Transactions on Computational and Scientific Methods | Vo. 4, No. 7, 2024
ISSN: 2998-8780
https://pspress.org/index.php/tcsm
Pinnacle Science Press

Gradient-Guided Adversarial Sample Construction for Robustness Evaluation in Language Model Inference

Yichen Wang
Georgia Institute of Technology, Atlanta, USA
yichenk.wang@gmail.com

Abstract: This study addresses the challenge of adversarial robustness faced by large language models in natural language inference tasks. It proposes a gradient-guided adversarial sample generation method. The method introduces an inference sensitivity scoring mechanism, which uses internal gradient information to precisely identify input regions most sensitive to reasoning outcomes. This enables the selection of efficient perturbation positions. At the same time, a semantics-preserving perturbation strategy is designed. It aims to achieve the attack objective while preserving the semantic consistency and contextual coherence of the original text. The method extracts embedding representations from the input text and constructs a perturbation priority ranking by combining gradient magnitude with semantic attention weights. High-quality adversarial samples are generated through dual constraints of semantic similarity and contextual consistency. Under various input conditions, including perturbation position strategies, text length, and multilingual scenarios, the method demonstrates strong attack efficiency, semantic preservation, and generalization stability. Experimental results show that the proposed approach significantly improves attack success rates while maintaining a low perturbation rate. The generated texts remain highly natural and readable. These findings validate the effectiveness and applicability of the proposed mechanisms in text-level adversarial sample construction.

Keywords: Adversarial sample generation, language model inference, gradient guidance, semantic preservation

1. Introduction

With the rapid advancement of artificial intelligence, large language models (LLMs) have demonstrated unprecedented capabilities in the field of natural language processing. They are widely applied in tasks such as text generation, sentiment analysis, question answering, and automatic summarization[1,2]. By learning from large-scale corpora, these models possess strong abilities in language understanding and generation, making them a core component of modern intelligent systems. However, despite their impressive performance across various tasks, concerns remain regarding their robustness and safety. In particular, their vulnerability to adversarial examples has become increasingly evident. Adversarial examples are malicious inputs crafted by introducing subtle perturbations to the original text, leading the model to produce incorrect outputs. These examples expose potential flaws in the reasoning process and pose challenges to the model's reliability in critical applications[3].

The reasoning mechanism of large language models is fundamentally based on probabilistic language modeling and contextual understanding. This makes them highly sensitive to small perturbations in the input. In computer vision, research on adversarial examples is relatively mature[4]. However, in natural language processing, due to the discrete nature of text and the complexity of language structures, generating adversarial examples is more difficult and remains an open research area. As LLMs continue to scale in parameter size and improve in semantic modeling capacity, systematically evaluating their adversarial robustness and designing effective attack methods have become essential prerequisites for improving model security[5]. Adversarial examples not only reveal hidden model weaknesses but also serve as tools for developing more robust training mechanisms. This promotes the evolution of language models toward greater reliability and interpretability[6].

Gradient information serves as a crucial link between model input and output and plays a central role in adversarial example generation. Gradient-based attack methods analyze the model's response to input directions and generate targeted perturbations to mislead the model's reasoning. In natural language processing, the integration of gradient-guided strategies into text adversarial example generation has gained significant attention[7]. These methods improve attack efficiency and align with the complex semantic representation mechanisms within LLMs. By using gradient information to guide perturbations, researchers can more accurately identify sensitive regions in the reasoning process. This provides new insights into the semantic decision-making of language models. Thus, gradient-guided mechanisms form an important theoretical foundation for constructing efficient and representative adversarial examples for language models[8].

Most large language models currently lack mechanisms for adversarial robustness evaluation during real-world deployment. This limitation restricts their use in safety-critical domains. In fields such as law, healthcare, and finance, biased reasoning or inappropriate outputs from models may lead to severe consequences. Enhancing model robustness is not only a technical requirement for ensuring output stability but also a key issue in addressing the social responsibility and ethical constraints of AI systems. Research on adversarial examples, especially those based on gradient-guided attacks, helps uncover vulnerabilities in model reasoning. It also supports the development of more robust and trustworthy language models. In addition, such studies can improve generalization under diverse input conditions and expand the practical adaptability of models in complex scenarios[9].

In conclusion, the study of adversarial example generation for large language models based on gradient guidance holds significant theoretical and practical value. On the one hand, it advances understanding of the reasoning mechanisms of LLMs and lays a foundation for research on interpretability, safety, and controllability. On the other hand, the outcomes of this research are expected to be widely applied in model testing, robustness evaluation, and defense mechanism design. This will facilitate the safe deployment of natural language processing systems in real-world settings. As LLMs increasingly become the core of intelligent systems, systematically exploring their performance in adversarial contexts will be essential for building trustworthy artificial intelligence.

2. Related work

2.1 Large Language Model Inference

With the development of large language models, their reasoning ability has become an important indicator of intelligence. Reasoning in large language models typically refers to the prediction and generation of the next word or entire text under given contextual conditions. It is essentially a process of probabilistic modeling over language sequences using deep neural networks. Through training on large-scale corpora, these models gradually learn lexical, syntactic, and semantic patterns in language. As a result, they acquire the ability to handle complex reasoning tasks. In particular, large language models show near-human performance in open-ended question answering, logical inference, and text summarization. This marks a new stage in the progress of natural language processing[10].

The reasoning process of large language models usually relies on their capacity to model contextual information. This makes them highly dependent on the structure and semantics of the input. The Transformer architecture, as the primary framework, uses self-attention mechanisms to capture long-range dependencies within the input sequence. This allows the model to integrate global information during reasoning. Such global modeling ability is crucial for reasoning tasks, which often require combining information from multiple text fragments to produce coherent and logical outputs. However, despite their advanced structural design, large language models can still produce uncertain or incorrect reasoning results when faced with ambiguous, vague, or out-of-distribution inputs[11].

