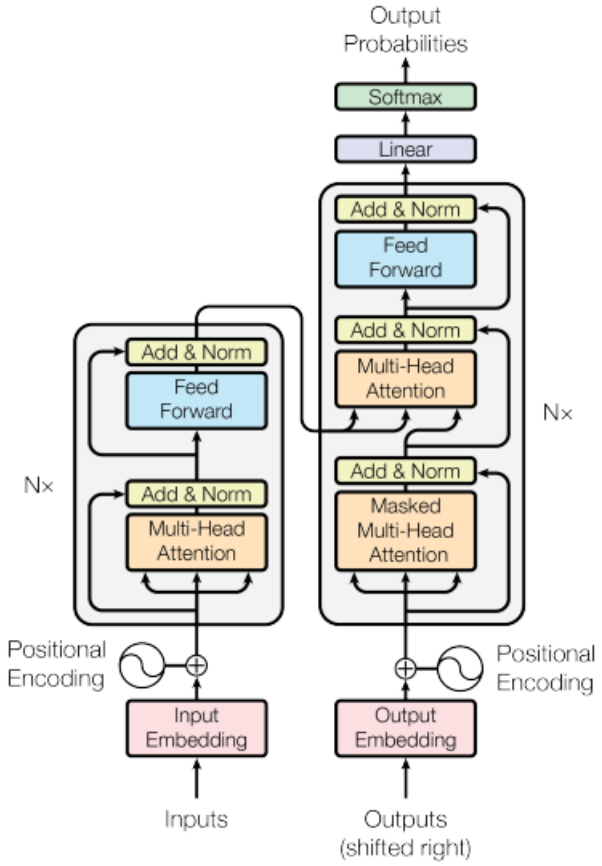
Bezpieczeństwo aplikacji opartych na sztucznej inteligencji

Jacek Wojcieszyński, kwiecień 2024

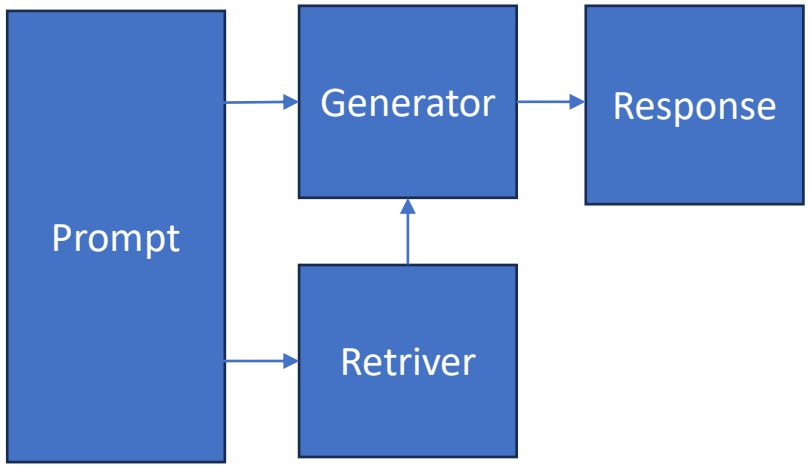
Krótkie wprowadzenie do wielkich modeli językowych LLM (Large Language Model)



- GPT (Generative Pre-trained Transformer) to rodzina modeli językowych (LLM) zbudowanym w architekturze transformer'a
- Modele GPT zostały wytrenowane na danych pozyskanych z Internetu, w tym z Wikipedii
- GPT generuje słowo po słowie w kontekście poprzednich słów i całego tekstu, dzięki mechanizmowi samouwagi (self attention)
- Transformer składa się z warstwy zanurzania przekształcającej słowa na wektory zwane embeddings, warstwy dekodera z mechanizmem samouwagi, która analizuje kontekst i generuje sekwencje, oraz warstwy wyjściowej przypisującej prawdopodobieństw kolejnym słowom w sekwencji

Transformer. Źródło: "Attention is All you Need" (Vaswani, i inni, 2017)

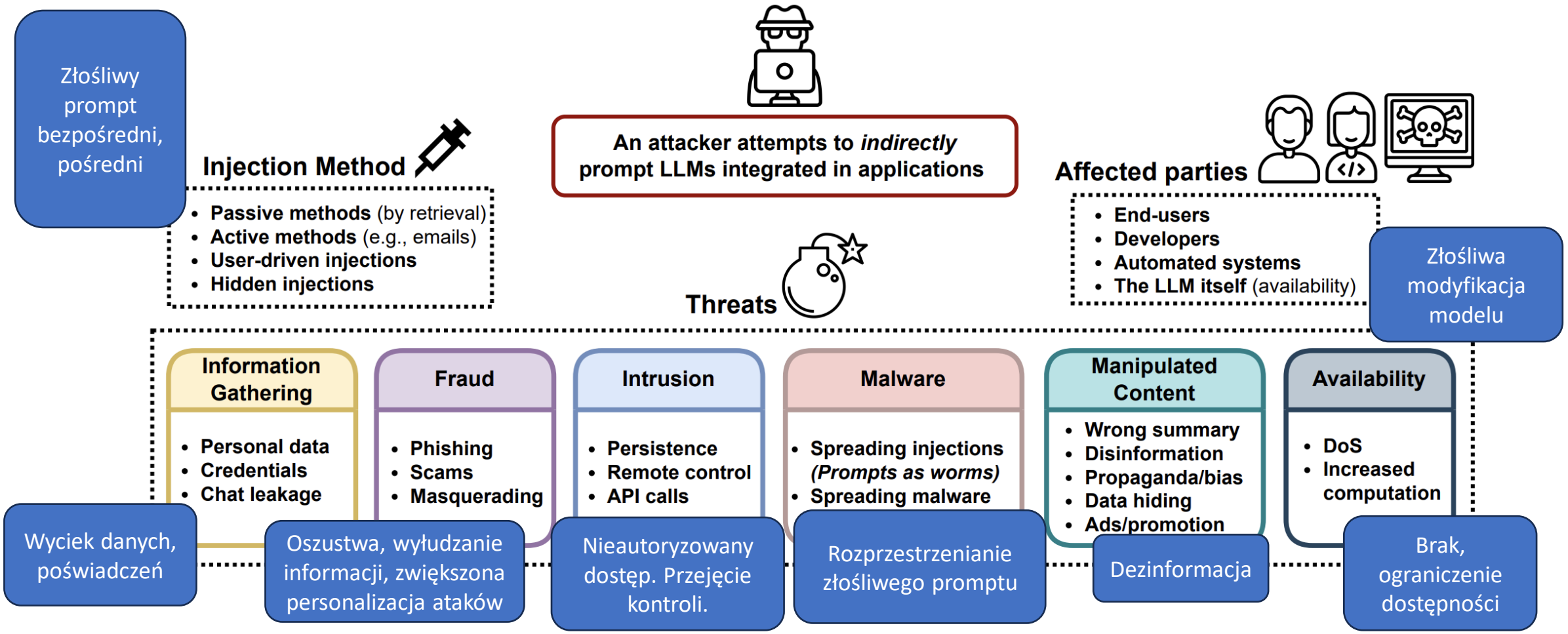
Architektura aplikacji LLM na przykładzie RAG (Retrieval-Augmented Generation)



