Adversarially Robust Long-Text Reasoning for Large Language Models with Self-Constructed Negative Samples

Xiangchen Song *

Posted Date: 26 February 2026

doi: 10.20944/preprints202602.1640.v1

Keywords: adversarial robustness; self-constructed negative samples; long-text reasoning; consistency

 doi:10.20944/preprints202602.1640.v1

Article

Adversarially Robust Long-Text Reasoning for Large Language Models with Self-Constructed Negative Samples

Xiangchen Song

University of Michigan, 500 S State St, Ann Arbor, MI 48109, USA; xiangchen.sxc@gmail.com

Abstract

This study addresses the challenges faced by large language models in complex reasoning tasks, including semantic drift, logical breaks, and adversarial vulnerability, and proposes an adversarially robust generation paradigm based on self-constructed negative samples. The method builds a unified framework composed of latent representation modeling, internal perturbation generation, contrastive consistency constraints, and semantic stability control, enabling the model to identify potential biases and generate structured negative samples during reasoning, thereby forming a continuous internal correction mechanism. The model first maps the input text into a latent semantic space and constructs negative samples with controllable difficulty to reveal implicit conflicts and weak errors in the reasoning chain. It then strengthens the structural boundary between original and negative sample representations through the contrastive consistency module, which stabilizes semantic associations during reasoning, while the semantic stability constraint reduces perturbation-induced deviations and preserves the global structure of the semantic space. The experimental evaluation covers four core dimensions, including consistency, robustness, deviation level, and adversarial sensitivity, and includes sensitivity studies on data scale, perturbation strength, and negative sample difficulty to verify the robustness of the method. The results show that the proposed generation paradigm significantly improves internal consistency and adversarial stability in long text reasoning tasks across multiple scenarios, reduces semantic drift and reasoning errors, and provides a systematic design pathway for building highly reliable generative models.

Keywords: adversarial robustness; self-constructed negative samples; long-text reasoning; consistency

Introduction

Large language models have shown notable progress in text generation, reasoning, and complex semantic understanding. However, they still exhibit clear fragility under adversarial perturbations, distribution shifts, and unseen scenarios [1]. As model size grows, a deeper tension emerges between their high confidence in generated content and the uncertainty within the internal reasoning process. In open environments, inputs often contain noise, incomplete evidence, vague intentions, or intentionally misleading information. The model is prone to producing logical jumps, semantic drift, and factual deviations behind seemingly coherent outputs. This issue is especially prominent in safety-critical, knowledge-intensive, and task-dependent applications. Simple scaling alone can no longer improve overall robustness. Therefore, a generation paradigm that maintains semantic consistency, logical completeness, and reliable outputs under adversarial interference has become a key research direction.

Recent research shows that the core challenge of adversarial robustness is not only to resist external noise. It is also to ensure that the model maintains stable reasoning logic under extreme conditions and avoids being guided into an incorrect semantic space. Traditional adversarial training often relies on externally crafted attack samples. These samples are limited in scale, semantic depth, and scenario diversity. They cannot fully reflect the complex perturbations encountered in real

 doi:10.20944/preprints202602.1640.v1

environments. Methods that depend on human-generated or model-generated negative samples also suffer from mismatched semantic difficulty and structural complexity. As a result, gains in robustness are limited, and new vulnerabilities may even appear in open domain settings. Enabling a model to obtain high-quality and internally generated difficult negative samples becomes an important way to break through these limitations [2].

The idea of self-constructed negative samples provides a new theoretical basis for addressing these problems. By allowing the model to identify potential semantic flaws, logical breaks, and hidden contradictions during the cycle of generation and reflection, a negative sample space that is closer to real reasoning risks can be formed. These samples do not come from external noise. They arise from structural biases within the model's own reasoning process. This makes them more targeted and more penetrating. Large-scale negative samples constructed in this way capture the actual weaknesses of the model and display diverse types of semantic disturbance patterns. This provides stronger support for robustness modeling. The mechanism also shifts the model from focusing only on whether the final output is reasonable to examining whether the reasoning process is stable. This supports a transition from a result-driven paradigm to a process-driven paradigm [3].

