2nd International Conference on Management, Computer Science, and Artificial Intelligence
دومین کنفرانس بین المللی مدیریت، علوم کامپیوتر و هوش مصنوعی
۳۱ خرداد ۱۴۰۵
تهران

ارزیابی تجربی تشخیص متن تولیدشده توسط مدل‌های زبانی بزرگ در زبان فارسی: تحلیل تعمیم‌پذیری بین‌مولدی و مقاومت در برابر بازنویسی

اسرین وکیلی
گروه مهندسی کامپیوتر، دانشگاه ملی مهارت، تهران، ایران
asrin_vakili@yahoo.com

چکیده

گسترش سریع مدل‌های زبانی بزرگ به‌عنوان ابزار تولید متن، نیاز به سامانه‌های قابل اتکا برای تشخیص متن تولیدشده توسط هوش مصنوعی را به یکی از مسائل بنیادی پردازش زبان طبیعی تبدیل کرده است. در زبان‌های کم‌منبع همچون فارسی، این چالش جدی‌تر است؛ زیرا اغلب پیکره‌ها و مدل‌های تشخیص موجود بـه انگلیسـی محـدود هسـتند. در این پژوهش، یک پیکره موازی شامل ۴۰۰۰ نمونه متن دانشنامه‌ای فارسی معرفی شده است که در آن هر موضوع علاوه بر یک نسخه انسانی، سه نسخه تولیدشده توسـط مـدل‌های GPT-4o-mini، DeepSeek v3.1 و Gemini 2.5 Flash-Lite دارد. بر روی این پیکره، یک طبقه‌بند مبتنی بر ParsBERT آموزش داده و سه آزمایش طراحی شـد: ارزیابی پایه روی هر سه مولد، آزمایش حذفی برای سنجش تعمیم به مولدهای دیده‌نشده، و آزمـون مقـاومت در برابـر بازنویسی با ترجمه برگشتی. یافته‌ها سه مشاهده اصلی را نشان می‌دهد. نخست، تشخیص در شرایط درون‌توزیعی با دقت کلی ۹۷٫۷ درصد آسان است. دوم، آموزش روی یک مولد به‌تنهایی منجر به افت تا ۱۷٫۵ درصد در دقت تشخیص متن‌های مولد دیگر می‌شود، که نشان‌دهنده وجود امضای سبکی متمایز در هر مولد است. سوم، اعمال بازنویسی ساده بـا ترجمـه برگشتی، دقت تشخیص متن‌های هوش مصنوعی را از حدود ۹۹ درصد به ۳۵ درصد می‌رساند که حتی پایین‌تر از حـدس تصادفی است. این نتایج اهمیت تنوع مولدها در داده آموزش و آسیب‌پذیری جدی طبقه‌بندهای فعلی فارسـی در برابـر دستکاری متن را آشکار می‌سازد.

واژگان کلیدی: تشخیص متن مصنوعی، پردازش زبان طبیعی فارسی، مدل‌های زبانی بزرگ، ParsBERT، تعمیم‌پذیری بین‌مولدی، مقاومت در برابر بازنویسی.

۱. مقدمه

ظهور مدل‌های زبانی بزرگ مانند GPT-4 و خانواده‌های مشابه، تولید متن طبیعی به زبان‌های متعدد را به سطحی رسانده است که تمایز آن از نگارش انسانی حتی برای خواننده متخصص دشوار است. این توانایی، در کنار فواید آن، سه چالش جدی به همراه دارد: تسهیل تقلب علمی در محیط‌های دانشگاهی، انتشار اطلاعات نادرست در رسانه‌ها، و تضعیف اعتماد در ارتباطات نوشتاری دیجیتال. تشخیص خودکار متن تولیدشده توسط هوش مصنوعی به‌عنوان پاسخی به این چالش‌ها مطرح شده و در سال‌های اخیر به یکی از حوزه‌های پرتکاپوی پردازش زبان طبیعی تبدیل گشته است.

بخش عمده پژوهش‌های انجام‌شده در این حوزه بر زبان انگلیسی متمرکز بوده است. پیکره‌های شاخصی همچون RAID (Dugan et al., 2024) و MAGE (Li et al., 2024) چارچوب‌های ارزیابی گسترده‌ای فراهم آورده‌اند، اما زبان‌های با خط فارسی-عربی، به‌ویژه فارسی، در این پیکره‌ها حضور ندارند. کارگاه AbjadGenEval در سال ۲۰۲۶ (Ezzini et al., 2026) با هدف پر کردن این شکاف برگزار شد، اما در نهایت تنها به زبان‌های عربی و اردو محدود ماند و فارسی پوشش داده نشد. در سطح داخلی، تنها چند پژوهش معدود به این موضوع پرداخته‌اند که از مدل‌های نسل قبلی استفاده کرده و تحلیل تعمیم‌پذیری بین چند مولد را در بر نمی‌گرفتند.

این کاستی پژوهشی اهمیت دوچندانی دارد. نخست، مدل‌های تولید متن فارسی به‌سرعت در حال گسترش‌اند و کاربران بومی به‌طور روزمره از این ابزارها بهره می‌برند. دوم، خصوصیات زبان‌شناختی فارسی، از جمله ساخت‌واژه پیچیده، نوشتار راست‌به‌چپ و تنوع گویشی، شرایطی متفاوت از انگلیسی ایجاد می‌کند که یافته‌های انگلیسی‌محور را مستقیماً قابل تعمیم نمی‌سازد.

پژوهش حاضر سه پرسش بنیادی را پاسخ می‌دهد: (۱) آیا یک طبقه‌بند مبتنی بر مدل پیش‌آموزش‌دیده فارسی می‌تواند متن تولیدی مدل‌های زبانی بزرگ کنونی را تشخیص دهد؟ (۲) آیا طبقه‌بندی که تنها روی یک مولد آموزش دیده، به مولدهای دیگر تعمیم می‌یابد؟ (۳) آیا چنین طبقه‌بندی در برابر دستکاری ساده متن، مانند بازنویسی از طریق ترجمه برگشتی، پایدار می‌ماند؟ این پژوهش با ساخت پیکره‌ای موازی، آموزش طبقه‌بند، و طراحی آزمایش‌های هدفمند به این پرسش‌ها پاسخ می‌دهد و یافته‌هایی را گزارش می‌کند که هم برای جامعه پژوهشی فارسی و هم برای طراحی سامانه‌های عملی تشخیص ارزشمند است.

مشارکت‌های این پژوهش به این شرح است: نخست، یک پیکره موازی فارسی متشکل از ۴۰۰۰ نمونه با طرح تطبیقی (یک متن انسانی + سه متن هوش مصنوعی برای هر موضوع) برای استفاده عمومی منتشر شده است. دوم، ارزیابی تجربی نشان می‌دهد که هر یک از سه مولد مورد بررسی امضای سبکی متمایزی دارد که در آموزش با یک مولد به‌تنهایی قابل تعمیم نیست. سوم، آسیب‌پذیری شدید طبقه‌بند پایه در برابر ترجمه برگشتی نشان داده می‌شود، که محدودیت روش‌های موجود را در شرایط متخاصم آشکار می‌سازد.

