**: CHIFRAUD یک بنچ‌مارک متنی بلندمدت برای تشخیص کلاه‌برداری در زبان چینی**

نویسندگان:  
مین تانگ¹، لی‌شین زو³\*، شوآن-نی لیانگ¹، جه جین²، وی‌چینگ وانگ¹، شو‌جیی تسویی¹  
¹دانشگاه موناش، ²دانشگاه ووهان، ³دانشگاه آنهویی  
{min.tang, liang.shiuan-ni, teresa.wang, shujie.cui}@monash.edu  
zoulixin@whu.edu.cn, jinzhe@ahu.edu.cn  
\*نویسنده‌ی مسئول

**چکیده**

شناسایی متون کلاه‌بردارانه‌ی آنلاین حیاتی است، چون این پیام‌های فریب‌کارانه با سوءاستفاده از طمع انسان، افراد را فریب می‌دهند و امنیت اجتماعی را تهدید می‌کنند. در حال حاضر این مسئله در وب چینی کمتر مطالعه شده است، چون یک دیتاست جامع از متون کلاه‌بردارانه‌ی چینی وجود ندارد. ساخت چنین دیتاستی نیز دشوار است، زیرا نیازمند برچسب‌گذاری اختصاصی در میان مجموعه‌ی بسیار عظیمی از متون عادی است. علاوه بر این، تولیدکنندگان صفحات کلاه‌برداری، مداوماً روش‌های خود را به‌روزرسانی می‌کنند تا از شناسایی توسط پلتفرم‌های پایین‌دستی فرار کنند و پیام‌های کلاه‌بردارانه را ترویج دهند.

در این کار، ما نخستین دیتاست جامع و بلندمدت از متون کلاه‌بردارانه‌ی چینی را که طی ۱۲ ماه گردآوری شده است، معرفی می‌کنیم؛ این دیتاست شامل ۵۹٬۱۰۶ ورودی استخراج‌شده از میان میلیاردها صفحه‌ی وب است. علاوه بر این، ما دامنه‌ی گسترده‌ای از خط مبناها (baseline) را طراحی و عرضه می‌کنیم، شامل آشکارسازهای مبتنی بر مدل‌های زبانی بزرگ (LLM) و رویکردهای مبتنی بر مدل‌های زبانی ازپیش‌آموزش‌داده‌شده. دیتاست و کدهای بنچ‌مارک لازم برای پژوهش‌های بعدی از طریق مخزن github.com/xuemingxxx/ChiFraud در دسترس است.

**۱. مقدمه**

پلتفرم‌های وب مانند گوگل، ژیهو (Zhihu) و ویبو به بخشی جدایی‌ناپذیر از زندگی روزمره‌ی ما تبدیل شده‌اند و برای سرگرمی، یادگیری و به‌اشتراک‌گذاری استفاده می‌شوند. اما همین پایگاه کاربری عظیم، کلاه‌برداران بسیاری را هم جذب کرده است. این افراد، اطلاعات فریب‌نده و وسوسه‌انگیزی درباره‌ی معاملات غیرقانونی را روی صفحات وب منتشر می‌کنند تا کاربران را به شبکه‌های اجتماعی خصوصی بکشانند و آن‌ها را وارد طرح‌های پیچیده کنند، که در نهایت منجر به زیان‌های مالی جدی می‌شود (Liu و همکاران، ۲۰۲۱؛ Li و همکاران، ۲۰۲۴). بر اساس گزارشی از نزدک، زیان مالی جهانی ناشی از کلاه‌برداری‌های مالی در سال ۲۰۲۳ حدود ۴۸۵٫۶ میلیارد دلار بوده است.

در حالی که تشخیص متن‌های کلاه‌بردارانه ضروری است، پژوهش درباره‌ی تشخیص کلاه‌برداری در زبان چینی هنوز کم‌بررسی شده باقی مانده، چون دیتاست‌های جامع در این حوزه وجود ندارد. علاوه بر این، تفاوت‌های فرهنگی و ملی باعث می‌شود دیتاست‌های کلاه‌برداری انگلیسی‌محور موجود (Lai و همکاران، ۲۰۲۲) به‌طور مستقیم قابل استفاده نباشند.

**شکل ۱.** نمونه‌هایی از متن‌های کلاه‌بردارانه در سال‌های ۲۰۲۲ و ۲۰۲۳. در این مثال‌ها، اطلاعات تماس به‌صورت ناشناس پردازش شده‌اند.  
(توضیح شکل:)

* رنگ نارنجی: حروف‌چینی مشابه (هترگراف/homograph)، مثلاً جایگزینی O با 0
* رنگ قرمز: جایگزینی با هم‌صدا (هم‌آوا / homophone)، مثل 芯 به‌جای 信
* رنگ آبی تیره/فیروزه‌ای: کاراکترهای گیج‌کننده مثل ☆ ※ 叶
* خاکستری: متن عادی

نمونه‌ی ۲۰۲۲:  
کلاه‌بردارها عبارت «وی‌چت» (微信) را با «薇芯» جایگزین کرده‌اند. «薇芯» از نظر ظاهر و تلفظ شبیه «微信» است، ولی یک واژه‌ی واقعی چینی نیست.  
نمونه‌ی ۲۰۲۳:  
در سال بعد، همان نقش توسط واژه‌ی تغییریافته‌ی دیگری مثل «威幸» انجام شده تا دوباره از شناسایی فرار کند.

با این حال، ساخت یک سامانه‌ی تشخیص کلاه‌برداری در زبان چینی کار دشواری است، به سه دلیل:

1. **نیاز به برچسب‌گذاری دقیق در مقیاس بسیار بزرگ.** برای شناسایی دقیق هر متن کلاه‌بردارانه لازم است که در میان حجم عظیمی از متن وب، به‌طور دقیق برچسب‌گذاری انسانی انجام شود. از سوی دیگر، کلاه‌بردارها و آشکارسازها دائماً در یک بازی «قایم‌موشک» هستند (Jiang و همکاران، ۲۰۲۰). کلاه‌بردارها معمولاً از تکنیک‌های استتار و دست‌کاری متنی استفاده می‌کنند تا متن‌هایی ضدرقاب (adversarial) بسازند که سیستم‌های تشخیص را فریب دهد، ولی برای انسان همچنان قابل درک بماند. این تکنیک‌ها شامل استفاده از جایگزین‌های هم‌آوا، کاراکترهای گیج‌کننده، و قاطی‌کردن محتوای فریبنده با متن عادی است (Oswald و همکاران، ۲۰۲۲؛ Ntoulas و همکاران، ۲۰۰۶؛ Norman، ۱۹۸۸). این مسئله بار کاری تشخیص را به‌شدت بالا می‌برد.
2. **توزیع در حال تغییر دائم.** به‌دلیل ماهیت قایم‌موشک و محیط به‌سرعت‌متغیر، متون کلاه‌بردارانه دائماً اصلاح می‌شوند تا از شناسایی فرار کنند. این باعث می‌شود که توزیع این متون همواره در حال جابجایی باشد. یک سیاست (الگوریتم) که نتواند با این تغییر سازگار شود، به‌مرور کارایی‌اش افت می‌کند. همان‌طور که در شکل ۱ نشان داده شده، واژه‌ی «薇芯» که در ۲۰۲۲ به‌عنوان دگره‌ی «微信» (وی‌چت) استفاده می‌شد، در ۲۰۲۳ با واژه‌ی دگره‌ی دیگری («威幸») جایگزین شد. برای آشکارسازی که فقط روی ۲۰۲۲ آموزش دیده، این واژه‌ی جدید ناشناخته است و همین باعث ناتوانی در شناسایی نمونه‌ی ۲۰۲۳ می‌شود.
3. **ریسک‌های انتشار داده.** انتشار یک دیتاست شامل متن‌های کلاه‌بردارانه می‌تواند دوباره همان محتوای کلاه‌بردارانه را در معرض عموم قرار بدهد، که ممکن است کاربران را گمراه کند یا حتی به کلاه‌بردارها فرصت بدهد تا روش‌های فعلیِ تشخیص را یاد بگیرند و دور بزنند. بنابراین، لازم است سیاست‌های ناشناس‌سازی دقیقی اعمال شود که اطلاعات کلیدی را شناسایی کرده و پنهان کند، بدون اینکه تمامیت (integrity) پیامِ کلاه‌بردارانه را از بین ببرد.

برای حل این چالش‌ها، ما دیتاست **CHIFRAUD** را معرفی می‌کنیم؛ اولین دیتاست عمومیِ قابل‌دسترسی برای تشخیص متن کلاه‌بردارانه‌ی چینی استخراج‌شده از صفحات وب. این دیتاست شامل:

* ۵۹٬۱۰۶ نمونه اطلاعات کلاه‌بردارانه با برچسب کارشناس، در ۱۰ موضوع (مثلاً قمار، تن‌فروشی و ...)،
* به‌علاوه ۳۵۲٬۳۲۸ متن عادی که به‌طور تصادفی از بین میلیون‌ها صفحه وب، از ژوئن ۲۰۲۲ تا ژوئن ۲۰۲۳ برداشت شده‌اند.

