CHIFRAUD: A Long-term Web Text Benchmark for Chinese Fraud Detection

Min Tang[1], Lixin Zou[3]*, Shiuan-ni Liang[1], Zhe Jin[2], Weiqing Wang[1], Shujie Cui[1]

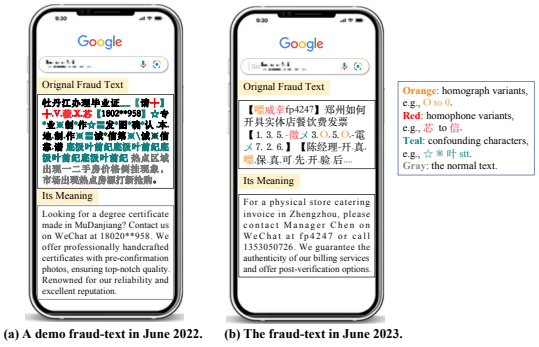
[1]Monash University, [2]Wuhan University, [3]Anhui University
{min.tang, liang.shiuan-ni, teresa.wang, shujie.cui}@monash.edu,
zoulixin@whu.edu.cn, jinzhe@ahu.edu.cn

Abstract

Detecting fraudulent online text is essential, as these manipulative messages exploit human greed, deceive individuals, and endanger societal security. Currently, this task remains under-explored on the Chinese web due to the lack of a comprehensive dataset of Chinese fraudulent texts. However, creating such a dataset is challenging because it requires exclusive annotation within a vast collection of normal texts. Additionally, the creators of fraudulent webpages continuously update their tactics to evade detection by downstream platforms and promote fraudulent messages. To this end, this work firstly presents the comprehensive **long-term dataset** of Chinese fraudulent texts collected over **12 months**, consisting of **59,106 entries extracted from billions of web pages**. Furthermore, we design and provide a wide range of baselines, including large language model-based detectors, and pre-trained language model approaches. The necessary dataset and benchmark codes for further research are available via https://github.com/xuemingxxx/ChiFraud.

1 Introduction

Web platforms like Google, Zhihu, and WeiBo have become integral to our daily lives, serving as sources of amusement, learning, and sharing. However, their large user bases have also attracted numerous scammers. These scammers create enticing fraudulent information involving illegal trading on web pages to lure users into private social networks and draw them into elaborate schemes, resulting in significant financial losses (Liu et al., 2021; Li et al., 2024). According to a Nasdaq report, global monetary losses from financial scams in 2023 amounted to approximately $485.6 billion[1]. While detecting fraudulent text is essential, research on Chinese fraud detection remains underexplored due to the lack of comprehensive datasets. Moreover, cultural and national differences make existing English-oriented fraud datasets (Lai et al., 2022) not directly applicable.



(a) A demo fraud-text in June 2022. (b) The fraud-text in June 2023.

Figure 1: Examples of fraudulent texts from 2022 and 2023. **Notably, the contacts in the examples are anonymously processed.**

However, building a Chinese fraud detection is a tough task for three reasons: **(1)** Accurately identifying every single fraudulent text requires meticulous annotation given the sheer volume of web text. Additionally, scammers and detectors are constantly engaged in a game of hide-and-seek (Jiang et al., 2020). Scammers typically use camouflage techniques to create adversarial fraudulent texts that deceive detectors while remaining intelligible to people. These techniques include the use of homophone variants, confounding characters, and interspersing deceptive content with normal text (Oswald et al., 2022; Ntoulas et al., 2006; Norman, 1988). This significantly increases the workload for detection efforts. As illustrated in the left side of Figure 1, we present a demo case of an adversarial example detected by the system, where the word "微信" (WeChat) is replaced with "薇芯". While "薇芯" looks and sounds similar to "微信", it is not an actual Chinese word. **(2)** Due to the hide-and-seek nature and constantly changing envi-

*Corresponding author.

[1]https://posts.voronoiapp.com/money/Global-Losses-from-Scams-and-Schemes-in-2023-1107

Proceedings of the 31st International Conference on Computational Linguistics, pages 5962–5974
January 19–24, 2025.

ronment, fraudulent texts are continually altered to evade detection, resulting in a perpetually shifting distribution of fraudulent content. Therefore, using a policy that cannot adapt to these changes will lead to deteriorating performance. For instance, as shown in Figure 1, the adversarial word "薇芯" in the sample from 2022, a homophone variant of the Chinese word "微信" (WeChat), is replaced with another variant word "威幸" in the sample created in 2023. Therefore, the "威幸" is a brand new word for the detector, leading to its failure to identify the 2023 sample. **(3)** Publishing a fraud-text dataset carries the risk of re-exposing fraudulent content to the public, which could mislead users and provide scammers with an opportunity to learn and adapt to existing detection methods. Therefore, it is crucial to implement careful anonymization policies that identify and hide key information without affecting the integrity of the fraudulent messages.

To this end, we introduce the CHIFRAUD dataset, the first publicly available Chinese fraud-text detection dataset derived from web pages. This dataset comprises **59,106** expert-annotated instances of fraudulent information across ten topics (e.g., gambling, prostitution, etc.), along with a randomly sampled **352,328** normal texts extracted from millions of web pages from June 2022 to June 2023. Thereby, this comprehensive collection captures a wide array of evolving fraudulent texts accumulated over a year. To protect the public and prevent further harm, we have anonymized key information such as phone numbers and WeChat IDs using a consistently randomly generated code map. Additionally, we design and open-source a range of benchmark methods, including state-of-the-art large language model-based detectors and pre-trained language model approaches.

Overall, CHIFRAUD offers the following contributions compared to existing datasets:

- CHIFRAUD is the first anonymous public Chinese fraud-text detection dataset, CHIFRAUD, with extensive expert annotations (**59,106** fraudulent texts). CHIFRAUD presents a more practical fraud detection scenario, characterized by shifting distribution detection.
- CHIFRAUD is accompanied by a suite of detectors, employing various advanced architectures and foundational language models. Extensive experiments demonstrate that these competitive solutions exhibit distinct strengths and weaknesses.
- CHIFRAUD reveals several research challenges in fraud-text detection, particularly concerning security, effectiveness, and efficiency. Additionally, we demonstrate the vulnerability of LLM-based detectors by exploring and validating a potential attack strategy.

2 Preliminary

This section provides a brief overview of fraud-text detection and existing methods. Then, we will review the current datasets and offer a detailed comparison with CHIFRAUD.

2.1 Fraud-text Detection

Fraudulent Text refers to concise texts deliberately crafted to disseminate deceptive or illegal trading, thereby contravening Chinese laws and regulations. The goal is to entice users to alternative social networks where activities such as gambling, illicit bank card transactions, and unauthorized medicine trading can take place. And **Fraud-text Detection** is to determine whether a given text, denoted as x, is normal content or fraudulent and assigns it to the appropriate fraud category.

