چالش‌ها و تهدیدهای امنیت داده در LLMs

دکتر عذرا دائی
استادیار گروه کتابداری و اطلاع رسانی پزشکی
دانشکده مدیریت و اطلاع رسانی پزشکی تبریز
مرداد ۱۴۰۴

مقدمه

- **LLMها به دلایل ساختاری و داده‌ای در معرض حملات و افشای اطلاعات هستند.**

Adversarial Attacks

1. Prompt Injection
2. Evasion Attacks
3. Poisoning Attacks
4. Model Inversion Attacks
5. Model Stealing Attacks
6. Membership Inference Attacks

تزریق دستور Prompt Injection

- حملات تزریق دستور شامل طراحی ورودی‌های خاصی هستند که هدفشان **دستکاری رفتار LLMها با بهره‌گیری از وابستگی آن‌ها به دستورات کاربران** است. این ورودی‌ها به گونه‌ای ساخته می‌شوند که به‌طور ظریف پاسخ‌های مدل را تحت تأثیر قرار دهند و ممکن است موجب تولید خروجی‌های مضر یا غیرمنتظره شوند.

حملات فرار Evasion Attacks

- حملات فرار شامل **تغییرات ظریف در ورودی‌ها (تصاویر، فایل‌های صوتی، اسناد و غیره) برای گمراه کردن مدل‌ها در زمان استنتاج** هستند و به همین دلیل، روشی مخفی و مؤثر برای دور زدن سیستم‌های کنترلی محسوب می‌شوند.
- این حملات به گونه‌ای طراحی می‌شوند که **ورودی‌ها را به‌طور جزئی تغییر دهند تا مدل‌های هوش مصنوعی را در زمان استنتاج گمراه کنند** و باعث دسته‌بندی نادرست ورودی‌های خاص شوند. برای مثال:

حملات مسموم‌سازی Poisoning Attacks

- حملات مسموم‌سازی **فاز آموزش مدل‌های هوش مصنوعی را هدف قرار می‌دهند و با وارد کردن داده‌های مخرب به مجموعه داده‌های آموزشی، رفتار مدل را به خطر می‌اندازند.**
- مسموم‌سازی داده‌ها نوعی حمله علیه مدل‌های هوش مصنوعی است که در آن **حمله‌کنندگان به‌طور عمدی داده‌های خراب یا گمراه‌کننده را به مجموعه داده‌های آموزشی مدل وارد می‌کنند.**

حملات وارون‌سازی مدل Model Inversion Attacks

- حملات وارون‌سازی مدل با هدف مهندسی معکوس مدل‌های هوش مصنوعی برای بازیابی اطلاعات حساس مربوط به داده‌های آموزشی انجام می‌شوند. در این حملات، عاملان مخرب پیش‌بینی‌های مدل را در پاسخ به ورودی‌های مختلف تحلیل می‌کنند.

حملات استخراج مدل
Model Extraction Attacks

- حمله‌گر با ارسال **درخواست‌های متعدد** به مدل، ساختار آن را کپی‌برداری می‌کند.
- حملات استخراج مدل با هدف تقلید عملکرد یک مدل اختصاصی انجام می‌شوند، از طریق **ارسال تعداد زیادی ورودی و مشاهده خروجی‌های مدل**.

حملات استنتاج عضویت
Membership Inference Attacks

- **حملات استنتاج عضویت** تلاش می‌کنند تا مشخص کنند آیا **داده‌های خاصی در آموزش مدل استفاده شده است یا نه.**
- یک رکورد داده خاص، در مجموعه آموزش یک مدل هوش مصنوعی استفاده شده است یا خیر.
- این حملات شامل تلاش‌ها برای استنتاج اطلاعات حساس از یک مدل هوش مصنوعی از طریق بررسی **خروجی‌ها و رفتارهای آن** است.

