API4:2023 Unrestricted Resource Consumption

Grzegorz Koperwas

30 listopada 2023

W sprawozdaniu omówimy sposoby przeciwdziałania podatności `API4:2023 Unrestricted Resource Consumption`. Będziemy omawiać je w kontekście aplikacji wykorzystujących rozwiązania oparte na LLM.

LLM, z racji ogromnej ilości parametrów, wymagają dużych ilości pamięci. Są również nie przewidywalne, podatne na ataki *prompt injection* oraz są nowym trendem, przez co nie koniecznie mamy zawsze do czynienia z przemyślanym rozwiązaniem.

Z tego powodu będziemy się skupiać na rozwiązaniach, które nie tylko ograniczają prawdopodobieństwo zablokowania się systemu, ale również ograniczających jego podatności na ataki *prompt injection*.
Będziemy ewaluowali rozwiązania w ramach ich przydatności do zabezpieczenia przykładowej aplikacji realizującej *Retrieval Augmented Generation*, w skrócie `RAG`.

Aplikacje te wykorzystują wektorowe bazy danych, gdzie przechowywane są dane. W ramach zapytania baza taka potrafi dobrać informacje, które pomagają odpowiedzieć `LLM` na zadane pytania. Korzystają one z wyspecjalizowanych sieci zwanych *embeddings*, które zamieniają tekst na wektor. Mając nasz tekst w formie wektorowej, możemy łatwo porównywać dystans między różnymi elementami, czyli pośrednio ich podobieństwo.

Podstawową linią obrony przeciwko zbyt dużemu zużyciu zasobów jest ograniczenie tego, ile pojedyncze zapytanie do systemu może wywoływać modele `LLM`. Innym czynnikiem, który musimy kontrolować, jest *kontekst* oraz jego rozmiar. Modele `LLM` operują nie na nieskończonej ilości danych, tylko na pewnym wycinku tekstu, jaki jest im zadany jako argument.

Te przysłowiowe okienko, przez które na problem spoglądają `LLM` nazywamy kontekstem, nic poza nim nie ma wpływu na sieć. Jego rozmiar jest mierzony w *tokenach*, które mniej-więcej odpowiadają słowom lub częściom słów. Każdy model ma określony rozmiar kontekstu, przykładowo dla modeli `LLama` mamy do czynienia z kontekstem wielkości 2048 tokenów, a niektóre z modeli `GPT` mają konteksty rozmiaru nawet 128 tysięcy tokenów, dla modelu `GPT-4 Turbo`.

Jednak musimy pamiętać, że często koszt użycia modelu jest wyznaczany na podstawie ilości tokenów, które zostają przekazana jako kontekst, jak i ilości wygenerowanych tokenów.

Podczas budowania naszej aplikacji wykorzystującej `RAG` możemy chcieć dać modelowi dostęp do całej naszej bazy danych. Jednak przekazanie w kontekście wszystkiego jak leci, nie zadziała z następujących powodów:

- Będzie to kosztowne, ponieważ płacimy za każdy token.
- Będzie to nieefektywne, ponieważ nasze dane nie zmieszczą się w kontekście.

W celu rozwiązania tego problemu możemy skorzystać z biblioteki *langchain*, która pomaga nam zarządzać kontekstem.
Możemy skorzystać z jej pomocy w celu integracji z wektorową bazą danych, na przykład *chromadb*. Pozwoli nam to do naszego kontekstu wprowadzać tylko dokumenty (lub ich fragmenty) faktycznie związane z naszym problemem.
Z użyciem takiej bazy wektorowej możemy ograniczyć przetwarzane informacje tylko do konkretnych dokumentów lub ich fragmentów, co pozwala nam używać mniejszych kontekstów, co ogranicza koszty pojedynczego zapytania.

Dodatkowo możemy zastosować prosty limit ilości tokenów, które może model wygenerować. Inną opcją kontroli wyników modeli jest, na przykład w narzędziu `llama.cpp`, jest zadawanie gramatyki, jaką musi spełniać odpowiedź. Przykładowo, zamiast ograniczać ilość tokenów, możemy modelowi uniemożliwić wygenerowanie więcej niż dwóch zmian. Możemy również nakazać modelowi generowanie odpowiedzi w konkretnym formacie, na przykład `yaml`, zamiast opisywać mu w kontekście, jakie warunki ma spełniać jego odpowiedź.

Dodatkowo możemy stosować tradycyjne metody kontroli ilości zapytań, na przykład:

- Ograniczenia ilości zapytań dla klienta.
- Ograniczenia rozmiaru zapytania.

Jednak te metody nie są specyficzne dla modeli `LLM`, więc nie chcę ich szczegółowo omawiać w tym referacie.

- *Podsumowanie* - Modele `LLM` mogą korzystać zarówno z tradycyjnych metod zabezpieczania `API`, jak i z innych technik pozwalających optymalizować ich działanie.
- *Wnioski* - Koszt wykorzystania modeli `LLM` jest zależny od rozmiaru kontekstu, dlatego musimy kontrolować jego rozmiar. Jest to podobne do tego, czemu stosujemy paginację.