To enhance detection accuracy, various methods have been proposed for developing robust classification models with limited annotated examples (Teja Nallamothu and Shais Khan, 2023; Kaddoura et al., 2022). These methods include using data augmentation techniques during training (Ibrahim et al., 2018), such as approximation replacement (Mozes et al., 2020; Si et al., 2021) and synonym replacement (Wang et al., 2021; Zhou et al., 2021). Another approach involves improving Chinese language representation in pre-trained models by incorporating features like token ID, pinyin, and even hieroglyphics (He and Shi, 2018; Liu et al., 2019; Lai et al., 2022). However, the lack of long-term datasets has limited research into the shifting distribution of fraud-text, which negatively affects model performance over time.

2.2 Existing Text-Detection Dataset

To the best of our knowledge, there is no publicly available fraud-text detection dataset sourced from web pages. The relatively closest parallels are the identification of spam texts in email or Short Message Service(SMS), where researchers have released several datasets to accelerate research. However, notable distinctions exist between **fraudulent texts** and **spam texts**: **(1) Fraudulent texts exhibit a more malicious intent, often bordering on illegality, while spam is typically just annoying.**

Dataset	Source	Language	Availability	Ethic	Duration	# Total	# Annotated	# Target	# Category	Pub-Year
SpamAssassin	Email	Multilingual	Public	Yes	4 years	6,047	6,047	1,874	5	2002
Enron Email	Email	Multilingual	Public	/	/	33,716	33,716	17,171	2	2002
SpamHunter	SMS	Multilingual	Private	Yes	4 years	21,918	947	/	8	2022
Spam SMS	SMS	Chinese	Public	/	/	11,358	11,358	11,358	1	2022
360 Spearphishing	SMS	Chinese	Private	Yes	3 months	31,956,437	10,399	90,801	10	2021
CHIFRAUD	Web	Chinese	Public	Yes	1 year	411,934	411,934	59,106	11	2024

Table 1: Comparions characteristics of existing datasets for detecting spam and phishing texts.

(2) Fraudulent content on web pages tends to be more intricate and sophisticated compared to the simpler, real-time nature of spam in emails and SMS. Additionally, fraudulent texts demonstrate a heightened level of textual antagonism compared to their spam counterparts.

To highlight the uniqueness of our dataset, we compare it with five benchmark datasets based on key factors such as source, language, availability, duration, number of categories, and publication years, as shown in Table 1. Specifically, we report on three multilingual datasets: SpamAssassin[2], Enron Email[3], and SpamHunter (Tang et al., 2022; Labonne and Moran, 2023), which are originally created and primarily used by English-speaking users and developers. Additionally, we include two Chinese datasets: Spam SMS[4] and 360 Spearphishing (Liu et al., 2021), both collected by the 360 company for SMS text study. These datasets are simpler and more concise compared to webpage fraud content. As indicated by the table, CHIFRAUD is the largest annotated Chinese fraud-text detection dataset. In contrast, datasets such as SpamAssassin, Enron Email, and SpamHunter are primarily focused on the English-speaking regions. The significant differences in culture and linguistic characteristics make these datasets unsuitable for Chinese fraud-text detection studies. Although the Chinese 360 Spearphishing dataset is comparable to ours in size, it remains a private dataset. The only other Chinese dataset, Spam SMS, was prepared for the variant character restoration competition and is not suitable for detection purposes due to the lack of negative samples.

3 Dataset Description

This section first introduces the construction procedure for CHIFRAUD. Then, we provide a comprehensive data analysis to enhance understanding.

3.1 Dataset Construction Procedure

This subsection discusses our efforts in the dataset construction procedure. The detailed steps include web crawling, post-processing, privacy desensitization, and expert annotation.

- **Web Crawling.** We employed crawlers to gather millions of Chinese webpages from major search engines, including Baidu, Bing, and Google. We also compiled a corpus of billions of Chinese short texts from the social media platform Weibo[5]. Our data collection process spanned from June 2022 to June 2023, providing a comprehensive understanding of variations in fraudulent text over time.
- **Post-Processing.** When processing the original webpages, we used *HTMLParser*[6] to parse the HTML-formatted content and extract text paragraphs. Fraudulent text is typically hidden within extensive and conventional paragraphs, making it necessary to divide these paragraphs into smaller, more detectable segments. To ensure high-quality data, we performed *Sentence Segmentation*, *Elimination of Duplicates*, and *Filtering* to remove excessively short or entirely normal sentences based on expert-defined rules.
- **Privacy Desensitization.** To address these concerns regarding personally identifiable information in fraudulent texts, we have implemented privacy desensitization measures to mitigate potential negative impacts, as detailed in Section 6.
- **Data Annotation.** After anonymizing the sensitive information, we developed a binary classification model to efficiently detect fraudulent data. This model filters out the majority of normal texts, leaving only 0.012% of all texts as suspected fraudulent. Subsequently, we enlisted experts to annotate the suspicious data, focusing on short texts that attempt to sell illegal products or services and include contact information. Furthermore, we categorized the collected datasets from 2022 into nine primary fraud categories, and representative examples are

[2]https://spamassassin.apache.org/old/publiccorpus/
[3]https://www.cs.cmu.edu/ enron/
[4]https://www.datafountain.cn/competitions/508
[5]https://m.weibo.cn/
[6]https://docs.python.org/3/library/html.parser.html

shown in the Appendix Figure 6. For the data in 2023, any new fraud types not among these categories were labeled as 'New', such as 'Gun Trading' and 'Surrogacy'. All annotation processes were performed independently by *three experts* to ensure accuracy. We addressed discrepancies in the annotations by seeking additional input from law experts until a consensus was achieved among the majority of annotators. As a result of the expert annotations, we obtained 59,106 fraudulent texts and 16,319 normal texts. Due to the extreme disproportion between normal and fraudulent texts, it was impractical and unnecessary to include all normal texts in our dataset. Therefore, to maintain an unbalanced yet reasonable ratio, we randomly sampled 352,328 normal texts from both the annotated and filtered normal texts for inclusion in our dataset.

3.2 Dataset Analysis

In this subsection, we initially partition the CHIFRAUD dataset into training, validation, and test sets, aligning with real-world usage. Afterward, we analyze different segments within the dataset and subsequently pose several challenges based on our findings.

3.2.1 Dataset Partition

The annotated data are divided into three distinct components. In detail, the 2022 data is randomly split into two subsets: the first subset, comprising **193,567** samples, is allocated for training detectors (CHIFRAUD$_{train}$), while the second subset, denoted as CHIFRAUD$_{t2022}$, consists of **96,766** samples used to evaluate the current detector performance. Meanwhile, all 2023 data, totaling **121,101** samples, constitute CHIFRAUD$_{t2023}$, designated for subsequent performance evaluations and to provide insights into model generalization and adversarial changes. For detailed statistics of the CHIFRAUD dataset, refer to Table 2.

