CATS: Category-Aware Token-level Steering for Training-Free Redundancy Reduction in Large Reasoning Models

Anonymous Author(s)
Affiliation
Address
email

Abstract

While Large Reasoning Models (LRMs) exhibit remarkable capabilities in complex tasks, they often suffer from excessive redundancy in their chain-of-thought reasoning. This significantly reduces inference efficiency and increases computational costs. We identify that LRM redundancy is not uniformly homogeneous but can be taxonomized according to whether it is destructive to the final answer: destructive redundancy (e.g., logical drift, hallucination amplification) versus non-destructive redundancy (e.g., repetition, over-elaboration). Moreover, LRM's redundant and concise responses exhibit a significant distinction in their hidden layer representation spaces. Based on these insights, we propose CATS (Category-Aware Token-Level Steering), a training-free and lightweight method to reduce the redundancy phenomenon. CATS decomposes redundancy into six semantically interpretable characteristic dimensions. By flexibly weighting and combining the differential vectors corresponding to these dimensions, CATS synthesizes a composite intervention vector, enabling zero-parameter intervention in the hidden layers. Experiments across three LRM models and five mathematical reasoning datasets demonstrate that CATS reduces reasoning length by an average of 25% while maintaining or even slightly improving task accuracy. CATS offers a pluggable, training-free, and lightweight solution, making it particularly beneficial for users in low-resource environments.Our code can be found at https://anonymous.4open.science/r/cats-63B6

1 Introduction

Large Reasoning Models (LRMs) have demonstrated powerful capabilities in complex reasoning tasks through Chain-of-Thought (CoT) prompting [1, 2] . However, pervasive reasoning redundancy in CoT paths poses a significant challenge. Such redundancy not only increases inference latency and computational costs but may also introduce logical errors or hallucinations. Existing redundancy elimination methods, such as Supervised Fine-Tuning (SFT) [3–5], Reinforcement Learning (RL) [6, 7], or Prompt Engineering [8–10], often demand substantial computational resources and struggle with fine-grained control over redundancy types, thereby limiting their deployment flexibility.

We observe that redundancy in LRMs is not a uniformly homogeneous phenomenon. Instead, it can be taxonomized into six semantically interpretable characteristic dimensions. These dimensions can be further categorized into destructive redundancy (e.g., logical drift, hallucination amplification, internal contradiction) and non-destructive redundancy (e.g., repetition, over-elaboration, over-caution). More crucially, by analyzing the hidden layer representation space, we discover significant divergences between concise paths and various redundant paths in the representation space (see Appendix A), providing a theoretical basis for precise intervention.

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

Inspired by these insights, we proposes CATS (Category-Aware Token-level Steering), a training-free and lightweight framework designed to systematically reduce reasoning redundancy through zero-parameter intervention in LRM hidden layers. The core idea of CATS is to extract differential vectors corresponding to each redundancy type and dynamically combine and apply these vectors to the model's hidden layers during inference, thereby precisely suppressing specific types of redundancy. Our main contributions include:

- We are the first to categorize LRM redundancy into six semantically interpretable types and demonstrate their distinguishable representations in hidden layers.
- We propose CATS, a novel training-free intervention framework that enables fine-grained control over LRM redundancy through flexibly weighted and combined differential vectors.
- Extensive experiments across three LRM models and five mathematical reasoning datasets demonstrate that CATS reduces reasoning length by an average of 25% while maintaining or even slightly improving task accuracy, and significantly reducing the incidence of destructive redundancy by 21%.

2 Method

This section introduces the CATS framework, which focuses on reducing LRM reasoning redundancy through interventions in the model's hidden layers.
Given an input query x, a LRM generates a multi-step reasoning output $y = \{t_1, t_2, \ldots, t_n\}$. Our objective is to minimize the length of the reasoning chain $|y|$ without compromising the accuracy of the final answer:

$$\min |y| \quad s.t. \quad acc(y) \geq acc(y_0) \tag{1}$$

where $acc(y_0)$ is the accuracy of the original model output.