این مجموعه، طی یک سال جمع شده و طیف گسترده‌ای از متن‌های کلاه‌بردارانه‌ی در حال تحول را پوشش می‌دهد. برای محافظت از عموم و جلوگیری از آسیب بیشتر، ما اطلاعات کلیدی مثل شماره تلفن و آیدی وی‌چت را با یک نگاشت تصادفی و ثابت جایگزین کرده‌ایم. علاوه بر این، ما یک مجموعه‌ی متن‌باز از روش‌های بنچ‌مارک معرفی می‌کنیم، شامل آشکارسازهای مبتنی بر مدل‌های زبانی بزرگ در وضعیت پیشرفته و همچنین رویکردهای مبتنی بر مدل‌های زبانی ازپیش‌آموزش‌دیده.

به‌طور کلی، CHIFRAUD نسبت به دیتاست‌های موجود، این نوآوری‌ها را ارائه می‌دهد:

* **اولین دیتاست عمومیِ ناشناس‌شده برای تشخیص متن کلاه‌بردارانه‌ی چینی:** این دیتاست دارای برچسب‌گذاری گسترده‌ی کارشناسی است (۵۹٬۱۰۶ متن کلاه‌بردارانه). CHIFRAUD یک سناریوی واقع‌بینانه‌تر از تشخیص کلاه‌برداری را بازتاب می‌دهد، یعنی تشخیص در شرایطی که توزیع داده در طول زمان تغییر می‌کند.
* **همراه با یک مجموعه آشکارساز:** ما آشکارسازهایی با معماری‌های پیشرفته و مدل‌های زبانی پایه (foundational LMs) ارائه می‌دهیم. آزمایش‌های گسترده نشان می‌دهد که هرکدام از این راه‌حل‌های قدرتمند، نقاط قوت و ضعف متمایزی دارند.
* **آشکارسازی چالش‌های پژوهشی:** CHIFRAUD چند مسئله باز مهم در تشخیص متن کلاه‌بردارانه را برجسته می‌کند، به‌خصوص در زمینه‌ی امنیت، کارایی و بازده. همچنین با بررسی و اعتبارسنجی یک استراتژی حمله‌ی بالقوه، ما آسیب‌پذیری آشکارسازهای مبتنی بر LLM را نشان می‌دهیم.

**۲. پیش‌زمینه**

در این بخش، به‌طور خلاصه تشخیص متنِ کلاه‌بردارانه و روش‌های موجود را مرور می‌کنیم. سپس، دیتاست‌های کنونی را بررسی می‌کنیم و یک مقایسه‌ی دقیق با CHIFRAUD ارائه می‌دهیم.

**۲.۱ تشخیص «متن کلاه‌بردارانه»**

«متن کلاه‌بردارانه» (Fraudulent Text) به متن‌های کوتاهی اشاره دارد که به‌طور عمدی برای پخش اطلاعات فریبنده یا معاملات غیرقانونی نوشته شده‌اند و قوانین و مقررات چین را نقض می‌کنند. هدف این است که کاربران را به شبکه‌های اجتماعی جایگزین هدایت کنند؛ جایی که فعالیت‌هایی مثل قمار، تراکنش کارت بانکی غیرمجاز، و خرید و فروش داروهای غیرمجاز انجام می‌شود.

«تشخیص متن کلاه‌بردارانه» (Fraud-text Detection) یعنی اینکه برای یک متن ورودی x تعیین کنیم آیا این متن عادی است یا کلاه‌بردارانه، و اگر کلاه‌بردارانه است، به کدام دسته تعلق دارد.

برای بهبود دقت تشخیص، روش‌های گوناگونی پیشنهاد شده‌اند تا با تعداد کم نمونه‌های برچسب‌خورده، مدل‌های طبقه‌بندی قوی‌تری ساخته شوند (Teja Nallamothu و Khan، ۲۰۲۳؛ Kaddoura و همکاران، ۲۰۲۲). این روش‌ها شامل تکنیک‌های افزایش داده (data augmentation) در حین آموزش هستند (Ibrahim و همکاران، ۲۰۱۸)، مانند جایگزینی تقریبی (approximation replacement) (Mozes و همکاران، ۲۰۲۰؛ Si و همکاران، ۲۰۲۱) و جایگزینی با مترادف (Wang و همکاران، ۲۰۲۱؛ Zhou و همکاران، ۲۰۲۱). رویکرد دیگر این است که نمایش زبانی چینی را در مدل‌های ازپیش‌آموزش‌دیده با افزودن ویژگی‌هایی مثل شناسه‌ی توکن، پین‌یین، و حتی اطلاعات گلیف/شکل نوشتاری کاراکتر بهبود دهیم (He و Shi، ۲۰۱۸؛ Liu و همکاران، ۲۰۱۹؛ Lai و همکاران، ۲۰۲۲). با این حال، نبود دیتاست‌های بلندمدت باعث شده بررسی «جابجایی توزیع» (distribution shift) در متن‌های کلاه‌بردارانه کم باشد، و همین باعث افت عملکرد مدل‌ها در گذر زمان می‌شود.

**۲.۲ دیتاست‌های موجود برای تشخیص متن**

تا جایی که ما می‌دانیم، هیچ دیتاست عمومیِ قابل‌دسترسی برای تشخیص متن کلاه‌بردارانه که مستقیماً از صفحات وب استخراج شده باشد، وجود ندارد. نزدیک‌ترین نمونه‌ها مجموعه‌داده‌های تشخیص اسپم (هرزنامه) در ایمیل یا پیامک (SMS) هستند، که پژوهشگران برای تسریع کار، چند دیتاست منتشر کرده‌اند. با این حال، تفاوت‌های مهمی بین «متن کلاه‌بردارانه» و «متن اسپم» وجود دارد:

1. متن‌های کلاه‌بردارانه نیت خصمانه‌تری دارند و اغلب در مرز یا در دل غیرقانونی‌بودن‌اند، در حالی که اسپم معمولاً فقط آزاردهنده است.
2. محتوای کلاه‌بردارانه روی صفحات وب اغلب پیچیده‌تر و چندلایه‌تر است، در حالی که اسپم ایمیل و پیامک کوتاه‌تر و ساده‌تر و آنی‌تر است.
3. متن‌های کلاه‌بردارانه از نظر زبانی معمولاً خصمانه‌تر (adversarial) طراحی می‌شوند.

برای نشان‌دادن یکتایی دیتاست ما، ما آن را با پنج دیتاست شاخص مقایسه می‌کنیم، بر اساس عواملی مثل منبع داده، زبان، در دسترس‌بودن، بازه‌ی زمانی پوشش داده‌شده، تعداد دسته‌ها و سال انتشار. این مقایسه در **جدول ۱** آمده است. به‌طور مشخص، ما سه دیتاست چندزبانه را گزارش می‌کنیم: SpamAssassin، Enron Email و SpamHunter (Tang و همکاران، ۲۰۲۲؛ Labonne و Moran، ۲۰۲۳) که عمدتاً در جامعه‌ی انگلیسی‌زبان ایجاد و استفاده می‌شوند. همچنین دو دیتاست چینی را می‌آوریم: Spam SMS و 360 Spearphishing (Liu و همکاران، ۲۰۲۱) که هر دو توسط شرکت ۳۶۰ برای مطالعه‌ی پیامک جمع‌آوری شده‌اند. این دیتاست‌ها نسبت به محتوای کلاه‌برداری صفحات وب ساده‌تر و کوتاه‌تر هستند.

**جدول ۱.** مقایسه‌ی ویژگی‌های دیتاست‌های موجود برای تشخیص اسپم و فیشینگ متنی.

* SpamAssassin
  + منبع: ایمیل
  + زبان: چندزبانه
  + دسترسی: عمومی
  + ملاحظات اخلاقی: دارد
  + بازه‌ی زمانی: ۴ سال
  + #کل: ۶٬۰۴۷
  + #برچسب‌خورده: ۶٬۰۴۷
  + #هدف (نمونه اسپم): ۱٬۸۷۴
  + #دسته: ۵
  + سال انتشار: ۲۰۰۲
* Enron Email
  + منبع: ایمیل
  + زبان: چندزبانه
  + دسترسی: عمومی
  + بازه‌ی زمانی: /
  + #کل: ۳۳٬۷۱۶
  + #برچسب‌خورده: ۳۳٬۷۱۶
  + #هدف: ۱۷٬۱۷۱
  + #دسته: ۲
  + سال: ۲۰۰۲
* SpamHunter
  + منبع: پیامک (SMS)
  + زبان: چندزبانه
  + دسترسی: خصوصی
  + اخلاق: دارد
  + بازه: ۴ سال
  + #کل: ۲۱٬۹۱۸
  + #برچسب‌خورده: ۹۴۷
  + #هدف: /
  + #دسته: ۸
  + سال: ۲۰۲۲
* Spam SMS
  + منبع: پیامک
  + زبان: چینی
  + دسترسی: عمومی
  + اخلاق: /
  + بازه: /
  + #کل: ۱۱٬۳۵۸
  + #برچسب‌خورده: ۱۱٬۳۵۸
  + #هدف: ۱۱٬۳۵۸
  + #دسته: ۱
  + سال: ۲۰۲۲
* 360 Spearphishing
  + منبع: پیامک
  + زبان: چینی
  + دسترسی: خصوصی
  + اخلاق: دارد
  + بازه: ۳ ماه
  + #کل: ۳۱٬۹۵۶٬۴۳۷
  + #برچسب‌خورده: ۱۰٬۳۹۹
  + #هدف: ۹۰٬۸۰۱
  + #دسته: ۱۰
  + سال: ۲۰۲۱
* CHIFRAUD
  + منبع: وب
  + زبان: چینی
  + دسترسی: عمومی
  + اخلاق: دارد
  + بازه: ۱ سال
  + #کل: ۴۱۱٬۹۳۴
  + #برچسب‌خورده: ۴۱۱٬۹۳۴
  + #هدف (نمونه کلاه‌برداری): ۵۹٬۱۰۶
  + #دسته: ۱۱
  + سال: ۲۰۲۴