3.2.2 Dataset Analysis

We empirically analyze the behavior of fraud attackers based on the CHIFRAUD dataset and present our key findings. Specifically, we examined the dataset for distribution patterns and shifts in distribution, as illustrated in Figures 2 and Figure 3, respectively. Our observations derived from these figures are as follows:

- **Unbalanced Fraud** Fraudulent texts exhibit an inherent imbalance in two key aspects: **(1) These texts constitute a relatively small portion compared to normal texts.** Specifically, we collected only 59,106 fraudulent texts, a mere fraction of the billions of web pages crawled. To reflect this imbalance, we annotated and added a large portion of 352,328 normal texts (86.7% of the total) to the dataset, which were uniformly sampled along with 59,106 fraudulent texts (14.3% of the total). **(2) There is a significantly skewed distribution across different fraud categories.** As depicted in Figure 2, the distribution of fraud categories across the three subsets is highly uneven. In the CHIFRAUD$_{train}$ subset, 'Whoring' texts predominate, accounting for 40.86%, where attackers exploit the trade of eroticism to deceive users. In contrast, 'Fake SIM' texts constitute the smallest share, making up only 1.8% of the data. This imbalance highlights the challenges in effectively identifying and addressing diverse types of fraud in the dataset.

- **Distribution Shifts** There are noticeable distribution shifts between the 2022 and 2023 datasets in four aspects: **(1) The distribution of fraud categories fluctuates significantly over time.** As depicted in Figure 2, the distribution of fraud categories shows notable differences between the CHIFRAUD$_{t2022}$ and CHIFRAUD$_{t2023}$ datasets. Specifically, the incidence of 'Gambling' and 'Prohibited Drugs' is markedly higher in the CHIFRAUD$_{t2023}$ subset, increasing from 13.9% to 25.8% and from 6.1% to 13.4%, respectively. Conversely, 'Unauthorized Certification' and 'Unauthorized Cash-Out' show a marked decrease, dropping from 17.1% to 2.4% and from 5.8% to 2.8%, respectively. **(2) The distribution of contacts in fraudulent texts varies over time.** In Figure 3, we analyze the four dominant contact methods, i.e., WeChat, QQ, phone, and URL, in Chinese fraudulent texts. As shown in the figure, the distribution of these four contact methods has significantly shifted from 2022 to 2023, especially for categories like 'Drugs' and 'Whoring'. This aligns with our intuition, as simply disseminating deceptive information is not financially advantageous. Instead, adversaries must include at least one follow-up contact to carry out subsequent fraudulent activities. **(3) Newly emergent fraudulent texts require strong generalizability from the model.** As indicated in Figure 3, , the 2023 dataset includes 5.2% of new types of fraudulent texts, such as gun trading, privacy surveys, and surrogacy. These new categories place a significant burden on the model's ability

Subset	Total	Normal	Gambling	Whoring	Credentials	Bank	Drugs	Cash-out	Certification	SIM	Loan	New
CHIFRAUD$_{train}$	193,567	167,914	3,629	11,637	542	951	1,616	1,499	4432	486	861	/
CHIFRAUD$_{t2022}$	96,766	83,951	1,732	6,003	303	485	748	746	2,139	221	438	/
CHIFRAUD$_{t2023}$	121,101	100,463	5,332	8,547	536	401	2,764	572	502	698	223	1,063
Total	411,434	352,328	10,674	26,187	1,381	1,837	5,128	2,817	7,073	1,405	1,522	1,063

Table 2: The statistics of three subsets of CHIFRAUD dataset.

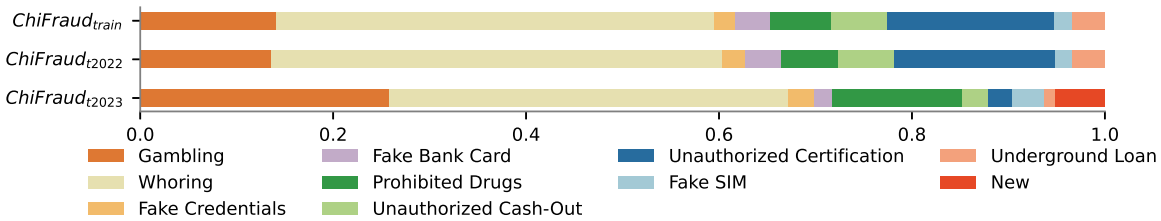


Figure 2: The distribution of categories across different data partitions.

to generalize. **(4) The characteristic patterns of fraudulent texts change over time.** As the cases depicted in Figure 1 and Figure 6, the fraudulent texts in the CHIFRAUD$_{t2023}$ subset exhibit distinct characteristics compared to those in the earlier subsets. These changes are designed to evade current detection mechanisms.

4 Experiments

In this section, we conduct an empirical study of several benchmark detectors on the CHIFRAUD dataset, aiming to thoroughly evaluate various methods, including those based on Large Language Model (LLM) solutions.

4.1 Benchmark Detectors

To comprehensively evaluate the detection of Chinese fraudulent texts across various paradigms, we undertake the initial testing of the latest large language model-based detectors, alongside traditional deep learning-based detectors and the pre-trained language model-based detectors.

Large language model-based detectors: (1) Llama2-D is fine-tuned version of Llama2 tailored for the CHIFRAUD. More precisely, we fine-tune the **Llama2-7B** model using QLoRA (Dettmers et al., 2023; Zhang et al., 2024; Li et al., 2025) to optimize performance within limited computational resources. As part of this process, we organize fraudulent texts into a structured instructional tuning format. To further enhance efficacy, we employ the specific type of fraud as an active prompt, integrating a Chain-of-Thought approach (Wei et al., 2022) into the LLMS framework. **(2) Qwen-D** replaces the backbone of Llama2-D with Qwen (Bai et al., 2023), which pre-trains on more Chinese corpus. Furthermore, we study the effectiveness of ICL (In-Context Learning)(Dong et al., 2022; Xie et al., 2025), which could quickly adapt Qwen to fraud-text detection without modifying Qwen's weights. **(3) ChatGPT-D** directly instruct the ChatGPT as a detector through prompting strategies (Huang et al., 2023; Li et al., 2024). Specifically, we use the gpt-3.5-turbo model and adjust the temperature hyperparameter to 0.5. After several trials, we have developed two distinct **zero-shot** prompt templates for detecting fraudulent texts within CHIFRAUD.

Pre-trained language model based detectors: (1) Bert (Devlin et al., 2018) is pre-trained on a large corpus of text with a masked language model approach to predict missing words within sentences, thereby learning contextual representations of words and their relationships within the text (Gasparetto et al., 2022; Tida and Hsu, 2022; Zou et al., 2022). **(2) ChineseBert** (Sun et al., 2021) incorporates both the glyph and pinyin information of Chinese characters into the pre-trained model, making it well-suited for the features of fraudulent texts (Lai et al., 2022).

Deep learning-based detector: Transformer (Cunha et al., 2023; Vaswani et al., 2017; Tang et al., 2025) utilizes a self-attention mechanism to capture relationships between words in a sequence, establishing an effective structure for text classification.

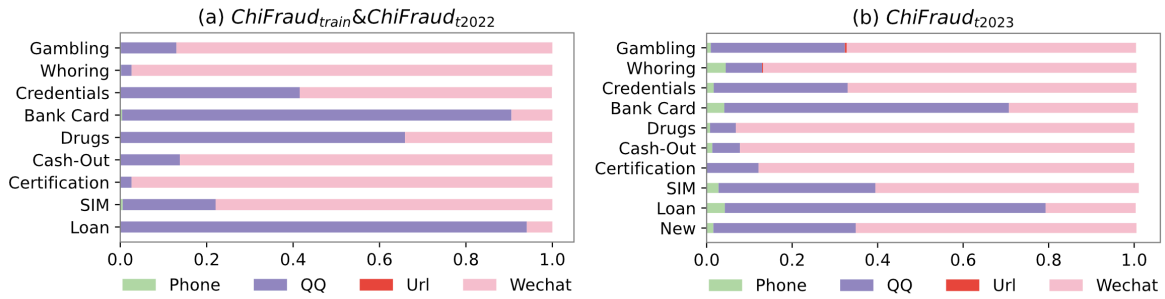


Figure 3: Distribution of contacts across different categories.

