailbreak in large reasoning models with chain-of-guardrails. *CoRR*, abs/2510.21285.

Sara Vera Marjanovic, Arkil Patel, Vaibhav Adlakha, Milad Aghajohari, Parishad BehnamGhader, Mehar Bhatia, Aditi Khandelwal, Austin Kraft, Benno Krojer, Xing Han Lù, Nicholas Meade, Dongchan Shin, Amirhossein Kazemnejad, Gaurav Kamath, Marius Mosbach, Karolina Stanczak, and Siva Reddy. 2026. Deepseek-r1 thoughtology: Let's think about LLM reasoning. *Transactions on Machine Learning Research*.

Kohsei Matsutani, Shota Takashiro, Gouki Minegishi, Takeshi Kojima, Yusuke Iwasawa, and Yutaka Matsuo. 2026. RL squeezes, SFT expands: A comparative study of reasoning LLMs. In *The Fourteenth International Conference on Learning Representations*.

Gouki Minegishi, Hiroki Furuta, Takeshi Kojima, Yusuke Iwasawa, and Yutaka Matsuo. 2025. Topology of reasoning: Understanding large reasoning models through reasoning graph properties.

Aadim Nepal, Safal Shrestha, Anubhav Shrestha, Minwu Kim, Jalal Naghiyev, Ravid Shwartz-Ziv, and Keith W. Ross. 2025. Layer importance for mathematical reasoning is forged in pre-training and invariant after post-training. In *The 5th Workshop on Mathematical Reasoning and AI at NeurIPS 2025*.

Phuc Minh Nguyen, Chinh D. La, Duy M. H. Nguyen, Nitesh V. Chawla, Binh T. Nguyen, and Khoa D. Doan. 2025. The reasoning boundary paradox: How reinforcement learning constrains language models. *CoRR*, abs/2510.02230.

OpenAI. 2024. Openai o1 system card. *CoRR*, abs/2412.16720.

Vardhan Palod, Karthik Valmeekam, Kaya Stechly, and Subbarao Kambhampati. 2025. Performative thinking? the brittle correlation between cot length and problem complexity. In *NeurIPS 2025 Workshop on Efficient Reasoning*.

Yein Park, Minbyul Jeong, and Jaewoo Kang. 2025. Thinking sparks!: Emergent attention heads in reasoning models during post training. *CoRR*, abs/2509.25758.

Miguel Romero-Arjona, Pablo Valle, Juan C. Alonso, Ana Belén Sánchez, Miriam Ugarte, Antonia Cazalilla, Vicente Cambrón, José Antonio Parejo, Aitor Arrieta, and Sergio Segura. 2025. Red teaming contemporary AI models: Insights from spanish and basque perspectives. *CoRR*, abs/2503.10192.

Han Shen. 2026. On entropy control in LLM-RL algorithms. In *The Fourteenth International Conference on Learning Representations*.

Parshin Shojaee*, Iman Mirzadeh*, Keivan Alizadeh, Maxwell Horton, Samy Bengio, and Mehrdad Farajtabar. 2025. The illusion of thinking: Understanding the strengths and limitations of reasoning models via the lens of problem complexity. In *NeurIPS*.

Viacheslav Sinii, Nikita Balagansky, Yaroslav Aksenov, Vadim Kurochkin, Daniil Laptev, Alexey Gorbatovski, Boris Shaposhnikov, and Daniil Gavrilov. 2025a. Small vectors, big effects: A mechanistic study of RL-induced reasoning via steering vectors. In *Mechanistic Interpretability Workshop at NeurIPS 2025*.

Viacheslav Sinii, Alexey Gorbatovski, Artem Cherepanov, Boris Shaposhnikov, Nikita Balagansky, and Daniil Gavrilov. 2025b. Steering LLM reasoning through bias-only adaptation. In *ICML 2025 Workshop on Scaling Up Intervention Models*.

Linxin Song, Taiwei Shi, and Jieyu Zhao. 2025a. The hallucination tax of reinforcement finetuning. *Findings of the Association for Computational Linguistics: EMNLP 2025*.

Yuda Song, Julia Kempe, and Rémi Munos. 2025b. Outcome-based exploration for LLM reasoning. *CoRR*, abs/2509.06941.