In real-world applications, the reasoning output of large language models is influenced not only by the quality of training data but also by slight variations in input. This sensitivity introduces a potential risk. The model may generate entirely different outputs due to minor textual perturbations. On the one hand, this reflects the model's sensitivity to linguistic detail. On the other hand, it reveals instability in the reasoning mechanism. In safety-critical scenarios, such inconsistency and unreliability in reasoning can lead to untrustworthy outputs, limiting the feasibility of deployment. Therefore, studying the stability and robustness of the reasoning process is essential for improving language model performance[12].

Moreover, although large language models acquire broad general knowledge during pretraining, their reasoning ability can still fluctuate significantly across specific tasks and contexts. This issue has drawn attention to the model's response mechanisms to semantic variations in input. Researchers aim to explore how small perturbations while preserving original meaning, can lead to different reasoning outcomes. This is not only important for understanding internal representations and reasoning paths but also provides direction for building more robust natural language systems. Overall, while large language models possess strong potential in reasoning, their sensitivity to input perturbations and the uncertainty of output still pose challenges to their safety and trustworthiness.

2.2 Research on adversarial sample generation methods

As a key approach to improving model security and robustness, adversarial example generation has attracted growing attention in the field of natural language processing. Unlike continuous perturbations in the image domain, the discrete nature of the text and its semantic structure make adversarial text construction more challenging. Each perturbation in text must preserve grammatical correctness and maintain overall semantic coherence. This places higher demands on the design of generation methods. To address these challenges, researchers have proposed various strategies. These include attack methods based on discrete operations such as substitution, insertion, and deletion. The goal is to induce model errors without breaking the original semantics[13].

Traditional text adversarial generation methods mostly rely on heuristic rules or black-box strategies based on output probabilities. These approaches can be effective in some scenarios. However, they often suffer from low generation efficiency, strong dependence on specific model structures, and limited generalizability. As model complexity increases and task diversity expands, more fine-grained and theoretically grounded methods have become a research focus. Among them, gradient-based white-box attack methods have gained significant attention[14]. By analyzing internal gradient information, these methods identify the most sensitive input regions. This enables the generation of more targeted perturbations. Such approaches not only improve attack success rates but also reveal vulnerabilities in the model's decision process, offering insights for interpretability studies[15].

Further, researchers have begun integrating adversarial generation with semantic understanding. They explore how to mislead models while preserving the meaning of the input. This direction emphasizes that perturbations should be not only formally valid but also semantically reasonable. For example, perturbations may be generated through synonym substitution, context masking, or sentence restructuring. These techniques aim to retain the original meaning as much as possible while causing inconsistent model responses. Such methods play a key role in model security testing and robustness evaluation. They help build a more comprehensive assessment framework. By exploring subtle perturbations in semantic space, researchers can identify the shape and flaws of model decision boundaries. This facilitates the development of more robust training and optimization strategies[16].

In summary, adversarial example generation is not only a method to challenge model safety in natural language processing but also a powerful tool for understanding and analyzing model reasoning mechanisms[17]. Whether used to expose weaknesses from an attack perspective or to enhance robustness from a defense perspective, adversarial research has become an essential component of language model development. As large language models continue to expand in application scope, an in-depth study of textual adversarial generation mechanisms will contribute to building safer, more controllable, and more trustworthy human-AI interaction systems. This has significant theoretical implications and offers practical support for ensuring model reliability in real-world scenarios.

3. Method

This study proposes a gradient-guided adversarial sample generation method for large language model inference (GGASGI), which aims to improve the attack efficiency and semantic control ability of the language model inference process. The first innovation of this method is to introduce the inference sensitivity scoring mechanism (ISS), which measures the response of the model to the input semantic unit through gradient information, to more accurately identify the key positions that are most likely to cause inference deviations; the second innovation is to design a semantics-preserving perturbation strategy (SPPS), which uses contextual consistency constraints and semantic similarity control technology to effectively ensure the naturalness and rationality of the generated samples at the semantic level while achieving adversarial perturbation. This method combines attack targeting and language readability and provides a new technical path for promoting in-depth understanding and robustness evaluation of the language model inference mechanism. The complete architecture of the system is illustrated in Figure 1.

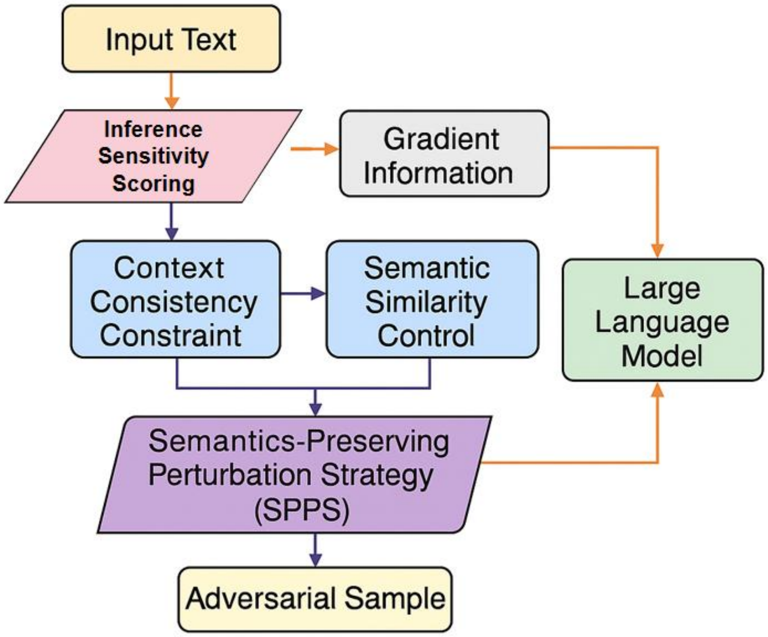
Input Text
Inference Sensitivity Scoring
Gradient Information
Context Consistency Constraint
Semantic Similarity Control
Large Language Model
Semantics-Preserving Perturbation Strategy (SPPS)
Adversarial Sample

