پیشنهاد پروژه ارزیابی امنیتی و بهره‌برداری از مدل‌های زبان بزرگ (LLM) در حوزه امنیت سایبری

مقدمه:

در سال‌های اخیر، مدل‌های زبان بزرگ (LLM) مانند BERT ،GPT-4 و T5 به یکی از پیشرفته‌ترین دستاوردهای هوش مصنوعی تبدیل شده‌اند. این مدل‌ها توانسته اند با استفاده از معماری‌های پیچیده یادگیری عمیق، در حوزه‌هایی مانند پردازش زبان طبیعی، ترجمه ماشینی، تولید محتوا و تحلیل داده‌ها تحولات چشمگیری ایجاد کنند. با این حال، قدرت و انعطاف‌پذیری بالای این مدل‌ها چالش‌های امنیتی جدیدی را نیز به همراه آورده است. تهدیداتی مانند استخراج اطلاعات حساس، دستکاری داده‌ها و تولید محتوای مخرب می‌توانند امنیت کاربران و سازمان‌ها را به خطر می اندازند. از سوی دیگر، توانمندی‌های این مدل‌ها می‌تواند به ابزاری قدرتمند در تقویت امنیت سایبری نیز تبدیل شود. بنابراین، ضمن توجه به فرصت ها و تهدیدات این فناوری جدید، مسیرهای پژوهشی زیر در حوزه‌های مرتبط با امنیت نرم‌افزار و LLM را پیشنهاد می‌دهد:

1. ایجاد پایگاه دانش آسیب‌پذیری‌های رایج در سیستم‌های مبتنی بر LLM
2. ارزیابی و برقراری امنیت در نرم‌افزارهای مبتنی بر LLM
3. استفاده از LLM در تشخیص آسیب‌پذیری‌ها

در ادامه، هر یک از این مسیرهای پژوهشی را بیشتر بررسی می کنیم.

1. ایجاد پایگاه دانش آسیب‌پذیری‌های رایج در برنامه‌های مبتنی بر LLM:

با توجه به قابلیت های بالا در پردازش و تولید زبان های طبیعی و دسترسی این مدل ها به داده های حساس، ممکن است در برنامه های مبتنی بر مدل های LLM آسیب پذیری های جدیدی نظیر افشای اطلاعات محرمانه، تزریق دستور، دستکاری داده ها و غیره ایجاد شود. ممکن است عامل هوشمند LLM مانند یک کارمند ناآگاه، هدف حمله مهندسی اجتماعی قرار گرفته و امنیت کل داده های یک سامانه را دچار مخاطره کند. به جهت اهمیت موضوع، سازمان OWASP پروژه معرفی مجموعه‌ی ده آسیب‌پذیری حیاتی LLM را شروع کرده و هر سال خطرناک ترین آسیب پذیری های ممکن در برنامه های مبتنی بر LLM را گزارش می دهد. به همین جهت، این پژوهش بر شناسایی کامل آسیب پذیری های ممکن در برنامه های مبتنی بر LLM، نحوه سواستفاده از آن ها و روش‌های جلوگیری از این آسیب‌پذیری‌ها تمرکز خواهد داشت.

خروجی این پژوهش می‌تواند شامل موارد زیر باشد:

- **شناسایی آسیب‌پذیری‌ها**: بررسی آسیب‌پذیری‌های رایج در سیستم‌های مبتنی بر LLM، از جمله استخراج اطلاعات، دستکاری داده‌ها و تولید محتوای مخرب.

- **مستندسازی**: ایجاد یک پایگاه داده جامع از انواع آسیب‌پذیری‌ها، شامل توضیحات، روش‌های سوءاستفاده و راهکارهای پیشگیری، ارائه آموزش و گزارش به توسعه دهندگان و متخصصان امنیت.
- **پیاده‌سازی برنامه‌های آموزشی**: طراحی برنامه‌های شبیه‌سازی شده با آسیب‌پذیری‌های عمدی برای آموزش متخصصان امنیتی.

