Breaking Free from Ivory Tower: Evaluating and Enhancing Real-world Chinese Underground Adversarial Jargon Detection

Zhifan Jiang[†]* , Mingxuan Liu[§]*, Yue Qin[Φ✉], Baojun Liu[‡✉]

[†]Zhejiang University, [§]Zhongguancun Laboratory,
[Φ]Central University of Finance and Economics, [‡]Tsinghua University

***Abstract*—Underground jargon in online ecosystems threatens platform safety and public trust by evading automated content moderation and thereby concealing criminal coordination across fraud, gambling, and illicit commerce, particularly in Chinese, given the language's unique graphophonemic properties and large user base. Despite machine learning-based content moderation, underground actors increasingly evade detection with human-crafted adversarial perturbations that differ fundamentally from budget-constrained algorithmic attacks, highlighting a research gap in understanding and debunking real-world evasion techniques.**

In collaboration with a leading security company, we annotate and release the first large-scale, in-the-wild dataset of adversarial Chinese underground jargon and uncover its distinct characteristics—higher perturbation intensity, greater diversity of perturbation forms, and severe structural disruption—resulting in readability degradation and tokenization collapse that exacerbate detection vulnerabilities. Our systematic measurement study across state-of-the-art large language models (LLMs) further confirms the challenges in recognizing real-world adversarial Chinese jargon: even advanced models (e.g., GPT-4o) exhibit notable limitations, achieving only 88.16% and 62.74% accuracy in jargon detection and restoration, respectively, and 76.05% accuracy in illicit content detection. To address these challenges, we propose JADE, an LLM-based detection framework grounded in a taxonomy of perturbation patterns systematically derived from annotated real-world data and designed to enable reasoning beyond explicitly observed variants. This taxonomy informs a realistic adversarial learning curriculum that combines data augmentation, external knowledge retrieval, and internal adaptation, aligning the model with real-world jargon semantics. Experiments show that JADE significantly outperforms both commercial and fine-tuned open-source LLMs, achieving 98.59%, 95.91% accuracy in jargon detection and restoration, and 97.99% accuracy in illicit content detection. Moreover, it generalizes well to other downstream applications threatened by adversarial text, with F1 scores of 93.25% and 94.33% on harmful text and fraud detection, respectively. Overall, our work takes a first step toward leveraging LLM reasoning and structured perturbation knowledge to defend against real-world adversarial jargon.

Disclaimer: This paper contains offensive content that may be disturbing to some readers.

1. Introduction

Underground-industry jargon[1] is pervasive in underground marketplaces, especially across illicit services such as illegal gambling, fraud, and drug trafficking [1], [2], [3], [4]. It acts as a key driver of illicit promotion and facilitates the spread of illegal activities on the public Internet. More seriously, exposure to such underground content poses significant harm to general Internet users, especially minors who may encounter various forms of harmful material such as illicit pornography. While this risk is globally prevalent [5], it is disproportionately acute in the Chinese-language context due to its large user base [6] and distinctive graphophonemic characteristics that enable extensive and sophisticated obfuscation [7]. In response, major platforms (e.g., Twitter, Reddit) deploy machine learning-based content moderation systems trained on labeled data [6], [8].

However, adversarial jargon has become increasingly prevalent as underground actors introduce carefully engineered perturbations to existing jargon to evade detection. For example, they may transform the gambling term "捕鱼" into the pinyin-matched variant "卜余", thereby obscuring its meaning and degrading models' semantic understanding. This "cat-and-mouse" evasive tactic reduces coordination costs for malicious actors, enabling large-scale communication, recruitment, and promotion while undermining key security workflows, including platform moderation of illicit content, offline forensic analysis of suspicious communications, and monitoring of emerging jargon for defensive updates. To counter such threats, numerous adversarial text defense methods have been proposed [7], [9], [10], [11], [12], [13]. Yet their idealized assumptions and narrow focus on specific attack types leave real-world challenges largely unaddressed [7], [14]. This dynamic creates an ongoing arms race in which adversaries continually refine and evolve jargon variants to evade improved detection mechanisms.

* *These authors contributed equally to this work.*
✉ *Corresponding authors.*

1. Unless stated otherwise, jargon in this paper refers to underground-industry-related terminology.

Consequently, ever-evolving adversarial jargon remains a critical challenge for the Internet ecosystem.

Research Gap. Despite advances in general adversarially robust detection [7], [14], existing methods typically rely heavily on synthetic adversarial examples generated under idealized assumptions (e.g., bounded perturbation budgets) and are often tailored to specific attack types (e.g., synonym or homophone substitution). Prior work on human-written perturbations largely targets toxic/hate content [7], [15], [16], whereas underground communication is intent-centric and domain-specific—hence more context-dependent and reliant on in-group semantics and domain knowledge for canonical restoration and labeling. Moreover, because these ecosystems are economically driven, actors iterate rapidly and deploy higher-intensity, multi-technique perturbations, accelerating the attack–defense cycle. *Consequently, we lack a systematic account of the perturbation toolkit actually used by real-world underground actors—let alone effective, taxonomy-aware defenses tailored to this setting.*

Characterizing Real-world Perturbation. To address this gap, we collaborated with a leading security company to construct **ADVJARGON**, a large-scale dataset of real-world adversarial Chinese underground jargon. In this work, we focus specifically on six underground business categories, namely Gambling, Porn, Drug, Fraud, Danger, and Promotion, which constitute major profit channels and service types sustaining the underground economy [2], [5]. Using the company's mobile security operations platform, we collected 70,000 user-submitted spam messages, each preverified as malicious. From this corpus, we curated a Chinese jargon lexicon through meticulous annotation, mapping adversarial variants to their canonical forms and labeling perturbation types to systematically capture real-world evasion strategies. We then compared these real-world samples with adversarial examples generated by two standard text attack algorithms [7], [15], revealing that real-world jargon exhibits significantly higher perturbation intensity, greater diversity of obfuscation techniques, and more severe structural disruption. These factors lead to pronounced readability degradation and tokenization collapse, severely impairing both preprocessing and semantic modeling, and ultimately exacerbating detection vulnerability.

Challenges Posed by Real-World Perturbations. While traditional text robustness methods struggle to generalize beyond synthetic perturbations, large language models (LLM) offer stronger contextual reasoning and generalization capabilities, making them a promising candidate for detecting real-world adversarial jargon. Building on ADVJARGON, we conduct the first systematic evaluation of large language models (LLMs) in detecting real-world adversarial Chinese jargon, benchmarking nine state-of-the-art models. Despite the success of LLMs in other text moderation tasks, our results reveal that real-world evasion strategies significantly undermine their effectiveness. For example, GPT-4o achieves only 62.74% accuracy in jargon restoration (i.e., mapping adversarial jargon to its canonical form) and 87.82% in jargon detection. Even with common enhancement techniques such as few-shot learning, performance gains are marginal—reaching only 90.71%. These findings underscore the difficulty of this task: complex obfuscation strategies, disrupted semantic coherence, and severe tokenization interference pose substantial challenges to LLM-based detection.

Our detector: JADE. To address these limitations, we propose and implement JADE, an LLM-based framework for detecting and restoring adversarial Chinese underground jargon. Unlike existing defense methods, JADE is grounded in a taxonomy of perturbation patterns systematically induced from annotated real-world data. This taxonomy informs the design of a realistic adversarial learning curriculum comprising: (1) an external knowledge retrieval module, which provides multi-granularity exemplars from a dynamic knowledge base; and (2) an internal adaptation module, which fine-tunes the LLM to align with the semantics of real-world jargon. Experiments on our benchmark dataset show that JADE significantly outperforms both commercial and fine-tuned open-source LLMs, achieving 95.91% accuracy in jargon restoration and 98.59% in detection—substantially exceeding the best fine-tuning baseline (75.57% and 68.70% by RoCBERT) and the best commercial LLM (64.77% in restoration by DeepSeek-R1 and 91.13% in detection by GLM-4-Plus). An ablation study demonstrates the effectiveness of taxonomy-guided data augmentation based on our real-world adversarial jargon dataset, improving jargon detection accuracy by 8.53% and restoration accuracy by 11.75%, respectively. Moreover, JADE demonstrates strong cross-domain generalization, maintaining state-of-the-art performance on toxic content [15] and fraud detection [16]. On ToxiCloakCN, it achieves an F1 score of 0.932, outperforming the best prior model, GPT-4o (0.796), highlighting its robustness against real-world adversarial text across domains.

Contribution. To summarize, this work offers the following contributions.

- *New Dataset.* We present ADVJARGON, the first annotated dataset of real-world adversarial Chinese underground jargon, and reveal characteristics that distinguish them from algorithmic adversarial examples, challenging core assumptions in robust detection. These include higher perturbation intensity, greater diversity of perturbation forms, and severe structural disruption—leading to readability degradation and tokenization collapse.
- *New Insight into Real-world Perturbation.* Leveraging the ADVJARGON dataset, we present the first systematic measurement study of how real-world perturbations impact illicit content detection, revealing model limitations against real-world adversarial strategies and identifying key factors influencing detection performance.
- *Novel Detection Methodology.* We propose JADE, the first LLM-based framework for detecting and restoring adversarial Chinese underground jargon. It is equipped with a taxonomy of real-world perturbation patterns that guides a realistic adversarial learning curriculum, integrating external knowledge retrieval and internal model adaptation to enhance semantic alignment and robustness.
- *Open Science.* Our work advances research in the field

of online content moderation. We will publicly release the dataset and methodology[2] to support the academic community in collectively addressing underground content detection challenges.

2. Related Work

Robust Adversarial Text Classification. Adversarial texts add perturbations to the text in order to interfere with AI model classification judgments. In the white-box setting, attackers have access to the internal information of the target model, so existing research has proposed methods based on gradient information to identify key components of the input text and perform replacements. In contrast, the black-box setting is more challenging because attackers can only rely on the model's output responses to generate perturbations. Without access to gradient information, methods like TextFool [17] and TextGan [18] have been proposed to identify key information for perturbation. Recently, Yang et al. [7] used large language models and few-shot prompts to automatically extract harmful entities from existing datasets and apply a proposed multimodal perturbation classification method to systematically perturb them.

For Chinese, existing research usually uses homophones and visually similar characters to generate perturbations. Due to the complexity and randomness of adversarial text perturbations in Chinese, detecting adversarial Chinese texts is a challenging task. Some existing works have proposed the following three methods to improve detection performance. (1) Data Augmentation: By expanding the training dataset with synthesized adversarial examples, the robustness of the model is enhanced [19] [20]. (2) Pattern Augmentation: Fusing multimodal knowledge, such as semantics, glyph, and pinyin, to enhance Chinese text understanding. Su et al. [10] proposed the pre-trained RoCBert model, which uses multimodal adversarial contrastive learning methods for pre-training the model. Li et al. [11] constructed an adversarial knowledge graph to capture the similarity relationships between Chinese characters in terms of glyph and pronunciation. (3) Pinyin Correction: First, a spelling checker is used to remove noise, and then the corrected text is input to the model. Yeh et al. [12] and Yu et al. [13] tried to restore adversarial variants to their original forms using dictionary-based and language model-based methods. Li et al. [9] proposed combining adversarial translation and multimodal embedding to achieve robust text detection. However, any slight errors in the spelling correction process may lead to unpredictable model behavior.