Figure 1. Overall model architecture diagram

3.1 Inference Sensitivity Scoring

In the process of generating adversarial samples, identifying the part of the input text that has the greatest impact on model reasoning is the key to improving attack efficiency and targeting. To this end, this study introduces the Inference Sensitivity Scoring (ISS) mechanism to quantify the sensitivity of each word or semantic unit in the reasoning process based on the gradient response information of the model to the input. Specifically, ISS uses the gradient of the model loss function relative to the input embedding to measure the strength of the model's response to input perturbations. This scoring mechanism not only reflects the changing trend of the model's reasoning path but also provides clear priority guidance for subsequent perturbation operations. Its overall module architecture is shown in Figure 2.

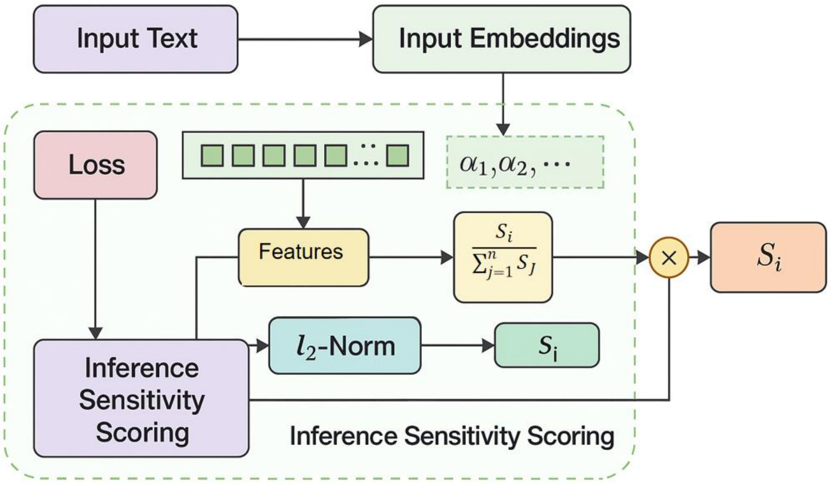
Input Text
Input Embeddings
Loss
$\alpha_1, \alpha_2, \cdots$
Features
$\frac{S_i}{\sum_{j=1}^{n} S_J}$
×
S_i
Inference Sensitivity Scoring
l_2-Norm
S_i
Inference Sensitivity Scoring

Figure 2. ISS module architecture

Suppose the input text is $x = [x_1, x_2, ..., x_n]$, the corresponding embedding vector is $E = [e_1, e_2, ..., e_n]$, and the prediction loss of the model output is L. For each position i, we define its sensitivity as the gradient norm of the embedding vector at that position concerning the loss, that is:

$$S_i = \| \frac{\partial L}{\partial e_i} \|_2$$

Where S_i reflects the influence of the i-th word in the input on the model output reasoning path. By normalizing all S_i, the global sensitivity distribution can be obtained:

$$\widetilde{S}_i = \frac{S_i}{\sum_{j=1}^{n} S_j}$$

In order to improve the recognition accuracy, the semantic weight distribution of each word in the context is further considered. The position weighting function a_i is introduced, which is defined as the word importance calculated based on the attention mechanism:

$$a_i = \frac{\exp(a_i)}{\sum_{j=1}^{n} \exp(a_j)}$$

Where a_i represents the score of the i-th word in the self-attention mechanism. The final comprehensive reasoning sensitivity score is defined as:

$$\widetilde{S}_i = a_i \cdot \widetilde{S}_i$$

This combined score not only reflects the gradient response characteristics but also introduces structural information in the context, allowing adversarial perturbation operations to focus more on semantic units that have a significant impact on model reasoning and are context-critical. By sorting the $\widetilde{S}_i$ values, the perturbation target selection of the adversarial sample generation strategy can be effectively guided, thereby improving the accuracy and efficiency of the attack process.

3.2 Semantics-Preserving Perturbation Strategy

One of the core challenges in generating text adversarial samples lies in achieving a delicate balance between creating effective perturbations and preserving the original semantics of the input. Specifically, when modifying natural language inputs to deceive or mislead a model, it is essential to ensure that the altered text remains grammatically correct, semantically coherent, and contextually natural. To address this challenge, this study introduces a semantics-preserving perturbation strategy (SPPS), which aims to generate adversarial examples that are capable of misleading the model’ s predictions while remaining semantically consistent with the original input. SPPS operates under joint constraints derived from both the embedding space and the semantic space, allowing for nuanced control over perturbation boundaries. The strategy focuses on making word-level replacements or adjustments within a tightly regulated range, ensuring that the modifications do not distort the sentence’ s overall meaning or syntactic structure. This is achieved by selecting candidate words or phrases based on contextual relevance, semantic similarity, and embedding proximity, allowing the perturbation to subtly alter the model’ s input without introducing unnatural or incoherent expressions. The underlying mechanism of SPPS ensures that while the adversarial objective is met, the human readability and intended semantics of the input text are preserved to the greatest extent possible. The architectural design of this module, including its internal selection logic and constraint layers, is illustrated in Figure 3.