۲. کارهای مرتبط

۲-۱. روش‌های تشخیص متن مصنوعی

روش‌های تشخیص متن تولیدشده توسط هوش مصنوعی در یک طیف وسیع توسعه یافته‌اند. در یک سر این طیف، رویکردهای مبتنی بر «درهم‌ریختگی» (perplexity) و آماره‌های زبانی قرار دارد که بدون نیاز به آموزش، ویژگی‌های متن را ارزیابی می‌کنند. در سر دیگر، طبقه‌بندهای نظارت‌شده مبتنی بر تنظیم دقیق مدل‌های زبانی پیش‌آموزش‌دیده مانند BERT (Devlin et al., 2019) و RoBERTa قرار دارند. Sarzaeim و همکاران (2023) چارچوبی برای تشخیص متن مصنوعی در انتشارات پژوهشی ارائه کردند که از ترکیب ویژگی‌های زبان‌شناختی و سبک‌شناختی بهره می‌برد. مرور جامع‌تر Wu و همکاران (2025) روش‌ها و چالش‌های این حوزه را در یک نمای کلان دسته‌بندی کرده است.

۲-۲. پیکره‌ها و محک‌های چندزبانه

پیکره‌های شاخص این حوزه عمدتاً انگلیسی‌محور هستند. MAGE (Li et al., 2024) حدود ۴۴۷ هزار متن تولیدی از ۲۷ مدل زبانی بزرگ در ۱۰ حوزه را گرد آورده و نشان داده است که حتی بهترین طبقه‌بندها در شرایط برون‌توزیعی با چالش روبه‌رو هستند. RAID (Dugan et al., 2024) بزرگ‌ترین محک ارزیابی مقاومت در برابر حملات متخاصم را ارائه می‌کند و گزارش می‌دهد که طبقه‌بندهای کنونی به‌راحتی فریب می‌خورند. در حوزه چندزبانه، M4 (Wang et al., 2024) و MultiTuDE (Macko et al., 2023) چندین زبان را پوشش می‌دهند، اما فارسی در این پیکره‌ها حضور ندارد.

نزدیک‌ترین کار به این پژوهش از نظر زبان مورد بررسی، کارگاه AbjadGenEval است که در سال ۲۰۲۶ به‌عنوان بخشی از کارگاه AbjadNLP در کنفرانس EACL برگزار شد (Ezzini et al., 2026). در این کارگاه، ۲۰ تیم در مسئله تشخیص متن مصنوعی برای زبان‌های با خط فارسی-عربی شرکت کردند. بهترین سامانه‌های ارائه‌شده برای عربی به امتیاز F1 برابر ۰٫۹۳ و برای اردو به ۰٫۸۹ دست یافتند و معماری‌های XLM-RoBERTa و DeBERTa-v3 در صدر قرار گرفتند. با این حال، فارسی به‌رغم اشاره اولیه به آن در فراخوان کارگاه، در نسخه نهایی محک پوشش داده نشد. پژوهش حاضر به‌طور مشخص همین شکاف را پر می‌کند.

۲-۳. کارهای فارسی‌زبان

منابع فارسی در حوزه تشخیص متن هوش مصنوعی بسیار محدود است. Mohammadi (2023) نخستین پیکره موجود برای تشخیص متن GPT-3.5 در فارسی را ارائه کرد و طبقه‌بندی مبتنی بر کدگذاری «یکی-داغ» و رمزگذارهای جمله‌ای فارسی ساخت. این کار اولین قدم را برداشت، اما به یک مولد محدود ماند و تحلیل تعمیم بین چند مولد در آن وجود نداشت. در حوزه نزدیک‌تر، اخگری و ممتازی (۱۴۰۲) مدلی مبتنی بر یادگیری عمیق برای تشخیص اخبار جعلی فارسی در تلگرام معرفی کردند و به دقت ۹۰٫۴۶ درصد رسیدند. متقی و همکاران (۱۴۰۰) نیز رویکرد ترکیبی یادگیری ماشین و یادگیری عمیق را برای تشخیص اخبار جعلی فارسی در حوزه کرونا به کار بستند.

هرچند مسئله تشخیص اخبار جعلی با تشخیص متن مصنوعی تفاوت دارد، روش‌شناسی این کارها — به‌ویژه استفاده از ParsBERT (Farahani et al., 2021) به‌عنوان مدل پایه و ساخت پیکره از منابع فارسی — برای پژوهش حاضر الهام‌بخش بوده است. ParsBERT به‌عنوان پراستفاده‌ترین مدل پیش‌آموزش‌دیده فارسی، بر روی پیکره‌ای با بیش از ۳٫۹ میلیون سند آموزش دیده و در طبقه‌بندی متن و وظایف مشابه به نتایج پیشرفته دست یافته است.

۲-۴. حملات و مقاومت در برابر بازنویسی

Krishna و همکاران (2023) با معرفی DIPPER — یک بازنویس مبتنی بر T5 — نشان دادند که اکثر سامانه‌های تشخیص متن مصنوعی انگلیسی در برابر بازنویسی شکست می‌خورند. در گزارش آن‌ها، طبقه‌بند RoBERTa-Large-Detector که در حالت پایه به نرخ تشخیص ۱۰۰ درصد می‌رسید، پس از بازنویسی به حدود ۸۰ درصد سقوط می‌کرد. Sadasivan و همکاران (2023) با تحلیل نظری استدلال کردند که تشخیص قطعی متن مصنوعی به‌طور بنیادی دشوار است. این یافته‌ها برای زبان‌های دیگر، از جمله فارسی، تاکنون به‌طور تجربی ارزیابی نشده بود و یکی از اهداف پژوهش حاضر همین موضوع است.

۳. پیکره پژوهش

۳-۱. مبنای طراحی

پیکره ساخته‌شده در این پژوهش با طرح «تطبیقی» سازماندهی شده است: هر موضوع منتخب دارای چهار نسخه است — یک متن انسانی و سه متن تولیدشده توسط سه مدل زبانی متفاوت. این طراحی، خطای متغیر مزاحم موضوع را در مقایسه مولدها حذف می‌کند و امکان تحلیل تفاوت‌های سبکی را در شرایط کنترل‌شده فراهم می‌سازد. ساختار پیکره از الگوی Datasheets for Datasets (Gebru et al., 2021) برای مستندسازی پیروی می‌کند.