The CATS framework comprises three key stages (as illustrated in Figure 1):

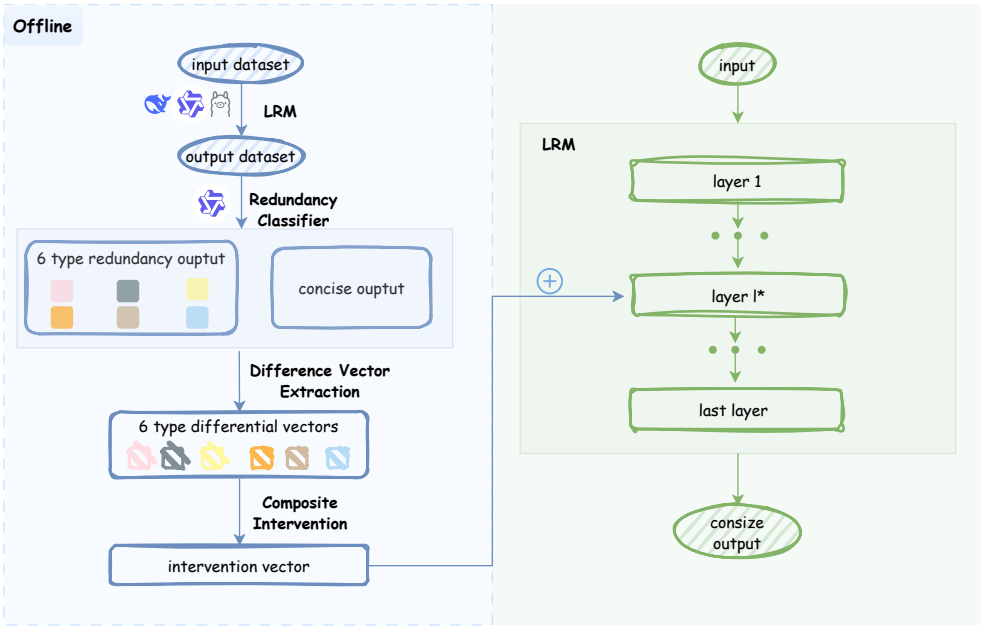

Figure 1: Overview of the CATS (Category-Aware Token-level Steering) Framework

2.1 Redundancy Classification

We classify LRM reasoning redundancy into six semantic categories by combining large language model annotation with human verification. These categories are primarily divided into two types:

- **Destructive Redundancy:** Includes logical drift, hallucination amplification, and internal contradiction. These redundancies can directly lead to reasoning failures or incorrect results.
- **Non-Destructive Redundancy:** Includes repetition, over-elaboration, and over-caution. These redundancies primarily increase reasoning length without directly causing logical errors.

This fine-grained categorization allows for a more precise understanding and suppression of different types of redundancy. Detailed classification rules and classifier evaluation are provided in the Appendix B.

2.2 Difference Vector Extraction

We extract difference vectors by analyzing the hidden status differences in LRM between concise reasoning paths and redundant reasoning paths. We selected 1,000 mathematical problems from the GSM8K dataset and generated 10 reasoning paths for each problem. Through annotation by Qwen3-235B-A22B, we constructed a concise sample set S_c and a redundant sample set $S_{r,k}$ for each redundancy type k.

For layer l of the LRM, we compute the average hidden state of the concise samples $\mu^l_{S_c}$ and the average hidden state for each redundancy type k, $\mu^l_{S_{r,k}}$. The difference vector v^l_k for each redundancy type k is defined as the difference between the average hidden states of the concise samples and the samples of that redundancy type:

$$v^l_k = \mu^l_{S_c} - \mu^l_{S_{r,k}} \tag{2}$$

These vectors capture the semantic information for the model in the direction of suppressing specific redundancy types.

2.3 Composite Intervention

During the inference phase, CATS linearly combines the extracted differential vectors and adds them to the hidden state of a specific model layer, thereby achieving redundancy suppression. The intervention formula is as follows:

$$h'_{l^*,-1} = h_{l^*,-1} + \sum_{k=1}^{6} w_k \cdot v^{l^*}_k \tag{3}$$

where $h_{l^*,-1}$ is the hidden state of the original LRM at the last token of the input to layer l^*; $v^{l^*}_k$ is the differential vector for the k-th redundancy type; and w_k are user-configurable weights used to adjust the suppression strength for each redundancy type. This process involves only a single forward pass, requiring no additional training, making it a zero-parameter intervention with extremely high deployment efficiency and flexibility.

