ExplainableGuard: Interpretable Adversarial Defense using Chain-of-Thought Reasoning with DeepSeek-Reasoner

Anonymous ACL submission

Abstract

Large Language Models (LLMs) are increasingly vulnerable to adversarial attacks that can subtly manipulate their outputs. While various defense mechanisms have been proposed, many operate as black boxes, lacking transparency in their decision-making. This paper introduces ExplainableGuard, an interpretable adversarial defense framework leveraging the chain-of-thought (CoT) reasoning capabilities of DeepSeek-Reasoner. Our approach not only detects and neutralizes adversarial perturbations in text but also provides step-by-step explanations for each defense action. We demonstrate how tailored CoT prompts guide the LLM to perform a multi-faceted analysis (character, word, structural, semantic) and generate a purified output along with a human-readable justification. Preliminary results on BLUE and IMDB show promising defense efficacy while offering crucial insights into the attack vectors and defense rationale, paving the way for more trustworthy LLM deployments.

1 Introduction

Large Language Models (LLMs) like GPT-4 (Achiam et al., 2023), Llama (Touvron et al., 2023), and others have demonstrated remarkable capabilities across diverse NLP tasks. However, their widespread adoption is hampered by their susceptibility to adversarial attacks (Goodfellow et al., 2015; Jin et al., 2020). These attacks involve crafting subtle, often human-imperceptible perturbations to input text, causing LLMs to produce erroneous, biased, or harmful outputs.

Existing defense strategies range from input preprocessing and adversarial training to detection mechanisms (Jia et al., 2019; Zhu et al., 2020). While effective to some extent, many of these methods lack transparency. Understanding "why" a specific input was flagged or modified is crucial for building trust, debugging models, and iterating on security measures. This is particularly important in high-stakes applications.

To address this gap, we propose ExplainableGuard, a novel defense mechanism that utilizes the advanced reasoning abilities of a powerful LLM, DeepSeek-Reasoner, to not only defend against adversarial attacks but also to explain its defense process. Our core contribution lies in designing a structured, Chain-of-Thought (CoT) (Wei et al., 2022) prompting strategy that elicits detailed reasoning from the defense LLM. This reasoning breaks down the analysis into character, word, structural, and semantic levels, culminating in a decision, a purified text, and a comprehensive explanation.

This paper details the architecture of ExplainableGuard, the CoT prompting methodology, and presents its potential for robust and interpretable defense. We aim to demonstrate that such a system can effectively mitigate various attack types while providing valuable insights into its operational logic.

2 Related Work

2.1 Adversarial Attacks on LLMs

Adversarial attacks on LLMs involve subtle perturbations of the input text that alter the predictions of the model without affecting the human-perceived meaning of the text (Xu et al., 2023). Adversarial attacks can be categorized into several levels. Character-level attacks involve manipulations such as typos, homoglyphs, or invisible characters (Ebrahimi et al., 2018). Word-level attacks involve replacing words with synonyms, paraphrasing sentences, or inserting/deleting words (Alzantot et al., 2018). More sophisticated attacks target sentence structure or semantics, including prompt injection and jailbreak techniques (Zou et al., 2023). These types of attacks expose the vulnerability of current LLMs, highlighting the need for more robust defenses (Wang et al., 2023).

2.2 Adversarial Defense Mechanisms

Defense usually involves input cleaning (e.g., filtering special characters, spell checking), adversarial training (fine-tuning models of adversarial examples)(Goyal et al., 2023; Jia et al., 2019), and in the field of image classification, adversarial defense also involves authentication defense that provides robustness guarantees (Croce et al., 2020). Some methods employ detector models to flag malicious inputs (Mozes et al., 2023). However, many of these defenses do not clearly explain why an input is considered adversarial or how an attack is eliminated.

2.3 Explainable AI (XAI) in NLP

Explainability in natural language processing seeks to render model predictions and internal reasoning processes transparent and interpretable (Danilevsky et al.). Traditional approaches include attention visualization such as AttentionViz (Yeh et al., 2023) and feature-attribution methods such as LIME (Ribeiro et al., 2016) and SHAP (Lundberg and Lee, 2017) More recently, *natural language explanations* have been introduced to generate human-readable justifications alongside model outputs, often by training on datasets annotated with explanatory comments (Danilevsky et al.).