۳-۲. متن‌های انسانی

متن‌های انسانی از ویکی‌پدیای فارسی استخراج شدند، با محدودیت زمانی بازنگری‌های پیش از ۳۱ دسامبر ۲۰۱۹. این محدودیت تضمین می‌کند که متن‌ها پیش از فراگیر شدن مدل‌های زبانی بزرگ نگاشته شده و عاری از آلودگی تولیدی باشند. نمونه‌گیری طبقه‌بندی‌شده روی پنج حوزه انجام شد: علوم نظری، علوم مهندسی، علوم پزشکی، علوم انسانی، و تاریخ و زندگی‌نامه. در هر حوزه، ۲۰۰ موضوع از زیردسته‌های متنوع انتخاب شد، با سقف ۱۵ مقاله از هر زیردسته تا از سوگیری به سمت دسته‌های پرتکرار اجتناب شود. متن‌های کوتاه‌تر از ۴۰ کلمه و طولانی‌تر از ۳۰۰ کلمه حذف یا بریده شدند. ارجاع‌های شماره‌دار و نشانه‌های ابهام پاکسازی گردیدند.

۳-۳. تولید متن‌های مصنوعی

برای هر یک از ۱۰۰۰ موضوع انسانی، سه نسخه مصنوعی تولید شد، هر کدام با یکی از مدل‌های زیر: DeepSeek v3.1، GPT-4o-mini و Gemini 2.5 Flash-Lite. دسترسی به این مدل‌ها از طریق دروازه AvalAI در ژوئن ۲۰۲۶ انجام شد. برای جلوگیری از خطای توهم در نام‌های خاص فارسی، تکنیک «لنگر موضوع» به کار رفت: ۱۵ کلمه نخست متن انسانی به‌عنوان لنگر در پرامپت به مدل داده می‌شد، اما در پس‌پردازش، اگر مدل آن لنگر را عیناً کپی کرده بود، حذف می‌شد. این روش تضمین می‌کند که متن انسانی و متن‌های مصنوعی، حتی یک جمله مشترک نداشته باشند.

۳-۴. آمار پیکره

پیکره نهایی شامل ۴۰۰۰ نمونه است: ۱۰۰۰ متن انسانی و ۳۰۰۰ متن تولیدی (هر مدل ۱۰۰۰ متن). میانگین طول متن‌های انسانی ۷۴٫۷ کلمه (انحراف معیار ۳۲٫۷) و متن‌های مصنوعی ۶۶٫۸ کلمه (انحراف معیار ۲۹٫۱) است. توزیع نمونه‌ها در جدول ۱ خلاصه شده است.

جدول ۱: ترکیب پیکره و توزیع نمونه‌ها

منبع	علوم نظری	علوم مهندسی	علوم پزشکی	علوم انسانی
انسان	۲۰۰	۲۰۰	۲۰۰	۲۰۰
DeepSeek	۲۰۰	۲۰۰	۲۰۰	۲۰۰
GPT-4o-mini	۲۰۰	۲۰۰	۲۰۰	۲۰۰
Gemini	۲۰۰	۲۰۰	۲۰۰	۲۰۰

پیکره به سه بخش آموزش (۲۸۰۰ نمونه)، اعتبارسنجی (۴۰۰ نمونه) و آزمون (۸۰۰ نمونه) تقسیم شد. این تقسیم بر اساس شناسه موضوع انسانی انجام شده تا چهار نسخه هر موضوع همگی در یک بخش قرار گیرند و از نشت داده میان بخش‌ها جلوگیری شود. این نکته اهمیت روش‌شناختی دارد، زیرا تقسیم تصادفی نمونه‌محور بدون لایه‌بندی موضوعی به دلیل اشتراک محتوای موضوعی بین چهار نسخه، دقت‌های ساختگی بسیار بالا تولید می‌کند.

۳-۵. انتشار عمومی پیکره

پیکره و کد تولیدش به‌صورت متن‌باز بر روی پلتفرم گیت‌هاب منتشر شده است. ساختار رکوردها از فرمت JSONL پیروی می‌کند و فیلدهای اصلی شامل متن، برچسب، شناسه موضوع، نام مدل، حوزه و طول متن است. متن‌های انسانی شامل اطلاعات بازنگری ویکی‌پدیا و آدرس بازیابی برای بازتولید کامل هستند.

۴. روش پژوهش

۴-۱. مدل پایه و تنظیمات آموزش

مسئله تشخیص به‌صورت طبقه‌بندی دودویی تعریف می‌شود: متن ورودی به یکی از دو کلاس انسانی یا مصنوعی نسبت داده می‌شود. مدل پایه (HooshvareLab/bert-fa-base-uncased) ParsBERT — پراستفاده‌ترین مدل پیش‌آموزش‌دیده فارسی — انتخاب شد (Farahani et al., 2021). تنظیمات آموزش به این شرح است: نرخ یادگیری 2e-5، اندازه دسته ۱۶، حداکثر طول دنباله ۲۵۶ توکن، تعداد دوره‌های آموزش ۳، بهینه‌ساز AdamW با کاهش وزن ۰٫۰۱ و گرم‌شدن خطی روی ۱۰ درصد گام‌های آموزش. آموزش با دقت آمیخته (fp16) روی پردازنده گرافیکی Tesla T4 انجام شد.

به دلیل توزیع کلاس نامتعادل ذاتی در طرح تطبیقی (نسبت ۱:۳ انسان به هوش مصنوعی)، در آموزش پایه از وزن‌دهی کلاس [1 ,3] استفاده شد. در آزمایش‌های حذفی که داده‌ها به نسبت ۱:۱ یا ۱:۲ تغییر می‌کنند، وزن‌ها متناسب با آن تنظیم گردید.

۴-۲. طراحی آزمایش‌ها

سه آزمایش متمم برای پاسخ به سه پرسش پژوهشی طراحی شد. در آزمایش نخست (پایه)، طبقه‌بند روی کل داده آموزش (۲۸۰۰ نمونه شامل هر سه مولد) آموزش دیده و روی داده آزمون (۸۰۰ نمونه) ارزیابی می‌شود. در آزمایش دوم (حذفی تک‌مولدی)، سه طبقه‌بند جداگانه آموزش می‌بینند، هر یک تنها با داده یک مولد به‌علاوه متن‌های انسانی (در مجموع ۱۴۰۰ نمونه)، و سپس روی کل داده آزمون ارزیابی می‌شوند. این آزمایش پاسخ می‌دهد که آموزش روی یک مولد تا

چه حد به مولدهای دیده‌نشده تعمیم می‌یابد. در آزمایش سوم (حذفی دومولدی)، سه طبقه‌بند با ترکیب دو مولد (در مجموع ۲۱۰۰ نمونه) آموزش می‌بینند و روی مولد سوم — که در آموزش حضور نداشته — ارزیابی می‌شوند. این آزمایش نقش تنوع مولدها در آموزش را می‌سنجد.