Disclosure Concerns in Large Language Models (LLMs)

تزریق دستور
Prompt Injection

- این حمله با دستکاری **ورودی‌های کاربر (Prompts)**باعث می‌شود LLMها پاسخ‌های نادرست یا مخربی تولید کنند.
- وارد کردن اطلاعات محرمانه از منابع بیرونی (فایل‌ها، صفحات وب، تقویم‌ها)
- اگر مدل آن را بخواند، ممکن است محتوای حساس را افشا کند.

نشت اطلاعات در حافظه گفتگو
Multi-turn Memory Leakage

- برخی مدل‌ها تاریخچه گفتگو را حفظ می‌کنند.
- اطلاعات مربوط به مکالمات قبلی به طور ناخواسته در حافظه گفتگو باقی می‌مانند و در مکالمات بعدی قابل دسترسی هستند، حتی اگر این اطلاعات دیگر مرتبط نباشند. این موضوع می‌تواند منجر به افشای اطلاعات حساس، کاهش حریم خصوصی و ایجاد مشکلات امنیتی شود.

به‌یادسپاری بیش از حد
Output Memorization

- اگر مدل بیش‌ازحد یاد بگیرد overfitting، ممکن است عبارات حساس را حفظ کرده و تکرار کند.
- وقتی یک مدل به جای **یادگیری الگوها و تعمیم آن‌ها**، به شکل مستقیم و مکانیکی **خروجی‌های آموزش training data outputs** را حفظ می‌کند و در هنگام تولید متن یا پاسخ، همان داده‌ها را تقریباً **کلمه به کلمه** تکرار می‌کند، به این حالت **Output Memorization**می‌گویند.

آموزش روی داده‌های خام
Uncleaned Fine-tuning

- استفاده از داده‌های خام پزشکی بدون پاک‌سازی، در فرآیند fine-tune
- به حالتی گفته می‌شود که یک مدل **بدون پالایش یا پاک‌سازی داده‌های آموزشی**، مستقیماً روی همان داده‌های خام موجود **ری‌فاین‌تیون Fine-tune** می‌شود.

نشت در زنجیره پرسش‌ها (روش پرامپ‌نویسی) Prompt Chaining Leakage

- حمله گر با طرح سوالات خاص، موفق می‌شود بخش‌هایی از پرامپت سیستم یا دستورالعمل‌های مخفی مدل را درخواست کرده و به دست بیاورد.

بازپخش ناخواسته
Shadow Prompt Replay

- مهاجم ابتدا یک Promptدستور ورودی می‌سازد که ممکن است فیلترهای امنیتی را دور بزند.
- سپس همان Promptیا نسخه‌های بازنویسی‌شده آن Shadow Promptsرا بارها به مدل می‌دهد.

نگه‌داری داده بدون رضایت
Data Retention without Consent

- داده‌های کاربران (Promptها، مکالمات، خروجی‌ها) توسط سیستم ذخیره یا حتی برای آموزش/بهبود مدل استفاده بشوند، **بدون اینکه کاربر آگاه باشد یا رضایت داده باشد.**

حملات بازشناسی مجدد
Re-identification Attacks

- داده‌های پزشکی حتی پس از حذف نام و شماره ملی می‌توانند قابل شناسایی باقی بمانند.
- LLMهایی که با داده‌های نیمه‌ناشناس آموزش دیده‌اند، ممکن است الگوهای خاص بیماران را حفظ کرده باشند.