2. ارزیابی و برقراری امنیت در نرم‌افزارهای مبتنی بر LLM:

با توجه به اینکه به کارگیری LLM در آینده در انواع برنامه ها و سامانه‌ها به وضوح قابل پیش بینی است، تحلیل امنیت این برنامه‌ها از نیازهای مهم آینده خواهد بود. تاکنون روش های پیشرفته مختلفی جهت تحلیل و تشخیص خودکار آسیب‌پذیری در انواع نرم‌افزارها، مانند برنامه های وب، موبایل، برنامه های embedded، ارائه شده است. در این مسیر پژوهشی، منطبق سازی روش‌های موجود تحلیل و تشخیص آسیب پذیری و توسعه روش‌های جدید برای این برنامه‌ها مورد توجه قرار خواهد گرفت.

خروجی این مسیر پژوهشی شامل موارد زیر خواهد بود:

- تدوین چک‌لیست و فرآیند ارزیابی امنیتی این برنامه‌ها.
- ارائه الگوریتم‌های ایستا و پویا (مانند فازینگ) مبتنی بر ویژگی‌های ساختاری برنامه‌های مبتنی بر LLM.
- پیاده سازی الگوریتم های پیشنهادی و ارائه ابزارهای خودکار تحلیل و تست برنامه‌های مبتنی بر LLM.

3. استفاده از LLM در تشخیص آسیب‌پذیری‌ها:

مدل‌های LLM با قابلیت‌های پردازش الگوها و تحلیل داده‌های حجیم، می‌توانند ابزارهای مؤثری برای تشخیص آسیب‌پذیری‌ها در نرم‌افزارهای بزرگ و تحلیل سریعتر شرایط پیچیده باشند. این مدل‌ها قادرند از طریق تحلیل کدها، مستندات و گزارش‌های آسیب‌پذیری‌های قبلی به شناسایی آسیب پذیری های جدید کمک کنند. در این مسیر پژوهشی، روش‌های مختلف استفاده از LLM برای تشخیص آسیب‌پذیری‌ها بررسی خواهد شد.

روش‌های استفاده از LLM در تشخیص آسیب‌پذیری‌ها:

1. **تحلیل ایستای کد برای شناسایی آسیب‌پذیری‌ها**: با توجه به قدرت تحلیل الگوهای زبانی، مدل‌های LLM می توانند الگوی آسیب‌پذیری‌های رایج مانند Buffer Overflow، XSS، SQL Injection و Race Conditions را دریافت و در متن برنامه ها جستجو کنند. تعریف این الگوها و تنظیم و آماده سازی مدل های LLM بدین منظور، هدف مورد نظر این مسیر پژوهشی است.

2. **شبیه‌سازی حملات و شناسایی آسیب‌پذیری‌های ناشناخته**: با ارائه الگوی حمله و سواستفاده از یک کلاس آسیب پذیری به LLM، می توان به شبیه‌سازی سناریوهای حملات جدید پرداخته و پایگاهی از الگوهای حمله جهت تحلیل پویای برنامه ها و تشخیص آسیب‌پذیری‌های ناشناخته بدست آورد. در این مسیر پژوهشی کارگیری LLM در تحلیل پویای برنامه ها و تشخیص آسیب پذیری مورد مطالعه قرار می گیرد.
3. **استفاده از LLM در روش فازینگ برای شناسایی آسیب‌پذیری‌ها**: مدل‌های LLM قادرند با ایجاد ورودی‌های هوشمندتر و پیچیده‌تر منطبق مطابق با سناریوهای حملات واقعی، به بهبود فرآیند فازینگ کمک کنند. این ورودی‌ها نه تنها تصادفی بلکه دقیقاً هدفمند هستند و می‌توانند بر اساس تحلیل آن ها از کد یک برنامه و شرایط پیمایش یک مسیر اجرایی خاص ایجاد شده باشند. ترکیب روش های قبلی فازینگ با ایجاد ورودی تست توسط LLM به منظور گسترش دقت و سرعت تشخیص آسیب پذیری در این مسیر پژوهشی مورد نظر خواهد بود.

جمع بندی:

پروپوزال حاضر در نظر دارد که با استفاده از توانمندی‌های مدل‌های زبان بزرگ (LLM) در تحلیل کدها، شبیه‌سازی حملات و شناسایی آسیب‌پذیری‌ها، امنیت سامانه‌های مبتنی بر این مدل‌ها را تقویت کند. همچنین، با توسعه روش‌ها و ابزارهای خودکار، آموزش متخصصان امنیتی و مستندسازی آسیب‌پذیری‌ها، قدم‌های مؤثری در زمینه امنیت برنامه های مبتنی بر LLM برداشته خواهد شد.