3 Experiments

This section details the experimental configuration used to validate the effectiveness of the CATS method, including selected models, datasets, evaluation metrics, and baselines.

3.1 Setup

We evaluated CATS on DeepSeek-R1-Distill-Qwen-1.5B, DeepSeek-R1-Distill-Qwen-7B, and DeepSeek-R1-Distill-Llama-8B models [1]. Experiments were conducted on five mathematical reasoning datasets: MATH-500[11], AMC, AMC23, AIME 2024, and AIME 2025[12]. We primarily assessed the method's effectiveness using task accuracy, average token length, and incidence rate of each redundancy type. All experimental details are available in the Appendix C.

3.2 Baselines

We compared CATS against the following baselines:

- **Original LRM:** The raw model without any intervention.
- **Generic Intervention:** This baseline uses only a single "longest-concise" differential vector for intervention, not distinguishing between redundancy types. This aims to simulate approaches that treat all redundancy as homogeneous, highlighting the advantages of fine-grained classification.

3.3 Main Results

As shown in tabel 1, CATS consistently demonstrated significant redundancy reduction capabilities across all models and datasets. On average, CATS reduced the reasoning path length by 25.0% while maintaining or even slightly improving task accuracy in most cases.
Moreover, CATS outperforms the generic intervention baseline in both length compression and accuracy. This highlights the critical role of fine-grained redundancy classification, enabling CATS to precisely target different redundancy types for a more intelligent and safe efficiency optimization.

Table 1: Overview of Main Results of CATS across Different Models and Datasets

	MATH-500		AMC		AMC23		AIME 2024		AIME 2025	
	len(↓)	acc(↑)	len(↓)	acc(↑)	len(↓)	acc(↑)	len(↓)	acc(↑)	len(↓)	acc(↑)
DeepSeek-R1-Distill-Qwen-1.5B	4528.31	83.21	8982.86	58.52	8697.42	55.42	11758.16	27.54	11709.96	25.43
+ Generic Intervention	3935.10	79.55	6902.43	57.12	7338.45	54.25	10637.61	26.54	10503.83	24.80
+ CATS	**3577.36**	82.01	**6108.34**	**60.13**	**6523.07**	**56.22**	**9759.27**	**27.94**	**9133.77**	**25.83**
DeepSeek-R1-Distill-Qwen-7B	3817.48	92.96	6572.61	79.07	6204.70	81.03	10220.63	51.64	10051.13	36.85
+ Generic Intervention	3169.55	91.34	4814.84	76.20	5425.86	78.15	9390.91	48.66	8661.53	34.55
+ CATS	**2855.45**	**93.20**	**4115.25**	**79.37**	**4559.55**	**82.26**	**7825.76**	**52.32**	**7278.60**	**37.15**
DeepSeek-R1-Distill-Llama-8B	3863.48	89.73	6878.65	77.80	6639.62	79.21	11404.50	44.80	11579.90	30.84
+ Generic Intervention	3256.30	87.81	5175.50	76.65	5969.68	78.81	10318.79	43.83	9944.82	32.17
+ CATS	**2936.24**	**90.06**	**4539.91**	**78.21**	**5378.09**	**80.01**	**8895.51**	**46.14**	**8800.72**	**34.22**
avg. Δlen = -25%, avg. Δacc = 1.2%										

Table 2: Occurrence Rates of Redundancy Categories in Model Responses

	Destructive Redundancy			Non-Destructive Redundancy		
	Log. Drift	Hall. Amp.	Int. Contr.	Rep.	Over-Elab.	Over-Caut.
DeepSeek-R1-Distill-Qwen-1.5B	39.09	3.20	4.90	15.81	13.65	13.70
+ CATS	**31.27**	**2.62**	**3.43**	**11.38**	**10.51**	**10.41**
DeepSeek-R1-Distill-Qwen-7B	31.50	3.60	4.40	15.94	17.82	16.57
+ CATS	**26.15**	**2.81**	**3.34**	**11.64**	**12.47**	**12.10**
DeepSeek-R1-Distill-Llama-8B	32.20	2.70	5.40	16.72	14.57	17.37
+ CATS	**25.44**	**1.89**	**3.24**	**11.70**	**11.22**	**12.16**
avg. ΔOccurrence Rate of Destructive Redundancy = -21%						
avg. ΔOccurrence Rate of Non-Destructive Redundancy = -27%						