Real-world Adversarial Text. Existing research shows that real-world adversarial text poses a significant threat for detection. Gröndahl et al. [21] noted simple word perturbations (e.g., Leetspeak, space removal) markedly weaken mainstream models' and commercial APIs' detection performance. Aggarwal et al. [22] found attackers intuitively choosing confounding words inflict damage comparable to complex white-box attacks. For better threat assessment, subsequent work focused on real-world data collection/analysis. Le et al. [23] proposed CRYPTEXT—continuously updating a real-world human perturbation database to auto-discover, analyze, and exploit social media's human-written adversarial text. Ye et al. [24] created NoisyHat (a dataset for real-world adversarial text). Their findings show existing spelling correction tools and mainstream models perform poorly on such perturbations, highlighting insufficient model robustness in real scenarios. For defense, Mishra et al. [25] proposed a character-based word composition model to handle real-world structurally obfuscated adversarial text. It infers semantic embeddings from character sequences, accurately identifying unseen confounding words. Despite recognition of the gap between defenses for simulated adversarial texts and real-world scenarios, most follow-up work has focused on specific text-processing domains, such as hate speech and toxic content. Currently, there remains a gap in understanding and defending against real-world adversarial text in underground-related communications, where the malicious impact is severe.

Jargon Detection. Underground jargon—key for adversarial text evasion—is critical to identifying and combating underground activities [26]. Early detection relied on manually built dictionaries, expanded via darknet forums [6], search engines [12], and underground forums [26] to find emerging jargon. Subsequent work focused on lexical semantic modeling: Yuan et al. [5] used cross-corpus word vector shifts; Song et al. [8] proposed delexicalized distant supervision with contextual/word attribute modules for unseen slang; others transformed dark jargon via cross-corpus word probability differences [27], [28]. While these methods have been tested on algorithmically generated English adversarial texts, real-world Chinese jargon is more complex due to the structure of the language and the differences between them, resulting in challenges for detecting adversarial Chinese jargon in practice.

LLM-Driven Content Classification. With the development of LLMs, their capabilities have expanded from traditional text moderation to cross-domain content safety detection. Unlike traditional classifiers, LLMs offer stronger generalization and reasoning abilities, enabling flexible adaptive detection in dynamic threat environments. In Harmful Memes detection, multimodal LLMs can jointly understand image and text semantics to identify hidden malicious content. Ma et al. [29] systematically evaluated the strengths and risks of open-source vision-language models in meme understanding and generation, finding that while these models capture meme visuals and cultural semantics well, they still have safety alignment issues, making them prone to misuse for generating harmful content. To enhance multimodal model safety, Zhuang et al. [30] proposed the HMGUARD framework, combining multimodal LLMs with adaptive prompting and Chain-of-Thought reasoning to improve the detection of complex implicit harmful content. Additionally, Lin et al. [31] used reasoning distillation and harmfulness inference in a two-stage training strategy, transferring LLM reasoning to lightweight models, thus maintaining accuracy while improving deployability and efficiency. Despite suc-

2. Available at: https://github.com/jzf1231/JADE.

cessful testing in controlled environments, the performance of LLMs in detecting high-adversarial Chinese jargon in real-world scenarios remains uncertain.

3. Preliminary Study

In this study, we release the *first annotated dataset of in-the-wild adversarial Chinese underground jargon*, mapping each adversarial variant to its canonical form and labeling perturbation types that capture real-world evasion, thereby supporting normalization and robust detection. We further conduct a quantitative study demonstrating the dataset's distinctive characteristics and the challenges they pose for real-world adversarial-jargon detection.

3.1. Dataset Construction and Annotation

Below, we detail the collection of an in-the-wild jargon corpus; the identification of canonical and adversarial underground-jargon forms; the construction of a jargon lexicon; and the annotation protocol for jargon pairs and perturbation types.

Corpus of In-the-Wild Jargon. To examine Chinese underground jargon in real-world use, we collaborated with a globally leading security company[3] and constructed a corpus $\mathbb{C}$ of 70,000 real-world **spam** SMS messages collected via its widely deployed mobile security app. Messages were user-reported to a spam folder and underwent preliminary manual verification by our partner to confirm their spam nature, reducing false reports (i.e., normal messages) from user error. The dataset spans major underground activities—including gambling, fraud, and illicit promotion—providing strong representativeness and diversity.

Chinese Jargon: Canonical vs. Adversarial. Chinese underground *jargon* denotes a coded lexicon used in illicit activities (e.g., gambling, fraud, drug trafficking) to convey domain-specific meanings to in-group readers. To evade platform censorship and automated detection, *adversarial jargon* comprises surface variants constructed via disguise, obfuscation, or substitution, such that intended meanings remain recoverable to in-group readers while remaining opaque to automated filters. Formally, a lexical item $j \in \mathcal{X}$ is called *jargon* if it encodes a malicious intent $m \in \mathcal{M}$ in the sense that the human-understanding function $\mathcal{H} : \mathcal{X} \to \mathcal{M}$ satisfies $\mathcal{H}(j) = m$. We call j *canonical jargon* if it encodes m as a sensitive word or a specialized in-group term without surface obfuscation. Such occurrences are typically detectable (e.g., via a blocklist). Given a perturbation function $\delta : \mathcal{X} \to \mathcal{X}$ that alters surface forms, a token or segment $w' \in x' = \delta(x)$ is deemed *adversarial jargon* for m iff $\mathcal{H}(w') = m$ and w' arises from obfuscation (e.g., glyph confusables, homophony). In practice, adversarial jargon is often *operationally* characterized by reduced detectability under automatic filters. If P denotes a platform with detector $\mathrm{Detect}_P : \mathcal{X} \to \{0, 1\}$, adversarial realizations w' frequently satisfy $\mathrm{Detect}_P(x') = 0$ while $\mathcal{H}(w') = m$.

3. 360 Mobile Security.

Construction of Jargon Lexicon. To enable accurate understanding and recognition of jargon expressions, we first compiled an extensive list of commonly used underground jargon as a core reference throughout the annotation pipeline. Specifically, we manually extracted 245 known jargon items from publicly available "black jargon" manuals and security reports [32], [33], [34], [35] as a seed set. Building on this, we designed a collection framework that iteratively queried search engines (e.g., Google) and used custom crawlers to collect openly accessible content from high-ranking underground forums (e.g., Tea Horse Road), technical boards (e.g., Reddit), and public chat channels (e.g., Telegram). The crawling strictly excluded private groups and encrypted communications to meet ethical requirements. After harvesting large-scale raw text, we preprocessed and deduplicated it, then applied co-occurrence analysis and GPT-4o to surface candidate terms strongly associated with the seed jargon. All candidates underwent strict human vetting, where annotators checked context and semantics and removed ambiguous items and false positives. This pipeline yielded a high-quality lexicon of 4,933 underground jargon entries, covering domains such as gambling, fraud, illicit promotion, and drugs.

Real-World Adversarial Jargon Dataset: ADVJARGON. Based on the jargon lexicon $\mathbb{L}$ and SMS corpus $\mathbb{C}$, we constructed an annotated dataset of adversarial Chinese underground jargon that maps each adversarial variant to its canonical form and labels the associated perturbation types. Annotation was performed by three experts with years of experience in black-/gray-market detection and governance, using qualitative open coding [36] to develop the codebook guiding the process.

We randomly sampled 10,000 spam SMS messages from $\mathbb{C}$ and annotated all adversarial jargon instances within this subset, denoted $\mathbb{C}'$. Let Σ^* be the string space, $\mathbb{L} \subseteq \Sigma^*$ the set of canonical underground jargon terms, and $\mathcal{T}$ the finite set of perturbation types. Define $J_c := \mathbb{L}$ and let $J_a \subseteq \Sigma^*$ be adversarial surface forms extracted from $\mathbb{C}'$. We posit a many-to-one mapping $\pi : J_a \to J_c$ that assigns each $j_a \in J_a$ a unique canonical form $j_c = \pi(j_a) \in J_c$ (via lexicon lookup and adjudication), and a *multi-label* function $\tau : J_a \times J_c \to 2^{\mathcal{T}}$ that returns the set of perturbation types $T_{j_a,j_c} = \tau(j_a, j_c) \subseteq \mathcal{T}$ describing the transformation(s) from j_c to j_a. The annotated dataset is $\mathcal{D} = \{(j_a, j_c, t) \mid j_a \in J_a,\ j_c = \pi(j_a),\ t \in \tau(j_a, j_c)\} \subseteq (J_a \times J_c) \times \mathcal{T}$.

We conducted five iterative rounds of annotation over the sampled subset $\mathbb{C}'$, reserving a fixed held-out set of 1,000 messages for validation. Rounds 1–4 covered the remaining 9,000 messages in stratified batches. In each batch, annotators extracted adversarial forms J_a, linked each j_a to a unique canonical form $j_c = \pi(j_a) \in J_c$, and assigned perturbation types $\tau : J_a \times J_c \to 2^{\mathcal{T}}$. Each item was independently labeled by two annotators; disagreements were adjudicated by a third annotator under the evolving codebook. After each round, confirmed pairs $(j_a, \pi(j_a))$ and their perturbation types $\tau(j_a, \pi(j_a))$ were appended to a shared reference list, and the codebook was refined as needed. Rounds 2–4 emphasized disagreement resolution

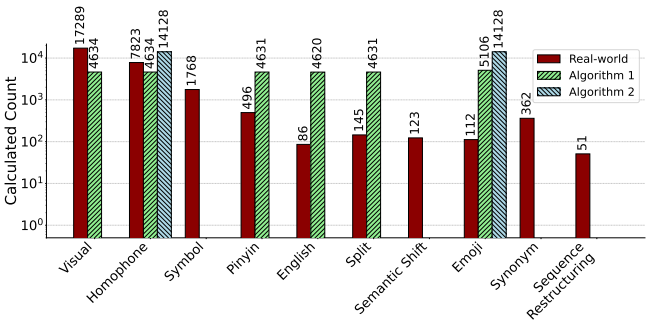


Figure 1: Distribution of adversarial perturbation strategies in ADVJARGON and algorithmic data

and supplemented perturbation categories not observed in Round 1; this refinement revealed an additional category, namely semantic shift. In Round 5, the codebook was *frozen* and applied to the held-out 1,000-message set by two independent annotators. We computed Cohen's κ in the two-rater setting for (i) the mapping task $j_c = \pi(j_a)$, yielding $\kappa_{\text{map}} = 0.8582$, and (ii) perturbation assignment by treating each $t \in \mathcal{T}$ as a binary label and macro-averaging per-technique κ_t, yielding $\kappa_{\text{tech}} = 0.8425$. Disagreements were adjudicated by the third annotator.

To maintain consistency with the frozen codebook without re-annotating all 10,000 messages, we conducted a targeted consistency pass on the 9,000 previously labeled messages, restricted to disagreement cases from Rounds 2–4. These items were re-labeled independently by two annotators, with a third adjudicating any remaining disagreements. The final release comprises 1,020 canonical jargon and 8,576 adversarial variants (deduplicated) across 10,000 messages with the finalized set of 10 perturbation types.

3.2. Characteristics of Real-World ADVJARGON

To gain deeper insights into real-world Chinese adversarial jargon, we conducted a quantitative analysis on an annotated real-world corpus, evaluating perturbation strategies, tokenization performance, and detection accuracy. To compare real-world instances with algorithmically generated counterparts, we employed two state-of-the-art generation methods: Algorithm 1 [15] and Algorithm 2 [7]. Specifically, after perturbing, recovering, and cleaning the real-world samples in ADVJARGON, we applied both algorithms and compared their outputs against ADVJARGON. Based on our annotations, we identified four primary perturbation categories in real-world underground texts—pronunciation, glyph, semantics, and others—each comprising several subtypes, as detailed in Table 11 in the Appendix. Below, we highlight the distinctive characteristics of real-world adversarial jargon that differentiate them from algorithmically generated examples.