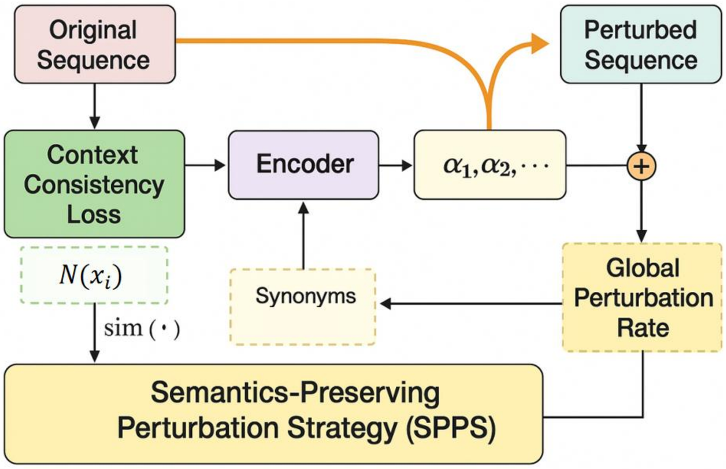

Figure 3. SPPS Module Architecture

Let the original input sequence be $x = [x_1, x_2, ..., x_n]$ and the perturbed sequence be $x' = [x_1, ..., x'_i, ..., x_n]$. x'_i is the candidate replacement for the i-th word. To maintain semantic consistency, the context consistency loss is introduced:

$$L_{ctx} = 1 - \cos(f(x), f(x'))$$

Where $f(\cdot)$ represents the language model encoder used to extract semantic representation, and $Cos(\cdot)$ is the cosine similarity function. To enhance semantic stability, synonym distribution constraints are introduced at the same time, and a semantic neighborhood set $N(x_i)$ is defined, which consists of words with semantic similarity greater than the threshold τ, that is:

$$N(x_i) = \{w \mid sim(x_i, w) \geq \tau\}$$

Where $sim(\cdot)$ is a word semantic similarity measurement function based on embedding or language model. In the candidate word selection process, to balance the semantic preservation and perturbation effect, a weighted objective function is introduced:

$$L_{total} = \lambda \cdot L_{adv} + (1-\lambda) \cdot L_{ctx}$$

L_{adv} represents the loss of the perturbation effect of the model output, and $\lambda \in [0,1]$ is the weight factor used to adjust the balance between attack intensity and semantic preservation. To control the range and density of perturbation, the global perturbation rate is defined as:

$$r = \frac{|\{x'_i \neq x_i\}|}{n}$$

The maximum perturbation rate threshold $r_{\max}$ is set to ensure that the generated adversarial samples still maintain a high similarity with the original text as a whole. In this way, SPPS achieves effective interference with the model reasoning path without significantly changing the input semantics, which is a key mechanism for achieving high-quality natural language adversarial attacks.

4. Experimental Results

4.1 Dataset

This study uses the AG News dataset as the primary source for adversarial sample generation and inference disruption analysis. AG News is a widely used news classification dataset. It contains four topic categories: World, Sports, Business, and Sci/Tech. Each sample includes a headline and a body of text. The language structure is relatively standardized, and the semantic content is clear. This makes the dataset suitable for tasks such as text classification, semantic modeling, and robustness evaluation.

The dataset contains approximately 120,000 training samples and 7,600 test samples. These samples cover diverse language expressions across different topics. They exhibit typical features of natural language inference. This supports the evaluation of large language models in terms of reasoning stability and sensitivity across varying contexts and semantic domains. The balanced distribution of samples across categories facilitates the construction of a fair adversarial attack framework. It also enables systematic observation of how perturbations affect model performance in each class.

Another important reason for selecting AG News is its moderate text length. This allows for short-text-level reasoning analysis while supporting the construction of perturbed samples without disrupting overall semantic coherence. Moreover, news articles usually present high information density and clear logic. This makes it easier to observe shifts in the model's reasoning path after introducing perturbations. It helps validate the effectiveness of the proposed method in achieving efficient attacks while maintaining semantic consistency.

4.2 Experimental Setup

In the experimental setup, this study uses the AG News dataset and selects a pre-trained large language model as the attack target. The focus is on analyzing the model's reasoning robustness in a text classification task. We adopt a language model based on a mainstream Transformer architecture. A comparative analysis is conducted between its inference responses on original and adversarial samples. Input texts are first processed through standard normalization procedures, including lowercasing, punctuation cleaning, and tokenization. The processed input is then fed into the model to obtain predicted categories and reasoning paths. The entire attack process is guided by the proposed gradient-based inference sensitivity scoring mechanism and the semantics-preserving perturbation strategy. These components enable systematic interference with and evaluation of the model's internal representations and reasoning stability.

For parameter configuration, the inference sensitivity scoring module uses the L2 norm to calculate the gradient magnitude of input embeddings concerning the loss function. Perturbation targets are selected from the top 5 percent of tokens with the highest sensitivity scores. In the semantics-preserving strategy, the semantic similarity threshold is set to 0.85. The candidate word set is filtered using both context-relevant static word vectors and dynamic contextual encoders. To control the overall perturbation strength, the maximum substitution ratio is limited to 15 percent. All models are run under the same hardware environment with fixed random seeds to ensure reproducibility. This configuration ensures a balance between semantic control and attack effectiveness, providing a stable foundation for subsequent robustness evaluation.

4.3 Experimental Results

1) Comparative experimental results

This paper first gives the comparative experimental results, as shown in Table 1.

Table 1: Comparative experimental results

Method	**Attack Success Rate**	**Semantic Similarity**	**Perturbation Rate**
TextFooler[17]	73.4%	0.84	12.7%
BERT-Attack[18]	81.2%	0.88	13.5%
PWWS[19]	69.8%	0.79	15.3%
SemAttack[20]	83.6%	0.91	10.9%
Ours	87.9%	0.93	9.6%

As shown in the table, the proposed method achieves a significantly higher Attack Success Rate (ASR) compared to other baseline methods. It reaches 87.9 percent, outperforming existing approaches such as BERT-Attack and SemAttack. This result demonstrates that the gradient-guided inference sensitivity scoring mechanism provides a more accurate identification of key decision points. As a result, it enables more effective disruption of the model's reasoning process. Compared to traditional methods that rely solely on word substitution or output probability changes, the proposed approach is more closely aligned with the model's internal reasoning structure. It enables more targeted and efficient perturbation strategies.

