Article

Evidence-Budgeted Commit–Check Reasoning for Deterministic Selective Chain-of-Thought

Firstname Lastname [1], Firstname Lastname [2] and Firstname Lastname [2,*]

[1] Affiliation 1; e-mail@e-mail.com
[2] Affiliation 2; e-mail@e-mail.com
[*] Correspondence: e-mail@e-mail.com; Tel.: (optional; include country code; if there are multiple corresponding authors, add author initials) +xx-xxxx-xxx-xxxx (F.L.)

Abstract

Prompt-only Chain-of-Thought (CoT) can improve multi-step reasoning, yet a critical deployment failure persists: when wrong, models often remain maximally confident and supply fluent rationales without a decision rule for when to revise or refuse. We study a deterministic, prompt-only reliability protocol that reframes self-critique as constrained verification under greedy decoding. Evidence-Budgeted Commit-Check CoT (EB-C3oT) uses two model calls. First, the model commits to 3-6 structured, checkable claims with coarse confidences and minimal evidence handles. Second, the same model is reprompted as a strict verifier that labels each claim SUPPORTED or BROKEN, computes BROKEN_COUNT and an evidence-weighted risk score, and outputs either a certified final answer or ABSTAIN subject to a hard revision budget and risk threshold. We evaluate EB-C3oT on the first 200 GSM8K test items with Llama-3.1-8B-Instruct (temperature 0). Baseline single-pass CoT attains 87.0% accuracy at full coverage (174/200). EB-C3oT (B=1, risk threshold 0.5) abstains on 70% of items, achieves 68.3% accuracy on answered items (41/60), and emits fewer wrong answers overall than baseline (19 vs. 26) but with substantially reduced coverage. These results expose a sharp utility-safety trade-off and motivate improved verifier alignment and calibration for prompt-only selective answering.

Keywords: keyword 1; keyword 2; keyword 3 (List three to ten pertinent keywords specific to the article; yet reasonably common within the subject discipline.)

1. Introduction

Large language models (LLMs) can be prompted to produce intermediate reasoning traces that improve performance on multi-step tasks, including grade-school mathematics [1]. This style of prompting, commonly called Chain-of-Thought (CoT), is attractive because it is training-free and can be applied in a few lines of text. However, the same fluency that makes CoT outputs easy to read also creates a deployment risk: when the final answer is incorrect, the model can still produce a persuasive rationale that is difficult for a downstream user to locally falsify. In high-stakes settings, a system that answers incorrectly with confidence can be more harmful than one that refuses.

A large body of prompt-based reasoning work proposes templates that encourage planning, decomposition, critique, or refinement, and surveys emphasize both the breadth of these strategies and their fragility under prompt variation [2,3]. A recurring practical limitation is that many prompt-only "self-critique" patterns do not define an explicit, enforceable decision rule for when the model must revise its reasoning, when it should

Received:
Revised:
Accepted:
Published:

https://doi.org/10.3390/1010000

keep the original solution, and when it should abstain. Under deterministic decoding, critique can become narrative behavior: the model may generate text that sounds like verification, but still returns the same incorrect final answer.

This paper explores a lightweight protocol intended to close this gap in a way that remains compatible with strict deployment constraints: no fine-tuning, no external tools, and no additional sampling beyond fixed extra calls. We study Evidence-Budgeted Commit–Check CoT (EB-C3oT), a deterministic two-call prompting procedure that turns self-audit into a constrained certification problem. The core idea is to require the model to represent its reasoning as a small set of checkable claims, then to re-invoke the same model in a verifier role that attempts to falsify those claims while obeying a hard revision budget. If verification fails under the budget, the system must abstain rather than guess.

EB-C3oT targets three deployment-relevant goals. First, it aims to reduce wrong answers without relying on sampling-based methods such as generating many candidate rationales and selecting among them, which can be costly in latency and compute. Second, it produces explicit, step-level signals (SUPPORTED versus BROKEN labels plus an aggregate risk score) that can be logged and audited. Third, it supports selective answering: by adjusting the budget and risk threshold, a practitioner can navigate a coverage–accuracy trade-off and potentially route abstentions to escalation pathways.

