Precision vs. Perturbation: Robustness Analysis of Synonym Attacks in Ukrainian NLP

Volodymyr Mudryi
Ukrainian Catholic University,
Lviv, Ukraine
mudryi.pn@ucu.edu.ua

Oleksii Ignatenko
Ukrainian Catholic University,
Lviv, Ukraine
o.ignatenko@ucu.edu.ua

Abstract

Synonym-based adversarial tests reveal fragile word patterns that accuracy metrics overlook, while virtually no such diagnostics exist for Ukrainian, a morphologically rich and low-resource language. We present the first systematic robustness evaluation under synonym substitution in Ukrainian. Adapting TEXTFOOLER and BERT-ATTACK to Ukrainian, we (i) adjust a 15000-entry synonym dictionary to match proper word forms; (ii) integrate similarity filters; (iii) adapt masked-LM search so it generates only valid inflected words. Across three text classification datasets (reviews, news headlines, social-media manipulation) and three transformer models (Ukr-RoBERTa, XLM-RoBERTa, SBERT), single-word swaps reduce accuracy by up to 12.6, while multi-step attacks degrade performance by as much as 40.27 with around 112 model queries. A few-shot transfer test shows GPT-4o, a state-of-the-art multilingual LLM, still suffers 6.9–15.0 drops on the same adversarial samples. Our results underscore the need for sense-aware, morphology-constrained synonym resources and provide a reproducible benchmark for future robustness research in Ukrainian NLP.

1 Introduction

Natural Language Processing (NLP) has undergone a rapid transformation over the past decade. Early systems were built on rule-based heuristics and simple statistical models, where model behavior was largely transparent and evaluation relied on basic accuracy or coverage metrics. As these models lacked generalization power, error patterns were relatively easy to identify and correct.

With the introduction of neural networks—particularly transformer architectures (Vaswani et al., 2017) trained on large-scale text corpora—modern NLP systems have achieved remarkable performance across diverse tasks such as machine translation, sentiment analysis, and question answering. These models can capture complex semantic and syntactic patterns, often surpassing human-level benchmarks. However, the internal behavior of these models is difficult to interpret, and traditional metrics such as accuracy or F1 score often overestimate performance and fail to reveal model vulnerabilities, motivating the need for more comprehensive evaluation methods (Ribeiro et al., 2020).

This has motivated the development of stress tests and behavioral diagnostics to probe how models behave under controlled perturbations. One such technique is the synonym substitution attack, which evaluates whether a model's prediction is sensitive to small, meaning-preserving changes in the input text. These attacks are appealing because they preserve grammaticality and semantics from a human perspective while often revealing inconsistent model behavior.

While synonym substitution attacks have been widely studied in English, their applicability to low-resource and morphologically rich languages remains underexplored. Ukrainian, for example, poses additional challenges: it exhibits complex inflectional morphology, has limited lexical resources, and lacks large-scale evaluation benchmarks for adversarial robustness. As a result, it is unclear how well multilingual or Ukrainian-specific language models perform under such perturbations.

To our knowledge, this work presents the first systematic evaluation of synonym substitution attacks in Ukrainian. We explore whether current models—both monolingual and multilingual—are robust to these types of perturbations, and we assess whether the vulnerability persists in modern LLMs.

Contributions. Our main contributions are:

- We implement and adapt two state-of-the-art

Proceedings of the Fourth Ukrainian Natural Language Processing Workshop (UNLP 2025), pages 131–146

adversarial attack frameworks—TextFooler and BERT-Attack—for the Ukrainian language, addressing issues of morphological agreement and synonym quality.

- We evaluate the robustness of three models (Ukr-RoBERTa, XLM-RoBERTa, and SBERT) across three Ukrainian text classification datasets spanning different domains.
- We measure the transferability of attacks to a modern instruction-tuned LLM (GPT-4o), providing insights into cross-model robustness in Ukrainian.

Resources. We release the complete codebase - including dataset loaders, fine-tuning scripts, and attack pipelines—in a single public repository so that other researchers can benchmark the robustness of their own Ukrainian or multilingual models with minimal effort: `https://github.com/Mudryi/ukr-synonym-robustness`.

2 Related Work

Adversarial attacks were first explored in computer vision, where imperceptible pixel-level perturbations can drastically alter model outputs (Goodfellow et al., 2014). In contrast, NLP inputs are discrete, so crafting adversarial examples requires careful preservation of grammar and semantics.

Adversarial attacks in NLP are commonly categorized into character-level, word-level, sentence-level, and syntactic-level perturbations, depending on the granularity and linguistic structure affected.

Character-level attacks such as HotFlip use white-box gradients to identify single-character edits that maximally increase loss, demonstrating that even single letter change can mislead a classifier (Ebrahimi et al., 2018). In the black-box setting, DeepWordBug applies simple heuristics—swaps, deletions, or insertions—to high-saliency tokens, achieving substantial accuracy drops with minimal edit distance (Gao et al., 2018).

Word-level synonym substitution attacks replace important words with context-preserving alternatives. Early genetic algorithm approaches by (Alzantot et al., 2019) and the PWWS method by (Ren et al., 2019) ranked words by importance before substituting them using WordNet. TextBugger (Li et al., 2018) combined both character-level and word-level perturbations and introduced semantically similar replacements using embedding-based nearest neighbors. TextFooler later showed that a small number of carefully chosen synonym swaps can reduce BERT's accuracy by over 50% while keeping the text fluent (Jin et al., 2019). Building on this, BERT-Attack leverages masked language model infilling to generate higher-quality substitutes with fewer queries, further exposing model brittleness (Li et al., 2020).

Going beyond individual words, sentence-level testing frameworks probe models' sensitivity to diverse linguistic phenomena. (Iyyer et al., 2018) introduced Syntactically Controlled Paraphrase Networks (SCPN) to generate paraphrases under alternate parse templates, revealing that models often fail on syntactic variations despite preserved meaning. (Ribeiro et al., 2020) proposed CheckList, a behavioral testing framework that uses task-agnostic "capabilities" and targeted test suites, such as minimum functionality tests, invariance tests, and directional expectation tests, to uncover fine-grained weaknesses in NLP models beyond traditional accuracy metrics.

