MetaCoT-CTRL: Metacognitive Control for Reliable Chain-of-Thought Reasoning in Large Language Models

Zichen Song(sls530@skku.edu; songzch21@lzu.edu.cn)
School of Applied Artificial Intelligence, Sungkyunkwan University, Seoul, 07930, South Korea
School of Information Science and Engineering, Lanzhou University, Gansu, 730000, CHINA

Abstract

Chain-of-thought (CoT) prompting improves the accuracy of large language models (LLMs) by externalizing intermediate reasoning steps, yet these traces are often unreliable, containing local inconsistencies or spurious steps that undermine interpretability and trust. We propose **MetaCoT-CTRL**, a metacognitive control framework that treats CoT as an explicit cognitive trajectory subject to monitoring and correction. MetaCoT-CTRL first generates a low-cost System-1 draft reasoning chain, then applies a conflict monitor to localize unreliable steps via contradiction, entailment, and task-specific consistency signals. A System-2 targeted repair module selectively revises only high-conflict segments, preserving stable reasoning. To ensure robustness, the repaired chains are further evaluated using causal consistency and counterfactual validation under semantically preserving input perturbations. A cognitive cost regularizer discourages redundant or overly long reasoning. Experiments on ANLI, SVAMP, and CommonQA show that MetaCoT-CTRL improves reasoning reliability and stability over existing CoT-based methods while maintaining efficient inference.

Keywords: chain-of-thought; metacognition; cognitive control; reasoning reliability; counterfactual robustness; large language models; interpretability

Introduction

Large language models (LLMs) are increasingly used as general-purpose decision assistants in tasks requiring multi-step reasoning, such as natural language inference, mathematical word problems, and commonsense question answering. Chain-of-Thought (CoT) prompting has become a widely adopted technique for improving performance by externalizing intermediate reasoning steps **?**. Beyond accuracy gains, CoT is often viewed as a pathway to transparency, enabling users to inspect how a conclusion is reached. However, growing evidence suggests that CoT traces are not inherently reliable explanations. Intermediate steps may contain local inconsistencies, causal reversals, or redundant justifications, even when the final answer is correct. Such *pseudo-aligned* reasoning traces can inflate user trust while masking unstable or erroneous reasoning processes.

From a cognitive science perspective, this limitation reflects a gap between generating a reasoning trace and exercising control over the reasoning process. Human cognition does not rely solely on producing candidate solutions; it also involves monitoring for conflict, detecting anomalies, and allocating additional cognitive resources to repair uncertain steps. Classical dual-process accounts distinguish fast, heuristic processing (System 1) from slower, deliberative processing (System 2), coordinated by metacognitive control mechanisms. In contrast, standard CoT prompting typically produces a single, static reasoning trace without explicit mechanisms for error localization or correction. Existing reliability-enhancement methods, such as self-consistency or post-hoc filtering, often operate at the level of whole chains, either aggregating or discarding them, rather than identifying and repairing specific faulty steps. Moreover, robustness is usually assessed only at the level of final answers, despite the fact that effective cognition should also exhibit stability of intermediate claims under semantically preserving perturbations.

To address these limitations, we propose **MetaCoT-CTRL**, a metacognitive control framework that treats CoT as an externally observable cognitive trajectory subject to monitoring, correction, and validation. MetaCoT-CTRL begins with a low-cost System-1 draft CoT that produces an initial answer and a set of explicit intermediate claims. A conflict monitor then assigns step-wise conflict scores using contradiction signals, entailment checks, and task-specific invariance constraints, identifying localized high-risk spans within the chain. Guided by these signals, a System-2 targeted repair module selectively revises only the high-conflict segments, preserving stable reasoning steps rather than regenerating the entire chain. To ensure that repaired traces reflect coherent reasoning, MetaCoT-CTRL further applies causal consistency validation and counterfactual robustness checks under semantically preserving input perturbations. A cognitive cost regularizer discourages unnecessary length and redundancy, reflecting bounded rationality. We evaluate MetaCoT-CTRL on three benchmarks spanning adversarial reasoning (ANLI), mathematical consistency (SVAMP), and commonsense inference (CommonQA). Across tasks, MetaCoT-CTRL improves not only answer accuracy but also the reliability and stability of reasoning trajectories, demonstrating the value of explicit metacognitive control for trustworthy LLM reasoning.