We evaluate EB-C3oT on GSM8K, a standard benchmark for grade-school math reasoning [1]. Using meta-llama/Llama-3.1-8B-Instruct with greedy decoding (temperature 0) and a fixed evaluation slice of the first 200 test items, we compare a strong baseline single-pass CoT prompt against EB-C3oT configured with revision budget B=1 and risk threshold 0.5. The results are mixed and, importantly, highlight failure modes of prompt-only verification. The baseline answers all items and reaches 87.0

These findings underline a central challenge: verification must be accurate enough to filter errors but not so over-strict that it rejects many solvable questions. In the tested configuration, abstention appears to be driven primarily by the broken-claim constraint (average broken count 1.63 with B=1), suggesting either that the claim format is not sufficiently "checkable" for the verifier or that the verifier is overly conservative.

Our contributions are as follows:

- We formalize EB-C3oT, a prompt-only, deterministic two-call protocol with an explicit accept-versus-abstain decision rule based on a hard revision budget and an evidence-weighted risk score. - We provide an end-to-end implementation in a lightweight inference pipeline and evaluate it against standard single-pass CoT on 200 GSM8K test items under greedy decoding. - We report empirical evidence that EB-C3oT reduces the unconditional count of wrong emitted answers but at the cost of severe coverage loss and lower answered-item accuracy in the tested operating point, clarifying limitations and setting up concrete next steps.

Future work should emphasize operating-point sweeps (budget and risk threshold) to map the full accuracy–coverage curve, prompt and parsing refinements to reduce spurious BROKEN markings, and evaluation of whether the risk score provides a meaningful ranking signal for selective answering rather than only inducing high rejection.

2. Related Work

CoT prompting is a foundational prompt-only technique for eliciting multi-step reasoning and has shown large gains on arithmetic and commonsense benchmarks, including GSM8K, particularly at large model scales [1]. At the same time, CoT traces are neither guaranteed correct nor guaranteed faithful, and a fluent rationale can conceal incorrect intermediate steps or unsupported assumptions. EB-C3oT is motivated by this reliability

https://doi.org/10.3390/1010000

gap: it treats the reasoning trace not merely as an explanation but as an object to be checked and potentially rejected.

Surveys of reasoning with language model prompting organize a broad spectrum of techniques, including direct step-by-step prompting, decomposition strategies, and critique-and-refine templates [2,3]. EB-C3oT is best understood as orthogonal to most of these strategies. It does not propose a new decomposition heuristic; instead, it introduces a decision-procedure layer that can wrap around an existing reasoning style. The key difference is the explicit, deterministic accept-or-abstain rule, which forces the system to act on its own verification output.

A common family of prompt-only reliability techniques generates an initial answer and then asks the model to critique or revise it. In practice, these templates often lack a formal stopping condition: critique may not translate into changes to the final answer, and deterministic decoding can lead to repeated failure on the same trajectory. EB-C3oT differs by introducing a hard resource constraint, the revision budget B, which limits how many steps can be corrected. This aims to prevent the check pass from turning into an unconstrained second attempt while still allowing targeted repairs.

Another major direction for improving reasoning reliability uses multiple samples and selection, such as methods that generate several candidate chains and pick the most consistent one. EB-C3oT explicitly avoids this regime: it uses temperature 0 and exactly two calls, making it suitable when sampling is disallowed for cost, latency, or policy reasons. This strict constraint also makes the method vulnerable when the initial commitment is flawed and cannot be repaired within the budget.

Finally, iterative reprompting methods can be viewed as inference-time search over reasoning trajectories, sometimes explicitly using repeated prompting loops and selection rules [4]. EB-C3oT is more constrained than such approaches: it performs a single commit and a single check. This design choice is deliberate for determinism and bounded compute, but it limits the system's ability to explore alternative reasoning paths and may partly explain the high abstention rate observed in our experiments.

Overall, EB-C3oT contributes a simple but explicit certification framing under deterministic decoding: rather than trusting that self-critique will implicitly improve outcomes, it defines a rule that can be evaluated, tuned, and audited.

3. Background

Chain-of-Thought prompting asks an LLM to produce intermediate reasoning steps prior to a final answer, often improving accuracy on tasks that require multi-step transformation [1]. Under deterministic decoding, CoT corresponds to committing to a single reasoning trajectory. This has two implications for reliability. First, if the chosen trajectory contains an early error, the model may not recover. Second, absent an explicit refusal mechanism, the system will still output a final answer even when internal uncertainty is high.

We study selective question answering under a prompt-only, inference-time protocol. Each example consists of a question q and a gold answer y. In GSM8K, y is a numeric value extracted from the dataset answer field. A method produces either an answer a or an abstention decision.

