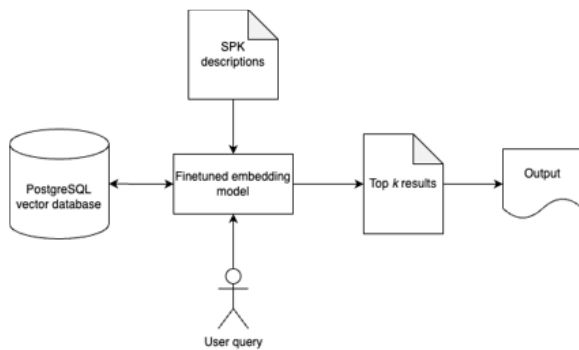


## Kopija \_\_Design



Čia pavaizduota šiuo metu naudojama vieno embedding modelio architektūra. Ji yra greita, reikalaujanti pakankamai nedaug resursų tiek mokymo, tiek naudojimo metu. Iki šiol tinkamo įterpinių modelio panaudojimas ir eksperimentai parodė, kad tokios architektūros užtenka geram modelio veikimui, o tolesnis tobulinimas gali vykti gerinant duomenų (aprašymų ir paieškų pavyzdžių) kokybę.

Kaip "**Finetuned embedding model**" naudojamas atvirų svorių iš anksto apmokytas kalbos modelis [intfloat/multilingual-e5-large-instruct](https://huggingface.co/intfloat/multilingual-e5-large-instruct), pilnai valdomas mūsų infrastruktūroje.

**SPK descriptions** – tai iš turimų paslaugų duomenų sugeneruoti tekstai pasitelkiant Llama-3-70B kalbos modelį. Paduodant paslaugos aprašymą, pavadinimą bei teikėjus, šio LLM užklausiama sugeneruoti ~100 žodžių ilgio paslaugos aprašymą, tinkamą semantinei paieškai. Toks automatizuotas generavimas reikalingas todėl, kad pateikti paslaugų aprašymai bei pavadinimai yra labai skirtingų ilgių/kokybės (tam tikrais atvejais jų išvis nėra), o tai kenkia modelio našumui. Verta pabrėžti, kad tai yra vienkartinis duomenų paruošimo procesas.

**DeepL ar bet kokie kitokie vertimai šiuo metu nėra naudojami.** Kadangi naudojamas embedding kalbos modelis yra daugiakalbis (iš anksto apmokytas 100 kalbų, įskaitant ir lietuvių), poreikio naudoti vertimus nematome., nes paieška veikia gerai ir be jų.