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

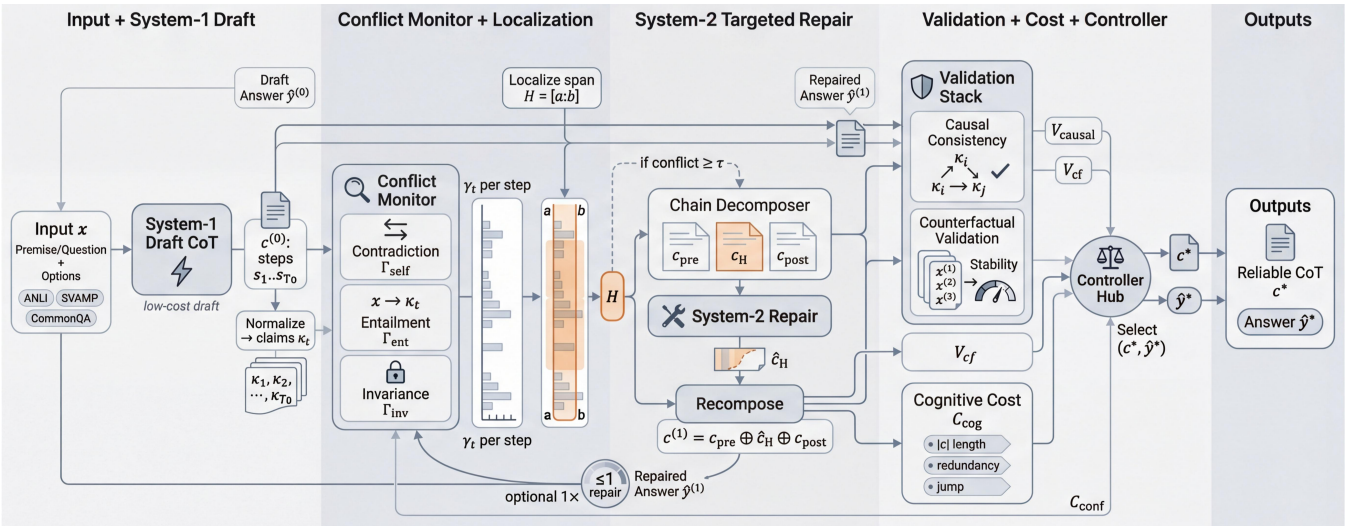

Figure 1: **Overview of the MetaCoT-CTRL framework.** Given an input x, a low-cost System-1 process generates a draft chain-of-thought $c^{(0)}$ and initial answer $\hat{y}^{(0)}$. A conflict monitor evaluates step-wise contradictions, entailment violations, and invariance constraints to localize a high-risk span H, which triggers targeted System-2 repair. The repaired chain is validated for causal consistency and counterfactual robustness, while cognitive cost penalizes unnecessary length and redundancy. A controller then selects the reasoning trajectory and answer that jointly optimize reliability and bounded cognitive cost.

Our contributions are threefold:

- We propose **MetaCoT-CTRL**, a metacognitive control framework that treats CoT as a cognitive trajectory and improves reasoning reliability through step-wise conflict monitoring, targeted repair, and validation.
- We introduce a *localized* conflict monitoring mechanism that identifies high-risk spans in a reasoning chain using contradiction/entailment signals and task-specific invariances, enabling selective System-2 repair rather than full regeneration.
- We provide an empirical evaluation on ANLI, SVAMP, and CommonQA demonstrating consistent improvements in reasoning reliability and counterfactual stability over representative CoT-based baselines, while controlling cognitive cost (verbosity and redundancy).

