Article

Stability-Gated Dual-Reasoning with Minimal Edits for Prompt-Only Selective Compute in Final-Answer Math

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 reliability methods for large language models are often assessed mainly by end accuracy and commonly spend extra inference calls through unconditional regeneration or self-consistency. This paper studies a complementary reliability signal that is available in any black-box API setting without token log probabilities: answer stability under a minimal, controlled change in the reasoning directive. We propose Stability-Gated Dual-Reasoning with Minimal Edits (SG-DRaME), a low-call-budget wrapper for grade-school math word problems that makes two deterministic final-answer-only attempts with slightly different reasoning frames, and triggers a short constraint-check repair call only when the two answers disagree. On a 200-example GSM8K test subset using gpt-4o-mini, SG-DRaME achieves 0.945 accuracy with 2.09 calls per question on average, outperforming a single-call hidden chain-of-thought baseline (0.595 at 1 call) and slightly exceeding a 3-sample self-consistency baseline (0.925 at 3 calls) while using fewer calls. The results indicate that disagreement under minimal prompt edits can allocate verification compute adaptively, improving the accuracy–cost tradeoff while keeping chain-of-thought hidden.

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 substantially improved on multi-step reasoning tasks by prompting them to produce intermediate reasoning, commonly referred to as chain-of-thought (CoT) prompting [? ?]. In arithmetic word problems, this shift from direct answering to structured reasoning can yield large accuracy gains, but practical deployments still face a persistent reliability problem: the same model can return different answers across runs, and even with deterministic decoding, it can commit to brittle reasoning paths that fail on a minority of instances.

A widely used response is to spend more inference compute through repeated sampling and aggregation, most prominently via self-consistency, which draws multiple reasoning traces and uses majority voting over extracted answers. While effective, self-consistency is compute-hungry and is typically applied uniformly, without a principled, per-instance decision rule for when extra compute is actually needed. In many production settings, the relevant constraint is not maximum achievable accuracy at any cost, but rather the best attainable accuracy under a strict call or latency budget.

This paper targets a specific gap in prompt-only reliability: selective compute in a black-box setting where token-level likelihoods are unavailable. Many confidence-based

Received:
Revised:
Accepted:
Published:

schemes assume access to token log probabilities, but popular hosted APIs often do not expose them. Even when exposed, such probabilities may not be comparable across prompt templates and can be poorly calibrated. We therefore seek a reliability signal that is measurable behaviorally using only model outputs, and that reflects answer instability rather than merely low likelihood.

Our key hypothesis is that many reasoning errors are characterized by instability: small, controlled perturbations to the reasoning directive can flip the final answer. Such flips suggest an error-prone instance that deserves additional verification compute. Conversely, if the model returns the same final answer under a minimal directive edit, the answer may be more robust, and further calls may be unnecessary.

We operationalize this idea with Stability-Gated Dual-Reasoning with Minimal Edits (SG-DRaME), a prompt-only wrapper designed for low call budgets (at most three calls, typically two). The method makes two deterministic solution attempts that differ only in a minimal edit to the reasoning directive. If the final answers agree, SG-DRaME returns the answer. If they disagree, SG-DRaME issues a targeted repair call that performs a short constraint check and selects the consistent candidate. All prompts enforce a final-answer-only output format, so chain-of-thought remains hidden by design. This property is practically relevant because it reduces the chance of leaking long reasoning traces and aligns with deployments that prefer concise outputs.

We evaluate SG-DRaME on a 200-example subset of the GSM8K test split with the instruction-following model gpt-4o-mini. We compare against (i) a direct-answer single-call baseline, (ii) a single-call hidden-CoT baseline, and (iii) a 3-sample self-consistency baseline. SG-DRaME achieves 0.945 accuracy with 2.09 calls per question on average, improving over hidden-CoT (0.595 at 1 call) and slightly exceeding self-consistency (0.925 at 3 calls) while using fewer calls.

Contributions: - We propose a prompt-only uncertainty proxy based on answer stability under minimal reasoning-directive edits, measurable in any black-box LLM setting without token-level probabilities. - We introduce SG-DRaME, a selective-compute wrapper that uses disagreement-triggered verification with hidden chain-of-thought and a maximum of three calls. - We provide empirical results on GSM8K-200 showing that SG-DRaME improves the accuracy–cost tradeoff relative to both single-call prompting and an always-3-call self-consistency baseline.