A particularly promising paradigm is *chain-of-thought* (CoT) prompting, whereby large language models are induced to produce intermediate reasoning steps that lead to a final answer (Wei et al., 2022). This mechanism not only provides the final output of the model, but also generates verifiable decision sequences that can be carefully examined by researchers. Our work utilizes CoT to generate explanations in the context of adversarial defense.

3 ExplainableGuard: Methodology

Our proposed system, ExplainableGuard, employs DeepSeek-Reasoner (DeepSeek-AI et al., 2025) as a security analyst LLM. Given potentially adversarial input text T_{adv}, the goal is to produce a cleaned version T_{clean} and a human-readable short explanation E detailing the purification content. Additionally, our system will produce a reasoning process R containing how our LLM analysis the text, detect adversarial patterns and do the purification. The workflow of ExplainableGuard is illustrated in Figure 1.

3.1 System Overview

Our method can be formalized as a function $D : \mathcal{T} \to (\mathcal{T}, \mathcal{E}, \{0,1\}, \mathcal{R})$, where $\mathcal{T}$ denotes the space of texts, $\mathcal{E}$ denotes the space of explanations, and $\mathcal{R}$ denotes the space of reasoning contents. Given an adversarial input T_{adv}, the function outputs a tuple:

$$(T_{clean}, E, \text{is\_adv}, R) = D(T_{adv})$$

Here, T_{clean} is the purified text, E is a concise human-readable explanation, is_adv is a boolean indicating whether the input was identified as adversarial, and R is the detailed reasoning content generated by the model during the analysis and purification process.

3.2 Chain-of-Thought Prompting for Defense

The interpretability of ExplainableGuard is achieved through a carefully designed chain-of-thought (CoT) prompt, P_{CoT}, which systematically guides DeepSeek-Reasoner through a sequence of analytical steps. As illustrated in Figure 1, the prompt instructs the LLM to conduct a comprehensive assessment at multiple levels: starting with character-level inspection (e.g., detecting homoglyphs, invisible characters, typos, leetspeak), followed by word-level analysis (such as identifying unusual synonym usage or suspicious insertions/deletions), then examining structural aspects (like sentence structure anomalies or embedded commands), and finally performing semantic and contextual checks (to uncover subtle meaning shifts or indirect prompt injections).

After these analyses, the model determines whether the input is adversarial and formulates an appropriate purification strategy. The LLM then applies this strategy to generate the cleaned text T_{clean}. Finally, it produces a structured summary that includes the adversarial judgment (is_adv), the purified text, a concise explanation E, and a reasoning process R.

4 Experimental Setup

4.1 Dataset

We conducted experiments on both short and long text datasets.

GLUE Benchmark: For short text evaluation, we selected three representative tasks from the GLUE benchmark (Wang et al., 2018): SST-2, RTE, and QQP. These datasets are widely used

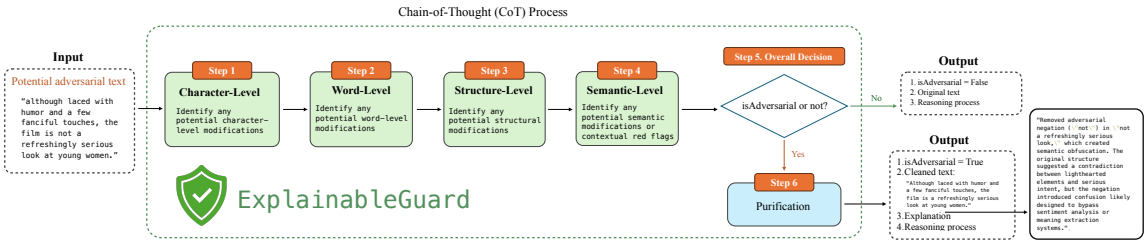

Figure 1: Overview of the ExplainableGuard Chain-of-Thought (CoT) workflow.