نشت داده‌ها Prompt Leakage

- **Prompt Leakage**
- گاهی اطلاعات وارد شده توسط کاربران قبلی در حافظه موقتی مدل باقی می‌ماند این اطلاعات ممکن است در پاسخ به کاربر جدید ظاهر شود.
- **Training Data Leakageنشت داده از طریق آموزش**

نمونه

- **نشت اطلاعات در ChatGPT**
- در سال ۲۰۲۳، برخی کاربران گزارش کردند که مدل اطلاعات گفتگوهای دیگر کاربران را نشان می‌دهد
- این مشکل ناشی از cachingاشتباه در prompt history بود
- در حوزه سلامت، چنین نشت‌هایی می‌تواند به فاجعه تبدیل شود

Inference / Extraction Attacks حملات

- Inference or Extraction Attacksحملاتی که کاربر مهاجم تلاش می‌کند با ارسال promptهای خاص، داده‌های آموزش دیده را استخراج کند. معمولاً شامل سوالات باز و هدف‌دار است حتی اگر داده‌ها ناشناس شده باشند، اطلاعات ساختاریافته قابل بازیابی‌اند

افشای ناخواسته اطلاعات حساس
Unintended Disclosure

- به وضعیتی گفته می‌شود که **اطلاعات حساس یا محرمانه** به‌طور غیرعمدی در خروجی یک سیستم یا مدل هوش مصنوعی آشکار می‌شود.

HIPAA Risks for Large Language Models (LLMs)

1. Data Privacy and Confidentiality
حریم خصوصی و محرمانگی داده‌ها

- یکی از نگرانی‌های اصلی در مورد LLMها، خطر نقض حریم خصوصی داده‌ها است. این مدل‌ها بر روی مجموعه‌داده‌های عظیمی آموزش می‌بینند که می‌توانند شامل اطلاعات حساس سلامت باشند. در صورتی که این داده‌ها به‌درستی ناشناس‌سازی نشوند، احتمال افشای ناخواسته آن‌ها وجود دارد. حتی اگر داده‌ها ناشناس‌سازی شوند، تکنیک‌های بازشناسی مجدد Re-identificationدر برخی موارد می‌توانند فرآیند ناشناس‌سازی را معکوس کرده و منجر به نقض حریم خصوصی بیماران شوند.
- **راهکارهای کاهش ریسک**
- **Data Anonymization** اطمینان حاصل شود که تمامی اطلاعات سلامت فردی PHIبه‌طور کامل ناشناس‌سازی شده باشند، پیش از آن‌که برای آموزش یا تعامل با LLMها مورد استفاده قرار گیرند.
- **Data Minimization** استفاده از حداقل داده‌های لازم برای دستیابی به نتیجه مطلوب.
- **Access Controls** پیاده‌سازی کنترل‌های دسترسی سختگیرانه به‌منظور اطمینان از اینکه تنها افراد مجاز بتوانند به داده‌ها دسترسی داشته باشند.

7. Patient Consent
رضایت بیمار

- دریافت رضایت بیمار یک جزء حیاتی از رعایت HIPAAاست. استفاده از LLMها در حوزه بهداشت و درمان می‌تواند فرآیند دریافت و مدیریت رضایت بیمار را پیچیده کند.
- **راهکارهای کاهش ریسک**
- **ارتباط شفاف:** به‌طور شفاف به بیماران توضیح داده شودکه داده‌هایشان چگونه استفاده خواهد شد و رضایت صریح آن‌ها دریافت شود.
- **سیستم‌های مدیریت رضایت:** سیستم‌هایی برای مدیریت و پیگیری مؤثر رضایت بیماران پیاده‌سازی شود.
- **شفافیت:** با بیماران درباره استفاده از LLMها و پیامدهای آن برای حریم خصوصی و امنیت داده‌ها شفاف باشد.

Best Practices for HIPAA Compliance of Large Language Models (LLMs)

1. Data Anonymization and De-identification

ناشناس‌سازی و حذف هویت داده‌ها

یکی از اقدامات اساسی برای رعایت HIPAA این است که هرگونه PHIمورد استفاده در آموزش یا تعامل با LLMها به‌طور صحیح ناشناس‌سازی یا هویت‌زدایی شود.