We use two core evaluation quantities. Let N be the number of questions and let A denote the set of answered (non-abstained) examples. Coverage is defined as $|A|/N$. Answered-item accuracy is defined as $(1/|A|)$ times the number of answered examples for which the normalized predicted answer equals the normalized gold answer. We also report accuracy_all, defined as the number of correct answers divided by N, which is sensitive to abstention.

EB-C3oT uses two deterministic model calls per example. The first call produces a structured set of intermediate claims, intended to be checkable. The second call reprompts the same model as a verifier that labels each claim as SUPPORTED or BROKEN based on the original question and the committed claims.

The protocol's key novelty is an explicit accept-or-abstain decision rule based on two signals. The first is a hard revision budget B: at most B claims may be BROKEN (or corrected) while still allowing a final answer. The second is an evidence-weighted risk score. In the conceptual specification, each claim i includes a discrete confidence p$_i intheset 0.55, 0.70, 0.85. Theriskscoreisdefinedasrisk = sumoverio f(1 - p_i) timestheindicatorthatclaimiisBROKEN. Thesystemoutputsafinalansweronlyifbothconstraintsares$ $BROKEN\_COUNTisatmostBandriskisatmostathresholdtau; otherwiseitoutputsABSTAIN.$

This framing turns self-critique into a deterministic certification procedure. Its effectiveness depends on two factors that are easy to state but difficult to ensure with prompting alone: (i) whether the commit pass produces genuinely checkable claims with evidence handles that constrain interpretation, and (ii) whether the verifier reliably detects incorrect steps rather than producing superficial agreement or over-conservative rejection. These dependencies motivate reporting not just accuracy and coverage but also verification diagnostics such as average broken_count and average risk_score.

4. Method

EB-C3oT is a prompt-only, deterministic two-call protocol intended to provide bounded-compute verification and selective abstention without fine-tuning, external tools, or sampling. The method consists of a Commit pass followed by a Check pass, with an explicit decision rule.

In the Commit pass, the model is prompted to produce a short plan containing 3–6 numbered steps. Each step must be written as a checkable claim under a restricted schema. The implementation uses the following step types: ARITH (an arithmetic equation or transformation), FACT (an atomic factual relation), and CHOICE (an elimination or selection rule). Each step is annotated with a coarse confidence chosen from 0.55, 0.70, 0.85 and a minimal evidence handle such as a short equation or a quoted phrase from the question. The conceptual design separates commitments from final answers; however, in the GSM8K implementation the commit prompt ends by asking for "Final Answer: <numeric value only>" so that the verifier can reference a concrete candidate answer.

In the Check pass, the same model is reprompted as a strict verifier. The verifier receives the original question and the full commit response. It must evaluate each step and label it SUPPORTED or BROKEN, providing a one-line reason. It then computes two scalar diagnostics: BROKEN_COUNT and RISK_SCORE, where the risk score is defined as the sum over BROKEN steps of (1 - confidence). Finally, the verifier applies the decision rule: it outputs ABSTAIN if BROKEN_COUNT exceeds a fixed revision budget B or if RISK_SCORE exceeds a fixed threshold tau; otherwise it outputs "FINAL ANSWER <numeric value>".

The revision budget B is intended to prevent the second pass from becoming an unconstrained second solution attempt and to provide a controllable knob for coverage. In principle, increasing B should reduce abstention, potentially at the cost of letting more errors through.

The risk threshold tau provides a second knob that is sensitive to the commit pass confidences. The coarse confidence bins are intended to reduce noise relative to unconstrained probability outputs and to discourage spurious precision in self-reported confidence.

Beyond the final decision, EB-C3oT produces interpretable artifacts useful for auditing and downstream policy. Even when abstaining, the output includes which steps were

https://doi.org/10.3390/1010000

judged BROKEN and aggregate verification statistics. These can support escalation policies, for example routing abstentions to a human reviewer or a stronger model, and they provide a structured substrate for error analysis.

5. Experimental Setup

We evaluate EB-C3oT on GSM8K using the HuggingFace dataset "gsm8k" with the "main" configuration and the test split. To match a lightweight pilot design, we use the first 200 test items.

All methods use meta-llama/Llama-3.1-8B-Instruct served via vLLM. Decoding is deterministic with temperature 0.0, top_p 1.0, and max_tokens 2048. The baseline method uses one model call per question, while EB-C3oT uses exactly two calls per question (commit then check).