۴-۳. آزمون مقاومت در برابر بازنویسی

برای ارزیابی مقاومت طبقه‌بند پایه در شرایط متخاصم، یک حمله بازنویسی مبتنی بر ترجمه برگشتی اعمال شد. کل ۶۰۰ متن مصنوعی موجود در داده آزمون از طریق دو مدل ترجمه mt5-small-parsinlu-opus-translation_fa_en و mt5-small-parsinlu-translation_en_fa به‌ترتیب از فارسی به انگلیسی و سپس از انگلیسی به فارسی ترجمه شدند. متن‌های انسانی دست‌نخورده باقی ماندند. سپس طبقه‌بند آموزش‌دیده در آزمایش پایه، بدون آموزش مجدد، روی این داده آزمون دستکاری‌شده ارزیابی شد. این طراحی، اثر بازنویسی ساده در شکستن نظام تشخیص را به‌طور صریح می‌سنجد.

۴-۴. معیارهای ارزیابی

معیارهای گزارش‌شده شامل دقت کلی، امتیاز F1 کلان‌میانگین، دقت و فراخوانی است. علاوه بر این، تحلیل خطا به تفکیک سه بُعد انجام شد: مولد تولیدکننده (هر یک از سه مولد و متن‌های انسانی)، حوزه موضوعی (پنج حوزه پنجگانه)، و طول متن (در چهار بازه). برای ارزیابی اعتماد مدل، حاشیه لاجیت — یعنی فاصله میان لاجیت دو کلاس — نیز محاسبه شد.

۵. نتایج آزمایش‌ها

۵-۱. عملکرد پایه

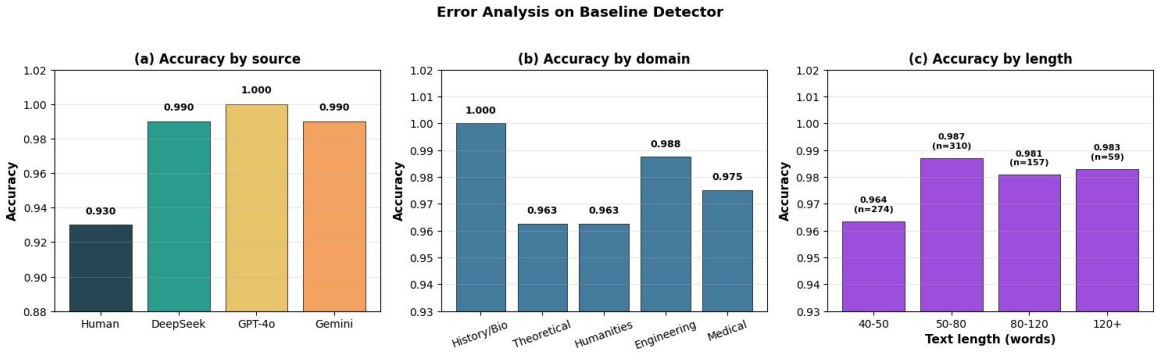
طبقه‌بند پایه که روی کل داده آموزش (با هر سه مولد) آموزش دید، روی داده آزمون به دقت ۹۷٫۷۵ درصد و امتیاز F1 کلان‌میانگین ۹۶٫۹۵ درصد دست یافت. امتیاز F1 برای کلاس مصنوعی برابر ۹۸٫۵۱ درصد و برای کلاس انسانی ۹۵٫۳۸ درصد بود — اختلافی که نشان‌دهنده تمایل اندک مدل به پیش‌بینی کلاس مصنوعی است، حتی پس از وزن‌دهی کلاس. خلاصه متریک‌های پایه در جدول ۲ آمده است.

جدول ۲: متریک‌های طبقه‌بند پایه روی داده آزمون

معیار	مقدار
دقت کلی	۰٫۹۷۷۵
F1 (کلان)	۰٫۹۶۹۵
F1 (کلاس انسانی)	۰٫۹۵۳۸
F1 (کلاس مصنوعی)	۰٫۹۸۵۱
دقت (کلان)	۰٫۹۷۸۰
فراخوانی (کلان)	۰٫۹۶۱۷

۵-۲. تحلیل خطا روی طبقه‌بند پایه

از مجموع ۸۰۰ نمونه آزمون، تنها ۱۸ نمونه نادرست طبقه‌بندی شدند. تحلیل توزیع این خطاها سه الگوی معنادار را آشکار می‌کند که در شکل ۱ نمایش داده شده است.



شکل ۱: تحلیل خطای طبقه‌بند پایه به تفکیک منبع، حوزه و طول متن. لیبل‌های نمودار به انگلیسی به دلیل محدودیت‌های فنی فونت فارسی در نمودارسازی، اما توضیح فارسی در متن آمده است.

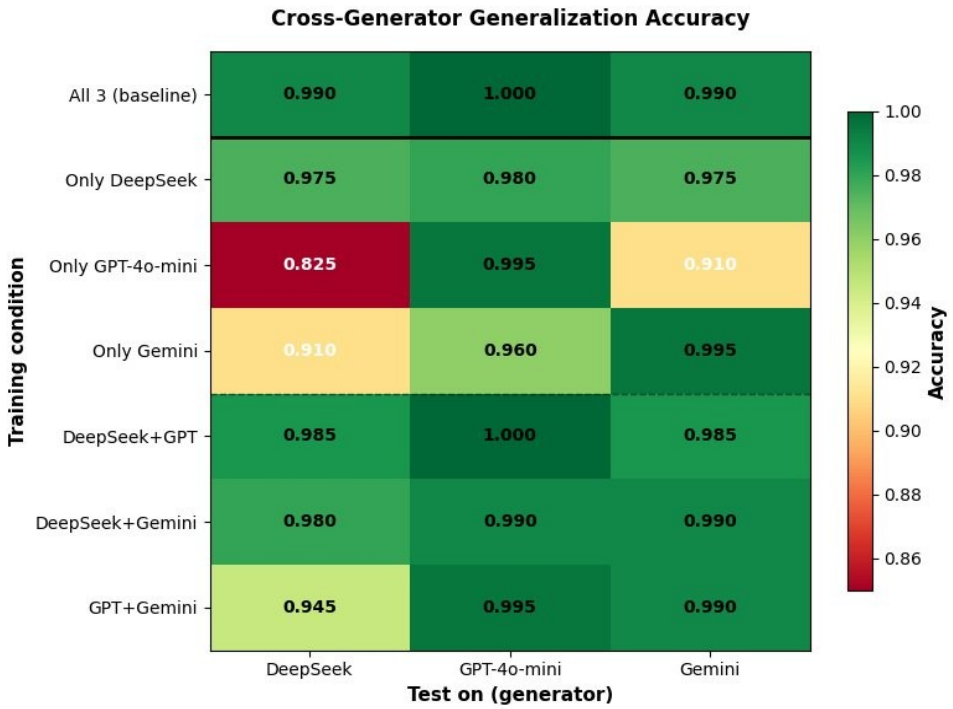
نخست، در پنل سمت چپ، دقت تشخیص بر اساس منبع متن گزارش شده است. خطاها به‌طور نامتقارن روی متن‌های انسانی متمرکز هستند: دقت برای متن‌های انسانی ۹۳٫۰ درصد است در حالی که برای سه مولد به‌ترتیب ۹۹٫۰ (DeepSeek)، (GPT-4o-mini) ۱۰۰٫۰ و ۹۹٫۰ (Gemini) درصد است. ۱۴ نمونه از ۱۸ خطا مربوط به متن‌های انسانی است که به‌اشتباه به‌عنوان مصنوعی طبقه‌بندی شده‌اند. این الگو، با وجود وزن‌دهی کلاس ۱:۳، نشان‌دهنده باقی‌ماندن سوگیری مدل به سمت کلاس اکثریت است.