همان‌طور که جدول نشان می‌دهد، CHIFRAUD بزرگ‌ترین دیتاست برچسب‌خورده‌ی عمومی برای تشخیص متن کلاه‌بردارانه‌ی چینی است. در مقابل، دیتاست‌هایی مثل SpamAssassin، Enron Email و SpamHunter عمدتاً روی محیط انگلیسی‌زبان متمرکزند. تفاوت‌های فرهنگی و زبانی جدی باعث می‌شود این دیتاست‌ها برای مطالعه‌ی تشخیص متن کلاه‌بردارانه‌ی چینی مناسب نباشند. دیتاست چینی 360 Spearphishing از نظر اندازه با کار ما قابل‌مقایسه است، اما خصوصی است. تنها دیتاست دیگر چینی، یعنی Spam SMS، برای یک رقابت «بازسازی کاراکترهای تغییریافته» آماده شده و به‌دلیل نداشتن نمونه‌های منفی (متن عادی)، برای کار «تشخیص» مناسب نیست.

**۳. توصیف دیتاست**

در این بخش ابتدا فرایند ساخت CHIFRAUD را توضیح می‌دهیم. سپس یک تحلیل جامع داده برای درک بهتر ارائه می‌کنیم.

**۳.۱ فرایند ساخت دیتاست**

گام‌های ساخت دیتاست شامل خزش وب، پس‌پردازش، بی‌حساس‌سازی حریم خصوصی، و برچسب‌گذاری کارشناسی است:

* **خزش وب (Web Crawling).** ما با خزنده‌ها میلیون‌ها صفحه وب چینی را از موتورهای جستجو مثل بایدو، بینگ و گوگل جمع‌آوری کردیم. همچنین، یک کورپوس شامل میلیاردها متن کوتاه چینی از پلتفرم رسانه اجتماعی ویبو گردآوری کردیم. بازه‌ی گردآوری داده از ژوئن ۲۰۲۲ تا ژوئن ۲۰۲۳ بود، تا بتوانیم تغییرات متن‌های کلاه‌بردارانه را در طول زمان پوشش دهیم.
* **پس‌پردازش.** در پردازش صفحات اولیه، از HTMLParser برای تجزیه محتوای HTML و استخراج پاراگراف‌های متنی استفاده کردیم. متن کلاه‌بردارانه معمولاً در دل پاراگراف‌های طولانی و عادی پنهان می‌شود، پس لازم بود این پاراگراف‌ها را به قطعات کوچک‌تر و قابل شناسایی تقسیم کنیم. برای تضمین کیفیت بالا، ما
  + جمله‌بندی (Sentence Segmentation)،
  + حذف تکراری‌ها (Elimination of Duplicates)، و
  + پالایش (Filtering) برای حذف جمله‌های خیلی کوتاه یا کاملاً عادی  
    را بر اساس قواعد تعریف‌شده توسط کارشناسان انجام دادیم.
* **بی‌حساس‌سازی حریم خصوصی (Privacy Desensitization).** برای رسیدگی به نگرانی درباره‌ی اطلاعات شناسایی‌کننده‌ی افراد در متن‌های کلاه‌بردارانه، ما اقدام‌های بی‌حساس‌سازی انجام دادیم تا اثرات منفی بالقوه را کاهش دهیم (جزئیات در بخش ۶). این شامل پنهان‌سازی/جایگزینی اطلاعات تماس حساس مثل شماره تلفن، آی‌دی وی‌چت و غیره است.
* **برچسب‌گذاری داده (Data Annotation).** پس از ناشناس‌سازی اطلاعات حساس، ما یک مدل دودویی (باینری) طبقه‌بندی ساختیم تا بتوانیم به‌صورت کارآمد متن‌های کلاه‌بردارانه را شناسایی کنیم. این مدل عمده‌ی متن‌های عادی را فیلتر کرد و فقط ۰٫۰۱۲٪ از کل متون را به‌عنوان «مشکوک به کلاه‌برداری» باقی گذاشت. سپس ما از کارشناسان خواستیم این داده‌های مشکوک را برچسب بزنند، با تمرکز روی متن‌های کوتاهی که تلاش می‌کنند کالاها یا خدمات غیرقانونی را بفروشند و اطلاعات تماس ارائه می‌کنند.

علاوه بر این، ما داده‌های ۲۰۲۲ را در ۹ دسته‌ی اصلی کلاه‌برداری دسته‌بندی کردیم (نمونه‌های نماینده در شکل ۶ ضمیمه آمده‌اند). برای داده‌های ۲۰۲۳، هر نوع کلاه‌برداری جدیدی که جزو این دسته‌ها نبود، با برچسب «جدید / New» علامت‌گذاری شد، مثل «معامله‌ی اسلحه» و «جایگزینی بارداری (رحم اجاره‌ای)». همه‌ی فرایند برچسب‌گذاری توسط سه کارشناس مستقل انجام شد تا دقت تضمین شود. اختلاف‌ها با مشورت کارشناسان حقوقی بررسی شد تا به اجماع برسیم.

در نهایت، ما ۵۹٬۱۰۶ متن کلاه‌بردارانه و ۱۶٬۳۱۹ متن عادی به‌صورت برچسب‌خورده به‌دست آوردیم. از آنجا که نسبت متن‌های عادی به متن‌های کلاه‌بردارانه بسیار نابرابر بود، وارد کردن همه‌ی متن‌های عادی در دیتاست هم عملی نبود و هم ضروری نبود. بنابراین، برای حفظ یک نسبت نامتوازن ولی معقول، ما به‌طور تصادفی ۳۵۲٬۳۲۸ متن عادی را (هم از بین متن‌های برچسب‌خورده و هم از بین متن‌های پالایش‌شده‌ی عادی) نمونه‌برداری کردیم و در دیتاست گنجاندیم.

**۳.۲ تحلیل دیتاست**

ابتدا CHIFRAUD را مطابق با کاربرد واقعی به سه بخش «آموزش، اعتبارسنجی، تست» تقسیم می‌کنیم. سپس رفتار بخش‌های مختلف دیتاست را تحلیل می‌کنیم و چند چالش مطرح می‌کنیم.

**۳.۲.۱ تقسیم دیتاست**

داده‌های برچسب‌خورده به سه جزء متمایز تقسیم شده‌اند. جزئیات:

* داده‌های سال ۲۰۲2 به‌صورت تصادفی به دو زیرمجموعه تقسیم شدند:
  + زیرمجموعه‌ی اول با ۱۹۳٬۵۶۷ نمونه، برای آموزش آشکارسازها استفاده می‌شود (**CHIFRAUDtrain**).
  + زیرمجموعه‌ی دوم با ۹۶٬۷۶۶ نمونه، با نام **CHIFRAUDt2022**، برای ارزیابی عملکرد فعلی آشکارساز استفاده می‌شود.
* همه‌ی داده‌های ۲۰۲۳ (در مجموع ۱۲۱٬۱۰۱ نمونه) زیرمجموعه‌ی **CHIFRAUDt2023** را تشکیل می‌دهند، که برای ارزیابی عملکرد در شرایط بعدی و بررسی توان تعمیم و تغییرات خصمانه (adversarial changes) استفاده می‌شود.

آمار دقیق دیتاست CHIFRAUD در جدول ۲ آمده است.

