

On the Paradoxical Interference between Instruction-Following and Task Solving

Yunjia Qi[♣], Hao Peng[♣], Xintong Shi[♣], Amy Xin[♣],
Xiaozhi Wang[♣], Bin Xu[♣], Lei Hou[♣], Juanzi Li[♣]

[♣]Department of Computer Science and Technology, BNRist;

[♣]Shenzhen International Graduate School;

Tsinghua University

qyj23@mails.tsinghua.edu.cn

Abstract

Instruction following aims to align Large Language Models (LLMs) with human intent by specifying explicit constraints on how tasks should be performed. However, we reveal a counterintuitive phenomenon: instruction following can paradoxically interfere with LLMs’ task-solving capability. We propose a metric, SUSTAINSCORE, to quantify the interference of instruction following with task solving. It measures task performance drop after inserting into the instruction a self-evident constraint, which is naturally met by the original successful model output and extracted from it. Experiments on current LLMs in mathematics, multi-hop QA, and code generation show that adding the self-evident constraints leads to substantial performance drops, even for advanced models such as Claude-Sonnet-4.5. We validate the generality of the interference across constraint types and scales. Furthermore, we identify common failure patterns, and by investigating the mechanisms of interference, we observe that failed cases allocate significantly more attention to constraints compared to successful ones. Finally, we use SUSTAINSCORE to conduct an initial investigation into how distinct post-training paradigms affect the interference, presenting empirical observations on current alignment strategies. We will release our code and data to facilitate further research¹.

1 Introduction

Large Language Models (LLMs) have shown remarkable performance across diverse tasks and complex agentic workflows (Achiam et al., 1923; Guo et al., 1925; Li et al., 1924). One important factor behind this progress is instruction following, which allows human to align model behavior with their intent through explicit constraints and task specifications (Lou et al., 2024; Sun et al., 2024; Qi et al., 2024). Consequently, recent alignment

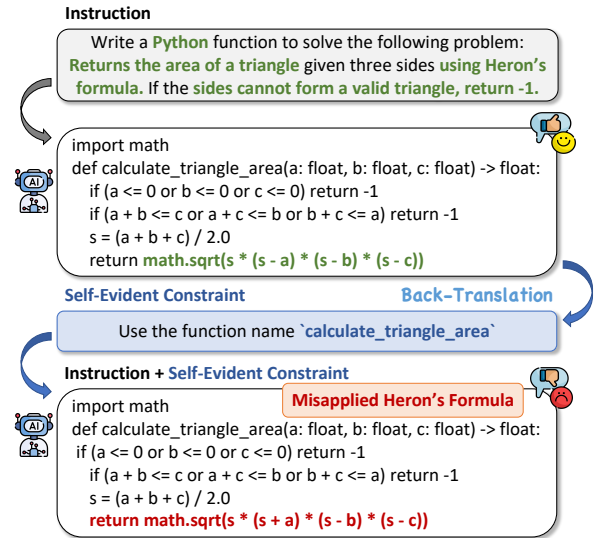


Figure 1: LLMs may fail at a task merely by adding a self-evident constraint that is already being met in their originally successful, unconstrained outputs.

pipelines and evaluation frameworks increasingly emphasize constraint satisfaction, driving the development of numerous instruction-following benchmarks (Zhou et al., 2023; Qin et al., 2024; He et al., 2024b) and targeted optimizations in advanced models (e.g., GPT-5 (OpenAI, 2025), Qwen3 (Yang et al., 2025), and GLM-4.6 (Zeng et al., 2025)). However, while instruction following is crucial for alignment, its potential impact on the model’s general task-solving abilities remains underexplored. In particular, *does adding self-evident constraints affect the model’s task-solving performance?*

In this paper, we formalize this problem as the interference of instruction following with LLMs’ task-solving capabilities, and define the underlying capability as the robustness of a model’s task-solving abilities under self-evident constraints. We introduce an automated evaluation framework and present the first systematic empirical study of this interference. Specifically, we propose a metric, SUSTAINSCORE, which evaluates this interfer-

¹<https://github.com/kijlk/IF-Interference>

ence by measuring how much task performance is sustained after self-evident constraints are imposed. To isolate the interference of capability limitations, we utilize only instances that the model solves successfully without constraints and construct self-evident constraints by deriving them directly from its own successful outputs, ensuring that the model inherently possesses the capability to solve the task under these constraints. We consider five common and widely used forms of constraints, including *Method*, *Length*, *Style*, *Keyword*, and *Structure* constraints (Zhou et al., 2023). We evaluate LLMs of varying sizes and paradigms across Mathematics, Multi-hop QA, and Code. Through extensive experiments, we uncover a paradoxical interference: *adding even self-evident constraints significantly degrades general task-solving performance*. Figure 1 illustrates a concrete instance of this fragility, where a self-evident constraint derived from a model’s own successful solution leads to unexpected task failure. These performance drops are substantial. Even advanced models such as Claude-Sonnet-4.5 (Anthropic, 2025) show measurable losses, whereas many strong models with 30B to 70B parameters retain only around 65% to 85% of their original performance. These results highlight a general lack of robustness in current models when executing tasks with constraints. This gap is not fully captured by instruction-following and task-solving benchmarks, implying that benchmark success does not guarantee proficiency in realistic, constrained scenarios.

To investigate whether this paradoxical interference is prevalent, how it impacts model performance, and what its possible underlying reasons are, we conduct three comprehensive analyses. **First**, we show that this performance degradation is consistent across experimental setups, supporting the reliability of SUSTAINSCORE. We first confirm that the observed performance decline is not a mere artifact of instruction design, such as increased context length or structural changes. Furthermore, by analyzing constraint types and scaling effects, we find that performance degrades consistently across diverse constraint categories. In terms of constraint scaling, the performance exhibits a sharp initial decline with the addition of the first few constraints, followed by a plateau phase where the degradation stabilizes. **Second**, we investigate the failure patterns and potential mechanisms behind this performance degradation. At the behavioral level, we observe two dominant error modes: (1) *Reasoning Error*, where the imposition of constraints causes the model to deviate from the correct reasoning trajectory; and (2) *Output Specification Error*, where the model derives the correct core solution but fails to satisfy specific output or completeness requirements. To investigate the underlying mechanisms, we introduce a *Constraint Attention Score*, measuring the proportion of attention focused on constraints during generation. Our comparison indicates that failed cases tend to exhibit higher constraint attention scores than successful ones, suggesting that the performance degradation may stem from an excessive attentional focus on constraint parts. **Finally**, we conduct a preliminary analysis of possible underlying reasons rooted in the training process. Our observations suggest that RL may simultaneously enhance both task performance and robustness under constraints. In contrast, while supervised fine-tuning on long chain-of-thought data boosts general performance, it appears significantly more susceptible to this degradation.

In summary, our contributions are threefold: (1) We identify the paradoxical interference between instruction following and task-solving, and formalize it as task robustness under constraints. (2) We introduce SUSTAINSCORE to purify this interference and demonstrate pervasive performance drops in current LLMs. (3) We analyze this phenomenon from empirical validation to mechanistic interpretation, and provide practical insights for model alignment. We encourage the community to emphasize task robustness under constraints to meet realistic, constrained scenarios. We also advise caution in instruction design, as excessive constraints may inadvertently degrade core task performance.

In summary, our contributions are threefold: (1) We identify the paradoxical interference between instruction following and task-solving, and formalize it as task robustness under constraints. (2) We introduce SUSTAINSCORE to purify this interference and demonstrate pervasive performance drops in current LLMs. (3) We analyze this phenomenon from empirical validation to mechanistic interpretation, and provide practical insights for model alignment. We encourage the community to emphasize task robustness under constraints to meet realistic, constrained scenarios. We also advise caution in instruction design, as excessive constraints may inadvertently degrade core task performance.

