Reasoning Beyond Language: A Comprehensive Survey on Latent Chain-of-Thought Reasoning

Xinghao Chen[1,2*], **Anhao Zhao**[2*], **Heming Xia**[1], **Xuan Lu**[2], **Hanlin Wang**[1],
Yanjun Chen[1,2], **Wei Zhang**[2], **Jian Wang**[1†], **Wenjie Li**[1], **Xiaoyu Shen**[2†]

[1]Department of Computing, The Hong Kong Polytechnic University
[2]Ningbo Digital Twin Institute, Eastern Institute of Technology, Ningbo, China
xing-hao.chen@connect.polyu.hk   plclmezboss@gmail.com
jian51.wang@polyu.edu.hk   xyshen@eitech.edu.cn

Abstract

Large Language Models (LLMs) have achieved impressive performance on complex reasoning tasks with Chain-of-Thought (CoT) prompting. However, conventional CoT relies on reasoning steps explicitly verbalized in natural language, introducing inefficiencies and limiting its applicability to abstract reasoning. To address this, there has been growing research interest in *latent* CoT reasoning, where inference occurs within latent spaces. By decoupling reasoning from language, latent reasoning promises richer cognitive representations and more flexible, faster inference. Researchers have explored various directions in this promising field, including training methodologies, structural innovations, and internal reasoning mechanisms. This paper presents a comprehensive overview and analysis of this reasoning paradigm. We begin by proposing a unified taxonomy from four perspectives: token-wise strategies, internal mechanisms, analysis, and applications. We then provide in-depth discussions and comparative analyses of representative methods, highlighting their design patterns, strengths, and open challenges. We aim to provide a structured foundation for advancing this emerging direction in LLM reasoning. [1]

1 Introduction

"Whereof one cannot speak, thereof one must be silent."
      — Ludwig Wittgenstein

Large Language Models (LLMs) have demonstrated remarkable capabilities on complex reasoning tasks (Guo et al., 2025; OpenAI, 2025; Qwen, 2025) via Chain-of-Thought (CoT) reasoning (Wei et al., 2022; Chen et al., 2025b), which encourages models to reason step-by-step through natural language. This approach not only improves
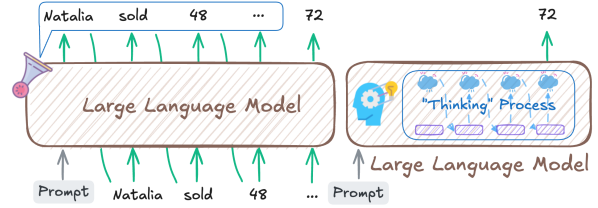

Figure 1: Explicit CoT (*left*) generates reasoning steps with natural language, while latent CoT (*right*) allows the model to reason internally in latent spaces.

interpretability but often leads to better task performance (Kojima et al., 2022; Chu et al., 2024).

Despite its utility, explicit CoT reasoning is inherently constrained by its reliance on natural language for representing each step. This linguistic mediation leads to two primary challenges. First, it introduces *computational inefficiency* (Lin et al., 2025b; Feng et al., 2025; Qu et al., 2025; Sui et al., 2025; Wang et al., 2025a; Liu et al., 2025), as not all tokens in the articulated thought process carry informative content. Secondly, human thinking often transcends the limits of language. There are other aspects of cognition—such as abstract insights, intuitive leaps, or highly compositional thoughts—that resist complete or precise verbalization (Wittgenstein, 1922; Pinker, 1994). For these tasks, as noted by Hao et al. (2024), *forcing the verbalization of every step can be not only difficult but also an unnatural constraint on the reasoning process itself.*

These inherent limitations of natural language and explicit reasoning have directly motivated a shift towards **Latent Chain-of-Thought reasoning**. As illustrated in Figure 1, models reason not through language tokens but in **latent spaces**, offering a more abstract and efficient medium for a thought-like process. This process can be viewed as "de-linguistified" reasoning, enabling richer thought representations, faster inference through compressed computation, and greater flexibility for non-verbal cognitive patterns (Lindsey et al., 2025).
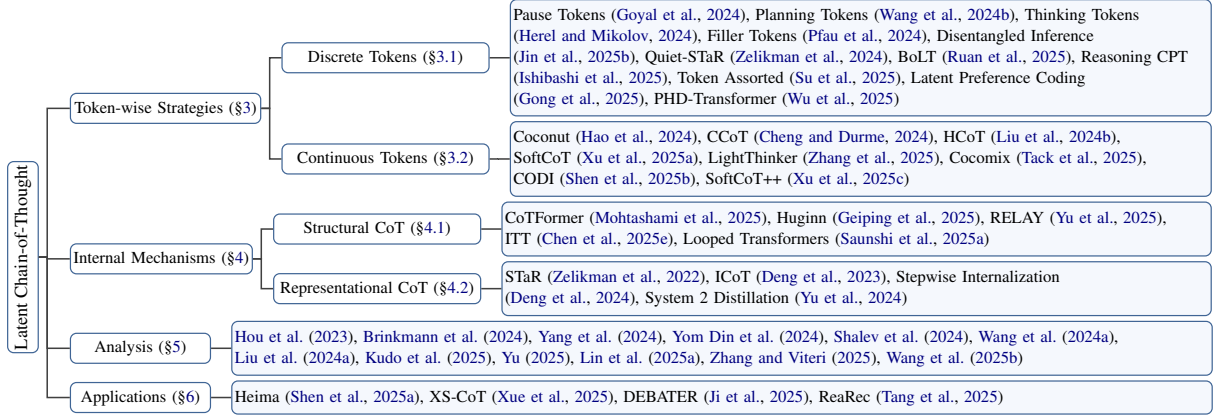
---

*Equal Contributions.
†Corresponding Authors.
[1]The relevant papers will be regularly updated at https://github.com/EIT-NLP/Awesome-Latent-CoT.

Latent Chain-of-Thought

Token-wise Strategies (§3)
- Discrete Tokens (§3.1): Pause Tokens (Goyal et al., 2024), Planning Tokens (Wang et al., 2024b), Thinking Tokens (Herel and Mikolov, 2024), Filler Tokens (Pfau et al., 2024), Disentangled Inference (Jin et al., 2025b), Quiet-STaR (Zelikman et al., 2024), BoLT (Ruan et al., 2025), Reasoning CPT (Ishibashi et al., 2025), Token Assorted (Su et al., 2025), Latent Preference Coding (Gong et al., 2025), PHD-Transformer (Wu et al., 2025)
- Continuous Tokens (§3.2): Coconut (Hao et al., 2024), CCoT (Cheng and Durme, 2024), HCoT (Liu et al., 2024b), SoftCoT (Xu et al., 2025a), LightThinker (Zhang et al., 2025), Cocomix (Tack et al., 2025), CODI (Shen et al., 2025b), SoftCoT++ (Xu et al., 2025c)

Internal Mechanisms (§4)
- Structural CoT (§4.1): CoTFormer (Mohtashami et al., 2025), Huginn (Geiping et al., 2025), RELAY (Yu et al., 2025), ITT (Chen et al., 2025e), Looped Transformers (Saunshi et al., 2025a)
- Representational CoT (§4.2): STaR (Zelikman et al., 2022), ICoT (Deng et al., 2023), Stepwise Internalization (Deng et al., 2024), System 2 Distillation (Yu et al., 2024)

Analysis (§5): Hou et al. (2023), Brinkmann et al. (2024), Yang et al. (2024), Yom Din et al. (2024), Shalev et al. (2024), Wang et al. (2024a), Liu et al. (2024a), Kudo et al. (2025), Yu (2025), Lin et al. (2025a), Zhang and Viteri (2025), Wang et al. (2025b)

Applications (§6): Heima (Shen et al., 2025a), XS-CoT (Xue et al., 2025), DEBATER (Ji et al., 2025), ReaRec (Tang et al., 2025)

Figure 2: Taxonomy of Latent Chain-of-Thought (CoT) reasoning.