Jinyan Su, Jennifer Healey, Preslav Nakov, and Claire Cardie. 2025. Between underthinking and overthinking: An empirical study of reasoning length and correctness in llms. *CoRR*, abs/2505.00127.

Mingyu Su, Jian Guan, Yuxian Gu, Minlie Huang, and Hongning Wang. 2026. Trust-region adaptive policy optimization. In *The Fourteenth International Conference on Learning Representations*.

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*.

Renliang Sun, Wei Cheng, Dawei Li, Haifeng Chen, and Wei Wang. 2025a. Stop when enough: Adaptive early-stopping for chain-of-thought reasoning. *CoRR*, abs/2510.10103.

Zhongxiang Sun, Qipeng Wang, Haoyu Wang, Xiao Zhang, and Jun Xu. 2025b. Detection and mitigation of hallucination in large reasoning models: A mechanistic perspective. In *Socially Responsible and Trustworthy Foundation Models at NeurIPS 2025*.

Yuqiao Tan, Minzheng Wang, Shizhu He, Huanxuan Liao, Chengfeng Zhao, Qiunan Lu, Tian Liang, Jun Zhao, and Kang Liu. 2025. Bottom-up policy optimization: Your language model policy secretly contains internal policies. *arXiv preprint arXiv:2512.19673*.

Miles Turpin, Andy Arditi, Marvin Li, Joe Benton, and Julian Michael. 2025. Teaching models to verbalize reward hacking in chain-of-thought reasoning. In *ICML 2025 Workshop on Reliable and Responsible Foundation Models*.

Karthik Valmeekam, Kaya Stechly, Vardhan Palod, Atharva Gundawar, and Subbarao Kambhampati. 2025. Beyond semantics: The unreasonable effectiveness of reasonless intermediate tokens. *Preprint*, arXiv:2505.13775.

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. In *Workshop on Reasoning and Planning for Large Language Models*.

Changyue Wang, Weihang Su, Qingyao Ai, and Yiqun Liu. 2025a. Joint evaluation of answer and reasoning consistency for hallucination detection in large reasoning models. *CoRR*, abs/2506.04832.

Chen Wang, Zhaochun Li, Jionghao Bai, Yuzhi Zhang, Shisheng Cui, Zhou Zhao, and Yue Wang. 2025b. Arbitrary entropy policy optimization: Entropy is controllable in reinforcement fine-tuning. *CoRR*, abs/2510.08141.

Haozhe Wang, Qixin Xu, Che Liu, Junhong Wu, Fangzhen Lin, and Wenhu Chen. 2026. Emergent hierarchical reasoning in llms through reinforcement learning. *The Fourteenth International Conference on Learning Representations*.

Jiayu Wang, Yifei Ming, Zixuan Ke, Caiming Xiong, Shafiq Joty, Aws Albarghouthi, and Frederic Sala. 2025c. Beyond accuracy: Dissecting mathematical reasoning for LLMs under reinforcement learning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Kai Wang, Yihao Zhang, and Meng Sun. 2025d. When thinking llms lie: Unveiling the strategic deception in representations of reasoning models. *CoRR*, abs/2506.04909.

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. 2025e. Thoughts are all over the place: On the underthinking of o1-like llms. *CoRR*, abs/2501.18585.

Zengzhi Wang, Fan Zhou, Xuefeng Li, and Pengfei Liu. 2025f. Octothinker: Mid-training incentivizes reinforcement learning scaling. In *2nd AI for Math Workshop @ ICML 2025*.

Jake Ward, Chuqiao Lin, Constantin Venhoff, and Neel Nanda. 2025a. Reasoning-finetuning repurposes latent representations in base models. *ICML 2025 Workshop on Actionable Interpretability*.

Jake Ward, Paul M. Riechers, and Adam Shai. 2025b. Rank-1 reasoning: Minimal parameter diffs encode interpretable reasoning signals. In *Mechanistic Interpretability Workshop at NeurIPS 2025*.

Fang Wu and Yejin Choi. 2025. The invisible leash: Why rlvr may not escape its origin. In *2nd AI for Math Workshop @ ICML 2025*.

Zhen Xiong, Yujun Cai, Zhecheng Li, and Yiwei Wang. 2025. Mapping the minds of LLMs: A graph-based analysis of reasoning LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 17751–17763, Suzhou, China. Association for Computational Linguistics.

