Precise Language Deception: XAI Driven Targeted Adversarial Examples with Restricted Knowledge

Mateusz Gniewkowski[1 [0000−0002−0620−8123]],
Paweł Walkowiak[1 [0009−0008−0381−9202]],
Marek Klonowski[1 [0000−0002−3141−8712]], and
Tomasz Walkowiak[1 [0000−0002−7749−4251]]

Wrocław University of Science and Technology
mateusz.gniewkowski@pwr.edu.pl

Abstract. In this paper, we propose a novel approach for crafting targeted adversarial examples (attacks) using explainable artificial intelligence (XAI) techniques. Our method leverages XAI to identify key input elements that, when altered, can mislead NLP models, such as BERT and large language models (LLMs), into producing specific incorrect outputs. We demonstrate the effectiveness of our targeted attacks across a range of NLP tasks and models, even in scenarios where internal model access is restricted. Our approach is particularly effective in zero-shot learning settings, underscoring its adaptability and transferability to both traditional and conversational AI systems. In addition, we outline mitigation strategies, demonstrating that adversarial training and fine-tuning can enhance model defenses against such attacks. Although our work highlights the vulnerabilities of LLMs and BERT models to adversarial manipulation, it also lays the groundwork for developing more robust models, advancing the dual goal of understanding and securing black-box NLP systems. Through targeted adversarial examples and SHAP-based techniques, we not only expose the weaknesses of existing models but also propose strategies to enhance AI's resilience to deceptive linguistic input.

Keywords: LLM · BERT · Adversarial Example · XAI

1 Introduction

Adversarial Examples (AE), known also as adversarial attacks, are considered to be one of the most important obstacles to the further development of advanced deep learning methods and the implementation of trustworthy Artificial Intelligence (AI). Although tens of thousands of papers have been presented on this topic over the last decade, many AE issues in the field of Natural Language Processing (NLP) still seem underinvestigated. In this paper, we show that by using relatively simple methods of Explainable Artificial Intelligence (XAI), we are able not only to confuse the language model by constructing a wide class of various AEs but also to conduct a **targeted attack**. We can force the system to provide a pointed, specified (incorrect) answer. We refine attack strategies to control output classes across various NLP tasks. Beyond traditional

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97626-1_4

transformer-based classifiers (BERT [6]), we also explore vulnerabilities in Large Language Models (LLMs) including ChatGPT and OpenChat [19], focusing on zero-shot learning scenarios.

Our main contributions are the following. We demonstrate a novel targeted attack method that requires only a small number of examples to mislead the model. The methods we propose are effective even with **restricted knowledge** of the model. That is, we do not require access to the model weights, making our approach applicable to models accessible through APIs, such as ChatGPT. We also examine mitigation strategies for BERT model, showing that standard fine-tuning and adversarial training can effectively defend against such attacks. Our findings contribute to both understanding and improving the security of black-box NLP models. In general, our result can be seen as an argument for the weakness of language models (in particular LLMs) facing AE. On the other hand, it seems that our results can also contribute to the building of more robust language models.

2 Our Approach and Related Work

Our work concerns the security of language models (esp. LLMs), explainable artificial intelligence (XAI), and so-called adversarial examples (AE). Each of these three areas has many thousands of relevant works and has already spawned numerous meta-surveys, so it is difficult to even list most important related works. Due to space limitations, in the following, we mention only the most important works and some results closest to our findings presented in this paper.

AE are carefully modified inputs to AI models that cause their incorrect/dangerous responses. In the current form they have been introduced in the seminal paper [16]. However, similar concepts have been explored earlier (e.g. [2]). Since AEs are considered to be one of the most serious threats to building trustworthy AI systems, tens of thousands of papers have been written on them in recent years. Most of them can be found in a constantly updated list [5] currently containing several thousand works. Nevertheless, few works have been written on AE for LLM and language models in general. The work [10] from a few years is the first paper to present AE in LLMs. An interesting approach to crafting AEs was proposed in [17]. This method allows generating a much richer class of AEs (compared to previous algorithms) in a semi-automatic model (with a human in the loop). An overview of security threats in LLMs, with a particular focus on AEs from a different perspective, can be found in [11]. Several works have shown that AEs are one of the main reasons for limiting LLMs in a number of applications where reliability is critical (i.e. [1]).