From Table 2, CATS intervention effectively reduced the occurrence rate of all redundancy categories. Specifically, the average occurrence rate of destructive redundancy decreased by 21%, while that of non-destructive redundancy decreased by 27%, further enhancing the conciseness and reliability of model outputs.Further analysis on ablation study is show in Appendix D.

4 Conclusion

This paper presents CATS, a novel training-free framework for category-aware, token-level redundancy reduction in large reasoning models. By decomposing redundancy into six semantically interpretable types and leveraging differential vectors in hidden layers for precise intervention, CATS successfully reduced reasoning length by an average of 25% while maintaining or slightly improving task accuracy, and effectively suppressing destructive redundancy. CATS's lightweight, training-free, and highly interpretable nature makes it a powerful tool for optimizing LRM inference efficiency. Future work will explore adaptive weight optimization and extend its application to larger-scale models and broader multimodal reasoning scenarios.

References

[1] Daya Guo et al. “Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning”. In: *arXiv preprint arXiv:2501.12948* (2025).

[2] Aaron Jaech et al. “Openai o1 system card”. In: *arXiv preprint arXiv:2412.16720* (2024).

[3] Tergel Munkhbat et al. “Self-training elicits concise reasoning in large language models”. In: *arXiv preprint arXiv:2502.20122* (2025).

[4] Heming Xia et al. “Tokenskip: Controllable chain-of-thought compression in llms”. In: *arXiv preprint arXiv:2502.12067* (2025).

[5] Yu Kang et al. “C3ot: Generating shorter chain-of-thought without compromising effectiveness”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 39. 23. 2025, pp. 24312–24320.

[6] Haotian Luo et al. “O1-pruner: Length-harmonizing fine-tuning for o1-like reasoning pruning”. In: *arXiv preprint arXiv:2501.12570* (2025).

[7] Bairu Hou et al. “Thinkprune: Pruning long chain-of-thought of llms via reinforcement learning”. In: *arXiv preprint arXiv:2504.01296* (2025).

[8] Matthew Renze and Erhan Guven. “The benefits of a concise chain of thought on problem-solving in large language models”. In: *2024 2nd International Conference on Foundation and Large Language Models (FLLM)*. IEEE. 2024, pp. 476–483.

[9] Tingxu Han et al. “Token-budget-aware llm reasoning”. In: *arXiv preprint arXiv:2412.18547* (2024).

[10] Silei Xu et al. “Chain of draft: Thinking faster by writing less”. In: *arXiv preprint arXiv:2502.18600* (2025).

[11] Hunter Lightman et al. “Let’s verify step by step”. In: *The Twelfth International Conference on Learning Representations*. 2023.

[12] Mislav Balunović et al. “Matharena: Evaluating llms on uncontaminated math competitions”. In: *arXiv preprint arXiv:2505.23281* (2025).

[13] Weixiang Zhao et al. “Exploring and Exploiting the Inherent Efficiency within Large Reasoning Models for Self-Guided Efficiency Enhancement”. In: *arXiv preprint arXiv:2506.15647* (2025).

