اثربخشی حملات فیشینگ انسان ساخت در برابر حملات فیشینگ تولیدشده با هوش مصنوعی

Effectiveness of Human-Crafted vs AI-Generated Phishing Attacks

Kalira Kawal (Mohammed Jowkari) محمد جوکاری
British Columbia Institute of Technology (BCIT), Vancouver, BC, Canada

چکیده: فیشینگ یکی از مهم ترین حملات در امنیت سایبری است و می تواند به زیان مالی، نقض داده و سرقت هویت منجر شود. با ظهور مدل های زبانی بزرگ، مهاجمان می توانند ایمیل های فیشینگ را سریع تر و در مقیاس بسیار بزرگ تر تولید کنند. با این حال، هنوز روشن نیست که برای یک کاربر عادی، ایمیل های فیشینگ انسان ساخت قانع کننده تر هستند یا ایمیل های تولیدشده با هوش مصنوعی. در این پژوهش، یک آزمایش کنترل شده با ۱۲ ایمیل فیشینگ شامل ۶ ایمیل انسان ساخت و ۶ ایمیل تولیدشده با هوش مصنوعی و ۵۰ شرکت کننده انجام شد. هدف، مقایسه باورپذیری این دو نوع ایمیل برای عموم مردم بود. نتایج نشان داد که ایمیل های انسان ساخت در این مطالعه از نظر آماری قانع کننده تر از ایمیل های تولیدشده با هوش مصنوعی بودند. این یافته می تواند برای بهبود آگاهی امنیتی، طراحی آموزش های ضد فیشینگ و درک بهتر تفاوت میان مقیاس پذیری هوش مصنوعی و ظرافت اقناعی انسان مفید باشد.

کلیدواژه ها: فیشینگ، امنیت سایبری، مهندسی اجتماعی، هوش مصنوعی، ایمیل، رفتار کاربر

۱ مقدمه

فیشینگ یکی از رایج ترین حملات سایبری است که برای افراد و سازمان ها زیان مالی و نقض داده به همراه دارد. در گذشته، این حملات عمدتا توسط انسان نوشته می شدند و ساخت آنها به مهارت، زمان و تلاش نیاز داشت. همین موضوع تا حدی مقیاس حمله را محدود می کرد. امروز با مدل های زبانی بزرگ مانند ChatGPT، مهاجمان می توانند ایمیل های فیشینگ را در حجم بالا و با سرعت زیاد تولید و ارسال کنند. این تغییر، یک ریسک جدید برای سازمان ها و کاربران عادی ایجاد کرده است.

در عین حال، فیشینگ انسان ساخت هنوز می تواند ظرافت، شخصی سازی و درک روان شناسی بیشتری داشته باشد. به همین دلیل این پرسش مطرح می شود که کدام نوع پیام برای یک کاربر عادی قانع کننده تر است: ایمیل انسان نوشته یا ایمیل تولیدشده با هوش مصنوعی. بیشتر پژوهش های موجود یا روی عبور از فیلترهای اسپم تمرکز دارند یا روی تشخیص فنی پیام، اما مقایسه مستقیم این دو نوع پیام در یک شرایط یکسان کمتر بررسی شده است. این پژوهش برای پر کردن همین شکاف انجام شد.

۲ مرور ادبیات

مطالعات پیشین نشان داده اند که پیام های تولیدشده با هوش مصنوعی می توانند از نظر زبانی بسیار روان و قانع کننده باشند. برخی پژوهش ها گزارش کرده اند که این نوع پیام ها حتی از بعضی فیلترهای هرزنامه نیز عبور می کنند. در مقابل، پژوهش های دیگر نشان می دهند که پیام های انسان ساخت هنوز در استفاده از جزئیات اجتماعی، لحن طبیعی و نشانه های ظریف اعتماد برتری دارند.

