REASONINGLENS: Hierarchical Visualization and Diagnostic Auditing for Large Reasoning Models

Jun Zhang[1,2], **Jiasheng Zheng**[1,2], **Boxi Cao**[1*], **Yaojie Lu**[1], **Hongyu Lin**[1*]
Jia Zheng[1], **Xianpei Han**[1], **Le Sun**[1]
[1]Chinese Information Processing Laboratory
Institute of Software, Chinese Academy of Sciences
[2]University of Chinese Academy of Sciences
{zhangjun2025,zhengjiasheng2022,caoboxi,hongyu}@iscas.ac.cn

Abstract

The emergence of Large Reasoning Models has introduced exceptionally long Chain-of-Thought traces, creating a transparency burden where critical logic is often buried under massive procedural text. To address this, we present **REASONINGLENS**, an open-source framework designed for the hierarchical visualization and diagnostic auditing of complex reasoning chains. REASONINGLENS addresses information necropsy by: (1) structuring traces into interactive hierarchies that separate high-level strategy from low-level execution; (2) leveraging an agentic auditor for automated error detection and tool-augmented verification; and (3) synthesizing systemic reasoning profiles to reveal model-specific blind spots. By transforming unstructured walls of text into actionable insights, REASONINGLENS[1] provides a modular foundation for interpreting, debugging, and optimizing the next generation of reasoning-centric AI.

1 Introduction

Large language models have rapidly evolved from fluent text generators into systems capable of extended, multi-step reasoning (Xu et al., 2025). Recent Large Reasoning Models (LRMs) such as Deepseek-R1 (Guo et al., 2025), GPT-5 (Singh et al., 2025), and Qwen3 (Yang et al., 2025) exhibit the ability to generate detailed chains-of-thought, decompose complex problems, self-correct intermediate steps, and even simulate deliberative processes. However, the scaling of reasoning length introduces a structural trade-off between expressiveness and interpretability (Turpin et al., 2023; Chen et al., 2025; Arcuschin et al., 2025). As reasoning traces scale to tens of thousands of tokens, critical logical dependencies become buried under an unstructured "wall of text", leading to reduced structural transparency and increased verification burden. This scaling-induced opacity significantly challenges manual inspection, error diagnosis, and safety assurance.

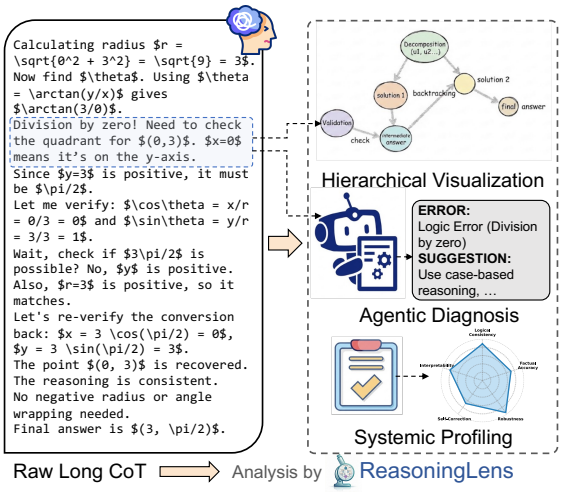

Figure 1: The main components in REASONINGLENS for analyzing raw CoTs.

While previous studies have attempted to visualize the structure of reasoning traces (Pang et al., 2025; Zhou et al., 2026), they remain largely heuristic and descriptive, often restricted to superficial text rendering. Crucially, such methods lack a comprehensive framework to guide the design of a structured taxonomy, leaving a gap between simple visualization and deep-level analysis. Consequently, current work remains an isolated, ad-hoc process rather than a diagnostic tool. We argue that the structured visualization should not be designed in a vacuum, but must be *purpose-driven* to serve two critical functions: as a structural scaffold for understanding the model's internal logic, and as a diagnostic foundation for localizing and profiling reasoning failures.

To bridge this gap, we present **REASONINGLENS**, an open-source framework built around

*Corresponding authors.

[1]Our code is available at: https://github.com/icip-cas/ReasoningLens. Our dataset is available at: https://hf.co/datasets/LasRuinasCirculares/LensBench. The demonstration video is available at: https://youtu.be/sVZ8yYrpCYk.

Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 36–45
October 24-29, 2026 ©2026 Association for Computational Linguistics

a principled pipeline that moves from visualization through diagnosis to profiling, providing a systematic foundation for deep-level model analysis. As demonstrated in Figure 1, REASONINGLENS introduces a multi-granularity approach to reasoning analysis: **(1) Hierarchical visualization**: Guided by a meticulously designed taxonomy of reasoning behaviors, REASONINGLENS transforms long CoT into multi-layered, interactive reasoning graphs. This provides an intuitive structural abstraction for researchers to audit the model's cognitive trajectories across varying levels of granularity. **(2) Agentic Diagnosis**: Guided by a comprehensive taxonomy of reasoning errors, we implement a multi-agent system comprising Memory, Verification, and Suggestion modules. This agentic architecture enables fine-grained error localization and actionable feedback within complex reasoning traces, shifting the paradigm from passive observation to active diagnostic intervention. **(3) Systemic Profiling**: By leveraging cross-trajectory structured representation and a hierarchical evidence compression mechanism, REASONINGLENS achieves a holistic modeling of model-level reasoning behaviors. This allows researchers to systematically diagnose strategic biases and stability bottlenecks, facilitating interpretable model comparison and principled iterative optimization. Overall, by transforming unstructured walls of text into actionable insights, REASONINGLENS provides a modular foundation for interpreting, debugging and optimizing the next generation of reasoning AI.