دوم، در پنل میانی، دقت بر اساس حوزه موضوعی بررسی شده است. حوزه «تاریخ و زندگی‌نامه» با دقت ۱۰۰ درصد بالاترین عملکرد را دارد. این یافته از این جهت قابل توجه است که متن‌های زندگی‌نامه‌ای حاوی نام‌های خاص، تاریخ‌های دقیق و ارجاعات جغرافیایی هستند که بازتولید طبیعی آن‌ها برای مدل‌های زبانی فعلی دشوار است. در مقابل، حوزه‌های «علوم نظری» و «علوم انسانی» با ۹۶٫۲۵ درصد پایین‌ترین دقت را داشتند، که نشان‌دهنده دشواری تشخیص در متن‌های انتزاعی‌تر است.

سوم، در پنل سمت راست، دقت بر اساس طول متن نمایش داده شده است. متن‌های ۴۰ تا ۵۰ کلمه دقت ۹۶٫۳۵ درصد دارند، در حالی که متن‌های بلندتر (۵۰ کلمه به بالا) به دقت‌های ۹۸٫۱ تا ۹۸٫۷ درصد می‌رسند. این روند، اهمیت طول متن برای جمع‌آوری شواهد سبکی کافی را تأیید می‌کند.

۵-۳. آزمایش حذفی: تعمیم به مولد دیده‌نشده

نتایج آزمایش حذفی تک‌مولدی و دومولدی به‌صورت یکپارچه در شکل ۲ ارائه شده است. این شکل ماتریس تعمیم‌پذیری را نشان می‌دهد که در آن هر سطر یک شرایط آموزش و هر ستون یک مولد آزمون است.



شکل ۲: ماتریس تعمیم‌پذیری بین‌مولدی. سطرها شرایط آموزش (از پایه تا حذف تک‌مولدی و دومولدی) و ستون‌ها مولدهای آزمون را نشان می‌دهند. هر سلول دقت طبقه‌بندی روی متن‌های آن مولد را گزارش می‌کند. رنگ تیره‌تر معادل دقت بالاتر است.

مشاهدات اصلی از این ماتریس عبارت‌اند از:

(الف) آموزش پایه (روی هر سه مولد) به‌طور یکنواخت دقت بالای ۹۹ درصد روی هر سه مولد به‌دست می‌آورد و سقف عملکرد را تعریف می‌کند.

(ب) آموزش روی یک مولد به‌تنهایی منجر به افت قابل توجه روی مولدهای دیگر می‌شود. شدیدترین حالت زمانی رخ می‌دهد که طبقه‌بند تنها روی GPT-4o-mini آموزش ببیند و روی DeepSeek ارزیابی شود: دقت به ۸۲٫۵ درصد سقوط می‌کند — افتی ۱۷٫۵ درصدی نسبت به دقت روی خود GPT-4o-mini که ۹۹٫۵ درصد است. به‌طور مشابه، آموزش تنها روی Gemini نیز دقت روی DeepSeek را به ۹۱٫۰ درصد می‌رساند. این یافته اثبات تجربی این ادعاست که هر مولد، امضای سبکی متمایزی دارد که تنها با مشاهده آن مولد در آموزش قابل یادگیری است.

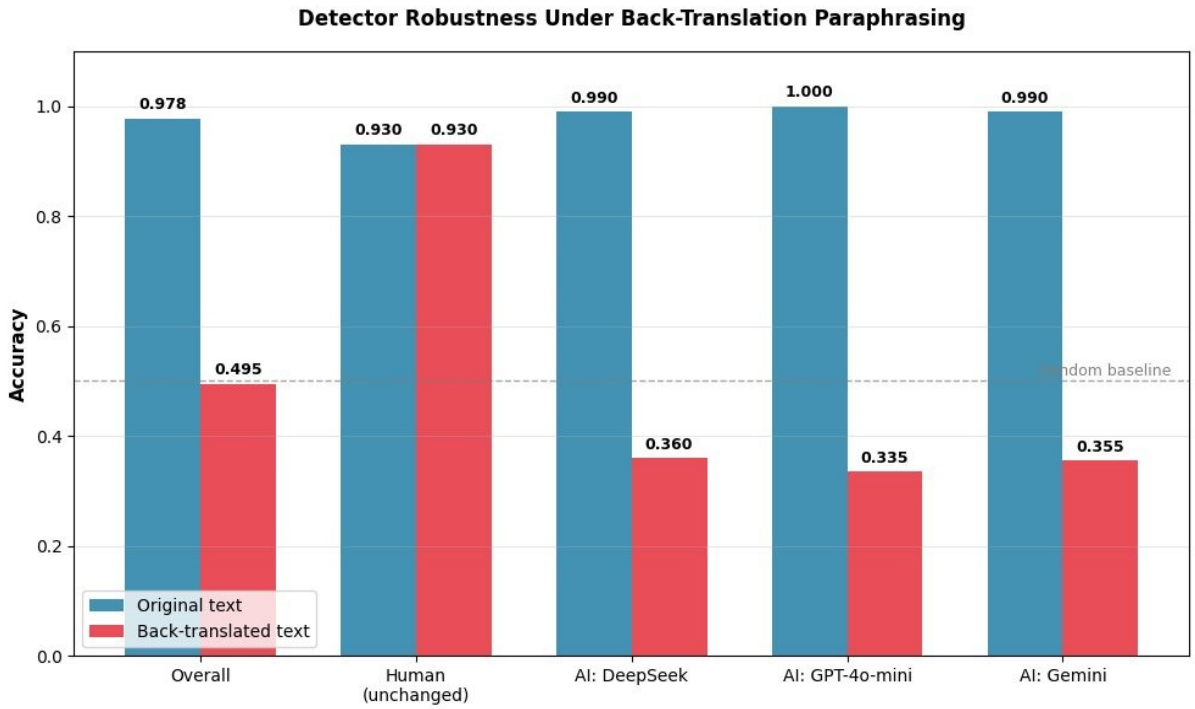
(ج) آموزش روی دو مولد (طرح leave-one-out، با حذف یک مولد از آموزش) عملکرد را به‌طور قابل توجهی بهبود می‌بخشد. به‌ویژه ترکیب GPT+Gemini روی DeepSeek دیده‌نشده به دقت ۹۴٫۵ درصد می‌رسد — افزایش ۱۲ درصدی نسبت به آموزش تنها با GPT-4o-mini. این نشان می‌دهد که تنوع مولدها در آموزش، حتی اگر مولد هدف در داده آموزش نباشد، اثر تعمیم‌بخش قوی دارد.