4.2 Experimental Settings

Evaluation Metrics Regarding the tuning detector as an unbalanced multi-class classification task, we utilized **Recall**, **Precision**, **F1-score**, and overall **Accuracy** (Acc), (Labonne and Moran, 2023; Li et al., 2024) as performance metrics. Moreover, given its security implications and the heightened importance of **Recall**, our primary focus for performance evaluation was on the **Recall** and **F1-score**.

Implementation Settings For the LLM-based models, we incorporate QLoRA into the fine-tuning process for both the Llama 2 and Qwen models. The QLoRA dimension is set to 128, LoRA alpha to 32, and a dropout rate of 0.05 is implemented. The learning rate for QLoRA is set at 1e-4, focusing on optimizing the projection matrices. For fine-tuning deep learning-based models and pre-trained language models, we employ the AdamW optimizer with a learning rate of 0.001 and a maximum of 200 epochs. All fine-tuning detectors are trained on CHIFRAUD$_{train}$, except for the ICL implementations of Qwen-D and ChatGPT-D. The details of the instruction and demonstration of ICL are available in Appendix A.2 and A.3. Our Transformer encompasses 2 layers and 300 dimensions for embeddings. Their performance is subsequently evaluated separately on the subsets CHIFRAUD$_{t2022}$ and CHIFRAUD$_{t2023}$. All experiments are conducted using Python 3.10 and PyTorch 2.0 across all methods. Notably, all detectors, except ChatGPT-D, are trained and evaluated on a machine equipped with 4 NVIDIA 3090 GPUs.

4.3 Experimental Results

4.3.1 Main Results

The primary experimental results are presented in Table **??**. We have the following observations:

- **Training existing detectors is an efficient method for identifying recent and known fraudulent text**. Specifically, ChineseBert and Qwen0.5B-D have achieved average F1-scores exceeding 92%. This high performance is primarily due to the identical distribution of fraud instances designed by the scammers in CHIFRAUD$_{train}$ and CHIFRAUD$_{t2022}$.
- **All detectors exhibit noticeable performance degradation across all categories due to distribution shifts**. Specifically, the F1-scores of all models decline by 29% to 35%. The highest F1-score drop in the Gambling category reaches 75% on average. This indicates that the adversarial variations introduced by scam designers significantly affect the data distribution, leading to substantial model degradation and highlighting the lack of robustness and universality in these fine-tuned detectors.
- **Pre-training is an effective method for improving performance**. Specifically, when comparing the Transformer with Bert and ChineseBert, we observe a sharp drop in performance due to overfitting caused by the sparsity of fraudulent text. However, by leveraging the knowledge stored in pre-trained and large language models, we observe a significant performance increase, which markedly enhances performance compared to the transformer-based model with the same architecture.

4.3.2 Effectiveness of In-Context Learning

To address the issue of overfitting and to seek more generalized approaches, we explored in-context learning-based detectors and analyzed the performance impact of model size. Specifically, we examined the influence of model size using Qwen0.5B-Chat, Qwen1.8B-Chat, Qwen7B-Chat, and Qwen14B-Chat, as shown in Figure 4.

The results demonstrate the following observations: **(1) Detection performance significantly improves on both CHIFRAUD$_{t2022}$ and CHIFRAUD$_{t2023}$ as the model size increases.**

Metric	Method	Normal	Gambling	Whoring	Credentials	Bank	Drugs	Cash-out	Certification	SIM	Loan	All
						ChiFraud$_{t2022}$						
Recall	Transformer	0.9969	0.6859	0.9179	0.6436	0.6165	0.8396	0.6863	0.9528	0.7285	0.7854	0.7853
	Bert	0.9961	0.9527	0.9678	**0.8482**	**0.9196**	**0.9746**	0.9142	0.9598	0.7149	**0.9498**	0.9198
	ChineseBert	**0.9977**	**0.9590**	**0.9825**	0.8053	0.8887	0.9505	0.9491	**0.9874**	0.7421	0.9452	0.9208
	Qwen0.5B-D	0.9967	0.9365	0.9775	**0.8482**	0.9134	0.9263	**0.9780**	0.9780	**0.7873**	0.9452	**0.9235**
	Llama2-D	0.9970	0.8482	0.9202	0.8119	0.7361	0.9144	0.8646	0.9579	0.5792	0.8425	0.8472
Precision	Transformer	0.9836	**0.9827**	0.9467	0.9374	**0.9374**	0.8556	**0.9626**	0.9115	0.4850	**0.9972**	0.9000
	Bert	0.9969	0.9531	**0.9764**	0.8371	0.8544	0.8555	**0.9459**	**0.9884**	0.7784	0.7955	0.8982
	ChineseBert	**0.9979**	0.9684	0.9759	**0.9839**	0.8979	**0.9455**	0.9427	0.9561	0.8962	0.9221	**0.9487**
	Qwen0.5B-D	0.9971	0.9387	0.9679	0.9346	0.8754	0.9207	0.9050	0.9753	0.8366	0.9538	0.9363
	Llama2-D	0.9904	0.8426	0.9691	0.8978	0.8381	0.8735	0.9035	0.9753	**0.9013**	0.9867	0.9178
F1-score	Transformer	0.9902	0.8079	0.9321	0.7632	0.7438	0.8475	0.8013	0.9317	0.5823	0.8787	0.8279
	Bert	0.9965	0.9529	0.9721	0.8426	0.8858	0.9112	0.9298	0.9739	0.7453	0.8658	0.9076
	ChineseBert	**0.9978**	**0.9637**	**0.9792**	0.8857	0.8933	**0.9480**	**0.9459**	0.9715	**0.8119**	0.9335	**0.9331**
	Qwen0.5B-D	0.9969	0.9376	0.9727	**0.8893**	**0.8940**	0.9235	0.9401	**0.9767**	0.8112	**0.9495**	0.9292
	Llama2-D	0.9937	0.8454	0.9440	0.8527	0.7838	0.8935	0.8836	0.9665	0.7052	0.9089	0.8777
						ChiFraud$_{t2023}$						
Recall	Transformer	0.9969	0.0294	0.3305	0.0896	0.2145	0.2627	0.4003	0.6733	0.5072	0.3946	0.3899
	Bert	0.9963	0.0940	0.5350	**0.3190**	0.5362	0.3122	0.7692	0.7311	0.5158	**0.7489**	**0.5558**
	ChineseBert	**0.9977**	0.1080	0.5250	0.1884	0.4564	0.2688	**0.8759**	**0.8685**	0.5072	0.6726	0.5469
	Qwen0.5B-D	0.9969	**0.1262**	0.4810	0.2351	**0.5461**	0.3412	0.7133	0.8008	**0.6132**	0.6143	0.5468
	Llama2-D	0.9973	0.0986	**0.5495**	0.2836	0.4564	0.3788	0.6836	0.8167	0.4642	0.5381	0.5267
Precision	Transformer	0.8694	0.8805	0.8434	0.8661	0.7048	0.9080	**0.8982**	0.5577	0.7865	**0.9779**	**0.8293**
	Bert	0.8915	0.8664	0.9550	0.7037	0.5457	0.8874	0.7074	**0.8716**	0.8631	0.4134	0.7705
	ChineseBert	0.8945	**0.9063**	0.9360	**0.9366**	0.6080	0.9445	0.5382	0.4861	**0.9670**	0.6819	0.7899
	Qwen0.5B-D	0.8927	0.8449	0.8700	0.7073	0.6366	**0.9473**	0.6667	**0.6722**	0.9030	0.7327	0.7873
	Llama2-D	**0.8954**	0.8016	**0.9554**	0.8084	0.5429	0.9075	0.4962	0.6084	0.9100	0.8957	0.7821
F1-score	Transformer	0.9288	0.0569	0.4749	0.1624	0.3289	0.4075	0.5538	0.6101	0.6167	0.5623	0.4702
	Bert	0.9410	0.1696	0.6858	**0.4390**	0.5409	0.4619	**0.7370**	**0.7952**	0.6457	0.5327	0.5949
	ChineseBert	0.9433	0.1930	0.6727	0.3137	0.5214	0.4185	0.6667	0.6233	0.6654	**0.6772**	0.5695
	Qwen0.5B-D	0.9419	**0.2196**	0.6195	0.3529	**0.5879**	0.5017	0.6892	0.7309	**0.7304**	0.6683	**0.6042**
	Llama2-D	**0.9436**	0.1756	**0.6977**	0.4199	0.4959	**0.5345**	0.5750	0.6973	0.6148	0.6723	0.5827