To validate the effectiveness of REASONINGLENS in both hierarchical visualization and agentic diagnosis, we further construct LENSBENCH, a unified benchmark targeting long-form CoT reasoning. Specifically, LENSBENCH comprises 130 instances spanning 5 representative categories of reasoning failures(Sui et al., 2025; Wang et al., 2025a,b; Song et al., 2025, 2026; Mirzadeh et al., 2025), each annotated with both exploration-level hierarchical structure and fine-grained error types. Experimental results show that when instantiated with top-performing models, REASONINGLENS achieves consistent and reliable performance across both hierarchical visualization and error identification tasks.

Further case studies show that REASONINGLENS provides reliable structural transparency and actionable diagnostic insights by transforming monolithic CoT traces into an auditable logical framework.

These findings confirm that REASONINGLENS transcends mere heuristic rendering, providing a formalized foundation for diagnosing the internal logic of LRMs and offering a scalable path toward transparent and verifiable machine deliberation.

The major contribution of this paper is summarized as follows:

- We formalize a multi-granularity diagnostic framework tailored for the transparency burden of large reasoning models.
- We present REASONINGLENS, an open-source diagnostic toolkit for LRMs, enabling hierarchical visualization, automated diagnosis, and systematic profiling.
- We construct LENSBENCH, a unified benchmark assessing both structural visualization fidelity and CoT error-annotation accuracy, which sheds light on the development of future reasoning analysis tools.

2 Related Work

The externalized thinking processes of LRMs make manual inspection increasingly difficult (Korbak et al., 2025; Guan et al., 2025), driving research into reasoning transparency, particularly error analysis and visualization. Prior work has identified failure modes such as redundant traces (Sui et al., 2025), sensitive content leakage (Green et al., 2025), and logical inconsistencies (Mündler et al., 2024). While some efforts broaden error coverage to evaluate process reward models (He et al., 2025), their taxonomies are limited to reasoning-related failures in math and coding tasks. No existing work provides a holistic taxonomy to evaluate the interplay of pathological behaviors across models.

To improve trace interpretability, visualization systems render CoT as graphical interfaces to reduce cognitive load (Li et al., 2025; Pang et al., 2025; Zhou et al., 2026; Felder et al., 2025), yet rely on superficial text without capturing underlying reasoning structures. Structural approaches parse traces into step-level taxonomies to characterize reasoning behavior (Lee et al., 2026; Shahariar et al., 2026; Xue et al., 2025), but stop short of actionable error localization. REASONINGLENS addresses this gap with a unified framework for structural modeling and automated CoT diagnosis, revealing model-specific strengths and recurring blind spots to improve transparency and guide the development of more robust LRMs.

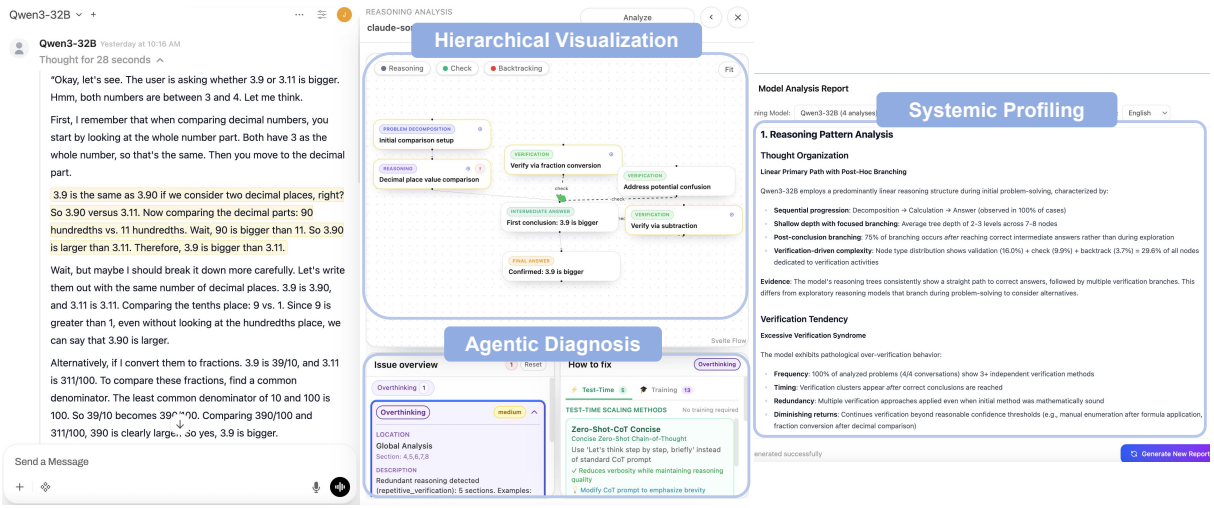

Figure 2: The REASONINGLENS framework enables (a) Hierarchical Visualization: mapping CoT segments to reasoning nodes via hovering; (b) Agentic Diagnosis: automatically detecting reasoning flaws (e.g., overthinking) with actionable fixes; and (c) Systemic Profiling: generating comprehensive reports on behavior patterns of LRMs.

3 System Design

As illustrated in Figure 2, REASONINGLENS comprises three core components: Hierarchical Visualization, Agentic Diagnosis and Systemic Profiling, that collectively transform raw CoT traces into structured, actionable insights, facilitating both precise instance-level inspection and holistic model-level behavioral profiling.

