Pokročilé





Univerzitní AI Sandbox je postavený na platformě OpenWebUI. To je naše lokálně hostovaná platforma postavená na open-source softwaru s širokou škálou funkcí. Uživatelé mohou „chatovat“ s lokálními i cloudovými modely umělé inteligence, vytvářet specializované AI agenty, pracovat se znalostní bází a další.

Přístup k modelům je směrován přes bránu LiteLLM, která zajišťuje autentizaci, směrování požadavků a limity využití. Sada modelů dostupných jednotlivým uživatelům závisí na jejich afiliaci (roli a příslušnosti k fakultě).

Platforma i modely jsou hostovány lokálně na serverech MFF UK, více o univerzitním clusteru najdete zde .

Pro přístup k modelům přes API klíče prosím kontaktuje naše administrátory na ai.sandbox@cuni.cz

AI Sandbox je momentálně v testovacím provozu.

Pokročilá funkcionalita: Nástroje (Tools)
Nástroj je funkce v Pythonu, která rozšiřuje schopnosti modelu tím, že v reálném čase **volá externí služby.**

Nástroje se vytvářejí ve vestavěném editoru kódu v části **Workspace > Tools**. Každá funkce musí mít docstring s popisem svého účelu a parametrů. Model se podle něj rozhoduje, kdy a jak funkci zavolat.

Po uložení je nástroj k dispozici napříč konverzacemi a v jednotlivých chatech se zapíná **ikonou nástrojů** v poli pro zadání zprávy. Model nástroje volá samostatně podle kontextu promptu.

Nástroje lze používat v kanálech i v soukromých chatech.

V tuto chvíli nejsou nástroje ve výchozím nastavení povoleny. O povolení k jejich vytváření a používání požádejte **administrátora (ai.sandbox@cuni.cz).**

Pokročilá funkcionalita: Kanály (Channels)
Kanál je trvalá **skupinová konverzace**, kterou sdílí více uživatelů a model.

Kanály se vytvářejí z postranního panelu: zadáte název, vyberete model a později přidáte členy ze své skupiny. Všechny zprávy a odpovědi modelu vidí každý člen a zůstávají uložené i mezi jednotlivými relacemi.

Můžete také vytvořit přímou zprávu v kanálu, kterou uvidíte vy, jeden další uživatel a jeden jazykový model.

Pokročilá funkcionalita: Webhooky (Webhooks)
Webhook je odchozí **požadavek odeslaný na externí URL adresu**, když nastane určitá interní událost.
Webhooky najdete v části **Settings > Notifications > Notification Targets.**

Cíl (target) je URL adresa webhooku spojená s událostí, k jejichž odběru se přihlásíte: dokončení chatu, selhání chatu, zpráva v kanálu nebo jiné.

Pokročilá funkcionalita: Vlastní modely (Custom models)
Vlastní model je **uložená konfigurace** postavená na existujícím základním modelu.

Modely se vytvářejí v části **Workspace > Models:** vyberete základní model, přiřadíte mu název, napíšete systémový prompt a upravíte parametry generování, například teplotu a délku kontextu. Systémový prompt určuje chování modelu. Po uložení se model zobrazí vedle výchozích modelů platformy.

V tuto chvíli nejsou vlastní modely ve výchozím nastavení povoleny. O povolení k jejich vytváření a používání požádejte **administrátora (ai.sandbox@cuni.cz).**

Sledování využití přes LiteLLM
LiteLLM umožňuje uživatelům kontrolovat útratu v tokenech pomocí vlastního API klíče. Klíč uživatelé získají od administrátora sandboxu na adrese ai.sandbox@cuni.cz.