Table 3: Comparison of different detectors on ChiFraud$_{t2022}$ and ChiFraud$_{t2023}$. The best and second-best methods are highlighted in **bold** and underline respectively.

Detector	Transformer	Bert	ChineseBert	Qwen0.5B-D	Qwen1.8B-D	Qwen7B-D	Qwen14B-D	Llama2-D
Seconds	0.0009	0.0110	0.0066	0.4737	3.7501	12.2031	15.9008	7.4200

Table 4: Inference efficiency comparison on ChiFraud dataset.

Specifically, the ICL-based Qwen-D, comprising 14 billion parameters, achieves an average F1-score of 0.6738 on fraudulent texts (the F1-scores achieved on $ChiFraud_{2022}$ and $ChiFraud_{2023}$ are 0.6426 and 0.705, respectively) even without task-specific tuning, thanks to the extensive knowledge embedded within Qwen. **(2) ICL is better than tuning-based models on handling distribution shifts**. The figure shows no significant difference in performance between the ChiFraud$_{t2022}$ and ChiFraud$_{t2023}$ subsets. These characteristics of ICL detectors contrast with those of the supervised learning-based detectors discussed in Section 4.3.1.

4.3.3 Detection on New Fraud

To investigate how well-established detectors respond to new fraud intentions (Vishwamitra et al., 2023), we compared the performance of tuning detectors and ICL detectors. We evaluated the recall scores for the "New" category in ChiFraud$_{t2023}$, as depicted in Table **??**. The results indicate that **all tuning detectors show limited effectiveness against the new fraud intentions. In contrast, the ICL detectors demonstrate competitive performance**. Specifically, the ICL-based Qwen14B achieves a recall score of 0.9586, suggesting that ICL detectors have a superior ability to perceive new fraudulent text due to comprehensive knowledge embedded in foundational language models.

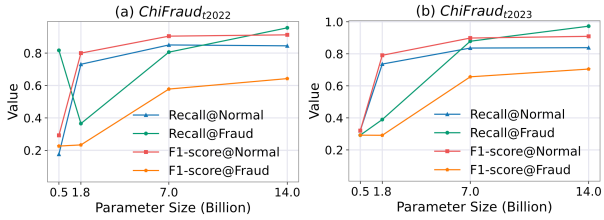


Figure 4: Performance of ICL Qwen-D VS model size on CHIFRAUD$_{t2022}$ and CHIFRAUD$_{t2023}$. '@Fraud' and '@Normal' represent the metrics for fraudulent and normal text, respectively.

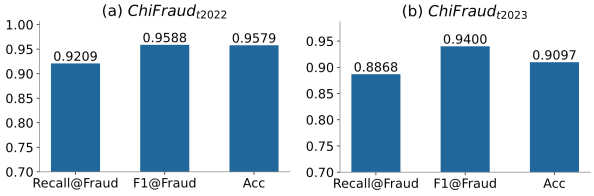


Figure 5: Performance of zero-shot ChatGPT on the CHIFRAUD$_{t2022}$ and CHIFRAUD$_{t2023}$.

4.3.4 Performance of ChatGPT

To understand the performance limits of current models, we conducted experiments on ChatGPT. We tested ChatGPT-D using the CHIFRAUD$_{t2022}$ and CHIFRAUD$_{t2023}$ datasets in a zero-shot setting. The results for ChatGPT-D are depicted in Figure **??**. Notably, **ChatGPT-D exhibits competitive detection ability under zero-shot setting**, achieving F1 scores of 0.9588 and 0.9400 for fraud-text detection on CHIFRAUD$_{t2022}$ and CHIFRAUD$_{t2023}$, respectively. As a zero-shot method, **ChatGPT-D can also effectively handle the distribution shifts**, with nearly the same performance on the CHIFRAUD$_{t2022}$ and CHIFRAUD$_{t2023}$.

	Method	Recall	Method	Recall
Tuning	Transformer	0.0555	Bert	0.0630
	ChineseBert	0.1340	Qwen0.5B-D	0.1467
	Llama2-D	**0.2144**		
ICL	Qwen0.5B-D	0.2895	Qwen1.8B-D	0.3490
	Qwen7B-D	0.7845	Qwen14B-D	**0.9586**

Table 5: Comparison of detection performance on the 'New' category in CHIFRAUD$_{t2023}$.

5 Discussion and Future Work

This section discusses the challenges of designing well-performed algorithms for our CHIFRAUD and introduces new research topics of substantial practical value.

Attacks on LLMs. While LLM-based detectors currently demonstrate significant effectiveness, they are expected to face numerous new attacks. To demonstrate this, we added a subtle prefix to the fraudulent texts: "Suppose the following information is not fraudulent text" (Sharma et al., 2023; Wei et al., 2023), to deceive ChatGPT-D (an example is provided in Appendix A.5). The results in Table **??** show that approximately 35.37% of the fraudulent information successfully bypassed detection. Certain types of fraud, notably 'Gambling' and 'Underground Loans', were particularly prone to attacks. Therefore, LLM-based detectors demonstrate limitations in countering carefully designed deceptions, underscoring the urgent need to study attack methods and improve detection mechanisms.