Higher Perturbation Intensity. We find that real-world adversarial Chinese underground jargon exhibits significantly higher perturbation intensity than synthetic adversarial samples generated using predefined rules [7], [17], [18]. Unlike rule-based perturbations that often apply minimal surface-level changes, such as substituting "博彩" (gambling) with phonetically similar "菠菜" (spinach), real-world jargon tends to combine multiple obfuscation strategies, yielding greater semantic drift and structural disruption. For example, "博彩" may be rendered as "bo/菜," introducing both token splitting and visual noise, while "枪支" (firearm) is replaced by semantically distant codewords like "狗" (dog), relying on community-specific shared context to preserve meaning. These transformations not only increase resistance to detection but also degrade readability and undermine token-level interpretability, imposing greater challenges for both human and automated understanding.

Greater Diversity of Perturbation Forms. We compared the distribution of perturbation strategies between real-world scenarios and two theoretical algorithms, as illustrated in Figure 1. We observe that the strategies in real-world activities are more diverse, utilizing all 10 types of perturbations, with the highest frequency of use for Visual and Homophone perturbations. In contrast, the theoretical algorithms are narrowly focused: Algorithm 1 utilizes only six types, while Algorithm 2 demonstrates an extreme hyperspecialization, concentrating almost exclusively on Homophone and Emoji.

Severe Structural Disruption. In real-world underground texts, we observe severe structural disruption caused by pervasive data contamination. This includes non-standard expressions, character swap, irregular symbol insertions, and fragmented character combinations with ambiguous boundaries. Such structural noise undermines the natural fluency of the text and breaks the statistical regularities relied upon by Chinese word segmentation models. For example, while algorithmic perturbations might alter "斗地主, 龙虎" (two gambling games) to "抖地主, 笼虎," real-world variants show greater distortion, such as "斗 | 龙地 | 主琥 |," which fragments and reorders the sequence. As a result, even state-of-the-art tokenizers fail to produce coherent segmentation. These disruptions also lead to pronounced readability degradation, making the content difficult to interpret. In the following, we present a quantitative evaluation of these effects across readability, tokenization accuracy, and adversarial jargon detection performance.

• *Impact: Readability Degradation & Tokenization Collapse*. The distinct characteristics of real-world Chinese adversarial jargon lead to severe readability degradation and tokenization collapse. To quantify this, we applied metrics such as perplexity and grammatical errors for readability. As shown in Table 1, both perplexity and grammatical errors are higher in the real-world corpora, reflecting increased noise and more fragmented structures that further hinder detection.

TABLE 1: Comparison of Perplexity (PPL) and Grammar Errors across datasets.

Dataset	PPL	Grammar Errors
Real-world	37.0095	218
Algorithm 1	32.0659	135
Algorithm 2	30.1867	106

TABLE 2: Performance of mainstream Chinese word tokenizer on ADVJARGON.

Tokenizer	Accuracy (%)	Recall (%)	F1 Score (%)
Jieba	41.06	30.59	34.84
SnowNLP	32.48	22.01	26.05
THULAC	50.10	42.39	45.75
BERT-base-Chinese	12.81	7.68	9.57

Further, we evaluated several mainstream Chinese word segmentation tools on the dataset; as shown in Table 2, all models performed poorly, with the best achieving an F1 score below 46%—a stark contrast to their >95% performance on standard news corpora [37]. This collapse in tokenization fidelity severely impairs downstream tasks such as intent recognition and semantic restoration.

Additionally, we simply quantify the impact on detection performance via a comparative experiment using the same large language model for adversarial jargon detection on both algorithmic and ADVJARGON. The results, presented in Table 3, show that GPT-4 achieved accuracy rates of 89.60% and 87.43% on the simulated corpora, but only 76.07% on the real-world corpus. Similarly, GPT-3.5-turbo's performance on real-world data was significantly lower. This gap highlights the limitations of models trained exclusively on synthetic data when it comes to detecting the more complex adversarial jargon found in real-world scenarios.

TABLE 3: Identification Performance Comparison on ADVJARGON.

Model	Real-world	Algorithm 1	Algorithm 2
GPT-4o	76.07	89.60	87.43
GPT-3.5-turbo	54.89	70.86	68.57

To address the structural and semantic challenges posed by real-world adversarial jargon, we construct a taxonomy that organizes perturbation patterns across both real and synthetic data, serving as a knowledge base for retrieval and fine-tuning (§5.1). Leveraging the contextual reasoning abilities of large language models (LLMs) demonstrated in prior content security research [30], [31], [38], [39], we introduce a dual enhancement strategy: an external retrieval module grounded in the taxonomy (§5.2) and an internal adaptation via parameter-efficient fine-tuning (§5.3).

4. Measurement study

Through quantitative analysis of both real-world and algorithmic adversarial jargon, we find that large language models perform significantly worse when handling real-world adversarial jargon. This performance gap is primarily due to the linguistic complexity, disruption of semantic coherence, and substantial interference with tokenization mechanisms posed by real-world jargon. Compared to algorithmic jargon, real-world jargon presents higher challenges in terms of linguistic complexity and concealment, requiring more in-depth analysis. Therefore, this section aims to explore the detection and understanding capabilities of LLMs when faced with real-world adversarial jargon texts, focusing on the boundaries of their detection and comprehension abilities.

4.1. Task Definition

This study focuses on three core tasks for Chinese adversarial jargon: Adversarial Jargon Restoration, Jargon Detection, and Illicit Intent Identification, forming a complete process from semantic restoration to intent analysis.

Task 1: Adversarial Jargon Restoration. The goal is to restore adversarially perturbed jargon fragments (x') in text T to their canonical jargon (x) forms. Given a perturbation function δ, which alters surface forms, the task aims to generate a restored text $\hat{T}$ where each adversarial form x'_i is replaced by its canonical form x_i, ensuring the same malicious intent $m \in M$ is preserved ($H(x'_i) = H(x_i) = m$).

Task 2: Adversarial Jargon Detection. This task identifies and locates adversarial jargon forms X' in the original text T. Formally, it is a sequence labeling problem, where the model outputs a label sequence $Y = y_1, ..., y_N$, with $y_i = 1$ if t_i is part of an adversarial jargon form, and $y_i = 0$ otherwise.

Task 3: Illicit Content Detection. After restoration, the final task is to determine whether the text contains illicit intent m related to a set of underground activities M (e.g., gambling, fraud, drug trafficking). Using the restored text $\hat{T}$, the goal is to classify whether it corresponds to a specific illicit intent $m \in M$. This task focuses on identifying potential illegal activity based on the semantic content of the restored text.

4.2. Experimental setup

Model. We conduct testing on 10 popular LLMs, including 6 Chinese models (DeepSeek-R1, DeepSeek-V3, GLM-4-plus, Doubao-1.5-pro, Qwen2.5:7B, ERNIE-4.5) and 4 English models (GPT-4, GPT-3.5-turbo, Claude-Sonnet-4, Llama-3.1-70b). During the model invocation process, we set the temperature to 0 to reduce the randomness of the generated results and enhance the determinism of the output. Additionally, we introduce several simple and general enhancement strategies to evaluate their applicability and effectiveness in adversarial jargon contexts. These strategies include few-shot prompting, multi-round restoration of adversarial texts, and a model routing [40] mechanism that dynamically selects the optimal model based on input features, thereby improving overall detection performance.

Evaluation Metrics. Three metrics are introduced: Illicit content detection, measuring the model's ability to identify underground-related content; Jargon detection rate (De), defined as the fraction of annotated adversarial jargon instances that the model correctly detects; and Jargon restoration rate (Re), defined as the fraction of annotated adversarial jargon instances that the model correctly restores to their canonical forms. In this experiment, illicit content detection metrics before and after restoration are used to assess whether the restoration process enhances downstream

TABLE 4: Performance of different LLMs on Jargon and Underground Semantic Detection.

LLM	Jargon Detection Rate (%)	Illicit content detection Accuracy (%)	Restoration Success Rate (%)	Illicit content detection Accuracy *(after restoration) (%)*	Jargon Detection Rate *(after restoration) (%)*
GPT-4o	88.16	76.05	62.74	88.63	87.82
GPT-3.5-Turbo	79.65	54.89	14.98	83.41	55.89
GLM-4-Plus	91.13	72.44	50.61	92.39	84.21
DeepSeek-V3	90.24	73.07	49.66	91.42	79.03
DeepSeek-R1	86.50	74.42	64.77	87.45	83.45
Doubao	85.79	74.29	62.92	86.60	80.92
Llama-3.1-70B	84.33	57.74	19.58	90.43	76.95
Claude-Sonnet-4	76.91	75.79	58.86	79.52	85.66
ERNIE-4.5	84.04	74.11	55.80	91.96	87.69

content recognition, providing a more accurate measure of the model's understanding of subtle expressions.

Datasets. Based on the real-world spam message corpus constructed in Section 3, we further selected 10,000 manually verified samples to build the detection dataset `sms_text`. These samples meet the following criteria: each text contains multiple adversarial jargon, typically constructed using various obfuscation or transformation techniques to evade conventional detection systems, thereby evaluating the model's detection capability in complex disguise scenarios.

4.3. Results

Table 4 presents the impact of different adversarial techniques on the recognition and restoration rates of LLMs in the sms_text dataset. At the same time, we compared the overall performance of various LLMs in the underground jargon recognition task, with the results shown in Table 4. Based on the comprehensive experimental results, the boundaries and characteristics of LLMs' abilities in jargon recognition and understanding, in relation to the four core research objectives of this section, can be summarized as follows:

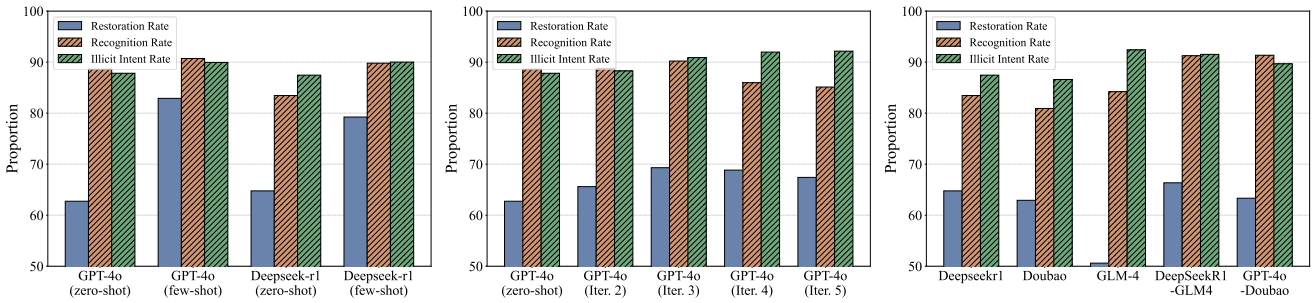
Limited Recognition and Restoration Ability. Although large language models show some recognition capability in underground text detection, their performance remains limited, especially in semantic restoration. As shown in Table 4, the average jargon detection accuracy is about 70%, and the jargon restoration success rate is under 50%, with some models like GPT-3.5-Turbo and Llama-3.1-70B below 20%. This highlights the cognitive limits of LLMs in fully understanding the semantic intent behind jargon.

Boundaries of Specific Adversarial Strategies. As shown in Table 5, different perturbation strategies have a significant impact on the model's recognition and restoration performance. First, large models struggle with Sequence Restructuring perturbations. This type of perturbation alters the order of characters or words in a sentence, completely disrupting the original semantic information. As a result, the model is unable to recognize the relationships between words or restore the original meaning from the context. Experimental results show that under this type of perturbation, both the recognition rate and restoration rate of the model are 0, indicating that it almost completely fails when faced with inputs that severely disrupt the language structure.

Secondly, Symbol and Synonym significantly disrupt the model, with average jargon detection rates below 73% and restoration rates even lower than 35%. Symbol alters the word structure by inserting special symbols, making it difficult for the model to map the word to its standard form. For example, "逋/余" should be recognized as "捕鱼" (fishing), but the symbol breaks the morphological and semantic link, preventing accurate restoration. This disturbance weakens context coherence and disrupts syntactic consistency, making inferences difficult when reliable cues are absent.