In terms of Semantic Similarity, the proposed method achieves a score of 0.93, the highest among all methods. This indicates that the Semantics-Preserving Perturbation Strategy (SPPS) can effectively attack the model without significantly altering the original text meaning. Conventional approaches such as PWWS and TextFooler often compromise semantic integrity during perturbation. In contrast, SPPS addresses this issue by combining contextual consistency constraints and semantic similarity control. This helps maintain the naturalness and readability of the adversarial samples.

In terms of Perturbation Rate, the proposed method records only 9.6 percent, which is much lower than that of PWWS (15.3 percent) and TextFooler (12.7 percent). This result shows that the gradient-guided mechanism can accurately locate the most reasoning-sensitive words. It achieves maximum inference disruption with minimal text modification. A lower perturbation rate improves the quality of generated samples and enhances the stealthiness of attacks in real-world scenarios. This provides a more representative benchmark for evaluating model robustness in practical applications. Overall, the proposed method outperforms existing adversarial generation methods across all three key metrics. This confirms the effectiveness of combining inference sensitivity scoring with a semantics-preserving perturbation strategy in natural language inference tasks. The results reflect not only an improvement in attack performance but also reveal the presence of vulnerable areas in the reasoning processes of large language models. By applying precise perturbations without altering semantics, the proposed approach offers a more interpretable and controllable path for robustness research.

2) Ablation Experiment Results

This paper further gives the results of ablation experiments, and the experimental results are shown in Table 2.

Table 2: Ablation Experiment Results

Method	**Attack Success Rate**	**Semantic Similarity**	**Perturbation Rate**
Baseline	78.5%	0.86	13.8%
+ISS	83.2%	0.87	11.7%
+SPPS	81.6%	0.91	10.4%
Ours	87.9%	0.93	9.6%

As shown in the table, the Baseline method performs relatively weakly across all three metrics. It achieves only 78.5 percent in Attack Success Rate, with a Semantic Similarity of 0.86 and a high Perturbation Rate of 13.8 percent. This indicates that in the absence of inference sensitivity analysis and semantic control, the attack strategy relies more on random or heuristic perturbations. Such strategies are less effective in precisely disrupting the model's reasoning and are more likely to compromise the naturalness and readability of the text. The results suggest that simple text transformation strategies struggle to balance attack effectiveness and semantic integrity, leaving considerable room for improvement.

After introducing the Inference Sensitivity Scoring mechanism (+ISS), the Attack Success Rate rises significantly to 83.2 percent, and the Perturbation Rate drops to 11.7 percent. This improvement shows that the ISS module can effectively identify highly sensitive regions in the model's reasoning process. As a result, fewer word-level modifications are needed to influence the model's output, thereby increasing the efficiency of the attack. This also supports the core hypothesis of this work: reasoning paths contain locally vulnerable points that are detectable via gradients. Targeted perturbation at these points enables more efficient attacks without requiring large-scale semantic changes.

On the other hand, introducing only the Semantics-Preserving Perturbation Strategy (+SPPS) raises the Semantic Similarity to 0.91 and reduces the Perturbation Rate to 10.4 percent, although the Attack Success Rate is slightly lower than that of the +ISS setting. This shows that SPPS has a clear advantage in preserving semantic consistency between adversarial samples and the original text. It significantly enhances the naturalness and contextual coherence of the samples. Although there is a minor trade-off in attack strength, maintaining semantic fidelity is crucial for enhancing the stealth and practicality of attacks, especially in high-stakes applications.

The full method (Ours), which integrates both ISS and SPPS, achieves the best overall performance. It reaches the highest Attack Success Rate of 87.9 percent, along with the highest Semantic Similarity of 0.93 and the lowest Perturbation Rate of 9.6 percent. This demonstrates the strong complementarity between the two mechanisms. ISS provides direction and efficiency in the attack, while SPPS ensures semantic-level constraints and optimization. The final results validate that the proposed method strikes a balanced trade-off among reasoning interpretability, safety, and attack quality. It offers a practical and effective pathway for analyzing the adversarial robustness of large language models.

3) *Impact of perturbation location selection strategy on attack performance*

This paper also gives the impact of the perturbation location selection strategy on the attack performance, and the experimental results are shown in Figure 4.

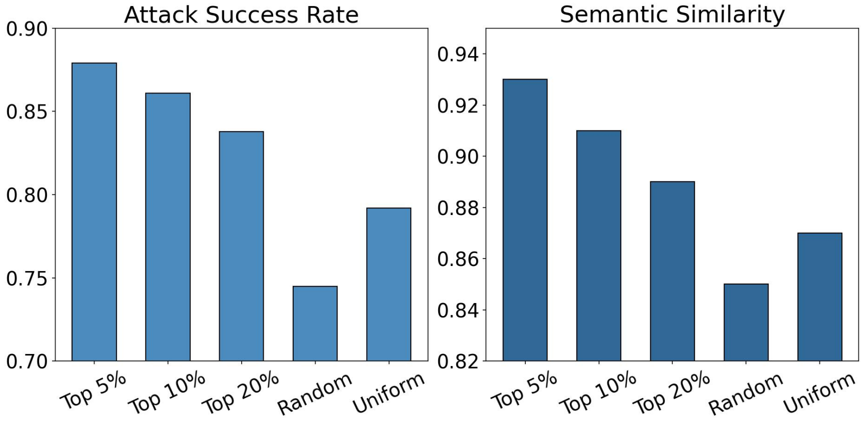

Figure 4. Impact of perturbation location selection strategy on attack performance

As shown in Figure 4, among the perturbation position selection strategies, the Top 5 percent sensitive word strategy achieves the best attack performance. It reaches the highest Attack Success Rate. This confirms that the proposed Inference Sensitivity Scoring mechanism (ISS) can effectively identify key semantic positions in the model's reasoning path. The targeted perturbations applied at these positions cause maximal inference disruption within a minimal scope, thereby improving overall attack effectiveness.