2 SUSTAINSCORE

In this section, we first provide the formal definition of SUSTAINSCORE. This metric evaluates the impact of instruction following on task-solving capabilities, i.e., the robustness of task performance under constraints. Then, we describe an automated evaluation framework to calculate SUSTAINSCORE based on existing datasets, allowing for scalability across domains without additional human effort.

2.1 Definition

Current benchmarks primarily evaluate instruction following by assessing whether a model’s output strictly adheres to a set of explicit constraints (Zhou et al., 2023; Qin et al., 2024). This evaluation pro-

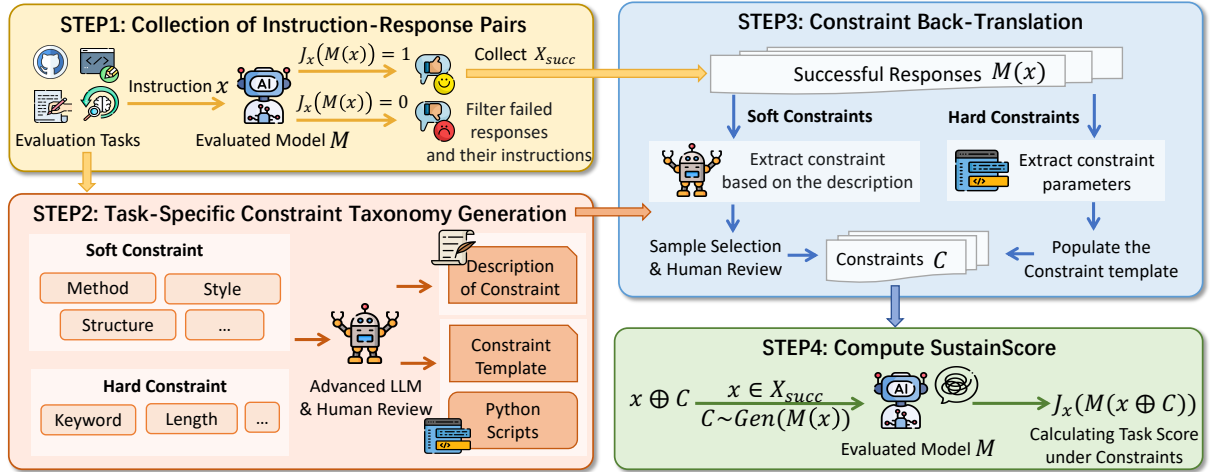


Figure 2: An overview of the evaluation framework for computing SUSTAINSCORE.

cess can be formalized by considering a core task instruction x and a set of constraints $\mathcal{C} = \{c_i\}_{i=1}^k$. The model M receives the concatenated input as instruction $x \oplus \mathcal{C}$ (where \oplus denotes concatenation) and produces an output $M(x \oplus \mathcal{C})$. The instruction following success is determined by verifying whether $M(x \oplus \mathcal{C})$ satisfies all constraints in \mathcal{C} .

While this paradigm effectively measures explicit constraint adherence, it overlooks the impact of constraints on core task performance, i.e., whether the output $M(x \oplus \mathcal{C})$ can sustain the performance of $M(x)$. To bridge this gap, we propose SUSTAINSCORE to evaluate task performance under constraints, serving as a measure of task robustness. A critical challenge for measuring this robustness is ensuring that any observed performance drop is caused by the interference of instruction following, rather than the model’s inability to solve the core task itself, or imposed constraints that render the task logically unsolvable. Therefore, it is essential to ensure the solvability of the core task for model M under the imposed constraints. Specifically, this requires satisfying two conditions: (1) **Inherent Task Capability**: the model must demonstrate the ability to solve the core task; and (2) **Constraint Reasonableness**: the imposed constraints must be *self-evident*, meaning they align with a valid solution path and do not introduce conflicts that exceed the model’s capabilities.

Accordingly, we formalize the calculation of SUSTAINSCORE as follows. Let $J_x(y) = 1$ indicate that an output y correctly solves the core task x . For a batch of instructions \mathcal{X} , we construct the valid subset $\mathcal{X}_{\text{succ}} = \{x \in \mathcal{X} \mid J_x(M(x)) = 1\}$, which contains only the tasks that the model can

successfully solve without constraints. Then, for each $x \in \mathcal{X}_{\text{succ}}$, we generate a set of *self-evident* constraints \mathcal{C} by deriving them from the successful output $M(x)$ (denoted as $\mathcal{C} \sim \text{Gen}(M(x))$), thereby ensuring that the model is inherently capable of satisfying both the task and the constraints. We define SUSTAINSCORE for model M as:

$$\text{SUSTAINSCORE}_M(\mathcal{X}) = \mathbb{E}_{x \sim \mathcal{X}_{\text{succ}}, \mathcal{C} \sim \text{Gen}(M(x))} \left[J_x(M(x \oplus \mathcal{C})) \right] \quad (1)$$

where \mathbb{E} denotes the expected task performance averaged over the evaluation dataset. This metric directly quantifies a model’s ability to retain its task success under self-evident constraints.

2.2 An automated evaluation framework

In this work, we empirically implement SUSTAINSCORE based on the formalization in Equation 1. Following prior work on instruction evaluation (Zhou et al., 2023; Peng et al., 2025), we consider five common constraint types in two categories. The first category, hard constraints, targets syntactic or lexical forms, including *Length* (e.g., word or sentence count) and *Keyword* (the inclusion or exclusion of specific terms). The second category, soft constraints, targets semantic content and organizational logic, including *Method* (the core reasoning approach), *Style* (the semantic tone), and *Structure* (the output’s organization).

The full evaluation pipeline is illustrated in Figure 2. For a given task dataset, the process begins by (1) collecting solvable instruction-response pairs while filtering out failures. Next, (2) we construct a task-specific constraint taxonomy based on

the characteristics of the evaluated task. This taxonomy then guides (3) the constraint back-translation stage, where explicit constraints are derived from each successful response. Finally, (4) compute SUSTAINSCORE. We detail the core stages below.

Collection of Instruction-Response Pairs. We generate model responses to the evaluation task dataset and filter out any failed instances. We collect only the pairs of instructions and responses where the model successfully solves the core task.

Task-Specific Constraint Taxonomy Generation. This process is tailored to the evaluation task domain (e.g., mathematics, code generation) and handles hard and soft constraints differently. For hard constraints like *Length* and *Keyword*, which can be extracted by a rule-based approach, we use a two-step procedure. First, GPT-5 (OpenAI, 2025) generates a diverse taxonomy of sub-types and phrasing templates (e.g., for *Length*, sub-types may include “word count,” “number of paragraphs,” or “sentence count”). Second, for each sub-type, the LLM drafts a Python script to extract the needed concrete parameters (e.g., a specific word count) from a given model response. All templates and scripts are then manually verified for correctness. For soft constraints such as *Method*, *Structure*, and *Style*, where a rule-based approach is infeasible, we prompt GPT-5 to produce fine-grained, task-specific descriptions for each soft type (e.g., for math, the *Method* description enumerates valid reasoning paths such as “use a system of linear equations in two variables”). These descriptions serve as rubrics for the subsequent constraint generation step, enabling targeted extraction of complex, semantic constraints from model outputs.

Constraint Front-Translation. With the taxonomy in place, we convert implicit properties of a successful response into an explicit constraint c , inspired by Qi et al. (2024). For hard constraints, we execute the Python scripts on each successful output to extract concrete parameters and instantiate the corresponding template. For soft constraints, we employ Qwen3-32B (Yang et al., 2025) as a semantic extractor, which is guided by the taxonomy’s fine-grained descriptions and few-shot examples. The LLM analyzes a response to identify its specific method, style, or structure, and formulates this into a natural-language constraint. To validate this process, we randomly sampled 200 generated constraints and conducted a manual verification;

results showed that 195 (97.5%) were consistent with the original response. Because each constraint is grounded in behaviors the model has already demonstrated, the resulting set \mathcal{C} is self-evident.