Yet, latent CoT also raises critical challenges: (1) *unsupervisable processes*, as their internal reasoning processes occur in latent spaces that are not directly interpretable by humans (Lindsey et al., 2025); (2) *evaluation gaps*, with no clear metrics to distinguish deep latent reasoning from input-output shortcuts (Ameisen et al., 2025); and (3) *alignment risks*, where the inability to inspect or constrain latent trajectories complicates ethical control (Xu et al., 2025b; Ruan et al., 2025).

Despite these open questions, the rapid yet fragmented development of latent reasoning research highlights the pressing need for a clear and structured understanding within the research community. In this work, we present the first comprehensive survey of latent Chain-of-Thought reasoning. Our key contributions are threefold: (1) **Systematic taxonomy**: We introduce a structured taxonomy of latent CoT research, dividing existing work into four distinct categories. Within each, we organize representative studies into a coherent framework that clarifies their methodological assumptions and innovations (as illustrated in Figure 2); (2) **In-depth analysis**: Building on this taxonomy, we conduct a comprehensive analysis of representative works in each category, comparing training strategies, design paradigms, supervision signals, and efficiency trade-offs; and (3) **Challenge identification and research frontiers**: We identify critical open problems and outline promising directions for future research.

We aim to consolidate the fragmented landscape of latent reasoning and facilitate future developments in this emerging direction.

2 Overview

This paper presents a comprehensive survey of latent CoT reasoning in LLMs. We begin by examining **methodological advances**, which fall into two major categories: *Token-wise strategies* (§3), including both *discrete tokens* (§3.1) and *continuous tokens* (§3.2); and *Internal mechanisms* (§4), which divide into *structural* and *representational* forms. Beyond design mechanisms, we review a growing body of work on the **analysis and interpretability** of latent reasoning (§5). Finally, we discuss real-world applications (§6), challenges and future directions (§7).

3 Token-wise Strategies

While explicit CoT has significantly enhanced the reasoning capabilities of LLMs by generating reasoning steps, it often increases computational costs and inference latency. To mitigate these limitations and further extend the expressive capacity of reasoning models, recent work has explored the use of *token-wise strategies*, which are designed not only to streamline reasoning but also to unlock more abstract and compact cognitive processes. We categorize these external tokens into two primary types: **Discrete Tokens**, which are symbolic, and often serve as explicit control cues; and **Continuous Tokens**, which are learned embeddings in latent spaces and facilitate implicit reasoning.

3.1 Discrete Tokens

Discrete tokens, which serve as symbolic representations of intermediate reasoning steps or cognitive operations, have emerged as a promising paradigm for enhancing the reasoning capabilities of LLMs. They significantly contribute to improved task performance and greater efficiency.

Early studies in exploring discrete tokens introduced simple markers such as "[pause]" or ellipses ("...") to segment reasoning steps, which has significantly improved multi-step task performance

(Pfau et al. (2024), Herel and Mikolov (2024)). Prior to these efforts, Goyal et al. (2024) proposed adaptive and learnable "pause tokens," which dynamically allocate computational resources. These tokens enable delayed prediction, allowing models to perform additional internal computation before generating outputs, thereby enhancing accuracy for logic-intensive tasks. Beyond these pioneering exploration, researchers developed more sophisticated tokens to encode complex reasoning structures. For example, Wang et al. (2024b) introduced "planning tokens" derived from heuristics or variational autoencoders (VAEs) to improve coherence and precision in reasoning. To disentangle cognitive processes and enhance interpretability, Jin et al. (2025b) proposed specialized tokens such as "memory" and "reason", which modularize reasoning by isolating specific cognitive operations.

To further advance modularized reasoning, Zelikman et al. (2024) introduced *Quiet-STaR*, a method that uses learnable tokens to mark the boundaries of internal rationales. This approach enables language models to infer unstated reasoning steps, leading to improved generalization on challenging tasks without requiring task-specific fine-tuning. Building on this foundation, Ruan et al. (2025) proposed *BoLT*, which models the thought process as a trainable latent variable. This innovation allows models to infer and refine sequences of cognitive steps during pretraining, enhancing their ability to tackle complex reasoning tasks. Ishibashi et al. (2025) expanded on *BoLT* by introducing continual pretraining (CPT) with synthetic data containing hidden thought processes. Their reasoning CPT framework reconstructed the implicit cognitive steps underlying texts, significantly improving reasoning across diverse domains. These advancements were particularly impactful in specialized areas such as STEM and law, demonstrating notable performance gains on challenging tasks and showcasing the transferability of reasoning skills across domains.

Pfau et al. (2024) pointed out that the structural organization of tokens is more critical than their semantic content. Surprisingly, replacing meaningful tokens with neutral placeholders yields negligible performance loss, underscoring the importance of token structure. Inspired by this finding, compression-based approaches have emerged to address computational inefficiencies. For example, Su et al. (2025) employed vector-quantized VAEs (VQ-VAEs) to condense reasoning steps into dis-

crete latent tokens, reducing computational costs while maintaining performance. To further enhance token-based frameworks, Gong et al. (2025) extended this compression-based strategy to preference modeling, leveraging a learnable codebook of latent codes to align reasoning outputs with human expectations. The Parallel Hidden Decoding Transformer (PHD-Transformer) series introduced a pivotal innovation by utilizing hidden decoding tokens for efficient length scaling (Wu et al., 2025). This method achieves deeper reasoning and better task performance without increasing the size of the key-value (KV) cache, addressing long-context reasoning and enhancing the utility of discrete tokens.

Overall, discrete tokens have progressed from simple markers to versatile tools for abstract cognitive modeling. They serve as powerful mechanisms that advances LLM reasoning capabilities, improving both efficiency and interpretability.

3.2 Continuous Tokens

In contrast to discrete tokens, a growing body of research investigates latent reasoning through continuous representations, where reasoning processes are modeled as trajectories within high-dimensional embedding spaces rather than explicit textual sequences. This shift reflects a significant transition from *hard, discrete tokens* to *soft, continuous tokens*, offering more flexible and compact representations of intermediate reasoning states. We categorize existing methods based on whether the latent reasoning is integrated during *post-training* or *pre-training*.