در یک خط پژوهشی، مقایسه میان ایمیل های کاملا انسان نوشته، پیام های کمک گرفته از اینترنت و پیام های تولیدشده با LLM انجام شده و نتایج نشان داده اند که هوش مصنوعی می تواند از نظر اقناع عملکرد بالایی داشته باشد. در خط دیگری از پژوهش، تمرکز روی عبور از فیلترهای ارائه دهندگان ایمیل و محدودیت های سبک سنجی بوده است. مجموعه دیگری از مطالعات نیز روی حملات ترکیبی انسان و هوش مصنوعی کار کرده اند و نشان داده اند که ترکیب مقیاس پذیری ماشین با قضاوت انسانی می تواند نرخ کلیک را بیشتر کند.

با وجود این تنوع، هنوز این پرسش به طور روشن باقی مانده است که در یک مقایسه رو در رو و با یک طراحی ساده، یک کاربر معمولی کدام نوع ایمیل را باورپذیرتر می بیند. بنابراین، این پژوهش تمرکز خود را روی سنجش مستقیم باورپذیری گذاشت و از پیچیده کردن سناریو خودداری کرد تا پاسخ به پرسش اصلی روشن تر باشد.

۳ روش شناسی

این پژوهش از روش کمی و آزمایشی استفاده کرد. کمی بود، زیرا همه شرکت کنندگان به هر ایمیل امتیاز عددی دادند و این امتیازها با هم مقایسه شدند. آزمایشی بود، زیرا محیط مشاهده و محتوای ارائه شده به شرکت کنندگان کنترل شد و تنها عامل اصلی که تغییر می کرد، منبع تولید ایمیل بود: انسان یا هوش مصنوعی.

در این مطالعه، ۱۲ ایمیل فیشینگ طراحی شد. شش ایمیل توسط اعضای تیم نوشته شد و شش ایمیل دیگر با کمک هوش مصنوعی تولید شد. سپس این ایمیل ها از طریق یک فرم برای ۵۰ شرکت کننده نمایش داده شدند. از شرکت کنندگان خواسته شد به هر ایمیل طوری نگاه کنند که گویی در صندوق ورودی واقعی خودشان قرار دارد، نه به عنوان یک سوال درسی. این کار برای نزدیک تر کردن واکنش ها به دنیای واقعی انجام شد.

نمونه گیری با استفاده از کد QR در رستوران یکی از اعضای تیم انجام شد. این روش باعث شد شرکت کنندگان فقط محدود به دانشجویان امنیت سایبری نباشند و افراد عادی، خانواده ها، کارکنان و افراد مسن نیز در مطالعه حضور پیدا کنند. البته این روش بی نقص نبود و احتمال سوگیری خودانتخابی وجود داشت، زیرا افرادی که با فناوری راحت تر بودند ممکن بود بیشتر در نظرسنجی شرکت کنند.

۴ مدل اندازه گیری و تحلیل داده

پاسخ ها از Google Forms دریافت و به فایل CSV تبدیل شدند. سپس با استفاده از Python و Jupyter Notebook، داده ها تمیزسازی و تحلیل شدند. برای هر ایمیل، میانگین، انحراف معیار و تعداد پاسخ ها محاسبه شد. شش سوال اول در گروه هوش مصنوعی و شش سوال آخر در گروه انسانی قرار گرفتند.

برای مقایسه کلی، میانگین شش ایمیل هوش مصنوعی و میانگین شش ایمیل انسانی محاسبه شد. همچنین میانگین هر شرکت کننده در دو گروه جداگانه بررسی شد تا مشخص شود آیا بیشتر افراد در سطح فردی نیز ایمیل های انسانی را باورپذیرتر می بینند یا نه. برای بررسی قابلیت اعتماد داده ها، آلفای کرونباخ نیز محاسبه شد. این شاخص کمک کرد مشخص شود پاسخ دهی شرکت کنندگان در کل مجموعه، همگن و قابل اتکا بوده است.