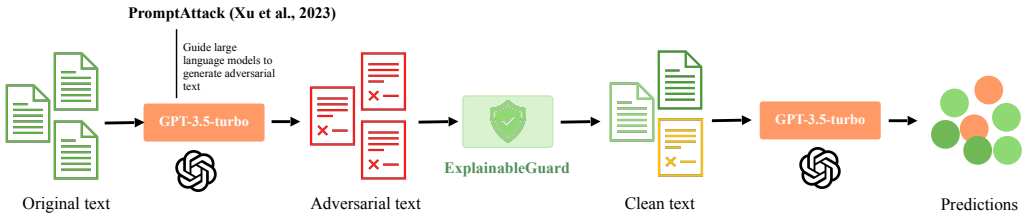

Figure 2: The workflow of the experiments, including the generation of adversarial examples and the defense of ExplainableGuard.

for assessing natural language understanding and are characterized by relatively short input texts.

IMDB Movie Reviews: For long text evaluation, we included the IMDB movie review dataset (Maas et al., 2011), which contains lengthy user-written reviews labeled for sentiment.

Together, these datasets allow us to evaluate our method's robustness and interpretability across diverse text lengths and task types. Further information about these datasets is available in the Appendix A.

4.2 Baselines

We compare ExplainableGuard with a baseline where no defense is applied. Specifically, we evaluate the performance of a target LLM (GPT-3.5-turbo) on adversarial inputs generated by the PromptAttack method (Xu et al., 2023), without any additional defense mechanism. The detailed attack strategies are described in the Appendix B.

4.3 Evaluation Metrics

We use two main evaluation metrics in our experiments, each serving a distinct purpose:

Attack Success Rate (ASR): ASR directly measures whether our defense can help the model make correct predictions in the presence of adversarial attacks. A lower ASR indicates that the defense is more effective at preventing the model from being fooled by adversarial inputs. It is defined as (Lin and Zhao, 2024):

$$\text{ASR} = \frac{|\{x \in D_{\text{correct}} : f(x') \neq y\}|}{|D_{\text{correct}}|} \quad (1)$$

where D_{correct} represents the set of samples that are correctly classified by the model on the original test dataset, x' is the corresponding adversarial example, $f(x')$ is the model's prediction result on the adversarial example, y is the true label.

BLEU Score: The BLEU score is used to measure the similarity between the purified text and the original, unperturbed text (Papineni et al., 2002). In our evaluation, we compute a weighted average of the 1-gram and 2-gram BLEU scores for each example, and then report the mean over all successfully defended examples. Let $C = \{1, 2\}$ denote the set of n-gram orders considered, and N be the total number of successful defense examples. The BLEU score is defined as:

$$\text{BLEU} = \frac{1}{N} \sum_{i=1}^{N} \sum_{n \in C} w_n \cdot \text{BLEU}_i^{(n)} \quad (2)$$

where $w_n = 0.5$ for $n = 1, 2$, and $\text{BLEU}_i^{(n)}$ is the n-gram BLEU score for the i-th example. A higher BLEU score indicates that our purification process better preserves the original content.

5 Preliminary Results and Analysis

5.1 Attack Success Rate

Table 1 presents the ASR against PromptAttack-EN (PA-EN) and PromptAttack-FS-EN (PA-FS-EN) attacks (Xu et al., 2023), comparing the performance of ExplainableGuard (EG) with the baseline model (GPT-3.5-turbo) without any defense. Across the three datasets, we observe a notable reduction in ASR when EG is applied. For instance, under the PA-EN attack, the ASR for RTE drops from 34.30% to 13.18%. The average ASR across all datasets is reduced from 37.27% to 24.21% (PA-EN) and 42.87% to 24.31% (PA-FS-EN). This indicates that EG effectively mitigates the adversarial attacks, enhancing the model's robustness. Additionally, for IMDB dataset, the ASR without defense is 38.71%, while applying EG reduces it to 30.11% as shown in Table 2. This indicates that our method is also effective for long-text adversarial defense.

