Seznam LLM: od prototypu až k produktu

Marek Šimůnek

Lead Software Engineer

Agenda

Vektorové modely v Seznamu

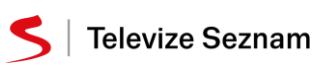
LLM v Seznamu

Zrychlujeme inference

Ponaučení

O Seznamu


Seznam.cz


Televize Seznam


Super.cz


stream


Sreality.cz


Sport.cz


Smoto.cz


Sklik.cz


Email


Sdovolená.cz


Czech Properties


Sblog


Sbazar.cz


Sauto.cz


Proženy


Počasí.cz


Novinky.cz


Mapy.cz


Lidé


Kupi.cz


Hry.cz


Horoskopy.cz


Garáž.cz


Zboží.cz


Firmy.cz


Seznam Zprávy


classic
praha


EXPRESFM

O Seznamu


Seznam.cz
Televize Seznam
Super.cz
stream
Sreality.cz
Sport.cz
Smoto.cz
Sklik.cz
Email
Sdovolená.cz
Czech Properties
Sblog
Sbazar.cz
Sauto.cz
Proženy
Počasí.cz
Novinky.cz
Mapy.cz
Lidé
Kupi.cz
Hry.cz
Horoskopy.cz
Garáž.cz
Zboží.cz
Firmy.cz
Seznam Zprávy
classic praha
Expres FM

O Seznamu


Seznam.cz
Televize Seznam
Super.cz
stream
Sreality.cz
Sport.cz
Smoto.cz
Sklik.cz
Email
Sdovolená.cz
Czech Properties
Sblog
Sbazar.cz
Sauto.cz
Proženy
Počasí.cz
Novinky.cz
Mapy.cz
Lidé
Kupi.cz
Hry.cz
Horoskopy.cz
Garáž.cz
Zboží.cz
Firmy.cz
Seznam Zprávy
classic praha
EXPRESFM

O mně

40,000,000,000 klíčů v databázi
300,000,000 stažení denně

- Big data
- MLOps
- LLM

1000 fyzických serverů
15 PB úložného prostoru
65 TB paměti

Platforma pro 5 výzkumnických týmů
25+ lidí
100+ modelů

Seznam LLM co umí česky

Vektorové modely v Seznamu

- BERT = **B**idirectional **E**ncoder **R**epresentations from **T**ransformers

Vektorové modely v Seznamu

- BERT = **B**idirectional **E**ncoder **R**epresentations from **T**ransformers
- Small-E-Czech [čti: smolíček] - Github[1] a HuggingFace[2]

Zdroj:

[1] https://github.com/seznam/small-e-czech

[2] https://huggingface.co/Seznam/small-e-czech

[3] https://github.com/seznam/czech-semantic-embedding-models

Vektorové modely v Seznamu

- BERT = **B**idirectional **E**ncoder **R**epresentations from **T**ransformers
- Small-E-Czech [čti: smolíček] - Github[1] a HuggingFace[2]
  - zvýšení kvality vyhledávání a opravy překlepů

Zdroj:

[1] https://github.com/seznam/small-e-czech

[2] https://huggingface.co/Seznam/small-e-czech

[3] https://github.com/seznam/czech-semantic-embedding-models

Vektorové modely v Seznamu

- BERT = **B**idirectional **E**ncoder **R**epresentations from **T**ransformers
- Small-E-Czech [čti: smolíček] - Github[1] a HuggingFace[2]
  - zvýšení kvality vyhledávání a opravy překlepů
- RetroMAE-Small, Dist-MPNet-ParaCrawl
  - high-quality sentence embeddings (similarity search, retrieval, clustering or classification)

Zdroj:

[1] https://github.com/seznam/small-e-czech

[2] https://huggingface.co/Seznam/small-e-czech

[3] https://github.com/seznam/czech-semantic-embedding-models

Large Language Models (LLM) hype

ChatGPT 3.5

ChatGPT 4

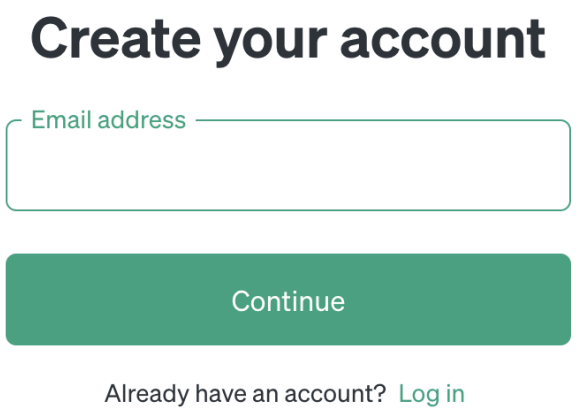
LLM v Seznamu

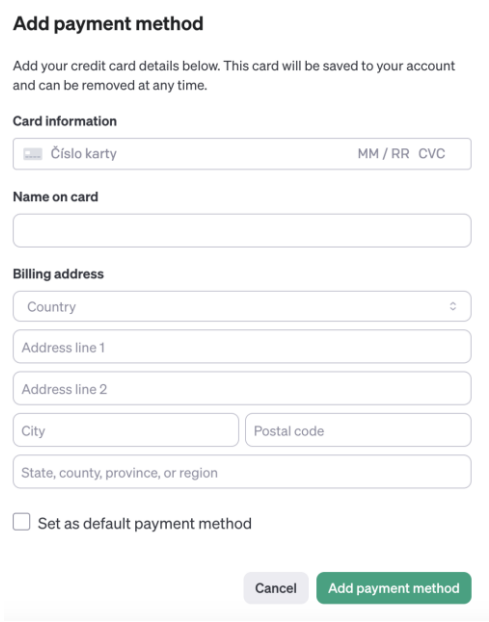
Máme produktovou strategii

Máme výzkumné know how

Jak to rozšířit po firmě

OpenAI





Security

Budgety pro týmy

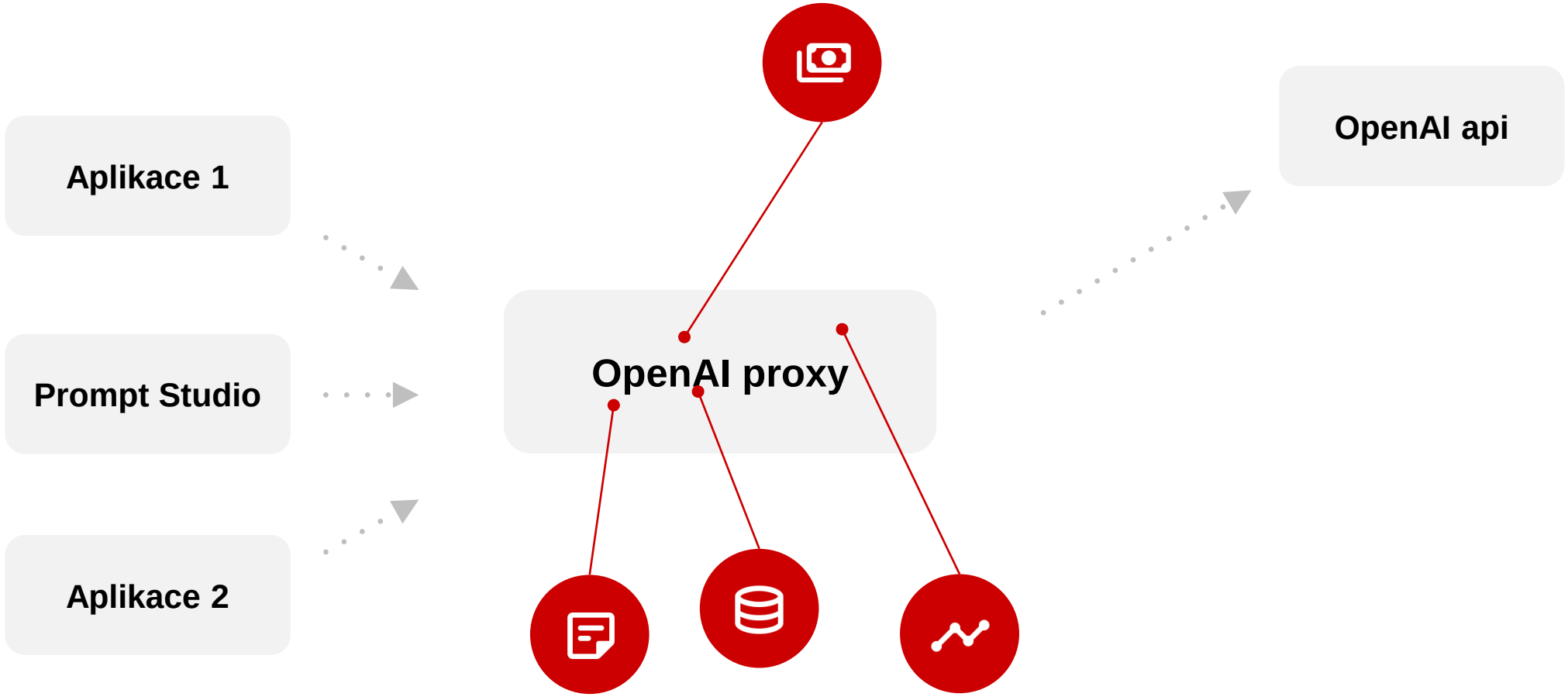
Měření kvality

OpenAI proxy

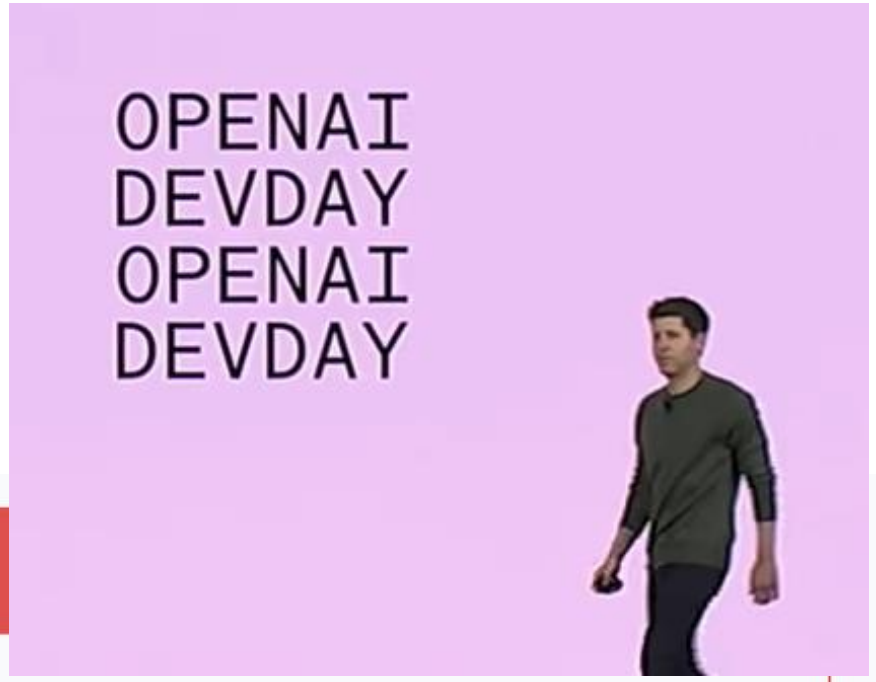
Interní token

URL na naší proxy

OpenAI proxy



OpenAI výpadek



Major Outage across ChatGPT and API