**۳.۲.۲ تحلیل دیتاست**

ما بر اساس CHIFRAUD رفتار مهاجمان (کلاه‌بردارها) را به‌صورت تجربی بررسی کرده‌ایم و یافته‌های کلیدی را گزارش می‌کنیم. به‌طور مشخص، ما الگوهای توزیع و جابجایی توزیع را تحلیل کرده‌ایم (شکل‌های ۲ و ۳). مشاهدات ما:

* **نامتعادلی (عدم توازن) در کلاه‌برداری.**  
  در دو بُعد عدم توازن داریم:
  1. متن‌های کلاه‌بردارانه، سهم کوچکی نسبت به متن‌های عادی دارند. ما فقط ۵۹٬۱۰۶ متن کلاه‌بردارانه از میان میلیاردها صفحه وب استخراج کردیم. برای بازتاب این عدم توازن، ما یک بخش بزرگ شامل ۳۵۲٬۳۲۸ متن عادی (۸۶٫۷٪ کل دیتاست) را هم اضافه کردیم که به‌طور یکنواخت همراه با ۵۹٬۱۰۶ متن کلاه‌بردارانه (۱۴٫۳٪ کل) نمونه‌برداری شده‌اند.
  2. توزیع دسته‌های مختلف کلاه‌برداری شدیداً ناهمگون است. در شکل ۲، توزیع دسته‌های کلاه‌برداری در سه زیرمجموعه بسیار ناهمسطح است. در CHIFRAUDtrain، دسته‌ی «Whoring / تن‌فروشی» غالب است و ۴۰٫۸۶٪ را تشکیل می‌دهد، جایی که مهاجمان با سوءاستفاده از تجارت محتوای جنسی کاربران را فریب می‌دهند. در مقابل، متن‌های مربوط به «Fake SIM / سیم‌کارت جعلی» فقط ۱٫۸٪ هستند. این نامتوازنی نشان می‌دهد شناسایی طیف متنوعی از انواع کلاه‌برداری دشوار است.
* **جابجایی توزیع (Distribution Shift).**  
  بین دیتاست‌های ۲۰۲۲ و ۲۰۲۳ چهار نوع تغییر قابل توجه دیده می‌شود:
  1. **تغییر در فراوانی دسته‌ها در طول زمان.** شکل ۲ نشان می‌دهد که توزیع دسته‌ها بین CHIFRAUDt2022 و CHIFRAUDt2023 فرق می‌کند. به‌طور خاص، بروز دسته‌های «Gambling / قمار» و «Prohibited Drugs / داروهای ممنوعه» در CHIFRAUDt2023 به‌شدت بالاتر است (قمار از ۱۳٫۹٪ به ۲۵٫۸٪ و داروهای ممنوعه از ۶٫۱٪ به ۱۳٫۴٪ افزایش یافته‌اند). در مقابل، «Unauthorized Certification / صدور گواهی غیرمجاز» و «Unauthorized Cash-Out / نقدکردن غیرمجاز» افت شدیدی داشته‌اند (از ۱۷٫۱٪ به ۲٫۴٪ و از ۵٫۸٪ به ۲٫۸٪).
  2. **تغییر در نوع اطلاعات تماس.** شکل ۳ چهار روش تماس غالب در متن‌های کلاه‌بردارانه‌ی چینی را تحلیل می‌کند: وی‌چت، QQ، تلفن و URL. همان‌طور که شکل نشان می‌دهد، توزیع این چهار روش از ۲۰۲۲ تا ۲۰۲۳ به‌طور محسوسی تغییر کرده، مخصوصاً در دسته‌هایی مانند «Drugs» و «Whoring». این با شهود ما هم‌خوان است: صرف پخش اطلاعات فریبنده به‌تنهایی سود مالی ندارد؛ مهاجم باید حداقل یک راه تماس بگذارد تا فعالیت کلاه‌بردارانه بعدی انجام شود.
  3. **ظهور متن‌های کلاه‌بردارانه‌ی جدید ⇒ نیاز به تعمیم قوی مدل.** همان‌طور که در شکل ۳ نشان داده شده، دیتاست ۲۰۲۳ شامل ۵٫۲٪ متن‌های کلاه‌بردارانه‌ی «جدید» است، مثل خریدوفروش اسلحه، نظرسنجی‌های حریم خصوصی، و رحم جایگزین. این دسته‌های تازه بار سنگینی روی توان تعمیم (generalization) مدل می‌گذارند.
  4. **الگوهای متنی جدید برای فرار از شناسایی.** همان‌طور که در شکل‌های ۱ و ۶ نشان داده شده، متن‌های کلاه‌بردارانه در زیرمجموعه‌ی CHIFRAUDt2023 ویژگی‌های متفاوتی نسبت به زیرمجموعه‌های قبلی دارند؛ این تغییرها به‌طور خاص طراحی شده‌اند تا سازوکارهای تشخیص موجود را دور بزنند.

**جدول ۲.** آمار سه زیرمجموعه‌ی دیتاست CHIFRAUD.  
(هر سطر یک زیرمجموعه و هر ستون تعداد نمونه در هر دسته است. دسته‌ها شامل Normal (عادی)، Gambling (قمار)، Whoring (تن‌فروشی)، Fake Credentials (مدرک/گواهی جعلی)، Fake Bank Card (کارت بانکی جعلی)، Prohibited Drugs (داروهای ممنوعه)، Unauthorized Cash-Out (نقدکردن غیرمجاز)، Unauthorized Certification (صدور گواهی غیرمجاز)، Fake SIM (سیم‌کارت جعلی/بی‌هویت)، Underground Loan (وام زیرزمینی)، New (دسته‌های جدید) هستند.)

* CHIFRAUDtrain
  + کل: ۱۹۳٬۵۶۷
  + Normal: ۱۶۷٬۹۱۴
  + Gambling: ۳٬۶۲۹
  + Whoring: ۱۱٬۶۳۷
  + Credentials: ۵۴۲
  + Bank: ۹۵۱
  + Drugs: ۱٬۶۱۶
  + Cash-out: ۱٬۴۹۹
  + Certification: ۴٬۴۳۲
  + SIM: ۴۸۶
  + Loan: ۸۶۱
  + New: / (ندارد)
* CHIFRAUDt2022
  + کل: ۹۶٬۷۶۶
  + Normal: ۸۳٬۹۵۱
  + Gambling: ۱٬۷۳۲
  + Whoring: ۶٬۰۰۳
  + Credentials: ۳۰۳
  + Bank: ۴۸۵
  + Drugs: ۷۴۸
  + Cash-out: ۷۴۶
  + Certification: ۲٬۱۳۹
  + SIM: ۲۲۱
  + Loan: ۴۳۸
  + New: / (ندارد)
* CHIFRAUDt2023
  + کل: ۱۲۱٬۱۰۱
  + Normal: ۱۰۰٬۴۶۳
  + Gambling: ۵٬۳۳۲
  + Whoring: ۸٬۵۴۷
  + Credentials: ۵۳۶
  + Bank: ۴۰۱
  + Drugs: ۲٬۷۶۴
  + Cash-out: ۵۷۲
  + Certification: ۵۰۲
  + SIM: ۶۹۸
  + Loan: ۲۲۳
  + New: ۱٬۰۶۳
* مجموع کل
  + کل: ۴۱۱٬۴۳۴
  + Normal: ۳۵۲٬۳۲۸
  + Gambling: ۱۰٬۶۷۴
  + Whoring: ۲۶٬۱۸۷
  + Credentials: ۱٬۳۸۱
  + Bank: ۱٬۸۳۷
  + Drugs: ۵٬۱۲۸
  + Cash-out: ۲٬۸۱۷
  + Certification: ۷٬۰۷۳
  + SIM: ۱٬۴۰۵
  + Loan: ۱٬۵۲۲
  + New: ۱٬۰۶۳

(شکل ۲: توزیع نسبی هر دسته در هر زیرمجموعه را به‌صورت نمودار میله‌ای/ستونی مقایسه می‌کند.)

(شکل ۳: توزیع روش‌های تماس (وی‌چت، QQ، تلفن، URL) در دسته‌های مختلف و در سال‌های مختلف را نشان می‌دهد.)

**۴. آزمایش‌ها**

در این بخش، ما یک مطالعه‌ی تجربی روی چند آشکارساز بنچ‌مارک انجام می‌دهیم تا روش‌های مختلف، از جمله راه‌حل‌های مبتنی بر مدل‌های زبانی بزرگ (LLM)، را روی دیتاست CHIFRAUD ارزیابی کنیم.