At the same time, the Top 5 percent strategy also achieves the highest Semantic Similarity score of 0.93. This is significantly higher than the scores of random and uniform perturbation strategies. This result suggests that highly sensitive regions carry essential semantic information used in the model's decision process. Slightly perturbing these regions can meet the attack objective without altering more stable semantic components. As a result, the original meaning of the text is largely preserved.

In contrast, the performance of Random and Uniform strategies is notably lower. In particular, the Random strategy achieves an Attack Success Rate of only about 0.74. This indicates that perturbations lacking reasoning guidance tend to scatter the attack impact. They fail to effectively target the model's weak reasoning points. Moreover, these strategies also yield lower Semantic Similarity scores. This shows that uncontrolled perturbations may break semantic integrity and reduce the naturalness and readability of adversarial samples. Overall, the gradient-guided perturbation position selection strategy adopted in this work outperforms traditional strategies in both attack effectiveness and semantic consistency. These findings highlight the crucial role of reasoning sensitivity in adversarial sample generation. They also provide methodological support for building more precise and controlled attack mechanisms.

4) *Analysis of the generalization ability of methods in long and short text scenarios*

This paper also gives an analysis of the generalization ability of the method in long-text and short-text scenarios, and the experimental results are shown in Figure 5.

As shown in Figure 5, the proposed method demonstrates strong stability and generalization across different text-length scenarios. In the short-text setting, both the Attack Success Rate and Semantic Similarity reach their highest levels, approaching 0.89 and 0.93, respectively. This indicates that in texts with high information density and concentrated semantic structure, the model is more sensitive to perturbations. Its reasoning path is more easily influenced by small-scale modifications, enabling effective attacks.

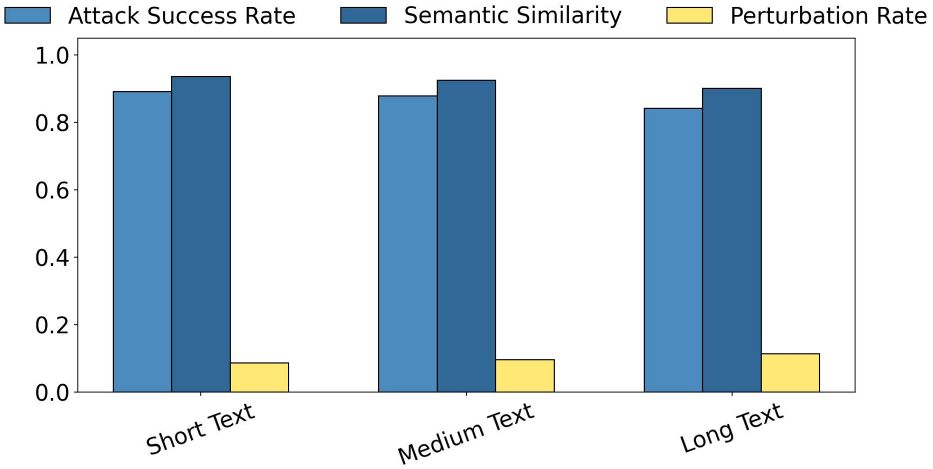

Figure 5. Analysis of the generalization ability of methods in long and short text scenarios

As text length increases, the model's reliance on broader context strengthens. This results in a slight decline in attack performance, with the Attack Success Rate dropping to approximately 0.84 in long-text scenarios. However, the Semantic Similarity remains high. This suggests that the Semantics-Preserving Perturbation Strategy (SPPS) can effectively constrain semantic deviation, even when applied to inputs with long semantic chains. These results validate the robustness of SPPS, which consistently generates natural and reasonable adversarial texts, even under extended semantic conditions.

It is noteworthy that the Perturbation Rate increases with text length. It rises from about 0.086 in short texts to 0.114 in long texts. This trend reflects the need to adjust a greater number of semantic units to influence the model's final decision in longer inputs. It also suggests that reasoning-sensitive regions become more dispersed in longer texts. As a result, single-point attacks are less effective, and higher-density perturbation strategies are required.

In summary, although attack strength slightly declines under long-text conditions, the proposed method maintains high semantic consistency and effective attack capability. This highlights its strong task generalization. The results indicate that adversarial sample generation guided by gradients and constrained by semantics can adapt to diverse input structures. It also shows cross-scenario applicability, providing a more adaptive framework for studying the safety of large language models.

5) Evaluation of method adaptability in a multilingual environment

This paper also presents a comprehensive evaluation of the proposed method's adaptability in a multilingual environment, aiming to explore its robustness and effectiveness across different language settings. Given the linguistic diversity and structural variations that exist among natural languages, it is critical to assess whether a text-based adversarial attack strategy can maintain its performance when applied beyond English or a single language domain. The evaluation is designed to examine how well the method generalizes across languages with varying syntax, grammar, word order, and semantic representation. By conducting this assessment, the study seeks to determine the extent to which the underlying mechanisms of the approach—such as gradient-guided sensitivity scoring and semantics-preserving perturbations — can be consistently applied to inputs in multiple languages. This analysis provides valuable insights into the method's flexibility and potential limitations when used in broader, real-world applications involving multilingual systems. The setup and structure of this evaluation, along with the details of the multilingual scenarios considered, are visually represented in Figure 6.