- **بهترین شیوه‌ها:**
- **حذف شناسه‌ها:** مجموعه داده‌ها از شناسه‌های مستقیم مانند نام، آدرس و شماره تامین اجتماعی پاک‌سازی می‌شوند.
- **تجمیع داده‌ها:** از داده‌های تجمیع‌شده استفاده می‌شود که قابل ردیابی به بیماران منفرد نباشد.
- **پنهان‌سازی داده‌ها:** تکنیک‌های پنهان‌سازی داده برای مخفی کردن اطلاعات شناسایی‌کننده اعمال می‌شوند.
- **حسابرسی‌های منظم:** حسابرسی‌های منظم انجام می‌شود تا اطمینان حاصل شود که داده‌های هویت‌زدایی‌شده همچنان غیرقابل شناسایی باقی می‌مانند.

10. Legal and Regulatory Compliance

رعایت قوانین و مقررات

اطلاع از تغییرات قوانین و مقررات برای رعایت HIPAAحیاتی است.

- **بهترین شیوه‌ها:**
- **مشاوره حقوقی:** از مشاوره حقوقی با تخصص در HIPAAو هوش مصنوعی استفاده می‌شود تا رعایت مقررات تضمین شود.
- **برنامه‌های رعایت مقررات:** برنامه‌های جامع رعایت مقررات تدوین می‌شوند و به‌طور منظم به‌روزرسانی می‌شوند.
- **به‌روزرسانی سیاست‌ها:** سیاست‌ها و رویه‌ها به‌طور منظم به‌روزرسانی می‌شوند تا با مقررات و استانداردهای جدید هم‌راستا باشند.

منابع

1. Abdur Rahman M, Barek MA, Riad AB, Mostafizur Rahman M, Bajlur Rashid M, Ambedkar S, Raihan Miaa M, Wu F, Cuzzocrea A, Ahamed SI. Embedding with Large Language Models for Classification of HIPAA Safeguard Compliance Rules. arXiv e-prints. 2024 Oct:arXiv-2410.
2. Ogunde F. Navigating the legal landscape: large language models and the hesitancy of legal professionals. International Journal of the Legal Profes
3. Moreno JG, Lovon J, Robin-Charlet MR, Damase-Michel C, Tamine L. PatientDx: Merging Large Language Models for Protecting Data-Privacy in Healthcare. arXiv preprint arXiv:2504.17360. 2025 Apr 24.sion. 2024 Sep 1;31(3):311-22.
4. Neupane S, Mittal S, Rahimi S. Towards a hipaa compliant agentic ai system in healthcare. arXiv preprint arXiv:2504.17669. 2025 Apr 24.
5. Rezaeikhonakdar D. AI Chatbots and Challenges of HIPAA Compliance for AI Developers and Vendors. J Law Med Ethics. 2023;51(4):988-995. doi: 10.1017/jme.2024.15. Epub 2024 Mar 13. PMID: 38477276; PMCID: PMC10937180.
6. Rabin R, McGregor S, Judd N. Malicious and Unintentional Disclosure Risks in Large Language Models for Code Generation. arXiv preprint arXiv:2503.22760. 2025 Mar 27.
7. Hui B, Yuan H, Gong N, Burlina P, Cao Y. Pleak: Prompt leaking attacks against large language model applications. InProceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security 2024 Dec 2 (pp. 3600-3614).
8. Sternak T, Runje D, Granoša D, Wang C. Automating Prompt Leakage Attacks on Large Language Models Using Agentic Approach. arXiv preprint arXiv:2502.12630. 2025 Feb 18.
9. Agarwal D, Fabbri AR, Risher B, Laban P, Joty S, Wu CS. Prompt leakage effect and defense strategies for multi-turn llm interactions. arXiv preprint arXiv:2404.16251. 2024 Apr 24.
10. Morris JX, Campion TR, Nutheti SL, Peng Y, Raj A, Zabih R, Cole CL. Diri: Adversarial patient reidentification with large language models for evaluating clinical text anonymization. AMIA Summits on Translational Science Proceedings. 2025 Jun 10;2025:355.


THANK YOU