۵ پرسش پژوهش و فرضیه

پرسش اصلی پژوهش این بود که آیا ایمیل های فیشینگ انسان ساخت از ایمیل های فیشینگ تولیدشده با هوش مصنوعی موثرتر هستند یا نه. در این مطالعه، موثرتر بودن به معنی باورپذیرتر بودن و احتمال بیشتر کلیک کردن توسط یک کاربر عادی تعریف شد.

فرضیه ما این بود که با وجود مقیاس پذیری بالای هوش مصنوعی، ایمیل های انسان ساخت در صورتی که با دقت و توجه کافی نوشته شوند، از نظر اقناع فردی قوی تر خواهند بود. به بیان ساده، انتظار می رفت میانگین شش ایمیل انسانی از میانگین شش ایمیل هوش مصنوعی بیشتر باشد و این برتری نه فقط در سطح میانگین گروهی، بلکه در سطح بیشتر شرکت کنندگان نیز دیده شود.

۶ شاخص های ارزیابی

باورپذیری هر ایمیل با امتیازهای ۱ تا ۱۰ سنجیده شد. در گزارش، بازه ۱ تا ۳ به عنوان کم باورپذیر، ۴ تا ۷ به عنوان باورپذیر و ۷ تا ۱۰ به عنوان بسیار باورپذیر در نظر گرفته شد. برای هر ایمیل، میانگین، انحراف معیار و تعداد پاسخ ها گزارش شد. در سطح کلی نیز اختلاف میانگین دو گروه، آزمون تفاوت، اندازه اثر و فاصله اطمینان ۹۵ درصد بررسی شد. برای اطمینان از پایداری مقیاس، آلفای کرونباخ برای کل ۱۲ آیتم و نیز برای هر دو زیرمجموعه محاسبه شد.

۷ نتایج

تحلیل داده ها نشان داد که ایمیل های انسان ساخت در این مطالعه باورپذیرتر از ایمیل های تولیدشده با هوش مصنوعی بودند. در نسخه اصلی گزارش، میانگین گروه انسانی ۴۷.۲ و میانگین گروه هوش مصنوعی ۶۳.۱ گزارش شد. اختلاف این دو مقدار ۸۴.۰۰ بود. مقدار p برابر با 0.000000000032304 گزارش شد که بسیار کمتر از آستانه رایج 0.05 است و بنابراین تفاوت مشاهده شده از نظر آماری معنادار بود.

فاصله اطمینان ۹۵ درصد نیز نشان داد که برتری ایمیل های انسانی در این مطالعه تقریبا بین ۶۴.۰ تا ۰۳.۱ امتیاز قرار دارد. نمودارهای تحلیل همچنین نشان دادند که در سطح بیشتر شرکت کنندگان، میانگین ایمیل های انسانی از میانگین ایمیل های هوش مصنوعی بیشتر بوده است. در بعضی از مثال های فردی، علت این برتری به طبیعی تر بودن لحن، نبود لینک مستقیم، یا استفاده بهتر از روان شناسی اعتماد نسبت داده شد.

از طرف دیگر، بعضی از پیام های تولیدشده با هوش مصنوعی به دلیل کوتاهی متن، لحن بیش از حد عمومی یا استفاده نامناسب از نشانه هایی مانند ایموجی، کمتر متقاعدکننده دیده شدند. این موضوع نشان می دهد که تولید سریع پیام به تنهایی برای اقناع کافی نیست.

۸ بحث

نتایج این مطالعه از فرضیه اصلی پشتیبانی می کنند. هرچند هوش مصنوعی در مقیاس پذیری، سرعت تولید و حتی عبور از بعضی فیلترها قدرت بالایی دارد، اما پیام های انسان ساخت هنوز می توانند در جلب اعتماد یک فرد عادی موفق تر باشند. این برتری احتمالا از توجه بیشتر انسان به نشانه های اجتماعی، زمینه واقعی و لحن طبیعی ناشی می شود.