Compute SUSTAINSCORE. Finally, we append five randomly ordered constraints (one per category) to the original instruction. SUSTAINSCORE is then computed as the model’s task success rate on these new, constrained instructions, serving as a direct measure of its task-solving robustness under instruction following. See Appendix A for details.

3 Experiments

3.1 Experimental Setup

Evaluation Benchmarks Our evaluation covers three core domains: mathematical reasoning, multi-hop question answering, and code generation. (1) **Mathematical reasoning:** We use a total of 1,000 samples, randomly drawing 250 from each of the following four datasets: GSM8K (Cobbe et al., 2021), SVAMP (Patel et al., 2021), Olympiad-Bench (He et al., 2024a), and MATH500 (Lightman et al., 2023). (2) **Multi-hop QA:** For this reasoning task, we compiled a set of 1,200 samples, with 400 randomly drawn from each of HotpotQA (Yang et al., 2018), 2WikiMultiHop (Ho et al., 2020), and Musique (Trivedi et al., 2022). (3) **Code generation:** Our test set consists of 319 samples, including the complete HumanEval benchmark (Chen et al., 2021) (164 samples) and 150 from MCEval (Chai et al., 2024) (50 each for C, C++, and Python).

Investigated Models Our evaluation covers a diverse spectrum of models, categorized into two primary tiers based on their capabilities and scale: (1) **Flagship Models:** To benchmark the state-of-the-art (SOTA) models, we evaluate leading proprietary models including Claude-Sonnet-4.5 (Anthropic, 2024), Grok-4-Fast (xAI, 2025), Gemini-2.5-Flash (Comanici et al., 2025), and GPT-4.1-MINI (Achiam et al., 2023). Crucially, we also include large open-weights models that are comparable to these closed systems, specifically Qwen3-235B-A22B-Instruct (Yang et al., 2025), and DeepSeek-V3.1 (Guo et al., 2025). (2) **Efficient Models:** Spanning the 1.5B to 72B range, this category targets standard-scale and resource-efficient deployment scenarios, including models from the Qwen3 series (Yang et al., 2025), Llama series (Grattafiori et al., 2024), distilled versions of the DeepSeek series (Guo et al., 2025), GLM

Model	IF	Mathematics		Multi-Hop QA		Code Generation		Overall AVG	
		Accuracy	SUSTAINSCORE	Accuracy	SUSTAINSCORE	Accuracy	SUSTAINSCORE	Accuracy	SUSTAINSCORE
Qwen2.5-1.5B-Instruct	53.8	33.1	65.3	17.7	65.6	31.5	61.6	27.4	64.1
Qwen2.5-7B-Instruct	80.6	67.3	86.5	18.8	58.4	65.3	51.2	50.5	65.4
DeepSeek-R1-Distill-Qwen-7B	71.6	68.8	80.5	13.0	58.3	63.4	86.4	48.4	75.1
Meta-Llama-3.1-8B-Instruct	83.5	56.1	87.2	19.6	69.8	45.9	55.6	40.5	70.8
DeepSeek-R1-Distill-Llama-8B	69.9	64.8	75.2	18.2	58.4	60.2	77.2	47.7	70.3
Qwen2.5-14B-Instruct	86.7	73.3	90.3	28.7	70.6	66.2	42.8	56.1	67.9
OpenReasoning-Nemotron-14B	66.4	72.3	91.1	13.2	65.4	52.2	61.6	45.9	72.7
DeepSeek-R1-Distill-Qwen-14B	82.8	76.7	88.8	28.5	67.0	74.8	81.7	60.0	79.1
QwQ-32B	90.7	83.7	90.6	31.2	74.1	62.4	50.0	59.1	71.5
Qwen2.5-32B-Instruct	87.4	73.5	92.8	26.8	76.4	72.3	55.9	57.5	75.0
DeepSeek-R1-Distill-Qwen-32B	82.4	76.1	92.2	32.8	62.4	76.8	92.1	61.9	82.3
Qwen3-32B	90.2	85.9	91.3	34.2	73.9	48.1	78.1	56.1	81.1
GLM-Z1-32B	90.5	85.7	89.9	35.3	66.5	43.3	38.2	54.8	64.9
Meta-Llama-3.1-70B-Instruct	90.7	70.0	92.9	35.2	78.9	62.4	61.2	55.9	77.7
Qwen2.5-72B-Instruct	90.8	74.5	90.9	32.2	80.1	72.9	64.2	59.9	78.4
Gemini-2.5-Flash	92.9	72.7	84.5	42.9	73.0	76.1	86.6	63.9	81.4
Grok-4-Fast	88.8	68.5	80.7	34.9	74.5	77.1	74.8	60.2	76.7
GPT-4.1-MINI	90.9	78.9	94.3	40.2	83.6	77.1	50.8	65.4	76.3
Deepseek-V3.1	92.5	83.2	94.8	33.5	77.4	77.1	90.1	64.6	87.4
Qwen3-235B-A22B-Instruct	94.0	84.9	97.8	40.6	83.2	77.1	84.7	67.5	88.5
Claude-Sonnet-4-5	93.5	84.8	96.7	45.1	85.0	77.1	92.6	69.0	91.4

Table 1: Model performance (%) across IF (instruction following), Accuracy (task performance evaluated on the original datasets), and SUSTAINSCORE (robustness under constraints). Overall AVG reports the macro-average. The background color indicates the distance from 100% (perfect sustain), where darker shades indicate a more significant performance degradation.

series (Zeng et al., 2025), and their post-trained derivatives (Hu et al., 2025; Chen et al., 2025).

3.2 Experimental Results

Our primary experimental results, presented in Table 1, evaluate model robustness of task performance under instruction following using SUSTAINSCORE. We analyze SUSTAINSCORE alongside two reference metrics: the instruction following (IF) score, computed as the instruction-level accuracy from IFEval (Zhou et al., 2023) (the most widely used instruction-following benchmark), and the original unconstrained task accuracy (Accuracy), which reflects the model’s inherent capability on the evaluated tasks. Analysis of our main findings reveals three critical insights:

(1) All models demonstrate suboptimal performance under SUSTAINSCORE. Even the state-of-the-art model, Claude-Sonnet-4.5, retains only 85.0% of its capability on Multi-Hop QA. The degradation is more drastic for other models. For instance, GLM-Z1-32B retains merely 38.2% of its original performance on Code. In particular, many competitive models in the 32B to 72B parameter range retain only 64.9% to 82.3% of their performance on average. This widespread inability to maintain performance highlights the necessity of SUSTAINSCORE for quantifying this gap.

(2) High scores on instruction following and task-solving benchmarks do not guarantee high SUSTAINSCORE. An example is GPT-4.1-MINI,

which achieves excellent instruction-following (IF: 90.9%) and strong task performance (Code Accuracy: 77.1%), yet its Code SUSTAINSCORE is only 50.8%. Furthermore, even among strong models with IF scores exceeding 90.0% and average Accuracy above 65.0% (peaking at 69.0%), the Overall SUSTAINSCORE still spans a significant range, from 76.3% to 91.4%. This disconnection demonstrates that high task capability and strong instruction following do not necessarily translate into robust task execution under constraints.

(3) Code generation task exhibits the most severe degradation, with seven models retaining less than 60% of their performance. This likely stems from the strict nature of execution-based evaluation, coupled with the code’s structural fragility, where constraints can disrupt global logic rather than just local phrasing. Given the central role of code generation in agentic systems, this fragility highlights the importance of explicitly measuring robustness under constraints.