In a complementary line of research, several studies have shown that models often rely on unintended lexical artifacts present in the training data itself, e.g. annotation artifacts in NLI (Gururangan et al., 2018), heuristic "competency problems" (Gardner et al., 2021), and surface-cue reliance in reading-comprehension benchmarks (Ray Choudhury et al., 2022). Such lexical shortcuts further motivate synonym-substitution tests, because they imply that changing a single word can flip a prediction even outside an explicit adversarial setting.

While most of these methods target English, recent work extends robustness evaluation to low-resource and morphologically rich languages. (Alshahrani et al., 2024) adapted masked-LM synonym attacks to Arabic, finding that BERT-based classifiers can be even more vulnerable than traditional models. For Chinese, (Zhang et al., 2021) Argot framework uses homophone and look-alike character substitutions to generate readable, high-success-rate adversarial examples.

Recent work such as PromptRobust (Zhu et al., 2023) demonstrates that even advanced LLMs are sensitive to minor textual perturbations—including synonyms, typos, and rephrasings—across a range of tasks. However, such evaluations remain largely limited to high-resource languages like English. Robustness diagnostics for morphologically rich, low-resource languages such as Ukrainian are still lacking, motivating the need for adapted adversarial benchmarks.

In this work, we focus exclusively on synonym substitution attacks. We select this approach because it generates fluent, semantically-preserving perturbations that are both realistic and challenging for models to detect, offering a clear benchmark for word-level robustness while maintaining the original intent of the text.

3 Synonym Substitution Attack Formulation

Adversarial attacks in NLP aim to slightly perturb a valid input $\mathbf{x}_{\text{orig}} = [w_1, w_2, \ldots, w_n]$, where each w_i denotes a word token, to generate an adversarial counterpart $\mathbf{x}_{\text{adv}}$ such that:

$$f_{\text{human}}(\mathbf{x}_{\text{adv}}) \approx f_{\text{human}}(\mathbf{x}_{\text{orig}}),$$
$$f_{\text{model}}(\mathbf{x}_{\text{adv}}) \neq f_{\text{model}}(\mathbf{x}_{\text{orig}})$$

where f_{model} is the prediction function of the target model, and f_{human} reflects the perceived semantic meaning by a human reader. The goal is to fool the model while keeping the input interpretable and natural.

To maintain plausibility, the adversarial perturbation is constrained by a predefined budget, typically limiting the number of modified tokens:

$$\|\mathbf{x}_{\text{adv}} - \mathbf{x}_{\text{orig}}\|_0 \leq k,$$

where $\|\cdot\|_0$ denotes the number of word substitutions and k is a small constant bounding the allowable number of changes.

In the case of Synonym Substitution Attacks (SSA), the perturbation involves replacing one or more words with contextually appropriate synonyms. An adversarial example takes the form $\mathbf{x}_{\text{adv}} = [w_1, w_2', \ldots, w_n]$, where w_2' is a synonym of w_2 selected to preserve fluency and meaning. The candidate set for substitution is typically constrained by part-of-speech tags, semantic similarity, or language model likelihood.

Synonym substitution attacks (SSA) challenge models by preserving surface structure and meaning while revealing overreliance on specific tokens. Unlike character-level or noise-based attacks, synonym perturbations generate more realistic inputs, making them ideal for evaluating semantic robustness.

We define SSA as a sequence of word replacements aimed at flipping the model's prediction while not changing the overall meaning of the sequence and maintaining its grammatical and semantic validity. This involves identifying important words, selecting appropriate synonyms, and substituting them sequentially until misclassification or a stopping condition occurs. Sections 6.1–6.2 detail our adaptations of TextFooler and BERT-Attack for Ukrainian.

4 Experimental Setup

To measure how robust modern models are to Ukrainian synonyms substitutions, we need to select a dataset where we want to measure robustness and models themselves.

4.1 Datasets

Our study focuses on text classification tasks in Ukrainian. After reviewing available datasets, we selected three diverse benchmarks that vary in domain, task complexity, and language style. All datasets were randomly partitioned into training, validation, and test sets with an 80/10/10 split. summary statistics of the each dataset are provided in the Table 1.

Cross_Domain_UA_Reviews.(Kovenko, 2021) Dataset of Ukrainian-language user reviews from various online platforms, including Rozetka, Tripadvisor, and others. Each review is associated with a score from 1 to 5. We filtered the dataset to include only Ukrainian-language entries, resulting in approximately 15k samples. The dataset exhibits a slight class imbalance, with more reviews labeled as 5 (very positive) (≈71% of the filtered set), followed by score 4 (≈13%), while each of the remaining three classes accounts for ≤7%.

UA News Classification. (Ivanyuk-Skulskiy et al., 2021) This dataset is a part of the **UA-datasets** collection, which contains over 150k news articles collected from more than 20 Ukrainian news portals. Each article is labeled with one of five high-level topics: бізнес (business), новини (news), політика (politics), спорт (sports), and технології (technology). All classes are balanced. To simplify inputs and reduce text length, we use the article titles for classification. This also helps isolate the impact of synonym substitution on domain-specific keywords.

UNLP 2025: Detecting Social Media Manipulation (UNLP Workshop Organizers, 2025). This dataset, curated by Texty.org.ua for the UNLP 2025 shared task, includes 3,8k Telegram posts, manually labeled for the presence of manipulation techniques, such as appeals to fear or loaded language. Approximately 67% of the posts contain manipula-

Dataset	Task	Size	Avg. len ± std
UA Reviews	Sentiment	15k	25.1 ± 24.9
UA News	Multiclass	150k	10.7 ± 2.9
UNLP 2025	Binary	3.8k	82.6 ± 77.7

Table 1: Summary of the Ukrainian text-classification datasets used in our experiments.

tion. Since the corpus includes both Ukrainian and Russian entries, we partition the data so that the final test set contains only Ukrainian-language posts, which prevents cross-lingual artifacts during synonym substitution attacks. We simplify the original multilabel/span-detection tasks into a binary classification ("manipulative" vs "non-manipulative"), allowing us to study how synonym substitutions affect sensitivity to manipulative language, which often relies on certain phrasing.