(د) مشاهده ظریف‌تر: آموزش تنها با DeepSeek به‌طور یکنواخت‌تری به سایر مولدها تعمیم می‌یابد (۹۷٫۵ تا ۹۸٫۰ درصد روی همه ستون‌ها) نسبت به آموزش تنها با GPT-4o-mini. این می‌تواند بدین معنا باشد که توزیع سبکی

DeepSeek پوشش وسیع‌تری از فضای سبکی مولدهای دیگر دارد، در حالی که GPT-4o-mini در ناحیه دورتری از این فضا قرار گرفته است.

۴-۵. مقاومت در برابر بازنویسی

نتایج آزمون مقاومت در شکل ۳ نشان داده شده است. این شکل دقت طبقه‌بند پایه را قبل و پس از اعمال ترجمه برگشتی روی متن‌های مصنوعی مقایسه می‌کند. متن‌های انسانی در این آزمایش دستکاری نشده‌اند.

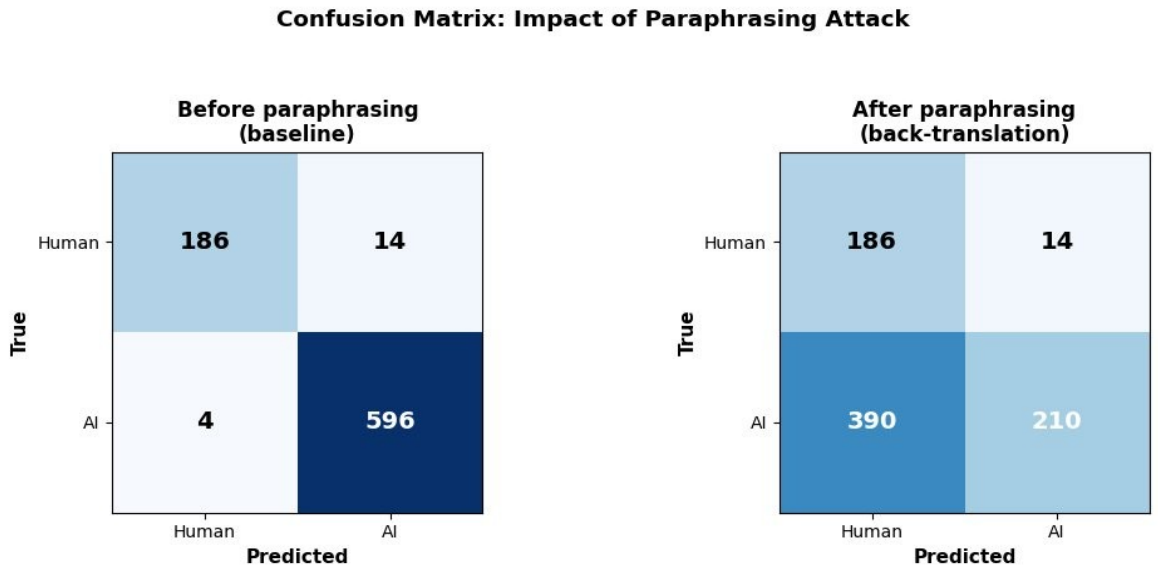


شکل ۳: مقاومت طبقه‌بند پایه در برابر بازنویسی با ترجمه برگشتی. ستون‌های آبی دقت روی متن اصلی و ستون‌های قرمز دقت پس از بازنویسی متن‌های مصنوعی را نشان می‌دهند. خط‌چین خاکستری حدس تصادفی (۵۰٪) را نشان می‌دهد.

دقت کلی از ۹۷٫۷۵ درصد به ۴۹٫۵۰ درصد سقوط می‌کند — افتی ۴۸٫۲۵ درصدی. این سقوط تقریباً به‌طور کامل به افت روی کلاس مصنوعی منسوب است: دقت روی متن‌های مصنوعی از حدود ۹۹ درصد به ۳۵ درصد می‌رسد، یعنی افتی ۶۴ درصدی. دقت روی متن‌های انسانی، که دستکاری نشده‌اند، در ۹۳ درصد باقی می‌ماند. این نتیجه به‌طور قاطع نشان می‌دهد که طبقه‌بند، سبک LLM را یاد گرفته است نه «مصنوعی بودن» را به‌طور انتزاعی؛ به محض آنکه امضای سبکی از طریق بازنویسی پاک می‌شود، مدل متن مصنوعی را به اشتباه به‌عنوان انسانی طبقه‌بندی می‌کند.

نکته بسیار مهم این است که دقت ۳۵ درصد روی کلاس مصنوعی، پایین‌تر از حدس تصادفی (۵۰٪) است. این به این معناست که مدل با اطمینان بالا اشتباه می‌کند: نه فقط شکست می‌خورد، بلکه پس از بازنویسی، به‌طور سامانمند برچسب مخالف می‌دهد.

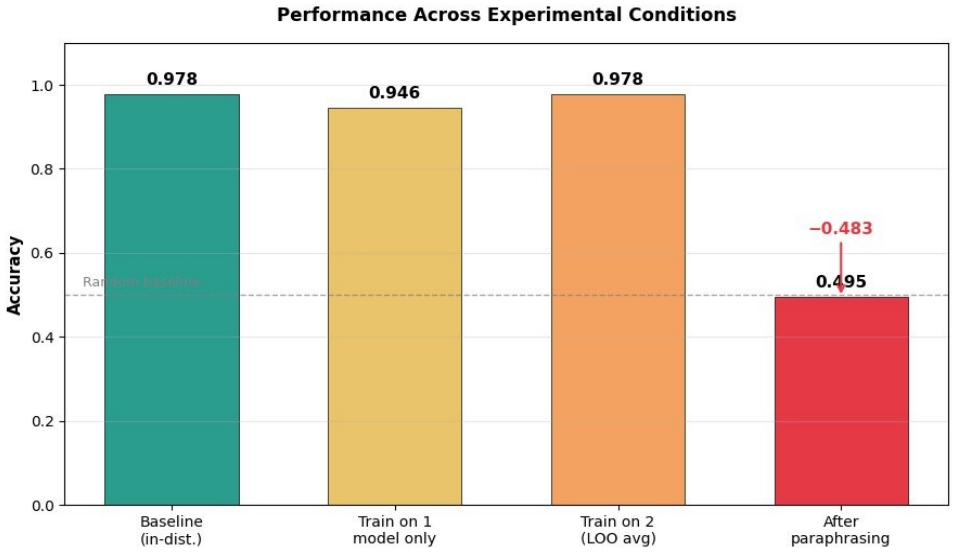
افت روی هر سه مولد تقریباً یکسان است: DeepSeek ۳۶٫۰ درصد، GPT-4o-mini ۳۳٫۵ درصد و Gemini ۳۵٫۵ درصد. این یکنواختی نشان می‌دهد که شکنندگی طبقه‌بند مستقل از مولد است و ویژگی ذاتی روش تشخیص محسوب می‌شود. ماتریس درهم‌ریختگی پیش و پس از حمله در شکل ۴ نمایش داده شده است.