Post-training methods offer an efficient way to equip LLMs with latent reasoning capabilities using minimal additional data. Based on whether an LLM both generates the final output and is responsible for producing and consuming the continuous tokens, we categorize existing methods into two types: 1) Intrinsic methods keep the whole pipeline inside a single LLM; and 2) Auxiliary methods introduce a separate module that generates continuous tokens, which are then injected into the main model. Both methods aim to address the key question: *how can we guide continuous tokens toward the correct reasoning direction?* Figure 3 provides a comparative illustration of these approaches.

Among intrinsic methods, *COCONUT* (Hao et al., 2024) made pioneering efforts to enable internal reasoning by feeding the model's last hidden states into its next input embedding, effectively allowing for latent iteration without producing ex-
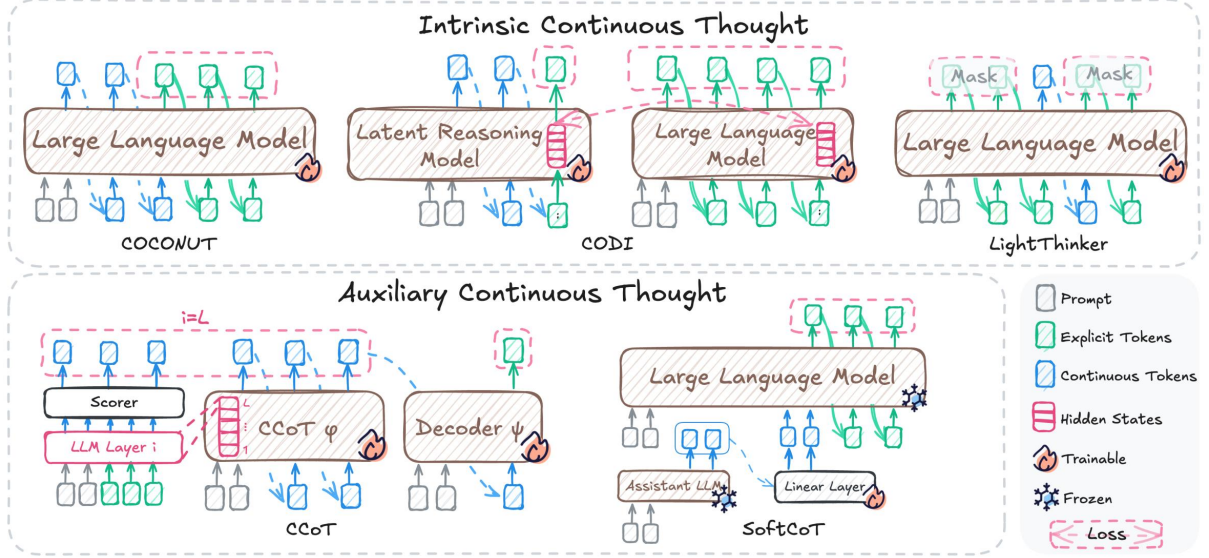
Figure 3: llustration of representative *Continuous Tokens*-based methods. <u>Intrinsic</u> methods generate and consume continuous tokens within a single LLM. <u>Auxiliary</u> methods use external modules to generate continuous tokens.

plicit rationales. This recurrent reuse of internal states supports breadth-first exploration and improves efficiency. To improve the semantic directionality of these latent trajectories, *CODI* (Shen et al., 2025b) introduced a self-distillation loss to force the hidden activations of the specific position token of the student model to mimic the teacher model's hidden activations under explicit CoT supervision. *LightThinker* (Zhang et al., 2025) trained the model to decide *when* and *how* to compress reasoning into latent "gist" tokens, using strategically placed masking to reduce KV cache usage. These studies show that *intrinsic latent representations can elicit viable reasoning behavior*. The addition of structural priors or alignment objectives significantly stabilizes learning and improves generalization, demonstrating that *internal trajectories benefit from consistent directional guidance*.

Among auxiliary methods, *HCoT* (Liu et al., 2024b) trained a dedicated auxiliary CoT model to generate and compress the full thought process into a compact special token representation, which was then passed to the main model as input for answer generation. Following a similar process, *CCoT* (Cheng and Durme, 2024) encoded complete reasoning sequences into variable-length latent embeddings using a trained *CCoT* model φ, replacing explicit chains with dense, semantically rich contemplation tokens. The contemplation tokens were supervised to match a subset of hidden states precomputed from concatenated input. A subset was selected via a scorer, and subsequently fed into a trained decoder ψ to generate final an-

swers. To reduce training cost and ensure stability and generalization across different domains, *SoftCoT* (Xu et al., 2025a) combined a frozen assistant model with a trained projection layer to generate "soft tokens" that plug directly into a frozen LLM. *SoftCoT++* (Xu et al., 2025c) extended SoftCoT to the test-time scaling paradigm by enabling diverse explorations in the continuous space. SoftCoT++ perturbs the latent space using multiple specialized initial tokens, and applies contrastive learning to promote diversity among soft thoughts.

While post-training methods consistently yield improvements in efficiency, reducing token usage and latency, their reasoning performance often matches, rather than exceeds, that of explicit CoT prompting on standard benchmarks. The ceiling suggests that, without deeper objectives that sculpt latent trajectories, continuous-token reasoning may continue to lean on capabilities learnt in text space.

Pre-training methods take a step further by embedding latent reasoning directly into the model's cognitive prior during the pre-training phase. Rather than treating reasoning as a generative process, these methods model it as an internalizable, optimizable process within the latent space of representations.

CoCoMix (Tack et al., 2025) introduced this idea by mixing continuous, high-level "concepts" into the model's hidden states during pre-training. These concepts were extracted using a sparse autoencoder trained on the activations of a pretrained model and selected based on their causal influence on the next-token prediction. CoCoMix enhanced

LLMs by interleaving predicted concepts alongside token embeddings, creating a latent scaffold that improves both performance and interpretability. Unlike post-training strategies that treat latent reasoning as a side effect, pre-training embeds it as a native cognitive faculty, potentially yielding more generalizable and cognitively aligned models.

4 Internal Mechanisms

Recent research has explored the internal computational mechanisms that underlie reasoning within LLMs. These internal mechanisms focus on how reasoning can emerge implicitly through internal architectures and representations, without relying on explicit token-level traces. We categorize this line of work into two main directions: (1) **Structural CoT**, which examines how architectural depth, recurrence, and looping computations support latent reasoning; and (2) **Representational CoT**, which explores how intermediate reasoning processes can be embedded directly into the model's hidden states, without requiring explicit intermediate outputs.