We compare two methods. The baseline is standard single-pass CoT prompting, which instructs the model to answer the question step-by-step and provide a final numerical answer. The proposed method is EB-C3oT configured with revision budget B=1 and risk threshold tau=0.5, implemented via commit and check prompt templates stored in configuration.

The baseline prompt is a single instruction-following template that requests step-by-step reasoning and a final numeric answer. The EB-C3oT commit prompt requests 3–6 steps in a fixed per-step format that includes type, confidence, and evidence, followed by a numeric final answer. The EB-C3oT check prompt requests SUPPORTED/BROKEN judgments for each step and requires the verifier to report BROKEN_COUNT, RISK_SCORE, and a final DECISION following the explicit abstention rule.

For correctness evaluation, GSM8K gold answers are parsed from the dataset's answer field by extracting the numeric value after the "####" marker and storing it as a float. Predicted answers are normalized by extracting (in priority order) a boxed number pattern if present, then common final-answer patterns, and otherwise the last number in the output text. This normalization is designed to avoid incorrectly selecting step indices such as "Step 1". A prediction is judged correct if the absolute difference between predicted and gold numeric values is less than 1e-3.

We report total_samples, num_answered, num_abstained, abstain_rate, accuracy_all (correct answers divided by total samples), and accuracy_answered (correct answers divided by answered samples). For EB-C3oT we additionally report avg$_{b}roken\_countandavg_{r}isk\_score,$

The evaluation pipeline exports per-run metrics as JSON and produces PDF figures summarizing per-run metrics and cross-run comparisons. All provided figures are included in the Results section, and we restrict quantitative claims to values present in the saved metrics.

6. Results

This section reports results from two deterministic runs on the first 200 GSM8K test items: baseline-cot (single-pass CoT) and proposed-ebc3ot (EB-C3oT with B=1 and risk threshold 0.5). All numbers are taken from saved metrics logs, and we do not extrapolate beyond them.

![Key metrics for the baseline single-pass CoT run on GSM8K (first 200 test items); higher accuracy values are better, while lower abstain rate is better.](images/baseline-cot$_{m}etrics.pdf)fig : baseline - metrics$

![Key metrics for the proposed EB-C3oT run on GSM8K (first 200 test items); higher accuracy values are better, while lower abstain rate is better.](images/proposed-ebc3ot$_{m}etrics.pdf)fig : proposed - metrics$

![Comparison of abstain rate across runs; lower values indicate better coverage.](images/compa

![Comparison of accuracy_all across runs; higher values indicate better overall correctness on the full set (including abstentions as incorrect by definition).](images/comparison$_{a}ccuracy\_a$

![Comparison of accuracy_answered across runs; higher values indicate better correctness conditional on answering.](images/comparison$_{a}ccuracy\_answered.pdf)fig : comparison - acc$

![Tabular summary comparing runs; higher accuracy values and lower abstain rate indicate better performance.](images/comparison$_{s}ummary_{t}able.pdf)fig : comparison - summary$

![Methodology comparison between baseline CoT and EB-C3oT; this is a conceptual diagram and has no direct performance metric.](images/.research/diagrams/methodology$_{c}omparison$

![System architecture of the experimental pipeline; this is a conceptual diagram and has no direct performance metric.](images/.research/diagrams/system$_{a}rchitecture_{a}iras_{p}ipeline.pdf)fi$

Experiment 1: Baseline CoT versus EB-C3oT (accuracy and coverage)

The baseline CoT answers all 200 questions (abstain_rate 0.0) and achieves accuracy_all = 0.87, which equals accuracy_answered = 0.87 because coverage is 1.0 (174 correct out of 200). EB-C3oT abstains on 140 questions (abstain_rate 0.70), answers 60 questions (answer$_{r}ate0.30), andachievesaccuracy\_answered = 0.6833(41correctoutof60answered).Becauseabste$ $C3oT'saccuracy\_allis0.205(41correctoutof200).$

Under the study's primary metric, accuracy_answered, EB-C3oT underperforms the baseline by 0.1867 absolute (0.6833 versus 0.87), as summarized in @fig:comparison-accuracy-answered and @fig:comparison-summary.