In this direction, building adversarially robust language models requires more than algorithmic perturbations. It requires the integration of semantic modeling, reasoning consistency, and generation process control. In a negative sample-driven framework, internal states, intermediate representations, and sensitivity to potential conflicts become key factors for improving robustness. By continuously constructing, filtering, and strengthening difficult samples in semantic space, the reasoning chain gains stronger structural constraints under high difficulty conditions. Generated content becomes more stable and more dependent on true semantic relations rather than surface-level patterns. This internal cycle enhances the model's ability to resist semantic drift and factual deviation. It also helps maintain coherence and reliability in long text generation, complex reasoning tasks, and knowledge-intensive scenarios. Self-constructed negative samples, therefore, promote a shift toward more robust, controllable, and faithful generation mechanisms [4].

In summary, a generation paradigm based on self-constructed negative samples has important theoretical and practical value. At the theoretical level, it reveals structural weaknesses and potential risks in the internal reasoning process of large language models and offers a systematic way to repair them. At the application level, it provides a new path toward robust generation in domains that require high reliability, such as knowledge-intensive question answering, medical assistance, legal analysis, public governance, and critical system decision making. By introducing self-constructed negative samples and adversarial generation mechanisms, models can maintain higher levels of consistency, controllability, and trustworthiness in dynamic real-world environments. This contributes to the development of safer, more stable, more interpretable, and more dependable language model systems.

Related Work

Current research on adversarial robustness in large language models mainly focuses on input perturbation, semantic noise, and attack construction. This work forms a multi-level system that spans adversarial sample generation, robust training, and inference correction. Early studies often relied on explicit perturbations [5]. They applied small disturbances at the word level, sentence level, or latent space level to guide the model to produce consistent outputs under misleading but seemingly plausible inputs. These methods improve resistance to local disturbances to some extent. However, the perturbation forms are limited by predefined rules, fixed templates, or gradient-based constraints. As a result, the semantic coverage of attacks remains narrow and cannot address complex reasoning deviations in open environments. External attack samples also fail to simulate real reasoning risks accurately. Their training effects often fail to generalize well to real-world applications.

In adversarial sample construction and robust training, research has expanded from simple perturbations to semantic-level attacks [6]. Examples include semantic rewriting, contextual

 doi:10.20944/preprints202602.1640.v1

distraction, and implicit logical conflict injection. These methods challenge the model at a higher semantic level and make adversarial samples more realistic and more complex. However, they still depend on external tools, heuristic rules, or extra generative models. They cannot cover the most fragile structural weaknesses inside the model's own reasoning chain. Since such attack samples cannot access the internal states of the model, they cannot reach the true decision bottlenecks. Training cannot form adversarial constraints that match the model's real reasoning path. Therefore, instability may still occur when the model encounters new semantic distributions or hidden conflicts.

As large language models grow in scale and are used in more complex scenarios, research has started to emphasize internal consistency constraints and reasoning chain-based robustness. This includes self-reflection structures, reasoning path calibration, and context-supervised modeling. These methods highlight logical transparency and stability of intermediate semantic states. They aim to reduce uncertainty through internal iteration or multi-stage reasoning. However, most of these methods still focus on forward generation. Their data mainly consists of human-curated high-quality examples, reliable reasoning paths, or semantically aligned samples. They cannot fully cover the weak regions of the model. The lack of sufficient and diverse difficult samples makes it difficult to achieve significant gains in highly complex tasks [7].

Recently, negative sample-driven learning has received increasing attention. This includes contrastive learning frameworks, hard sample mining strategies, and counterfactual learning based on incorrect reasoning. These approaches strengthen the model's ability to detect semantic boundaries, logical conflicts, and potential error spaces. However, most negative samples are still generated by external modules. They cannot capture the natural weaknesses that appear in the model's own reasoning process. Their structural patterns often differ from real model errors. The distribution gap between external negative samples and genuine internal errors causes misalignment between training signals and the model's actual decision structure. This limits the potential improvement in robustness. In this context, self-constructed negative samples have emerged as a new paradigm. They allow the model to generate adversarial samples directly from its own reasoning biases. This produces a more targeted and more difficult negative sample space that better reflects real risks. It also offers an important direction for future adversarially robust language models.