4 Analysis

In this section, we present three progressive analyses. First, we validate the generality of this paradoxical interference between instruction following and task solving (§ 4.1). Second, we investigate how this interference leads to performance degradation from the perspectives of behavioral patterns and internal mechanisms (§ 4.2). Finally, we conduct a preliminary analysis of the underlying causes

Model	x_{long}	$Inst_0$	$Inst_1$	$Inst_2$	$Inst_3$
Qwen2.5-7B-Instruct	92.7	86.5	80.5	86.3	84.4
Meta-Llama-3.1-8B-Instruct	92.6	87.2	86.8	90.0	88.9
OpenReasoning-Nemotron-14B	95.7	91.1	94.6	92.3	90.5
Qwen2.5-14B-Instruct	93.2	90.3	90.6	90.9	89.8
GLM-Z1-32B-0414	97.9	89.9	87.9	88.7	77.9
Qwen3-32B	95.5	91.3	91.4	91.2	91.3

Table 2: Performance (%) on instruction length and structure.

rooted in the training process (§ 4.3). We hope our analysis provides a few fresh perspectives on constructing reliable models.

4.1 Analysis on Prevalence of the Interference

To ensure the reliability of our findings, we validate SUSTAINSCORE across a wide range of experimental setups. (1) **Validity:** We first disentangle the impact of surface-level factors, confirming that the observed degradation is not merely an artifact of instruction length or structure. (2) **Scalability:** We further conduct analyses over different constraint quantities and types to verify the robustness of SUSTAINSCORE under varying levels of complexity.

Instruction Length. A potential confounding variable in our main experiment is the increased token count of the constrained instruction ($x \oplus \mathcal{C}$) compared to the original (x). We conduct an experiment where we paraphrase x to match the word count of constrained instruction $x \oplus \mathcal{C}$, creating a longer instruction x_{long} that is semantically identical to x . The experiment is conducted on the math task, and results are detailed in Table 2. We find that model performance on x_{long} shows only a negligible drop, suggesting that the significant performance degradation measured by SUSTAINSCORE is attributable to the interference of instruction following, not merely from the instruction length.

Instruction Structure. We also test whether our findings are sensitive to the structural arrangement of the task instruction and the constraint. We evaluate three alternatives to our default “instruction + constraint” format ($Inst_0$): (1) **Constraint-First Priming ($Inst_1$):** The constraint is placed before the instruction to test for order-dependence effects. (2) **Task Prioritization ($Inst_2$):** A verbose template that explicitly instructs the model to prioritize the correct solution first and the constraint second. (3) **Step-by-Step Guidance ($Inst_3$):** A template that instructs the model to think first and then write an answer that meets the constraint. Table 2 shows that the performance degradation persists across

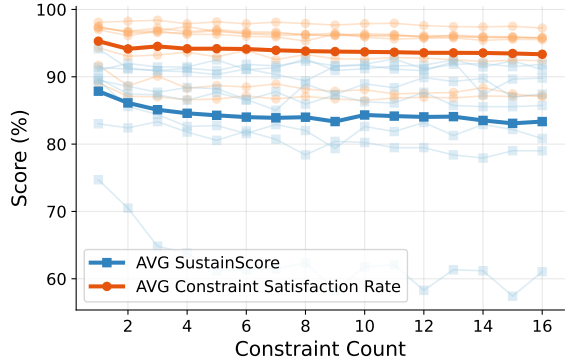


Figure 3: Comparison of SUSTAINSCORE and Constraint Satisfaction Rate as the number of constraints increases.

these structural variations. No single template alleviates the poor performance. This confirms that the robustness failure captured by SUSTAINSCORE is an intrinsic property of the model’s capability, rather than an artifact of instruction design.

Constraint Quantity. To understand SUSTAINSCORE’s scalability to the number of constraints, we conducted a constraint quantity analysis on the Math task across eight representative models using *Keyword*. We increase the number of required keywords from 1 to 16, as shown in Figure 3. The most important finding is the significant gap between a model’s ability to follow instructions and its ability to maintain core task performance. While models consistently demonstrate a near-perfect ability to adhere to the constraints (orange line, averaging over 94% satisfaction), their SUSTAINSCORE, which measures core task success, is significantly lower (blue line, averaging 84%). This gap demonstrates that existing instruction-following benchmarks, which primarily measure constraint satisfaction, fail to capture the hidden cost of constraints, treating format adherence as success even when the core task logic is broken. Furthermore, the SUSTAINSCORE curve exhibits a sharp initial decline as constraints are added. Specifically, the most informative degradation occurs within the first 5 constraints, after which the curve flattens, and the score largely stabilizes. This finding underscores the efficiency and practicality of SUSTAINSCORE, suggesting that an evaluation using just 5 constraints is sufficient to capture the model’s robustness.

Constraint Types. To investigate the impact of different constraint categories on SUSTAINSCORE, we modified our standard procedure. We applied only a single constraint from each of the five cat-

Task	Length	Keyword	Style	Method	Structure	Δ
Mathematics	87.5	89.0	89.6	90.4	90.3	2.9
Multi-Hop QA	70.1	73.2	74.2	74.9	77.4	7.3
Code Generation	73.9	70.0	83.0	86.9	87.4	17.4
AVG	77.2	77.4	82.3	84.1	85.0	7.8

Table 3: SUSTAINSCORE on different constraint types. Δ indicates the max-min gap across constraint types.

egories in isolation, allowing us to measure how models respond to different constraint types. The results, averaged across all “efficient models” in our main experiment, are presented in Table 3. First, performance is largely consistent across constraint types on average. The average performance gap between the best and worst-performing types is small (7.8), indicating that SUSTAINSCORE is not overly sensitive to the specific choice of constraint type. In practice, this flexibility allows the selection of constraint types that are easiest to implement without significantly skewing the final evaluation. Second, the Code domain exhibits higher sensitivity to *Keyword* (70.0). This aligns with the structural fragility discussed earlier: restricting specific keywords in code may risk disrupting global logic and execution flow, thereby amplifying the difficulty of the task.

4.2 Analysis on Underlying Mechanisms

To investigate the causes of model failure **at the behavioral level**, we manually inspected 50 failed cases across domains and identified two dominant error types: (1) *Reasoning Error*: The model fails to establish a correct reasoning chain, leading to incorrect mathematical derivations, erroneous multi-hop inferences, or flawed algorithmic decisions in code. (2) *Output Specification Error*: The model derives the correct core solution but fails to satisfy specific output or completeness requirements. Failures range from formatting mismatches (e.g., missing units or pattern deviations) to incomplete implementations (e.g., missing boundary cases), despite having the capability to handle them. We further analyze the distribution of these two error types across evaluated models on math. We observe that direct-answer models are more prone to Reasoning Errors, whereas CoT models frequently suffer from Output Specification Errors. This aligns with prior findings that while CoT enhances reasoning, it often exhibits weaker instruction following capabilities (Li et al., 2025). Examples and distributional statistics are provided in the Appendix B.1.

To investigate these failures **at the mechanism**

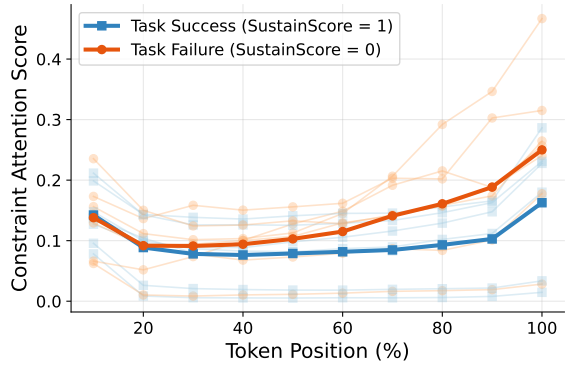


Figure 4: Comparison of constraint attention scores between successful and failed generations.

level, we conduct an attention-based analysis that tracks the model’s attention on constraint tokens throughout response generation. We define *Constraint Attention Score* to quantify how prominently constraint tokens appear among the tokens that carry most of the model’s attention mass. Specifically, at each generation step and for each attention head, we identify the smallest set of tokens whose cumulative attention weight reaches 50% proportion of the total attention mass. We then compute the fraction of these high-attention tokens that belong to the constraint and normalize it by the proportion of constraint tokens in the entire context, yielding a head-level constraint attention score. To obtain a token-level score, we average this head-level score over all attention heads in the last 4 transformer layers at the same generation step. More details are provided in Appendix B.2.