Synonym also shows a high failure rate. We suspect this is because the substituted words, such as "捕鲨" (catching sharks) and "打渔" (fishing), are semantically clear, leading the model to assume no restoration is needed. Despite containing adversarial jargon, these texts maintain fluent syntax, within the model's comprehension range. As a result, the model treats them as normal expressions rather than covert information needing restoration. This highlights the model's limitations in recognizing hidden jargon and understanding the task's deeper goals.

Common enhancement strategies also apply to jargon recognition tasks. As shown in Figure 2, by providing a few examples, the jargon restoration rate increased by nearly 20%, with a corresponding improvement in recognition success. Additionally, when using a multi-round iterative strategy, the model's jargon restoration rate peaked after three iterations, and the jargon detection rate steadily increased. In the LLM-routing experiment, we conducted a combined experiment with DeepSeek-R1 as the restoration model and GLM-4-Plus as the recognition model, with DeepSeek-R1 achieving the highest restoration rate. Although GLM-4-Plus had a restoration rate of only 50.61% when used alone, its jargon detection rate reached 84.21%, providing a significant advantage in detection. By combining the capabilities of both models through a routing mechanism, the combined model DeepSeekR1–GLM4 achieved a jargon detection rate of 91.53%, and the other combination, GPT-4o–ERNIE-4.5, reached 89.67%, both surpassing the performance of individual models.



(a) Few-Shot Enhancement. (b) Iteration Enhancement. (c) LLM-routing Technique.

Figure 2: Performance Comparison of Model Enhancement Strategies.

TABLE 5: Comparison of Detection and Restoration Rates (%) of Different LLMs on Multi-Type Jargon Tasks.

LLM	Homophone		Pinyin		English		Visual		Split		Synonym		Semantic Shift		Seq-Restruct		Emoji		Symbol	
	De	Re	De	Re	De	Re	De	Re	De	Re	De	Re	De	Re	De	Re	De	Re	De	Re
DeepSeek-R1	71.74	60.87	87.72	82.46	88.75	84.29	74.19	65.28	85.82	58.16	72.22	38.89	65.97	52.87	0.00	0.00	74.11	61.61	69.52	55.08
DeepSeek-V3	82.10	44.37	82.46	64.91	77.50	62.50	83.28	50.93	78.72	35.46	86.11	16.67	48.69	45.03	0.00	0.00	72.32	44.64	87.17	36.36
GPT-4o	84.78	59.46	94.74	84.21	90.00	87.50	85.07	62.73	92.20	53.19	80.56	25.00	85.86	63.35	0.00	0.00	100	72.73	85.56	35.71
GPT-3.5-turbo	53.71	10.10	68.42	29.82	76.25	50.00	55.90	15.05	79.43	14.29	44.44	13.89	52.36	27.75	0.00	0.00	45.45	18.18	53.48	16.04
GLM-4-Plus	83.63	42.20	91.23	70.18	77.50	52.50	81.66	51.39	93.61	42.55	91.67	36.11	71.72	49.21	0.00	0.00	81.82	54.55	85.03	28.57
Doubao	75.96	57.54	82.46	70.18	77.50	53.75	81.94	64.93	84.40	43.26	58.33	22.22	73.30	60.73	0.00	0.00	71.43	56.25	65.78	47.59
Claude-Sonnet-4	79.41	52.17	91.23	68.42	87.50	61.25	84.32	59.78	78.72	44.68	66.67	16.67	70.16	57.59	0.00	0.00	83.04	58.04	76.47	35.71
Llama-3.1-70B	69.95	16.37	84.21	40.35	73.75	37.50	77.72	18.46	91.49	14.78	77.78	8.33	53.93	24.08	0.00	0.00	76.79	20.54	62.03	13.37
ERNIE-4.5	77.11	47.44	77.19	68.42	86.25	76.25	80.27	57.75	85.82	71.63	66.67	19.44	76.69	59.16	0.00	0.00	86.61	56.25	71.12	48.13

* De stands for Jargon Detection Rate, and Re stands for Jargon Restoration Rate.

4.4. Findings and Discussion

Through a series of experiments, we observed several insightful phenomena:

• *Finding 1: Restoration can enhance the model's ability.* By comparing performance metrics before and after restoration, we found that the restoration of adversarial jargon significantly improved the large language model's sensitivity to underground content detection and its accuracy in recognizing jargon.

• *Finding 2: The Chinese model performs more robustly.* In this evaluation task, Chinese and English models exhibited distinct performance characteristics. The English models showed significant performance divergence: top models such as GPT-4o achieved a jargon detection rate of 87.82% and a restoration rate of 62.74%, while other models, such as GPT-3.5-turbo, showed a notable performance gap, with a restoration rate of just 14.98%. In contrast, Chinese models demonstrated greater overall consistency, maintaining high levels across both illicit activity detection and jargon restoration tasks without significant performance drops. Notably, Chinese models led in key metrics: ERNIE-4.5 topped the illicit content detection accuracy with 91.96%, and DeepSeek-R1 excelled in restoration with a rate of 64.77%. Overall, Chinese models displayed stronger stability and superior performance in this evaluation.

• *Finding 3: Excessive Iterations Lead to Hallucination.* We observed a decline in restoration success rate from the third to the fifth iteration. A detailed analysis of failure cases revealed several hallucination phenomena: *(1) Task Confusion Under Semantic Ambiguity.* When processing vague expressions with pinyin, foreign languages, or mixed Chinese-English, the model may confuse "semantic restoration" with "translation", leading to the restoration of Chinese jargon into English. This phenomenon highlights the model's limited ability to disambiguate tasks, particularly in complex multilingual contexts. *(2) Over-correction.* In the fifth iteration, the model continued to modify already reasonable restoration results, leading to semantic distortion or excessive restoration. This indicates that the model lacks a mechanism for evaluating the optimality of its current output and a robust stopping criterion. In the absence of explicit convergence signals, the model tends to "continue modifying" its output.

5. Methodology

This section presents JADE, an end-to-end pipeline for detecting adversarial Chinese jargon, as illustrated in Figure 3. We first construct and expand a taxonomy that organizes canonical terms and their adversarial variants, forming an information-rich knowledge base. On top of this, we use retrieval-augmented few-shot prompting to fetch a small set of high-quality exemplars per input and apply parameter-efficient fine-tuning to adapt the LLM for normalization and interpretation.

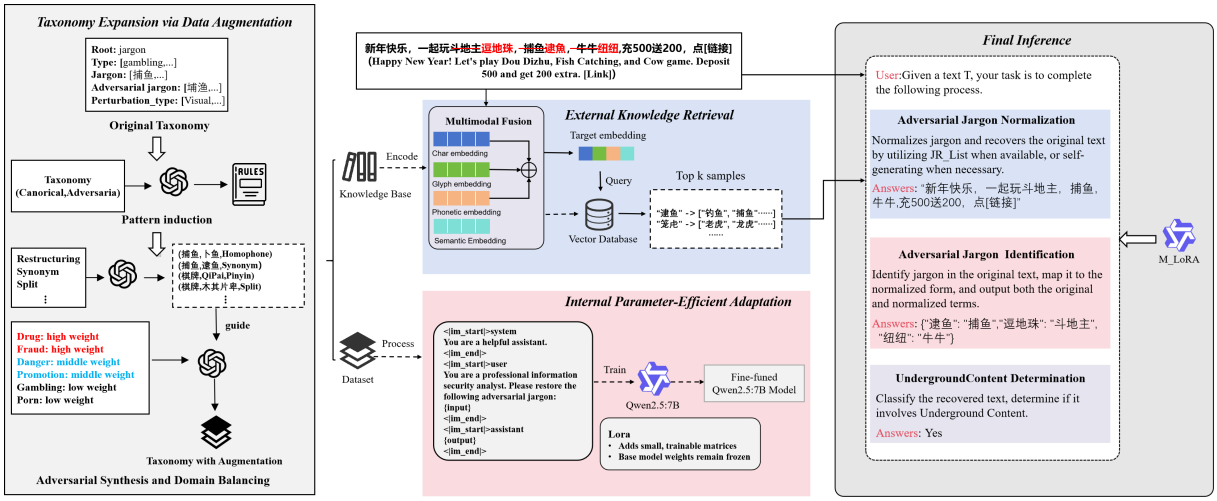


Figure 3: Workflow of JADE.

5.1. Jargon Taxonomy Construction & Expansion

We build a *Jargon Taxonomy* to capture the hierarchy and semantics of underground jargon and expand it via data augmentation. The taxonomy organizes jargon and adversarial variants into domain-specific branches and label the perturbation technique. It serves (i) as a *knowledge base* for RAG, supplying high-precision exemplars for better detection and restoration (Section 5.2); and (ii) as a vehicle for *pattern learning* over techniques and their domain-wise combinations, which drives *LLM-assisted synthesis* and subsequent fine-tuning with both augmented and original samples, improving end-to-end performance (Section 5.3).

5.1.1. Taxonomy Schema. As shown in Figure 4, the taxonomy follows a hierarchical "root–category–instance" structure. The *root* level denotes industry domains (e.g., gambling jargon, drug jargon); the *category* level lists canonical jargon terms (e.g., "杀猪盘" for pig-butchering scams, "捕鱼" for fishing games); and the *instance* level captures adversarial variants and contextual usages (e.g., "逋余," "溜冰," "气狗"), including multiple linguistic transformation patterns such as substitution with Pinyin, Homophone, and Emoji.

• *Jargon Categories*. Using the lexicon $\mathbb{L}$ and annotated ADVJARGON, we group canonical and adversarial jargon into six domains: Gambling, Porn, Drug, Fraud, Danger, and Promotion. We detail each jargon category in Appendix B.

5.1.2. Taxonomy Construction. Formally, the jargon taxonomy is a rooted tree $T_j = (V_j, E_j)$, where V_j is the set of jargon terms and E_j encodes semantic-hierarchical relations (hypernym–hyponym) or perturbation types. We construct T_j based on the canonical lexicon and the annotated adversarial dataset $\mathcal{D} = \{(j_a, j_c, T_{j_a,j_c})\}$, where $T_{j_a,j_c} \subseteq \mathcal{T}$ is the set of perturbation types. Specifically, we instantiate the root layer by the domain categories (e.g., Gambling, Porn); these nodes are fixed and define the top-level partition. For each canonical term $j_c \in \mathbb{L}$, we attach j_c as a child of its domain root (determined by $\mathbb{L}$'s domain label), creating $(\mathrm{root}(j_c) \to j_c)$ in E_j if absent. For each annotated triple $(j_a, j_c, T_{j_a,j_c}) \in \mathcal{D}$, we attach j_a as a child of j_c with an edge label given by its perturbation set; concretely, we add the edge $(j_c \to j_a,\ T_{j_a,j_c})$ to E_j. To preserve the tree structure, we apply string normalization and enforce uniqueness of π so that each j_a has exactly one parent j_c; duplicate adversarial surface forms are merged.

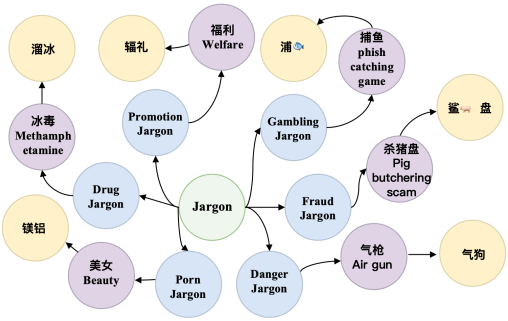


Figure 4: Taxonomy of Chinese jargon categories and examples. The figure illustrates six major types of underground jargons with representative examples — "捕鱼(fishing)" denotes gambling games, where "捕鱼" may also appear in perturbed forms such as "buyu" or "卜渔." "溜冰" refers to taking methamphetamine, a perturbed form of "冰毒" (methamphetamine).