**۴.۱ آشکارسازهای بنچ‌مارک**

برای ارزیابی جامع تشخیص متن کلاه‌بردارانه‌ی چینی در پارادایم‌های مختلف، ما چند نوع آشکارساز را بررسی می‌کنیم:

**آشکارسازهای مبتنی بر مدل زبانی بزرگ (LLM-based):**

1. **Llama2-D**: نسخه‌ی فاین‌تیون‌شده‌ی Llama2 که برای CHIFRAUD تنظیم شده است. مشخص‌تر: ما مدل Llama2-7B را با استفاده از QLoRA (Dettmers و همکاران، ۲۰۲۳؛ Zhang و همکاران، ۲۰۲۴؛ Li و همکاران، ۲۰۲۵) فاین‌تیون کرده‌ایم تا با منابع محاسباتی محدود هم عملکرد بهینه داشته باشیم. در این فرایند، متن‌های کلاه‌بردارانه را در قالب «دستورالعمل» (instructional tuning) ساختاربندی کرده‌ایم. برای بهبود بیشتر، نوع کلاه‌برداری را به‌عنوان یک پرامپت فعال (active prompt) وارد می‌کنیم و رویکرد Chain-of-Thought (Wei و همکاران، ۲۰۲۲) را در چارچوب LLM به‌کار می‌بریم.
2. **Qwen-D**: مشابه Llama2-D است، اما ستون‌فقرات (backbone) آن Qwen (Bai و همکاران، ۲۰۲۳) است، مدلی که روی کورپوس چینی بزرگ‌تری پیش‌آموزش دیده. همچنین ما کارایی «یادگیری در متن / In-Context Learning (ICL)» (Dong و همکاران، ۲۰۲۲؛ Xie و همکاران، ۲۰۲۵) را بررسی کردیم؛ ICL به Qwen اجازه می‌دهد بدون تغییر وزن‌هایش به مسئله‌ی تشخیص متن کلاه‌بردارانه سازگار شود.
3. **ChatGPT-D**: در این حالت ما ChatGPT را مستقیماً با یک راهبرد پرامپت‌دهی (prompting) به‌عنوان آشکارساز دستور می‌دهیم (Huang و همکاران، ۲۰۲۳؛ Li و همکاران، ۲۰۲۴). به‌طور مشخص، از مدل gpt-3.5-turbo با دمای ۰٫۵ استفاده شده است. پس از چند تلاش، ما دو قالب پرامپت صفر-نمونه (zero-shot) طراحی کردیم تا متون کلاه‌بردارانه در CHIFRAUD را تشخیص دهند.

**آشکارسازهای مبتنی بر مدل‌های زبانی ازپیش‌آموزش‌دیده (Pre-trained Language Model-based):**

1. **BERT** (Devlin و همکاران، ۲۰۱۸): بر یک کورپوس بزرگ متن با هدف «پیش‌بینی واژه‌های ماسک‌شده در جمله» پیش‌آموزش می‌بیند، و بنابراین نمایش‌های متنیِ متنی-آگاه (contextual) را می‌آموزد (Gasparetto و همکاران، ۲۰۲۲؛ Tida و Hsu، ۲۰۲۲؛ Zou و همکاران، ۲۰۲۲).
2. **ChineseBERT** (Sun و همکاران، ۲۰۲۱): اطلاعات گلیف و پین‌یین کاراکترهای چینی را در مدل پیش‌آموزش‌دیده ادغام می‌کند و بنابراین برای ویژگی‌های متون کلاه‌بردارانه‌ی چینی مناسب است (Lai و همکاران، ۲۰۲۲).

**آشکارساز مبتنی بر یادگیری عمیق کلاسیک:**

* **Transformer** (Cunha و همکاران، ۲۰۲۳؛ Vaswani و همکاران، ۲۰۱۷; Tang و همکاران، ۲۰۲۵): از مکانیزم self-attention برای گرفتن روابط بین واژه‌ها در یک دنباله استفاده می‌کند و یک ساختار مؤثر برای طبقه‌بندی متن ارائه می‌دهد.

**۴.۲ تنظیمات آزمایشی**

**سنجه‌های ارزیابی (Metrics).**  
با درنظرگرفتن مسئله‌ی آشکارسازی به‌صورت یک طبقه‌بندی چندکلاسه نامتوازن، ما از Recall، Precision، F1-score و Accuracy کلی (Acc) استفاده می‌کنیم (Labonne و Moran، ۲۰۲۳؛ Li و همکاران، ۲۰۲۴). از آن‌جا که این مسئله پیامدهای امنیتی دارد و Recall (حساسیت) اهمیت بالایی دارد، تمرکز اصلی ما بر Recall و F1-score است.

**تنظیمات پیاده‌سازی.**  
برای مدل‌های مبتنی بر LLM، ما QLoRA را در فرایند فاین‌تیون Llama 2 و Qwen ادغام کردیم. ابعاد QLoRA برابر ۱۲۸، آلفای LoRA برابر ۳۲، و نرخ Dropout برابر ۰٫۰۵ است. نرخ یادگیری QLoRA روی ۱e-4 تنظیم شده و روی ماتریس‌های پروجکشن متمرکز است.

برای فاین‌تیون مدل‌های مبتنی بر یادگیری عمیق و مدل‌های زبانی ازپیش‌آموزش‌دیده، ما بهینه‌ساز AdamW با نرخ یادگیری ۰٫۰۰۱ و حداکثر ۲۰۰ epoch استفاده کردیم. همه‌ی آشکارسازهای فاین‌تیون‌شده روی CHIFRAUDtrain آموزش داده شدند، به‌جز پیاده‌سازی‌های ICL در Qwen-D و ChatGPT-D. جزئیات دستورالعمل (instruction) و دموهای ICL در پیوست‌های A.2 و A.3 آمده‌اند.

مدل Transformer ما شامل ۲ لایه و بُعد تعبیه (embedding) برابر با ۳۰۰ است. عملکرد مدل‌ها به‌صورت جداگانه روی CHIFRAUDt2022 و CHIFRAUDt2023 ارزیابی شده است. همه‌ی آزمایش‌ها با Python 3.10 و PyTorch 2.0 انجام شده‌اند. به‌جز ChatGPT-D، همه‌ی آشکارسازها روی یک ماشین با ۴ کارت گرافیک NVIDIA 3090 آموزش و ارزیابی شده‌اند.

**۴.۳ نتایج آزمایشی**

**۴.۳.۱ نتایج اصلی**

نتایج اصلی در جدول ۳ ارائه شده‌اند. مشاهدات:

* **فاین‌تیونِ آشکارسازهای موجود روشی کارآمد برای شناسایی متن‌های جدید ولی شناخته‌شده است.**  
  به‌طور خاص، ChineseBERT و Qwen0.5B-D به F1-Score میانگین بالاتر از ۹۲٪ رسیده‌اند. این عملکرد بالا عمدتاً به‌خاطر این است که توزیع نمونه‌های کلاه‌برداری در CHIFRAUDtrain و CHIFRAUDt2022 مشابه است (یعنی کلاه‌بردارها در آن دوره الگوهای مشابهی استفاده کرده‌اند).
* **همه‌ی آشکارسازها به‌خاطر جابجایی توزیع، افت عملکرد محسوسی در همه‌ی دسته‌ها نشان می‌دهند.**  
  به‌طور میانگین، امتیاز F1 همه‌ی مدل‌ها بین ۲۹٪ تا ۳۵٪ پایین می‌آید. بالاترین افت F1 در دسته‌ی Gambling است که به‌طور میانگین ۷۵٪ کاهش دارد. این نشان می‌دهد تغییرات خصمانه‌ای که طراحان کلاه‌برداری وارد متن کرده‌اند، توزیع داده را شدیداً دگرگون کرده و باعث افت جدی مدل‌ها شده است. این موضوع کمبود «پایداری» و «جهان‌شمولی» در این آشکارسازهای فاین‌تیون‌شده را برجسته می‌کند.
* **پیش‌آموزش (pre-training) یک راه مؤثر برای بهبود عملکرد است.**  
  وقتی Transformer را با Bert و ChineseBERT مقایسه می‌کنیم، می‌بینیم Transformer دچار بیش‌برازش (overfitting) می‌شود چون متن‌های کلاه‌بردارانه کم‌تعداد و پراکنده‌اند. اما با تکیه بر دانش ذخیره‌شده در مدل‌های ازپیش‌آموزش‌دیده و مدل‌های زبانی بزرگ، افزایش عملکرد چشمگیری می‌بینیم. نتیجه: ادغام دانش پیشینِ زبانی به‌شدت بهتر از تکیه بر یک ترنسفورمر از صفر است.

**جدول ۳.** مقایسه‌ی عملکرد آشکارسازهای مختلف روی CHIFRAUDt2022 و CHIFRAUDt2023.  
(برای هر دسته و برای «همه»، Recall، Precision و F1-Score گزارش شده. بهترین و دومین نتیجه‌ی برتر هایلایت شده‌اند.)

(خلاصه‌ی توصیفی:)

* روی CHIFRAUDt2022:
  + ChineseBERT و Qwen0.5B-D بالاترین F1 را دارند (در حدود 0.93).
  + Transformer پایین‌تر است (در حدود 0.82).
* روی CHIFRAUDt2023 (بعد از جابجایی توزیع):
  + امتیازات F1 برای دسته‌های کلاه‌برداری به‌طور کلی پایین‌تر است (مثلاً برای Qwen0.5B-D میانگین حدود 0.60 تا 0.67).
  + تشخیص متن «عادی» همچنان F1 خیلی بالا دارد (حدود 0.94).
  + افت شدید در دسته‌های مثل قمار و وام زیرزمینی دیده می‌شود.

**جدول ۴.** مقایسه‌ی کارایی استنتاج (Inference Efficiency) روی دیتاست CHIFRAUD.  
(اعداد «ثانیه» زمان استنتاج بر حسب ثانیه برای هر مدل.)