A Hidden State Visualization

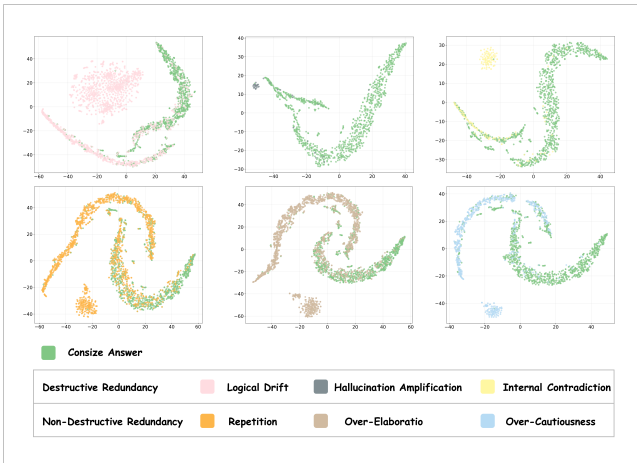

Figure 2: Hidden State Differences between Redundancy Categories and Concise Paths (t-SNE Visualization). This figure presents the t-SNE dimensionality reduction visualization of the hidden states from the 15th layer of the DeepSeek-R1-Distill-Qwen-7B model. It clearly shows a distinct separation of the six redundancy categories from concise paths in the representation space, validating the discriminability of our category-specific difference vectors. Similar patterns were observed in other models and layers.

B Redundancy Classification Details

B.1 Redundancy Definition

We categorize redundant content into two main types, encompassing six specific kinds of redundancy (as shown in Figure 3):

Destructive Redundancy

This type of redundancy interferes with the original reasoning path and reduces the correctness of the final answer.

- **Logical Drift**: Deviation from the task objective during reasoning, producing reasoning steps in an unrelated direction.
- **Hallucination Amplification**: Continuously expanding reasoning based on false or unestablished premises, constructing seemingly plausible but actually incorrect paths.
- **Internal Contradiction**: The reasoning path contains logically inconsistent statements, where the subsequent reasoning attempts to reconcile or proceeds despite these inconsistencies.

Non-Destructive Redundancy

This type of redundancy does not affect the correctness of the final answer but significantly increases output length and reduces reasoning efficiency.

- **Repetition**: Repeatedly expressing existing information or rephrasing the same content with different wordings.
- **Over-Elaboration**: Introducing definitions, background, or redundant explanations unrelated to problem-solving, thereby extending the reasoning chain.
- **Over-Cautiousness**: Using vague expressions such as "might", "probably", or "seems like" to describe definitive information.

Classification Process: We utilize the Qwen3-235B-A22B model as a classifier to automatically label redundancy types in model outputs. For each category, 50 samples were randomly selected and manually reviewed (two independent annotators, majority voting for decisions, Kappa coefficient of 0.86) to ensure classification quality. The distribution of each redundancy category in DeepSeek-R1-Distill-Qwen-7B model outputs is shown in Figure 4.

Figure 3: Examples of Redundancy Categorization. This figure illustrates our proposed six semantically interpretable redundancy categories using a mathematical reasoning problem.

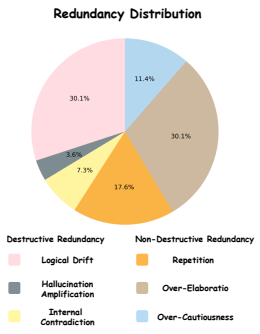

Figure 4: Distribution of Redundancy Categories in DeepSeek-R1-Distill-Qwen-7B Model Outputs