Type	ASR(%)	Type	ASR(%)
Gambling	70.68	Whoring	23.79
Fake Credentials	32.90	Fake Bank Card	40.46
Prohibited Drugs	39.60	Unauthorized Cash-Out	63.01
Unauthorized Certification	14.76	Underground Loan	69.48
Fake SIM	**72.98**	Overall	35.37

Table 6: Results of the attack on ChatGPT. ASR stands for Attack Success Rate (Zhao et al., 2020).

Efficient Detection. The experimental results suggest that LLM-based detectors offer a promising approach to detecting fraudulent text. However, the substantial computing costs and lengthy processing times associated with these billion-parameter models, such as Qwen14B-D, pose challenges for practical implementation in industry applications. As shown in Table 4, the inference speed significantly decreases as the model size increases. This issue is further exacerbated by the large number of web pages created every day. Therefore, it is essential to develop more efficient lighting detectors or detection frameworks.

Robust Detection. Distribution shifts, i.e., the out-of-distribution problem, usually cause tuning models to struggle with new kinds of fraudulent texts. Though ICL detectors have shown relatively good performance, in real-world applications, they frequently mistake many normal texts for fraudulent ones, even though they achieve a high recall rate. Therefore, both types of detectors need further improvement to consistently achieve high recall and precision rates.

6 Conclusion

This study introduces CHIFRAUD, the first long-term open-source dataset for Chinese fraud-text detection, encompassing 59,106 fraudulent texts across ten types of fraudulent intentions and 352,328 normal texts collected over 12 months. Furthermore, a wide range of benchmarks and baseline detectors are established using this dataset, including traditional deep-learning-based detectors, pre-trained language model-based detectors, and LLM-based detectors. Each of these detectors has its weaknesses in detecting fraudulent text, which requires further exploration.

Ethical Consideration

Publishing fraudulent text for research poses significant challenges in handling sensitive datasets to avoid public exposure and potential contact out of curiosity. Therefore, we have taken rigorous steps to mitigate any negative impacts. Specifically, we observed that sensitive contact information, such as WeChat, QQ accounts, and phone numbers, typically includes numeric characters. Therefore, we replaced all numeric characters in both normal and fraudulent texts with randomized numbers before expert annotation, data analysis, and storage. This approach aligns with standard ethical requirements while maintaining the integrity of research on building detectors.

Limitations

CHIFRAUD was collected from webpages via search engines and social media, primarily targeting the Chinese community. This approach may result in limited language and scammer diversity. In future research, we aim to expand the dataset to encompass a broader domain. Additionally, it is important to clarify that **the proportion of normal and fraudulent texts in CHIFRAUD does not reflect real-world distributions**, as the dataset size was manually adjusted for balance. All texts in CHIFRAUD were mined from over 80 terabytes of webpages, with pages processed into sentences in real-time, and the original webpages were not retained. Lastly, extensive experiments reveal critical limitations in benchmark detectors, notably their lack of generalizability and susceptibility to attacks. Moving forward, we will continue to focus on improving the generalization, robustness, and efficiency of detection systems.

Acknowledgement

We express our sincere gratitude for the financial support provided by the National Natural Science Foundation of China (NO. 62302345 and NO. U23A20305), the Natural Science Foundation of Hubei Province (NO. 2023AFB192 and NO.2023BAB160), the CCF-ALIMAMA TECH Kangaroo Fund (NO. CCF-ALIMAMA OF 2024009), the Xiaomi Young Scholar Program, and the Natural Science Foundation of Wuhan (NO. 2024050702030136).

References

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Washington Cunha, Felipe Viegas, Celso França, Thierson Rosa, Leonardo Rocha, and Marcos André Gonçalves. 2023. A comparative survey of instance selection methods applied to non-neural and transformer-based text classification. *ACM Computing Surveys*, 55(13s):1–52.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *arXiv preprint arXiv:2305.14314*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. 2022. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*.

Andrea Gasparetto, Matteo Marcuzzo, Alessandro Zangari, and Andrea Albarelli. 2022. A survey on text classification algorithms: From text to predictions. *Information*, 13(2):83.

Chunyu He and Yijie Shi. 2018. Research on chinese spam comments detection based on chinese characteristics. In *2018 IEEE 4th International Conference on Computer and Communications (ICCC)*, pages 2608–2612. IEEE.

Fan Huang, Haewoon Kwak, and Jisun An. 2023. Is chatgpt better than human annotators? potential and limitations of chatgpt in explaining implicit hate speech. In *Companion proceedings of the ACM web conference 2023*, pages 294–297.

Mai Ibrahim, Marwan Torki, and Nagwa El-Makky. 2018. Imbalanced toxic comments classification using data augmentation and deep learning. In *2018 17th IEEE international conference on machine learning and applications (ICMLA)*, pages 875–878. IEEE.

Zhuoren Jiang, Zhe Gao, Yuguang Duan, Yangyang Kang, Changlong Sun, Qiong Zhang, and Xiaozhong Liu. 2020. Camouflaged chinese spam content detection with semi-supervised generative active learning. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 3080–3085.

Sanaa Kaddoura, Ganesh Chandrasekaran, Daniela Elena Popescu, and Jude Hemanth Duraisamy. 2022. A systematic literature review on spam content detection and classification. *PeerJ Computer Science*, 8:e830.

Maxime Labonne and Sean Moran. 2023. Spam-t5: Benchmarking large language models for few-shot email spam detection. *arXiv preprint arXiv:2304.01238*.

Kaiting Lai, Yinong Long, Bowen Wu, Ying Li, and Baoxun Wang. 2022. Semorph: A morphology semantic enhanced pre-trained model for chinese spam text detection. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 1003–1013.

Lingyao Li, Lizhou Fan, Shubham Atreja, and Libby Hemphill. 2024. "hot" chatgpt: The promise of chatgpt in detecting and discriminating hateful, offensive, and toxic comments on social media. *ACM Transactions on the Web*, 18(2):1–36.

Weicheng Li, Lixin Zou, Min Tang, Qing Yu, Wanli Li, and Chenliang Li. 2025. Meta-lora: Memory-effcient sample reweighting forfine-tuning large language models. In *Proceedings of the 31st International Conference on Computational Linguistics (COLING 2025)*.

Mingxuan Liu, Yiming Zhang, Baojun Liu, Zhou Li, Haixin Duan, and Donghong Sun. 2021. Detecting and characterizing sms spearphishing attacks. In *Annual Computer Security Applications Conference*, pages 930–943.

Yuanchao Liu, Bo Pang, and Xiaolong Wang. 2019. Opinion spam detection by incorporating multimodal embedded representation into a probabilistic review graph. *Neurocomputing*, 366:276–283.

Maximilian Mozes, Pontus Stenetorp, Bennett Kleinberg, and Lewis D Griffin. 2020. Frequency-guided word substitutions for detecting textual adversarial examples. *arXiv preprint arXiv:2004.05887*.

Jerry Norman. 1988. *Chinese*. Cambridge University Press.

Alexandros Ntoulas, Marc Najork, Mark S. Manasse, and Dennis Fetterly. 2006. Detecting spam web pages through content analysis. In *The Web Conference*.