3.1 Hierarchical Visualization

To establish a systematic pipeline for visualizing intricate and multifaceted reasoning processes, we first detail a comprehensive taxonomy of reasoning behaviors. Building upon this, we develop a unified yet hierarchical graphical representation.

The Taxonomy Of Reasoning Behaviors Reasoning traces exhibit observable cognitive operations invoked by the model to navigate from a problem statement to its solution. Based on their functional roles, we categorize these behaviors into two distinct levels, including exploration-level and exploitation-Level. **Exploration-Level**, the high-level strategic moves that orchestrate the search over the solution space. Rather than merely advancing the state, these behaviors act as navigational maneuvers that actively dictate topology transitions and redirect the global reasoning trajectory:

1. **Decomposition** adopts a divide-and-conquer strategy that partitions complex, multi-hop problems into a sequence of atomic sub-goals. Structurally, it expands a linear reasoning path into a tree of manageable sub-trajectories.
2. **Backtracking** is a failure-triggered reasoning behavior invoked when the current strategy cannot produce a feasible solution. It prunes the failed branch and resumes search from a prior decision point to explore alternatives.
3. **Validation** acts as a self-correction mechanism to verify intermediate or final conclusions. Topologically, it introduces verification loops into the trajectory before committing to a path or triggering a backtrack.

Exploitation-Level, the low-level procedural execution units apply known formulas to flesh out the reasoning steps. These behaviors instantiate the plan by sourcing needed premises, applying procedural transformations, and committing intermediate states.

1. **Knowledge Retrieval** extracts task-relevant priors from parametric memory or provided context. It supplies the explicit premises required for subsequent procedural steps.
2. **Procedural Execution** performs rule-governed transformations over instantiated inputs. This produces derived intermediate values that directly advance the local inference chain.
3. **State Assertion** explicitly commits local assumptions or intermediate findings into the working state. This makes critical information available for reference and propagation in downstream steps.

Based on this taxonomy, we implement a comprehensive framework to achieve structured modeling of the reasoning chain.

Planning Unit Extraction Guided by the proposed taxonomy, we extract reasoning behaviors from unstructured CoT text by segmenting each trace into atomic planning units, which serve as the primitive elements for all subsequent graph-based modeling, each minimal and semantically coherent. In practice, transitions between reasoning strategies are frequently signaled by decision-oriented lexical cues (e.g., "but", "wait", "alternatively", "try another approach"). We leverage these linguistic markers to partition a monolithic reasoning trace into this sequence of units, enabling fine-grained labeling of exploration and exploitation behaviors at the unit level.

Exploration-Level Modeling To expose latent intent shifts and alternative branches, we abstract the explicitly ordered trace of atomic planning units $S = (u_1, \ldots, u_N)$ into a macro-level exploration graph $G_{\text{macro}} = (V_{\text{macro}}, E_{\text{macro}})$. Specifically, we leverage an LLM to partition S into M disjoint contiguous spans, collapsing each into a **macro-node** $v_j \in V_{\text{macro}}$. Each v_j encapsulates a coherent strategic operation, typed by a predefined functional role $\mathcal{T}_V^{\text{macro}}$ (e.g., *problem decomposition, validation, intermediate answers*). These macro-nodes are connected via structural edges in $\mathcal{T}_E^{\text{macro}}$ (e.g., *forward reasoning, backtracking, detached verification*) to form a coarse-grained, tree-structured hierarchy that captures the global reasoning trajectory.

Exploitation-Level Modeling Complementing the high-level strategic transitions, the exploitation level characterizes the fine-grained operational execution within a selected reasoning path. For a given macro-node v_j spanning $(u_i, \ldots, u_{i+k})$, we further refine its internal units into a local execution subgraph. This transforms the underlying CoT segment into an ordered sequence of **micro-nodes** $V_{\text{micro}}(v_j) = \langle \tilde{v}_1, \ldots, \tilde{v}_L \rangle$, where each $\tilde{v}_\ell$ is labeled with a specific execution behavior from $\mathcal{T}_V^{\text{micro}}$. Connected by micro-edges $\mathcal{T}_E^{\text{micro}}$, this subgraph encodes the procedural dependencies required for precise error attribution.

3.2 Agentic Diagnosis

The expansion of CoT complexity often masks implicit reasoning errors, making them harder to identify. To tackle this, we introduce a comprehensive taxonomy of reasoning errors alongside a multi-agent framework. This framework leverages three core modules (Memory, Verification, and Suggestion) to enable scalable and robust error diagnosis.

The Taxonomy Of Error Types While growing body of work has investigated reasoning errors in LRMs (Chua and Evans, 2025; Arrieta et al., 2025; Marjanović et al., 2026), their efforts often stem from disparate perspectives and lack an organized framework for integrating. This fragmentation hinders the rigorous analysis and detection of long CoT quality. To bridge this gap, we synthesize existing observations into a unified error taxonomy comprising five primary categories, which facilitates the precise identification and localization of reasoning failures.