Background and Motivation

Chain-of-Thought and Its Limitations

Chain-of-Thought (CoT) prompting encourages LLMs to express intermediate reasoning steps in natural language, often yielding substantial performance gains on tasks requiring multi-step inference **?**. By transforming an implicit computation into an explicit textual trajectory, CoT is frequently interpreted as improving transparency: intermediate steps appear to provide a rationale that can be inspected by users or downstream verifiers. In practice, however, the epistemic status of CoT remains ambiguous. A reasoning trace may be fluent and structured yet contain incorrect intermediate claims, unsupported assumptions, or causal reversals, while still producing the correct final label or option. This phenomenon resembles *pseudo-alignment*, where surface-consistent explanations coexist with unstable internal reasoning **??**. As a result, CoT can conflate *legibility* with *validity*: steps are readable, but not necessarily justified.

Existing reliability-oriented approaches typically address CoT errors through global selection or aggregation. Self-consistency and diverse sampling methods generate multiple independent chains and select the most frequent answer, improving accuracy but providing limited guarantees about step-level correctness. Post-hoc filtering and curation methods remove low-confidence steps or discard entire chains based on heuristics such as likelihood, entropy, or similarity to reference patterns. While these strategies can be effective in aggregate, they often lack *error localization*: failures are treated at the level of the whole chain rather than pinpointing which step is problematic. Consequently, they may discard chains that are mostly correct but locally flawed, or retain chains that appear coherent globally yet contain a critical unsupported leap. Moreover, most evaluations focus on final-answer accuracy and do not quantify whether intermediate claims are stable under semantically preserving perturbations (e.g., paraphrases), even though

such stability is central to the notion of robust reasoning. These limitations motivate a shift from treating CoT as a static product to treating it as a process that can be monitored and corrected.

Cognitive Control and Metacognition

Cognitive science has long distinguished between generating candidate responses and controlling the reasoning process that produces them. Dual-process theories describe fast, heuristic processing (often associated with System 1) and slower, deliberative processing (System 2), with behavior emerging from their interaction rather than from either system alone. Crucially, System 2 is not always engaged; it is recruited when the agent detects conflict, uncertainty, or a mismatch between expectations and ongoing computation. This recruitment is often framed as *cognitive control*, supported by metacognitive mechanisms that monitor internal states (e.g., conflict, confidence) and regulate strategy selection.

In human reasoning, metacognition provides at least three functions that are directly relevant to CoT reliability. First, *conflict monitoring* identifies when current reasoning is inconsistent with prior knowledge, contextual constraints, or intermediate commitments. Second, *control allocation* determines whether additional computation is warranted, trading off expected benefits against cognitive costs. Third, *error correction* enables local repairs: rather than restarting from scratch, reasoners often revise a specific step (e.g., correcting an arithmetic operation or replacing a faulty premise) while preserving the remainder of the reasoning trajectory. These properties suggest an actionable analogy for LLM reasoning: CoT generation can be treated as System 1 drafting, while reliability requires System 2-style monitoring and correction guided by explicit metacognitive signals.

However, standard CoT prompting lacks an explicit control loop. The model produces a chain as a one-shot artifact, and any subsequent verification is typically external and global. To align CoT prompting with cognitive accounts of reliable reasoning, a framework should (i) produce explicit *monitoring signals* that localize likely errors, (ii) perform *selective intervention* rather than full regeneration, and (iii) evaluate outcomes in ways that reflect robust cognition, including stability under counterfactual variations and sensitivity to causal structure. This motivates MetaCoT-CTRL as a metacognitive control approach to CoT.

Design Principles for Cognitive Reasoning

Our goal is to operationalize *effective cognitive reasoning* for LLMs: reasoning that is not only correct at the final answer, but also coherent, stable, and efficient. We therefore adopt three design principles that guide MetaCoT-CTRL.

(1) Localize, do not only score. Reliability failures in CoT are often local: a single unsupported inference or mis-bound variable can invalidate an otherwise sensible chain. Thus, effective control requires step-wise monitoring signals that identify *where* a chain becomes unreliable. We prioritize conflict measures that support localization (e.g., contradiction across steps, entailment violations with respect to the input, and task-specific invariance checks) over monolithic chain-level scores.