One of the key ideas behind the targeted attacks presented in this paper is to leverage XAI methods to generate effective adversarial examples. In our approach, explainability techniques are used to identify the most crucial elements of the input that influence the system's behavior. These elements are then modified in various (potentially subtle) ways to achieve the desired outcome with a sufficiently high probability in practice. To our knowledge, the connection between XAI and adversarial examples was first identified in [7], where the authors hypothesized *a deep relationship between model explainability and adversarial examples*. In [9], the authors demonstrated that the widely

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97626-1_4

used game-theoretic SHAP method [13] can be used to mitigate adversarial attacks. On the other hand, the work [22] shows that AEs can be used to effectively manipulate SHAP scores and extract sensitive data. Similar results for manipulating LIME-based explanations ([15]) of text classification can be found in [4].

In the context of LLMs, significant research has focused on "jailbreaking" ([20]), i.e., techniques for extracting restricted content from a model, such as instructions for potentially harmful actions public safety. The study in [23] shows that jailbreak attacks can be effectively automated and, more surprisingly, exhibit a degree of transferability between models. Other important works include [21], where general mechanisms of protection against jailbreak attacks are considered. Although jailbreak can be seen as a form of adversarial attack, it remains unclear how these methods could be applied to the LLM classification problem explored in our work. Recent efforts have specifically targeted adversarial prompt generation to induce harmful behaviors in LLMs. Much of this research highlights potential risks or proposes mitigation strategies, such as output filtering ([18]) or modifying training data to reduce vulnerabilities ([12]). Only a few studies have attempted to explain the underlying mechanisms that enable these attacks. A notable finding is presented in [20], where the authors identify competing objectives and mismatched generalization as key factors contributing to the existence of adversarial examples in LLMs.

Our work is most closely related to [8], which explores methods to effectively find adversarial examples using SHAP functions. In particular, in the current paper we also make extensive use of SHAP. However, the key distinction is that our focus is on targeted attacks, specifically constructing adversarial examples where the incorrect output follows predefined properties (for example, ensuring that an element from class A is always misclassified as B). Naturally, this constraint limits the generation of adversarial examples and requires a different methodological approach. Moreover, we have implemented the proposed methods in state-of-the-art NLP solutions, specifically LLMs.

3 Methods

The goal of the proposed method is to execute targeted attacks, forcing the AI system to produce a specific, predefined (incorrect) response. To achieve this, we utilize differentiated ranking lists to steer the attack toward designated classes. By manipulating importance-based rankings (obtained by the SHAP method), we can precisely influence the attack trajectory, enhancing control over the final output of the model. This refined approach enables the generation of more precise adversarial examples, making it particularly effective for tasks that require class-specific misclassifications.

Let $R_A = \{(w, S_A(w))\}$ and $R_B = \{(w, S_B(w))\}$ be the ranked lists of tokens w with their corresponding SHAP importance scores $S_A(w)$ and $S_B(w)$ for classes A and B, respectively. To execute a targeted adversarial attack aimed at shifting the prediction from class A to class B, we compute the differential importance score for a token w as $\Delta S(w) = S_A(w) - S_B(w)$. The resulting list is defined as $R_{A\to B} = ((w, \Delta S(w)) \mid w \in R_A \cup R_B)$ sorted in descending order with respect to $\Delta S(w)$. This ranking highlights tokens that are highly influential for class A while minimally supporting class B, making them ideal candidates for modification. Intuitively, if $S_A(w) >= S_B(w)$, the

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97626-1_4

token w supports class A and is prioritized for alteration. If $S_B(w) > S_A(w)$, modifying w is less favorable as it already aligns with the class B. Using tokens with the highest positive $\Delta S(w)$, the attack effectively reduces the influence of class A while steering the model toward class B.