یکی از محدودیت های پژوهش، شیوه نمونه گیری بود. چون شرکت کنندگان از یک محیط عمومی و از طریق کد QR جذب شدند، احتمال سوگیری خودانتخابی وجود داشت. همچنین ترتیب نمایش ایمیل ها می توانست روی برداشت افراد اثر بگذارد، زیرا شش ایمیل اول مربوط به هوش مصنوعی و شش ایمیل آخر مربوط به انسان بودند. در کارهای آینده، تصادفی سازی ترتیب نمایش می تواند دقت مطالعه را بیشتر کند.

با وجود این محدودیت ها، یافته های پژوهش از نظر کاربردی اهمیت دارند. برای سازمان ها، این نتایج نشان می دهد که دفاع در برابر فیشینگ نباید فقط بر تشخیص خودکار تکیه کند. آموزش کاربر، طراحی برنامه های آگاهی امنیتی و تحلیل عوامل انسانی همچنان نقش مهمی دارند.

۹ نتیجه گیری و کار آینده

این پژوهش یک مقایسه مستقیم و کنترل شده میان فیشینگ انسان ساخت و فیشینگ تولیدشده با هوش مصنوعی ارائه کرد. نتایج نشان دادند که در این مطالعه، ایمیل های انسان ساخت از نظر باورپذیری برای شرکت کنندگان قوی تر از ایمیل های هوش مصنوعی بودند. بنابراین، اگرچه هوش مصنوعی می تواند حجم حمله را افزایش دهد، اما کیفیت اقناعی انسان همچنان یک مزیت جدی محسوب می شود.

در آینده می توان این پژوهش را با گروه های کاربری متفاوت، فیلترهای واقعی ایمیل، ارائه دهندگان واقعی سرویس و همچنین سناریوهای ترکیبی انسان و هوش مصنوعی توسعه داد. همچنین مقایسه کاربران فنی و غیر فنی و ارزیابی اثر آموزش امنیتی نیز می تواند به گسترش این خط پژوهش کمک کند.

یادداشت نمایه سازی

برای دیده شدن بهتر این کار در جستجوهای فارسی و انگلیسی، بهتر است نسخه فارسی در کنار نسخه انگلیسی روی یک صفحه عمومی قرار گیرد و نام نویسنده به هر دو صورت Mohammed Jowkari و محمد جوکاری در آن صفحه دیده شود.

مراجع

[۱] internet-aided، human-written، “Comparing Ekeihal، A. out- Persuasion emails: phishing LLM-generated and Skövde، of University survey،” controlled a in comes ۲۰۲۴.

[۲] phishing “AI-generated Lee، K. and Smith، R. Hazell، J. ۲۰۲۴ study،” empirical An evasion: spam-filter and

[۳] hybrid “Human+AI Zhou، Y. and Patel، S. Heiding، M. ۲۰۲۳ factors،” risk and rates Click-through phishing:

[۴] SMS Phishing Spear Human-Authored vs AI “Assessing ۲۰۲۵ Attacks،”

[۵] LLM for prompting Structured Detector: Spam “Chat ۲۰۲۴ classfiication،” email

[۶] baselines،” learning machine and corpus “AI-phish ۲۰۲۴

[۷] phish- AI-generated vs filters provider and “Stylometry ۲۰۲۵ ing،”

[۸] NDSS LLMs،” with spear-phishing “Instagram-based ۲۰۲۵ Poster،

[۹] phishing for agents model language large vs. “Detectors ۲۰۲۴ rewrites،”

[۱۰] phishing،” human expert vs GPT-۴ “Head-to-head: ۲۰۲۳

[۱۱] ۲۰۲۵ Co-created،” vs GPT-۴ vs Human “Training:

[۱۲] ۲۰۲۴ Emails،” Marketing vs Detectors “LLM

[۱۳] ۲۰۲۵ GPT-۴o،” vs Stylometry + “Providers

[۱۴] ۲۰۲۴ ACM، Efficacy،” Email “Phishing

[۱۵] Behav- Comparing Attacks: Phishing “Eye-Tracking Emails،” Human-Crafted and AI- to Responses ioral ۲۰۲۴/۲۰۲۵