(2) Repair selectively, do not always regenerate. Because many chains contain a mixture of correct and incorrect steps, full regeneration is inefficient and may discard useful partial structure. Inspired by human error correction, we treat revision as a targeted intervention: modify only the high-conflict span and its minimal dependencies while preserving stable segments. This selective repair supports both efficiency and interpretability, since the differences between pre- and post-repair trajectories can be inspected.

(3) Validate beyond the final answer. A correct answer alone does not guarantee reliable reasoning. We incorporate validators that test whether repaired chains exhibit (i) *causal coherence*—whether cause–effect relations implied by steps are directionally plausible and consistent with context—and (ii) *counterfactual robustness*—whether key intermediate claims and the final answer remain stable under semantically preserving perturbations. These validators provide process-level evidence that the reasoning trajectory reflects an effective cognitive strategy rather than brittle pattern matching.

Finally, effective cognition must respect bounded resources. We therefore explicitly model *cognitive cost* by penalizing unnecessary length, redundancy, and long-range inferential jumps. This cost-sensitive view connects reliability with the classical notion of bounded rationality: reliable reasoning should not only be correct but also economical and controlled. Together, these principles motivate a metacognitive control loop for CoT that can monitor, repair, and validate reasoning trajectories while balancing accuracy against cognitive cost.

The MetaCoT-CTRL Framework

Metacognitive Control Pipeline

MetaCoT-CTRL formulates Chain-of-Thought (CoT) reasoning as a controlled cognitive process rather than a one-shot explanatory artifact. Given an input instance x, the framework treats a reasoning trace as a sequence of explicit intermediate states,

$$c = \langle s_1, s_2, \ldots, s_T \rangle, \tag{1}$$

where each step s_t corresponds to a natural-language claim produced by the model. The goal of MetaCoT-CTRL is not only to predict a correct answer $\hat{y}$, but to select a reasoning trajectory that is reliable, stable, and economical.

The framework introduces an explicit *metacognitive control loop* that operates over candidate reasoning trajectories. Starting from an initial draft $c^{(0)}$ generated by a fast heuristic process (System-1), MetaCoT-CTRL conditionally engages deliberative processes (System-2) based on monitored conflict signals. Formally, the control objective is defined as maximizing a utility function:

$$U(c, \hat{y} \mid x) = V(c, \hat{y} \mid x) - \lambda C_{\text{conf}}(c \mid x) - \mu C_{\text{cog}}(c), \quad (2)$$

where $V(\cdot)$ measures reasoning validity and robustness, C_{conf} captures detected conflicts within the chain, and C_{cog} encodes cognitive cost. The hyperparameters λ and μ regulate the trade-off between reliability and efficiency. This formulation reflects bounded rationality: additional computation is allocated only when expected gains outweigh cognitive cost.

Conflict Monitoring and Targeted System-2 Repair

The first stage of metacognitive control is conflict monitoring, which evaluates the initial draft chain $c^{(0)}$ at the level of individual steps. Each step s_t is mapped to a normalized propositional claim κ_t, yielding a claim set

$$\mathcal{K}(c^{(0)}) = \{\kappa_1, \kappa_2, \ldots, \kappa_{T_0}\}. \quad (3)$$

For each step, a conflict score γ_t is computed by aggregating three complementary signals:

$$\gamma_t = \alpha_1 \Gamma_{\text{self}}(t) + \alpha_2 \Gamma_{\text{ent}}(t) + \alpha_3 \Gamma_{\text{inv}}(t), \quad (4)$$

where Γ_{self} captures internal inconsistency with prior steps, Γ_{ent} measures lack of entailment support from the input x, and Γ_{inv} encodes task-specific invariance violations.