Fengli Xu, Qianyue Hao, Zefang Zong, Jingwei Wang, Yunke Zhang, Jingyi Wang, Xiaochong Lan, Jiahui Gong, Tianjian Ouyang, Fanjin Meng, Chenyang Shao, Yuwei Yan, Qinglong Yang, Yiwen Song, Sijian Ren, Xinyuan Hu, Yu Li, Jie Feng, Chen Gao, and Yong Li. 2025. Towards large reasoning models: A survey of reinforced reasoning with large language models. *CoRR*, abs/2501.09686.

Ge Yan, Chung-En Sun, and Tsui-Wei Weng. 2025. Reflctrl: Controlling LLM reflection via representation engineering. In *Mechanistic Interpretability Workshop at NeurIPS 2025*.

Shu Yang, Junchao Wu, Xin Chen, Yunze Xiao, Xinyi Yang, Derek F. Wong, and Di Wang. 2025a. Understanding aha moments: from external observations to internal mechanisms. *CoRR*, arXiv:2504.02956.

Wenkai Yang, Shuming Ma, Yankai Lin, and Furu Wei. 2025b. Towards thinking-optimal scaling of test-time compute for LLM reasoning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Xinhao Yao, Lu Yu, Xiaolin Hu, Fengwei Teng, Qing Cui, Jun Zhou, and Yong Liu. 2025a. The debate on rlvr reasoning capability boundary: Shrinkage, expansion, or both? a two-stage dynamic view. *Preprint*, arXiv:2510.04028.

Yang Yao, Xuan Tong, Ruofan Wang, Yixu Wang, Lujundong Li, Liang Liu, Yan Teng, and Yingchun Wang. 2025b. A mousetrap: Fooling large reasoning models for jailbreak with chain of iterative chaos. In *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 7837–7855. Association for Computational Linguistics.

Zijun Yao, Yantao Liu, Yanxu Chen, Jianhui Chen, Junfeng Fang, Lei Hou, Juanzi Li, and Tat-Seng Chua. 2025c. Are reasoning models more prone to hallucination? *CoRR*, abs/2505.23646.

Qingyu Yin, Chak Tou Leong, Linyi Yang, Wenxuan Huang, Wenjie Li, Xiting Wang, Jaehong Yoon, YunXing, XingYu, and Jinjin Gu. 2025. Refusal falls off a cliff: How safety alignment fails in reasoning? *CoRR*, abs/2510.06036.

Zonghao Ying, Guangyi Zheng, Yongxin Huang, Deyue Zhang, Wenxin Zhang, Quanchen Zou, Aishan Liu, Xianglong Liu, and Dacheng Tao. 2025. Towards understanding the safety boundaries of deepseek models: Evaluation and findings. *CoRR*, abs/2503.15092.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, YuYue, Weinan Dai, Tiantian Fan, Gaohong Liu, Juncai Liu, LingJun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, and 17 others. 2025. DAPO: An open-source LLM reinforcement learning system at scale. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Lifan Yuan, Weize Chen, Yuchen Zhang, Ganqu Cui, Hanbin Wang, Ziming You, Ning Ding, Zhiyuan Liu, Maosong Sun, and Hao Peng. 2026. From f(x) and g(x) to f(g(x)): Llms learn new skills in RL by composing old ones.

Cai Yuchen, Ding Cao, Xin Xu, Zijun Yao, Yuqing Huang, Benyi Zhang, Zhenyu Tan, Guiquan Liu, and Junfeng Fang. 2026. On predictability of reinforcement learning dynamics for large language models. In *The Fourteenth International Conference on Learning Representations*.

Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Yang Yue, Shiji Song, and Gao Huang. 2025. Does reinforcement learning really incentivize reasoning capacity in LLMs beyond the base model? In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Yuchen Zeng, Shuibai Zhang, Wonjun Kang, Shutong Wu, Lynnix Zou, Ying Fan, Heeju Kim, Ziqian Lin, Jungtaek Kim, Hyung Il Koo, Dimitris Papailiopoulos, and Kangwook Lee. 2025. Rejump: A tree-jump representation for analyzing and improving llm reasoning. *Preprint*, arXiv:2512.00831.