We compare the constraint attention scores of successful versus failed generations on math tasks across seven representative models. As shown in Figure 4, failed cases consistently exhibit higher constraint attention scores than successful ones across most decoding steps. This gap becomes increasingly pronounced toward the later stages of generation: while both curves rise as decoding proceeds, the failure curve grows much more sharply, indicating that unsuccessful generations allocate a disproportionate amount of attention to constraint tokens near the end of the sequence. This suggests that excessive attention to constraints may interfere with the model’s reasoning, offering a plausible explanation for the observed performance decline.

4.3 Analysis on Post-Training Factors

A critical application of any evaluation metric is to diagnose the shortcomings of existing techniques,

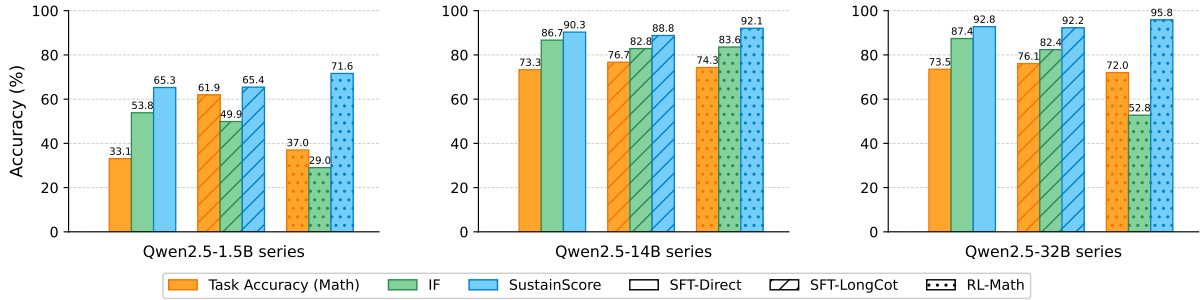


Figure 5: SUSTAINSCORE for different post-training strategies within the same model family on the Math task.

thereby guiding future model development (Zheng et al., 2023; Zhou et al., 2023). Hence, we employ SUSTAINSCORE to explore the potential impact of post-training paradigms. We compare models within the Qwen2.5 ecosystem that are built upon the same base architecture but differentiated by three alignment strategies: (1) SFT-Direct: Supervised Fine-Tuning with direct answers, the standard approach for most instruction-tuned models (e.g., the Qwen2.5-Instruct series). (2) SFT-LongCoT: Supervised Fine-Tuning on long CoT data, such as the DeepSeek-R1-Distill series (Guo et al., 2025). (3) RL-Math: Reinforcement Learning specifically for mathematical reasoning, such as the OpenReasoner-Zero series (Hu et al., 2025). As shown in Figure 5, our preliminary observations suggest that RL-based models tend to maintain better robustness alongside task performance. Conversely, SFT-LongCoT, while strong in task capability, appears more susceptible to performance degradation under constraints. This robustness gap is not fully reflected by standard metrics like IF, which focus on constraint satisfaction but overlook the preservation of core task capabilities, highlighting the complementary value of SUSTAINSCORE.

5 Related Work

Instruction Following. Instruction following constitutes a cornerstone of Large Language Model (LLM) alignment. Current instruction-following benchmarks typically assess whether a model’s response satisfies given constraints in a binary manner (Zhou et al., 2023; Qin et al., 2024; He et al., 2024b; Qi et al., 2025; Wen et al., 2024; Fu et al., 2025). However, these approaches face two significant limitations. First, constructing such benchmarks often requires extensive manual annotation and verification (Qi et al., 2025; Qin et al., 2024), thereby limiting their scalability. Second, assessing adherence to complex constraints, such as stylis-

tic conformity, frequently relies on LLM-based automatic evaluation (Jiang et al., 2024), which inevitably introduces noise and potential bias (Ye et al., 2025). More critically, effective alignment should entail not merely adhering to superficial instructions but also successfully accomplishing the core task. To this end, we propose SUSTAINSCORE, designed to capture a crucial dimension of alignment overlooked by prior instruction following works: the capability to maintain core task performance under self-evident constraints.

Prompt Sensitivity. Prompt sensitivity is commonly defined as the phenomenon where semantically equivalent reformulations of an instruction lead to substantial variations in model behavior or task performance (Chatterjee et al., 2024; Polo et al., 2024; Mizrahi et al., 2024; Voronov et al., 2024). Most prior work studies this effect through surface-level perturbations, such as changes in punctuation, formatting, or phrasing, or through paraphrasing that preserves task semantics (Razavi et al., 2025; Mizrahi et al., 2024). In contrast, we move beyond surface-level sensitivity to evaluate robustness under functional constraints (e.g., formatting or tool usage), which are common in real-world use. These constraints preserve the task goal but alter how it should be carried out. We quantify a critical robustness gap: the inability to sustain performance under constraints, a dimension not captured by classical prompt sensitivity benchmarks.

6 Conclusion

In this paper, we reveal a counterintuitive phenomenon: instruction following can paradoxically interfere with LLMs’ task-solving capability. We introduce SUSTAINSCORE to quantify this interference. Our results reveal that this interference is severe and prevalent. Additionally, attention-based analyses suggest that the performance drop

may stem from an excessive focus on constraints. We further provide preliminary insights into how different post-training paradigms influence this interference. We caution that excessive instructions may degrade model performance, and encourage future work to adopt SUSTAINSCORE to foster the development of more capable and reliable models.

Limitations

The main limitations lie in the implementation of SUSTAINSCORE: (1) Our validation is confined to mathematical reasoning, code generation, and multi-hop QA. While effective in these domains, the generalizability of SUSTAINSCORE across more diverse tasks remains to be explored. While our evaluation is fully automated and easy to scale, we encourage the community to apply our metric to a broader range of scenarios. (2) Our experiments are conducted in English. We recognize that linguistic differences could impact the results and welcome further investigation across different languages. In conclusion, we acknowledge these limitations not as shortcomings, but as important directions for future work. We believe that SUSTAINSCORE is a significant contribution, and we look forward to future research that builds more reliable models for real-world scenarios.

Ethical Statements

(1) Datasets: Our research utilizes widely-used evaluation datasets, and we strictly adhere to their licensing terms. We believe the original data is properly anonymized, and our work does not introduce any new sensitive information. (2) Methodology Risks: Our proposed metric, SUSTAINSCORE, relies on an LLM generator for soft constraint creation. It therefore inherits the inherent risks of the generator model. However, our approach does not introduce additional risks. A manual review of 200 randomly sampled instances confirmed the absence of sensitive information. (3) AI Assistance: We used AI assistants (ChatGPT and Gemini) for language refinement and editing of this manuscript.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Anthropic. 2024. [Introducing claude 3.5 sonnet](#). Accessed: 2025-05-15.

Anthropic. 2025. [Introducing claude sonnet 4.5](#). Accessed: 2025-09-29.

Linzhen Chai, Shukai Liu, Jian Yang, Yuwei Yin, Ke Jin, Jiaheng Liu, Tao Sun, Ge Zhang, Changyu Ren, Hongcheng Guo, and 1 others. 2024. Mceval: Massively multilingual code evaluation. *arXiv e-prints*, pages arXiv–2406.