3.1 Computing Importance for LLM

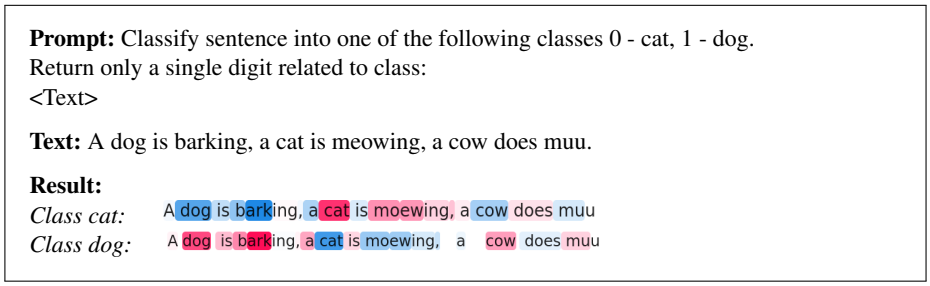

Fig. 1: Classification example. SHAP results for GPT-4o-mini, shows the local importance of each token for the respective classes. Red indicates higher importance, while blue represents lower importance.

To compute token-level importance scores for LLMs, particularly those accessible via API, we employ a modified SHAP-based approach tailored for black-box settings. We begin with a text sample and a corresponding prompt crafted to guide the LLM's response. Consider the example from Figure 1.

To compute SHAP values for individual tokens, we modify only the input text (using any tokenizer) while keeping the prompt constant. Each modified version is then sent to the remote LLM using OpenAI API. The model needs to return two things: a list of tokens and log probabilities (`logprobs`) for its generated outputs. These log probabilities can be converted into standard probabilities, allowing us to assess how changes in the text influence the likelihood of each class. In binary classification tasks, we concentrate on the top-n = 2 most relevant tokens. This approach enables us to calculate Shapley values by observing how the removal or alteration of specific tokens affects the model's predictions. If the returned token does not correspond to any predefined class label, it is simply excluded from the explanation process.

3.2 Attack Methodology

To investigate the behavior of large language models (LLMs) under adversarial conditions, we adapted the attack methodology from [8], incorporating explainable AI (XAI) techniques to test the models. The original approach leveraged SHAPley value-based global explanations (computed on a separate dataset split) to identify the most influential parts of the victim text for classification. Additionally, we employed the following

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97626-1_4

text-disrupting methods:

– **WordNet-XAI**, which replaces selected words in the text with their synonyms retrieved from plWordNet. Candidate synonyms are further filtered based on cosine similarity, computed using FastText [3] word embeddings. Only candidates with a similarity score that exceeds the threshold ϵ_w (set to 65%) are considered for substitution.

– **WordNet-XAI-CharDiscard** (WordNet-XAI-ChD), which introduces perturbations by randomly deleting letters from a word w_i with a given probability p (set to 0.4%).

The candidate attack sentences, after applying the substitutions, are filtered based on a cosine similarity threshold ϵ (set to 95%), using sentence embeddings generated by the Sentence Transformer [14]. For LLM evaluation, a zero-shot prompting approach was chosen. Furthermore, the original class labels were represented as digits (see Figure 1), allowing the extraction of SHAP values from the models. The results presented were obtained using the same algorithmic parameters as those in [8].

4 Results

Table 1: The characteristics of the dataset used: language of the dataset, number of labels, sizes of dataset parts, average length of the texts in words. All datasets and their parts are balanced.

Dataset	Lang	No. of classes	Train size	Test size	XAI size	Aver. len
AG_News	EN	4	120,000	6,840	100	38
Wiki_PL	PL	4	801	358	40	186

In reported experiments, we tested the BERT model, which we trained for straightforward text classification. Furthermore, we compared its performance with classification results from the OpenChat and GPT-4o-mini models using a zero-shot learning approach. The LLM prompts followed the structure illustrated in Figure 2. This figure demonstrates that even minor modifications to a sample can sometimes be enough to mislead these classifiers. Moreover, it highlights that the constraints imposed on the generation methods ensure that the modified samples remain highly similar to the original ones.

Table 1 provides a summary of the datasets used in our study, namely AG_News[1] and Wiki_PL[2]. Each dataset contains four distinct classes and has been divided into

[1] http://groups.di.unipi.it/~gulli/AG_corpus_of_news_articles.html
[2] http://hdl.handle.net/11321/216

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97626-1_4

Prompt: Classify sentence into one of the following classes:
0: Articles related to international events, global news, and world affairs. This category includes stories on political events, international conflicts, diplomacy, and relations between countries.
1: This category includes news related to sports events, athletes, match outcomes, developments across various sports, as well as updates on teams and sporting events.
2: Articles in this category cover financial, economic, and market-related news. It includes content on companies, market trends, investments, financial matters, and economic topics.
3: Articles focused on science and technology news. Topics include new technologies, scientific research, discoveries, and trends or innovations in fields such as medicine, IT, computers, and beyond.
Return only a single digit related to class:
<Text>

Text: Panel Urges N.Y. to Pay $14 Billion More for City Schools Court appointed referees recommended the state → commonwealth pay an additional $14 billion over four years to improve New York City schools.