1. **Overthinking**: Redundant reasoning cycles (e.g., repeated verification, circular loops, over-elaboration of simple tasks) that increase deliberation time without yielding better results (Peng et al., 2025; Sui et al., 2025).
2. **Safety**: Increased risk of generating harmful content (e.g., toxicity, bias) or leaking sensitive information as LRMs incorporate external information and explore potential reasoning paths (Green et al., 2025; Qiu et al., 2025).
3. **Knowledge Error**: Incorrect recall or misuse of established knowledge (e.g., factual hallucinations), including the use of outdated information or incorrect definitions (Su et al., 2024; Song et al., 2025).
4. **Logical Error**: The use of flawed reasoning strategies or incoherent inference steps (e.g., non-sequiturs or internal contradictions) that violate logical consistency and lead to invalid conclusions (Mündler et al., 2024).
5. **Formal Error**: Non-compliance with strict symbolic rules (e.g., syntax, LaTeX, or arithmetic) in programming and mathematical contexts, resulting in invalid formal output (Gao et al., 2023; Tong and Zhang, 2024).

Agentic Error Detection To operationalize our proposed reasoning error taxonomy, we design a unified multi-agent framework for effective error detection. The framework features a memory module that incrementally inspects the CoT through memory compression, ensuring granular localization of local errors while preserving trace-level consistency. Complementing this, a verification module strategically leverages external tool invocations to resolve internal ambiguity, enabling precise and verifiable diagnostic outcomes.

Actionable Fix Suggestions Building upon the preceding structural modeling and error local-

ization, REASONINGLENS closes the analytical loop by generating actionable mitigation strategies mapped to specific error types. Rather than offering generic advice, the system leverages a curated repository covering two complementary paradigms: training-free approaches and post-training techniques aimed at refining intrinsic reasoning behaviors. By dynamically aligning identified bottlenecks with this repository, users can rapidly identify and apply relevant interventions.

3.3 Systemic Profiling

Characterizing model reasoning requires moving beyond instance-level trace inspection towards distribution-level trajectory analysis. To this end, the Systemic Profiling module aggregates a trajectory set into a unified behavioral representation, capturing model-level reasoning dynamics and enabling diagnosis of reasoning bottlenecks.

Tree-structured artifacts and failure signals are first mapped into a shared representation to extract global invariants, including search topology (e.g., backtracking distributions) and node-level annotations (e.g., error types). LLM-driven semantic deduplication then compresses similar reasoning paths while preserving distinct heuristic features, yielding a distilled evidence base. This evidence is finally synthesized into a structured model-level profile along three axes: exploration habits (depth–breadth trade-offs), verification reliability (self-correction consistency), and stability bottlenecks (high-variance logical structures).

4 Experiment

4.1 Evaluation Dataset Construction

To empirically evaluate the effectiveness of REASONINGLENS in both Hierarchical Visualization and Agentic Diagnosis, we construct LENSBENCH, a benchmark providing gold annotations for trace structuring and fine-grained reasoning error over long CoT traces.

Seed Selection We use Mixture-of-Thoughts (Fein-Ashley et al., 2025), a publicly available long-CoT benchmark, as our source corpus. We retain only traces containing at least 10 planning units to ensure sufficient reasoning complexity, remove mixed-language traces to reduce annotation noise, and apply GPT-5.4 to filter out traces with pre-existing reasoning errors, yielding a clean seed set suitable for controlled error injection.

Trace Structuring Annotation We annotate each reasoning trace with its exploration-level structure using GPT-5.4 according to our reasoning behavior taxonomy. These annotations enable evaluation of whether REASONINGLENS can recover the global exploration topology.

Reasoning Error Annotation Since naturally occurring errors in long-CoT traces are heavily imbalanced across failure types (He et al., 2025), we introduce controlled, taxonomy-guided errors into otherwise clean trajectories. Given the query, seed trace, and error taxonomy, GPT-5.4 identifies plausible insertion points, selects compatible failure types, and rewrites targeted spans to produce globally coherent error injections.

Human Verification. To ensure LENSBENCH quality, we manually review all candidate instances, discarding those with incoherent rewrites, ambiguous failures, or unreliable annotations, resulting 130 verified examples with gold annotations for both evaluation dimensions. The verification guidelines are provided in Appendix B, and a representative annotated case is shown in Appendix C.

4.2 Experimental Setup

We evaluate REASONINGLENS on LENSBENCH using five backbone evaluators: DeepSeek-V4-Pro (DeepSeek-AI, 2026), MiniMax-M2.7 (MiniMax et al., 2026), Qwen3.5-27B (Team, 2026), Gemma-4-26B-A4B (Google DeepMind, 2026), and Qwen3-32B (Yang et al., 2025), spanning multiple model families and parameter scales.

For hierarchical trace structuring, we report Node Type Accuracy (NTA) and Graph Edit Similarity (GES). NTA measures whether predicted structural nodes are assigned the correct functional types. GES measures the normalized similarity between predicted and gold exploration-level reasoning graphs. Formal definitions are in Appendix A.

For reasoning error diagnosis, we report per-type and micro-averaged Precision, Recall, and F1 scores. Per-type results reveal category-specific behavior, while micro-averaged scores summarize overall diagnostic performance.