4.2 Models

For model selection, we chose three transformer-based architectures that are widely used in the Ukrainian NLP community.

UkrRoBERTa (Radchenko, 2021) is a Ukrainian-specific version of the RoBERTa model (Liu et al., 2019) trained on a large Ukrainian corpus. It uses a SentencePiece tokenizer specifically adapted to the Ukrainian language. Due to its language-specific pretraining, we expect it to be more robust and particularly well-suited for capturing Ukrainian morphology. It is also interesting to compare how it performs against more general-purpose multilingual models.

XLM-RoBERTa-base (Conneau et al., 2019) is a multilingual version of RoBERTa pretrained on 100 languages, including Ukrainian. This model has been widely adopted for Ukrainian-language tasks and has shown strong performance across various benchmarks, making it a reliable baseline for multilingual robustness.

Sentence Transformer (paraphrase-multilingual-mpnet-base-v2) aka SBERT (Reimers and Gurevych, 2019) is a sentence-level model trained for multilingual semantic similarity tasks. While it was not originally designed for token-level prediction, it has shown strong results on classification tasks in Ukrainian. We include it to assess whether sentence-level representation learning introduces additional robustness to synonym substitution.

In our experimental setup, we fine-tune each model separately on each of the three datasets by adding a single classification head on top of the transformer encoder. All models are trained using the same set of hyperparameters, which are detailed in Appendix A. This results in a total of nine fine-tuned models (three architectures applied to three datasets), which we evaluate for robustness under synonym substitution attacks. The clean performance of these models is summarized in Table 2.

Dataset	Ukr-RBT	XLM-RBT	SBERT
UA Reviews	76.28%	77.91%	77.58%
UA News	98.83%	93.52%	93.46%
UNLP 2025	81.41%	80.1%	81.67%

Table 2: Clean test performance of each model before any adversarial attack.

5 One-Word Replacement Baseline

Before implementing full synonym substitution attacks, we first evaluate a simple baseline to assess the robustness of each model to single-word replacements. Specifically, for each dataset, we identify the 1000 most frequent words in the training corpus and extract candidate synonyms from the publicly available Ukrainian synonym dictionary synonimy.info (Synonimy.info, 2025), which provides non-commercial use.

Although the dictionary offers broad coverage, it contains some outdated or overly specific entries and includes occasional mismatches that do not reflect true synonymy in modern usage (e.g., пес – посіпака). To improve substitution quality, we apply a multi-step filtering process: we discard badly formatted or duplicated items, remove words that differ from the original only in grammatical form, and eliminate antonyms by cross-referencing with an antonym dictionary (Antonimy.info, 2025). To further expand synonym coverage for high-impact words (identified using a leave-one-out strategy; see Section 6.1). To increase coverage we manually added several examples from the official online version of the Словник синонімів української мови (Наукова думка, 1999).

Finally, A portion of the resulting synonym sets was manually reviewed to confirm whether generated replacements preserved both grammatical compatibility and original meaning.

To ensure grammatical correctness, each synonym is morphologically transformed to match the original word form using `pymorphy2`. For each

word in the top 1000 most frequent words, we generated all valid one-word replacements by substituting it with each of its synonyms (if present in a given sentence). Each original test example, therefore, produces multiple perturbed variants, each containing exactly one synonym substitution.

We apply this procedure only to test-set samples that were correctly classified by the model, ensuring that we are measuring actual robustness rather than model errors. The total number of generated examples is calculated as the number of test samples times the number of overlapping top-1000 words times the number of valid synonyms per word.

Once all replacements are generated, we group them by their original (unperturbed) sample and select the one that causes the most harmful prediction change — defined as the replacement that leads to the largest drop in the target model's predicted probability for the original class (i.e., the highest reduction in confidence or a misclassification). This one-to-one mapping allows us to evaluate worst-case single-word synonym substitution per sentence. Examples of both successful and unsuccessful substitutions—along with model predictions—are provided in Appendix B.1.

We report the model's test-set accuracy after applying the most harmful replacement to each example. Table 3 shows the relative accuracy drop for each model-dataset pair.

Dataset	Ukr-RBT	XLM-RBT	SBERT
UA Reviews	-12.63%	-9.08%	-10.56%
UNLP 2025	-5.32%	-2.83%	-7.11%
UA News	-2.69%	-5.74%	-5.22%

Table 3: Test accuracy drop under one-word synonym substitution. Each model is evaluated on perturbed inputs with a single worst-case synonym replacement.

The results demonstrate that the three models used in this paper are not robust to even single-word substitutions in Ukrainian, with performance drops ranging from 2.69% to 12.63%. This variability reveals the presence of highly impactful words and motivates the development of more targeted, multi-step synonym substitution attacks.

In subsequent experiments, we improve the one-word synonym substitution attack by introducing attacks that apply sequential substitutions, continuing until the model changes its prediction or a stopping criterion is reached.

6 Synonym Substitution Attacks

To perform more advanced synonym substitution attacks that support multiple word replacements, we adapt two widely used adversarial frameworks: TextFooler (Jin et al., 2019) and BERT-Attack (Li et al., 2020). While both methods have demonstrated strong performance in English, they cannot be applied directly to Ukrainian due to limited language resources and morphological complexity. Each method requires adaptation with respect to the availability of synonym sources and Ukrainian linguistic characteristics. In particular, we integrate a dictionary-based synonym set into the TextFooler pipeline, combined with morphological transformations to ensure grammatical correctness. The BERT-Attack method, in contrast, relies on a masked language model for synonym generation, which we adjust to work effectively with Ukrainian inputs.

These two frameworks were selected due to their complementary strengths. TextFooler offers simplicity and transparency: it requires only a synonym list, allows precise control over POS and similarity constraints, and provides interpretability by clearly identifying which words trigger model changes. BERT-Attack, on the other hand, leverages a language model to propose replacements that better fit the surrounding context. It requires no explicit synonym dictionary and tends to generate more fluent, human-like paraphrases using the model's own learned vocabulary and semantics.