4.1 Structural CoT

Given the impressive reasoning capabilities exhibited by LLMs, recent work has attempted to investigate the scaling laws specific to reasoning tasks. Ye et al. (2025) suggested that scaling laws for reasoning were more nuanced than previously understood, with the model depth playing a critical role alongside parameters. At a fixed parameter budget, deeper—but—narrower models tend to outperform wider counterparts. This challenged the conventional wisdom of scaling laws, yet aligns with intuitive reasoning: the success of test-time scaling closely resembles shared-weight strategies (Lan et al., 2020; Dehghani et al., 2019), where reusing the same layers across multiple tokens effectively constructs deeper computational graphs. Further empirical evidence reinforced the importance of depth in reasoning. For example, Chen and Zou (2024) found that a minimal depth was a necessary condition for the emergence of CoT reasoning. While increasing depth presents a promising approach to enhancing reasoning, by enabling iterative refinement of latent representations, the continual addition of layers imposes substantial computational and memory overheads, thereby limiting scalability in practice.

Inspired by evidence from recurrent architectures in the "deep thinking" literature (Schwarzschild et al., 2021; McLeish and Tran-Thanh, 2023), which demonstrated inherent advantages in learning complex, iterative algorithms, recent research has shifted toward exploring recurrent methodologies for efficient latent reasoning, as illustrated in Figure 4. As an early exploration in this direction, Mohtashami et al. (2025) introduced *CoTFormer*, which emulates CoT reasoning by interleaving and looping representations. This approach maintains computational efficiency while mimicking the step-wise nature of human reasoning. To enable arbitrary computational depth at test time, Geiping et al. (2025) proposed *Huginn*, a novel recurrent framework that dynamically allocates resources through RNN-like iterative computations. Huginn achieves performance comparable to larger, static-depth models but with improved efficiency. Building upon the length generalization capability of looped architectures, Yu et al. (2025) proposed *RELAY*, which explicitly aligns CoT reasoning steps with loop iterations in a Looped Transformer. Intermediate supervision is applied during training to guide reasoning across steps, and the resulting reasoning chains are used to fine-tune an autoregressive model, enhancing performance on tasks that exceed training sequence lengths. To further improve reasoning on critical tokens, Chen et al. (2025e) introduced the Inner Thinking Transformer (*ITT*), where each Transformer layer is treated as a discrete reasoning step. By incorporating adaptive token routing
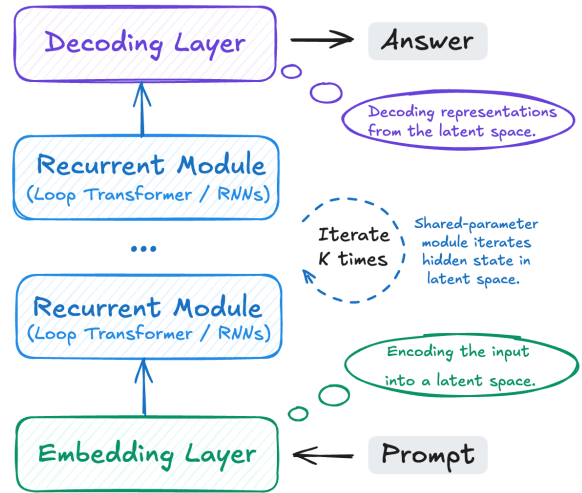

Figure 4: Illustration of structural CoT mechanisms, where latent reasoning emerges through iterative refinement of the hidden state via a recurrent module. Existing work commonly interprets **each recurrence as a discrete reasoning step** in the CoT.

and residual refinement, ITT dynamically allocates computation across tokens, achieving strong reasoning capabilities with fewer parameters and less training data. Finally, Saunshi et al. (2025b) empirically showed that deepening via recurrence, rather than increasing parameter count, can significantly enhance reasoning ability, reinforcing the trend toward recurrent strategies for latent reasoning.

These studies validate the potential of increased depth, achieved either through stacking or shared-weight mechanisms, to effectively support latent-space reasoning. This line of thinking drives research toward more computationally efficient ways that harness depth for reasoning-intensive tasks.

4.2 Representational CoT

In addition to the exploration of depth-driven reasoning, another promising avenue involves internalizing explicit CoT directly into the latent representations of LLMs. Early implementations of representational internalized CoT utilized rationale-augmented fine-tuning strategies, explicitly teaching models to predict intermediate reasoning outcomes without generating textual outputs (Zelikman et al., 2022). Subsequent advancements further refined this approach through sophisticated knowledge distillation methods, training student models to emulate hidden-state reasoning trajectories exhibited by teacher models performing explicit CoT (Deng et al., 2023). Additionally, phased fine-tuning paradigms (Deng et al., 2024) and self-distillation frameworks (Yu et al., 2024) enable LLMs to implicitly internalize complex reasoning pathways within their latent representations without explicitly articulating intermediate reasoning steps. Overall, this line of work shows that *it is effective to condense reasoning processes into compact and computationally efficient latent structures*.

In summary, structural and representational approaches offer two complementary pathways for internalizing reasoning within LLMs. Structural methods leverage architectural depth (such as via stacking, recurrence, or weight sharing) to support iterative computation, effectively simulating multi-step reasoning in a layer-wise manner. In contrast, representational methods encode reasoning processes directly within hidden states, enabling models to perform inference without emitting explicit intermediate steps. Together, these approaches underscore the dual importance of computational structure and internal representation in achieving efficient and powerful latent CoT reasoning.

5 Analysis and Interpretability

Since latent CoT decouples reasoning from explicit linguistic traces, it naturally raises the question: *do LLMs internally simulate step-by-step reasoning, or do they rely on shallow heuristics that only approximate such behavior?* This has encouraged analytical studies from various perspectives, including interpreting internal computation as evidence of structured reasoning, identifying shortcut mechanisms, and analyzing latent reasoning dynamics.

5.1 Internal Computation Interpretation

Several studies posit that LLMs can carry out multi-step reasoning implicitly within their hidden states, even without explicit CoT prompts is provided. These works attempt to uncover internal structures indicative of decompositional processes. Hou et al. (2023) recovered reasoning trees from attention patterns, revealing distributed latent inference across transformer layers. Brinkmann et al. (2024) dissected a transformer trained on symbolic logic tasks and revealed an emergent recurrent computation mechanism: the model reuses internal representations across depth to simulate iterative reasoning, despite lacking explicit recurrence in its architecture. Shalev et al. (2024) showed that hidden states simultaneously encode multiple intermediate reasoning paths, indicating parallel evaluation of latent inference options. Wang et al. (2024a) showed that grokked transformers shift from memorization to generalizable algorithmic patterns, forming implicit reasoning circuits that simulate step-by-step inference without explicit CoT, even in shallow models. Yang et al. (2024) demonstrated that LLMs can retrieve intermediate bridge facts without being prompted, providing behavioral evidence of latent multi-hop reasoning. All these findings support the view that reasoning can be internally enacted without the need for external verbalization.

5.2 Shortcut Mechanisms

A line of research argues that correct outputs may result not from latent reasoning, but from shortcut strategies acquired during pre-training. These studies highlight cases where models succeed by exploiting surface-level correlations or pattern completion, rather than engaging in true inference. Yom Din et al. (2024) demonstrated that final answers were often linearly decodable from early hidden layers via the logit lens, implying that later computations may simply rephrase an already-

available result. This challenges the assumption that depth corresponds to incremental reasoning. Liu et al. (2024a) showed that LLMs can learn expert-like shortcuts by skipping intermediate reasoning steps. Lin et al. (2025a) identified that reliance on token-level spurious associations, revealing fragile positional heuristics rather than compositional inference. Yu (2025) indicated LLMs dynamically alternate between shortcut mechanisms and latent multi-step reasoning depending on task complexity. These studies caution against interpreting accurate outputs as evidence of genuine reasoning. Instead, they highlight how shortcut mechanisms—rooted in superficial correlations and positional heuristics—can produce seemingly coherent answers without underlying inference, underscoring the importance of identifying when such shortcuts are at play.