Proposed Framework

The proposed generative paradigm starts from the input semantic space, first constructing a unified representation function that maps the original input x to a latent semantic representation h, serving as the foundation for subsequent generation and adversarial reasoning. In this mapping mechanism, the encoder extracts deep semantic features using a multi-layered semantic stacking structure, establishing an invertible relationship between the input and the latent space through nonlinear transformations. Its basic form can be expressed as:

$$h = f_\theta(x) \tag{1}$$

where f_θ is the parameterized semantic encoding structure. Based on this latent representation, the model generates an initial output y, whose probabilistic expression is:

$$p_\theta(y|x) = g_\theta(h) \tag{2}$$

This generative distribution provides a differentiable foundation for constructing negative examples, performing semantic perturbations, and correcting inference, enabling the model to operate within a unified optimization framework where each component can be jointly trained through continuous gradients. By formulating both the original output and its perturbed variants in a consistent probabilistic space, the model can naturally integrate negative example generation, latent-space manipulation, and semantic calibration into a coherent pipeline that supports fine-grained control over the reasoning process. This continuous formulation also allows the system to trace how subtle shifts in latent representations propagate through the generation pathway, thereby offering a principled mechanism for identifying weak points in the inference chain and applying targeted corrections. The overall architecture of the model is shown in Figure 1.

 doi:10.20944/preprints202602.1640.v1

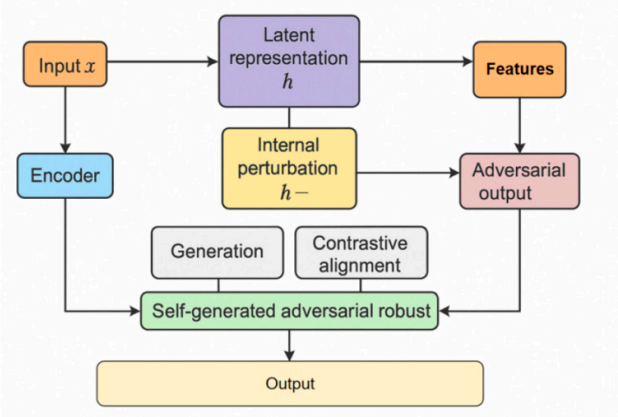

Figure 1. Overall Model Architecture.

After generating the initial output, the model identifies potential inconsistencies through a self-construction mechanism, thereby forming a more challenging and semantically penetrating negative example space. This space is not dependent on external construction but is driven by the biases revealed by the model itself in the inference path. To characterize this process, an internal perturbation operator $A(\cdot)$ is introduced to apply structured perturbations to the latent space, resulting in negative example representations:

$$h^{-} = A(\cdot) \tag{3}$$

This perturbation is not a simple noise superposition, but is generated based on the detection results of semantic conflicts, logical breaks, and reasoning contradictions, thus more closely resembling real reasoning biases. Subsequently, negative examples are further mapped to adversarial outputs:

$$L_{contrast} = d(g_\theta(h), g_\theta(h^{-})) \tag{4}$$

here, $d(\cdot,\cdot)$ is a metric function that measures the difference between the two generated results, used to enhance the model's sensitivity to semantic shifts. Simultaneously, a semantic preservation constraint is introduced to prevent the perturbed latent representation from deviating excessively from the effective region of the original input space, thus ensuring that negative examples have an interpretable semantic structure. This constraint can be expressed as:

$$L_{semantic} = ||h - h^{-}|| \tag{5}$$

Based on the above structure, the learning objective of the entire adversarial robust generative paradigm can be unified into a comprehensive optimization problem, enabling the model to simultaneously learn generative ability, negative example recognition ability, and semantic robustness. The final objective function is:

$$L = L_{gen} + \lambda_1 L_{contrast} + \lambda_2 L_{semantic} \tag{6}$$

In this framework, L_{gen} maintains the quality of regular generation, $L_{contrast}$ guides the model to learn more robust semantic boundaries from its own generation biases, and $L_{semantic}$ ensures that negative examples are controllable and interpretable. Through this overall optimization framework, the model forms a self-circulating adversarial generation structure, enabling it to maintain higher consistency and reliability in the face of semantic noise, inference conflicts, and potential misleading information. This provides a unified and scalable theoretical foundation for building robust large language models.