Experiment 2: Wrong-answer emission as a safety proxy

A key motivation for selective answering is reducing the number of wrong answers that are actually emitted. In absolute terms, baseline CoT emits 26 wrong answers (200 minus 174 correct). EB-C3oT emits 19 wrong answers among the 60 it answers (60 minus 41 correct). Relative to the full evaluation set, EB-C3oT emits wrong answers on 19/200 = 0.095 of questions, compared with 26/200 = 0.13 for the baseline. This reduction is achieved primarily by abstaining on many items, not by improving correctness on the items it answers.

Experiment 3: Verification diagnostics and abstention drivers

EB-C3oT reports avg$_{b}roken\_count = 1.6313andavg_{r}isk\_score = 0.3811attau = 0.5andB = 1.Giventheexplicitdecisionrule(abstainifBROKEN\_COUNT > 1orrisk\_score > 0.5), anaveragebrokencountabove1isconsistentwithBROKEN\_COUNTbeingafrequenttriggerforabst$ $stepbudgetisatleastasimportantastheriskthresholdindeterminingcoverage.$

Limitations of the current results

These results reflect a single EB-C3oT operating point and do not include sweeps over budgets or risk thresholds. As a result, we cannot use this run to characterize the full accuracy–coverage frontier or to evaluate ranking-based metrics for the risk score (for example, selective AUC or correlation with error). Additionally, the baseline CoT accuracy on this 200-item slice is unusually high relative to the expectations described in the research note, making improvements harder to demonstrate without additional tuning or alternative operating points.

7. Discussion

8. Conclusions

We studied Evidence-Budgeted Commit–Check CoT (EB-C3oT), a prompt-only, deterministic two-call protocol that converts self-critique into a constrained certification problem with an explicit accept-or-abstain decision rule. The protocol requires the model to commit to structured, checkable claims with coarse confidence levels and evidence handles, then reprompts the same model as a verifier constrained by a hard revision budget and an evidence-weighted risk threshold.

On the first 200 GSM8K test items using Llama-3.1-8B-Instruct with greedy decoding, standard single-pass CoT achieved 87.0

The results suggest that, in the tested form, EB-C3oT functions as a high-rejection safety filter rather than an accuracy-improving reasoning method. Future work should focus on mapping the coverage–accuracy trade-off by sweeping the budget and risk threshold, improving the claim and evidence format so that the verifier can more reliably distinguish real errors from ambiguous steps, and validating whether the risk score provides a smooth, actionable ranking signal for selective answering rather than inducing excessive abstention.

Author Contributions: For research articles with several authors, a short paragraph specifying their individual contributions must be provided. The following statements should be used "Conceptualization, X.X. and Y.Y.; methodology, X.X.; software, X.X.; validation, X.X., Y.Y. and Z.Z.; formal analysis, X.X.; investigation, X.X.; resources, X.X.; data curation, X.X.; writing—original draft preparation, X.X.; writing—review and editing, X.X.; visualization, X.X.; supervision, X.X.; project administration, X.X.; funding acquisition, Y.Y. All authors have read and agreed to the published version of the manuscript.", please turn to the CRediT taxonomy for the term explanation. Authorship must be limited to those who have contributed substantially to the work reported.

Funding: This research received no external funding.

Data Availability Statement: All resources used in this study are openly available at

Acknowledgments: In this study, we automatically carried out a series of research processes—from hypothesis formulation to paper writing—using generative AI.

Conflicts of Interest: The authors declare no conflicts of interest.

1. Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E.; Le, Q.; Zhou, D. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *arXiv preprint* **2022**.
2. Chen, M.; Guo, B.; Wang, H.; Li, H.; Zhao, Q.; Liu, J.; Ding, Y.; Pan, Y.; Yu, Z. Navigate through Enigmatic Labyrinth A Survey of Chain of Thought Reasoning: Advances, Frontiers and Future. *arXiv preprint* **2024**.
3. Leyva, E.S.; Lara, A.B.H.; Mateu, A.V.; Vivar, J.L.M. Reasoning with Language Model Prompting: A Survey. *arXiv preprint* **2024**.
4. Lynch, C.J.; Jensen, E.; Munro, M.H.; Zamponi, V.; Martinez, J.; O'Brien, K.; Feldhaus, B.; Smith, K.; Reinhold, A.M.; Gore, R. Reprompting: Automated Chain-of-Thought Prompt Inference Through Gibbs Sampling. *arXiv preprint* **2024**.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

https://doi.org/10.3390/1010000