Anwoy Chatterjee, HSVNS Kowndinya Renduchintala, Sumit Bhatia, and Tanmoy Chakraborty. 2024. Posix: A prompt sensitivity index for large language models. *arXiv preprint arXiv:2410.02185*.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, and 39 others. 2021. [Evaluating large language models trained on code](#).

Yang Chen, Zhuolin Yang, Zihan Liu, Chankyu Lee, Peng Xu, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2025. Acereason-nemotron: Advancing math and code reasoning through reinforcement learning. *arXiv preprint arXiv:2505.16400*.

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

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Tingchen Fu, Jiawei Gu, Yafu Li, Xiaoye Qu, and Yu Cheng. 2025. Scaling reasoning, losing control: Evaluating instruction following in large reasoning models. *arXiv preprint arXiv:2505.14810*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shiron Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

- Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, Jie Liu, Lei Qi, Zhiyuan Liu, and Maosong Sun. 2024a. *Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems*. *Preprint*, arXiv:2402.14008.
- Qianyu He, Jie Zeng, Wenhao Huang, Lina Chen, Jin Xiao, Qianxi He, Xunzhe Zhou, Jiaqing Liang, and Yanghua Xiao. 2024b. Can large language models understand real-world complex instructions? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18188–18196.
- Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. 2020. *Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps*. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6609–6625, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Jingcheng Hu, Yinmin Zhang, Qi Han, Daxin Jiang, Xianguy Zhang, and Heung-Yeung Shum. 2025. *Open-reasoner-zero: An open source approach to scaling reinforcement learning on the base model*. *Preprint*, arXiv:2503.24290.
- Yuxin Jiang, Yufei Wang, Xingshan Zeng, Wanjun Zhong, Liangyou Li, Fei Mi, Lifeng Shang, Xin Jiang, Qun Liu, and Wei Wang. 2024. *Follow-Bench: A multi-level fine-grained constraints following benchmark for large language models*. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4667–4688, Bangkok, Thailand. Association for Computational Linguistics.
- Xiaomin Li, Zhou Yu, Zhiwei Zhang, Xupeng Chen, Ziji Zhang, Yingying Zhuang, Narayanan Sadagopan, and Anurag Beniwal. 2025. When thinking fails: The pitfalls of reasoning for instruction-following in llms. *arXiv preprint arXiv:2505.11423*.
- Xinyi Li, Sai Wang, Siqi Zeng, Yu Wu, and Yi Yang. 2024. A survey on llm-based multi-agent systems: workflow, infrastructure, and challenges. *Vicinity*, 1(1):9.
- Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let’s verify step by step. *arXiv preprint arXiv:2305.20050*.
- Renze Lou, Kai Zhang, and Wenpeng Yin. 2024. Large language model instruction following: A survey of progresses and challenges. *Computational Linguistics*, 50(3):1053–1095.
- Moran Mizrahi, Guy Kaplan, Dan Malkin, Rotem Dror, Dafna Shahaf, and Gabriel Stanovsky. 2024. State of what art? a call for multi-prompt llm evaluation. *Transactions of the Association for Computational Linguistics*, 12:933–949.
- OpenAI. 2025. Introducing gpt-5. <https://openai.com/zh-Hans-CN/index/introducing-gpt-5/>. 2025-09-29.
- Arkil Patel, Satwik Bhattamishra, and Navin Goyal. 2021. Are NLP models really able to solve simple math word problems? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2080–2094, Online. Association for Computational Linguistics.
- Hao Peng, Yunjia Qi, Xiaozhi Wang, Bin Xu, Lei Hou, and Juanzi Li. 2025. Verif: Verification engineering for reinforcement learning in instruction following. *arXiv preprint arXiv:2506.09942*.
- Felipe Maia Polo, Ronald Xu, Lucas Weber, Mírian Silva, Onkar Bhardwaj, Leshem Choshen, Allysson Flavio Melo de Oliveira, Yuekai Sun, and Mikhail Yurochkin. 2024. Efficient multi-prompt evaluation of llms. *arXiv preprint arXiv:2405.17202*.
- Yunjia Qi, Hao Peng, Xiaozhi Wang, Amy Xin, Youfeng Liu, Bin Xu, Lei Hou, and Juanzi Li. 2025. Agentif: Benchmarking instruction following of large language models in agentic scenarios. *arXiv preprint arXiv:2505.16944*.
- Yunjia Qi, Hao Peng, Xiaozhi Wang, Bin Xu, Lei Hou, and Juanzi Li. 2024. Constraint back-translation improves complex instruction following of large language models. *arXiv preprint arXiv:2410.24175*.
- Yiwei Qin, Kaiqiang Song, Yebowen Hu, Wenlin Yao, Sangwoo Cho, Xiaoyang Wang, Xuansheng Wu, Fei Liu, Pengfei Liu, and Dong Yu. 2024. Infobench: Evaluating instruction following ability in large language models. *arXiv preprint arXiv:2401.03601*.
- Amirhossein Razavi, Mina Soltangheis, Negar Arabzadeh, Sara Salamat, Morteza Zihayat, and Ebrahim Bagheri. 2025. Benchmarking prompt sensitivity in large language models. In *European Conference on Information Retrieval*, pages 303–313. Springer.
- Haoran Sun, Lixin Liu, Junjie Li, Fengyu Wang, Baohua Dong, Ran Lin, and Ruohui Huang. 2024. Conifer: Improving complex constrained instruction-following ability of large language models. *arXiv preprint arXiv:2404.02823*.
- Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. MuSiQue: Multi-hop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*.
- Anton Voronov, Lena Wolf, and Max Ryabinin. 2024. Mind your format: Towards consistent evaluation of in-context learning improvements. *arXiv preprint arXiv:2401.06766*.

- Bosi Wen, Pei Ke, Xiaotao Gu, Lindong Wu, Hao Huang, Jinfeng Zhou, Wenchuang Li, Binxin Hu, Wendy Gao, Jiabin Xu, and 1 others. 2024. Benchmarking complex instruction-following with multiple constraints composition. *arXiv preprint arXiv:2407.03978*.
- xAI. 2025. Grok 4 fast model card. <https://data.x.ai/2025-09-19-grok-4-fast-model-card.pdf>. Last updated: September 19, 2025.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.
- Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner Geyer, Chao Huang, Pin-Yu Chen, and 1 others. 2025. Justice or prejudice? quantifying biases in llm-as-a-judge. In *International Conference on Learning Representations*.
- Aohan Zeng, Xin Lv, Qinkai Zheng, Zhenyu Hou, Bin Chen, Chengxing Xie, Cunxiang Wang, Da Yin, Hao Zeng, Jiajie Zhang, and 1 others. 2025. Glm-4.5: Agentic, reasoning, and coding (arc) foundation models. *arXiv preprint arXiv:2508.06471*.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. [Judging llm-as-a-judge with mt-bench and chatbot arena](#). *Preprint*, arXiv:2306.05685.
- Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*.

Appendices

A SUSTAINSCORE Details

Tables 5, 7, and 6 list the prompts for the *Method*, *Structure*, and *Style* constraint types in SUSTAINSCORE, while Table 8 provides those used for the instruction length analysis (§ 4.1). For constraint generation, we employ Qwen3-32B (Yang et al., 2025) with a maximum token length of 32,000 and temperature set to 0. Finally, across all evaluations, we use a maximum token length of 8,192 and temperature 0 to ensure reproducibility.

B Detailed Results

B.1 Error Analysis

Tables 9,10, provide representative examples of the two error types. Figure 6 reports their distribution across eight models on math tasks, including 4 direct-answer models and 4 chain-of-thought reasoning models. Because many output specification errors stem from formatting mismatches that cause exact-match or regex-based metrics to fail despite correct underlying solutions, we use an LLM-as-judge to determine semantic correctness. Outputs judged as correct by the LLM are labeled as output specification errors; otherwise, they are considered reasoning errors.