5.3 Latent Reasoning Dynamics

Bridging the two perspectives above, recent work has focused on representational analysis and controlled interventions to better characterize and steer latent reasoning dynamics. Kudo et al. (2025) used causal interventions to identify mixed reasoning strategies, showing that simple answers are computed prior to explicit reasoning, whereas harder tasks trigger active step-by-step inference. Zhang and Viteri (2025) discovered a latent CoT vector—an activation-space direction—that, when added to internal states, elicits CoT behavior without explicit prompts, revealing latent CoT as an internally accessible processing mode. Complementing this, Wang et al. (2025b) proposed CoE, a representation of hidden-state trajectories during reasoning, identifying distinct patterns linked to reasoning success that enable latent self-evaluation. Overall, latent reasoning leaves measurable traces in the activation space and may be controllable or interpretable through geometric and dynamic analysis, offering new avenues for understanding and harnessing latent CoT reasoning.

6 Applications

Latent CoT reasoning has been successfully applied in many domains due to its reasoning efficiency. Below, we discuss representative applications of latent CoT reasoning.

Textual Reasoning. Existing latent CoT methods have been systematically evaluated on natural-language reasoning tasks, including mathematical reasoning (Cobbe et al., 2021; Deng et al., 2023; Hendrycks et al., 2021b; Miao et al., 2020; Patel et al., 2021; Ling et al., 2017), general commonsense reasoning (Talmor et al., 2019; Suzgun et al., 2023; Rein et al., 2024; Hendrycks et al., 2021a), and logical multi-hop reasoning datasets (Yang et al., 2018; Geva et al., 2021; Saparov and He, 2023; Hao et al., 2024). However, latent reasoning methods have yet to be evaluated on several high-bar reasoning benchmarks that have become standard for assessing Large Reasoning Models (MAA, 2024), and code-centric datasets (Jimenez et al., 2024; Jain et al., 2025). Moreover, there remains a lack of benchmarks that are both aligned with real-world applications and specifically designed to showcase the advantages of latent reasoning.

Multimodal Reasoning and Generation. Latent reasoning has recently been extended to multimodal domains, where generating step-by-step explanations in natural language becomes both inefficient and semantically brittle. *Heima* (Shen et al., 2025a) introduces compact latent "thinking tokens" that summarize intermediate reasoning steps during multimodal tasks, cutting generation cost without hurting accuracy; *XS-CoT* (Xue et al., 2025) hides cross-lingual speech reasoning inside a semi-implicit token schedule that speeds non-core-language responses; and *LatentLM* (Sun et al., 2024) treats every modality as just another latent token, enabling a truly unified generative interface. They suggest that latent CoT reasoning is no longer confined to text. As modalities proliferate, the ability to steer and edit these hidden trajectories may become the key to controllable, efficient multimodal intelligence.

Retrieval-Augmented Generation and Recommendation. Recent work (Chen et al., 2025a; Song et al., 2025; Jin et al., 2025a) has integrated explicit reasoning mechanisms within Retrieval-Augmented Generation (RAG) frameworks, and compressing these retrieval–reasoning steps in latent space could further cut tokens and latency. Recent work on pluggable virtual tokens for RAG (Zhu et al., 2024) suggests that latent tokens can serve as lightweight carriers of external knowledge and implicit reasoning. DEBATER (Ji et al., 2025) incorporates a Chain-of-Deliberation (*CoD*) mechanism into dense retrieval. CoD introduces a sequence of prompt tokens to stimulate the latent reasoning capability of LLMs during document representation. It further employs self-

distillation to integrate multiple reasoning steps into a unified embedding. In the recommendation area, *ReaRec* (Tang et al., 2025) leverages latent reasoning to enhance user interest modeling, which recursively feeds the final hidden state of a user behavior back into the network for multiple rounds, using special positional embeddings to distinguish between original behavioral inputs and internal reasoning steps.

7 Challenges and Future Directions

In this section, we highlight key obstacles that hinder the full realization of latent reasoning's potential and outline critical areas for future research.

7.1 Challenges

Training Difficulties Despite its efficiency and inference speed, current latent reasoning methods still underperform explicit reasoning approaches in accuracy and problem-solving capability. This gap may stem from the difficulty of training, as current training methods typically optimize for explicit reasoning outputs, rather than directly supervising latent reasoning processes. There remains a key challenge in developing training methods that can fully activate LLMs' internal reasoning capabilities.

Generalization Issues The training methods for implicit reasoning demonstrate stability primarily on fixed patterns but exhibit poor generalization capabilities. Models trained with latent space reasoning techniques often struggle when faced with novel problem structures or reasoning patterns not encountered during training (Lin et al., 2025a). This fragility suggests that current approaches to latent reasoning may be learning to compress specific reasoning templates rather than developing truly flexible reasoning capabilities in abstract space.

Interpretability Concerns Recent studies suggest that models often perform reasoning in their "heads" that is not reflected in their verbalized CoTs, raising concerns about unfaithful or hidden internal processes (Chen et al., 2025d; Lindsey et al., 2025). The shift from explicit to implicit reasoning further introduces significant challenges for identifying errors and understanding how the model draws a particular conclusion.

7.2 Future Directions

To effectively advance latent reasoning, several promising directions merit exploration: (1) **Alter-native Architectures.** These may play a crucial role in enhancing the expressiveness and efficiency of latent reasoning. Beyond conventional Transformers, recurrent or looped Transformer variants, such as recurrent or looped Transformers (Saunshi et al., 2025c) enable reasoning through parameter reuse across multiple steps. In multimodal domains, diffusion model-based architectures present compelling alternatives, potentially due to their ability to model global dependencies and non-sequential reasoning in a parallel, noise-aware manner. Recent work has successfully demonstrated the effectiveness of the integration of diffusion models and latent CoT (Ye et al., 2024; Huang et al., 2025). (2) **Interpretability and Verification.** These are critical concerns that warrant further exploration in latent reasoning. Developing methods to probe, decode, or verify these latent representations is crucial for improving transparency and calibrating reasoning behavior (Chen et al., 2025c). (3) **Training Approaches.** Most existing training methods are insufficient to effectively shape latent reasoning capabilities. Reinforcement learning provides a promising paradigm for exploring the potential of LLMs to develop latent reasoning through self-evolution (Guo et al., 2025), using reward signals to implicitly sculpt a structured reasoning space aligned with task objectives. In addition, curriculum learning enables models to gradually acquire increasingly abstract reasoning skills via a simple-to-complex training process. (4) **LLM Agents.** These may benefit significantly from latent CoT reasoning, particularly in terms of inference efficiency. These agents often generate lengthy and verbose reasoning sequences, introducing substantial computational overhead (Zhou et al., 2025; Li et al., 2024; Zhang et al., 2024). With latent CoT reasoning, these agents are expected to perform more compact and faster planning and decision-making. (5) **Social Intelligence and Theory of Mind.** Latent reasoning provides a natural substrate for modeling nested mental states essential to *Theory of Mind*—the capacity to infer others' beliefs, desires, and intentions (Ma et al., 2023). Embedding latent belief modeling into reasoning pipelines could offer a scalable path toward socially competent AI.