Internal inconsistency is quantified as

$$\Gamma_{\text{self}}(t) = \max_{i<t} \textsc{Contradict}(\kappa_i, \kappa_t), \quad (5)$$

using a contradiction detector or NLI-based scorer. Entailment conflict is defined as

$$\Gamma_{\text{ent}}(t) = 1 - \textsc{Entail}(x, \kappa_t), \quad (6)$$

penalizing steps that are unsupported by the problem statement. Invariance conflict $\Gamma_{\text{inv}}(t)$ captures domain constraints, such as inconsistent number bindings or illegal operator transitions in arithmetic reasoning.

Rather than collapsing conflicts into a single scalar, MetaCoT-CTRL localizes a contiguous high-risk span,

$$H = \arg\max_{[a:b]} \sum_{t=a}^{b} \gamma_t, \quad (7)$$

subject to a maximum span length constraint. This localization identifies where System-2 intervention is most needed.

Targeted repair then revises only the localized span H. The original chain is decomposed as

$$c^{(0)} = c_{\text{pre}} \oplus c_H \oplus c_{\text{post}}, \quad (8)$$

and System-2 generates a repaired segment $\tilde{c}_H$ conditioned on x and c_{pre}. The repaired chain is

$$c^{(1)} = c_{\text{pre}} \oplus \tilde{c}_H \oplus c_{\text{post}}. \quad (9)$$

By constraining edits to a localized region, MetaCoT-CTRL preserves stable reasoning steps while correcting high-conflict inferences, analogous to human local error correction. Repair is applied for a bounded number of iterations to prevent overcorrection.

Validation, Cognitive Cost, and Control Policy

After repair, candidate reasoning trajectories are evaluated using two validators that assess process-level reliability. The first is *causal consistency validation*. Each chain c is interpreted as a set of directed relations between claims,

$$\mathcal{E}(c) = \{(\kappa_i \to \kappa_j)\}, \quad (10)$$

and assigned a causal coherence score

$$V_{\text{causal}}(c \mid x) = \frac{1}{|\mathcal{E}(c)|} \sum_{(i \to j) \in \mathcal{E}(c)} \textsc{Causal}(x, \kappa_i, \kappa_j), \quad (11)$$

which penalizes implausible or directionally reversed relations.

The second validator evaluates *counterfactual robustness*. Given a set of semantically preserving perturbations $\mathcal{P}(x) = \{x^{(1)}, \ldots, x^{(M)}\}$, robustness is measured as

$$V_{\text{cf}}(c, \hat{y} \mid x) = \mathbb{I}[\hat{y}^{(m)} = \hat{y}, \forall m] + \frac{1}{M} \sum_{m=1}^{M} \textsc{Sim}(\mathcal{K}(c), \mathcal{K}(c^{(m)})), \quad (12)$$

where $\textsc{Sim}(\cdot)$ measures stability of intermediate claims across perturbations. This criterion ensures that reasoning is not brittle to superficial rephrasings.

Finally, MetaCoT-CTRL introduces a cognitive cost term,

$$C_{\text{cog}}(c) = \eta_1 |c| + \eta_2 \textsc{Redundancy}(c) + \eta_3 \textsc{Jump}(c), \quad (13)$$

penalizing excessive length, repetition, and long-range inferential jumps. Given a small set of candidate trajectories $\mathcal{C}$ (typically the draft and one repaired version), the controller selects

$$(c^*, \hat{y}^*) = \arg\max_{(c, \hat{y}) \in \mathcal{C}} U(c, \hat{y} \mid x), \quad (14)$$

thereby implementing metacognitive control as a cost-sensitive selection process. This formulation allows MetaCoT-CTRL to improve reasoning reliability while respecting bounded computational resources.

Table 1: Overall performance and reasoning reliability. Higher is better (↑) except cognitive cost (CC, ↓).

Method	ANLI ↑	SVAMP ↑	CommonQA ↑	SSR ↑	CFS ↑	CC ↓
Step-by-Step (CoT)	49.8	66.1	61.4	0.71	0.62	1.00
Self-Consistency	51.3	69.4	62.7	0.73	0.64	1.48
Iterative Refinement	51.0	68.7	62.3	0.74	0.66	1.36
Effectiveness Ranking	50.9	68.2	62.1	0.75	0.65	1.18
MetaCoT-CTRL	**53.1**	**72.6**	**64.2**	**0.81**	**0.74**	**1.05**

Table 2: Statistics of benchmark datasets used in our experiments.