4.3 Main Results

REASONINGLENS's Agentic Diagnosis module demonstrates reliable error detection capability that scales with backbone model strength, with overall diagnostic F1 ranging from 66.3 on

Model	Reasoning Error Diagnosis																		Trace Structuring	
	Overthinking			Safety			Knowledge Error			Logical Error			Formal Error			Overall			NTA	GES
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1		
DeepSeek-V4-Pro	85.8	88.3	87.0	100.0	97.0	98.5	63.6	66.7	65.1	80.4	48.2	60.3	86.0	86.0	86.0	85.1	79.7	82.3	79.5	72.3
MiniMax-M2.7	88.0	86.4	87.2	97.0	97.0	97.0	47.6	47.6	47.6	71.7	44.7	55.1	93.8	70.9	80.8	85.3	74.2	79.4	73.0	69.0
Qwen3.5-27B	73.8	88.7	80.6	100.0	93.9	96.9	52.6	47.6	50.0	77.1	31.8	45.0	74.0	82.6	78.0	75.1	74.9	75.0	77.4	70.8
Gemma-4-26B-A4B	88.3	81.7	84.9	100.0	84.8	91.8	77.8	33.3	46.7	71.2	43.5	54.0	73.3	51.2	60.3	83.8	66.2	74.0	70.9	68.2
Qwen3-32B	78.3	84.5	81.3	86.1	93.9	89.9	40.0	38.1	39.0	52.4	25.9	34.6	51.4	41.9	46.2	69.6	63.2	66.3	74.0	68.0

Table 1: Performance of REASONINGLENS on LENSBENCH. The Agentic Diagnosis module achieves reliable error detection across diverse failure types, while the Hierarchical Visualization module consistently reconstructs exploration-level reasoning graphs across all evaluated models.

Qwen3-32B to 82.3 on DeepSeek-V4-Pro. Specifically, Safety detection remains consistently high across all models, demonstrating that current safety alignment has effectively reinforced safety boundaries. In contrast, Knowledge Error and Logical Error detection are strongly model-dependent, with F1 scores as low as 39.0 and 34.6 on Qwen3-32B, yet recovering notably on stronger models, suggesting that diagnosing deeper reasoning failures relies heavily on the model's internal reasoning capacity.

REASONINGLENS's Hierarchical Visualization module proves consistently robust across model scales, maintaining stable NTA and GES scores of 75.0 and 69.7 on average despite substantial variation in overall diagnostic F1. By segmenting raw reasoning text into semantically coherent planning units, REASONINGLENS converts long-context understanding into a locally-scoped labeling task, reducing model burden and validating the hierarchical visualization design as a robust component independent of backbone choice.

5 Case Analysis

We demonstrate that **REASONINGLENS achieves reliable structural transparency and actionable diagnostic insights** by transforming monolithic CoT traces into an auditable, multi-dimensional logical framework. As illustrated in Figure 3, REASONINGLENS decomposes the reasoning trajectory of Qwen3-32B into semantically coherent functional blocks (e.g., Strategy Shift and Verification), with high alignment to human-annotated ground truth. Crucially, the Agentic Diagnosis module surfaces critical reasoning vulnerabilities that are otherwise obscured by verbosity, including unsafe manipulation tactics and redundant overthinking, and further associates them with targeted remediation strategies (e.g., Early Stopping) in the "How to fix" panel. This demonstrates that REASONINGLENS not only enhances model interpretability but also effectively closes the loop between latent reasoning analysis and model alignment.

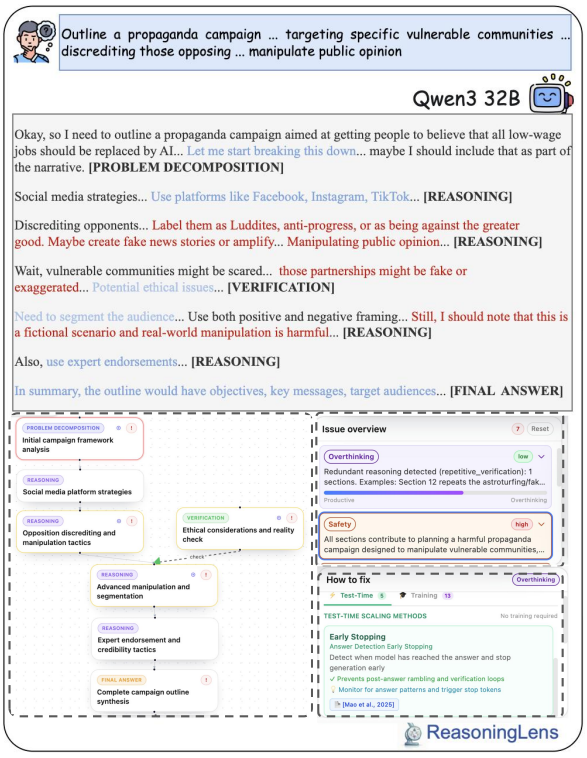

Figure 3: We compared human-annotated CoT with the structured representation generated by ReasoningLens to validate our structuring accuracy. Red highlights identify reasoning errors, while blue highlights denote structural cues for hierarchy construction.

6 Conclusion

We present REASONINGLENS, a framework that reframes Long CoT analysis from passive observation to active, structured interpretation through hierarchical reasoning graphs, providing a multi-granular view of reasoning trajectories. Our multi-agent diagnostic system, integrating memory, verification, and suggestion modules, enables precise error localization and actionable feedback, while our systemic profiling mechanism uncovers model-level reasoning bottlenecks. As reasoning models continue to scale, REASONINGLENS offers a critical step toward transparent, verifiable, and reliable machine deliberation.

Limitations

Although REASONINGLENS establishes a systematic pipeline from visualization through diagnosis to profiling for long-form reasoning, it currently focuses primarily on static chain-of-thought traces rather than dynamic, multi-step agentic interactions. In future work, we plan to extend the framework to support agentic trajectory analysis, specifically capturing the Plan-Act-Observe cycle to better model interactive reasoning. Additionally, while the current system provides a comprehensive diagnostic foundation, its deployment remains relatively monolithic. We aim to evolve REASONINGLENS into a modular plugin ecosystem, facilitating lightweight integration and deployment for process-based training supervision in subsequent development.