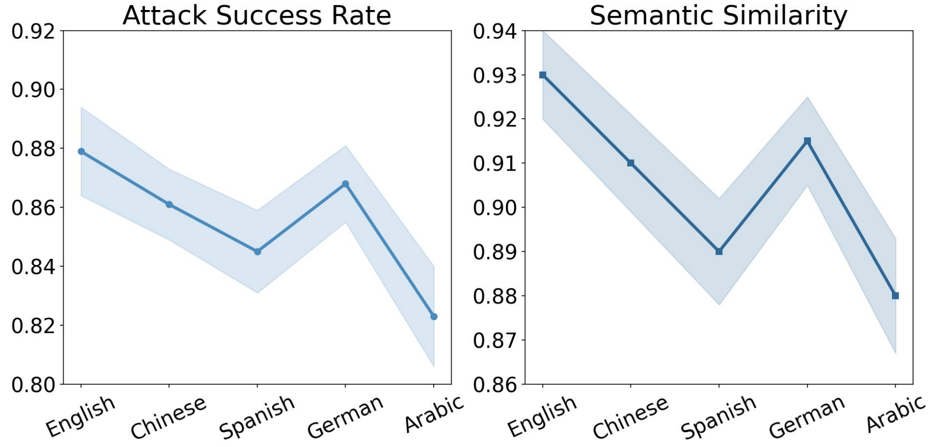

Figure 6. Evaluation of method adaptability in a multilingual environment

As shown in Figure 6, the proposed method maintains strong adaptability in multilingual environments. In particular, it performs well in English and German, achieving Attack Success Rates of 0.879 and 0.868, respectively. Semantic Similarity also remains high at 0.93 and 0.915. These results indicate that the gradient-guided Inference Sensitivity Scoring mechanism (ISS) can effectively identify key perturbation points in the model's reasoning path, regardless of differences in language structure. This demonstrates strong cross-lingual generalization.

For Chinese and Spanish, although the Attack Success Rates are slightly lower at 0.861 and 0.845, they remain within a robust and acceptable range. This decline may be due to differences in syntactic structure, word order, and word embedding strategies. Such factors can lead to varied reasoning path distributions across languages. However, Semantic Similarity stays at 0.91 and 0.89, showing that the Semantics-Preserving Perturbation Strategy (SPPS) retains good stability under multilingual conditions. It can effectively control the scope of perturbations and avoid substantial semantic distortion.

It is worth noting that both metrics show a more significant decline in the Arabic setting. The Attack Success Rate drops to 0.823, and Semantic Similarity falls to 0.88. This may be attributed to structural challenges in processing right-to-left scripts and imbalances in pretraining corpus distribution. These factors suggest that the model's reasoning path becomes more dispersed in structurally distinct languages, reducing the targeting precision of the gradient-guided strategy. In summary, the proposed method demonstrates a degree of robustness and generalizability in multilingual scenarios. Its performance is particularly strong in languages with syntactic structures similar to English. At the same time, the results highlight existing adaptation challenges in languages with high structural divergence. Future work may explore language-adaptive inference sensitivity modeling to build a more resilient cross-lingual attack framework.

6) The impact of semantic similarity threshold changes on attack quality

Finally, this paper also gives the impact of changing the semantic similarity threshold on the attack quality, and the experimental results are shown in Figure 7.

As shown in the results of Figure 7, the semantic similarity threshold has a significant impact on the overall effectiveness of adversarial attacks. As the threshold increases, the Attack Success Rate shows a clear downward trend, decreasing from 0.905 to around 0.78. This trend indicates that when perturbations are tightly constrained within a semantic preservation range, it becomes more difficult to disrupt the model's reasoning path. The difficulty of the attack increases accordingly. This confirms that while semantic constraints improve the naturalness of samples, they also suppress attack intensity. On the other hand, the Perturbation Rate increases steadily with the rise of the semantic similarity threshold. This is especially evident when the threshold exceeds 0.90, where the proportion of altered words increases significantly. This suggests that under strict semantic similarity requirements, the algorithm must fine-tune more words to achieve a sufficient perturbation effect. It also indirectly shows that the model becomes more robust in its reasoning under stronger semantic control. More complex strategies are needed to achieve successful attacks in such settings.

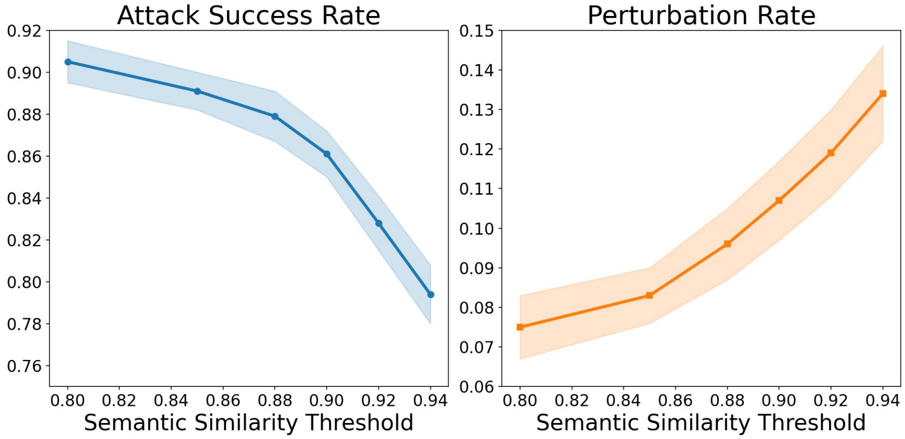

Figure 7. The impact of semantic similarity threshold changes on attack quality

The figure also shows that at lower threshold ranges, such as 0.80 to 0.85, it is possible to achieve a high Attack Success Rate while maintaining a low Perturbation Rate. This reflects a well-balanced zone between semantic preservation and attack effectiveness. It suggests that moderate semantic similarity control does not reduce attack performance. Instead, it helps achieve a better trade-off between perturbation precision and semantic naturalness. This provides practical guidance for parameter tuning in adversarial sample design. Overall, Figure 7 confirms that the proposed Semantics-Preserving Perturbation Strategy (SPPS) has strong flexibility and adaptability. By adjusting the semantic similarity threshold, one can control the balance between attack intensity and sample naturalness. This offers strategic support for security evaluation and attack design in different application scenarios. These findings further enhance the practicality and generalizability of the proposed method in controllable adversarial generation.