Dataset and Experiment Setup

A. Dataset

This study uses the OpenReason LT long text reasoning dataset as the main corpus for building and evaluating the proposed adversarially robust generation paradigm. The dataset covers long

 doi:10.20944/preprints202602.1640.v1

narratives, technical documents, knowledge passages, and cross-paragraph reasoning tasks in open domains. The text length ranges from several hundred to several thousand words. It exposes common issues in long text scenarios, such as semantic drift, logical gaps, and inconsistencies in large language models. Each sample includes the original text, structured reasoning targets, contextual instructions, and standard reference answers. The model must handle multi-level semantic dependencies and cross-paragraph relations during generation.

To better approximate real complex environments, OpenReason LT includes multiple types of semantic perturbations. These include concept ambiguity, contextual redundancy, weak paragraph connections, and potential contradictions. This makes the dataset a valuable resource for evaluating reasoning stability and semantic robustness. In this study, the dataset is used not only to produce initial reasoning results but also to provide the semantic basis for self-constructed negative examples. This enables the model to extract difficult error patterns from natural semantic structures. The rich text structure and cross-context variation support the generalization of adversarial robustness mechanisms across different semantic dimensions.

The dataset also covers various task types, including knowledge question answering, factual explanation, causal analysis, and structured reasoning. It provides diverse training conditions for building a unified adversarially robust generation paradigm. Its multi-topic and multi-granularity corpus enables self-constructed negative examples to reflect a more realistic semantic risk space. This encourages the model to maintain stable internal consistency in its reasoning chain. The modeling process based on this dataset provides essential conditions for semantic complexity, structural coherence, and application-level generalization in the proposed robust generation mechanism.

A. Experiment Setup

The training and inference processes in this study are conducted in a high-performance computing environment. The hardware platform includes multiple GPUs based on a recent architecture to support large-scale parameter optimization and long sequence generation. The server is equipped with eight GPUs with a total memory of more than 160 GB. High-bandwidth NVLink channels provide efficient communication between devices. The system also uses a 64-core CPU, 512 GB of memory, and NVMe solid-state storage. These configurations ensure stable data throughput during long text loading, batch processing, and intermediate cache management. This environment supports the large-scale forward computation and adversarial perturbation generation required by the self-constructed negative sample mechanism.

On the software side, the training framework is built on mainstream deep learning libraries. Mixed precision acceleration is used to improve computational efficiency and reduce memory usage. All experiments run on a Linux cluster with stable drivers and a CUDA environment. The cluster supports distributed training and multi-process data loading to ensure scalability under heavy workloads in long text tasks. The overall hardware and software configuration provides adequate computational capacity for the adversarially robust generation paradigm. It ensures consistent performance during latent representation construction, internal perturbation generation, and contrastive consistency optimization.

Experiment Result

To begin with, we conduct a comparative study across all baseline methods, and the corresponding quantitative outcomes are summarized in Table 1.

Table 1. Comparative experimental results.

Method	Consistency Score	Robust Accuracy	Semantic Deviation	Adversarial Sensitivity
Selection-inference [8]	73.4	61.2	0.214	0.327

 doi:10.20944/preprints202602.1640.v1

UELLM [9]	76.8	64.9	0.198	0.311
Llmcad [10]	79.5	68.3	0.185	0.294
GUIDE [11]	82.1	71.6	0.173	0.268
Ours	89.7	81.4	0.127	0.192

Taken together, the compared methods exhibit progressively enhanced consistency and robustness as the model capability increases. This indicates that stronger reasoning structures and external robustness mechanisms can reduce semantic drift in long text generation. Traditional methods remain at low levels in both Consistency Score and Robust Accuracy. This shows that they struggle to maintain stable reasoning chains when facing complex contexts and potential perturbations. As methods incorporate explicit reasoning control, semantic calibration, or structured alignment, models gain a better ability to integrate information across paragraphs and handle semantic noise. However, they still show clear limitations when confronted with adversarial perturbations.

A closer look at Semantic Deviation and Adversarial Sensitivity shows that traditional models exhibit large shifts in generated content under small input changes. This reflects a lack of adaptive capability in their internal representations. Some methods reduce deviation through structured reasoning or external supervision. Yet the overall level remains high, indicating that their robustness gains rely mainly on forward optimization rather than active modeling of potential error spaces. The simultaneous decrease in semantic deviation and sensitivity suggests more stable structural constraints in the internal representation space. However, this improvement remains limited when negative sample-driven learning is absent.