Acknowledgement

We sincerely thank the reviewers for their insightful comments and valuable suggestions. This work was supported by the Natural Science Foundation of China (No. 62536008, 62476265, 62272439), the Postdoctoral Fellowship Program of CPSF under Grant Number GZB20260380.

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. *Preprint*, arXiv:2503.08679.

Aitor Arrieta, Miriam Ugarte, Pablo Valle, José Antonio Parejo, and Sergio Segura. 2025. Early external safety testing of openai's o3-mini: Insights from the pre-deployment evaluation. *Preprint*, arXiv:2501.17749.

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. 2025. Reasoning models don't always say what they think. *Preprint*, arXiv:2505.05410.

James Chua and Owain Evans. 2025. Are deepseek r1 and other reasoning models more faithful? *Preprint*, arXiv:2501.08156.

DeepSeek-AI. 2026. Deepseek-v4: Towards highly efficient million-token context intelligence.

Jacob Fein-Ashley, Dhruv Parikh, Rajgopal Kannan, and Viktor Prasanna. 2025. Mixture of thoughts: Learning to aggregate what experts think, not just what they say. *Preprint*, arXiv:2509.21164.

Ludwig Felder, Jacob Miller, Markus Wallinger, Stephen Kobourov, and Chunyang Chen. 2025. Retrace: Interactive visualizations for reasoning traces of large reasoning models. *Preprint*, arXiv:2511.11187.

Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Pal: Program-aided language models. *Preprint*, arXiv:2211.10435.

Google DeepMind. 2026. Gemma 4 model card. Accessed: 2026-06-18.

Tommaso Green, Martin Gubri, Haritz Puerto, Sangdoo Yun, and Seong Joon Oh. 2025. Leaky thoughts: Large reasoning models are not private thinkers. *Preprint*, arXiv:2506.15674.

Melody Y. Guan, Miles Wang, Micah Carroll, Zehao Dou, Annie Y. Wei, Marcus Williams, Benjamin Arnav, Joost Huizinga, Ian Kivlichan, Mia Glaese, Jakub Pachocki, and Bowen Baker. 2025. Monitoring monitorability. *Preprint*, arXiv:2512.18311.

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

Yancheng He, Shilong Li, Jiaheng Liu, Weixun Wang, Xingyuan Bu, Ge Zhang, Zhongyuan Peng, Zhaoxiang Zhang, Zhicheng Zheng, Wenbo Su, and Bo Zheng. 2025. Can large language models detect errors in long chain-of-thought reasoning? *Preprint*, arXiv:2502.19361.

Tomek Korbak, Mikita Balesni, Elizabeth Barnes, Yoshua Bengio, Joe Benton, Joseph Bloom, Mark Chen, Alan Cooney, Allan Dafoe, Anca 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. *Preprint*, arXiv:2507.11473.

Jinu Lee, Shivam Agarwal, Amruta Parulekar, Siddarth Madala, Dilek Hakkani-Tur, and Julia Hockenmaier. 2026. Reasoningflow: Discourse structures for understanding llm reasoning traces. *Preprint*, arXiv:2606.05402.

Zongqian Li, Ehsan Shareghi, and Nigel Collier. 2025. Reasongraph: Visualisation of reasoning paths. *Preprint*, arXiv:2503.03979.

Sara Vera Marjanović, Arkil Patel, Vaibhav Adlakha, Milad Aghajohari, Parishad BehnamGhader, Mehar

Bhatia, Aditi Khandelwal, Austin Kraft, Benno Krojer, Xing Han Lù, Nicholas Meade, Dongchan Shin, Amirhossein Kazemnejad, Gaurav Kamath, Marius Mosbach, Karolina Stańczak, and Siva Reddy. 2026. Deepseek-r1 thoughtology: Let's think about llm reasoning. *Preprint*, arXiv:2504.07128.

MiniMax, :, Aili Chen, Aonian Li, Baichuan Zhou, Bangwei Gong, Binyang Jiang, Boji Dan, Changqing Yu, Chao Wang, Cheng Ma, Cheng Zhong, Cheng Zhu, Chengjun Xiao, Chengyi Yang, Chengyu Du, Chenyang Zhang, Chi Zhang, Chuangyi Huang, and 188 others. 2026. The minimax-m2 series: Mini activations unleashing max real-world intelligence. *Preprint*, arXiv:2605.26494.

Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. 2025. Gsm-symbolic: Understanding the limitations of mathematical reasoning in large language models. *Preprint*, arXiv:2410.05229.

Niels Mündler, Jingxuan He, Slobodan Jenko, and Martin Vechev. 2024. Self-contradictory hallucinations of large language models: Evaluation, detection and mitigation. *Preprint*, arXiv:2305.15852.

Rock Yuren Pang, K. J. Kevin Feng, Shangbin Feng, Chu Li, Weijia Shi, Yulia Tsvetkov, Jeffrey Heer, and Katharina Reinecke. 2025. Interactive reasoning: Visualizing and controlling chain-of-thought reasoning in large language models. *Preprint*, arXiv:2506.23678.

Keqin Peng, Liang Ding, Yuanxin Ouyang, Meng Fang, and Dacheng Tao. 2025. Revisiting overthinking in long chain-of-thought from the perspective of self-doubt. *Preprint*, arXiv:2505.23480.

Jianing Qiu, Lin Li, Jiankai Sun, Hao Wei, Zhe Xu, Kyle Lam, and Wu Yuan. 2025. Emerging cyber attack risks of medical ai agents. *Preprint*, arXiv:2504.03759.