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*.

Charlie Zhang, Graham Neubig, and Xiang Yue. 2025b. On the interplay of pre-training, mid-training, and rl on reasoning language models. *Preprint*, arXiv:2512.07783.

Chong Zhang, Yue Deng, Xiang Lin, Bin Wang, Dianwen Ng, Hai Ye, Xingxuan Li, Yao Xiao, Zhanfeng Mo, Qi Zhang, and Lidong Bing. 2025c. 100 days after deepseek-r1: A survey on replication studies and more directions for reasoning language models. *CoRR*, abs/2505.00551.

Honglin Zhang, Qianyue Hao, Fengli Xu, and Yong Li. 2026a. Reinforcement learning fine-tuning enhances activation intensity and diversity in the internal circuitry of LLMs. *The Fourteenth International Conference on Learning Representations*.

Jiajie Zhang, Nianyi Lin, Lei Hou, Ling Feng, and Juanzi Li. 2025d. AdaptThink: Reasoning models can learn when to think. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 3716–3730, Suzhou, China. Association for Computational Linguistics.

Jue Zhang, Qingwei Lin, Saravan Rajmohan, and Dongmei Zhang. 2025e. From reasoning to answer: Empirical, attention-based and mechanistic insights into distilled deepseek r1 models. *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*.

Junyu Zhang, Yifan Sun, Tianang Leng, Jingyan Shen, Liu Ziyin, Paul Pu Liang, and Huan Zhang. 2025f. When reasoning meets its laws. *arXiv preprint arXiv:2512.17901*.

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. 2025g. A survey of reinforcement learning for large reasoning models. *CoRR*, abs/2509.08827.

Wenhao Zhang, Yuexiang Xie, Yuchang Sun, Yanxi Chen, Guoyin Wang, Yaliang Li, Bolin Ding, and Jingren Zhou. 2026b. On-policy RL meets off-policy experts: Harmonizing supervised fine-tuning and reinforcement learning via dynamic weighting. In *The Fourteenth International Conference on Learning Representations*.

Wenjing Zhang, Xuejiao Lei, Zhaoxiang Liu, Limin Han, Jiaojiao Zhao, Beibei Huang, Zhenhong Long, Junting Guo, Meijuan An, Rongjia Du, Ning Wang, Kai Wang, and Shiguo Lian. 2025h. Safety evaluation and enhancement of deepseek models in chinese contexts. *CoRR*, abs/2503.16529.

Haoran Zhao, Yuchen Yan, Yongliang Shen, Haolei Xu, Wenqi Zhang, Kaitao Song, Jian Shao, Weiming Lu, Jun Xiao, and Yueting Zhuang. 2025a. Let LRMs break free from overthinking via self-braking tuning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

James Xu Zhao, Bryan Hooi, and See-Kiong Ng. 2025b. Test-time scaling in reasoning models is not effective for knowledge-intensive tasks yet. *CoRR*, abs/2509.06861.

Weixiang Zhao, Xingyu Sui, Jiahe Guo, Yulin Hu, Yang Deng, Yanyan Zhao, Xuda Zhi, Yongbo Huang, Hao He, Wanxiang Che, Ting Liu, and Bing Qin. 2025c. Trade-offs in large reasoning models: An empirical analysis of deliberative and adaptive reasoning over foundational capabilities. *Preprint*, arXiv:2503.17979.

Kaiwen Zhou, Chengzhi Liu, Xuandong Zhao, Shreedhar Jangam, Jayanth Srinivasa, Gaowen Liu, Dawn Song, and Xin Eric Wang. 2025. The hidden risks of large reasoning models: A safety assessment of r1. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 3250–3265, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

Rongzhi Zhu, Yi Liu, Zequn Sun, Yiwei Wang, and Wei Hu. 2025. When can large reasoning models save thinking? mechanistic analysis of behavioral divergence in reasoning. *CoRR*, abs/2505.15276.

A Taxonomy

We present the taxonomy of our paper in Figure 2. We follow Figure 1 to organize the research of various directions and list representative works accordingly.

B Training Methods

B.1 Combine SFT with RL

Running SFT and RL as two separate steps will let the bias introduced by SFT grow too large and degrade final performance. Therefore, some studies attempt to combine the two approaches into a unified single post-training step.