For each synonym substitution attack, we report the original and adversarial accuracy, the accuracy drop, the average word change rate, and the average number of model queries per sample. **Change Rate** denotes the percentage of words modified in the input, while **Queries** indicates the number of model forward passes required to construct the adversarial example. All results are in tables 5 and 6.

6.1 TextFooler

TextFooler is one of the most widely used frameworks for synonym substitution attacks in English. It operates in two main stages: (1) identifying important words for the model's prediction, and (2) replacing them with context-appropriate synonyms that preserve semantic meaning.

To estimate word importance, TextFooler applies a leave-one-out strategy: it replaces each word in the input with a mask token and computes the drop in prediction confidence. Words that cause the

largest change in the model's predicted probability are considered most important for the classification decision.

In the original implementation, candidate synonyms are retrieved using a precomputed similarity matrix based on counter-fitted FastText embeddings (Mrkšić et al., 2016), with additional POS filtering to ensure part-of-speech consistency. Counter-fitting is crucial: it repels antonyms and brings genuine synonyms closer, converting ordinary distributional vectors into a usable "synonym space." Such counter-fitted resources do not exist for Ukrainian. Off-the-shelf Ukrainian FastText vectors (Romanyshyn et al., 2023) provide only raw distributional similarity, which is not intended to model synonymy. In a brief evaluation, they often returned morphological variants or even antonyms for example, among the 15 nearest neighbors of хороший ("good") we found нехороший ("not good") and поганий ("bad"). As a result, we determined that raw FastText embeddings are unreliable for synonym discovery in Ukrainian. Retraining FastText using counter-fitting constraints would require significant resources and is out of the scope of this study.

To address this, we replace the FastText synonym source with our curated Ukrainian synonym dictionary (described in Section 5). Since most entries in the dictionary are in lemma form, we use `pymorphy2` to inflect each candidate replacement to match the original word's morphological features in context. On average, each word in the dictionary is associated with 29 synonyms, though this distribution is skewed by outliers—the median number of synonyms is 16. To ensure broad coverage while avoiding excessively long candidate lists, we limit the maximum number of synonyms per word to 200.

In the original TextFooler paper, the authors introduce an importance score threshold to pre-select the most influential words. In our adaptation, we instead run the attack across all words in the input, allowing us to evaluate the full vulnerability surface of the model rather than focusing only on highly weighted words.

Another key component of the original framework is the use of the Universal Sentence Encoder (USE) (Cer et al., 2018) to compute sentence similarity between the original and perturbed inputs. Since USE is not available for Ukrainian, we replace it with the multilingual sentence transformer model `paraphrase-xlm-r-multilingual-v1`, which effectively captures semantic similarity across languages. We retain only those substitutions that maintain a cosine similarity of at least 0.7 between the original and modified sentence.

After running the attack, we observe significant drops in model accuracy, often exceeding 30%, compared to baseline accuracy, demonstrating the effectiveness of this method even with constrained synonym sources. The detailed results, including accuracy drop, average number of queries per sample, and percentage of modified words, are summarized in Table 5. Additional qualitative analysis, including examples of successful and failed replacements as well as the most frequently substituted words are provided in Appendix B.2.

6.2 BERT-Attack

BERT-Attack is another widely used and effective approach for synonym substitution. Unlike TextFooler, which relies on a static synonym dictionary or embedding space, BERT-Attack generates substitutions using a masked language model (MLM). This allows it to produce contextually appropriate replacements that are more fluent and semantically aligned with the original sentence.

To adapt this method for Ukrainian, we use the `xlm-roberta-large` checkpoint as our MLM backbone. This model has shown strong performance on Ukrainian tasks and produces fluent, multilingual outputs due to its extensive training on over 100 languages.

In the original BERT-Attack implementation, the authors apply a byte-pair encoding (BPE) search to explore multi-token substitutions. However, in morphologically rich languages like Ukrainian, subword-level manipulations often result in grammatically invalid forms due to suffixation and complex inflectional endings. Despite extensive hyperparameter tuning, we found that BPE-level substitutions rarely produce valid or useful replacements in Ukrainian. Moreover, performing a full BPE search would significantly increase computational cost. As a result, we simplify the approach by disabling the BPE search and instead use the unmasked 'fill-mask' pipeline to directly suggest full-token replacements, even for multi-token targets.

We follow the original BERT-Attack method for word importance ranking: each word in the sentence is masked one at a time, and the change in the model's prediction probability is recorded. The

words that cause the largest drop in confidence are ranked as most important and are selected first for substitution.

To improve the quality and relevance of substitutions proposed by the masked language model (MLM), we apply several filtering steps. Candidates containing non-Ukrainian characters or invalid symbols are discarded. We then compute the cosine similarity between the original and candidate words using FastText embeddings, retaining only those with a similarity score above 0.33. Finally, to avoid trivial morphological variants, we compare the normal forms of the original and candidate words using `pymorphy2` and remove duplicates.

We configure the 'fill-mask' pipeline with `max_length=512` and enable truncation. For each masked word, we retrieve the top 128 candidate substitutions and keep only those with a confidence score above 0.04. The attack proceeds word by word according to the importance ranking, replacing words until the model's prediction changes or a predefined stop condition is reached. Specifically, we halt the attack if more than 40% of the words in the original text have been substituted, to prevent generating highly unnatural or adversarially overfit inputs.

With this modified setup, we observe a moderate drop in model accuracy, averaging around 12-22%. Although the degradation is not as strong as with TextFooler, the quality of the substitutions is generally higher in terms of fluency and contextual fit. We quantify this via human evaluation: as shown in Table 13, BERT-Attack produces a greater share of grammatically acceptable substitutions compared to TextFooler. Results are shown in Table 6, and examples of good and bad substitutions along with frequently replaced words are presented in Appendix B.3.