Result: 0 → 2

Fig. 2: Example of successful directed attacks for AG_News classification (Test set) using GPT-4o-mini model, achieved by altering just a single word in the sentence. The green indicates original form, the red one change after the attack. The importance score of the word 'state' is 0.021 for class "0" and -0.019 for class "2". The difference between these values (0.04) makes it the best candidate for replacement in class "0" to class "2" attack scenario.)

Table 2: Classification accuracy (ACC) for Wiki_PL and AG_News datasets. All datasets are balanced across classes to ensure fair evaluation. The BERT model is fine-tuned on the respective training sets, while OpenChat and GPT-4o-mini perform zero-shot classification using prompts that include natural language descriptions of each class.

Dataset	BERT ACC [%]	OpenChat ACC [%]	GPT-4o-mini ACC [%]
Wiki_PL	99.00	99.00	99.00
AG_News	95.00	85.00	85.00

three subsets. The first subset is a training set, used primarily for fine-tuning the BERT and OpenChat models. The second is a test set, which serves both for evaluating classifier accuracy and for executing adversarial attacks. Finally, the third subset, referred to as the XAI set, comprises a smaller sample of the data and is used to generate word-level importance rankings for modification purposes. The complete data processing pipeline is depicted in Figure 3.

All datasets are approximately balanced, with only minor deviations in the number of examples per class. Table 2 presents preliminary classification results for each of the

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97626-1_4

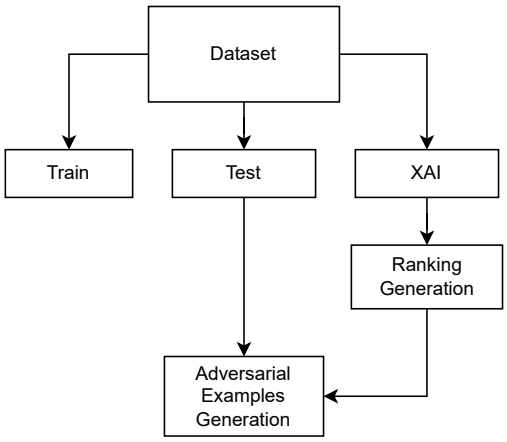

Fig. 3: General data flow used in the experimental setup. The original dataset is split into three parts: training data for fine-tuning BERT classifier, test data for evaluation (used also for OpenChat and GPT-4o-mini), and data used for generating explanations via an XAI method. The XAI component produces feature importance rankings, which guide the *Adversarial Examples Generation* module. This module uses the ranked features to create targeted perturbations on the test set, resulting in adversarial examples used for robustness evaluation of the models.

examined methods. As can be observed, all approaches achieve satisfactory accuracy scores, confirming their general effectiveness in tackling the classification task.

We began with a standard adversarial attack, where the goal was to alter the classification to any other class. This is illustrated in Table 3. In particular, attention should be paid to the results of methods applied to the XAI subset. These results demonstrate how an attack would perform on samples where we had prior information (the ranking remains global, averaged across the samples in the XAI subset). It is likely that slightly better results could be achieved by using the importance scores computed for individual samples. Although the data sets are not very large, they reveal that at least some of the samples are vulnerable to the attack (1% for the AG_News data set and 2.5% for Wiki_PL). Using the same rankings, we performed attacks on the test subsets. The results show a high effectiveness of the attacks, with the most successful method being the simple removal of characters from individual (most important) words. A slightly more sophisticated approach, replacing words with synonyms, also yields satisfactory results. An interesting observation is that methods based on LLMs seem to be effectively resistant to attacks for the Polish language. However, this is associated with the very high confidence of these models in basic classification tasks. Tables 4 and 5 illustrate the effectiveness of both targeted and untargeted attacks. Comparing these values allows

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97626-1_4

Table 3: Results of adversarial attacks. The **WordNet-XAI** method replaces important words with their synonyms, while **WordNet-XAI-ChD** introduces noise by randomly deleting characters from relevant words. BERT results are taken from [8]. The OpenChat and GPT-4o-mini results are based on zero-shot classification. An attack is considered successful if the model's prediction changes from a true positive to any other class.