B.2 Classification Prompt

```
You are a professional text analysis assistant, skilled at identifying redundancy phenomena in large language model (LLM) responses.
Your task is to analyze the LLM response I provide and categorize it into the redundancy categories I define.

Redundancy Category Definitions:
Detrimental Redundancy: This type of redundancy directly causes or exacerbates errors, inconsistencies, or unreliability in model outputs.
1. Logical Deviation: During reasoning, the model forgets what it was supposed to do and starts talking randomly about various things, causing redundancy and errors.
2. Hallucination Amplification: The model generates a non-existent false premise during reasoning, then continuously reasons based on this premise to reach a final "answer", causing redundancy of false premises and incorrect reasoning.
3. Internal Inconsistency: The model reaches two completely contradictory conclusions during reasoning, then continues to debate between these two conclusions, causing redundant argumentation.

Non-Detrimental Redundancy: This type of redundancy mainly causes inefficiency and information overload, but usually does not directly lead to core answer errors.
4. Repetition/Restatement: Repeatedly saying the same thing without providing new information or substantial progress.
5. Unnecessary Elaboration: Providing too many, overly specific, or unnecessary examples, analogies, background information, or detailed descriptions for a point or step. For example, when asked "what is 1+1?", starting from the definition of addition.
6. Over-Cautiousness/Defensiveness: Adding too many unnecessary qualifiers, possibility descriptions (like "maybe", "perhaps", "in some cases"), disclaimers, or listing all possible edge cases when they're not actually necessary.

Classification Instructions:
1. I will provide a "user question" with a "correct answer", and the LLM's "response" to that question.
2. Please carefully read and analyze the LLM's response.
3. Determine whether the response contains the redundancy types defined above.
4. If redundancy exists, please select at least one category that best represents the redundancy phenomenon. If multiple types of redundancy exist, please list all applicable categories and indicate the most prominent one.
5. If the response is very concise and contains no redundancy, please mark it as "none".
6. Your output format must be the following JSON format:
{
  "redundancy_type": ["main redundancy type, return redundancy type number (1-6)"],
  "secondary_redundancy_type": ["other applicable redundancy types, empty list if none"],
  "analysis_explanation": "Brief explanation of why it's categorized as this type, and point out specific redundant parts in the response."
}

Example:
User Question: "Is the Earth square?"
```

```
LLM Response: "The Earth is not square. From what we know about
    ↪ physics and astronomy, the Earth is an approximately
    ↪ spherical body, more accurately described as an oblate
    ↪ spheroid, because the equator is slightly bulged and the
    ↪ poles are slightly flattened. This shape is the result of
    ↪ the combined action of centrifugal force generated by
    ↪ Earth's rotation and its own gravity. Therefore,
    ↪ describing it as 'round' in everyday context is completely
    ↪  understandable and acceptable, but in the precise
    ↪ definition of science and astronomy, it is not a perfect
    ↪ geometric sphere. For example, Earth's ocean tidal
    ↪ movements, atmosphere, and internal geological structure
    ↪ all affect its precise shape, and these are all reasons
    ↪ why it's not completely a perfect sphere."
Your output:
{
  "redundancy_type": ["6"],
  "secondary_redundancy_type": ["5"],
  "analysis_explanation": "When answering a simple yes/no
      ↪ question, the LLM added too many unnecessary qualifiers
      ↪ and scientific details, such as 'strictly speaking it's
      ↪ an oblate spheroid', 'understandable in everyday context
      ↪ ' disclaimers, and excessive explanations for why its
      ↪ shape is imperfect, showing excessive caution and
      ↪ defensiveness. At the same time, these excessive
      ↪ scientific details also constitute unnecessary
      ↪ elaboration for a simple affirmative/negative answer."
}

User Question: "{input_text}"
Correct Answer: "{final_answer}"
LLM Response: "{generated_text}"
```

C Experimental Detail

Models

To evaluate the performance of CATS, we selected the following three representative reasoning-optimized models from the DeepSeek-R1-Distill series as experimental subjects:

- DeepSeek-R1-Distill-Qwen-1.5B
- DeepSeek-R1-Distill-Qwen-7B
- DeepSeek-R1-Distill-Llama-8B

These models are representative in their reasoning capabilities and encompass different underlying architectures, which helps to investigate CATS's effectiveness under various conditions.

Datasets

Experiments were primarily validated on the following five classic mathematical reasoning datasets:MATH-500 [11], AMC, AMC23, AIME 2024, AIME 2025 [12]. These datasets all require complex chain-of-thought reasoning, and the generated thought processes often contain a rich variety of redundancy types, making them ideal scenarios for observing, classifying, and quantifying redundancy phenomena.

For generation, we set the maximum generation length (including both reasoning trace and final answer) for all models to 16384 tokens. For each test question, we sample 4 to 16 outputs with a temperature of 0.7.

D Ablation Study

Per-Category Intervention Ablation
Within the CATS framework, we further individually removed the intervention vector for specific redundancy categories (i.e., setting their w_k to 0), observing the impact on the average change in accuracy and length, to reveal the unique contribution of each category's intervention, as shown in Figure 5.