Dataset	Total	Train	Dev	Test
ANLI R1	4,663	4,422	121	120
ANLI R2	9,637	9,393	121	123
ANLI R3	90,652	88,532	1,053	1,067
SVAMP	4,169	3,963	206	–
CommonQA	9,215	7,306	997	912

Experimental Setup

Benchmark Datasets

To evaluate whether metacognitive control improves both *answer correctness* and *process reliability*, we conduct experiments on three benchmark families spanning adversarial inference, arithmetic reasoning, and commonsense multiple-choice QA. These benchmarks are commonly used to probe different sources of reasoning failure and thus provide complementary stress tests for CoT control.

ANLI. We use the Adversarial Natural Language Inference dataset (ANLI), which is organized into three rounds (R1–R3) with increasing adversarial difficulty. Each instance consists of a premise and hypothesis with a three-way label (entailment/contradiction/neutral). Following standard practice, we report results on each round and their aggregate.

SVAMP. SVAMP consists of arithmetic word problems that require grounded quantity manipulation and compositional reasoning. Unlike purely symbolic math, SVAMP emphasizes language-induced confounds (e.g., distractors and paraphrased templates), making it suitable for assessing invariance violations such as number binding and operator consistency. As gold test labels are not publicly available, we report results on the development set following prior work.

CommonQA. We use a commonsense multiple-choice benchmark (CommonQA) in which each question is paired with a small set of candidate answers. This setting probes abductive and commonsense reasoning and exposes failure modes where models generate fluent but unjustified elimination rationales.

Dataset statistics (train/dev/test splits) are summarized in Table 2. Throughout, we keep dataset preprocessing minimal and follow the official splits.

Base Models and Implementation Details

Our primary evaluation uses a single instruction-following LLM as the base reasoner. MetaCoT-CTRL is *model-agnostic*: it requires no architectural modifications and is applied as a wrapper around any model capable of generating CoT traces. The framework produces an initial draft CoT (System-1) and conditionally triggers System-2 repair when monitored conflict exceeds a threshold.

For all experiments, the System-2 intervention threshold is set to the 75th percentile of conflict mass observed on the development set. In practice, more than 92% of instances require at most a single repair, and we disable further repair to enforce bounded deliberation.

CoT formatting. All methods output a step-wise reasoning trace followed by a final answer. Each reasoning step is canonicalized into a proposition-like claim κ_t using a lightweight normalization prompt. For SVAMP, the normalization additionally extracts quantity bindings (number, entity, and relation).

Conflict monitor. The monitor combines contradiction among claims, entailment support from the input, and task-specific invariance checks. Entailment and contradiction are computed using an off-the-shelf NLI scorer. Invariance checks for SVAMP include number introduction penalties, unit mismatches, and operator inconsistency.

Perturbation generator. For counterfactual validation, we generate $M = 3$ semantically preserving perturbations per instance using lexical substitution, clause reordering, and minimal paraphrasing, with at most one perturbation drawn from each category.

Baseline Methods

We compare MetaCoT-CTRL against representative CoT-based reasoning enhancements:

Step-by-Step (CoT): a single reasoning chain is generated.

Self-Consistency: multiple independent chains are sampled and the final answer is selected by majority vote.

Iterative Refinement: the model critiques and revises its own chain iteratively.

Effectiveness Ranking: a chain-level scoring method (ECCoT-style) selects the most effective reasoning trajectory without targeted repair.

All baselines use the same base model and decoding budget.

Table 3: Ablation results for MetaCoT-CTRL.