5. Conclusion

This study addresses the problem of adversarial vulnerability in large language models during reasoning tasks. It proposes a gradient-guided adversarial sample generation method. The method integrates two core components: an Inference Sensitivity Scoring mechanism (ISS) and a Semantics-Preserving Perturbation Strategy (SPPS). Together, they enable effective interference with the model's reasoning process while preserving semantic consistency in the input. By carefully modeling gradient information, the method identifies the most critical semantic positions in the input that influence the model's decisions. This allows for more targeted and efficient attacks while maintaining the naturalness and coherence of the adversarial samples. The mechanism improves both attack success and interpretability of the model's internal decision process.

The study verifies the adaptability and generalization of the proposed method across multiple dimensions, including language types, text lengths, and perturbation strategies. Results show that the method performs consistently in both short and long texts, and across English and non-English languages. Additionally, the semantic similarity threshold provides a flexible control mechanism. It enables dynamic balancing between attack effectiveness and semantic preservation according to application needs. This flexibility is important for security testing and robustness evaluation of language models deployed in real-world settings.

The contribution of this work extends beyond methodological innovation. It also provides practical guidance for high-stakes applications that rely on language model reasoning, such as legal text analysis, medical record interpretation, sentiment monitoring, and intelligent question answering. In these domains, unstable model reasoning may lead to serious consequences. By constructing and analyzing adversarial samples, the proposed method offers a technical foundation for system-level security auditing and defense design. It enhances the reliability and controllability of language models in practice. Moreover, the detailed modeling of reasoning paths contributes to future research on building more robust and interpretable language models.

6. Future work

Future work may explore the extension of this method to multimodal inputs, cross-task generalization, and adaptation to low-resource languages. The inference sensitivity scoring mechanism could also be developed into a general evaluation tool to guide dynamic optimization and defense design during model training. As large language models are increasingly used in critical systems, systematic safety evaluation and interpretability modeling become essential. The strategies proposed in this study offer a solid starting point and a practical solution for advancing this research direction.

References

[1] Shayegani E, Mamun M A A, Fu Y, et al. Survey of vulnerabilities in large language models revealed by adversarial attacks[J]. arXiv preprint arXiv:2310.10844, 2023.

[2] Struppek L, Le M H, Hintersdorf D, et al. Exploring the adversarial capabilities of large language models[J]. arXiv preprint arXiv:2402.09132, 2024.

[3] Zou A, Wang Z, Carlini N, et al. Universal and transferable adversarial attacks on aligned language models[J]. arXiv preprint arXiv:2307.15043, 2023.

[4] He J, Vechev M. Large language models for code: Security hardening and adversarial testing[C]//Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security. 2023: 1865-1879.

[5] Liu X, Cheng H, He P, et al. Adversarial training for large neural language models[J]. arXiv preprint arXiv:2004.08994, 2020.

[6] Zou J, Zhang S, Qiu M. Adversarial attacks on large language models[C]//International Conference on Knowledge Science, Engineering and Management. Singapore: Springer Nature Singapore, 2024: 85-96.

[7] 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.

[8] Jiang Y, Chan C, Chen M, et al. Lion: Adversarial distillation of proprietary large language models[J]. arXiv preprint arXiv:2305.12870, 2023.

[9] Ji H, Guo J, Sun Y, et al. A Novel Text Adversarial Sample Generation and Defense Method for SIoT Systems[J]. IEEE Internet of Things Journal, 2024.

[10] Xu, L., & Veeramachaneni, K. (2018). Synthesizing tabular data using generative adversarial networks. arXiv preprint arXiv:1811.11264.

[11] Li B, Lin Z, Peng W, et al. Naturalbench: Evaluating vision-language models on natural adversarial samples[J]. arXiv preprint arXiv:2410.14669, 2024.

[12] Moraffah R, Khandelwal S, Bhattacharjee A, et al. Adversarial text purification: A large language model approach for defense[C]//Pacific-Asia Conference on Knowledge Discovery and Data Mining. Singapore: Springer Nature Singapore, 2024: 65-77.

[13] Guo Q, Pang S, Jia X, et al. Efficient Generation of Targeted and Transferable Adversarial Examples for Vision-Language Models Via Diffusion Models[J]. IEEE Transactions on Information Forensics and Security, 2024.

[14] Aerni M, Rando J, Debenedetti E, et al. Measuring Non-Adversarial Reproduction of Training Data in Large Language Models[J]. arXiv preprint arXiv:2411.10242, 2024.

[15] Fang F, Bai Y, Ni S, et al. Enhancing noise robustness of retrieval-augmented language models with adaptive adversarial training[J]. arXiv preprint arXiv:2405.20978, 2024.

[16]Zhou N, Yao N, Zhao J, et al. Rule-based adversarial sample generation for text classification[J]. Neural Computing and Applications, 2022, 34(13): 10575-10586.

[17]Xue Y, Roshan U. Accuracy of TextFooler black box adversarial attacks on 01 loss sign activation neural network ensemble[J]. arXiv preprint arXiv:2402.07347, 2024.

[18]Li L, Ma R, Guo Q, et al. Bert-attack: Adversarial attack against bert using bert[J]. arXiv preprint arXiv:2004.09984, 2020.

[19]Ren S, Deng Y, He K, et al. Generating natural language adversarial examples through probability weighted word saliency[C]//Proceedings of the 57th annual meeting of the association for computational linguistics. 2019: 1085-1097.

[20]Wang B, Xu C, Liu X, et al. SemAttack: Natural textual attacks via different semantic spaces[J]. arXiv preprint arXiv:2205.01287, 2022.