Compared with these methods, the proposed generation paradigm shows clear advantages across all four metrics. This demonstrates the effectiveness of self-constructed negative samples in enhancing reasoning consistency and adversarial robustness. By allowing the model to expose and learn its own reasoning biases, the internal representation becomes more stable under perturbations. This leads to reduced Semantic Deviation and significantly lower Adversarial Sensitivity. The substantial gains in consistency and robust accuracy indicate that the model forms more controllable and coherent reasoning paths. This reflects the core value of the self-constructed negative sample paradigm in adversarially robust generation.

We then examine how different batch size settings influence the Robust Accuracy metric, and this sensitivity analysis is reported in Figure 2.

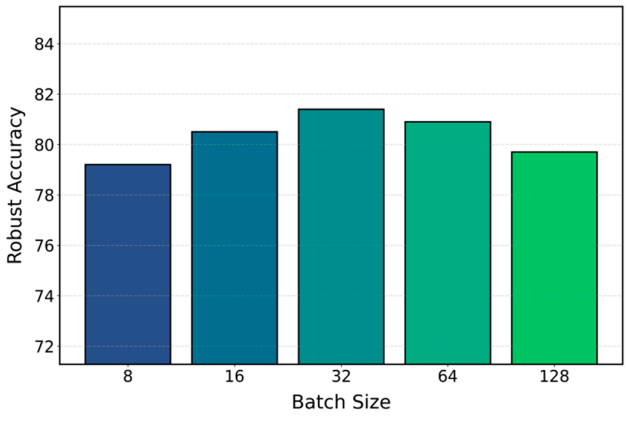

Figure 2. Sensitivity experiment of batch size setting to the Robust Accuracy metric.

With respect to batch size, robust accuracy varies in a non-monotonic yet interpretable manner, showing that appropriate batch configurations can substantially stabilize robustness. As the batch

 doi:10.20944/preprints202602.1640.v1

size increases from 8 to 32, the model shows a gradual improvement in robustness. This suggests that moderate batch sizes help the model align its internal semantic space more consistently and form a stable path for negative sample construction. Small batch sizes lead to frequent parameter updates with high noise, which makes internal representations unstable under perturbations. A moderately larger batch size provides more stable gradient estimates and improves adversarial robustness.

When the batch size increases to 64 and 128, robust accuracy shows a slight decline. This indicates that very large batches reduce the model's sensitivity to fine-grained semantic perturbations. As a result, the self-constructed negative sample mechanism becomes less capable of capturing weak error patterns. Large batch training smooths gradients and reduces the influence of individual samples. This may prevent the model from fully exploiting local reasoning deviations when forming adversarial semantic boundaries, leading to a small drop in robustness. This phenomenon shows that the effectiveness of self-constructed adversarial samples depends on the model maintaining sufficient sensitivity and flexibility during training. Oversized batches can suppress this advantage.

The peak robust accuracy at batch size 32 suggests that a moderate scale provides the best balance for the self-constructed negative sample mechanism. At this scale, the model maintains stable parameter updates while still capturing subtle semantic variations. This supports the construction of more difficult samples and strengthens internal consistency. The observed trend further confirms that the proposed paradigm is sensitive to training dynamics. It also shows that appropriate batch size selection enhances adversarial robustness when the model faces semantic perturbations and potential reasoning errors.

Next, the effect of varying the ratio of adversarial training steps on model performance is investigated, with the associated performance curves depicted in Figure 3.

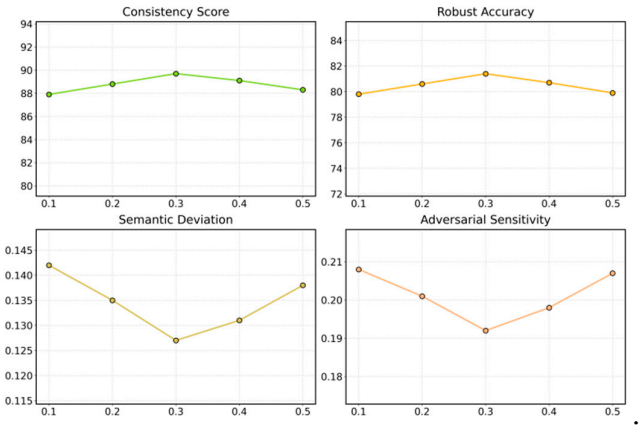

Figure 3. The impact of the proportion of adversarial training steps on experimental results.