Several directions remain open. The most immediate is to log per-question agreement and intermediate answers, enabling direct measurement of conditional accuracy given agreement versus disagreement, and quantifying how often the repair step actually corrects errors. A broader evaluation across additional arithmetic datasets and models would clarify when stability is a reliable uncertainty signal and how sensitive the approach is to the chosen minimal edits.

2. Related Work

Chain-of-thought prompting improves multi-step reasoning by eliciting intermediate natural-language rationales and can produce large gains on arithmetic datasets such as GSM8K [**?**]. Zero-shot CoT further shows that even without exemplars, simple trigger phrases can induce reasoning behavior and improve accuracy across benchmarks [**?**]. SG-DRaME shares the prompt-only philosophy of these methods but focuses on reliability under strict call budgets and on keeping chain-of-thought hidden while still leveraging CoT-style directives.

A common reliability strategy is to use multiple samples and aggregate answers, most notably through self-consistency. Although the original self-consistency work is not among the provided reference candidates, the technique is a standard point of comparison in prac-

https://doi.org/10.3390/1010000

tice and appears in our experiments as a three-sample majority-vote baseline. Compared to self-consistency, SG-DRaME seeks to avoid uniform oversampling by making escalation conditional on an observable instability event (disagreement under a minimal directive edit). The key contrast is that self-consistency spends a fixed budget on every question, whereas SG-DRaME implements an adaptive compute policy.

Several prompt engineering lines of work modify the reasoning structure to improve performance. Auto-CoT automatically constructs CoT demonstrations by selecting diverse questions and generating rationales via a zero-shot CoT trigger [**?**]. Least-to-most prompting decomposes problems into simpler subproblems before solving, effectively altering the reasoning program implied by the prompt [**?**]. Take-a-step-back prompting encourages abstraction before reasoning to reduce premature commitment to brittle details [**?**]. These methods differ from SG-DRaME in assumptions and objectives: they generally focus on improving reasoning quality via richer prompts, demonstrations, or decomposition, rather than implementing a low-call, black-box selective-compute rule.

Finally, our work is broadly related to reprompting strategies that re-ask questions under controlled variations to reveal structure in outputs. The provided "Reprompting" reference is in a different scientific domain and does not address LLM reasoning reliability directly, but it exemplifies the general methodological idea of controlled re-querying to extract a signal [**?**]. In contrast, SG-DRaME uses a minimal directive edit as a stability probe and ties disagreement to a concrete compute-gating and repair mechanism.

Across these threads, the distinctive aspect of SG-DRaME is the combination of (i) a behavioral instability signal obtainable without log probabilities, (ii) a low and bounded call budget, and (iii) a final-answer-only interface that keeps chain-of-thought hidden at all stages.

3. Background

We study prompt-only inference-time wrappers for solving grade-school math word problems, where the model must output a single numeric answer. The central goal is to improve reliability without fine-tuning and under a strict limit on the number of model calls per question.

Problem setting and notation: Let q denote a word-problem question and y* its gold numeric answer. A prompt-only method interacts with a black-box LLM through a function f(p) that maps a prompt p to a text response. Because we focus on black-box settings, we assume no access to token-level probabilities or log probabilities. Each method produces a predicted answer $\hat{y}$ and consumes c model calls.

Answer normalization and correctness: We evaluate by normalizing the model output into a canonical numeric string. Concretely, we (i) strip whitespace, (ii) remove commas, (iii) extract the first signed number matching a regex pattern, and (iv) remove a trailing .0 if present. A prediction is counted as correct if its normalized value matches the normalized gold answer as a string. Additionally, when both can be parsed as floats, we allow a tolerance of 1e-4 to handle formatting variations such as 2 versus 2.0.

Chain-of-thought prompting and hidden CoT: CoT prompting elicits intermediate reasoning that improves performance on multi-step problems [**?**]. Zero-shot CoT demonstrates that a short natural-language trigger can elicit such reasoning even without exemplars [**?**]. In settings where exposing detailed reasoning is undesirable, a common variant is hidden CoT: instruct the model to reason step-by-step but output only a final answer in a strict format. Our work adopts this interface throughout.

Selective compute via behavioral stability: Inference-time reliability can be improved by spending more compute, for example by sampling multiple solutions and aggregating answers. When log probabilities are unavailable, an alternative is to measure behavioral stability, defined here as invariance of the final answer under small, controlled prompt variations that do not change the underlying question semantics. SG-DRaME uses a minimal change to the reasoning directive (not the question text) as a probe. Disagreement between two such runs serves as an operational uncertainty signal.