5.2 BLEU Score

We evaluate the BLEU score of the purified text on successful defense results. Table 3 reports the BLEU scores for across all datasets. For SST-2, RTE, and QQP, both attack methods achieve high BLEU scores (>0.81), suggesting that ExplainableGuard can effectively clean adversarial inputs while maintaining semantic fidelity. On the IMDB dataset, the PA-EN BLEU score is 0.6195, indicating that EG can still effectively preserve much of the original content even on longer texts.

5.3 Explainability

The explanations generated by ExplainableGuard provide valuable insights into the defense process. For example, in the case of a PA-EN attack on SST-2, the model identifies specific character-level anomalies (e.g., "homoglyphs" or "typos") and word-level issues (e.g., "unusual synonym usage"). The explanation details how these factors contributed to the adversarial nature of the input and how they were addressed during purification. This level of transparency is crucial for understanding the model's decision-making process and building trust in its outputs. Some successful defense examples with explanation are provided in the Appendix C.

Attacks	EG	SST-2	RTE	QQP	Avg.
PA-EN	×	56.00	34.30	21.50	37.27
	✓	40.89	13.18	18.57	24.21
PA-FS-EN	×	75.23	36.12	17.26	42.87
	✓	48.61	10.32	14.01	24.31

Table 1: Performance comparison of defense methods against different attacks on SST-2, RTE, and QQP.

Attacks	EG	IMDB
PA-EN	×	38.71
	✓	30.11

Table 2: Performance comparison of defense methods against different attacks on IMDB.

Overall, these results demonstrate that ExplainableGuard substantially reduces the attack success rate across both short and long text datasets, while maintaining high similarity between the purified and original texts. This highlights the effectiveness and interpretability of our defense approach.

Method	SST-2	RTE	QQP	IMDB
PA-EN	0.82	0.8909	0.8626	0.6195
PA-FS-EN	0.85	0.81	0.8613	-

Table 3: BLEU scores for zeroshot and fewshot methods across different datasets.

6 Conclusion

We introduced ExplainableGuard, an adversarial defense system that utilizes DeepSeek-Reasoner and Chain-of-Thought prompting to detect, neutralize, and explain its actions against adversarial text. By guiding the LLM through a systematic analysis, our method provides not only a cleaned output but also a transparent rationale for its decisions. This approach enhances trustworthiness and provides valuable insights for users and security analysts. While further research is needed, ExplainableGuard demonstrates a promising direction for building more robust and understandable AI security systems.

Limitations

The preliminary results suggest that leveraging CoT reasoning in powerful LLMs like DeepSeek-Reasoner is a viable path towards interpretable adversarial defense. The structured analytical steps

(character, word, etc.) forced by the prompt not only improve detection but form the basis of the explanation. This transparency is a significant step over black-box defense models.

However, this approach is not without its limitations. First, the effectiveness of ExplainableGuard is inherently dependent on the capabilities of the underlying defense LLM, DeepSeek-Reasoner. If attackers develop more sophisticated adversarial strategies specifically targeting the weaknesses of the defense model, it may still be possible to bypass detection or purification. Second, the reliance on large-scale LLMs introduces practical concerns regarding latency and computational cost. In our experiments, processing a single input can take over 30 seconds, largely due to the complexity of generating detailed chain-of-thought reasoning. Such latency and resource demands may limit the applicability of this method in real-time or resource-constrained environments.

Future work may proceed in several directions. First, more rigorous evaluations could be conducted across a broader range of datasets and adversarial attack types to further validate the robustness of ExplainableGuard. Second, the quality of the generated explanations could be systematically assessed through comprehensive human studies. Additionally, approaches to distill the defense capabilities into smaller and more efficient models, while maintaining interpretability, are worth exploring. Finally, fine-tuning DeepSeek-Reasoner specifically for adversarial defense tasks may further enhance its effectiveness in this domain.

References

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

Moustafa Alzantot, Yash Sharma, Ahmed Elgohary, Bo-Jhang Ho, Mani Srivastava, and Kai-Wei Chang. 2018. Generating natural language adversarial examples. In *EMNLP*.

Francesco Croce, Maksym Andriushchenko, Vikram Sehwag, Edoardo Debenedetti, Nicolas Flammarion, Mung Chiang, Prateek Mittal, and Pascal Frossard. 2020. Robustbench: a standardized adversarial robustness benchmark. In *NeurIPS*.