B.2 Constraint Attention Score

Constraint Attention Score. Let $a_{t,l,h}(i)$ denote the attention weight assigned to the i -th token in the context at generation step t , layer l , and attention head h , where $i = 1, \dots, N_t$ and N_t is the number of tokens in the context. Let $S_2 \subseteq \{1, \dots, N_t\}$ be the set of constraint tokens.

Top- p attention set. Given a proportion threshold $p = 0.5$, we define the minimal top- p attention set as

$$\mathcal{T}_{t,l,h}^{(p)} = \arg \min_{S \subseteq \{1, \dots, N_t\}} \left\{ \sum_{i \in S} a_{t,l,h}(i) \geq p \sum_{j=1}^{N_t} a_{t,l,h}(j) \right\}. \quad (2)$$

Head-level score. We define the head-level constraint attention score as

$$c_{t,l,h}^{\text{global}} = \frac{|\mathcal{T}_{t,l,h}^{(p)} \cap S_2|}{|\mathcal{T}_{t,l,h}^{(p)}|} \bigg/ \frac{|S_2|}{N_t} \quad (3)$$

This score measures how much constraint tokens are over-represented in the high-attention region relative to their random baseline.

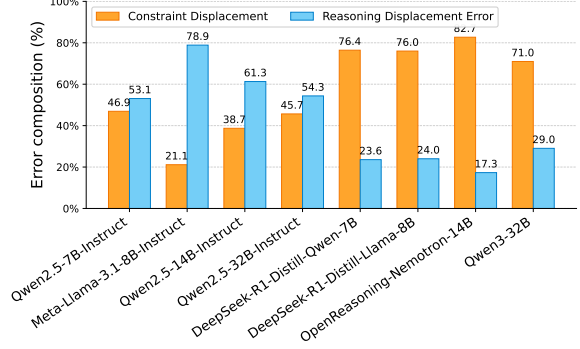


Figure 6: SUSTAINSCORE on different constraint types on Math. The score for each type represents the average performance across all models evaluated in our study.

Token-level score. Let \mathcal{L}_K denote the set of the last K transformer layers (we use $K = 4$), and let H_l be the number of attention heads in layer l . The token-level constraint attention score at generation step t is defined as

$$c_t^{\text{global}} = \frac{1}{K} \sum_{l \in \mathcal{L}_K} \left(\frac{1}{H_l} \sum_{h=1}^{H_l} c_{t,l,h}^{\text{global}} \right). \quad (4)$$

B.3 Results on Post-Training Analysis

Table 4 presents the results discussed in Section 4.3. We additionally include the LLaMA series in the table below. These models exhibit the same trends, further supporting our observations.

Series	Model	Task Accuracy	SUSTAINSCORE	IF
Qwen2.5-1.5B series				
SFT-Direct	Qwen2.5-1.5B-Instruct	33.1	65.3	53.8
SFT-LongCot	DeepSeek-R1-Distill-Qwen-1.5B	61.9	65.4	49.9
RL-Math	Open-Reasoner-Zero-1.5B	37.0	71.6	29.0
Qwen2.5-14B series				
SFT-Direct	Qwen2.5-14B-Instruct	73.3	90.3	86.7
SFT-LongCot	DeepSeek-R1-Distill-Qwen-14B	76.7	88.8	82.8
RL-Math	AceReason-Nemotron-14B	74.3	92.1	83.6
Qwen2.5-32B series				
SFT-Direct	Qwen2.5-32B-Instruct	73.5	92.8	87.4
SFT-LongCot	DeepSeek-R1-Distill-Qwen-32B	76.1	92.2	82.4
RL-Math	Open-Reasoner-Zero-32B	72.0	95.8	52.8
Llama3.1-8B series				
SFT-Direct	Llama3.1-8B-Instruct	56.1	87.2	83.5
SFT-LongCot	DeepSeek-R1-Distill-Llama-8B	64.8	75.2	69.9
RL-Math	Llama3.1-8B-Grpo-Reasoning	53.3	88.0	82.6

Table 4: Detailed results of the analysis of alignment techniques

You will be given a model response. Extract the core *answering approach* it uses and rewrite it as a single, imperative instruction that can be attached to future prompts.

Instructions

- Return **JSON only** with the key "approach_instruction".
- The value must be **one sentence** (≤ 25 words) starting with: **"Your answer should follow this approach: "**.
- Summarize the **strategy**, not problem-specific facts or solutions.
- **No explanations, no extra keys, no text outside JSON.**
- If the approach is implicit, infer the most plausible high-level method.

Example

Input Response: "Step 1: Kim's Contribution: Kim raised \$320 more than Alexandra, who raised \$430. \$430 + \$320 = \$750. Step 2: Maryam's Contribution: Maryam raised \$400 more than Sarah, who raised \$300. \$300 + \$400 = \$700. Step 3: Total Amount: We add: \$430 (Alexandra) + \$750 (Kim) + \$300 (Sarah) + \$700 (Maryam) = \$2180. Final Answer: 2180."

Output: {"approach_instruction": "Your answer should follow this approach: compute each person's amount from provided relations, then sum all contributions; show step-by-step arithmetic leading to the final total."}

Your Turn

Input Response: "{response}"

Output (Please directly return JSON only):

Table 5: Prompt for extracting Method constraints

You will be given a model response. Extract the core *language style* it uses and rewrite it as a single, imperative instruction that can be attached to future prompts.

Instructions

- Return **JSON only** with the key "style_instruction".
- The value must be **one sentence** (≤ 25 words) starting with: **"Your answer should adopt this style: "**.
- Summarize the **tone / style**, not content or problem-specific facts.
- **No explanations, no extra keys, no text outside JSON.**
- If the style is implicit, infer the most plausible high-level style.

Example

Input Response: "Don't worry, you can think of it step by step. Step 1: Kim's Contribution: Kim raised \$320 more than Alexandra, who raised \$430. \$430 + \$320 = \$750. Step 2: Maryam's Contribution: Maryam raised \$400 more than Sarah, who raised \$300. \$300 + \$400 = \$700. Step 3: Total Amount: To find the total, we add the amounts from all four girls: \$430 (Alexandra) + \$750 (Kim) + \$300 (Sarah) + \$700 (Maryam) = \$2180. Final Answer: #### 2180."

Output: {"style_instruction": "Your answer should adopt this style: patient and encouraging, guiding the reader step by step like a supportive tutor."}

Your Turn

Input Response: "{response}"

Output (Please directly return JSON only):

Table 6: Prompt for extracting Style constraints

You will be given a model response. Extract the core *answer structure* it uses and rewrite it as a single, imperative instruction that can be attached to future prompts.

Instructions

- Return **JSON only** with the key "structure_instruction".
- The value must be **one sentence** (≤ 25 words) starting with: **"Your answer should follow this structure: "**.
- Summarize the **structural format** (e.g., step numbering, bullet points, table layout, paragraph style, equation blocks), not content or style.
- **No explanations, no extra keys, no text outside JSON.**
- If the format is implicit, infer the most plausible high-level structure.

Example

Input Response: "Don't worry, you can think of it step by step. Step 1: Kim's Contribution: Kim raised \$320 more than Alexandra, who raised \$430. $\$430 + \$320 = \$750$. Step 2: Maryam's Contribution: Maryam raised \$400 more than Sarah, who raised \$300. $\$300 + \$400 = \$700$. Step 3: Total Amount: To find the total, we add the amounts from all four girls: \$430 (Alexandra) + \$750 (Kim) + \$300 (Sarah) + \$700 (Maryam) = \$2180. Final Answer: #### 2180."

Output: {"structure_instruction": "Your answer should follow this structure: sequential steps explicitly labeled as 'Step {{number}}: ...'."}

Your Turn

Input Response: "{response}"

Output (Please directly return JSON only):

Table 7: Prompt for extracting Structure constraints

You are an editor.