7 LLM Evaluation on Attacked Samples

Given the growing use of large language models (LLMs) in real-world NLP applications, we examine whether state-of-the-art LLMs remain vulnerable to synonym substitution attacks. While prior work has shown that such perturbations can mislead traditional models, it remains unclear whether modern instruction-tuned LLMs—especially those with advanced contextual reasoning—exhibit similar vulnerabilities, particularly in low-resource languages like Ukrainian.

Dataset	Orig. Acc.	Adv. Acc.	Drop
TextFooler Attack			
UA Reviews	44.00%	29.00%	-15.00
UA News	61.83%	61.00%	-0.83
UNLP 2025	80.00%	73.12%	-6.88
BERT-Attack			
UA Reviews	28.83%	21.00%	-7.83
UA News	62.83%	52.00%	-10.83
UNLP 2025	71.95%	64.20%	-7.75

Table 4: GPT-4o performance on original vs. adversarial inputs generated by synonym substitution attacks. Each score reflects accuracy over 600 examples.

We sample 200 adversarial examples for each combination of three datasets, three target models, and two attack strategies, yielding 3,600 examples in total (1,800 per attack type). All samples are selected from inputs that successfully fooled the original finetuned classifiers (XLM-RoBERTa, Ukr-RoBERTa, and SBERT), and are reused to test the robustness of GPT-4o - a strong, closed-source LLM with competitive multilingual capabilities, including Ukrainian.

We construct dataset-specific, few-shot prompts for GPT-4o (complete templates are listed in Appendix C). Although these prompts were not tuned to counter our attacks, a substantial portion of examples that fooled the finetuned classifiers likewise fooled GPT-4o. This finding indicates that even high-capacity, instruction-tuned LLMs remain vulnerable to meaning-preserving perturbations in morphologically rich, low-resource languages.

Table 4 summarizes GPT-4o's performance on clean versus adversarial inputs across datasets and attack types. The observed drops in accuracy confirm that synonym substitution remains a potent technique for evaluating the robustness of modern LLMs.

8 Analysis and Discussion

8.1 Overall Model Robustness

Overall, XLM-RoBERTa consistently demonstrated the highest resilience to both TextFooler and BERT-Attack across all three datasets, incurring an average accuracy drop of approximately 23.8 percentage points under TextFooler and 17.9 points under BERT-Attack, suggesting its byte-level BPE and multilingual pre-training lead to more robustness under synonym perturbations.

In contrast, Ukr-RoBERTa suffered the greatest

degradation under TextFooler (mean drop ≈ 29.8 points), and SBERT was the most vulnerable under BERT-Attack (mean drop of 21.5 points).

When comparing the two attacks directly, TextFooler proved more successful attacks - producing a mean accuracy decline of 26 points versus 19 points for BERT-Attack - largely because its dictionary search edits three times as many tokens (queries) on average.

At the dataset level, the UNLP 2025 dataset experienced the most painful impact from TextFooler (mean drop of 34.4 points), and the UA News dataset was hardest hit by BERT-Attack (mean drop of 23.1 points), while the News dataset with short input is the most robust.

8.2 Implications for Disambiguation

Our results underscore the important role of word sense disambiguation (WSD) in designing effective and interpretable synonym substitution attacks. One of the primary weaknesses of TextFooler is its reliance on surface-level synonym lists without accounting for sense disambiguation. This often leads to semantically incorrect substitutions that alter the original meaning. For example, as shown in Table 14, the phrase Команди Формули-1 (teams of Formula 1) was altered to Повеління Формули-1 (commands of Formula 1), where the replacement Повеління is indeed a synonym of Команда but in the sense of a directive or order, leading to incorrect replacement.[1]

Although BERT-Attack can potentially benefit from contextual awareness via MLM, it is still not immune to this issue. In some cases, the model inserts a distributionally similar but semantically unrelated token. For instance, in the sentence дуже класне печиво! свіженьке, ароматне. ("very nice cookie! fresh, aromatic."), the attack replaces печиво with молоко ("milk"), yielding дуже класне молоко! свіженьке, ароматне. - a fluent yet meaning-altering sentence. This illustrates that relying solely on distributional similarity, even with an MLM, is insufficient.

To address these issues, future synonym-substitution frameworks should incorporate sense-aware filtering using lexical-semantic resources, such as the Ukrainian Sense Dictionary, sense-annotated corpora, or the supervised WSD model for Ukrainian introduced by Laba et al. (2023). Such filtering would ensure that substitutions preserve the original meaning and help isolate true model errors from artifacts introduced by poor synonym choices.

[1]We adopted this intentionally naïve dictionary-first strategy because it mirrors the canonical TextFooler/BERT-Attack pipelines used in English, giving a direct cross-language baseline.

8.3 Quality of Synonym Substitutions

To estimate substitution quality, we manually reviewed 100 examples for each combination of dataset and attack method (900 in total), marking replacements as acceptable when they preserved the original word's grammatical form and meaning.

Human evaluation revealed notable differences in substitution quality across attack methods. The one-word baseline yielded the lowest quality, with only 23–40% of replacements rated as fluent and semantically correct, and up to 68% judged as meaning-altering (Table 8). TextFooler improved fluency but still suffered from semantic drift, with 38–51% good substitutions and high rates of incorrect meaning (Table 10). BERT-Attack achieved the highest overall quality (36–42% good), with fewer grammar issues, but semantic mismatches persisted (Table 13).

These results confirm that current synonym substitution attacks often alter sentence meaning and highlight the importance of integrating contextual or sense-aware filtering to improve semantic fidelity.

9 Conclusion and Future Work

We presented the first systematic evaluation of synonym substitution attacks for the Ukrainian language, demonstrating that both simple one-word replacements and advanced frameworks like TextFooler and BERT-Attack can significantly degrade model performance—causing accuracy drops of up to -40.2% with around 113 queries per sample. Few-shot evaluations with GPT-4o show that even large instruction-tuned LLMs remain susceptible, with accuracy declines ranging from 6.9% to 15.0%.