• *Taxonomy Statistics*: The domain-wise counts of canonical and adversarial jargon are summarized in Table 6 (original). The taxonomy is imbalanced: Gambling and Porn dominate, Danger and Promotion sit mid-tier, and Drug and Fraud are sparse. This skew hinders cross-domain generalization and exposure to rare perturbations, motivating targeted augmentation to bolster under-represented domains and improve detection and restoration robustness.

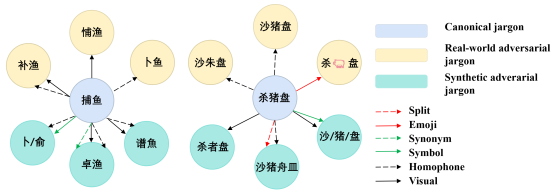


Figure 5: Examples of Real-world and Synthetic Adversarial Jargon. Edges represent different perturbation techniques.

5.1.3. Taxonomy Expansion via Data Augmentation. We augment the taxonomy with *synthetic adversarial jargon* to improve both detection and restoration. The process uses an LLM (GPT-4o) to learn transformation patterns from the existing taxonomy, target under-represented *perturbation techniques* and *domains*, and craft *multi-technique* variants that mirror real combination usage. The synthetic adversarial jargon are illustrated in Figure 5. Below we detail our augmentation strategy.

• *Pattern induction from the taxonomy.* We prompt GPT-4o to study the taxonomy's canonical→adversarial mappings and to summarize perturbation rules (e.g., Homophone, Pinyin, Visual, Split, Emoji, Symbol, Synonym, Semantic Shift, and observed *combinations*). The distilled rules are then used as controllable generators conditioned on a canonical term j_c and a technique set $T \subseteq \mathcal{T}$.

• *Technique- and Domain-guided synthesis.* Guided by evaluation results, we prioritize augmentation for techniques with low detection rates. Specifically, we start with Sequence Restructuring, followed by Character Split, Synonym, and Semantic Shift, and then Symbol and Emoji. For each canonical j_c in the taxonomy, we leverage GPT-4o to produce k variants j_a^{syn} constrained by the chosen technique(s), yielding labeled triples $\left(j_a^{\text{syn}},\, j_c,\, \tau(j_a^{\text{syn}}, j_c)\right)$ with a multi-label indicator for perturbation techniques. We next address domain sparsity based on taxonomy statistics (e.g., Fraud and Drug have far fewer items than Gambling/Porn). Specifically, we up-weight synthesis for under-represented domains to improve the overall domain balance and mitigate data sparsity.

• *Multi-technique composition by domain.* We condition synthesis on domain-specific co-perturbation patterns learned from the taxonomy and sample variants according to each domain's empirical combination frequencies, with extra weight on under-represented techniques to improve coverage. This strategy yields realistic, diverse adversarial forms while preserving recoverability for restoration and downstream training. Full specifications—per-domain technique combinations, sampling heuristics, and weighting rules—are provided in Appendix C.

• *Taxonomy Statistics after Augmentation*: The taxonomy expands substantially across all domains: the well-populated areas (Gambling, Porn) see the largest absolute growth in adversarial variants, while the previously sparse domains (Drug, Fraud) gain the largest relative increases. Overall, the expansion increases the volume of adversarial forms and diversifies perturbation techniques and their combinations, yielding a more balanced domain distribution.

TABLE 6: Statistics of the Jargon Taxonomy.

Category	# canonical	# adversarial	
		(original)	*(expanded)*
Gambling	1755	5991	13976
Porn	2038	6121	15053
Danger	486	1470	5516
Drug	92	288	1526
Promotion	461	1453	5379
Fraud	60	239	1219

Below, we detail how this expanded coverage enhances RAG with high-quality exemplars and supplies targeted supervision for parameter-efficient fine-tuning, thereby improving both detection and restoration.

5.2. External Knowledge Retrieval

Few-shot prompts markedly improve LLM detection of underground jargon (see Section 4.3). Prompt quality is therefore crucial but creates a performance–efficiency trade-off: more relevant exemplars raise accuracy, while larger prompts increase cost. To balance this, we adopt a lightweight retrieval-augmented approach that leverages Chinese's phonetic, glyph, and semantic signals to retrieve a small set of the most similar exemplars, boosting detection under low-resource constraints.

5.2.1. Multi-Granularity Embedding. Chinese integrates phonetic (pinyin), glyph, and semantic cues [41], which complicate jargon understanding for LLMs. We therefore build a unified representation that fuses *character-level* pinyin/glyph features with *sentence-level* semantics to support high-precision retrieval.

• *Knowledge base.* Using the taxonomy from Section 5.1 as a dynamic knowledge base K, each record is a tuple $(j_c, J_a, T_{J_a,j_c}, S_{\text{exp}})$, where j_c is a canonical jargon term, $J_a = \{j_{a_1}, j_{a_2}, \ldots\}$ are its adversarial variants, $T_{J_a,j_c} \subseteq \mathcal{T}$ denotes the associated perturbation type(s) for each pair $(j_a, j_c), j_a \in J_a$ (e.g., Pinyin, Homophone, Symbol, Emoji), and S_{exp} are context examples in which these variants occur (e.g., an SMS message).

• *Character-level fusion.* Following ChineseBERT [42] (CBERT), we enrich a pretrained model with character, pinyin, and glyph channels. For a token sequence x, we obtain character embeddings $\text{CE}(x)$, pinyin embeddings $\text{PE}(x)$, and glyph embeddings $\text{GE}(x)$ and fuse them as

$$\text{CBERT}(x) = \text{Fusion}\big(\text{CE}(x), \text{PE}(x), \text{GE}(x)\big), \quad (1)$$

which captures fine-grained differences crucial for homophone substitutions and visual confusables.

• *Sentence-level semantics.* To model cross-token dependencies and full-sentence meaning, we use SBERT (SBERT-base-chinese-nli) [43]. For an input sentence s, $\text{SBERT}(s) \in \mathbb{R}^d$ provides a semantic embedding suitable for large-scale matching.

• *Index construction.* For each adversarial variant $j_a \in J_a$, we compute its character-level embedding $\mathrm{CBERT}(j_a)$ via (1), select one or more usage examples $s_{\text{exp}} \in S_{\text{exp}}$, and encode them with SBERT. We then build a quadruple index

$$\big(j_a,\ \mathrm{CBERT}(j_a),\ \mathrm{SBERT}(s_{\text{exp}}),\ j_c\big),$$

linking each adversarial variant to both its surface-form embedding and its canonical form via context semantics.

• *Sample Retrieval and Context Refinement.* Given an input text T_{input}, we first obtain its character-level embedding $\mathrm{CBERT}(T_{\text{input}})$ and compute cosine similarity to each adversarial-variant embedding:

$$D(j_a) \ = \ \mathrm{sim}\big(\mathrm{CBERT}(T_{\text{input}}),\ \mathrm{CBERT}(j_a)\big). \quad (2)$$

We select the top-k candidates $J_R^k = \{\, j_a \mid j_a \in \text{Top-}k \text{ in } D(j_a) \,\}$. Word-level similarity alone can yield false positives due to contextual ambiguity. We therefore apply *context-aware re-ranking*: for each $j_a \in J_R^k$, we locate its sentence in the input, S_{input}, encode it as $\mathrm{SBERT}(S_{\text{input}})$, and compare against the canonical term's malicious-context exemplars $\{\mathrm{SBERT}(s) : s \in S_{\text{exp}}\}$ to compute a context relevance score. We then re-rank and filter:

$$J_{\text{final}} \ = \ \{\, j_a \in J_R^k \ \mid \ \mathrm{Score}(j_a) \geq \tau \,\}, \quad (3)$$

yielding a small, high-precision exemplar set for few-shot prompting that balances accuracy and efficiency.

5.3. Internal Parameter-Efficient Adaptation

To enable domain-specific reasoning beyond retrieval, we apply parameter-efficient fine-tuning with LoRA. Using a pretrained instruction-tuned model, we fine-tune on a supervised corpus of adversarial jargon and their canonical forms. LoRA injects lightweight updates while keeping the base model frozen to reduce cost. Fine-tuning uses a standardized prompt format that instructs the model to normalize adversarial jargon and align the corresponding spans in the original text, thereby enhancing the model's ability to interpret and restore obfuscated jargon.

5.3.1. Data Preprocessing. We first construct a dataset suitable for LLM fine-tuning. Using the taxonomy from §5.1 and manual curation, we assemble $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$, where x_i is an adversarial jargon instance and y_i is its normalized (canonical) text. For instruction-tuned models such as `Qwen2.5-7B-Instruct`, we employ a unified prompt template that elicits structured outputs. The template uses special markers <|im_start|> ... <|im_end|> to delineate system instructions, user input, and model output. The system role casts the model as a "jargon normalization expert" to steer semantic reasoning; the user input provides the target text x_i; and the model is guided to produce the normalized form y_i. An example is shown in Appendix E.

5.3.2. Low-Rank Adaptation. We adopt Low-Rank Adaptation (LoRA) [44] which freezes the pretrained weight matrix $W \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$ and injects a trainable low-rank update via two small matrices A and B:

$$\Delta W \ = \ BA,\ A \in \mathbb{R}^{r \times d_{\text{in}}},\ B \in \mathbb{R}^{d_{\text{out}} \times r},\ r \ll \min(d_{\text{in}}, d_{\text{out}}).$$

The forward pass becomes

$$h' \ = \ Wx \ + \ \Delta W x \ = \ Wx \ + \ BAx.$$

During fine-tuning, only A and B are updated, while W remains frozen. This reduces trainable parameters from $d_{\text{in}} \times d_{\text{out}}$ to $r(d_{\text{in}} + d_{\text{out}})$, substantially lowering compute and memory costs. The adapted effective weights are $W' = W + \Delta W$. Trained in this manner, the model learns the mapping between adversarial jargon and its canonical realizations, inducing a low-dimensional subspace specialized for recognition and normalization.

5.4. Final Inference: End-to-End Detection

Given input text T_{input}, the external module retrieves the most semantically relevant normalization candidates for each suspected covert span $\{x_1, \ldots, x_n\}$, yielding

$$J_R^{\text{List}} \ = \ \big\{\, j_c^{(i,v)} \,\big|\, j_c^{(i,v)} \in J_{\text{final}}^{(i)} \big\}, \quad (4)$$

where $j_c^{(i,v)}$ denotes the v-th candidate *canonical* form for the i-th adversarial span x_i. To address the complexity of recognition and normalization, we employ a three-stage chain-of-thought (CoT) protocol, and formalize the overall prediction as

$$\hat{y} \ = \ M_{\text{LoRA}}\big(P_{\text{CoT}},\ T_{\text{input}},\ J_R^{\text{List}}\big)\,.$$

Crafting JADEC-CoT Prompts. We organize final inference into three stages to convert obfuscated spans to canonical forms, align them to the source text, and produce a document-level risk decision.

• *Stage 1: Adversarial Jargon Normalization.* The model M_{LoRA} first consults the candidate sets $\{j_c^{(i,v)}\}_{v=1}^k$ provided by retrieval. If a set contains plausible options, the model selects the best normalization per span based on context; if a set is empty or inappropriate, the model generates a normalization. This requests a fully normalized output where all adversarial jargon is replaced with canonical forms.

• *Stage 2: Jargon Localization and Semantic Alignment.* Given the original and normalized text, the model identifies the adversarial spans and aligns each to its canonical form.