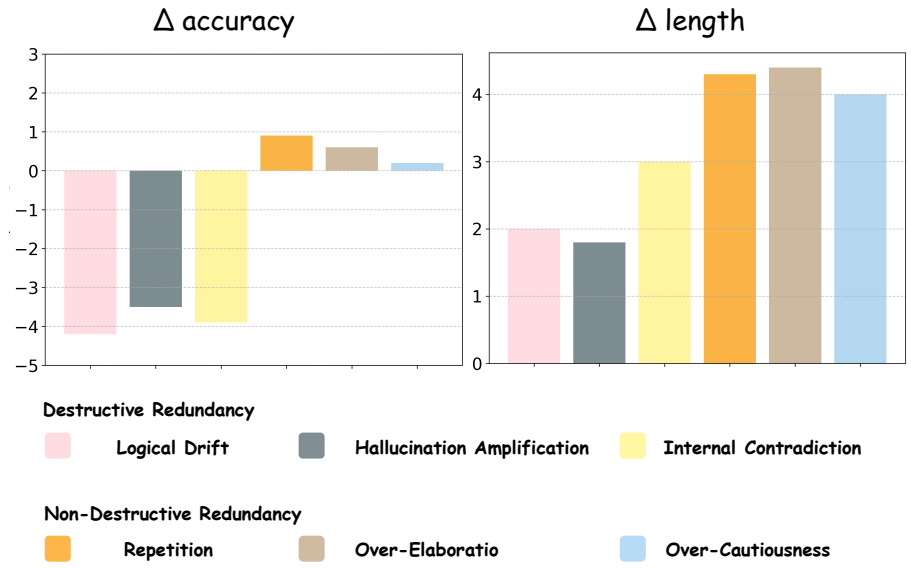

Figure 5: Impact of Removing Specific Redundancy Category Interventions on Core Metrics (Accuracy and Length)

Example Analysis:

- **Removing "Logical Drift" category intervention**: As seen in Figure 5, when intervention for the "Logical Drift" category is removed, the model's average accuracy decreases by 4.2%, and the average length only increases by 2.0%. This strongly proves the critical role of the "Logical Drift" category and its corresponding intervention vector in ensuring model output quality and suppressing destructive redundancy. It alerts us that not all redundancy is harmless, and identifying and intervening in specific destructive redundancies is crucial.
- **Removing "Repetition" category intervention**: Conversely, when we remove intervention for the "Repetition" category, the average length increases by 4.3%, but accuracy remains almost unchanged (change less than 1.0%). This indicates that intervention for non-destructive redundancies like "Repetition" primarily focuses on text conciseness and contributes significantly to length compression without affecting core performance.

These ablation results collectively demonstrate the rationality and effectiveness of our six-category redundancy classification system, as well as each category's unique contribution to achieving the overall optimization goals.

E Related work

E.1 Large Reasoning Models (LRMs) Background

Large Reasoning Models (LRMs), such as OpenAI's o1 [2] and DeepSeek-R1 [1], extend Large Language Models' reasoning ability by incorporating explicit Chain-of-Thought (CoT) mechanisms.

This simulates the human process of step-by-step problem decomposition, iterative verification, and idea refinement, significantly boosting model performance in mathematical, scientific, and commonsense reasoning tasks. Their typical inference process involves generating multi-step intermediate reasoning before outputting the final answer.

Despite the significant success of this strategy in improving accuracy, it also introduces substantial reasoning redundancy. To ensure answer reliability, LRMs tend to "over-explain for stability", leading to reasoning paths that can span hundreds to thousands of tokens for a single problem. Consequently, this results in issues such as increased inference latency and higher computational costs.

Existing redundancy mitigation strategies primarily fall into two categories: (1) training-based compression via Supervised Fine-Tuning (SFT) or Reinforcement Learning (RL), and (2) training-free compression through Prompt Engineering.

E.2 Supervised Fine-Tuning Strategies: Constructing "Verbose-Concise" Pairs for Compression

The core idea of these methods is to construct paired "verbose-concise" reasoning path data to train the model to generate more concise answers while maintaining reasoning validity. The specific process includes: collecting multiple reasoning paths for the same problem; selecting the shortest correct path as the gold standard; designing appropriate loss functions to guide the model to learn concise expressions during training.