Marina Danilevsky, Kunal Kashyap, Yannis Katsis, Griselda Q. Rushiti, Dennis Wei, and Oren Etzioni. A survey of the state of explainable ai for natural language processing. *AACL/IJCNLP*.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. 2018. Hotflip: White-box adversarial examples for text classification. In *ACL*.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*.

Shreya Goyal, Sumanth Doddapaneni, Mitesh M. Khapra, and Balaraman Ravindran. 2023. A survey of adversarial defenses and robustness in nlp. *ACM Comput. Surv.*, 55(14s).

Robin Jia, Aleksandar Risteski, and Dawn Song. 2019. Certified robustness to word substitutions with more character-level attacks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.

Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. 2020. BERT is not a panacea: Linguistic bias identification and multilingual attack on BERT-based text classification with TextFooler. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6030–6045.

Guang Lin and Qibin Zhao. 2024. Large language model sentinel: Llm agent for adversarial purification. *arXiv preprint arXiv:2405.20770*.

Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *NIPS*.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

Maximilian Mozes, Badr Alkhamissy, and Soroush Vosoughi. 2023. Use of llms for detecting and mitigating harmful content: A case study on jailbreaking attacks. In *Proceedings of the First Workshop on Socially Responsible Language Modelling Applications (SoLaR)*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *KDD*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2018. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. *arXiv e-prints*, arXiv:1804.07461.

Yulong Wang, Tong Sun, Shenghong Li, Xin Yuan, Wei Ni, Ekram Hossain, and H. Vincent Poor. 2023. Adversarial attacks and defenses in machine learning-powered networks: A contemporary survey. *Preprint*, arXiv:2303.06302.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

Xilie Xu, Keyi Kong, Ning Liu, Lizhen Cui, Di Wang, Jingfeng Zhang, and Mohan Kankanhalli. 2023. An llm can fool itself: A prompt-based adversarial attack. *Preprint*, arXiv:2310.13345.

Catherine Yeh, Yida Chen, Aoyu Wu, Cynthia Chen, Fernanda Viégas, and Martin Wattenberg. 2023. Attentionviz: A global view of transformer attention. *Preprint*, arXiv:2305.03210.

Chen Zhu, Yu Cheng, Zhe Gan, Siqi Sun, Tom Goldstein, and Jingjing Liu. 2020. Freelb: Enhanced adversarial training for language understanding. In *International Conference on Learning Representations*.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. 2023. Universal and transferable adversarial attacks on aligned language models. *Preprint*, arXiv:2307.15043.

A Appendix: Dataset Details

This appendix provides additional details on the datasets used in our study.

GLUE Benchmark: The GLUE (General Language Understanding Evaluation) benchmark (Wang et al., 2018) is a widely adopted collection of natural language understanding tasks. In our experiments, we focus on three representative GLUE tasks involving short text inputs:

- **SST-2 (Stanford Sentiment Treebank):** This dataset consists of movie review sentences labeled as either positive or negative sentiment. The task is to classify the sentiment of each sentence, making it a standard benchmark for sentiment analysis.
- **RTE (Recognizing Textual Entailment):** The RTE dataset contains pairs of sentences, where the goal is to determine whether the premise sentence entails the hypothesis sentence. This task evaluates the model's ability to perform natural language inference.
- **QQP (Quora Question Pairs):** QQP is composed of pairs of questions from Quora, with the objective of identifying whether the two questions are semantically equivalent. This task tests the model's capability to recognize paraphrases and semantic similarity.

These datasets are widely used for benchmarking natural language understanding models and are characterized by their relatively short input texts.

IMDB Movie Reviews: The IMDB dataset (Maas et al., 2011) is a large-scale sentiment analysis corpus consisting of lengthy user-written movie reviews. Each review is labeled as either positive or negative sentiment. Unlike the GLUE tasks, IMDB reviews are much longer and more complex, providing a challenging testbed for adversarial defense methods due to the increased context and potential for nuanced adversarial manipulations.