Variant	ANLI ↑	SVAMP ↑	CommonQA ↑
w/o Conflict Monitor	50.7	68.1	62.0
w/o Targeted Repair	51.4	69.2	62.8
w/o Causal Validation	51.9	70.3	63.1
w/o Counterfactual Validation	51.6	69.7	63.0
w/o Cognitive Cost	52.4	71.1	63.6
Full MetaCoT-CTRL	**53.1**	**72.6**	**64.2**

Evaluation Metrics

Primary task performance is measured by **Accuracy**. We additionally evaluate reasoning reliability using four metrics.

Step Support Rate (SSR).

$$\mathrm{SSR}(c \mid x) = \frac{1}{T}\sum_{t=1}^{T} \mathbb{I}\big[\textsc{Entail}(x, \kappa_t) \geq \tau\big], \tag{15}$$

where $\tau = 0.7$ is a fixed entailment confidence threshold.

Conflict Mass (CM).

$$\mathrm{CM}(c \mid x) = \frac{1}{T}\sum_{t=1}^{T} \gamma_t. \tag{16}$$

Counterfactual Stability (CFS).

$$\mathrm{CFS}(x) = \frac{1}{M}\sum_{m=1}^{M} \mathbb{I}\Big[\hat{y}^{(m)} = \hat{y}\Big] + \frac{1}{M}\sum_{m=1}^{M} \textsc{Sim}\Big(\mathcal{K}(c), \mathcal{K}(c^{(m)})\Big). \tag{17}$$

Cognitive Cost (CC).

$$\mathrm{CC}(c) = \eta_1 |c| + \eta_2\, \textsc{Redundancy}(c) + \eta_3\, \textsc{Jump}(c). \tag{18}$$

All sampling-based results are averaged over three random seeds.

Experimental Results

Overall Performance

Table 1 reports accuracy and reliability metrics. MetaCoT-CTRL consistently outperforms all baselines across datasets, with the largest gains on SVAMP, where errors are dominated by local arithmetic inconsistencies. Improvements in accuracy are accompanied by substantial gains in SSR and CFS, indicating more grounded and stable reasoning trajectories. Cognitive cost remains close to standard CoT, demonstrating that gains are not driven by increased verbosity.

Metacognitive Control and Robustness

Conflict monitoring localizes uncertainty to a small number of steps, and targeted repair reduces conflict mass while preserving low-conflict reasoning segments. Counterfactual analysis shows that MetaCoT-CTRL maintains both answer and claim stability under paraphrase, in contrast to standard CoT, which frequently alters intermediate commitments.

Ablation and Cognitive Interpretation

Ablation results (Table 3) show that removing conflict localization leads to the largest performance drop, highlighting the importance of identifying where reasoning fails. Removing validation components degrades stability before accuracy, while removing cognitive cost regularization increases chain length with limited benefit. These patterns support a cognitive interpretation in which metacognitive control enables local error correction under bounded resources.

Conclusion

We presented **MetaCoT-CTRL**, a metacognitive control framework that reconceptualizes Chain-of-Thought reasoning as a controlled cognitive process rather than a static explanatory artifact. By introducing conflict monitoring, localized System-2 repair, and process-level validation under bounded cognitive cost, MetaCoT-CTRL operationalizes core principles from cognitive control and metacognition within large language model reasoning. Across adversarial inference, arithmetic reasoning, and commonsense QA, the framework improves not only answer accuracy but also the stability and support of intermediate reasoning steps, without relying on extensive sampling or unrestricted refinement. Beyond performance gains, MetaCoT-CTRL enables a process-oriented evaluation of machine reasoning that aligns with cognitive theories emphasizing monitoring, control allocation, and bounded rationality. Metrics such as step support, counterfactual stability, and cognitive cost make it possible to analyze how and why a model arrives at an answer, not merely whether it is correct. We view this work as a step toward cognitively grounded reasoning systems in which transparency, robustness, and efficiency are jointly optimized, and as a foundation for future research on adaptive metacognitive control.

References

Aggarwal, S., Mandowara, D., Agrawal, V., Khandelwal, D., Singla, P., & Garg, D. (2021). Explanations for commonsenseqa: New dataset and models. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: Long papers)* (pp. 3050–3065). Association for Computational Linguistics.