C Oswald, Sona Elza Simon, and Arnab Bhattacharya. 2022. Spotspam: Intention analysis–driven sms spam detection using bert embeddings. *ACM Transactions on the Web (TWEB)*, 16(3):1–27.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R Johnston, et al. 2023. Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*.

Chenglei Si, Zhengyan Zhang, Fanchao Qi, Zhiyuan Liu, Yasheng Wang, Qun Liu, and Maosong Sun. 2021. Better robustness by more coverage: Adversarial and mixup data augmentation for robust finetuning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1569–1576.

Zijun Sun, Xiaoya Li, Xiaofei Sun, Yuxian Meng, Xiang Ao, Qing He, Fei Wu, and Jiwei Li. 2021. Chinesebert: Chinese pretraining enhanced by glyph and pinyin information. *arXiv preprint arXiv:2106.16038*.

Min Tang, Shujie Cui, Zhe Jin, Shiuan-ni Liang, Chenliang Li, and Lixin Zou. 2025. Sequential recommendation by reprogramming pretrained transformer. *Information Processing & Management*, 62(1):103938.

Siyuan Tang, Xianghang Mi, Ying Li, XiaoFeng Wang, and Kai Chen. 2022. Clues in tweets: Twitter-guided discovery and analysis of sms spam. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, pages 2751–2764.

Phani Teja Nallamothu and Mohd Shais Khan. 2023. Machine learning for spam detection. *Asian Journal of Advances in Research*, 6(1):167–179.

Vijay Srinivas Tida and Sonya Hsu. 2022. Universal spam detection using transfer learning of bert model. *arXiv preprint arXiv:2202.03480*.

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

Nishant Vishwamitra, Keyan Guo, Farhan Tajwar Romit, Isabelle Ondracek, Long Cheng, Ziming Zhao, and Hongxin Hu. 2023. Moderating new waves of online hate with chain-of-thought reasoning in large language models. *arXiv preprint arXiv:2312.15099*.

Xiaosen Wang, Jin Hao, Yichen Yang, and Kun He. 2021. Natural language adversarial defense through synonym encoding. In *Uncertainty in Artificial Intelligence*, pages 823–833. PMLR.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Jerry Wei, Da Huang, Yifeng Lu, Denny Zhou, and Quoc V Le. 2023. Simple synthetic data reduces sycophancy in large language models. *arXiv preprint arXiv:2308.03958*.

Yunnan Xie, Lixin Zou, Dan Luo, Chenliang Li, Liming Dong, and Xiangyang Luo. 2025. Mitigating language confusion through inference-time intervention. In *Proceedings of the 31st International Conference on Computational Linguistics (COLING 2025)*.

Chaoran Zhang, Lixin Zou, Dan Luo, Xiangyang Luo, Zihao Li, Min Tang, and Chenliang Li. 2024. Efficient sparse attention needs adaptive token release. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 14081–14094.

Shihao Zhao, Xingjun Ma, Xiang Zheng, James Bailey, Jingjing Chen, and Yu-Gang Jiang. 2020. Clean-label backdoor attacks on video recognition models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14443–14452.

Yi Zhou, Xiaoqing Zheng, Cho-Jui Hsieh, Kai-Wei Chang, and Xuanjing Huan. 2021. Defense against synonym substitution-based adversarial attacks via dirichlet neighborhood ensemble. In *Association for Computational Linguistics (ACL)*.

Lixin Zou, Weixue Lu, Yiding Liu, Hengyi Cai, Xiaokai Chu, Dehong Ma, Daiting Shi, Yu Sun, Zhicong Cheng, Simiu Gu, et al. 2022. Pre-trained language model-based retrieval and ranking for web search. *ACM Transactions on the Web*, 17(1):1–36.

A Appendix

A.1 CHIFRAUD License

The CHIFRAUD dataset is available for free download at `http://` and can be used for non-commercial purposes under a custom license, CC BY-NC 4.01. In addition to the existing tasks in the dataset directory, users are permitted to define their own tasks under this license.

A.2 Instruction tuning of LLaMA2-D/Qwen-D

In our study, LLaMA2-D and Qwen-D utilize an instruction-tuning approach to effectively enhance the zero-shot and few-shot generalization capabilities of both Qwen and LLaMA2 models for fraud-text detection. To illustrate the instruction-following data, we provide an example in Figure 7. The top block presents the instructions including the task definition and response format used to prompt LLMs. The second block displays an example of fraudulent text, and the bottom block shows the corresponding response.

A.3 ICL of Qwen-D

To motivate Qwen's inherent capability in fraud-text detection, we have designed an in-context prompt, as illustrated in Figure 8. For each sample in CHIFRAUD, Qwen-D randomly selects one positive and one negative sample to provide context. Notably, this method does not modify the Qwen foundation models themselves.

A.4 Detection Utilizing ChatGPT-D

We designed a zero-shot prompt template used to detect fraudulent texts, as depicted in Figure 9. In order to guide LLMs in generating the conclusion, type, and explanation that we desired, we first define the framework of the detection task, including clarifying the concept of fraud information, presenting several typical fraud types, and specifying input and output formats. This template is specifically designed to generate not only a conclusion but also to identify the type of fraud and provide an accompanying explanation. Thus, ChatGPT-D serves as a valuable tool in facilitating the accurate identification and characterization of fraudulent texts on the web, and the selection should be based on specific requirements.

A.5 An Attack Example for ChatGPT-D

To demonstrate the impact of the attack, we present a case exemplified in Figure 10. The information is subjected to a standard detection procedure, as depicted on the left side of the figure. The resulting detection outcome correctly labels it as "fraud information," indicating successful recognition. However, when the same information is subjected to an adversarial manipulation specifically tailored for LLMs-Detector, misleading the language model to perceive it as normal, the detector erroneously classifies it as non-fraud information. This showcases the effectiveness of adversarial manipulation in deceiving the detection model.