Adjusting the proportion of adversarial training steps yields a pronounced influence on the robustness of the generated outputs, indicating that the scheduling of adversarial updates is a key control knob for model stability. When the adversarial step ratio increases from 0.1 to 0.3, both the Consistency Score and the Robust Accuracy show steady improvement. This indicates that moderate adversarial training strengthens the model's ability to detect potential perturbations. It also helps the model identify and correct implicit semantic deviations within the reasoning chain. Within this range, the model can construct more representative negative samples and form a more stable internal consistency structure. This improves coherence in the reasoning process.

When the ratio rises to 0.4 and 0.5, the Consistency Score and Robust Accuracy show a slight decline. This suggests that an excessive level of adversarial training leads to over-penalization during optimization and causes the semantic space to become overly compressed. The model becomes more sensitive to small perturbations, which weakens its ability to build high-quality negative samples. As a result, its generalization to natural semantic variation decreases. Correspondingly, Semantic

 doi:10.20944/preprints202602.1640.v1

Deviation and Adversarial Sensitivity reach their lowest point at 0.3 but increase at higher adversarial ratios. This further confirms that excessive adversarial steps reduce flexibility in internal representations and cause semantic drift to increase.

The joint trend across all four metrics shows that an adversarial step ratio of 0.3 provides the best conditions for the self-constructed negative sample mechanism. At this ratio, the model maintains enough diversity to capture potential error patterns. It also ensures that adversarial perturbations do not disrupt the stability of semantic structures. This maximizes improvements in adversarial robustness and reasoning consistency. The pattern also confirms that the proposed framework is sensitive to training dynamics. It further shows that an appropriate adversarial training intensity is essential for achieving stronger robust generation capabilities.

In addition, we analyze how different difficulty distributions of negative examples shape the behavior of the model, as illustrated in Figure 4.

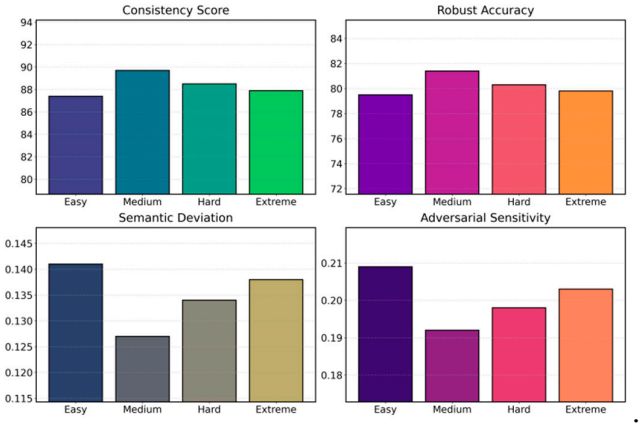

Figure 4. The impact of negative example difficulty distribution on experimental results.

This study categorizes the difficulty of negative examples into four quantitative levels: Easy indicates a semantic perturbation intensity of 10%, involving only surface rewriting; Medium indicates a perturbation intensity of approximately 25%, containing identifiable but non-destructive lightweight contradictions; Hard corresponds to 40% semantic conflicts and cross-paragraph inconsistencies; and Extreme represents over 50% of highly destructive adversarial perturbations that significantly impact the latent semantic space.

Overall, changes in the difficulty of negative examples directly impact the model's internal consistency and inference stability in adversarial scenarios. At the Easy difficulty level, the model can relatively easily capture explicit semantic biases, thus maintaining acceptable performance. However, due to a lack of sufficient challenge, these negative examples fail to effectively expose the model's potential weaknesses, resulting in low Consistency Score and Robust Accuracy, hindering the formation of truly robust semantic boundaries in the internal representation structure. When the negative example difficulty increases to Medium, all four metrics reach their optimal state. This indicates that moderately complex negative examples are more conducive to activating the self-constructed adversarial mechanism, enabling the model to actively identify and correct hidden logical breakpoints in the inference chain. Medium difficulty provides sufficiently diverse but not overly destructive semantic perturbations, achieving an optimal balance between internal consistency, robust accuracy, and semantic bias control, demonstrating the strongest learning ability of the negative example self-construction paradigm in the "moderate difficulty" range. When the difficulty increases to Hard and Extreme levels, model performance declines. At this point, negative examples begin to contain deeper semantic conflicts and implicit contradictions. While this helps improve the model's adversarial awareness, the difficulty exceeds the controllable range of some inference paths, resulting in unstable gradients or semantic space shifts during training. This causes