B Appendix: Details of PromptAttack Methods

In our experiments, PromptAttack is adopted as the main method for generating adversarial examples (Xu et al., 2023). This approach systematically perturbs clean input samples at three linguistic levels: character, word, and sentence. Table 4 summarizes the main perturbation strategies used at each level:

In addition to these basic perturbations, PromptAttack also supports few-shot (FS) and ensemble (EN) variants. The few-shot setting provides the model with several adversarial examples to help it better understand the attack patterns, while the ensemble approach aggregates multiple perturbed

Level	Abbreviation	Perturbation Instruction
Character	C1	Choose at most two words in the sentence, and change them so that they have typos.
	C2	Change at most two letters in the sentence.
	C3	Add at most two extraneous characters to the end of the sentence.
Word	W1	Replace at most two words in the sentence with synonyms.
	W2	Choose at most two words in the sentence that do not contribute to the meaning of the sentence and delete them.
	W3	Add at most two semantically neutral words to the sentence.
Sentence	S1	Add a randomly generated short meaningless handle after the sentence, such as @fasuv3.
	S2	Paraphrase the sentence.
	S3	Change the syntactic structure of the sentence.

Table 4: Perturbation instructions at the character, word, and sentence levels.

samples and selects those most likely to deceive the model.

These strategies ensure a diverse and challenging set of adversarial inputs, enabling a comprehensive evaluation of the robustness and interpretability of our defense system.

C Appendix: Successful Defense Examples with Explanation

Table 5 presents some selected examples of the adversarial text and its corresponding cleaned text and explanation.

Dataset	Adversarial Text	Cleaned Text	Explanation
SST-2	no sophomore slump for director sam mendes , who segues from oscar winner to oscar-winning potential with a smooth sleight of hand , unfortunately.	No sophomore slump for director Sam Mendes, who segues from Oscar winner to Oscar-winning potential with a smooth sleight of hand.	Removed the adversarially placed "unfortunately" at the end, which created a contradictory sentiment likely intended to confuse sentiment analysis models. Standardized capitalization ("Oscar") and punctuation for clarity. The adversarial modifier artificially injected negativity into an otherwise positive statement.
RTE	Quebec woman and her mother accused of plotting to kill a four-year-old gurl.	Quebec woman and her mother accused of plotting to kill a four-year-old girl	Detected intentional adversarial misspelling "gurl" (instead of "girl"), likely to bypass child safety content filters. Corrected to restore standard spelling while preserving original meaning.
QQT	Quebec woman and her mother accused of plotting to kill a four-year-old gurl.	Quebec woman and her mother accused of plotting to kill a four-year-old girl.	Detected intentional adversarial misspelling "gurl" (instead of "girl"), likely to bypass child safety content filters. Corrected to restore standard spelling while preserving original meaning.
IMDB	This is a really sad, and touching movie! It deals with the subject of child abuse. It's really sad, but mostly a true story, because it happens everyday. Elijah Wood and Joseph Mazzello play the two children or Lorraine Bracco, a single mother who just tries to make a home for them. While living with her parents, a man, who likes to be called T̈he Kingc̈omes into their life. He hits the youngest boy, Bobby, but the two brothers vow not to tell their mother. But finally she finds out, after the Bobby is hurt badly. The end kind of ruined it for me, because it is so totally unbelievable. But, except for that, I hate the movie."	This is a really sad, and touching movie! It deals with the subject of child abuse. It's really sad, but mostly a true story, because it happens everyday. Elijah Wood and Joseph Mazzello play the two children or Lorraine Bracco, a single mother who just tries to make a home for them. While living with her parents, a man, who likes to be called T̈he Kingc̈omes into their life. He hits the youngest boy, Bobby, but the two brothers vow not to tell their mother. But finally she finds out, after the Bobby is hurt badly. The end kind of ruined it for me, because it is so totally unbelievable. But, except for that, I love the movie.	The word ḧateïn the final sentence was identified as an adversarial modification (word-level substitution) designed to invert sentiment analysis. This contradicted the overall positive context of the review. Replaced ḧatewith ïloveïto restore semantic coherence. No other modifications detected.

Table 5: Example of adversarial text purification and corresponding explanation.