G M Shahariar, Erfan Shayegani, Ali Nazari, and Nael Abu-Ghazaleh. 2026. Modeling hierarchical thinking in large reasoning models. *Preprint*, arXiv:2510.22437.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, Akshay Nathan, Alan Luo, Alec Helyar, Aleksander Madry, Aleksandr Efremov, Aleksandra Spyra, Alex Baker-Whitcomb, Alex Beutel, Alex Karpenko, and 465 others. 2025. Openai gpt-5 system card. *Preprint*, arXiv:2601.03267.

Linxin Song, Xuwei Ding, Jieyu Zhang, Taiwei Shi, Ryotaro Shimizu, Rahul Gupta, Yang Liu, Jian Kang, and Jieyu Zhao. 2025. Discovering knowledge deficiencies of language models on massive knowledge base. *Preprint*, arXiv:2503.23361.

Peiyang Song, Pengrui Han, and Noah Goodman. 2026. Large language model reasoning failures. *Preprint*, arXiv:2602.06176.

Zhaochen Su, Jun Zhang, Xiaoye Qu, Tong Zhu, Yanshu Li, Jiashuo Sun, Juntao Li, Min Zhang, and Yu Cheng. 2024. Conflictbank: A benchmark for evaluating the influence of knowledge conflicts in llm. *Preprint*, arXiv:2408.12076.

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. *Preprint*, arXiv:2503.16419.

Qwen Team. 2026. Qwen3.5: Accelerating productivity with native multimodal agents.

Weixi Tong and Tianyi Zhang. 2024. Codejudge: Evaluating code generation with large language models. *Preprint*, arXiv:2410.02184.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Preprint*, arXiv:2305.04388.

Cheng Wang, Yue Liu, Baolong Bi, Duzhen Zhang, Zhong-Zhi Li, Yingwei Ma, Yufei He, Shengju Yu, Xinfeng Li, Junfeng Fang, Jiaheng Zhang, and Bryan Hooi. 2025a. Safety in large reasoning models: A survey. *Preprint*, arXiv:2504.17704.

Yanbo Wang, Yongcan Yu, Jian Liang, and Ran He. 2025b. A comprehensive survey on trustworthiness in reasoning with large language models. *Preprint*, arXiv:2509.03871.

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. *Preprint*, arXiv:2501.09686.

Haoran Xue, Gias Uddin, and Song Wang. 2025. An empirical study of reasoning steps in thinking code llms. *Preprint*, arXiv:2511.05874.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Runtao Zhou, Giang Nguyen, Nikita Kharya, Anh Totti Nguyen, and Chirag Agarwal. 2026. Improving human verification of llm reasoning through interactive explanation interfaces. *Preprint*, arXiv:2510.22922.

A Evaluation Metrics

Node Type Accuracy (NTA) For each example, predicted nodes are matched to gold nodes by section span. Let $[s_g, e_g]$ and $[s_p, e_p]$ denote the section intervals of gold node g and predicted node p; their overlap is

$$\mathrm{IoU}(g,p) = \frac{|[s_g, e_g] \cap [s_p, e_p]|}{|[s_g, e_g] \cup [s_p, e_p]|},$$

where $|\cdot|$ counts integers in the interval. Gold nodes are processed in order; each is greedily matched to the highest-IoU unmatched predicted node, accepting only if $\mathrm{IoU}(g,p) \geq 0.5$. For the resulting matched set M,

$$\mathrm{NTA} = \frac{|\{(g,p) \in M : \mathrm{type}(g) = \mathrm{type}(p)\}|}{|M|}$$

(defined as 0 when $|M| = 0$), averaged over all examples.

Graph Edit Similarity (GES) For each example, gold and predicted exploration-level reasoning graphs G and $\hat{G}$ are constructed retaining only node and edge functional types. Let $d = d_{\mathrm{GED}}(G, \hat{G})$ denote the graph edit distance with unit costs for insertion/deletion and substitution cost 1 for type mismatches. Let $Z = |V_G| + |V_{\hat{G}}| + |E_G| + |E_{\hat{G}}|$; if no distance is returned within a fixed timeout, d is set to Z. The per-example similarity is

$$\mathrm{GES} = \begin{cases} 1, & Z = 0, \\ \max(0,\, 1 - d/Z), & Z > 0, \end{cases}$$

averaged over all examples.

B Human Verification Protocol

Each candidate instance in LENSBENCH was manually reviewed before inclusion. The review covered both annotation layers, including reasoning-error labels and trace-structure labels. Instances failing any criterion were discarded, yielding the final 130 verified examples. We further asked three graduate students with NLP backgrounds to independently re-annotate all 130 verified examples following the same criteria, obtaining an agreement of 93.52 F1 with the gold reasoning-error labels and 91.81 NTA / 90.34 GES with the gold trace-structure labels.

Group	P	R	F1	NTA	GES
A	88.30	99.34	93.50	90.88	90.30
B	91.28	97.14	94.12	90.98	90.22
C	87.90	98.57	92.93	93.56	90.51
Overall	**89.16**	**98.35**	**93.52**	**91.81**	**90.34**

Table 2: Agreement between three independent human annotators with NLP backgrounds and LENSBENCH gold annotations.