Využití v USD zobrazíte takto:

1. Přejděte do LiteLLM UI.

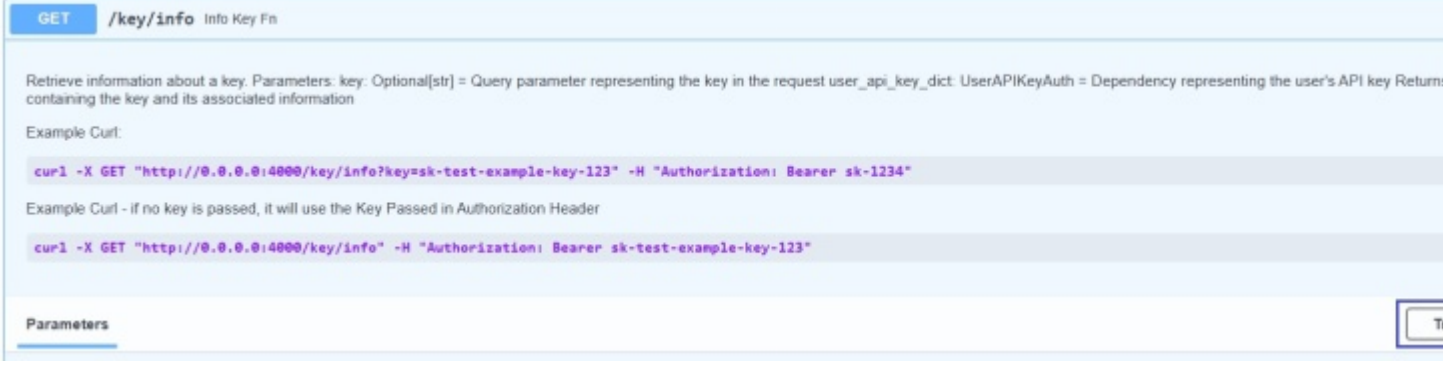
2. Přihlaste se výběrem možnosti **Authorize** a vložte svůj API klíč do pole **Value.**





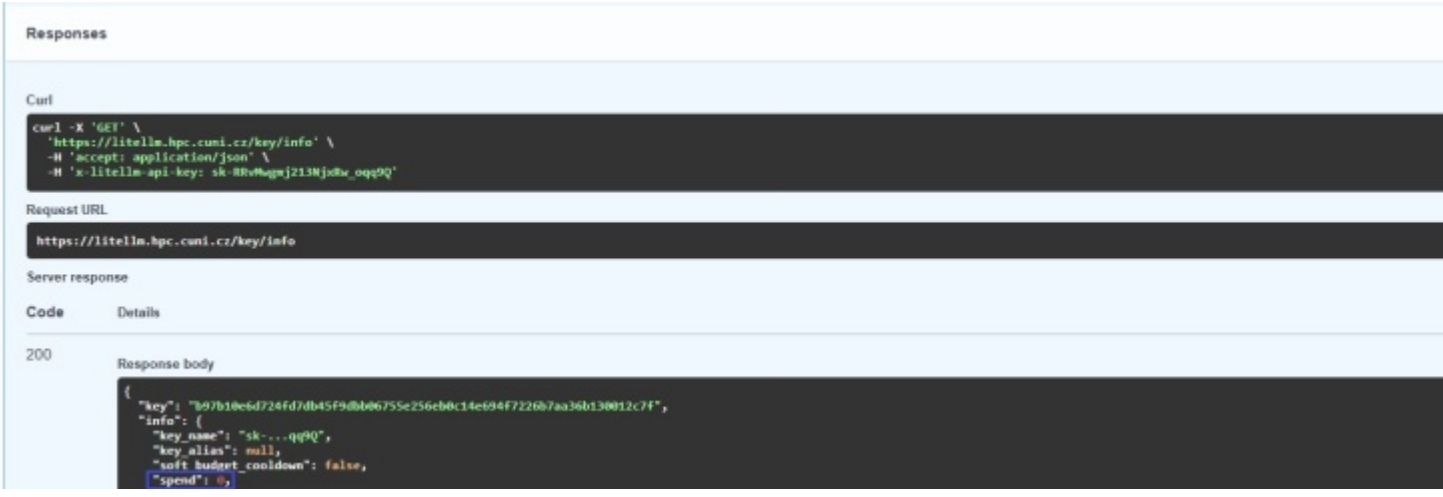
3. Přejděte na stránku key info .

4. Vyberte **Try it out** a poté volání spusťte tlačítkem **Execute.**





5. V detailech odpovědi uvádí **spend**, kolik USD jste celkem utratili, a **max_budget**, váš limit.



Chiméra HPC
Veškerá pokročilejší nastavení, jako práce s operátorem HPC najdete v dokumentaci Chiméra clusteru.