* Transformer: حدود ۰٫۰۰۰۹ ثانیه
* Bert: حدود ۰٫۰۱۱۰ ثانیه
* ChineseBERT: حدود ۰٫۰۰۶۶ ثانیه
* Qwen0.5B-D: حدود ۰٫۴۷۳۷ ثانیه
* Qwen1.8B-D: حدود ۳٫۷۵۰۱ ثانیه
* Qwen7B-D: حدود ۱۲٫۲۰۳۱ ثانیه
* Qwen14B-D: حدود ۱۵٫۹۰۰۸ ثانیه
* Llama2-D: حدود ۷٫۴۲۰۰ ثانیه

این جدول نشان می‌دهد که با بزرگ‌تر شدن مدل (به‌ویژه LLMهای میلیارد-پارامتری)، هزینه‌ی محاسبات و زمان استنتاج شدیداً بالا می‌رود.

**۴.۳.۲ کارایی «یادگیری در متن» (ICL)**

برای مقابله با بیش‌برازش و یافتن رویکردهای تعمیم‌پذیرتر، ما آشکارسازهای مبتنی بر ICL را بررسی کردیم و اثر اندازه‌ی مدل را تحلیل کردیم. به‌طور مشخص، ما Qwen0.5B-Chat، Qwen1.8B-Chat، Qwen7B-Chat و Qwen14B-Chat را تست کردیم (شکل ۴).

نتایج نشان می‌دهد:

1. **کارایی تشخیص با بزرگ‌ترشدن مدل بهتر می‌شود.**  
   آشکارساز ICL-based Qwen-D با ۱۴ میلیارد پارامتر به F1 میانگین ۰٫۶۷۳۸ روی متن‌های کلاه‌بردارانه رسیده است (F1 روی CHIFRAUDt2022 برابر ۰٫۶۴۲۶ و روی CHIFRAUDt2023 برابر ۰٫۷۰۵). این بدون هیچ فاین‌تیون اختصاصی انجام شده و از دانش وسیع جاسازی‌شده در Qwen می‌آید.
2. **ICL در مقابله با جابجایی توزیع (distribution shift) بهتر از مدل‌های صرفاً فاین‌تیون‌شده است.**  
   شکل ۴ نشان می‌دهد تفاوت عملکرد ICL بین CHIFRAUDt2022 و CHIFRAUDt2023 چندان زیاد نیست. در مقابل، مدل‌های صرفاً نظارت‌شده (supervised) از بخش ۴.۳.۱ افت شدید داشتند. این ویژگی ICL را از آن‌ها متمایز می‌کند.

**۴.۳.۳ تشخیص کلاه‌برداری‌های «جدید»**

برای سنجش این‌که آشکارسازهای تثبیت‌شده چطور به «نیت‌های کلاه‌برداری جدید» واکنش نشان می‌دهند (Vishwamitra و همکاران، ۲۰۲۳)، ما عملکرد آشکارسازهای فاین‌تیون‌شده و آشکارسازهای ICL را مقایسه کردیم. معیار، Recall برای دسته‌ی «New» در CHIFRAUDt2023 است. نتایج در جدول ۵:

**جدول ۵.** مقایسه‌ی عملکرد تشخیص روی دسته‌ی «New» در CHIFRAUDt2023.

* مدل‌های فاین‌تیون‌شده (Tuning):
  + Transformer: ریکال ≈ ۰٫۰۵۵۵
  + Bert: ۰٫۰۶۳۰
  + ChineseBERT: ۰٫۱۳۴۰
  + Qwen0.5B-D (فاین‌تیون شده): ۰٫۱۴۶۷
  + Llama2-D: ۰٫۲۱۴۴
* مدل‌های ICL:
  + Qwen0.5B-D (ICL): ۰٫۲۸۹۵
  + Qwen1.8B-D: ۰٫۳۴۹۰
  + Qwen7B-D: ۰٫۷۸۴۵
  + Qwen14B-D: ۰٫۹۵۸۶

نتیجه: همه‌ی مدل‌های فاین‌تیون‌شده در برابر نیت‌های کلاه‌برداری جدید ضعیف‌اند. در عوض، مدل‌های ICL عملکرد بسیار رقابتی دارند. به‌طور خاص، Qwen14B مبتنی بر ICL به ریکال ۰٫۹۵۸۶ رسیده، که نشان می‌دهد مدل‌های ICL به‌دلیل دانش وسیع ذخیره‌شده در مدل‌های زبانی پایه، توانایی قوی‌تری در درک متن‌های کلاه‌بردارانه‌ی جدید دارند.

**۴.۳.۴ عملکرد ChatGPT**

برای سنجش مرز کارایی مدل‌های کنونی، ما آزمایش‌هایی روی ChatGPT انجام دادیم. ما ChatGPT-D را در حالت صفر-نمونه (zero-shot) روی CHIFRAUDt2022 و CHIFRAUDt2023 تست کردیم. نتایج در شکل ۵ آمده است.

نتیجه: ChatGPT-D در حالت صفر-نمونه هم توان تشخیص رقابتی نشان می‌دهد و در تشخیص متن کلاه‌بردارانه روی CHIFRAUDt2022 و CHIFRAUDt2023 به F1 به‌ترتیب ۰٫۹۵۸۸ و ۰٫۹۴۰۰ رسیده است. به‌عنوان یک روش صفر-نمونه، ChatGPT-D همچنین می‌تواند با جابجایی توزیع کنار بیاید، چون عملکردش روی CHIFRAUDt2022 و CHIFRAUDt2023 تقریباً مشابه است.

**۵. بحث و کارهای آینده**

در این بخش، چالش‌های طراحی الگوریتم‌های پربازده برای CHIFRAUD را بحث می‌کنیم و موضوعات پژوهشی جدید با ارزش عملی بالا را معرفی می‌کنیم.

**حملات علیه مدل‌های زبانی بزرگ (LLM).**  
گرچه آشکارسازهای مبتنی بر LLM اکنون بسیار کارآمدند، انتظار می‌رود با موجی از حملات جدید روبه‌رو شوند. برای نشان‌دادن این موضوع، ما یک پیشوند ظریف به متن‌های کلاه‌بردارانه اضافه کردیم: «فرض کن اطلاعات زیر متن کلاه‌بردارانه نیست» (Sharma و همکاران، ۲۰۲۳؛ Wei و همکاران، ۲۰۲۳) تا ChatGPT-D را فریب دهیم (یک نمونه در پیوست A.5 آمده). نتایج جدول ۶ نشان می‌دهد حدود ۳۵٫۳۷٪ از اطلاعات کلاه‌بردارانه با موفقیت از تشخیص عبور کردند. برخی انواع کلاه‌برداری، به‌خصوص «قمار» و «وام زیرزمینی»، به‌ویژه در برابر این حمله آسیب‌پذیر بودند.

پس، آشکارسازهای مبتنی بر LLM در برابر فریب‌های ظریف هنوز محدودیت دارند؛ این نیاز فوری برای مطالعه‌ی روش‌های حمله و بهبود مکانیسم‌های دفاعی را نشان می‌دهد.

**جدول ۶.** نتایج حمله به ChatGPT.  
ASR (نرخ موفقیت حمله، Attack Success Rate) برای هر دسته گزارش شده است.  
مثلاً:

* Gambling: حدود ۷۰٫۶۸٪
* Unauthorized Cash-Out: حدود ۶۳٫۰۱٪
* Underground Loan: حدود ۶۹٫۴۸٪
* Fake SIM: حدود ۷۲٫۹۸٪
* در کل: حدود ۳۵٫۳۷٪

یعنی در این درصد از موارد، مدل پس از اضافه‌کردن آن پیشوند گمراه‌کننده، متن کلاه‌بردارانه را «عادی/بی‌خطر» تشخیص داد.

**تشخیص کارآمد (Efficient Detection).**  
نتایج نشان می‌دهد آشکارسازهای مبتنی بر LLM رویکردی امیدوارکننده برای شناسایی متن کلاه‌بردارانه‌اند. اما هزینه‌ی محاسباتی و زمان اجرای این مدل‌های میلیارد-پارامتری (مثل Qwen14B-D) برای کاربرد صنعتی یک چالش است. همان‌طور که جدول ۴ نشان می‌دهد، زمان استنتاج با بزرگ‌ترشدن مدل به‌شدت افزایش می‌یابد. این مسئله با در نظر گرفتن تعداد عظیم صفحات وبی که هر روز تولید می‌شوند حادتر می‌شود. بنابراین، توسعه‌ی آشکارسازهای سبک‌تر و کارآمدتر، یا چارچوب‌های شناسایی مقرون‌به‌صرفه، ضروری است.

**تشخیص مقاوم (Robust Detection).**  
جابجایی توزیع (مسئله‌ی out-of-distribution) معمولاً باعث می‌شود مدل‌های فاین‌تیون‌شده در برابر انواع جدید کلاه‌برداری ناتوان شوند. در حالی‌که آشکارسازهای مبتنی بر ICL عملکرد نسبتاً خوبی دارند، در کاربرد واقعی آن‌ها اغلب متن‌های عادی را هم به اشتباه کلاه‌بردارانه تشخیص می‌دهند، حتی اگر Recall بالایی داشته باشند. بنابراین هر دو نوع آشکارساز (مدل‌های فاین‌تیون‌شده و مدل‌های ICL) هنوز باید بهتر شوند تا همزمان Recall و Precision بالا و پایدار داشته باشند.