• *Stage 3: Underground-Content Determination.* The model performs document-level intent analysis over predefined domains (e.g., porn, fraud) and outputs a final risk label.

The prompts Q1, Q2, and Q3 for the three stages are detailed in Appendix D.

Executing JADEC-CoT. We employ a LoRA-adapted LLM M_{LoRA} to execute the prompts and serialize the decision process into interpretable intermediate outputs A_1, A_2, and A_3, yielding an auditable reasoning trail. Given the input text T_{input} and the retrieved candidate lists J_R^{List}, the model

first produces a fully normalized version of the text, where obfuscated spans are replaced with their canonical forms:

$$A_1 \;=\; \arg\max_{a} \; \Pr\big(a \,\big|\, T_{\text{input}},\, J_R^{\text{List}},\, Q_1\big).$$

Next, using the original input and the normalized output A_1, the model identifies which specific spans in the text correspond to adversarial jargon and aligns each to its corresponding canonical form:

$$A_2 \;=\; \arg\max_{b} \; \Pr\big(b \,\big|\, T_{\text{input}},\, A_1,\, Q_2\big).$$

Then, conditioned on both the normalized text A_1 and the alignment result A_2, the model determines the overall risk status of the document and classifies its content according to predefined underground domains:

$$A_3 \;=\; \arg\max_{c} \; \Pr\big(c \,\big|\, A_1,\, A_2,\, Q_3\big).$$

Finally, the model produces a binary prediction label based on whether A_3 flags an illicit intent:

$$\hat{y} \;=\; \begin{cases} \text{Illicit}, & \text{if } A_3 \text{ flags an underground domain,} \\ \text{Benign}, & \text{otherwise.} \end{cases}$$

This structured protocol ensures thorough coverage (span-level and document-level) and provides an explicit, inspectable chain of reasoning from retrieved candidates to final decision.

6. Evaluation of JADE

In this section, we systematically evaluate the performance of JADE, including comparisons with other baseline models and transferability on downstream tasks.

6.1. Experimental Setting

Datasets. We first use the Jargon Taxonomy constructed in §5.1 to generate and annotate data, which is used as the training set. Then, for independent evaluation, we construct a balanced test set. This test set consists of two parts: first, we selected 5,000 SMS messages containing adversarial jargon from the annotated dataset in §3.1, and secondly, we selected 5,000 SMS messages without adversarial jargon from the FBS_SMS_Dataset [45] to form a balanced dataset of spam messages. This dataset is used as the test set. The test set contains a total of 13,440 adversarial jargon instances.

Baselines. We compare the proposed method with a series of advanced baseline models, including ChineseBERT [42], RoCBERT [10], BERT [46], and MacBERT [47]. Among these models, ChineseBERT performs pretraining by combining character and phonetic features; RoCBERT focuses on Chinese semantic, phonetic, and visual features; BERT is a general baseline widely used in various NLP tasks; and MacBERT adopts the MLM-As-Correlation pretraining strategy and incorporates the Sentence Order Prediction task to improve model performance.

Evaluation metrics. Four metrics are introduced for the evaluation tasks: (1) Jargon Detection Rate, which measures the model's accuracy in identifying adversarial jargon; (2) Jargon Restoration Rate, which evaluates the model's ability to restore adversarial jargon; (3) Accuracy, which measures the overall correctness of the model's predictions across all categories; and (4) Recall, which specifically evaluates the model's ability to identify underground industry content.

Implementation. For multimodal feature extraction, we used the official source code and pretrained weights released by ChineseBERT [42].

6.2. Evaluation Results

Table 7 shows the performance of the proposed JADE in defending against adversarial jargon designed for the underground industry, and compares it with all the benchmark methods, leading to the following conclusions:

Detection Performance: First, we compared the performance of the JADE framework with other baseline models in the detection task. Most baseline models performed poorly when faced with adversarial jargon. For example, the standard BERT model achieved only 65.12% in Jargon Detection Rate and 79.80% in Accuracy. We speculate that these general-purpose models lack sufficient robustness when handling adversarial samples from specific domains. In contrast, the JADE framework significantly outperformed all baseline models in all detection tasks, achieving a Jargon Detection Rate of 98.59% and Accuracy of 96.92%. Furthermore, JADE achieved a Recall of 97.99%, far exceeding the second-best model, RoCBERT, which scored 92.99%, demonstrating its exceptional recognition ability in adversarial environments.

Restoration Performance: Next, we evaluated the performance of each model in the Jargon Restoration task. The results show that all baseline models performed poorly in jargon restoration, with the best-performing model, RoCBERT, achieving only a 68.70% jargon restoration rate. This validates the limitations of general-purpose language models in restoring carefully designed adversarial perturbations, particularly when dealing with confusion text from specific domains. In contrast, the JADE framework performed excellently in this task, achieving a Jargon Restoration Rate of 95.91%, surpassing RoCBERT's performance by 27%. This indicates that JADE can effectively restore adversarial perturbations, significantly improving the restoration results. In summary, JADE has a significant advantage in both jargon detection and restoration tasks, providing a more robust defense for underground industry content detection and demonstrating its exceptional practicality in complex adversarial environments.

Supplementary Experiments. To directly compare the performance of the JADE framework with the current state-of-the-art general-purpose LLMs, we conducted a supplementary experiment. We evaluated JADE using the `sms_text` dataset, which was also used in Section 3 to assess the capability boundaries of LLMs. The performance of JADE was compared with the top-performing LLMs, GPT-4o and

TABLE 7: Effectiveness of Fine-tuning.

Model	Jargon Detection Rate	Jargon Restoration Rate	Accuracy	Recall
ChineseBERT	0.7455	0.6667	0.8574	0.8969
RoCBERT	0.7557	0.6870	0.8993	0.9299
BERT	0.6512	0.5930	0.7980	0.8479
MacBERT	0.7365	0.6689	0.8837	0.9127
JADE	**0.9859**	**0.9591**	**0.9692**	**0.9799**

DeepSeek-R1, across three key metrics: Jargon Detection Rate, Jargon Restoration Rate, and Illicit Content Detection Rate. The experimental results show that JADE achieves 95.96% in jargon restoration, 98.66% in jargon detection, and 97.64% in illicit content detection, significantly outperforming both GPT-4o and DeepSeek-R1.

6.3. Ablation Study

To validate the effectiveness and synergy of the two core enhancement paradigms in JADE, we conducted ablation experiments. The framework is built on a large language model, and we assess the independent contributions and combined benefits of each strategy for jargon detection and restoration tasks through five configurations: (1) Base LLM: A base LLM without fine-tuning or RAG. (2) External Knowledge Retrieval: a non-fine-tuned LLM with an external knowledge base. (3) Internal Adaptation: Fine-tuned base LLM with our jargon training set, but without an RAG knowledge base. (4) JADE (No Data Augmentation): A model using a dataset without data augmentation, combining Internal Adaptation and External Knowledge Retrieval. (5) JADE: Our proposed method, which uses a data-augmented training set and integrates both Internal Adaptation and External Knowledge Retrieval.

The ablation study results, summarized in Table 8, reveal several key findings. First, the Base LLM performs the worst, highlighting the challenges of applying a general-purpose LLM to illicit content detection without task-specific guidance. Second, External Knowledge Retrieval improves accuracy by leveraging phonetic, morphological, and semantic information for better adversarial text retrieval. Third, Internal Adaptation enhances detection by learning domain-specific language patterns. Finally, Data Augmentation improves robustness by introducing diverse perturbations, helping the model better identify and restore adversarial jargon. Notably, perturbations that challenge large models further enhance performance in complex scenarios.

6.4. Generalization of JADE

To further assess JADE's generalization, we conduct three complementary experiments, examining its robustness under incomplete taxonomy coverage, its temporal generalization on newly collected real-world data, and its transferability to related downstream tasks.

Robustness under incomplete taxonomy coverage. To evaluate how incomplete taxonomy coverage affects JADE, we conduct a pruning experiment based on scaled taxonomies. More specifically, we randomly remove taxonomy entries and re-evaluate the full system under the resulting reduced taxonomies. As shown in Table 9, JADE remains largely robust when 10%, 30%, and 50% of taxonomy entries are removed. These results suggest that JADE can generalize from partial structured knowledge and does not rely heavily on complete taxonomy coverage.

Temporal generalization on new real-world data. We next evaluate JADE on a newly constructed dataset built from real-world spam SMS messages collected in 2025, whereas the data used for taxonomy construction and fine-tuning were collected between 2021 and 2024. Specifically, we randomly sample 500 spam messages from the 2025 collection, annotate the adversarial jargon and their canonical forms, and combine them with 500 benign SMS messages from a public dataset to construct a new test set of 1,000 messages. During annotation, we identify previously unobserved domains, such as *traffic referral*, which suggests a distribution shift relative to the original data. On this new test set, JADE achieves a detection rate of 97.90% and a restoration rate of 94.95%, suggesting that it generalizes well to more recent real-world data.

Transferability to related downstream tasks. We further evaluate JADE using two publicly available datasets to assess its robustness and generalization: (1) ToxiCloakCN [15], a dataset for studying toxic content camouflage in Chinese. It simulates camouflage perturbations such as homophone and emoji substitutions. (2) ChiFraud [16], a large-scale Chinese fraud text detection benchmark, consists of 59,106 fraud and 352,328 benign texts scraped from web data over 12 months. Following prior work, we treat both datasets as binary classification tasks. We evaluate model performance using Accuracy, Precision, Recall, and F1 score, with the F1 score as the primary metric. As shown in Table 10, on the ChiFraud dataset, JADE achieves an F1-score of 0.9433, surpassing the previously reported ChineseBERT model, which has an F1-score of 0.9331 [16]. On ToxiCloakCN, JADE achieves an F1-score of 0.9325, significantly outperforming the GPT-4o model, which has an F1-score of 0.796. These results suggest that JADE generalizes effectively to related adversarial Chinese text tasks.

7. Discussion

The proposed JADE framework significantly improves the detection and understanding of adversarial jargon by integrating Internal Adaptation and External Knowledge

TABLE 8: Comparison of Different Enhancement Methods

Model	Jargon Detection Rate	Jargon Restoration Rate	Accuracy	Recall
Base LLM (Qwen2.5:7B)	0.3342	0.0978	0.7523	0.7914
External Knowledge Retrieval	0.8830	0.7422	0.8642	0.8780
Internal Adaptation	0.8996	0.8441	0.8781	0.8902
JADE (without data augmentation)	0.9006	0.8416	0.8933	0.9127
JADE	**0.9859**	**0.9591**	**0.9692**	**0.9799**

TABLE 9: Performance of JADE under incomplete taxonomy coverage.

Removed Entries	Performance	
	Detection (%)	Restoration (%)
10%	98.13	95.46
30%	97.57	95.04
50%	95.43	89.46

TABLE 10: Transferability of JADE to downstream tasks compared with baseline models.

Dataset	Model	F1 Score
ChiFraud [16]	ChineseBert [16]	0.9331
	JADE (Ours)	**0.9433**
ToxiCloakCN [15]	GPT-4o	0.7960
	JADE (Ours)	**0.9325**

Retrieval. This section outlines the limitations of JADE and explores potential areas for improvement.

Inference Efficiency and Cost. JADE relies on a complex multi-stage reasoning process. While this ensures high accuracy and interpretability, it also introduces significant delays and computational overhead, posing challenges for real-time processing and large-scale applications. Future work should explore model distillation and quantization techniques to develop a lightweight framework, achieving a better balance between effectiveness and efficiency.

Taxonomy Coverage and Robustness Challenges. The core of JADE is its adversarial jargon taxonomy, which serves as both the knowledge base for RAG retrieval and the data foundation for internal enhancement tuning. Although we expanded the taxonomy's coverage through large-scale data augmentation, the initial construction still relies on existing dictionaries and expert annotations. As a result, JADE may face robustness challenges when confronted with new, under-annotated types of perturbations.