Koncepcja RAG

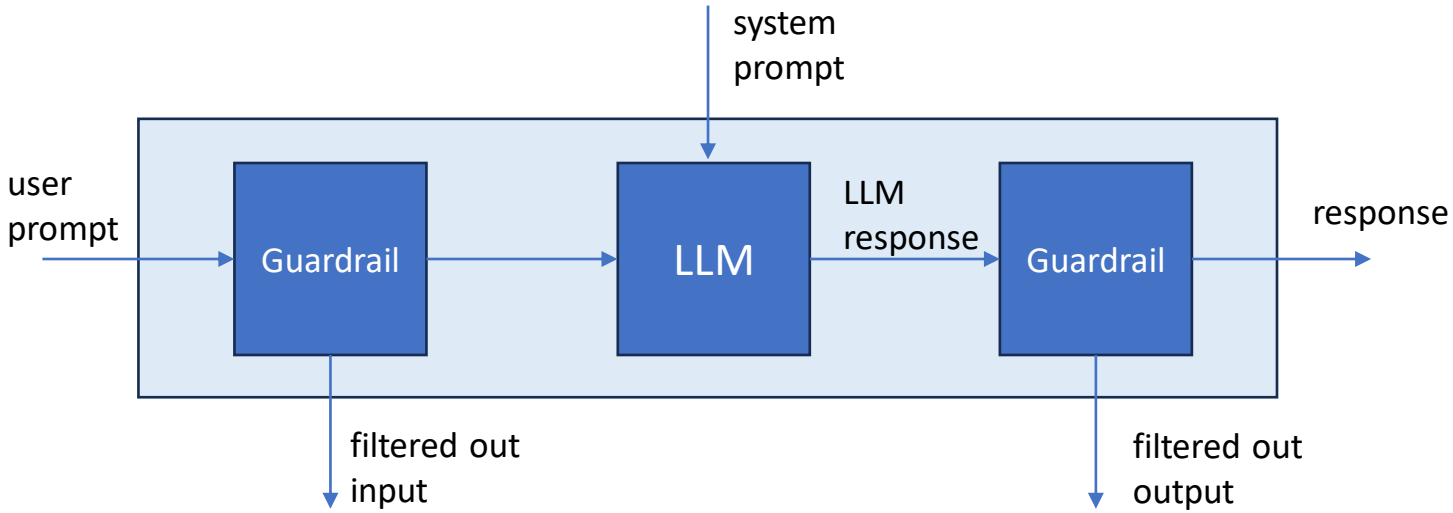
- Przygotowanie bazy wiedzy
  - Konwersja dokumentów na tokeny
  - Mapowanie tokenów na wektory (embeddings)
  - Składowanie embeddings w bazach wektorowych
- Przetwarzanie zapytania (prompt’u)
  - Zamiana zapytania na wektor (embedding)
  - Wyciągnięcie podobnych wektorów do zapytania z bazy wektorowej wraz z kontekstem
  - Wzbogacenie zapytania o pozyskane dane z bazy wektorowej
- Przygotowanie odpowiedzi
  - Przetwarzanie wzbogaconego zapytania przez LLM
  - Zwrócenie odpowiedzi do użytkownika

Mapa zagrożeń LLM



Źródło: Greshake i inni, „Not what you've signed up for”, 2023

Zabezpieczenia



Zabezpieczenia

- Dobry prompt systemowy
- Biała lista znaków dozwolonych
- Czarna lista znaków zakazanych
- Filtrowanie przy użyciu LLM
- Parametryzacja konwersacji
- Zasada braku zaufania do danych
- Ograniczanie dostępu do minimum
- Wykrywanie anomalii w odpowiedziach
- Wykrywanie stronniczości LLM
- Ograniczanie uprawnień LLM
- Badanie pochodzenia modelu
- Badanie podatności, patchowanie
- Monitorowanie aktywności LLM
- Badanie pochodzenia danych
- Monitorowanie przydzielonych zasobów
- Audytowanie bezpieczeństwa
- Podnoszenie świadomości użytkowników
- Wsparcie prawne

Podsumowanie