Some explorations primarily focused on interleaving the SFT and RL processes and on identifying appropriate switching points between them. Based on their research into RL training dynamics, Hu et al. (2025a) proposed the Annealed-RLVR algorithm, which introduces SFT for heating when accuracy is very low to disrupt the current suboptimal state, then continues RL to perform annealing. Ma et al. (2026) observes that RL excels at easy questions while SFT is better suited to hard ones; their *ReLIFT* pipeline automatically flags the hard instances during RL, collects corresponding expert demonstrations, and inserts an SFT update once enough difficult question–answer pairs have been accumulated. *TRAPO* (Su et al., 2026) interleaves SFT and RL within every training instance and sets up a mechanism that dynamically supplies expert-guided prefixes. SFT in *TRAPO* is constrained by trust-region gradient clipping to avoid distribution-blending.

Further more, some studies aim to fuse the loss functions of RL and SFT to achieve a truly unified post-training approach. *UFT* (Liu et al., 2025b) introduces an additional log-likelihood term to the objective function of RFT(RL Fine-Tuning), allowing the model to learn from the informative supervision signal and still benefit from the generalization of RFT. *HPT* (Lv et al., 2025) defines the total loss as a weighted sum of the SFT and RL losses and dynamically adjusts the weights of SFT and RL based on real-time performance. *CHORD* (Zhang et al., 2026b) treats SFT as a dynamically-weighted auxiliary objective within the RL process and introduces a token-level weighting function that up-weights the SFT component only when the model is uncertain about the answer. Going further, *SRFT* (Fu et al., 2026) incorporates demonstration