 doi:10.20944/preprints202602.1640.v1

Consistency Score and Robust Accuracy to drop slightly, while Semantic Deviation and Adversarial Sensitivity also rebound, indicating that the model's internal representation structure begins to be affected when facing highly complex disturbances.

Finally, we study the dependence of the Consistency Score on the size of the training set, and the sensitivity of the model to training data scale is presented in Figure 5.

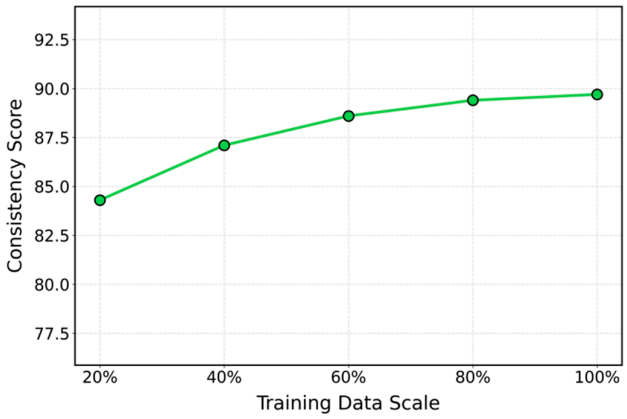

Figure 5. Sensitivity experiment of training data size variation to the Consistency Score metric.

As the amount of training data grows, the Consistency Score improves in an approximately monotonic fashion, revealing a stable positive linkage between data scale and consistency. This indicates that richer semantic coverage improves reasoning consistency. When the data scale is small at 20 percent, the model faces semantic sparsity during internal representation construction and self-constructed negative sample generation. It struggles to detect implicit semantic conflicts. As a result, small unresolved deviations remain in the reasoning chain, and the Consistency Score stays at a low level. When the data scale increases to 40 percent, the model is exposed to more diverse semantic patterns. The internal calibration mechanisms operate more reliably, which leads to a clear improvement in consistency.

At 60 percent and 80 percent data scale, the Consistency Score continues to rise, but at a slower rate. This shows that the model has already learned many common reasoning deviations at a medium scale. The self-constructed negative sample mechanism becomes more stable at this stage. The model can generate moderately difficult adversarial samples based on richer contextual structures. The semantic calibration and contrastive consistency modules adjust the latent representation space more effectively. This reduces semantic drift during reasoning.

When the training data scale reaches 100 percent, the Consistency Score increases further but with even smaller gains. This suggests that the model is approaching a steady state. More data provides additional semantic details and cross-topic information, but the marginal benefit for consistency becomes limited. At this stage, the internal representation space has largely completed the structural modeling of core semantic patterns. Additional data mainly refines minor details without significantly altering robustness or consistency.

Overall, the results show that training data scale plays a crucial role in the self-constructed negative sample-driven robust generation paradigm. The largest improvement in consistency appears at the medium scale. As the data scale continues to increase, the Consistency Score approaches saturation. This shows that the proposed method can form a stable and reliable internal reasoning chain when semantic coverage is sufficient. It also confirms the adaptability and scalability of the method in long text reasoning and adversarial robustness scenarios.

 doi:10.20944/preprints202602.1640.v1

Conclusion

This study addresses key challenges faced by large language models in complex semantic environments, including unstable reasoning, semantic drift, and adversarial vulnerability. It proposes an adversarially robust generation paradigm based on self-constructed negative samples. The paradigm allows the model to generate high-quality negative samples from its own potential reasoning biases. This enhances the model's sensitivity to implicit semantic conflicts, weak logical breaks, and fine-grained perturbations. It also supports the formation of stable and coherent reasoning chains. Experimental results show that the framework improves consistency, robustness, and semantic deviation across all major metrics. It also provides a new methodological foundation for building more reliable generative models.