8 Conclusion

This paper presents a comprehensive survey of latent CoT reasoning with LLMs. By moving reasoning beyond surface-level language into the latent

space, latent CoT reasoning enables more abstract, efficient, and scalable inference. We summarize the key methods, identify major challenges, and highlight promising future directions. We hope this survey serves as a foundation and offers valuable insights to support further exploration in this emerging field.

Limitations

This survey offers a comprehensive review of existing methodologies and analyses in the emerging field of latent reasoning with LLMs. However, due to the breadth and rapid evolution of related work, particularly in the areas of interpretability, internal analysis, and alignment, we may have inadvertently omitted other valuable contributions. We outline several promising future directions, including alternative architectures, training paradigms, LLM agents, and Theory-of-Mind modeling, which we highlight as areas for continued exploration. Additionally, as many surveyed works rely on small-scale models or limited benchmarks, there is a need for more up-to-date and rigorous empirical validation. We advocate for continued, in-depth research to provide practitioners with actionable and robust insights into the design and deployment of latent reasoning models.

Ethics Statement

This survey is based entirely on publicly available research papers, models, and datasets. All referenced works are properly cited and used in accordance with their respective licenses and intended purposes. While latent reasoning introduces novel challenges in interpretability and alignment, this survey aims to provide a neutral, structured overview of the field without promoting specific deployments. We emphasize the importance of future work addressing fairness, safety, and transparency in latent reasoning.

References

Emmanuel Ameisen, Jack Lindsey, Adam Pearce, Wes Gurnee, Nicholas L. Turner, Brian Chen, Craig Citro, David Abrahams, Shan Carter, Basil Hosmer, Jonathan Marcus, Michael Sklar, Adly Templeton, Trenton Bricken, Callum McDougall, Hoagy Cunningham, Thomas Henighan, Adam Jermyn, Andy Jones, Andrew Persic, Zhenyi Qi, T. Ben Thompson, Sam Zimmerman, Kelley Rivoire, Thomas Conerly, Chris Olah, and Joshua Batson. 2025. Circuit tracing: Revealing computational graphs in language models. Transformer Circuits Thread.

Jannik Brinkmann, Abhay Sheshadri, Victor Levoso, Paul Swoboda, and Christian Bartelt. 2024. A mechanistic analysis of a transformer trained on a symbolic multi-step reasoning task. In Findings of the Association for Computational Linguistics: ACL 2024, pages 4082–4102, Bangkok, Thailand. Association for Computational Linguistics.

Mingyang Chen, Tianpeng Li, Haoze Sun, Yijie Zhou, Chenzheng Zhu, Haofen Wang, Jeff Z. Pan, Wen Zhang, Huajun Chen, Fan Yang, Zenan Zhou, and Weipeng Chen. 2025a. Research: Learning to reason with search for llms via reinforcement learning. Preprint, arXiv:2503.19470.

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. Preprint, arXiv:2503.09567.

Runjin Chen, Zhenyu Zhang, Junyuan Hong, Souvik Kundu, and Zhangyang Wang. 2025c. Seal: Steerable reasoning calibration of large language models for free. Preprint, arXiv:2504.07986.

Xingwu Chen and Difan Zou. 2024. What can transformer learn with varying depth? case studies on sequence learning tasks. In Proceedings of the 41st International Conference on Machine Learning, ICML'24. JMLR.org.

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

Yilong Chen, Junyuan Shang, Zhenyu Zhang, Yanxi Xie, Jiawei Sheng, Tingwen Liu, Shuohuan Wang, Yu Sun, Hua Wu, and Haifeng Wang. 2025e. Inner thinking transformer: Leveraging dynamic depth scaling to foster adaptive internal thinking. Preprint, arXiv:2502.13842.

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

Zheng Chu, Jingchang Chen, Qianglong Chen, Weijiang Yu, Tao He, Haotian Wang, Weihua Peng, Ming Liu, Bing Qin, and Ting Liu. 2024. Navigate through enigmatic labyrinth a survey of chain of thought reasoning: Advances, frontiers and future. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1173–1203, Bangkok, Thailand. Association for Computational Linguistics.

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

Mostafa Dehghani, Stephan Gouws, Oriol Vinyals, Jakob Uszkoreit, and Lukasz Kaiser. 2019. Universal transformers. In International Conference on Learning Representations.

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

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

Sicheng Feng, Gongfan Fang, Xinyin Ma, and Xinchao Wang. 2025. Efficient reasoning models: A survey. Preprint, arXiv:2504.10903.

Jonas Geiping, Sean McLeish, Neel Jain, John Kirchenbauer, Siddharth Singh, Brian R. Bartoldson, Bhavya Kailkhura, Abhinav Bhatele, and Tom Goldstein. 2025. Scaling up test-time compute with latent reasoning: A recurrent depth approach. Preprint, arXiv:2502.05171.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. Transactions of the Association for Computational Linguistics, 9:346–361.

Zhuocheng Gong, Jian Guan, Wei Wu, Huishuai Zhang, and Dongyan Zhao. 2025. Latent preference coding: Aligning large language models via discrete latent codes. Preprint, arXiv:2505.04993.

Sachin Goyal, Ziwei Ji, Ankit Singh Rawat, Aditya Krishna Menon, Sanjiv Kumar, and Vaishnavh Nagarajan. 2024. Think before you speak: Training language models with pause tokens. In The Twelfth International Conference on Learning Representations.

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

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

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021a. Measuring massive multitask language understanding. In International Conference on Learning Representations.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021b. Measuring mathematical problem solving with the MATH dataset. In Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2).

David Herel and Tomas Mikolov. 2024. Thinking tokens for language modeling. Preprint, arXiv:2405.08644.

Yifan Hou, Jiaoda Li, Yu Fei, Alessandro Stolfo, Wangchunshu Zhou, Guangtao Zeng, Antoine Bosselut, and Mrinmaya Sachan. 2023. Towards a mechanistic interpretation of multi-step reasoning capabilities of language models. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 4902–4919, Singapore. Association for Computational Linguistics.

Zemin Huang, Zhiyang Chen, Zijun Wang, Tiancheng Li, and Guo-Jun Qi. 2025. Reinforcing the diffusion chain of lateral thought with diffusion language models. Preprint, arXiv:2505.10446.

Yoichi Ishibashi, Taro Yano, and Masafumi Oyamada. 2025. Mining hidden thoughts from texts: Evaluating continual pretraining with synthetic data for llm reasoning. Preprint, arXiv:2505.10182.

Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. 2025. Livecodebench: Holistic and contamination free evaluation of large language models for code. In The Thirteenth International Conference on Learning Representations.

Yifan Ji, Zhipeng Xu, Zhenghao Liu, Yukun Yan, Shi Yu, Yishan Li, Zhiyuan Liu, Yu Gu, Ge Yu, and Maosong Sun. 2025. Learning more effective representations for dense retrieval through deliberate thinking before search. Preprint, arXiv:2502.12974.

Carlos E Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik R Narasimhan. 2024. SWE-bench: Can language models resolve real-world github issues? In The Twelfth International Conference on Learning Representations.

Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and Jiawei Han. 2025a. Search-r1: Training llms to reason and leverage search engines with reinforcement learning. Preprint, arXiv:2503.09516.

Mingyu Jin, Weidi Luo, Sitao Cheng, Xinyi Wang, Wenyue Hua, Ruixiang Tang, William Yang Wang, and Yongfeng Zhang. 2025b. Disentangling memory and reasoning ability in large language models. Preprint, arXiv:2411.13504.

Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In Advances in Neural Information Processing Systems, volume 35, pages 22199–22213. Curran Associates, Inc.

Keito Kudo, Yoichi Aoki, Tatsuki Kuribayashi, Shusaku Sone, Masaya Taniguchi, Ana Brassard, Keisuke Sakaguchi, and Kentaro Inui. 2025. Think-to-talk or talk-to-think? when llms come up with an answer in multi-step arithmetic reasoning. Preprint, arXiv:2412.01113.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations. In International Conference on Learning Representations.

Yuanchun Li, Hao Wen, Weijun Wang, Xiangyu Li, Yizhen Yuan, Guohong Liu, Jiacheng Liu, Wenxing Xu, Xiang Wang, Yi Sun, Rui Kong, Yile Wang, Hanfei Geng, Jian Luan, Xuefeng Jin, Zilong Ye, Guanjing Xiong, Fan Zhang, Xiang Li, Mengwei Xu, Zhijun Li, Peng Li, Yang Liu, Ya-Qin Zhang, and Yunxin Liu. 2024. Personal llm agents: Insights and survey about the capability, efficiency and security. Preprint, arXiv:2401.05459.

Tianhe Lin, Jian Xie, Siyu Yuan, and Deqing Yang. 2025a. Implicit reasoning in transformers is reasoning through shortcuts. Preprint, arXiv:2503.07604.

Zicheng Lin, Tian Liang, Jiahao Xu, Qiuzhi Lin, Xing Wang, Ruilin Luo, Chufan Shi, Siheng Li, Yujiu Yang, and Zhaopeng Tu. 2025b. Critical tokens matter: Token-level contrastive estimation enhances llm's reasoning capability. Preprint, arXiv:2411.19943.

Jack Lindsey, Wes Gurnee, Emmanuel Ameisen, Brian Chen, Adam Pearce, Nicholas L. Turner, Craig Citro, David Abrahams, Shan Carter, Basil Hosmer, Jonathan Marcus, Michael Sklar, Adly Templeton, Trenton Bricken, Callum McDougall, Hoagy Cunningham, Thomas Henighan, Adam Jermyn, Andy Jones, Andrew Persic, Zhenyi Qi, T. Ben Thompson, Sam Zimmerman, Kelley Rivoire, Thomas Conerly, Chris Olah, and Joshua Batson. 2025. On the biology of a large language model. Transformer Circuits Thread.

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

Tengxiao Liu, Qipeng Guo, Xiangkun Hu, Cheng Jiayang, Yue Zhang, Xipeng Qiu, and Zheng Zhang. 2024a. Can language models learn to skip steps? In The Thirty-eighth Annual Conference on Neural Information Processing Systems.

Tianqiao Liu, Zui Chen, Zitao Liu, Mi Tian, and Weiqi Luo. 2024b. Expediting and elevating large language model reasoning via hidden chain-of-thought decoding. Preprint, arXiv:2409.08561.

Yue Liu, Jiaying Wu, Yufei He, Hongcheng Gao, Hongyu Chen, Baolong Bi, Jiaheng Zhang, Zhiqi Huang, and Bryan Hooi. 2025. Efficient inference for large reasoning models: A survey. Preprint, arXiv:2503.23077.

Ziqiao Ma, Jacob Sansom, Run Peng, and Joyce Chai. 2023. Towards a holistic landscape of situated theory of mind in large language models. In Findings of the Association for Computational Linguistics: EMNLP 2023, pages 1011–1031, Singapore. Association for Computational Linguistics.

MAA. 2024. American invitational mathematics examination - aime. Accessed in February 2024, from American Invitational Mathematics Examination - AIME 2024.

Sean Michael McLeish and Long Tran-Thanh. 2023. [re] end-to-end algorithm synthesis with recurrent networks: Logical extrapolation without overthinking. In ML Reproducibility Challenge 2022.

Shen-yun Miao, Chao-Chun Liang, and Keh-Yih Su. 2020. A diverse corpus for evaluating and developing English math word problem solvers. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 975–984, Online. Association for Computational Linguistics.

Amirkeivan Mohtashami, Matteo Pagliardini, and Martin Jaggi. 2025. CoTFormer: A chain of thought driven architecture with budget-adaptive computation cost at inference. In The Thirteenth International Conference on Learning Representations.

OpenAI. 2025. Learning to reason with llms. https://openai.com/index/learning-to-reason-with-llms/.

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

Jacob Pfau, William Merrill, and Samuel R. Bowman. 2024. Let's think dot by dot: Hidden computation in transformer language models. In First Conference on Language Modeling.

Steven Pinker. 1994. The Language Instinct: How the Mind Creates Language. Harper Collins, New York.

Xiaoye Qu, Yafu Li, Zhaochen 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. 2025. A survey of efficient reasoning for large reasoning models: Language, multimodality, and beyond. Preprint, arXiv:2503.21614.

Qwen. 2025. Qwen3: Think deeper, act faster. https://qwenlm.github.io/blog/qwen3/.

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

Yangjun Ruan, Neil Band, Chris J. Maddison, and Tatsunori Hashimoto. 2025. Reasoning to learn from latent thoughts. Preprint, arXiv:2503.18866.

Abulhair Saparov and He He. 2023. Language models are greedy reasoners: A systematic formal analysis of chain-of-thought. In The Eleventh International Conference on Learning Representations.

Nikunj Saunshi, Nishanth Dikkala, Zhiyuan Li, Sanjiv Kumar, and Sashank J. Reddi. 2025a. Reasoning with latent thoughts: On the power of looped transformers. In The Thirteenth International Conference on Learning Representations.

Nikunj Saunshi, Nishanth Dikkala, Zhiyuan Li, Sanjiv Kumar, and Sashank J. Reddi. 2025b. Reasoning with latent thoughts: On the power of looped transformers. Preprint, arXiv:2502.17416.