Attack type	Dataset	Part	Success % BERT	Success % OpenChat	Success % GPT-4o-mini
WordNet-XAI	AG_News	XAI	4.00	10.00	11.00
WordNet-XAI-ChD	AG_News	XAI	2.00	10.00	13.00
WordNet-XAI	AG_News	Test	2.27	7.44	7.65
WordNet-XAI-ChD	AG_News	Test	2.99	7.85	7.69
WordNet-XAI	Wiki_PL	XAI	5.00	7.50	5.00
WordNet-XAI-ChD	Wiki_PL	XAI	20.00	12.50	12.50
WordNet-XAI	Wiki_PL	Test	1.68	0.00	0.28
WordNet-XAI-Chd	Wiki_PL	Test	10.89	0.28	0.84

us to evaluate how well the proposed method directs the attacks. The only case where the method did not perform as expected was AG_News A → B, likely due to the strong separation between these classes. For the Wiki_PL dataset, we expected that targeted attacks, designed to highlight key modifications, would be more effective in misleading the model. However, our results indicate that the small changes introduced in the samples were insufficient, as successfully deceiving the model would require significantly larger alterations, making the modifications more noticeable to the reader. Additionally, it is important to consider that LLM models have probably encountered articles from this dataset in multiple languages, given that the data originate from Wikipedia.

Fine-tuning models on adversarially altered samples is a promising strategy for mitigating the impact of attacks. We successfully fine-tuned BERT, which led to improved results, although its performance declined slightly when using the character removal method. For OpenChat, Supervised Fine-Tuning (SFT) can be employed, and when executed properly, it should yield better results by adapting the model to a specific task. However, training such models is highly resource intensive, and performing SFT correctly presents challenges, particularly since task-specific specialization can degrade performance in other areas. A more efficient and cost-effective approach may involve detecting modifications before inputting the data into the LLM, thereby reducing the need for extensive fine-tuning. The classification stability of BERT remained unchanged after fine-tuning on the perturbed Wiki_PL dataset. Regarding adversarial attacks, the model showed increased resistance to WordNet-XAI attacks (0.28% success rate), likely due to its exposure to synonym substitutions during training. However, it struggled more with character removal attacks (7.54% success rate). This is because once a word is altered, it undergoes different tokenization and the modified tokens in the test data may not align with those seen during training.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97626-1_4

Table 4: Success rates of directed adversarial attacks on a BERT-based classifier. The **WordNet-XAI** method generates adversarial examples by replacing semantically important words with their synonyms using the WordNet. The **WordNet-XAI-ChD** method applies additional perturbations by randomly deleting characters from those important words. The notation $A \rightarrow B$ indicates an attempt to intentionally change a sample originally and correctly classified as class A into being misclassified as class B. Mean success rates are reported for each scenario.

Method	Attack type	Dataset	A → B	A → C	A → D	Mean
Targeted	WordNet-XAI	AG_News	2.28	2.34	2.4	2.34
	WordNet-XAI-ChD		2.75	3.45	2.75	2.98
Untargeted	WordNet-XAI		0.94	0.76	0.53	0.74
	WordNet-XAI-ChD		1.46	1.58	0.7	1.25
Targeted	WordNet-XAI	Wiki_PL	4.44	2.22	2.22	2.96
	WordNet-XAI-ChD		6.67	6.67	8.89	7.41
Untargeted	WordNet-XAI		0.00	0.00	1.11	0.37
	WordNet-XAI-ChD		1.11	0.00	1.11	0.74

Table 5: Success rates of directed adversarial attacks on large language models (LLMs): OpenChat and GPT-4o-mini. The **WordNet-XAI** method generates adversarial examples by replacing semantically important words with their synonyms using the WordNet. The **WordNet-XAI-ChD** method applies additional perturbations by randomly deleting characters from those important words. The notation $A \rightarrow B$ indicates an attempt to intentionally change a sample originally and correctly classified as class A into being misclassified as class B.