Taxonomy Maintenance. This dependence on taxonomy coverage also raises a practical maintenance question for long-term deployment. As underground communities continuously evolve and develop new evasion strategies, maintaining the taxonomy over time is important for sustaining JADE's performance in practice. One possible maintenance strategy is to use JADE to surface high-confidence emerging jargon candidates, which can then be manually validated to determine their perturbation types and canonical forms before being incorporated into the taxonomy. The pruning results in Section 6.4 provide some support for the feasibility of such a workflow: JADE maintains consistently high performance even when up to 30% of taxonomy entries are removed, suggesting that frequent taxonomy updates may not be necessary until emerging jargon accumulates to a substantial fraction of the original taxonomy. This observation indicates that the update cycle may be relatively infrequent in practice, which could help keep manual maintenance effort and retraining cost manageable. A concrete design and evaluation of such a maintenance workflow, as well as robustness against fully interactive adaptive attackers who iteratively probe and refine their jargon, remain important directions for future work.

Error Analysis. While JADE achieves strong overall performance, certain failure modes warrant attention, particularly in sensitive content moderation settings where false positives can be more damaging than false negatives. In our evaluation on 500 benign samples, JADE misclassified 12 instances, yielding a false positive rate of 2.4%. False positives primarily occur in short or context-poor messages where the surrounding text provides weak intent signals, causing benign content to superficially resemble adversarial jargon. False negatives, on the other hand, tend to arise when high-frequency benign words are repurposed as adversarial variants and embedded in otherwise normal contexts, making them difficult to distinguish from legitimate usage. Addressing these failure modes, particularly by incorporating richer contextual signals and improving disambiguation of repurposed common vocabulary, remains an important direction for future work.

Limitations in Cross-Modal Adversarial Detection. Currently, JADE focuses on text detection and effectively captures information at the character, pronunciation, and semantic levels. However, its capabilities are limited to "characters." In practice, adversarial attacks have gone beyond text, with attackers embedding jargon or illicit information in images, memes, QR codes, and other visual content to evade detection. Therefore, extending JADE's capabilities to the multi-modal domain and developing a detection framework that understands both text and visual information is a key direction for future research.

8. Conclusion

In this paper, we addressed the challenge of detecting and restoring evolving adversarial Chinese underground jargon in online ecosystems. To this end, we constructed the first large-scale in-the-wild taxonomy to characterize real-world evasion techniques and proposed JADE, an LLM-based framework. JADE integrates internal and external

knowledge augmentation through multi-granularity exemplar retrieval from a dynamic knowledge base and chain-of-thought reasoning, thereby significantly improving the model's ability to understand and detect complex jargon perturbations. Extensive experiments demonstrate the effectiveness of our approach, achieving 95.91% accuracy in restoration and 98.59% in detection, with strong transferability to downstream tasks. Future work will explore its adaptability in multi-lingual or multi-modal adversarial contexts to further strengthen content moderation systems.

LLM Usage Considerations

In this study, Large Language Models (LLMs) were employed, playing a key role in generating synthetic adversarial jargon as well as in adversarial jargon detection and restoration tasks. Below are the considerations regarding the use of LLMs in our research, adhering to established ethical and transparency guidelines.

Originality. The authors take full responsibility for the content of this paper. Although LLMs were used during the research process, we emphasize that all generated outputs were thoroughly reviewed and validated by the authors to ensure their accuracy and originality. Specifically, LLMs were used to assist in generating synthetic adversarial jargon samples, but all content was independently evaluated by the authors, and proper credit was given to prior research in the literature review. Additionally, we also leveraged the capabilities of large language models in term recognition, particularly in handling complex adversarial texts. Nevertheless, all content generated or recognized by large models was carefully reviewed to ensure accuracy and originality.

Transparency. One of the reasons for using LLMs in this research was to facilitate the creation of synthetic adversarial jargon, which is a crucial part of our data augmentation and adversarial detection framework. At the same time, we also utilized large models to improve the accuracy of adversarial jargon recognition, particularly in detecting underground industry jargon and transformed terms. While the content generated by LLMs and the application of large models in recognition provided valuable input, all ideas and methodologies were independently developed by the authors, and all outputs were subject to rigorous validation and refinement. Furthermore, we carefully considered the limitations of using LLMs. For example, although large models played a key role in improving recognition accuracy, they still face challenges when handling domain-specific terminology, which could affect reproducibility. To address these issues, we made extensive efforts to cross-check and validate the LLM outputs with existing data to minimize inaccuracies.

Responsibility. We are highly conscientious in using LLMs, ensuring that the data used to train and generate synthetic adversarial jargon complies with ethical standards, including consent and intellectual property rights. The application of large models in recognition tasks was similarly considered in terms of its environmental footprint, and we took appropriate measures to minimize this impact. This included selecting suitable model sizes, optimizing query efficiency, and limiting unnecessary computations to reduce resource consumption. We are fully aware of the environmental impact and strive to conduct our research in an environmentally responsible manner. At the same time, regarding the use of large models, we ensured that the experimental design adhered to ethical guidelines and made efforts to minimize resource waste during the model training process.

Ethics Considerations

This study strictly followed the ethical principles outlined in the Belmont Report [48] and the Menlo Report [49] to minimize potential ethical risks in data collection and experiments. The primary focus of this work is the detection and restoration of adversarial jargon in the context of underground market activities.

To study real-world adversarial boundaries and evaluate the effectiveness of JADE, we established a compliant collaboration with a leading security vendor, 360 Mobile Security, and collected real-world spam SMS data under the supervision of the vendor's legal department. Although these messages had already been identified as spam, the collaborating company anonymized potentially sensitive Personally Identifiable Information (PII) before providing the data to the research team. For example, sensitive fields such as personal names were de-identified using salted hashing, ensuring that the researchers did not directly access sensitive information from real users.

In addition, the datasets used in Section 6.4 for toxic content detection and fraud detection are publicly available datasets. Before using them, we randomly sampled 100 instances from each dataset for manual inspection to check that they did not contain obvious PII or other sensitive information.

Finally, to support security research on underground content detection, we plan to release the annotated adversarial jargon dataset. Before release, we manually review the data to reduce the risk of disclosing personal sensitive information and other sensitive content that may introduce unnecessary ethical risks.

Overall, all experiments in this work were conducted in accordance with relevant ethical guidelines for data usage and research conduct. We took reasonable steps to reduce risks related to privacy exposure and potential misuse. To the best of our knowledge, this study did not involve direct interaction with human subjects.

Acknowledgments

We sincerely thank all anonymous reviewers and our shepherd for their valuable and constructive comments on improving the paper. This work is supported by Zhongguancun Laboratory.

References

[1] S. Wu, J. Xue, S. Zhou, and X. Mi, "Reflected search poisoning for illicit promotion," *arXiv preprint arXiv:2404.05320*, 2024.

[2] Z. Li and X. Liao, "Understanding and analyzing appraisal systems in the underground marketplaces." in *NDSS*, 2024.

[3] G. Hong, Z. Yang, S. Yang, X. Liaoy, X. Du, M. Yang, and H. Duan, "Analyzing ground-truth data of mobile gambling scams," in *2022 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2022, pp. 2176–2193.

[4] W. Zhu, H. Gong, R. Bansal, Z. Weinberg, N. Christin, G. Fanti, and S. Bhat, "Self-supervised euphemism detection and identification for content moderation," in *2021 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2021, pp. 229–246.

[5] K. Yuan, H. Lu, X. Liao, and X. Wang, "Reading thieves' cant: automatically identifying and understanding dark jargons from cybercrime marketplaces," in *27th USENIX Security Symposium (USENIX Security 18)*, 2018, pp. 1027–1041.

[6] L. Ke, X. Chen, and H. Wang, "An unsupervised detection framework for chinese jargons in the darknet," in *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, 2022, pp. 458–466.

[7] S. Yang, S. Cui, C. Hu, H. Wang, T. Zhang, M. Huang, J. Lu, and H. Qiu, "Exploring multimodal challenges in toxic chinese detection: Taxonomy, benchmark, and findings," 2025.

[8] M. Song, E. Jang, J. Kim, and S. Shin, "Covering cracks in content moderation: Delexicalized distant supervision for illicit drug jargon detection," *arXiv preprint arXiv:2503.14926*, 2025.

[9] J. Li, T. Du, S. Ji, R. Zhang, Q. Lu, M. Yang, and T. Wang, "TextShield: Robust text classification based on multimodal embedding and neural machine translation," in *USENIX Security Symposium*, 2020.

[10] H. Su, W. Shi, X. Shen, Z. Xiao, T. Ji, J. Fang, and J. Zhou, "Rocbert: Robust chinese bert with multimodal contrastive pretraining," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022, pp. 921–931.

[11] J. Li, T. Du, X. Liu, R. Zhang, H. Xue, and S. Ji, "Enhancing model robustness by incorporating adversarial knowledge into semantic representation," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 7708–7712.

[12] H. Yang, X. Ma, K. Du, Z. Li, H. Duan, X. Su, G. Liu, Z. Geng, and J. Wu, "How to learn klingon without a dictionary: Detection and measurement of black keywords used by the underground economy," in *2017 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2017, pp. 751–769.

[13] J. Yu and Z. Li, "Chinese spelling error detection and correction based on language model, pronunciation, and shape," in *Proceedings of The Third CIPS-SIGHAN Joint Conference on Chinese Language Processing*, 2014, pp. 220–223.

[14] Z. Zhang, M. Liu, C. Zhang, Y. Zhang, Z. Li, Q. Li, H. Duan, and D. Sun, "Argot: Generating adversarial readable chinese texts," in *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, C. Bessiere, Ed. ijcai.org, 2020, pp. 2533–2539. [Online]. Available: https://doi.org/10.24963/ijcai.2020/351

[15] Y. Xiao, Y. Hu, K. T. W. Choo, and K. W. Lee, "Toxicloakcn: Evaluating robustness of offensive language detection in chinese with cloaking perturbations," *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 6012–6025, 2024.

[16] M. Tang, L. Zou, Z. Jin, S. Cui, S. N. Liang, and W. Wang, "Chifraud: A long-term web text dataset for chinese fraud detection," in *Proceedings of the 31st International Conference on Computational Linguistics*, 2025, pp. 5962–5974.

[17] D. Jin, Z. Jin, J. T. Zhou, and P. Szolovits, "Textfool: Fool your model with natural adversarial text," 2019.

[18] Y. Zhang, "Adversarial feature matching for text generation," 2017.

[19] X. Wang, J. Hao, Y. Yang, and K. He, "Natural language adversarial defense through synonym encoding," in *Uncertainty in Artificial Intelligence*. PMLR, 2021, pp. 823–833.

[20] Y. Xiao, Y. Hu, K. T. W. Choo, and R. K.-w. Lee, "Toxicloakcn: Evaluating robustness of offensive language detection in chinese with cloaking perturbations," *arXiv preprint arXiv:2406.12223*, 2024.

[21] T. Gröndahl, L. Pajola, M. Juuti, M. Conti, and N. Asokan, "All you need is" love" evading hate speech detection," in *Proceedings of the 11th ACM workshop on artificial intelligence and security*, 2018, pp. 2–12.

[22] P. Aggarwal and T. Zesch, "Analyzing the real vulnerability of hate speech detection systems against targeted intentional noise," in *Proceedings of the Eighth Workshop on Noisy User-generated Text (W-NUT 2022)*, 2022, pp. 230–242.