Error analysis shows that some successful attacks work by changing the meaning or producing grammatically invalid substitutions rather than truly revealing model vulnerabilities. This highlights the limitations of current synonym-substitution strategies. To address these, future work should explore hybrid adversarial pipelines that combine synonym dictionaries with masked language model proposals, integrate word sense

Dataset	Model	Orig. Acc.	Adv. Acc.	Drop	Change Rate	Queries
Reviews	Ukr-RoBERTa	76.28%	36.01%	-40.27	15.80%	112.9
	XLM-RoBERTa	77.91%	49.73%	-28.18	14.16%	126.8
	SBERT	77.58%	49.70%	-27.88	13.01%	122.1
UA News	Ukr-RoBERTa	88.55%	73.17%	-15.38	18.87%	50.5
	XLM-RoBERTa	93.52%	82.95%	-10.57	19.87%	52.0
	SBERT	93.46%	82.67%	-10.79	19.11%	51.8
UNLP	Ukr-RoBERTa	81.41%	47.65%	-33.76	11.62%	343.4
	XLM-RoBERTa	80.10%	47.38%	-32.72	10.57%	330.5
	SBERT	81.67%	45.03%	-36.64	9.95%	333.5

Table 5: TextFooler attack results across all datasets.

Dataset	Model	Orig. Acc.	Adv. Acc.	Drop	Change Rate	Queries
Reviews	Ukr-RoBERTa	76.28%	63.31%	-12.97	3.69%	29.7
	XLM-RoBERTa	77.91%	67.27%	-10.64	4.47%	31.9
	SBERT	77.58%	64.79%	-12.79	3.74%	30.5
UA News	Ukr-RoBERTa	98.83%	78.94%	-19.89	15.67%	17.9
	XLM-RoBERTa	93.52%	69.10%	-24.42	14.06%	16.9
	SBERT	93.46%	66.43%	-27.03	13.85%	16.6
UNLP	Ukr-RoBERTa	81.41%	58.38%	-23.03	5.73%	104.0
	XLM-RoBERTa	80.10%	61.52%	-18.58	8.08%	109.1
	SBERT	81.67%	57.07%	-24.60	4.90%	101.8

Table 6: BERT-Attack results across all datasets.

disambiguation to preserve meaning, and leverage LLMs to improve grammaticality and contextual alignment. Semi-supervised techniques may also help expand synonym resources with less manual effort.

Finally, establishing standardized Ukrainian adversarial benchmarks—evaluating not just prediction accuracy, but also fluency and semantic fidelity—will be key to enabling robust and reproducible evaluation of model resilience in low-resource settings.

Limitations

Our evaluation relies on a finite 15000-entry synonym lexicon and heuristic filters, which can potentially miss everyday speaking, domain-specific, or polysemous terms and may introduce semantic drift. Morphological agreement via `pymorphy2` is imperfect, occasionally producing ungrammatical variants. We focus solely on three text-classification tasks and encoder-only transformers, so robustness may differ for generative or sequence-to-sequence models. Human judgments cover only 100 samples per dataset and attack, and our GPT-4o probing used a single prompt template over 3600 cases. Finally, we limited attack budgets ($\leq$ 200 queries or 40% of the words changed), so stronger—but costlier—search strategies might reveal additional vulnerabilities.

Ethical Considerations

Adversarial synonym attacks can be abused to bypass moderation or disrupt Ukrainian NLP services; we release our study and code solely for research purposes and focus exclusively on open-source models. All datasets are publicly licensed - i.e., distributed under explicit open licences that permit research use without additional permission:

Cross_Domain_UA_Reviews (CC-BY-SA 4.0)[2], UA News Classification (MIT)[3], and UNLP 2025 Shared-Task Manipulation (CC-BY-NC-SA-4.0)[4].

Because language models and synonym resources reflect historical biases, perturbations could amplify unfair outcomes, so we recommend pairing this benchmark with fairness audits. GPT-4o evaluations were conducted via the official OpenAI API, in line with ACL ethics guidelines. Training and attacks consumed approximately 43 GPU-hours; we provide checkpoints and logs to avoid redundant computation.

References

Norah Alshahrani, Saied Alshahrani, Esma Wali, and Jeanna Matthews. 2024. Arabic synonym BERT-based adversarial examples for text classification. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 137–147, St. Julian's, Malta. Association for Computational Linguistics.

Moustafa Alzantot, Yash Sharma, Supriyo Chakraborty, Huan Zhang, Cho-Jui Hsieh, and Mani B. Srivastava. 2019. Genattack: practical black-box attacks with gradient-free optimization. In *Proceedings of the Genetic and Evolutionary Computation Conference*, GECCO '19, page 1111–1119, New York, NY, USA. Association for Computing Machinery.

Antonimy.info. 2025. Антоніми — онлайн словник українських антонімів. https://antonimy.info/.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, and 1 others. 2018. Universal sentence encoder. arXiv preprint arXiv:1803.11175.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. arXiv preprint arXiv:1911.02116.

Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. 2018. HotFlip: White-box adversarial examples for text classification. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 31–36, Melbourne, Australia. Association for Computational Linguistics.

[2] https://huggingface.co/datasets/vkovenko/cross_domain_uk_reviews

[3] https://github.com/fido-ai/ua-datasets

[4] https://github.com/unlp-workshop/unlp-2025-shared-task

Наукова думка. 1999. Словник синонімів української мови, volume 2 of Словники України. Наукова думка, Київ.

Ji Gao, Jack Lanchantin, Mary Lou Soffa, and Yanjun Qi. 2018. Black-box generation of adversarial text sequences to evade deep learning classifiers. In 2018 IEEE Security and Privacy Workshops (SPW), pages 50–56.

Matt Gardner, William Merrill, Jesse Dodge, Matthew Peters, Alexis Ross, Sameer Singh, and Noah A. Smith. 2021. Competency problems: On finding and removing artifacts in language data. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 1801–1813, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Bogdan Ivanyuk-Skulskiy, Anton Zaliznyi, Oleksand Reshetar, Oleksiy Protsyk, Bohdan Romanchuk, and Vladyslav Shpihanovych. 2021. ua_datasets: a collection of ukrainian language datasets.

Mohit Iyyer, John Wieting, Kevin Gimpel, and Luke Zettlemoyer. 2018. Adversarial example generation with syntactically controlled paraphrase networks. arXiv preprint arXiv:1804.06059.

Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. 2019. Is bert really robust? natural language attack on text classification and entailment. arXiv preprint arXiv:1907.11932.

Vadym Kovenko. 2021. Cross_domain_ua_reviews. https://huggingface.co/datasets/vkovenko/cross_domain_uk_reviews.

Yurii Laba, Volodymyr Mudryi, Dmytro Chaplynskyi, Mariana Romanyshyn, and Oles Dobosevych. 2023. Contextual embeddings for Ukrainian: A large language model approach to word sense disambiguation. In Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP), pages 11–19, Dubrovnik, Croatia. Association for Computational Linguistics.

Jinfeng Li, Shouling Ji, Tianyu Du, Bo Li, and Ting Wang. 2018. Textbugger: Generating adversarial text against real-world applications. arXiv preprint arXiv:1812.05271.

Linyang Li, Ruotian Ma, Qipeng Guo, Xiangyang Xue, and Xipeng Qiu. 2020. BERT-ATTACK: Adversarial attack against BERT using BERT. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 6193–6202, Online. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.

Nikola Mrkšić, Diarmuid Ó Séaghdha, Blaise Thomson, Milica Gašić, Lina M. Rojas-Barahona, Pei-Hao Su, David Vandyke, Tsung-Hsien Wen, and Steve Young. 2016. Counter-fitting word vectors to linguistic constraints. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 142–148, San Diego, California. Association for Computational Linguistics.

Vitalii Radchenko. 2021. ukr-roberta-base: A roberta model pretrained on ukrainian text. https://huggingface.co/youscan/ukr-roberta-base. Pretrained on Ukrainian Wikipedia, OSCAR, and social media corpora.

Sagnik Ray Choudhury, Anna Rogers, and Isabelle Augenstein. 2022. Machine reading, fast and slow: When do models "understand" language? In Proceedings of the 29th International Conference on Computational Linguistics, pages 78–93, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. arXiv preprint arXiv:1908.10084.

Shuhuai Ren, Yihe Deng, Kun He, and Wanxiang Che. 2019. Generating natural language adversarial examples through probability weighted word saliency. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 1085–1097, Florence, Italy. Association for Computational Linguistics.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 4902–4912, Online. Association for Computational Linguistics.

Nataliia Romanyshyn, Dmytro Chaplynskyi, and Kyrylo Zakharov. 2023. Learning word embeddings for Ukrainian: A comparative study of fastText hyperparameters. In Proceedings of the Second Ukrainian Natural Language Processing Workshop, pages 20–31, Dubrovnik, Croatia. Association for Computational Linguistics.

Synonimy.info. 2025. Синоніми — онлайн словник українських синонімів. https://synonimy.info/.

UNLP Workshop Organizers. 2025. UNLP 2025 shared task: Detecting social media manipulation. https://github.com/unlp-workshop/unlp-2025-shared-task. Licensed under CC BY-NC-SA 4.0.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. Advances in neural information processing systems, 30.

Zihan Zhang, Mingxuan Liu, Chao Zhang, Yiming Zhang, Zhou Li, Qi Li, Haixin Duan, and Donghong Sun. 2021. Argot: Generating adversarial readable chinese texts. In Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence, pages 2533–2539.

Kaijie Zhu, Jindong Wang, Jiaheng Zhou, Zichen Wang, Hao Chen, Yidong Wang, Linyi Yang, Wei Ye, Yue Zhang, Neil Gong, and 1 others. 2023. Promptrobust: Towards evaluating the robustness of large language models on adversarial prompts. In Proceedings of the 1st ACM Workshop on Large AI Systems and Models with Privacy and Safety Analysis, pages 57–68.

A Model Finetunning

All models were fine-tuned using the same training configuration across datasets and architectures. We used Hugging Face's Transformers library with the following hyperparameters:

Parameter	**Value**
Optimizer	AdamW
Learning rate	2×10^{-6}
Batch size	32
Max sequence length	512
Warm-up steps	10% of total steps
Learning rate scheduler	Linear
Epochs	up to 15
Hardware	NVIDIA RTX 3090 (24 GB)

Table 7: Fine-tuning hyperparameters used for all models.

B Attacks Results

B.1 One word

Dataset	Good	Semantic	Grammar
UA Reviews	40	48	12
UA News	28	64	8
UNLP 2025	23	68	9

Table 8: Human evaluation of One word-generated adversarial examples. Each row shows the percentage (%) of replacements judged as fluent and correct (*Good*), semantically incorrect (*Semantic*), or ungrammatical (*Grammar*) across 100 samples per dataset.

B.2 TextFooler

Original	Replacements
рекомендувати	відрекомендовувати
хороший	непоганий
якісний	тривкий
гарний	непоганий
зручно	вигідно
чудовий	доладний
якісний	доброякісний
працювати	мозолитися
товар	крам
сподобатися	пригледітися

Table 9: Top 10 word replacements that break a model for TextAttack

Dataset	Good	Semantic	Grammar
UA Reviews	51	40	9
UA News	38	56	6
UNLP 2025	38	57	5

Table 10: Human evaluation of TextFooler-generated adversarial examples. Each row shows the percentage (%) of replacements judged as fluent and correct (*Good*), semantically incorrect (*Semantic*), or ungrammatical (*Grammar*) across 100 samples per dataset.