Method	Attack type	Dataset	Openchat				GPT-4o-mini			
			A → B	A → C	A → D	Mean	A → B	A → C	A → D	Mean
Targeted	WordNet-XAI	AG_News	2.75	2.57	2.51	2.61	3.10	3.04	2.98	3.04
	WordNet-XAI-ChD		3.10	3.10	3.27	3.16	2.92	3.39	3.27	3.19
Untargeted	WordNet-XAI		3.74	2.87	1.35	2.65	3.16	3.22	1.17	2.52
	WordNet-XAI-ChD		4.33	3.04	1.64	3.00	4.09	3.10	1.46	2.88
Targeted	WordNet-XAI	Wiki_PL	0.00	1.11	0.00	0.37	0.00	0.00	0.00	0.00
	WordNet-XAI-ChD		1.11	0.00	1.11	0.74	0.00	0.00	0.00	0.00
Untargeted	WordNet-XAI		0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	WordNet-XAI-ChD		0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

5 Conclusion

In this study, we demonstrate that insights derived from explainable artificial intelligence (XAI) can be leveraged to craft targeted adversarial attacks on natural language processing models operating under black-box constraints. By observing the attribution scores returned by local explanation methods, we are able to focus on the input frag-

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97626-1_4

ments that matter most to the classifier and subsequently adjust them with minimal effort.

Our experiments show that inconspicuous edits can alter predictions even for LLMs such as GPT-4o-mini. Furthermore, a brute-force search can produce up to 500 times more candidates to be tested, and an XAI-driven strategy achieves similar success with far fewer queries. Targeted attacks are shown to be more effective than untargeted ones, highlighting the practical value of explanation-guided evaluation in real-world settings.

The results indicate that models such as BERT and OpenChat, although demonstrating promising performance in classification tasks, remain susceptible to adversarial attacks. In the context of targeted attacks, we found that all the analyzed models can be manipulated with subtle modifications. The effectiveness of these attacks varies; the removal of random characters has proven to be more effective, whereas other methods, such as synonym substitution, demonstrate greater robustness. This suggests that while targeted attacks are possible, they depend heavily on specific circumstances. Furthermore, using local explanation methods, we can identify key features that contribute to these vulnerabilities and potentially reduce the impact of such attacks.

Importantly, local explanation techniques expose tokens that contribute most strongly to model decisions. Knowledge of these fragile anchors can be used to both intensify attacks and design defenses. Future research should therefore explore mitigation strategies such as adversarial training, input sanitization, or confidence calibration that reduce vulnerability without eroding predictive accuracy.

Acknowledgments

Financed by: (1) CLARIN ERIC (2024–2026), funded by the Polish Minister of Science (agreement no. 2024/WK/01); (2) CLARIN-PL, the European Regional Development Fund, FENG program (FENG.02.04-IP.040004/24); (3) statutory funds of the Department of Artificial Intelligence, Wroclaw Tech; (4) the EU project 'DARIAH-PL', under investment A2.4.1 of the National Recovery and Resilience Plan. (5) the European Regional Development Fund as part of the 2014-2020 Smart Growth Operational Program (POIR.04.02.00-00C002/19); and (6) by the National Science Center, Poland, grant number 2018/29/B/HS2/02919.

References

1. Albrecht, J., Kitanidis, E., Fetterman, A.J.: Despite "super-human" performance, current llms are unsuited for decisions about ethics and safety (2022), https://arxiv.org/abs/2212.06295
2. Biggio, B., et al.: Evasion attacks against machine learning at test time. In: Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2013, Prague, Czech Republic, September 23-27, 2013, Proceedings, Part III. Lecture Notes in Computer Science, vol. 8190, pp. 387–402. Springer (2013). https://doi.org/10.1007/978-3-642-40994-3_25
3. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. Transactions of the association for computational linguistics **5**, 135–146 (2017)