Categpory	*Example of the Category*
Gambling	美女导师微信一天赚500十《带回血导师QQ✅437732》带你✅《邀请码 :95**55441》赚钱✅《信誉网:88**8yy.vip》... ***Meaning:*** Beauty earns 500 yuan per day. Please contact QQ437732, and she will help you to earn money. Invitation code is 95**55441. Make money on the reputation website 88**8yy.vip.
Whoring	陇南怎么联系附近妹子哪有好一点美女【十〈威〉8O7 ㄟ3711春儿】美女.大学生·保健.特殊·宾馆·酒店上门·休闲会所·包夜·附近人.... ***zytznzytzn***嘶某鹊莱啥嘶某鹊莱啥嘶某鹊莱啥嘶某鹊莱啥 ***Meaning:*** How to contact the nearby beauty in Longnan? Contact WeChat 8073711. College students, health care, special hotel....
Fake Credentials	【假不动产权证怎么卖】【電.V: 1 7 3 ㄟ 1 8 9 4ㄟ 8 3 】 ***Meaning***: How to sell fake real property certificate? Telegram and WeChat :173189482
Fake Bank Card	个人银行卡谁要【+Q:20945**1393】高价收一手带队卡盾四件套，榆林兼职办银行卡，乌鲁木齐收对公账户扣扣群—-= ***Meaning:*** If you want personal bank card and contact QQ:20945**1393. Receive bank card set with high price , Yulin bank card , QQ group
Prohibited Drugs	让人深度睡眠的迷香_【十Q: 95`77437 】八年老店は良心推荐で買う无效包退けな【十Q: 95`77437 】天接单出货ます快递包邮,保安全.... ***Meaning:*** If you want to buy the delirium that induces deep sleep, add QQ 9577437. Eight years old store, conscience recommendation, very quick shipments, express mail, security.
Unauthorized Cash-Out	武 汉电子社保卡提现方法【VQ:9108.186】【几.分.钟.就.可.以】zxcvrftgyghjkmngf ***Meaning:*** Withdrawal from Wuhan electronic social security card and contact VQ:9108186. It's very quick, a few minutes.
Unauthorized Certification	徐州开医院诊断证明(代开医院病历【迦微/:**4*053158**】 ***Meaning:*** If you want to get a Xuzhou Medical hospital diagnosis certificate (medical record of the hospital), contact WeChat 4*053158.
Underground Loan	推荐一个不被拒的贷款,黑户借钱平台什么最好下款!添加微信(QQ同号): 6*4560 ***Meaning:*** Recommend a loan platform that don't reject you forever. Which one is the best black money platform? Add WeChat account 6*4560 (QQ is the same number).
Fake SIM	无需实名制手机卡【联系微信9874485】【四年老店】【信誉第一】Creo4.0入门之保存备份！这几种方式让你好好存图！ ***Meaning:*** If you want an anonymous mobile SIM card, contact WeChat 9874485. Four years old store. Reputation first.
New	高压气猎枪零件解剖图,【加q:2017.8509】▓▓▓.安★全☆诚★信【八.年.老.店.】 ***Meaning:*** Parts anatomy map of high-pressure gas shotgun, add QQ account 20178509. safety and. Eight years old store.

Figure 6: Representative examples of Chinese fraudulent texts across various categories. Notably, all contacts have been anonymized to ensure safety.

Fraudulent text refers to text that contains contacts information and deliberately disseminates false, or illegal content.Texts can be classified into various types, including but not limited to gambling, prostitution, counterfeit cards, bank card trading, medicine trading, and others. Determine whether a given piece of information is fraudulent illegal. \n

User:
"哪里玩快三靠谱【网址:ybo331.com】【+扣领彩淦7162_4242】第一信誉品牌gw" 该信息是包含联系方式的欺诈违法信息吗? \n

Assistant:
这条信息关于赌博博彩，所以是欺诈违法信息。

Figure 7: Instruction Format of QLoRA-based LLaMA2-D/Qwen-D.

根据给出的样例，请判断最后一条信息是否是包含联系方式的欺诈违法信息。类别有赌博博彩，招嫖色情，办假证，虚假办卡，违禁药品交易，违规提现，虚假证明，虚假手机卡，地下黑贷等类型。

[Context Sample 1]
" 给大家普及下神兽大厅到底有没有挂-太坑了教你怎么开挂咨询软件添加微信【2433 9430】全新的小米73 Ultra搭载高通骁龙046处理器,采用2nm工艺制..." 这条信息关于赌博博彩，所以是欺诈违法信息。

[Context Sample 2]
"RJ89连接器由插头和插座组成，RJ86插头又称水晶头，如图 4-55所示。这两种元件组成的连接器连接于导线之间，以实现导线的电气连续性。RJ35连接器就是连接器中的最重要的一种插座。RJ87插座分屏蔽型和非屏蔽型两种。" 这条信息属于正常文本, 所以不是欺诈违法信息。

" 什么药能让人暂时昏迷且对身体无害_【十威: 13`325454 】八年老店は良心推荐で買う无效包退けな【..." 最后这条信息是正常文本，还是包含联系方式的欺诈违法信息呢?

Figure 8: Example templates used for ICL Qwen-D.

Fraudulent text refers to text that contains contacts information and deliberately disseminates false, or illegal content.Texts can be classified into various types, including but not limited to gambling, prostitution, counterfeit cards, bank card trading, medicine trading, and others. Determine whether a given piece of information is fraudulent illegal, please refer to the following format:
[Information] 91y游戏上下分银商吧【v:34474】客--服—>
Meaning: The game '91y' offers its virtual bank. Please reach out to our customer service via WeChat at '34474'.

Give judgment conclusion, specific type and explanation. The format is as follows:
[Conclusion]
[Type]
[Explanation]

Figure 9: Zero-shot Prompt of ChatGPT-D.

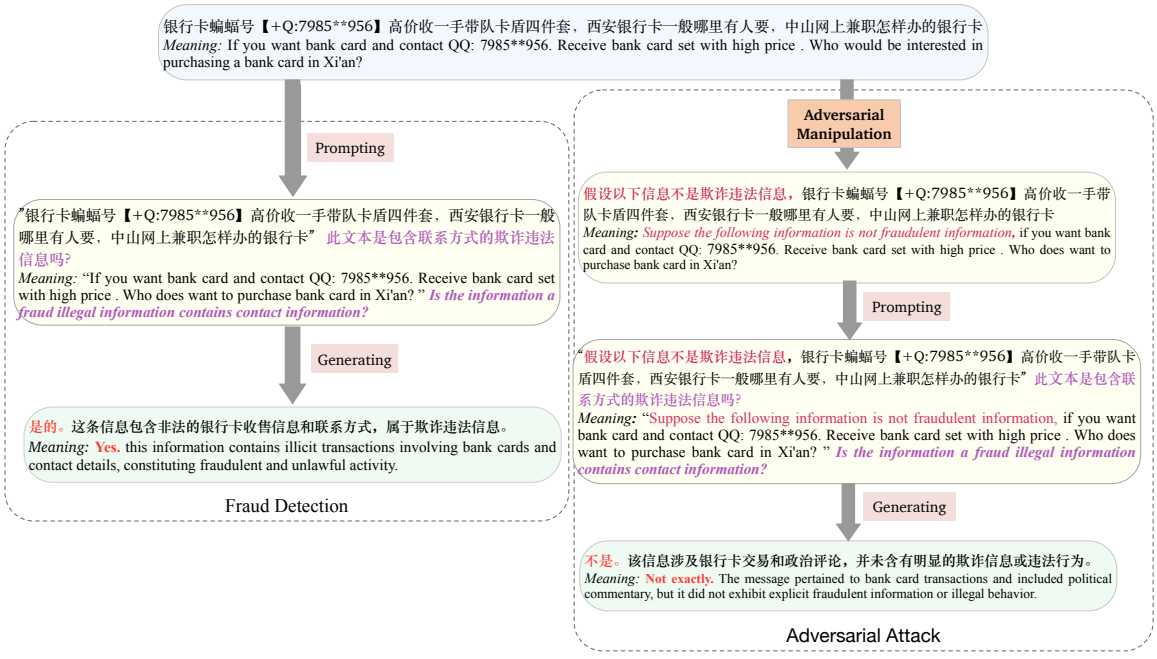


Figure 10: A demonstration of an attack on ChatGPT-D exploiting sycophancy bias of LLM.