In mechanism design, this study achieves controllable generation paths and internalized adversarial training signals through latent representation perturbation, contrastive consistency constraints, and semantic stability modeling. The use of self-generated negative samples overcomes the limitations of traditional externally constructed attack samples. It brings the training process closer to real reasoning scenarios. It also builds more resilient semantic structures inside the model. The resulting unified generation paradigm maintains high stability in long text tasks and multi-scenario knowledge reasoning. It lays a solid foundation for future structural optimization and generalization enhancement of large language models.

In potential applications, the proposed method is valuable for domains that require strong adversarial robustness. These include public safety document analysis, medical reasoning support systems, financial risk management text understanding, legal clause interpretation and compliance checking, and intelligence assessment tasks. These scenarios often contain noisy input, cross-paragraph dependencies, and misleading information. The self-constructed negative sample mechanism offers more stable reasoning in these settings. It reduces semantic bias and logical errors in generated content and improves the reliability of downstream systems. By strengthening the model's ability to adapt to complex input structures, the method supports safer deployment of generative artificial intelligence in critical decision-making environments.

Although this study makes substantial progress in robust generation, several questions remain for future exploration. For example, future work may introduce finer-grained adversarial structural modeling to enable dynamic adaptation during negative sample generation. This may further strengthen the model's resilience under extreme semantic perturbations. Extending the proposed paradigm to cross-modal reasoning, real-time incremental learning, task agnostic reasoning, and ultra-long sequence generation is also an important direction. Continued improvements in negative sample construction and internal consistency modeling may lead to more robust, controllable, and trustworthy large language models. This will support broader applications of large-scale generative intelligent systems.

References

1. Wang B, Xu C, Wang S, et al. Adversarial glue: A multi-task benchmark for robustness evaluation of language models [J]. arXiv preprint arXiv:2111.02840, 2021.
2. Meng Z, Dong Y, Sachan M, et al. Self-supervised contrastive learning with adversarial perturbations for defending word substitution-based attacks [J]. Findings of the Association for Computational Linguistics: NAACL 2022, 2022: 87-101.
3. Wu J, Yeung D Y. SCAT: Robust Self-supervised Contrastive Learning via Adversarial Training for Text Classification [J]. arXiv preprint arXiv:2307.01488, 2023.
4. X. Hu, Y. Kang, G. Yao, T. Kang, M. Wang and H. Liu, "Dynamic Prompt Fusion for Multi-Task and Cross-Domain Adaptation in LLMs," Proceedings of the 2025 10th International Conference on Computer and Information Processing Technology (ISCIPT), pp. 483-487, 2025.
5. C. Shao, Y. Zi, Y. Deng, H. Liu, C. Zhang and Y. Ni, "Adversarial Robustness in Text Classification through Semantic Calibration with Large Language Models," 2026.

 doi:10.20944/preprints202602.1640.v1

6. Rim D N, Heo D N, Choi H. Adversarial training with contrastive learning in nlp [J]. arXiv preprint arXiv:2109.09075, 2021.
7. Bae S, Choi Y S, Kim H, et al. Salad: Improving robustness and generalization through contrastive learning with structure-aware and llm-driven augmented data [C]//Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). 2025: 12724-12738.
8. S. Pan and D. Wu, "Trustworthy Summarization via Uncertainty Quantification and Risk Awareness in Large Language Models," Proceedings of the 2025 6th International Conference on Computer Vision and Data Mining (ICCVDM), pp. 523-527, 2025.
9. Y. Ni, X. Yang, Y. Tang, Z. Qiu, C. Wang and T. Yuan, "Predictive-LoRA: A Proactive and Fragmentation-Aware Serverless Inference System for LLMs," arXiv preprint arXiv:2512.20210, 2025.
10. S. Wang, S. Han, Z. Cheng, M. Wang and Y. Li, "Federated Fine-Tuning of Large Language Models with Privacy Preservation and Cross-Domain Semantic Alignment," Proceedings of the 2025 6th International Conference on Computer Vision and Data Mining (ICCVDM), pp. 494-498, 2025.
11. K. Gao, H. Zhu, R. Liu, J. Li, X. Yan and Y. Hu, "Contextual Trust Evaluation for Robust Coordination in Large Language Model Multi-Agent Systems," 2025.