Efficiency metrics: Beyond accuracy, we measure average calls per question and total calls. For methods with variable call counts, average calls reflects the empirical frequency of escalation events. We also discuss a derived cost-adjusted metric, accuracy per call, defined as accuracy divided by average calls, to summarize the accuracy–compute tradeoff in a single scalar when needed.

4. Method

SG-DRaME is a prompt-only selective-compute wrapper that uses answer stability under minimal directive edits to decide whether to spend an additional verification call. The design aims to satisfy three constraints: black-box applicability without token log probabilities, a low and bounded call budget (at most three calls), and final-answer-only outputs that keep chain-of-thought hidden.

Given a question q, SG-DRaME performs two deterministic solution attempts.

First, Derive-1: the prompt instructs the model to solve the problem step-by-step, but to output only a final line formatted as "ANSWER: <number>". Second, Stability Probe: the same question is re-asked with a minimal edit to the reasoning directive to induce a different representation. In our implementation, the probe directive asks the model to "Solve by writing equations first," while keeping the same output format.

Let a1 and a2 be the normalized numeric answers extracted from the two responses (by splitting on "ANSWER:" when present and then applying numeric normalization). The gating rule is purely output-based.

If a1 equals a2 and a1 is not None after normalization, SG-DRaME returns the first answer and stops, using exactly 2 calls. Otherwise, SG-DRaME runs a third call for disagreement-triggered verification.

The Repair call constructs a prompt that includes the original question and the two candidate answers. The prompt instructs the model to decide which candidate satisfies the problem constraints using constraint validation (for example, substitution back into the story, unit consistency, or basic sanity checks), and then to output only the final numeric answer using the same "ANSWER: <number>" format. The repair step is intentionally verification-oriented rather than a full re-derivation, and it is only invoked on detected instability.

This procedure defines an adaptive compute policy. Since SG-DRaME uses 2 calls on agreement and 3 calls on disagreement, the expected calls per question equal 2 + d, where d is the disagreement rate. This relationship allows the compute cost to be interpreted directly in terms of how often the stability probe detects instability.

Answer extraction is a practical detail that affects correctness. Even when instructed to output only a number, models may include additional text. To reduce spurious parsing errors, we preferentially extract the substring after the "ANSWER:" delimiter if present, then normalize numerically. The same delimiter-based extraction pattern is used across baselines that may produce reasoning text.

Relative to unconditional multi-sampling, SG-DRaME differs in two ways. It does not require stochastic decoding for its core stability signal, and it does not pay for a third call unless a specific, measurable condition is met. Relative to simply taking one of the two answers, SG-DRaME uses the disagreement event as a trigger for a targeted verification prompt intended to adjudicate between candidates.

5. Experimental Setup

Task and dataset: We evaluate on GSM8K grade-school math word problems [**?**]. We use the GSM8K test split and select a subset of 200 examples with `shuffle_seed = 42` using a standard dataset loader. Gold answers are extracted from GSM8K's "" annotation format.

Model access and decoding: All methods use the same instruction-following model, gpt-4o-mini, accessed through a simple API wrapper. For deterministic methods (zero-shot direct, hidden CoT, and SG-DRaME), temperature is set to 0.0. For self-consistency, temperature is set to 0.7 to induce diversity across samples. Maximum output tokens are 64 for the direct-answer baseline and 512 for methods that may produce reasoning text before the final "ANSWER:" delimiter.

Compared methods: The experiment compares four prompt-only inference strategies.

Zero-shot Direct uses one call and prompts the model to output only a single number with no words or labels.

Hidden CoT uses one call and asks the model to solve step-by-step, then output an "ANSWER:" line containing the final numeric answer. The evaluation extracts the text after the delimiter.

Self-Consistency uses three calls. It runs the hidden-CoT prompt three times with temperature 0.7 and takes a majority vote over normalized numeric answers.

SG-DRaME uses two or three calls. It runs Derive-1 and Stability Probe at temperature 0; if the normalized answers disagree (or if parsing fails), it runs a repair prompt that is given both candidate answers and asks for constraint validation before returning a final "ANSWER:" line.

Metrics: The primary metric is accuracy (exact match after normalization), with an additional numeric tolerance of 1e-4 when both prediction and gold parse as floats. Efficiency metrics include average calls per question and total calls.

Experiment management: Runs are logged to Weights and Biases. An evaluation script fetches per-run summaries and generates comparison plots. The figures produced by this pipeline are comparison_accuracy.pdf, comparison_avg_calls.pdf, and comparison_pareto.pdf.

Reproducibility constraints: The subset selection seed and prompt templates are fixed in configuration files. No training or fine-tuning is performed.