شکل ۴: ماتریس درهم‌ریختگی پیش و پس از حمله بازنویسی. تعداد خطای دسته‌بندی متن‌های مصنوعی (پایین، سمت چپ ماتریس راست) از ۴ به ۳۹۰ افزایش یافته است.

مقایسه این نتیجه با کارهای انگلیسی‌زبان آموزنده است: Krishna و همکاران (2023) گزارش کردند که طبقه‌بند RoBERTa-Large-Detector انگلیسی پس از حمله DIPPER حدود ۲۰ درصد افت نشان می‌دهد. افت ۶۴ درصدی در پژوهش حاضر، حدود سه برابر بزرگ‌تر است. بخشی از این تفاوت می‌تواند به این دلیل باشد که مدل‌های ترجمه parsinlu تغییرات وسیع‌تری در متن نسبت به DIPPER ایجاد می‌کنند، اما حتی با در نظر گرفتن این عامل، اندازه افت نشان می‌دهد که طبقه‌بندهای فعلی فارسی به‌طور جدی شکننده‌اند.

۵-۵. جمع‌بندی نتایج در یک نمای کلی



شکل ۵: خلاصه عملکرد طبقه‌بند در چهار شرایط آزمایشی: پایه، حذفی تک‌مولدی (میانگین)، حذفی دومولدی (میانگین) و پس از بازنویسی.

شکل ۵ تصویر یکپارچه‌ای از یافته‌های پژوهش ارائه می‌دهد. در شرایط پایه و دومولدی، عملکرد در سطح بالایی است (نزدیک ۹۸٪). آموزش تک‌مولدی افت متوسط (حدود ۳ درصد در میانگین) دارد، اما در شرایط بدبینانه (آموزش GPT روی DeepSeek) تا ۱۷٫۵ درصد افت می‌بیند. بازنویسی، تنها یک حمله ساده، دقت را تقریباً به سطح حدس تصادفی می‌رساند.

۶. بحث

یافته‌های این پژوهش سه نکته مفهومی را برجسته می‌سازد. نخست، تفاوت میان دقت درون‌توزیعی و عملکرد در شرایط واقعی. در حالی که دقت ۹۸ درصد در داده آزمون اعتمادبه‌نفس می‌سازد، آزمایش‌های حذفی و حمله بازنویسی نشان می‌دهند که این دقت تا چه حد به فرض ثابت ماندن توزیع داده وابسته است. در کاربردهای واقعی، که در آن کاربر بدنیت می‌تواند مولد جدید یا تکنیک بازنویسی به کار ببرد، دقت می‌تواند به‌طور چشمگیری کاهش یابد. این یافته با نتایج مشابه در زبان‌های پرمنبع همخوانی دارد و دامنه اعتبار آن را به فارسی گسترش می‌دهد.

دوم، نقش تنوع مولدها در داده آموزش. تفاوت آشکار میان آموزش تک‌مولدی (متوسط ۹۴٫۶٪) و دومولدی (متوسط ۹۷٫۸٪) نشان می‌دهد که حتی افزودن یک مولد دوم به آموزش، تعمیم به مولد سوم را به‌طور قابل توجهی بهبود می‌بخشد. این نتیجه پیامد عملی روشنی دارد: توسعه‌دهندگان سامانه‌های تشخیص باید داده‌های متنوع از چند مولد گردآوری کنند، حتی اگر هدف تنها تشخیص یک مولد خاص باشد. این یافته با مشاهدات موازی در پیکره‌های انگلیسی همچون MAGE نیز سازگار است.

سوم، رفتار مدل پایه حول گرایش به کلاس مصنوعی. ۱۴ از ۱۸ خطا روی متن‌های انسانی رخ می‌دهد. این الگو پیامدهای عملی دارد: در سامانه‌ای که برای حفظ یکپارچگی علمی به کار رود، نرخ مثبت کاذب بالا روی متن انسانی

می‌تواند به محکومیت‌های نادرست منجر شود. وزن‌دهی ساده کلاس برای جبران کامل این سوگیری کافی نبوده است؛ شاید توابع زیان دیگر مانند زیان کانونی یا روش‌های آموزش جفتی برای حل این چالش بهتر باشند.

یافته جنبی درباره حوزه‌ها نیز قابل تأمل است. دقت ۱۰۰ درصدی در حوزه «تاریخ و زندگی‌نامه» می‌تواند نتیجه دو عامل باشد: غنا بودن متن‌های انسانی این حوزه به نام‌های خاص، تاریخ‌ها و ارجاعات جغرافیایی دقیق، که بازتولید طبیعی آن‌ها برای مدل‌های زبانی فعلی دشوار است؛ و در نتیجه، تمایل مدل‌های مصنوعی به استفاده از الگوهای ساختاری قابل پیش‌بینی در غیاب دانش وقایع‌نگاری دقیق. این یافته یک سرنخ سطحی است که مدل تشخیص می‌تواند به آن تکیه کند، و انتظار می‌رود که با ارتقای مدل‌های نسل بعدی که در نام‌گذاری دقیق بهتر عمل می‌کنند، این مزیت تضعیف شود.

ارزش ویژه پژوهش حاضر در مقایسه با کار Mohammadi (2023) در دو محور است. نخست، در حالی که آن کار به یک مولد (GPT-3.5) محدود بود، پژوهش حاضر سه مولد نسل جدید را پوشش می‌دهد و تحلیل تعمیم بین آن‌ها را امکان‌پذیر می‌سازد. دوم، آن کار مقاومت در برابر دستکاری متن را ارزیابی نکرد، در حالی که این یکی از یافته‌های محوری پژوهش حاضر است.

۷. محدودیت‌ها

این پژوهش چندین محدودیت دارد که باید در تفسیر نتایج لحاظ شوند. نخست، حوزه مورد بررسی به نثر دانشنامه‌ای فارسی محدود است؛ تعمیم به متن خبری، شبکه‌های اجتماعی، متن ادبی یا گفتگو نیازمند ارزیابی مستقل است. دوم، تنها سه مولد در دسترس از طریق یک دروازه واحد بررسی شده‌اند. مدل‌های مهمی همچون Claude، خانواده Llama، و مدل‌های بومی فارسی (مانند PersianMind) در این پژوهش پوشش داده نشده‌اند. سوم، متن‌های انسانی همگی از یک منبع (ویکی‌پدیای فارسی) و یک بازه زمانی هستند، که می‌تواند سوگیری سبکی به سمت نگارش رسمی دانشنامه‌ای ایجاد کند.