Good replacements	
Dataset	**Example**
UNLP 2025	Orig (True): Роботине, запорізький напрямок, дуже файно працює наша артилерія. Adv. (False): Роботине, запорізький напрямок, дуже добре працює наша артилерія.
UA News	Orig (новини): В Україну повертається спека Adv. (політика): В Україну вернеться жара
UA Reviews	Orig (4/5): ... Ще не встановлював. Але низ (невидима сторона) дійсно якась дивна ... Adv. (3/5): ... Ще не встановлював. Але низ (невидима сторона) дійсно якась чудна ...
UA Reviews	Orig (5/5): Дійсно дуже якісний і теплий костюм. .... Повністю коштує своїх грошей Adv. (4/5): Дійсно дуже тривкий і теплий костюм. .... Абсолютно коштує своїх грошей
Bad replacements	
UNLP 2025	Orig (False): російська армія знову вдарила по Нікопольщині. Від ранку – двічі. Adv. (True): російська армія знову трафила по Нікопольщині. Від вавку – двічі.
UA News	Orig (спорт): У кого найкрасивіший болід? Команди Формули-1 показали нові машини Adv. (технології): У кого найкрасивіший болід? Повеління Формули-1 виставляли нові ...
UA Reviews	Orig (4/5): Набір хороший. Великі зручні маркери, олівці м'яко пишуть. Все пахне. Adv. (3/5): Набір нічогенький. Великі зручні маркери, олівці м'яко гилять. Все смердітиме.
UA Reviews	Orig. (5/5): Лосьон гарно зволожує шкіру дитини. Не викликає алергії. Хороший склад. Adv. (4/5): Лосьон ладно зволожує шкірку дитини. Не веселить алергії. Непоганий лад.

Table 11: Examples of good and bad adversarial replacements for the TextFooler attack across datasets. The labels in parentheses show the model's predicted label for each original (Orig) and adversarial (Adv.) example.

B.3 BERTAttack

Original	Replacements
ціна	вартість
чудовий	хороший
чудово	нормально
львівщина	Донбасі
млрд	млн
якісний	хороший
приємний	хороший
сподобатися	подобається
грн	гривень
гарний	хороший

Table 12: Top 10 word replacements that break a model for BertAttack

Dataset	Good	Semantic	Grammar
UA Reviews	40	54	6
UA News	36	61	3
UNLP 2025	42	58	0

Table 13: Human evaluation of BERT-Attack-generated adversarial examples. Each row shows the percentage (%) of replacements judged as fluent and correct (*Good*), semantically incorrect (*Semantic*), or ungrammatical (*Grammar*) across 100 samples per dataset.

Good replacements	
Dataset	**Example**
UNLP 2025	Orig (False): уряд чехії готується передати україні нову партію танків. Adv. (True): влада чехії хоче передати україні нову партію танків.
UA News	Orig (спорт): відео. дворічний син мессі показав, як потрібно качати прес Adv. (новини): відео. дворічний хлопчик мессі показав, як потрібно качати прес
UA Reviews	Orig (1/5): ... (або ж мені потрапив брак): фільтр абсолютно не працює - вода ... Adv. (2/5): ... (або ж мені потрапив дефект): фільтр практично не працює - вода ...
UA Reviews	Orig (3/5): вже кілька раз були поломки ... Adv. (2/5): вже кілька раз були проблеми ...
Bad replacements	
UNLP 2025	Orig (True): ворог не полишає спроб зруйнувати енергосистему, відправляючи десятки ... Adv. (False): ворог не робить спроб зруйнувати ситуацію, включаючи десятки ...
UA News	Orig (політика): польща і франція разом робитимуть новий танк Adv. (новини): Україна і франція спільно зробили новий танк
UA Reviews	Orig (5/5): поломалась присоска. підкажіть де купити нову. Adv. (3/5): поломалась камера. підкажіть де купити нову.
UA Reviews	Orig. (5/5): олія добра,смак легкий, зовсім трохи відчувається оликовий присмак ... Adv. (4/5): вода добра,смак легкий, зовсім трохи має оликовий присмак ...

Table 14: Examples of good and bad adversarial replacements for the BERT-Attack attack across datasets. The labels in parentheses show the model's predicted label for each original (Orig) and adversarial (Adv.) example.

C LLM Prompt Templates

C.1 UA Reviews

System Prompt

Ви — модель GPT-4o, мета якої — оцінити якість відгуку українською мовою за шкалою від 0 до 4, де:
0 — дуже погано
1 — погано
2 — посередньо
3 — добре
4 — дуже добре
Поверніть тільки JSON-об'єкт із ключем "predicted_label" без будь-яких трикрапок чи пояснень.

Few-Shot Examples

Приклад 1:
Вхід: {"review": "Я замовив доставку вчасно, але піца була холодною й пересоленою."}
Вихід: {"predicted_label": 1}

Приклад 2:
Вхід: {"review": "Чудовий сервіс, ввічливий персонал і дуже смачна їжа!"}
Вихід: {"predicted_label": 4}

Приклад 3:
Вхід: {"review": "Загалом непогано, але десерт міг бути солодшим."}
Вихід: {"predicted_label": 2}

Приклад 4:
Вхід: {"review": "Не рекомендую — замовлення загубили, потім переплутали страви."}
Вихід: {"predicted_label": 0}

C.2 UA News Classification

System Prompt

Ви — модель GPT-4o, мета якої — класифікувати українські заголовки новин за однією із п'яти категорій: «бізнес», «новини», «політика», «спорт», «технології». Поверніть тільки назву категорії.

Few-Shot Examples

Заголовок: "Уряд затвердив нову стратегію економічного розвитку"
Категорія: політика

Заголовок: "Apple анонсує новий iPhone з поліпшеною камерою"
Категорія: технології

Заголовок: "Шахтар перемагає у фіналі Ліги чемпіонів"
Категорія: спорт

C.3 UNLP: Manipulation Detection

System Prompt

Ви — модель GPT-4o, мета якої — визначити, чи містить український текст у соціальних мережах маніпулятивні риторичні чи стилістичні прийоми, спрямовані вплинути на аудиторію без чітких фактів.
Поверніть лише JSON-об'єкт із ключем "predicted_label":
1 — якщо маніпуляція є,
0 — якщо маніпуляції немає.

Few-Shot Examples

Вхід: "Всі нормальні люди вже бачать правду! Приєднуйтеся і ви, поки вам не пізно!"
Вихід: {"predicted_label": 1}

Вхід: "Згідно з офіційним звітом, кількість відвідувачів музею зросла на 15%."
Вихід: {"predicted_label": 0}

Вхід: "Уряд мовчить про реальні витрати — вони приховують від вас правду!"
Вихід: {"predicted_label": 1}

Вхід: "Не забудьте перевірити рівень масла перед довгою поїздкою."
Вихід: {"predicted_label": 0}