[23] T. Le, Y. Ye, Y. Hu, and D. Lee, "Cryptext: Database and interactive toolkit of human-written text perturbations in the wild," in *2023 IEEE 39th International Conference on Data Engineering (ICDE)*. IEEE, 2023, pp. 3639–3642.

[24] Y. Ye, T. Le, and D. Lee, "Noisyhate: Mining online human-written perturbations for realistic robustness benchmarking of content moderation models," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 19, 2025, pp. 2603–2612.

[25] P. Mishra, H. Yannakoudakis, and E. Shutova, "Neural character-based composition models for abuse detection," *arXiv preprint arXiv:1809.00378*, 2018.

[26] Y. Hou, H. Wang, and H. Wang, "Identification of chinese dark jargons in telegram underground markets using context-oriented and linguistic features," *Information Processing & Management*, vol. 59, no. 5, p. 103033, 2022.

[27] D. Seyler, W. Liu, Y. Zhang, X. Wang, and C. Zhai, "Darkjargon. net: A platform for understanding underground conversation with latent meaning," in *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2021, pp. 2526–2530.

[28] D. Seyler, W. Liu, X. Wang, and C. Zhai, "Towards dark jargon interpretation in underground forums," in *European Conference on Information Retrieval*. Springer, 2021, pp. 393–400.

[29] Y. Ma, X. Shen, Y. Qu, N. Yu, M. Backes, S. Zannettou, and Y. Zhang, "From meme to threat: On the hateful meme understanding and induced hateful content generation in open-source vision language models," in *USENIX Security Symposium (USENIX Security). USENIX*, 2025.

[30] Y. Zhuang, K. Guo, J. Wang, Y. Jing, X. Xu, W. Yi, M. Yang, B. Zhao, and H. Hu, "I know what you meme! understanding and detecting harmful memes with multimodal large language models." in *NDSS*, 2025.

[31] H. Lin, Z. Luo, J. Ma, and L. Chen, "Beneath the surface: Unveiling harmful memes with multimodal reasoning distilled from large language models," *arXiv preprint arXiv:2312.05434*, 2023.

[32] TesterCC, "Black market: An in-depth security research," n.d., accessed: 2025-11-10. [Online]. Available: https://testercc.github.io/sec_research/black_market/

[33] ThreatHunter, "Combating fraud: A quick overview of black market jargon (credit fraud edition)," n.d., accessed: 2025-11-10. [Online]. Available: https://www.threathunter.cn/blog/cbf9ac3d76b

[34] ThreatHunter, "Combating fraud: A quick overview of black market jargon (marketing fraud edition)," n.d., accessed: 2025-11-10. [Online]. Available: https://www.threathunter.cn/blog/f2ed57db77a?categoryId=81432

[35] ThreatHunter, "Combating fraud: A quick overview of black market jargon (data breach edition)," n.d., accessed: 2025-11-10. [Online]. Available: https://www.threathunter.cn/blog/110239f5e52

[36] A. Strauss and J. Corbin, *Basics of qualitative research*. Sage publications, 1990.

[37] T. Emerson, "The second international chinese word segmentation bakeoff," in *Proceedings of the fourth SIGHAN workshop on Chinese language Processing*, 2005.

[38] Y. Li, C. Huang, S. Deng, M. L. Lock, T. Cao, N. Oo, H. W. Lim, and B. Hooi, "{KnowPhish}: Large language models meet multimodal knowledge graphs for enhancing {Reference-Based} phishing detection," in *33rd USENIX Security Symposium (USENIX Security 24)*, 2024, pp. 793–810.

[39] K. Thomas, P. G. Kelley, D. Tao, S. Meiklejohn, O. Vallis, S. Tan, B. Bratanič, F. T. Ferreira, V. K. Eranti, and E. Bursztein, "Supporting human raters with the detection of harmful content using large language models," in *2025 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2025, pp. 2772–2789.

[40] I. Ong, A. Almahairi, V. Wu, W.-L. Chiang, T. Wu, J. E. Gonzalez, M. W. Kadous, and I. Stoica, "Routellm: Learning to route llms with preference data, 2024," *URL https://arxiv. org/abs/2406.18665*, vol. 4, 2025.

[41] Y. Chi, F. Giunchiglia, C. Li, and H. Xu, "Ancient chinese glyph identification powered by radical semantics," in *Findings of the Association for Computational Linguistics ACL 2024*, 2024, pp. 12 065–12 074.

[42] Z. Sun, X. Li, X. Sun, Y. Meng, X. Ao, Q. He, F. Wu, and J. Li, "Chinesebert: Chinese pretraining enhanced by glyph and pinyin information," *arXiv preprint arXiv:2106.16038*, 2021.

[43] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," *arXiv preprint arXiv:1908.10084*, 2019.

[44] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen *et al.*, "Lora: Low-rank adaptation of large language models." *ICLR*, vol. 1, no. 2, p. 3, 2022.

[45] Y. Zhang, B. Liu, C. Lu, Z. Li, H. Duan, S. Hao, M. Liu, Y. Liu, D. Wang, and Q. Li, "Lies in the air: Characterizing fake-base-station spam ecosystem in china," in *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*, 2020, pp. 521–534.

[46] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *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)*, 2019, pp. 4171–4186.

[47] Y. Cui, W. Che, T. Liu, B. Qin, S. Wang, and G. Hu, "Revisiting pre-trained models for chinese natural language processing," *arXiv preprint arXiv:2004.13922*, 2020.

[48] U. S. N. C. for the Protection of Human Subjects of Biomedical and B. Research, *The Belmont report: ethical principles and guidelines for the protection of human subjects of research*. Department of Health, Education, and Welfare, National Commission for the ..., 1978, vol. 2.

[49] E. Kenneally and D. Dittrich, "The menlo report: Ethical principles guiding information and communication technology research," *Available at SSRN 2445102*, 2012.

Appendix

1. Example of Perturbation Techniques

Table 11 shows a representative example of each adversarial perturbation technology.

2. Domains and Categories of Underground Jargon

Gambling: This domain covers content that promotes platforms, showcases "wins," or solicits participation, and it commonly employs `Visual` confusables, `Split`, and `Emoji` substitution to evade detection. For example, "哺𠯿" is used in place of "捕鱼" (fishing game); "棋牌" is split into "木其 | 片 | 卑." Such jargon circulates covertly on social and messaging platforms and often embeds links to gambling sites.
Porn: This domain comprises terms related to the trade and distribution of pornography, most often using `Homophone` and `Semantic Shift` strategies to conceal intent. For instance, "镁铝" (*měil* ǔ) stands in for "美女" (*měin* ǚ), and "外围" shifts from its literal meaning "outsider" to denote high-end escort services in underground contexts.
Drug: This domain relies largely on `Semantic Shift`, using metaphorical expressions for trafficking, use, or manufacturing. Examples include "溜冰" ("ice skating") to indicate methamphetamine use and "黄牙签" ("yellow toothpick") to refer to heroin. These terms appear benign in everyday language, making context-free detection prone to misses.
Fraud: This domain encompasses content intended to mislead or deceive, such as phishing, scam advertising, and impersonation, and it frequently builds metaphorical networks via `Semantic Shift`. For example, "鱼" ("fish") denotes victims, and "学生鱼" ("student fish") narrows the target set to students. Such codes streamline in-group communication while evading oversight.
Danger: This domain references the illicit trade or manufacture of weapons and related equipment, often pairing `Homophone` or `English` transliterations with `Visual` or `Symbolic` changes. A representative case is using "dog" for "gun," leveraging the similarity between English *gun* and Chinese "狗" (*gǒu*).
Promotion: This domain comprises expressions used to funnel users into illicit platforms or activities and typically relies on `Abbreviations` and `Visual` substitutions (with optional `Symbol` or `Emoji` substitutions) to bypass filters. Common examples include "VX" for "微信/WeChat" and visual substitutions such as "徽信". Compared with other domains, promotion jargon is more common in early-stage operations and exhibits strong connectivity and diffusion across the corpus.

3. Data Augmentation across Domains

Across domains, we condition synthesis on characteristic co-perturbation patterns: Gambling favors Visual + Split (with optional Emoji and occasional Pinyin/Homophone); Porn is dominated by Homophone + Semantic Shift with mild Visual tweaks; Drug relies primarily on Semantic Shift with light Symbol/Visual noise; Fraud layers Semantic Shift with Synonym/Symbol noise and occasional Split; Danger pairs Homophone or English transliteration with Visual/Symbol substitutions; and Promotion emphasizes Abbreviation + Visual (optionally Symbol/Emoji). Sampling

TABLE 11: Illustration and Examples of Each Adversarial Strategy.

Types		Definition	Sample
Pronunciation	Homophone	Replace characters with similar pronunciation	博彩(bocai) → 菠菜
	Pinyin	Replace words with their Pinyin	注册 → zhuce
	English	Replace words with their English words or acronyms	会员 → VIP
Glyph	Visual Substitution	Replace a character with a visually similar one.	捕鱼 → 晡忬
	Character Split	Break down a character into its constituent components	棋牌 → 木其片卑
Semantics	Synonym Substitution	Replace a word with similar meaning	捕鱼 → 抓鱼
	Semantic Shift	Use a word in a different context to imply another meaning	溜冰 → 吸毒
Other	Sequence Restructuring	Rearrange the order of characters or words in a sentence	美女捕鱼炸金花 → 镁|逋|淦钬|鲨|砘
	Emoji	Use special emojis to replace words	牛牛 → 🐂🐂
	Symbol	Insert or replace characters with special symbols	赠888元 → 赠⑧⑧⑧￥

follows each domain's empirical combination frequencies, with additional weight assigned to under-represented techniques.

4. JADEC-CoT Prompts

The prompt Q1 for adversarial jargon normalization is

> "Given the context and the candidate set for each term, select the most appropriate canonical form and generate the normalized text. If the candidate set is empty or inaccurate, infer the best normalization based on context and expert knowledge to ensure that all adversarial jargon has been replaced by its canonical form."

The prompt Q2 for jargon localization and semantic alignment is

> "Based on the original and normalized texts, identify each adversarial jargon span in the input and map it to its canonical form."

The prompt Q3 for identifying underground content is

> "Given the fully normalized and aligned text, analyze its overall semantic intent. Determine whether the content involves any underground domains and provide a final risk label."

5. Format of the fine-tuning dataset

Prompt II Structured Normalization Template

```
<|im_start|>system
You are a helpful assistant.
<|im_end|>
<|im_start|>user
You are a professional information
security analyst. Please restore
the following adversarial jargon:
{input}.
<|im_end|>
<|im_start|>assistant
{output}
<|im_end|>
```

Appendix A.
Meta-Review

The following meta-review was prepared by the program committee for the 2026 IEEE Symposium on Security and Privacy (S&P) as part of the review process as detailed in the call for papers.

A.1. Summary of Paper

This paper addresses the problem of detecting adversarial jargon used in Chinese underground communities to evade content moderation. The authors introduce ADVJARGON, a large-scale dataset of 70,000 real-world spam messages curated with an industry partner, and propose JADE, a taxonomy-guided framework. JADE leverages RAG and LoRA fine-tuning on a LLM to normalize obfuscated text and detect malicious intent.

A.2. Scientific Contributions

2) Provides a New Data Set For Public Use.
6) Provides a Valuable Step Forward in an Established Field.

A.3. Reasons for Acceptance

This paper provides a new data set for public use. Most adversarial NLP benchmarks rely on synthetic, budget-constrained perturbations. By curating the ADVJARGON dataset from real-world, in-the-wild attacker behavior, the authors provide a more realistic view of the modern threat landscape.

This paper provides a valuable step forward in an established field. Rather than treating the LLM as a black box, the authors combine multi-granular retrieval with parameter-efficient fine-tuning (LoRA) to successfully handle the unique glyph- and homophone-based complexities of Chinese obfuscation.