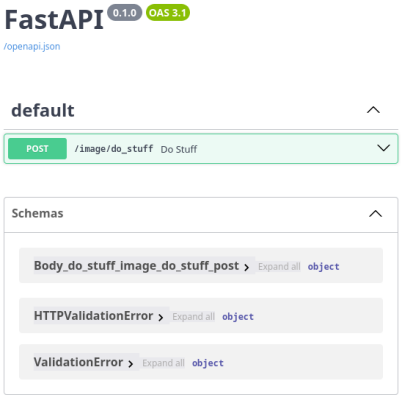
Zabezpieczanie webaplikacji opartych o LLM

Grzegorz Koperwas

Cel projektu

Celem projektu jest zabezpieczenie prostej aplikacji wykonanej w środowisku `FastAPI`, wykorzystującej model `LLava`.



Czym będziemy się zajmować?

W aplikacji będziemy zajmować się następującymi problemami, związanymi z `API4:2023 Unrestricted Resource Consumption`:

Czym będziemy się zajmować?

W aplikacji będziemy zajmować się następującymi problemami, związanymi z `API4:2023 Unrestricted Resource Consumption`:

- Podczas przetwarzania więcej niż dwóch zapytań aplikacja się zacina.

Czym będziemy się zajmować?

W aplikacji będziemy zajmować się następującymi problemami, związanymi z `API4:2023 Unrestricted Resource Consumption`:

- Podczas przetwarzania więcej niż dwóch zapytań aplikacja się zacina.
- Aplikacja może zjeść całą pamięć RAM

Czym nie będziemy się zajmować?

- Długość promptu - Model `LLava` nie przejmuje się promptem zbytnio, aplikacja była używana tylko jako wewnętrzne API ze stałym promptem.

Czym nie będziemy się zajmować?

- Długość promptu - Model `LLava` nie przejmuje się promptem zbytnio, aplikacja była używana tylko jako wewnętrzne API ze stałym promptem.
- Rozmiar załączników - Framework `FastAPI` nie ma dobrego mechanizmu na ich ograniczenie, więc tutaj najlepszym sposobem byłoby robienie tego za pomocą serwera proxy.

Live Demo

Live Demo

Podsumowanie

- Ograniczenie maksymalnej ilości instancji modelu gwarantuje nam pewien maksymalny poziom zużycia pamięci
- Podczas używania rozwiązań opartych o LLM musimy zdawać sobie sprawę, że czas oczekiwania na odpowiedź od takiego modelu może być długi