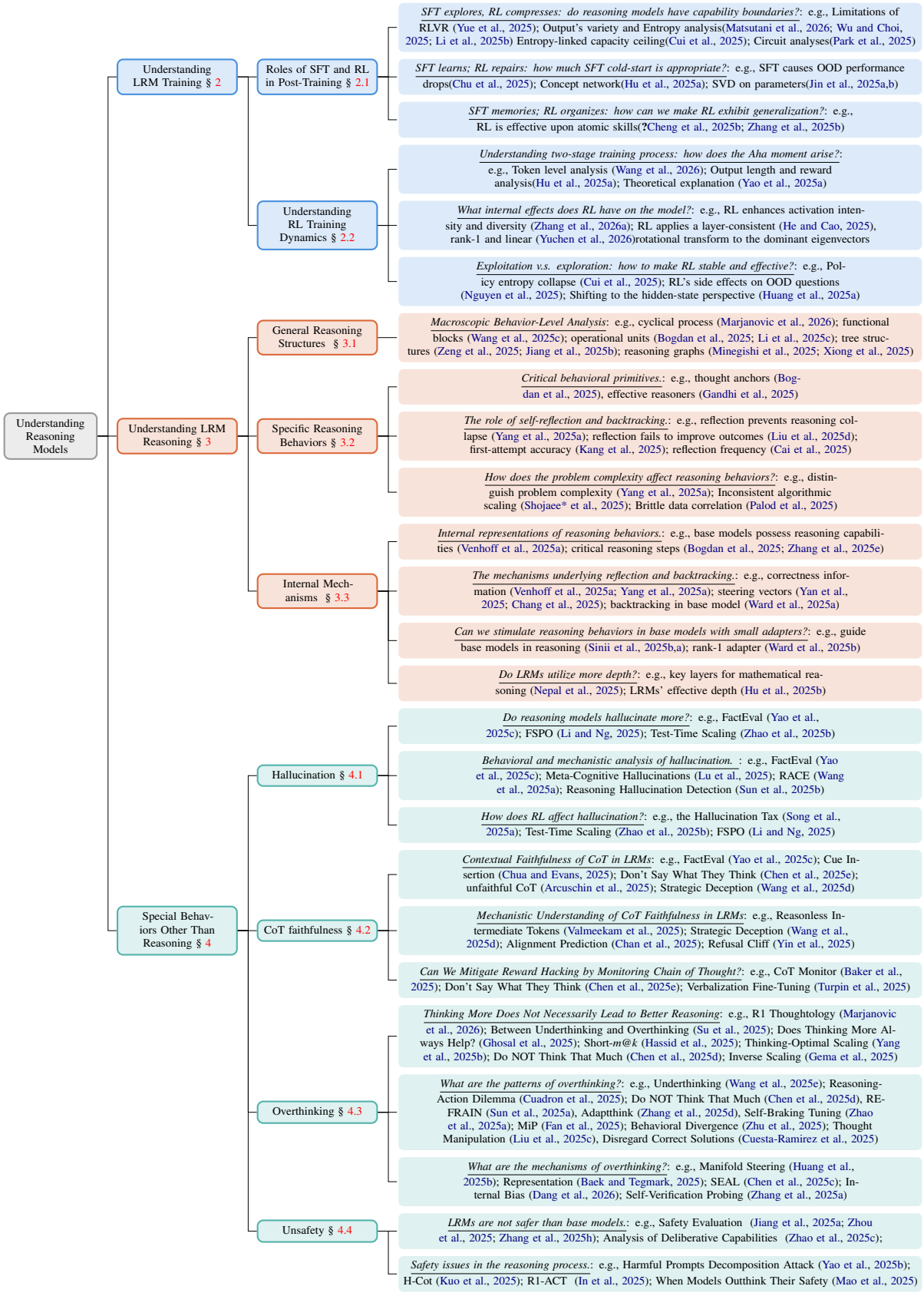

Figure 2: Taxonomy of our paper and representative works for each direction.

data into the RL training set and constructs the final loss as the entropy-weighted sum of four terms: SFT loss on demonstrations, RL loss on demonstrations, and RL loss on positive (entropy-weighted) and negative (without weighting) sampled rollouts. *BRIDGE* (Chen et al., 2025a) attaches LoRA fine-tuning blocks to the model architecture and formulates the post-training procedure as a bi-level optimization: SFT refines the model starting from the parameter optimum found by RL, optimizing solely over the LoRA weights, with an appropriate transformation eliminates the need for second-order gradients.

B.2 RL balancing exploration and exploitation

To address entropy collapse and balance exploration and exploitation for stable RL training, numerous studies have proposed solutions. *DAPO* (Yu et al., 2025) decouples the clipping bounds of PPO *into ϵ-high and ϵ-low*, raising ϵ-high to leave more head-room for boosting the probabilities of low-probability "exploratory" tokens. Cui et al. (2025) shows that the entropy change is governed by the covariance between the "action log-probabilities" and the "changes in action logits"; tokens with high covariance are the main drivers of entropy collapse. To counter this, they propose *Clip-Cov* which randomly truncates the gradients of high-covariance tokens, and *KL-Cov* which adds an extra KL-penalty to those tokens.*ProRL*(Liu et al., 2025a) adopts the Decoupled Clip technique from *DAPO*, and further equips the pipeline with KL regularization plus periodic reference-policy resets to avert entropy collapse, enabling effective RL training that continues for thousands of steps.

More recent studies have moved beyond simple clipping and experimented with additional mechanisms to further arrest entropy collapse.*CURE*(Li et al., 2025d) shows that the prevailing RLVR pipeline relies on sampling from a fixed initial state, biasing the model toward overly deterministic behavior and low diversity. They introduce a two-stage scheme to balance exploration and exploitation: in stage one they identify high-entropy tokens, truncate at those tokens, and then sample multiple continuations that are all used for updating, thereby intensifying exploration around high-entropy regions; in stage two they revert to the ordinary static-sampling *DAPO* routine. Similarly, Nguyen et al. (2025) advocates sampling problems that the base model still handles poorly, rather than repeatedly drawing those it already solves well.Song et al. (2025b) encourages historical exploration of rare answers through UCB-style rewards and fosters batch exploration at test time by penalizing duplicate answers within each batch.

Naive entropy regularization performs poorly when training reasoning models. Several works have designed regularization methods specifically tailored to reasoning models. Cheng et al. (2025a) injects a clipped, gradient-detached entropy term into the advantage function to encourage longer chains-of-thought. Wang et al. (2025b) stabilizes policy entropy by combining Policy-Gradient, Distribution, and Reinforce signals into a composite regularizer. Shen (2026); Jiang et al. (2025c) compute entropy only over the top-p tokens and adaptively rescale it for entropy regularization; the latter research further shows that regularizing those high-entropy tokens only can improve model performance.