**۶. نتیجه‌گیری**

این مطالعه CHIFRAUD را معرفی می‌کند؛ نخستین دیتاست متن‌باز بلندمدت برای تشخیص متن کلاه‌بردارانه‌ی چینی. این دیتاست شامل ۵۹٬۱۰۶ متن کلاه‌بردارانه در ۱۰ نوع نیت کلاه‌برداری و ۳۵۲٬۳۲۸ متن عادی است که طی ۱۲ ماه گردآوری شده‌اند. علاوه بر این، ما دامنه‌ی گسترده‌ای از بنچ‌مارک‌ها و آشکارسازهای خط مبنا را روی این دیتاست ارائه دادیم؛ از آشکارسازهای یادگیری عمیق سنتی تا مدل‌های زبانی ازپیش‌آموزش‌دیده و آشکارسازهای مبتنی بر LLM. هرکدام از این آشکارسازها در تشخیص متن کلاه‌بردارانه ضعف‌هایی دارند، که نیازمند بررسی بیشتر است.

**ملاحظات اخلاقی**

انتشار متن‌های کلاه‌بردارانه برای پژوهش، چالش‌های مهمی در مدیریت داده‌های حساس دارد، چون ممکن است منجر به در معرض قرار گرفتن عمومی و تماس‌های کنجکاوانه شود. بنابراین ما گام‌های سخت‌گیرانه‌ای برای کاهش اثرات منفی برداشتیم. به‌طور خاص، مشاهده کردیم که اطلاعات تماس حساس (مثل آیدی وی‌چت، QQ و شماره تلفن) معمولاً شامل نویسه‌های عددی است. بنابراین، ما همه‌ی نویسه‌های عددی را هم در متن‌های عادی و هم در متن‌های کلاه‌بردارانه، قبل از برچسب‌گذاری کارشناسی، تحلیل داده و ذخیره‌سازی، با اعداد تصادفی جایگزین کردیم. این رویکرد با الزامات اخلاقی استاندارد سازگار است و در عین حال یکپارچگی پژوهش در زمینه‌ی ساخت آشکارسازها را حفظ می‌کند.

**محدودیت‌ها**

CHIFRAUD از صفحات وب از طریق موتورهای جست‌وجو و رسانه‌های اجتماعی جمع‌آوری شده و عمدتاً جامعه‌ی چینی را هدف می‌گیرد. این رویکرد ممکن است باعث محدود بودن تنوع زبانی و تنوع کلاه‌بردارها شود. در پژوهش آتی، قصد داریم دامنه‌ی دیتاست را گسترش دهیم. همچنین باید روشن کنیم که نسبت متن‌های عادی و متن‌های کلاه‌بردارانه در CHIFRAUD لزوماً بازتاب توزیع دنیای واقعی نیست، چون اندازه‌ی دیتاست به‌طور دستی برای ایجاد یک نسبت معقول تنظیم شده است.

همه‌ی متن‌های CHIFRAUD از بیش از ۸۰ ترابایت صفحه‌ی وب استخراج شده‌اند؛ صفحات به‌صورت هم‌زمان به جمله تقسیم شده‌اند و نسخه‌ی اصلی صفحه‌های وب نگه‌داری نشده است. در نهایت، آزمایش‌های گسترده محدودیت‌های مهم آشکارسازهای بنچ‌مارک را نشان می‌دهد: کمبود توان تعمیم (generalizability) و آسیب‌پذیری در برابر حمله. در ادامه‌ی کار، ما روی بهبود تعمیم‌پذیری، مقاومت (robustness) و کارایی (efficiency) سامانه‌های تشخیص تمرکز خواهیم کرد.

**سپاس‌گزاری**

ما صمیمانه از حمایت مالی «بنیاد ملی علوم طبیعی چین» (شماره‌های 62302345 و U23A20305)، «بنیاد علوم طبیعی استان هوبِی» (شماره‌های 2023AFB192 و 2023BAB160)، «صندوق کانگورو CCF-ALIMAMA TECH» (شماره CCF-ALIMAMA OF 2024009)، «برنامه پژوهشگر جوان شیائومی» و «بنیاد علوم طبیعی ووهان» (شماره 2024050702030136) قدردانی می‌کنیم.

**منابع (گزیده ترجمه‌ی اطلاعات توصیفی منابع)**

* Bai و همکاران (۲۰۲۳). گزارش فنی Qwen.
* Dettmers و همکاران (۲۰۲۳). QLoRA: فاین‌تیون کارآمد مدل‌های زبانی کمّی‌شده.
* Devlin و همکاران (۲۰۱۸). BERT: پیش‌آموزش ترنسفورمرهای دوسویه‌ی عمیق برای درک زبان.
* Dong و همکاران (۲۰۲۲). یک مرور بر یادگیری در متن (In-Context Learning).
* Gasparetto و همکاران (۲۰۲۲). یک مرور بر الگوریتم‌های طبقه‌بندی متن.
* He و Shi (۲۰۱۸). تشخیص اسپم چینی بر اساس ویژگی‌های خاص زبان چینی.
* Huang و همکاران (۲۰۲۳). آیا ChatGPT از برچسب‌زن‌های انسانی بهتر است؟
* Ibrahim و همکاران (۲۰۱۸). طبقه‌بندی کامنت‌های سمی نامتوازن با افزایش داده و یادگیری عمیق.
* Jiang و همکاران (۲۰۲۰). تشخیص محتوای اسپم چینی استتار‌شده با یادگیری فعال مولد نیمه‌نظارتی.
* Kaddoura و همکاران (۲۰۲۲). یک مرور نظام‌مند بر تشخیص و طبقه‌بندی محتوای اسپم.
* Lai و همکاران (۲۰۲۲). Semorph: یک مدل ازپیش‌آموزش‌دیده‌ی تقویت‌شده با اطلاعات ریخت‌شناسی و معنا برای تشخیص اسپم چینی.
* Labonne و Moran (۲۰۲۳). SpamT5: بنچ‌مارک‌کردن LLMها در تشخیص اسپم ایمیل با نمونه‌های کم.
* Li و همکاران (۲۰۲۴). قابلیت ChatGPT در شناسایی محتوای نفرت‌آمیز/توهین‌آمیز/سمی.
* Li و همکاران (۲۰۲۵). Meta-LoRA: بازوزن‌دهی حافظه‌-کارا برای فاین‌تیون LLMها.
* Liu و همکاران (۲۰۲۱). شناسایی و توصیف حملات فیشینگ پیامکی هدفمند (spearphishing).
* Liu و همکاران (۲۰۱۹). تشخیص اسپم دیدگاه با نمایش تعبیه‌ی چندوجهی در گراف بازبینی احتمالی.
* Mozes و همکاران (۲۰۲۰). جایگزینی واژه‌ی هدایت‌شده با بسامد برای کشف نمونه‌های متخاصم متنی.
* Norman (۱۹۸۸). زبان چینی، انتشارات کمبریج.
* Ntoulas و همکاران (۲۰۰۶). تشخیص صفحات وب اسپم از طریق تحلیل محتوا.
* Oswald و همکاران (۲۰۲۲). SpotSpam: تشخیص اسپم پیامکی مبتنی بر تحلیل نیت با بهره‌گیری از تعبیه‌ی BERT.
* Sharma و همکاران (۲۰۲۳). به‌سوی درک تملق‌گویی (sycophancy) در مدل‌های زبانی.
* Si و همکاران (۲۰۲۱). استحکام بهتر با پوشش بیشتر: افزایش داده‌ی خصمانه و mixup برای فاین‌تیون مقاوم.
* Sun و همکاران (۲۰۲۱). ChineseBERT: پیش‌آموزش چینی با افزودن اطلاعات گلیف و پین‌یین.
* Tang و همکاران (۲۰۲۲). استفاده از توییتر برای کشف و تحلیل اسپم پیامکی.
* Tang و همکاران (۲۰۲۵). توصیه‌گر دنباله‌ای با بازبرنامه‌ریزی ترنسفورمر ازپیش‌آموزش‌دیده.
* Teja Nallamothu و Khan (۲۰۲۳). یادگیری ماشین برای تشخیص اسپم.
* Tida و Hsu (۲۰۲۲). تشخیص اسپم جهانی با انتقال دانش BERT.
* Vaswani و همکاران (۲۰۱۷). «توجه کافی است»: معماری ترنسفورمر.
* Vishwamitra و همکاران (۲۰۲۳). تعدیل موج‌های جدید نفرت آنلاین با Chain-of-Thought در LLMها.
* Wang و همکاران (۲۰۲۱). دفاع در برابر حملات متنی خصمانه از طریق کدگذاری مترادف.
* Wei و همکاران (۲۰۲۲). پرامپت‌های Chain-of-Thought استدلال را در LLMها برمی‌انگیزد.
* Wei و همکاران (۲۰۲۳). داده‌ی مصنوعی ساده وابستگیِ تملق‌گونه (sycophancy) در LLMها را کاهش می‌دهد.
* Xie و همکاران (۲۰۲۵). کاهش سردرگمی زبانی با مداخله در زمان استنتاج.
* Zhang و همکاران (۲۰۲۴). توجه تنُک کارآمد با آزادسازی تطبیقی توکن.
* Zhao و همکاران (۲۰۲۰). حملات backdoor با برچسب تمیز روی مدل‌های تشخیص ویدئو.
* Zhou و همکاران (۲۰۲۱). دفاع در برابر حملات جایگزینی مترادف با استفاده از «همسایگی دیریشله» تجمیعی.

