**Oto najważniejsze zagrożenia dla generatywnej sztucznej inteligencji. Zero Trust AI przyszłością.**

**Duże modele językowe (LLM) rewolucjonizują sposób, w jaki wchodzimy w interakcję z technologią. W rezultacie dostawcy SaaS, walcząc o przewagę konkurencyjną, oferują przedsiębiorstwom narzędzia AI, takie jak analitykę sprzedaży lub asystentów kodowania. Jak jednak zabezpieczać usługi oparte o sztuczną inteligencję?**

Tradycyjnie modele bezpieczeństwa o zerowym zaufaniu opierały się na wyraźnym rozróżnieniu między użytkownikami i aplikacjami. Jednak aplikacje zintegrowane z [LLM](https://en.wikipedia.org/wiki/Large_language_model) zakłócają to rozróżnienie, funkcjonując jednocześnie jako użytkownik i program. Ta rzeczywistość wprowadza nowy zestaw luk w zabezpieczeniach, narażających na wyciek danych, natychmiastowe wstrzykiwanie (prompt injection), ryzykowny dostęp do zasobów online, a nawet dostęp do zasobów korporacyjnych w imieniu pracowników – wymieniają specjaliści Check Point Software. Aby sprostać tym wyzwaniom, potrzebny jest unikalny zestaw środków typu „zero trust”.

**„Natychmiastowe wstrzyknięcie”**

Osoby atakujące nadużywają duże modele językowe, tworząc dane wejściowe umożliwiające manipulację zachowaniem LLM, bezpośrednio lub pośrednio, w celu wywołania szkodliwego lub nieetycznego zachowania.

Podpowiedzi mogą być wprowadzane bezpośrednio przez osobę atakującą lub pośrednio przez nieświadomego użytkownika korzystającego z aplikacji opartej na LLM w określonym przypadku.

Cztery rodzaje prompt injections to:

l **Direct Prompt Injection**, polega na tym, że atakujący wprowadzają określone podpowiedzi, aby zmienić zachowanie lub dane wyjściowe LLM w szkodliwy sposób. Osoba atakująca może bezpośrednio poinstruować LLM, aby odgrywała rolę nieetycznego modelu, udostępniać poufne informacje lub spowodować wykonanie przez model szkodliwego kodu.

l **Indirect Prompt Injection** jest bardziej subtelne i obejmuje manipulację źródłami danych, z których korzysta LLM, co czyni je znacznie bardziej niebezpiecznymi i trudniejszymi do wykrycia w środowiskach organizacyjnych.

l **Multimodalne** **Prompt Injection** umożliwia LLM odbieranie formatów takich jak obrazy, filmy i dźwięki jako dane wejściowe, z ukrytymi instrukcjami dodawanymi do danych wejściowych multimediów w celu zmiany zachowania bota aplikacji, [dzięki czemu może on rozmawiać jak Harry Potter](https://arxiv.org/pdf/2307.10490v4.pdf).

l **Ataki typu „odmowa usługi” (DoS)** można również przeprowadzać za pomocą natychmiastowych wstrzyknięć, co prowadzi do operacji obciążających zasoby na LLM aż do punktu przeciążenia, co prowadzi do degradacji usług lub wysokich kosztów.

**Wyciek wrażliwych danych – czy Twoja sztuczna inteligencja potrafi zachować tajemnicę?**

Modele można dostrajać lub rozszerzać dzięki dostępowi do danych, aby osiągnąć lepsze wyniki specyficzne dla danej dziedziny. Na przykład w przypadku bota obsługi klienta świetnie byłoby dostroić model na podstawie wcześniejszych zgłoszeń problemów. Ale czy Twoja sztuczna inteligencja potrafi zachować tajemnicę?

Eksperci Check Pointa podali przykład jednego z badań, w którym badacze wykorzystali mechanizm dostrajający ChatGPT do [wyodrębnienia nazwisk i adresów e-mail](https://www.nytimes.com/interactive/2023/12/22/technology/openai-chatgpt-privacy-exploit.html) ponad 30 pracowników New York Timesa. Przykład ten ukazuje, jak wrażliwe dane wykorzystywane do wstępnego szkolenia lub dostrajania LLM mogą wyciec. Okazuje się, że nie można ufać modelom LLM w zakresie ochrony wrażliwych danych przed wyciekiem.

Modele generatywnej sztucznej inteligencji przechodzą intensywne szkolenia na różnych zbiorach danych, często obejmujących większość treści internetowych. Proces szkolenia obejmuje wstępny trening na dużych zbiorach danych w celu uzyskania szerokiego zrozumienia języka i świata, a następnie dopracowanie pod kątem konkretnych celów przy użyciu wybranych zbiorów danych.

Zdaniem Check Pointa w przypadku zatruwania danych osoby atakujące mogą naruszyć bezpieczeństwo modeli, manipulując niewielką częścią, zaledwie 0,01% danych szkoleniowych. Ponieważ modelom i użytkownikom nie można ślepo ufać, nie można zakładać, że integralność i bezpieczeństwo danych szkoleniowych są również wiarygodne.

**Kontrola dostępu - Witamy na Dzikim Zachodzie**

Coraz większa liczba organizacji integruje LLM z aplikacjami wieloskładnikowymi, czyli „agentami”. Integracje te rozszerzają LLM o możliwości, takie jak dostęp do Internetu, odzyskiwanie zasobów korporacyjnych i wykonywanie na nich różnych działań. Warto zauważyć, że niedawne uruchomienie sklepu z wtyczkami [OpenAI ułatwia powszechny dostęp do rozszerzeń LLM.](https://openai.com/blog/introducing-the-gpt-store)

Pobieranie danych w czasie rzeczywistym z Internetu może być niezwykle cenne dla użytkowników. Te udoskonalenia umożliwiają LLM zapewnianie lepszych odpowiedzi na zapytania użytkowników w oparciu o aktualne informacje. Jednak zwiększenie dostępu LLM do Internetu stanowi dramatyczne wyzwanie, szczególnie w kontekście prompt injection. Ostatnio na przykład wstawianie złośliwych instrukcji do adresów URL powodowało, że [czat Bing namawiał użytkowników do odwiedzenia złośliwej witryny lub ujawnienia poufnych informacji](https://dl.acm.org/action/downloadSupplement?doi=10.1145%2F3605764.3623985&amp;amp;file=3623985-cvor.pdf) przesłanych na serwer zewnętrzny.

Aplikacje zintegrowane z LLM można zaprojektować tak, aby współpracowały z zasobami korporacyjnymi, takimi jak bazy danych lub aplikacje. Jednak ten rodzaj dostępu stwarza ryzyko, nawet w przypadku wewnętrznych użytkowników, ponieważ mogą oni przypadkowo uzyskać dostęp do wrażliwych danych i zasobów w wyniku interakcji z aplikacją zintegrowaną z LLM.

Podejście oparte na dostępie Zero Trust AI (ZTAI) proponuje postrzeganie aplikacji zintegrowanych z LLM jako podmiotów wymagających ścisłej kontroli dostępu, ochrony danych i zasad zapobiegania zagrożeniom – tworząc bardziej rygorystyczną linię obrony, niż byłoby to konieczne do zabezpieczenia przeciętnego pracownika. To wg analityków Check Pointa jedyny sposób na bezpieczeństwo organizacji aktywnie korzystającej z modeli AI wykorzystujących dane wewnętrzne.