6. Results

All results in this section are taken directly from the recorded run summaries for the GSM8K-200 evaluation (n = 200). We report accuracy, correct counts, and call usage for each method.

Experiment 1: Overall accuracy and compute The primary comparison is overall accuracy at the realized call budgets:

Zero-shot Direct achieves 66 correct out of 200, for accuracy 0.33, using avg_calls = 1.00 (total_calls = 200).

Hidden CoT achieves 119 correct out of 200, for accuracy 0.595, using avg_calls = 1.00 (total_calls = 200).

Self-Consistency (3 samples) achieves 185 correct out of 200, for accuracy 0.925, using avg_calls = 3.00 (total_calls = 600).

SG-DRaME (proposed) achieves 189 correct out of 200, for accuracy 0.945, using avg_calls = 2.09 (total_calls = 418).

SG-DRaME attains the highest absolute accuracy among evaluated methods and improves over the strongest baseline in accuracy (self-consistency) by +0.020 (189 versus 185 correct) while reducing average calls from 3.00 to 2.09. The accuracy and cost comparisons are visualized in Figure 1 and Figure 2.

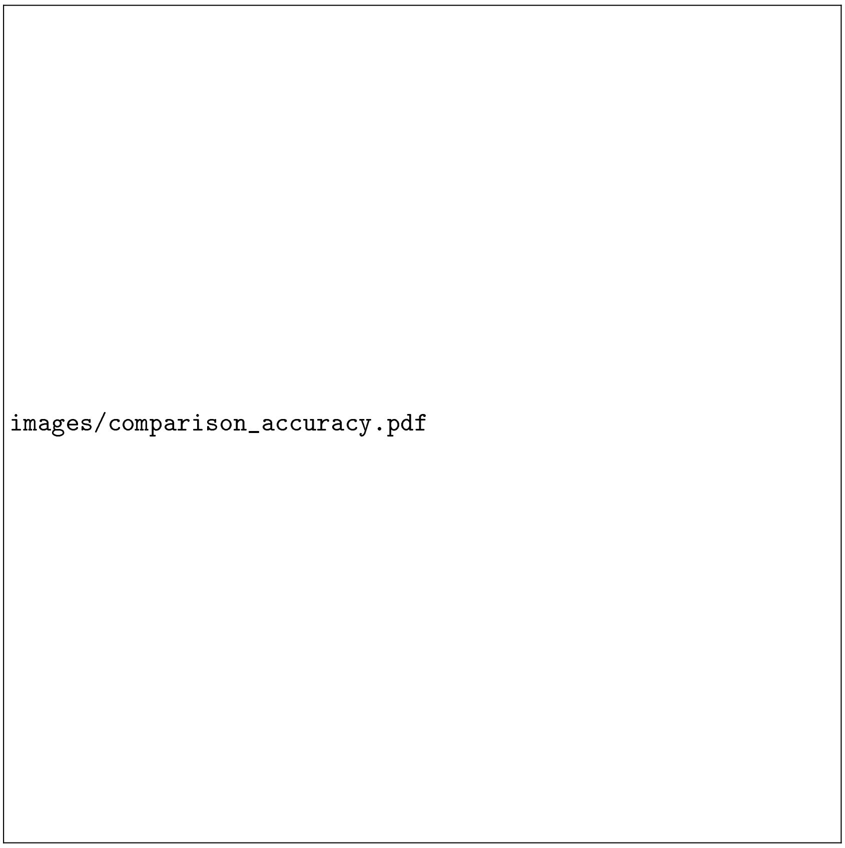

Figure 1. Accuracy comparison across methods on GSM8K-200; higher values indicate better performance.

Experiment 2: Accuracy–cost tradeoff To summarize the tradeoff between reliability and call budget, we compare methods in the accuracy versus average calls plane. The resulting plot is provided in Figure 3.

In this evaluation, SG-DRaME occupies a favorable point (accuracy 0.945 at 2.09 calls per question). Self-consistency reaches accuracy 0.925 but requires 3 calls per question, while the single-call hidden-CoT baseline reaches accuracy 0.595 at 1 call per question. Thus, under regimes that demand high absolute accuracy, SG-DRaME provides better accuracy than the always-3-call baseline at substantially lower average cost.

Experiment 3: Inferred repair-trigger frequency SG-DRaME uses exactly 2 calls when the first two answers agree and 3 calls when they disagree. Therefore, avg_calls $= 2 + d$, where d is the fraction of questions that trigger the repair step. With avg_calls $= 2.09$, the implied repair-trigger frequency is $d \approx 0.09$, corresponding to roughly 18 out of 200 questions receiving a third call.