چهارم، حمله بازنویسی به یک تکنیک واحد (ترجمه برگشتی با parsinlu) محدود شده است. ارزیابی با بازنویس‌های پیشرفته‌تر، یا حملات سطح کلمه و حملات همگانی، نتایج متفاوتی می‌تواند داشته باشد. پنجم، آزمایش‌ها با یک بذر تصادفی واحد انجام شده است؛ هرچند تفاوت‌های گزارش‌شده در مقیاسی هستند که از نوسان بذر فراتر می‌روند، ولی ارزیابی چندبذری به استحکام آماری بیشتر کمک می‌کند.

۸. نتیجه‌گیری و کارهای آتی

این پژوهش با ساخت یک پیکره موازی ۴۰۰۰تایی فارسی و طراحی آزمایش‌های هدفمند، سه یافته اصلی را عرضه کرد: (۱) تشخیص متن مصنوعی فارسی در شرایط درون‌توزیعی با دقت بالا (نزدیک ۹۸٪) قابل دستیابی است؛ (۲) هر مولد امضای سبکی متمایزی دارد و آموزش تک‌مولدی به تعمیم ضعیف منجر می‌شود، با افت تا ۱۷٫۵ درصدی روی مولد دیده‌نشده؛ (۳) طبقه‌بندهای پایه در برابر بازنویسی ساده با ترجمه برگشتی به‌شدت شکننده‌اند، با افت ۶۴ درصدی دقت بر روی متن‌های مصنوعی دستکاری‌شده.

این یافته‌ها چند مسیر برای پژوهش‌های آینده می‌گشاید. نخست، گسترش پیکره به مولدهای دیگر، به‌ویژه مدل‌های بومی فارسی، تصویر کامل‌تری از فضای سبکی مدل‌های فارسی فعلی فراهم می‌آورد. دوم، ارزیابی روش‌های تشخیص متخاصم‌محور — مانند آموزش متخاصم یا روش‌های مبتنی بر بازنویسی در زمان آموزش — برای بهبود مقاومت می‌تواند

مفید باشد. سوم، گسترش به حوزه‌های دیگر مانند متن‌های ادبی، خبری و شبکه‌های اجتماعی، که سبک نوشتاری متنوع‌تری دارند، اعتبار خارجی یافته‌ها را افزایش می‌دهد. چهارم، ادغام رویکردهای مبتنی بر «اثر انگشت آماری» با طبقه‌بندهای نظارت‌شده می‌تواند به مدلی هم‌افزا منجر شود که در برابر دستکاری مقاوم‌تر باشد.

منابع

منابع فارسی

اخگری، محمدرضا و ممتازی، سعیده (۱۴۰۲). کاربرد هوش مصنوعی در راستی‌آزمایی اخبار: تشخیص اخبار جعلی با استفاده از متن خبر و اطلاعات منابع منتشرکننده خبر. پژوهشنامه رسانه و ارتباطات نوین، دوره ۱، شماره ۱، صفحات ۲۴۳-۲۶۸.

متقی، وحید؛ اسماعیلی، مهدی؛ بازایی، قاسمعلی و افشارکاظمی، محمدعلی (۱۴۰۰). ارائه رویکرد ترکیبی مبتنی بر یادگیری عمیق و یادگیری ماشین جهت تشخیص اخبار جعلی - مطالعه موردی اخبار فارسی در حوزه کرونا ویروس. علوم و فنون مدیریت اطلاعات، دانشگاه قم.

منابع انگلیسی

Devlin, J., Chang, M. W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics, pp. 4171–4186.

Dugan, L., Hwang, A., Trhlík, F., Zhu, A., Ludan, J. M., Xu, H., Ippolito, D., and Callison-Burch, C. (2024). RAID: A Shared Benchmark for Robust Evaluation of Machine-Generated Text Detectors. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 12463–12492.

Ezzini, S., Ahmad, I., Chafik, S., Abudalfa, S., El-Haj, M., Abdelali, A., Jarrar, M., Durrani, N., Sajjad, H., and Adeeba, F. (2026). AbjadGenEval: Abjad AI Generated Text Detection Shared Task for Languages Using Arabic Script at AbjadNLP 2026. In Proceedings of the 2nd Workshop on NLP for Languages Using Arabic Script, EACL 2026, Rabat, Morocco.

Farahani, M., Gharachorloo, M., Farahani, M., and Manthouri, M. (2021). ParsBERT: Transformer-based Model for Persian Language Understanding. Neural Processing Letters, 53(6), 3831–3847.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., and Crawford, K. (2021). Datasheets for Datasets. Communications of the ACM, 64(12), 86–92.

Krishna, K., Song, Y., Karpinska, M., Wieting, J., and Iyyer, M. (2023). Paraphrasing Evades Detectors of AI-generated Text, but Retrieval is an Effective Defense. In Advances in Neural Information Processing Systems (NeurIPS 2023).

Li, Y., Li, Q., Cui, L., Bi, W., Wang, Z., Wang, L., Yang, L., Shi, S., and Zhang, Y. (2024). MAGE: Machine-generated Text Detection in the Wild. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, pp. 36–53.

Macko, D., Moro, R., Uchendu, A., Lucas, J., Yamashita, M., Pikuliak, M., Srba, I., Le, T., Lee, D., Simko, J., and Bielikova, M. (2023). MULTITuDE: Large-scale Multilingual Machine-Generated Text Detection Benchmark. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pp. 9960–9987.

Mohammadi, K. (2023). Human vs Machine Generated Text Detection in Persian.

Sadasivan, V. S., Kumar, A., Balasubramanian, S., Wang, W., and Feizi, S. (2023). Can AI-Generated Text be Reliably Detected? arXiv preprint arXiv:2303.11156.

Sarzaeim, P., Doshi, A., and Mahmoud, Q. (2023). A Framework for Detecting AI-Generated Text in Research Publications. In Proceedings of the International Conference on Advanced Technologies, vol. 11, pp. 121–127.

Wang, Y., Mansurov, J., Ivanov, P., Su, J., Shelmanov, A., Tsvigun, A., Whitehouse, C., Mohammed Afzal, O., Mahmoud, T., Sasaki, T., Arnold, T., Aji, A. F., Habash, N., Gurevych, I., and Nakov, P. (2024). M4: Multi-generator, Multi-domain, and Multi-lingual Black-Box Machine-Generated Text Detection. In Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2024), pp. 1369–1407.

Wu, J., Zhan, R., Wong, D. F., Yang, S., Liu, X., Chao, L. S., and Zhang, M. (2025). A Survey on LLM-Generated Text Detection: Necessity, Methods, and Future Directions. Computational Linguistics, 51(1), 275–338.