For example, Self-training [3] constructs a dataset with multiple solutions for the same problem and selects the shortest correct path for training to enhance the model's compression capability. The TokenSkip [4] method identifies and skips tokens that contribute minimally to the final answer, thereby compressing reasoning length while preserving semantic integrity. C3oT [5] designs a GPT-4-based compressor that generates shorter chain-of-thought paths by retaining critical reasoning steps.

E.3 Reinforcement Learning Strategies: Designing Dual-Objective Rewards for Length-Accuracy Trade-off

Reinforcement learning approaches focus on designing reward functions that guide the model to compress reasoning path length while ensuring output accuracy. This typically involves setting dual-objective rewards: a conciseness reward (penalizing redundant tokens) and an accuracy reward (ensuring the final answer's correctness).

For instance, O1-Pruner [6] uses length and accuracy as baselines for its reward function, encouraging the model to generate shorter reasoning paths without sacrificing precision. ThinkPrune [7] introduces a length-aware reward, requiring the model to complete correct reasoning within a given token budget, only receiving positive feedback when both objectives are met.

E.4 Prompt Engineering Strategies: Zero-Training Compression of Reasoning Paths

These methods do not rely on additional training. Instead, they guide the model to control reasoning length through prompt design, achieving immediate compression.

For example, CCoT [8] explicitly prompts the model to "Be concise". Token-Budget [9] sets a token usage limit in the prompt, guiding the model to complete reasoning tasks within the budget. Chain of Draft [10] requires the model to retain only core information in each reasoning step, even limiting the word count per step to reduce redundant descriptions.

In summary, while existing solutions can compress LRMs' lengthy reasoning to varying degrees, they each suffer from distinct shortcomings: SFT and RL require retraining, incurring high computational and human costs due to annotation or reward design; prompt engineering relies on manual prompt granularity, with effects drifting across tasks and lacking precise control; other hidden intervention methods [13] use only a single directional vector, overlooking the semantic diversity of redundancy. In contrast, our proposed CATS requires no training, achieving 25% reasoning compression on different reasoning models by weighting and fusing six categories of differential vectors. It balances accuracy with interpretability, providing a plug-and-play, fine-grained, and cost-effective redundancy mitigation solution for resource-constrained scenarios.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction of the paper accurately present its core contribution, which is proposing the CATS method to reduce redundancy in large inference models. Experimental results demonstrate that this method can shorten the inference length while maintaining or slightly improving accuracy. The paper clearly distinguishes between the classifications of "destructive redundancy" and "non-destructive redundancy," and it is noted that CATS is a lightweight method that does not require training, all of which are consistent with the experimental findings of the paper.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: It has been stated in the conclusion section.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: This paper is primarily an empirical study, without presenting rigorous theoretical results or providing formal mathematical proofs.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper elaborates on the CATS framework, the extraction process of differential vectors, the experimental settings (models and datasets used), and the evaluation metrics in the methodology and experimental sections. Such information is sufficient for replicating the core results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The code has been anonymously provided in the abstract section.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper elaborates on the experimental setup in Section 4 "Experiments", including three models used , five mathematical reasoning datasets , as well as evaluation metrics and baseline methods. Such information is sufficient for understanding the experimental results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: It has been stated in the Experiments section.

   Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: It has been explained in the appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: This study primarily focuses on model efficiency and redundancy, and does not involve sensitive data or applications harmful to humans. Therefore, it fully complies with the NeurIPS ethical guidelines.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

Justification: The introduction of the paper implies its positive social impact, i.e., improving the efficiency of LRMs by reducing computational redundancy, thereby lowering inference costs and energy consumption, which is particularly important for resource-constrained environments. The paper does not discuss potential negative impacts, and this part can be supplemented in the final version.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: The paper does not release any new datasets or models with a high risk of misuse.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The paper cites the original papers of the models used and datasets, and correctly attributes them to their creators.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: This paper provides relevant code, and all data are publicly available.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: This study does not involve crowdsourcing or human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

Justification: This study does not involve human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: The core research object of this paper is large reasoning models (LRMs), and an intervention method is proposed to improve their performance. Therefore, the use of LRMs is described in detail in both the methodology and experiments.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.