[Task]

Expand the input paragraph to approximately {TARGET_WORDS} words ($\pm 10\%$) while preserving its original meaning exactly.

#=====

[Input]

type: {TASK_TYPE} # one of: "math", "multi_hop_qa"
paragraph: {PARAGRAPH}

#=====

[Allowed expansion methods] (content-preserving only)

- Rephrase with synonyms or equivalent constructions that restate the same constraints.
- Add discourse connectives and hedging/flow words that do not introduce requirements (e.g., “in particular”, “more specifically”, “in other words”, “please”, “carefully”).
- Split or merge sentences; use parallel phrasing that reiterates existing points with equivalent wording.
- Expand abbreviations already present; add appositives that restate an existing noun phrase without adding attributes.
- Mild fluency padding (e.g., “clearly”, “precisely”), only if it does not add evaluation or new properties.

#=====

[Global hard constraints] (must all hold)

1) No new information:

- Do NOT add facts, entities, numbers, dates, locations, steps, options, claims, premises, constraints, caveats, or assumptions that are not already present.
- Do NOT change scope, difficulty, intent, or acceptance criteria.
- Do NOT add examples, citations, references, safety notes, or background knowledge.

2) Preserve every original detail AND the answer-format requirements:

- Keep all entities, quantities, units, variables, symbols, equations/inequalities, and conditions unchanged in meaning.
- **Do not omit, weaken, or alter any explicit answer-format or output-style instructions present in the paragraph.**
 - * Examples include (but are not limited to): “Your final answer must begin with ‘####’ ”, etc.
 - * You may restate these format directives verbatim or with equivalent wording, but their requirements must remain exactly the same.
 - * Do not introduce any new formatting requirements not in the input.
- Keep the same language as the input (English is expected).
- Maintain the same modality/tone (e.g., request/imperative/neutral).

3) Output format (your response to this prompt):

- Return ONLY a JSON object with exactly these keys: {{"expanded": string}}
- No additional fields, preface, or commentary outside the JSON.

#=====

[Type-specific constraints]

– If type == “math”:

- * Do not alter the mathematical setup, variable names, symbols, equations, inequalities, or numeric values.
- * Do not change what is being asked (objective) or any given constraints, bounds, or domain assumptions.
- * You may add connective phrases or restate the same condition, but must not derive intermediate results or add hints.

– If type == “multi_hop_qa”:

- * Preserve all entities, events, temporal and spatial qualifiers, and relations between them.
- * Do not add external facts, missing links, or bridging knowledge; only restate what is already present.
- * You may clarify with equivalent wording (e.g., restating a constraint in different words), but do not introduce new premises.

[Length target]

Aim for {TARGET_WORDS} words; acceptable range is $\pm 10\%$. If exact targeting would harm fidelity, prefer fidelity and stay inside the range.

#=====

Produce the JSON now.

Table 8: Prompt for paraphrasing an instruction with a length constraint

[Instruction]

Rani has ten more crabs than Monic, who has 4 fewer crabs than Bo. If Bo has 40 crabs, calculate the total number of crabs the three have together.

Your final answer must begin with '####' and should appear as the last part of your response, in accordance with the required format. Use: ####{Your Answer}.

#=====

[Constraints]

- Ensure that the keyword 'metal' is not present in your response.
- Apply this approach when formulating your answer: determine each person's quantity based on given relationships, then sum all quantities; show step-by-step calculations leading to the final total.
- Write at least 18 full sentences.
- Your answer should follow this style: structured and logical, using clear headings and step-by-step explanations.

#=====

[Model Response]

Solving the Crab Problem: A Step-by-Step Analysis

Understanding the Problem

Let me break down this problem systematically to find the total number of crabs owned by Rani, Monic, and Bo. We are given specific relationships between the quantities each person possesses, and we need to work through these relationships carefully.

Given Information

The problem provides us with the following key pieces of information. Bo has 40 crabs, which serves as our starting point. Monic has 4 fewer crabs than Bo, which means we need to subtract 4 from Bo's total. Rani has ten more crabs than Monic, so we'll need to add 10 to Monic's total.

Step 1: Determining Bo's Crab Count

Bo's crab count is explicitly stated in the problem. Bo has 40 crabs, which is our foundation for all subsequent calculations. This value will help us determine the quantities for both Monic and Rani.

Step 2: Calculating Monic's Crab Count

Now I need to find how many crabs Monic has. The problem states that Monic has 4 fewer crabs than Bo. Since Bo has 40 crabs, I subtract 4 from 40 to get Monic's total.

$$\text{Monic's crabs} = \text{Bo's crabs} - 4$$

$$\text{Monic's crabs} = 40 - 4$$

$$\text{Monic's crabs} = 36$$

Therefore, Monic has 36 crabs.

Step 3: Calculating Rani's Crab Count

Next, I need to determine Rani's crab count. The problem tells us that Rani has ten more crabs than Monic. Since we just calculated that Monic has 36 crabs, I add 10 to this amount.

$$\text{Rani's crabs} = \text{Monic's crabs} + 10$$

$$\text{Rani's crabs} = 36 + 10$$

$$\text{Rani's crabs} = 46$$

Therefore, Rani has 46 crabs.

Step 4: Finding the Total

Now that we have determined each person's individual crab count, we can calculate the total. We simply add together the crabs owned by Bo, Monic, and Rani.

$$\text{Total crabs} = \text{Bo's crabs} + \text{Monic's crabs} + \text{Rani's crabs}$$

$$\text{Total crabs} = 40 + 36 + 46 = 86$$

Verification of Our Answer

Let me verify this solution by checking our relationships. Bo has 40 crabs, Monic has 36 crabs (which is indeed 4 fewer than 40), and Rani has 46 crabs (which is indeed 10 more than 36). All relationships check out correctly, confirming our calculations are accurate.

The three individuals together have a combined total of 86 crabs.

####86

#=====

[Reference Answer]

122

Table 9: Reasoning Error example. A failure case from Claude-Sonnet-4.5 on GSM8K with additional constraints. The erroneous step is highlighted in red.

[Instruction]

Write a C function `int are_intervals_intersecting(int a, int b, int c, int d)` to solve the following problem: Determine if two closed intervals intersect.

Args: `a, b` represent the first closed interval $[a, b]$ where $0 \leq a \leq b \leq 1000$;
`c, d` represent the second closed interval $[c, d]$ where $0 \leq c \leq d \leq 1000$.

Returns: 1 if the intervals intersect, 0 otherwise.

#####

[Constraints]

- Do not let your code span more than 49 lines.
- Apply this approach when formulating your answer: apply logical conditions based on interval boundary comparisons to determine intersection.
- Please validate that your function or variable names include the keyword “close”.

#####

[Model Response]

```
/**
 * @brief Determines if two closed intervals intersect.
 *
 * This function applies logical conditions based on interval boundary comparisons
 * to determine if two closed intervals [a, b] and [c, d] have any common points.
 *
 * @param a The start point of the first closed interval [a, b].
 * @param b The end point of the first closed interval [a, b].
 * @param c The start point of the second closed interval [c, d].
 * @param d The end point of the second closed interval [c, d].
 * @return int 1 if the intervals intersect, 0 otherwise.
 */
int are_close_intervals_intersecting(int a, int b, int c, int d) {
    // Two closed intervals [a, b] and [c, d] intersect if:
    // 1. The end of the first interval (b) is not strictly before the start of the second (c). (b
    >= c)
    // AND
    // 2. The start of the first interval (a) is not strictly after the end of the second (d). (a
    <= d)
    // If both conditions are true, the intervals intersect.
    return (b >= c && a <= d) ? 1 : 0;
}
```

Table 10: *Output Specification Error* example. A failure case from Gemini-2.5-Flash on MCEval with additional constraints. The erroneous step is highlighted in red.