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97626-1_4

4. Burger, C., Chen, L., Le, T.: Are your explanations reliable? investigating the stability of lime in explaining text classifiers by marrying xai and adversarial attack (2023), https://arxiv.org/abs/2305.12351
5. Carlini, N.: A complete list of all (arxiv) adversarial example papers (2019-2025), https://nicholas.carlini.com/writing/2019/all-adversarial-example-papers.html
6. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Burstein, J., Doran, C., Solorio, T. (eds.) Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019). https://doi.org/10.18653/v1/N19-1423
7. Fidel, G., Bitton, R., Shabtai, A.: When explainability meets adversarial learning: Detecting adversarial examples using SHAP signatures. In: 2020 International Joint Conference on Neural Networks, IJCNN 2020, Glasgow, United Kingdom, July 19-24, 2020. pp. 1–8. IEEE (2020), https://doi.org/10.1109/IJCNN48605.2020.9207637
8. Gniewkowski, M., et al.: Do not trust me: Explainability against text classification. In: ECAI 2023 - 26th European Conference on Artificial Intelligence, September 30 - October 4, 2023, Kraków, Poland - Including 12th Conference on Prestigious Applications of Intelligent Systems (PAIS 2023). Frontiers in Artificial Intelligence and Applications, vol. 372, pp. 875–882. IOS Press (2023). https://doi.org/10.3233/FAIA230356
9. Hickling, T., Aouf, N., Spencer, P.: Robust adversarial attacks detection based on explainable deep reinforcement learning for uav guidance and planning. IEEE Transactions on Intelligent Vehicles **8**(10), 4381–4394 (2023)
10. Jia, R., Liang, P.: Adversarial examples for evaluating reading comprehension systems. CoRR **abs/1707.07328** (2017), http://arxiv.org/abs/1707.07328
11. Jia, X., Huang, Y., Liu, Y., Tan, P.Y., Yau, W.K., Mak, M.T., Sim, X.M., Ng, W.S., Ng, S.K., Liu, H., et al.: Global challenge for safe and secure llms track 1. arXiv preprint arXiv:2411.14502 (2024)
12. Lukas, N., et al.: Analyzing leakage of personally identifiable information in language models. In: 44th IEEE Symposium on Security and Privacy, SP 2023, San Francisco, CA, USA, May 21-25, 2023. pp. 346–363. IEEE (2023), https://doi.org/10.1109/SP46215.2023.10179300
13. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. Advances in neural information processing systems **30** (2017)
14. Reimers, N., Gurevych, I.: Sentence-bert: Sentence embeddings using siamese bert-networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics (11 2019), http://arxiv.org/abs/1908.10084
15. Ribeiro, M.T., Singh, S., Guestrin, C.: "why should I trust you?": Explaining the predictions of any classifier. In: Krishnapuram, B., Shah, M., Smola, A.J., Aggarwal, C.C., Shen, D., Rastogi, R. (eds.) Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016. pp. 1135–1144. ACM (2016). https://doi.org/10.1145/2939672.2939778, https://doi.org/10.1145/2939672.2939778
16. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R.: Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199 (2013)
17. Wallace, E., Rodriguez, P., Feng, S., Yamada, I., Boyd-Graber, J.L.: Trick me if you can: Human-in-the-loop generation of adversarial question answering examples. Trans. Assoc. Comput. Linguistics **7**, 387–401 (2019). https://doi.org/10.1162/TACL_A_00279, https://doi.org/10.1162/tacl_a_00279

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97626-1_4

18. Wang, B., et al.: Exploring the limits of domain-adaptive training for detoxifying large-scale language models. In: Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022 (2022), http://papers.nips.cc/paper_files/paper/2022/hash/e8c20cafe841cba3e31a17488dc9c3f1-Abstract-Conference.html
19. Wang, G., Cheng, S., Zhan, X., Li, X., Song, S., Liu, Y.: OpenChat: Advancing open-source language models with mixed-quality data. In: The Twelfth International Conference on Learning Representations (2024), https://openreview.net/forum?id=AOJyfhWYHf
20. Wei, A., Haghtalab, N., Steinhardt, J.: Jailbroken: How does LLM safety training fail? In: Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023 (2023), http://papers.nips.cc/paper_files/paper/2023/hash/fd6613131889a4b656206c50a8bd7790-Abstract-Conference.html
21. Wei, A., Haghtalab, N., Steinhardt, J.: Jailbroken: How does llm safety training fail? (2023), https://arxiv.org/abs/2307.02483
22. Yeghiazaryan, M., et al.: Texture- and shape-based adversarial attacks for vehicle detection in synthetic overhead imagery (2024), https://arxiv.org/abs/2412.16358
23. Zou, A., Wang, Z., Carlini, N., Nasr, M., Kolter, J.Z., Fredrikson, M.: Universal and transferable adversarial attacks on aligned language models. arXiv preprint arXiv:2307.15043 (2023)

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97626-1_4