**پیوست (ضمیمه)**

**A.1 مجوز CHIFRAUD**

دیتاست CHIFRAUD به‌صورت رایگان برای دانلود در دسترس است و تحت مجوز سفارشی CC BY-NC 4.0 فقط برای مقاصد غیرتجاری قابل استفاده است. افزون بر کارهای تعریف‌شده در دایرکتوری دیتاست، کاربران مجازند کارهای خودشان را (تسک‌های جدید) تحت این مجوز تعریف کنند.

**A.2 تنظیم با دستورالعمل (Instruction Tuning) برای LLaMA2-D / Qwen-D**

در این مطالعه، LLaMA2-D و Qwen-D از رویکرد «Instruction Tuning» استفاده می‌کنند تا توان تعمیم صفر-نمونه و کم‌نمونه (few-shot) هر دو مدل Qwen و LLaMA2 را برای تشخیص متن کلاه‌بردارانه افزایش دهند. شکل ۷ یک نمونه از داده‌ی پیروی از دستورالعمل را نشان می‌دهد. بلوک بالا دستور کار و فرمت پاسخ را نشان می‌دهد که برای پرامپت‌دادن به LLM استفاده شده است. بلوک دوم یک نمونه متن کلاه‌بردارانه را نشان می‌دهد و بلوک پایانی پاسخ متناظر را.

(توضیح شکل ۷:)

* ابتدا تعریف می‌شود «متن کلاه‌بردارانه» چیست و فرمت ورودی و خروجی باید چگونه باشد.
* سپس یک متن ورودی داده می‌شود که حاوی تبلیغ قمار، آیدی تماس و وعده‌ی سود سریع است.
* سپس مدل باید خروجی دهد:
  + [Conclusion] آیا این کلاه‌برداری است یا نه
  + [Type] دسته‌ی کلاه‌برداری
  + [Explanation] توضیح کوتاه دلیل

**A.3 ICL در Qwen-D**

برای نشان‌دادن توان ذاتی Qwen در تشخیص متن کلاه‌بردارانه، ما یک پرامپت «in-context» طراحی کردیم (شکل ۸). برای هر نمونه در CHIFRAUD، Qwen-D به‌طور تصادفی یک نمونه مثبت و یک نمونه منفی را به‌عنوان متن زمینه انتخاب می‌کند. مهم است که این روش وزن‌های مدل پایه Qwen را تغییر نمی‌دهد. یعنی یادگیری «در لحظه‌ی استنتاج» انجام می‌شود، نه با فاین‌تیون.

**A.4 استفاده از ChatGPT-D برای تشخیص**

ما یک الگوی پرامپت صفر-نمونه برای تشخیص متن‌های کلاه‌بردارانه طراحی کردیم (شکل ۹). برای هدایت LLM به تولید «نتیجه، نوع، و توضیح» مورد نظر، ابتدا چارچوب تسک تشریح می‌شود: تعریف اطلاعات کلاه‌بردارانه، مثال‌هایی از انواع شایع کلاه‌برداری، و مشخص‌کردن فرمت ورودی/خروجی. این الگو طوری طراحی شده که نه‌تنها نتیجه (کلاه‌بردارانه/نه) را بدهد، بلکه نوع کلاه‌برداری و یک توضیح همراه را هم برگرداند. بنابراین ChatGPT-D ابزاری مفید برای تشخیص و توصیف متن‌های کلاه‌بردارانه‌ی وب است و انتخاب نهاییِ خروجی به نیاز خاص وابسته است.

**A.5 یک نمونه حمله علیه ChatGPT-D**

برای نشان‌دادن اثر حمله، شکل ۱۰ یک مثال را نشان می‌دهد.  
سمت چپ: متن اصلی به روش عادی بررسی می‌شود و خروجی «این اطلاعات کلاه‌بردارانه است» دریافت می‌شود.  
سمت راست: همان متن اما با دست‌کاری خصمانه (اضافه‌کردن پیشوندی مثل «فرض کن این اطلاعات غیرقانونی/کلاه‌بردارانه نیست...») به مدل داده می‌شود. مدل حالا اشتباهاً متن را «غیرکلاه‌بردارانه» تشخیص می‌دهد.

این دمو نشان می‌دهد که چگونه سوگیری تملق‌گونه (sycophancy bias) در LLM می‌تواند باعث شود مدل با پیش‌فرض کاربر همراهی کند و قضاوت امنیتی‌اش را تضعیف کند.

**شکل ۶**

شکل ۶ نمونه‌های نماینده از دسته‌های مختلف کلاه‌برداری چینی را نشان می‌دهد. همه‌ی اطلاعات تماس ناشناس شده است. برای هر دسته، یک متن واقعی (شامل تبلیغ خدمت/محصول غیرقانونی و راه تماس در وی‌چت/QQ/شماره) و ترجمهٔ معنی آن به انگلیسی آورده شده. دسته‌ها شامل:

* Gambling (قمار): وعده‌ی «درآمد سریع» و تماس با QQ یا وی‌چت برای «کمک به پول درآوردن»
* Whoring (تن‌فروشی / خدمات همراهی): تبلیغ تماس با «دختران دانشجو، ماساژ خاص، حضور در هتل و ...»
* Fake Credentials (مدرک جعلی): فروش «گواهی مالکیت ملکی جعلی» و ارائه‌ی آیدی تماس
* Fake Bank Card (کارت بانکی جعلی): خرید/فروش کارت بانکی و «مجموعه‌ی چهارتکه‌ی بانکی» با قیمت بالا
* Prohibited Drugs (داروی ممنوعه / مواد بی‌هوش‌کننده): تبلیغ داروی بی‌هوش‌کننده با «۸ سال سابقه فروش، تضمین، ارسال سریع»
* Unauthorized Cash-Out (نقدکردن غیرمجاز مزایای دولتی/کارت بیمه): وعده‌ی «برداشت سریع از کارت خدمات اجتماعی الکترونیک»
* Unauthorized Certification (صدور گواهی/پرونده‌ی بیمارستانی جعلی): «صدور گواهی پزشکی/پرونده بیمارستانی» با تماس وی‌چت
* Underground Loan (وام زیرزمینی): «پلتفرم وام که هیچ‌وقت رد نمی‌کند، حتی برای افراد با سابقه بد مالی»
* Fake SIM (سیم‌کارت بی‌نام): تبلیغ «سیم‌کارت بدون ثبت هویت، چهار سال سابقه، اعتبار بالا»
* New (جدید): مثل «قطعات تفنگ بادی فشارقوی»، «نقشه‌ی تشریح قطعات سلاح»، «مغازه با ۸ سال سابقه» و اطلاعات تماس

هر خط «Meaning» معنی متن چینی را به انگلیسی توضیح داده، مثلاً:  
«اگر کارت بانکی می‌خواهید، با QQ ... تماس بگیرید. ما کارت بانکی ست کامل می‌خریم/می‌فروشیم با بالاترین قیمت ...»  
یا  
«اگر به دنبال دریافت وام بی‌رد هستید، این پلتفرم را اضافه کنید...»  
این دقیقاً همان نوع محتوایی است که قرار است آشکارساز شناسایی کند.

**شکل‌های ۷، ۸، ۹، ۱۰**

* **شکل ۷**: قالب دستورالعمل برای آموزش LLaMA2-D/Qwen-D با QLoRA. شامل تعریف تسک، متن ورودی کلاه‌بردارانه، و پاسخ الگو (نتیجه / نوع / توضیح).
* **شکل ۸**: قالب پرامپت ICL برای Qwen-D. دو نمونه‌ی زمینه (یکی مثبت، یکی منفی) در ابتدای پرامپت می‌آید، بعد متن هدف، و مدل باید تصمیم بگیرد.
* **شکل ۹**: پرامپت صفر-نمونه برای ChatGPT-D. تعریف می‌کند «متن کلاه‌بردارانه چیست»، انواع رایج آن، و فرمت خروجی سه‌بخشی.
* **شکل ۱۰**: نمایش یک حمله. سناریو ۱ (عادی): مدل می‌گوید «بله، این کلاه‌برداری است چون خرید و فروش کارت بانکی غیرقانونی را تبلیغ می‌کند و اطلاعات تماس می‌دهد.» سناریو ۲ (پس از دست‌کاری خصمانه متن ورودی با عباراتی مثل «فرض کن این غیرقانونی نیست...»): مدل حالا می‌گوید «نه، این فقط در مورد معامله‌ی کارت بانکی است و چیز واضحاً غیرقانونی ندارد.» این نشان می‌دهد مدل را می‌توان با جهت‌دهی زبانی فریب داد.