Nikunj Saunshi, Nishanth Dikkala, Zhiyuan Li, Sanjiv Kumar, and Sashank J. Reddi. 2025c. Reasoning with latent thoughts: On the power of looped transformers. In The Thirteenth International Conference on Learning Representations.

Avi Schwarzschild, Eitan Borgnia, Arjun Gupta, Furong Huang, Uzi Vishkin, Micah Goldblum, and Tom Goldstein. 2021. Can you learn an algorithm? generalizing from easy to hard problems with recurrent networks. In Advances in Neural Information Processing Systems.

Yuval Shalev, Amir Feder, and Ariel Goldstein. 2024. Distributional reasoning in llms: Parallel reasoning processes in multi-hop reasoning. Preprint, arXiv:2406.13858.

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

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

Huatong Song, Jinhao Jiang, Yingqian Min, Jie Chen, Zhipeng Chen, Wayne Xin Zhao, Lei Fang, and Ji-Rong Wen. 2025. R1-searcher: Incentivizing the search capability in llms via reinforcement learning. Preprint, arXiv:2503.05592.

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

Yang Sui, Yu-Neng Chuang, Guanchu Wang, Jiamu Zhang, Tianyi Zhang, Jiayi Yuan, Hongyi Liu, Andrew Wen, Shaochen Zhong, Hanjie Chen, and Xia Hu. 2025. Stop overthinking: A survey on efficient reasoning for large language models. Preprint, arXiv:2503.16419.

Yutao Sun, Hangbo Bao, Wenhui Wang, Zhiliang Peng, Li Dong, Shaohan Huang, Jianyong Wang, and Furu Wei. 2024. Multimodal latent language modeling with next-token diffusion. Preprint, arXiv:2412.08635.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc Le, Ed Chi, Denny Zhou, and Jason Wei. 2023. Challenging BIG-bench tasks and whether chain-of-thought can solve them. In Findings of the Association for Computational Linguistics: ACL 2023, pages 13003–13051, Toronto, Canada. Association for Computational Linguistics.

Jihoon Tack, Jack Lanchantin, Jane Yu, Andrew Cohen, Ilia Kulikov, Janice Lan, Shibo Hao, Yuandong Tian, Jason Weston, and Xian Li. 2025. Llm pretraining with continuous concepts. Preprint, arXiv:2502.08524.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Jiakai Tang, Sunhao Dai, Teng Shi, Jun Xu, Xu Chen, Wen Chen, Wu Jian, and Yuning Jiang. 2025. Think before recommend: Unleashing the latent reasoning power for sequential recommendation. Preprint, arXiv:2503.22675.

Boshi Wang, Xiang Yue, Yu Su, and Huan Sun. 2024a. Grokked transformers are implicit reasoners: A mechanistic journey to the edge of generalization. In Advances in Neural Information Processing Systems.

Rui Wang, Hongru Wang, Boyang Xue, Jianhui Pang, Shudong Liu, Yi Chen, Jiahao Qiu, Derek Fai Wong, Heng Ji, and Kam-Fai Wong. 2025a. Harnessing the reasoning economy: A survey of efficient reasoning for large language models. Preprint, arXiv:2503.24377.

Xinyi Wang, Lucas Caccia, Oleksiy Ostapenko, Xingdi Yuan, William Yang Wang, and Alessandro Sordoni.

2024b. Guiding language model reasoning with planning tokens. In First Conference on Language Modeling.

Yiming Wang, Pei Zhang, Baosong Yang, Derek F. Wong, and Rui Wang. 2025b. Latent space chain-of-embedding enables output-free llm self-evaluation. In The Thirteenth International Conference on Learning Representations.

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

Ludwig Wittgenstein. 1922. Tractatus Logico-Philosophicus. Annalen der Naturphilosophie.

Bohong Wu, Shen Yan, Sijun Zhang, Jianqiao Lu, Yutao Zeng, Ya Wang, and Xun Zhou. 2025. Efficient pre-training length scaling. Preprint, arXiv:2504.14992.

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

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

Yige Xu, Xu Guo, Zhiwei Zeng, and Chunyan Miao. 2025c. Softcot++: Test-time scaling with soft chain-of-thought reasoning. Preprint, arXiv:2505.11484.

Hongfei Xue, Yufeng Tang, Hexin Liu, Jun Zhang, Xuelong Geng, and Lei Xie. 2025. Enhancing non-core language instruction-following in speech llms via semi-implicit cross-lingual cot reasoning. Preprint, arXiv:2504.20835.

Sohee Yang, Elena Gribovskaya, Nora Kassner, Mor Geva, and Sebastian Riedel. 2024. Do large language models latently perform multi-hop reasoning? In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 10210–10229, Bangkok, Thailand. Association for Computational Linguistics.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Jiacheng Ye, Shansan Gong, Liheng Chen, Lin Zheng, Jiahui Gao, Han Shi, Chuan Wu, Xin Jiang, Zhenguo Li, Wei Bi, and Lingpeng Kong. 2024. Diffusion of thoughts: Chain-of-thought reasoning in diffusion language models. In Advances in Neural Information Processing Systems.

Tian Ye, Zicheng Xu, Yuanzhi Li, and Zeyuan Allen-Zhu. 2025. Physics of language models: Part 2.1, grade-school math and the hidden reasoning process. In The Thirteenth International Conference on Learning Representations.

Alexander Yom Din, Taelin Karidi, Leshem Choshen, and Mor Geva. 2024. Jump to conclusions: Short-cutting transformers with linear transformations. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pages 9615–9625, Torino, Italia. ELRA and ICCL.

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

Qifan Yu, Zhenyu He, Sijie Li, Xun Zhou, Jun Zhang, Jingjing Xu, and Di He. 2025. Enhancing auto-regressive chain-of-thought through loop-aligned reasoning. Preprint, arXiv:2502.08482.

Yijiong Yu. 2025. Do llms really think step-by-step in implicit reasoning? Preprint, arXiv:2411.15862.

Eric Zelikman, Georges Raif Harik, Yijia Shao, Varuna Jayasiri, Nick Haber, and Noah Goodman. 2024. Quiet-STar: Language models can teach themselves to think before speaking. In First Conference on Language Modeling.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. 2022. STar: Bootstrapping reasoning with reasoning. In Advances in Neural Information Processing Systems.

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

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

Yang Zhang, Shixin Yang, Chenjia Bai, Fei Wu, Xiu Li, Zhen Wang, and Xuelong Li. 2024. Towards efficient llm grounding for embodied multi-agent collaboration. Preprint, arXiv:2405.14314.

Xueyang Zhou, Guiyao Tie, Guowen Zhang, Weidong Wang, Zhigang Zuo, Di Wu, Duanfeng Chu, Pan Zhou, Lichao Sun, and Neil Zhenqiang Gong. 2025. Large reasoning models in agent scenarios: Exploring the necessity of reasoning capabilities. Preprint, arXiv:2503.11074.

Yutao Zhu, Zhaoheng Huang, Zhicheng Dou, and Ji-Rong Wen. 2024. One token can help! learning scalable and pluggable virtual tokens for retrieval-augmented large language models. Preprint, arXiv:2405.19670.