Allamanis, M., Panthaplackel, S., & Yin, P. (2025). Unsupervised evaluation of code llms with round-trip correctness. In *Proceedings of the 41st international conference on machine learning* (Vol. 235, pp. 1050–1066). JMLR.org.

Ang, Y., Bao, Y., Huang, Q., Tung, A. K. H., & Huang, Z. (2024). Tsgassist: An interactive assistant harnessing llms and rag for time series generation recommendations and benchmarking. *Proceedings of the VLDB Endowment*, *17*, 4309–4312. doi: 10.14778/3685800.3685862

Chen, Z., Chen, J., Singh, A., & Sra, M. (2024). Xplain-llm: A knowledge-augmented dataset for reliable grounded explanations in llms. In *Proceedings of the 2024 conference on empirical methods in natural language processing* (pp. 7578–7596). Association for Computational Linguistics.

Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tai, Y., Fedus, W., et al. (2024). Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, *25*(1), 1–53.

Fang, Z., He, Y., & Procter, R. (2024). Cwtm: Leveraging contextualized word embeddings from bert for neural topic modeling. In *Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation (lrec-coling 2024)* (pp. 4273–4286). ELRA and ICCL.

Greenblatt, R., Denison, C., Wright, B., Roger, F., MacDiarmid, M., Marks, S., & Hubinger, E. (2024). *Alignment faking in large language models.* (arXiv preprint)

Guo, R., Xu, W., & Ritter, A. (2024). Meta-tuning llms to leverage lexical knowledge for generalizable language style understanding. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 13708–13731). Association for Computational Linguistics.

Haffari, G. (2024). Direct evaluation of chain-of-thought in multi-hop reasoning with knowledge graphs. In *Findings of the association for computational linguistics: Acl 2024* (pp. 2862–2883). Association for Computational Linguistics.

Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 7132–7141).

Jiang, Y., Zhang, Y., Lin, C., Wu, D., & Lin, C.-T. (2020). Eeg-based driver drowsiness estimation using an online multi-view and transfer tsk fuzzy system. *IEEE Transactions on Intelligent Transportation Systems*, *22*(3), 1752–1764.

Lai, S., Li, J., Guo, G., Hu, X., Li, Y., Tan, Y., ... Liu, Z. (2024). Shared and private information learning in multimodal sentiment analysis with deep modal alignment and self-supervised multi-task learning. In *2024 international joint conference on neural networks (ijcnn)* (p. 1-8). doi: 10.1109/IJCNN60899.2024.10651442

Li, Y., Yang, Y., Zhu, J., Chen, H., & Wang, H. (2024). Llm-empowered few-shot node classification on incomplete graphs with real node degrees. In *Proceedings of the 33rd acm international conference on information and knowledge management* (pp. 1306–1315). ACM.

Li, Z., Fan, S., Gu, Y., Li, X., Duan, Z., Dong, B., & Wang, J. (2023). *Flexkbqa: A flexible llm-powered framework for few-shot knowledge base question answering.* (arXiv:2308.12060)

Lin, Z., Chen, H., Lu, Y., Rao, Y., Xu, H., & Lai, H. (2024). Hierarchical topic modeling via contrastive learning and hyperbolic embedding. In *Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation (lrec-coling 2024)* (pp. 8133–8143). ELRA and ICCL.

Mahmood, A., Wang, J., Yao, B., Wang, D., & Huang, C. M. (2025). User interaction patterns and breakdowns in conversing with llm-powered voice assistants. *International Journal of Human-Computer Studies*, 103406.

Mondal, D., Modi, S., Panda, S., Singh, R., & Rao, G. (2024). Kam-cot: Knowledge augmented multimodal chain-of-thoughts reasoning. In *Proceedings of the aaai conference on artificial intelligence.*

Pan, J., Cai, X., Mo, D., Yu, Y., & Li, Y. (2023). Residual attention capsule network for multimodal eeg-and eog-based driver vigilance estimation. *IEEE Transactions on Instrumentation and Measurement*, *72*, 3307756.