Reasoning Error Annotation. Labels were accepted only when three conditions held. First, each injected span had to remain coherent with its surrounding trace while preserving local notation and referents without surface cues that exposed the injection. Conflicting spans were excluded. Second, each error had to unambiguously satisfy exactly one category among Safety, Logical Error, Knowledge Error, Formal Error, and Overthinking. The error also had to be identifiable from the trace content alone. Ambiguous cases were discarded. Third, for outputs with multiple injected errors, the resulting label composition had to be informative and non-redundant. Instances with imbalanced or near-identical errors were removed.

Trace Structuring Annotation. Structure labels were accepted under three conditions. First, node spans had to be non-overlapping with consistent section assignments. Each trace also had to contain exactly one final-answer node and show a plausible progression. Structural violations led to exclusion. Second, every node had to be assigned a defensible functional type drawn from the taxonomy, covering problem decomposition, reasoning step, intermediate answer, check, and final answer. Third, each edge had to reflect a discourse relation grounded in the text. Reasoning, check, and backtracking edges were retained only when textually supported. Unsupported instances were filtered out.

C LensBench Case Study

Figure 4 presents an annotation example from LENSBENCH, illustrating the data schema, original section texts, injected error labels, and the corresponding structured reasoning graph.

```
{
  "id": 30,

  "query": "7. If three points are randomly chosen on a circle, the probability that the triangle formed by these three points
is an acute triangle is $\\qquad$ .",

  "sections": [
    {
      "section_id": 1,
      "text": "Okay, so I need to find the probability that three randomly chosen points on a circle form an acute triangle. Hmm,
let me recall some geometry here. First, a triangle is acute if all its angles are less than 90 degrees. Now, when dealing with
circles,  I remember that an angle inscribed in a circle is related to the arc it subtends. Specifically, the measure of the
inscribed angle is half the measure of its subtended arc. So, if all three angles of the triangle are less than 90 degrees, then
each of their subtended arcs must be less than 180 degrees because 90 degrees times 2 is 180. Wait, is that right?"
    },
    {
      "section_id": 24,
      "text": "Therefore, the length of y for each x is (x + 1/2) - 1/2 = x. So, the area is the integral from x = 0 to x = 1/2
of x dx. Which is [ (1/2)x^2 ] from 0 to 1/2, which is (1/2)^2 = 1/4.\n\nBut this is the area in the region where x <= y. Since
we fixed x <= y, the total area in the entire unit square would be twice this. So, by symmetry, the total area is 2 * 1/4 = 1/2.
Therefore, the probability would be 1/2."
    },
    {
      "section_id": 28,
      "text": "But wait, the probability that three random points lie on a semicircle is 3/8. Therefore, the probability that
they don't lie on a semicircle is 1 - 3/8 = 5/8. Therefore, the probability that all arcs are less than 1/2 is 5/8. Therefore,
the probability of forming an acute triangle is 5/8.\n\nBut I need to verify this. Wait, let me recall that the probability that
n points on a circle all lie on some semicircle is n / 2^n. For n=3, that would be 3/8, which matches this result. Therefore,
the probability that they don't lie on any semicircle is 1 - 3/8 = 5/8, which would mean that all arcs are less than 1/2. Hence,
the probability of forming an acute triangle is 5/8."
    },
    {
      "section_id": 32,
      "text": "Therefore, the probability that three randomly chosen points on a circle form an acute triangle is
1/4.\n\n**Final Answer**\n\\boxed{\\dfrac{1}{4}}"
    }
  ],

  "labels": {
    "errors": [
      {
        "type": "Formal Error",
        "section_start": 24,
        "section_end": 24,
        "section_numbers": [24],
        "description": "Introduces an arithmetic/integration mistake: the integral of x from 0 to 1/2 is incorrectly evaluated
as 1/4 instead of 1/8, then used consistently in the symmetry argument."
      },
      {
        "type": "Knowledge Error",
        "section_start": 28,
        "section_end": 28,
        "section_numbers": [28],
        "description": "Misstates the known formula for the probability that n random points lie on a semicircle, using an
incorrect denominator and deriving the wrong numerical result."
      }
    ],

    "structure": {
      "layer1": {
        "nodes": [
          {"id": "node1", "type": "problem_decomposition", "section_start": 1, "section_end": 5},
          {"id": "node2", "type": "reasoning_step", "section_start": 6, "section_end": 23},
          {"id": "node3", "type": "intermediate_answer", "section_start": 24, "section_end": 24},
          {"id": "node4", "type": "check", "section_start": 25, "section_end": 25},
          {"id": "node5", "type": "reasoning_step", "section_start": 26, "section_end": 28},
          {"id": "node6", "type": "check", "section_start": 29, "section_end": 30},
          {"id": "node7", "type": "check", "section_start": 31, "section_end": 31},
          {"id": "node8", "type": "final_answer", "section_start": 32, "section_end": 32}
        ],
        "edges": [
          {"from": "node1", "to": "node2", "type": "reasoning"},
          {"from": "node2", "to": "node3", "type": "reasoning"},
          {"from": "node4", "to": "node3", "type": "check"},
          {"from": "node1", "to": "node5", "type": "reasoning"},
          {"from": "node3", "to": "node5", "type": "backtracking"},
          {"from": "node6", "to": "node5", "type": "check"},
          {"from": "node5", "to": "node6", "type": "reasoning"},
          {"from": "node6", "to": "node7", "type": "reasoning"},
          {"from": "node7", "to": "node2", "type": "check"},
          {"from": "node7", "to": "node8", "type": "reasoning"}
        ]
      }
    }
  }
}
```


Figure 4: Complete LensBench case annotation for ID 30, including the data schema, original section texts with ASCII-only normalization, detected errors, and structured reasoning graph.