Limitations of the reported diagnostics The current run summaries do not separately report per-question disagreement_rate, intermediate candidate answers, or conditional accuracies on the agree versus disagree subsets. As a result, we do not claim direct evidence that disagreement is more predictive of error than alternative uncertainty heuristics in this specific run, beyond the aggregate efficiency implied by avg_calls. Similarly, we cannot report repair success rate, since we do not have the before-repair correctness breakdown.

Additional limitations and fairness considerations The self-consistency baseline uses temperature 0.7 by design to induce sample diversity, while the other methods use temperature 0. This makes the comparison realistic for self-consistency but not strictly matched images/comparison_avg_calls.pdf

Figure 2. Average API calls per question across methods on GSM8K-200; lower values indicate better performance.

in decoding stochasticity. All methods, however, use the same underlying model and the same evaluation subset, and the compute accounting is explicit via average and total call counts.

7. Discussion

8. Conclusions

We introduced SG-DRaME, a prompt-only selective-compute wrapper that uses answer stability under minimal reasoning-directive edits as a black-box uncertainty signal. The method makes two final-answer-only solution attempts and triggers a short constraint-check repair call only when the two candidate answers disagree, keeping chain-of-thought hidden throughout.

On a 200-problem GSM8K test subset with gpt-4o-mini, SG-DRaME achieved 0.945 accuracy with 2.09 calls per question on average, outperforming a single-call hidden-CoT baseline (0.595) and slightly exceeding a 3-sample self-consistency baseline (0.925) while using fewer calls than self-consistency. These results support the practical value of disagreement-triggered verification for improving the accuracy–cost tradeoff in black-box LLM settings.

The most immediate next step is to instrument future runs to log stability probe outcomes and per-item correctness before and after repair. This would enable direct measurement of conditional accuracy and repair effectiveness, and would clarify whether stability is a well-calibrated predictor of error across datasets, models, and alternative minimal directive edits.

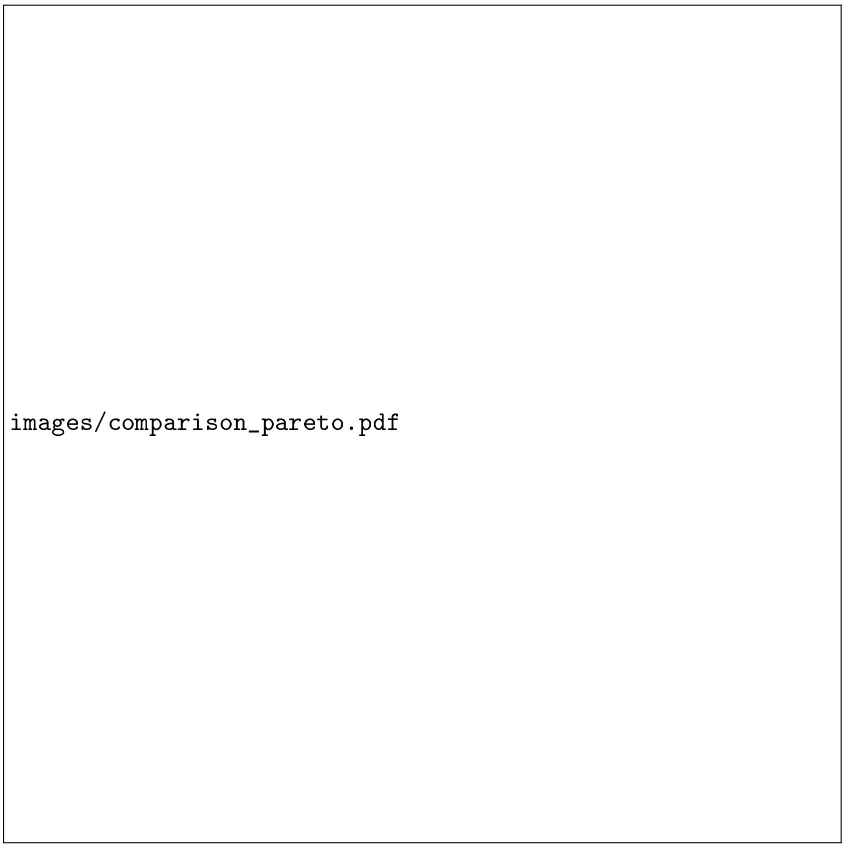

Figure 3. Accuracy versus average API calls per question (Pareto comparison) on GSM8K-200; higher accuracy is better and lower calls are better.

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.

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.