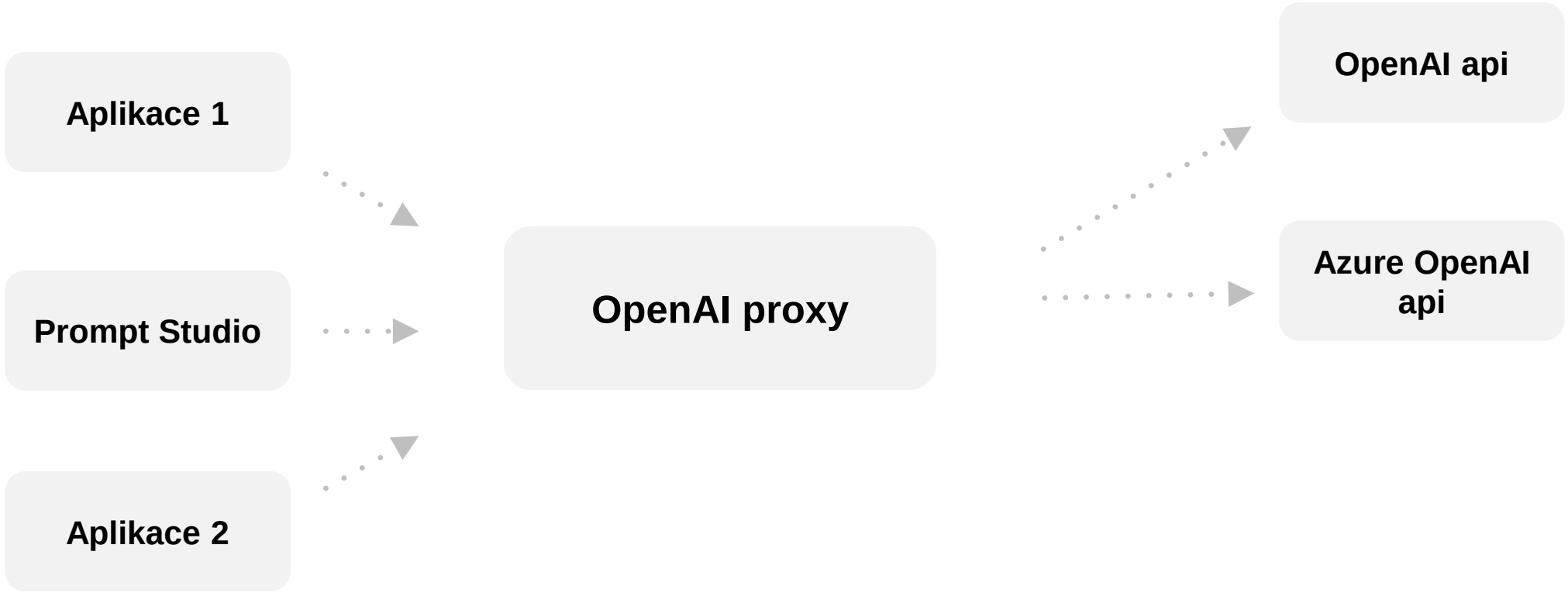
Identified - We've identified an issue resulting in high error rates across the API and ChatGPT, and we are working on remediation.
Nov 08, 2023 - 06:50 PST

Update - We are seeing high error rates across the API and ChatGPT and are actively investigating possible causes.
Nov 08, 2023 - 06:49 PST

Update - We are continuing to investigate this issue.
Nov 08, 2023 - 06:02 PST

Investigating - We are currently investigating this issue.
Nov 08, 2023 - 05:54 PST

OpenAI proxy



OpenAI drama

CEO Sam Altman vystoupil
a pak nastoupil

Naše filozofie a cíle pro Seznam LLM modely

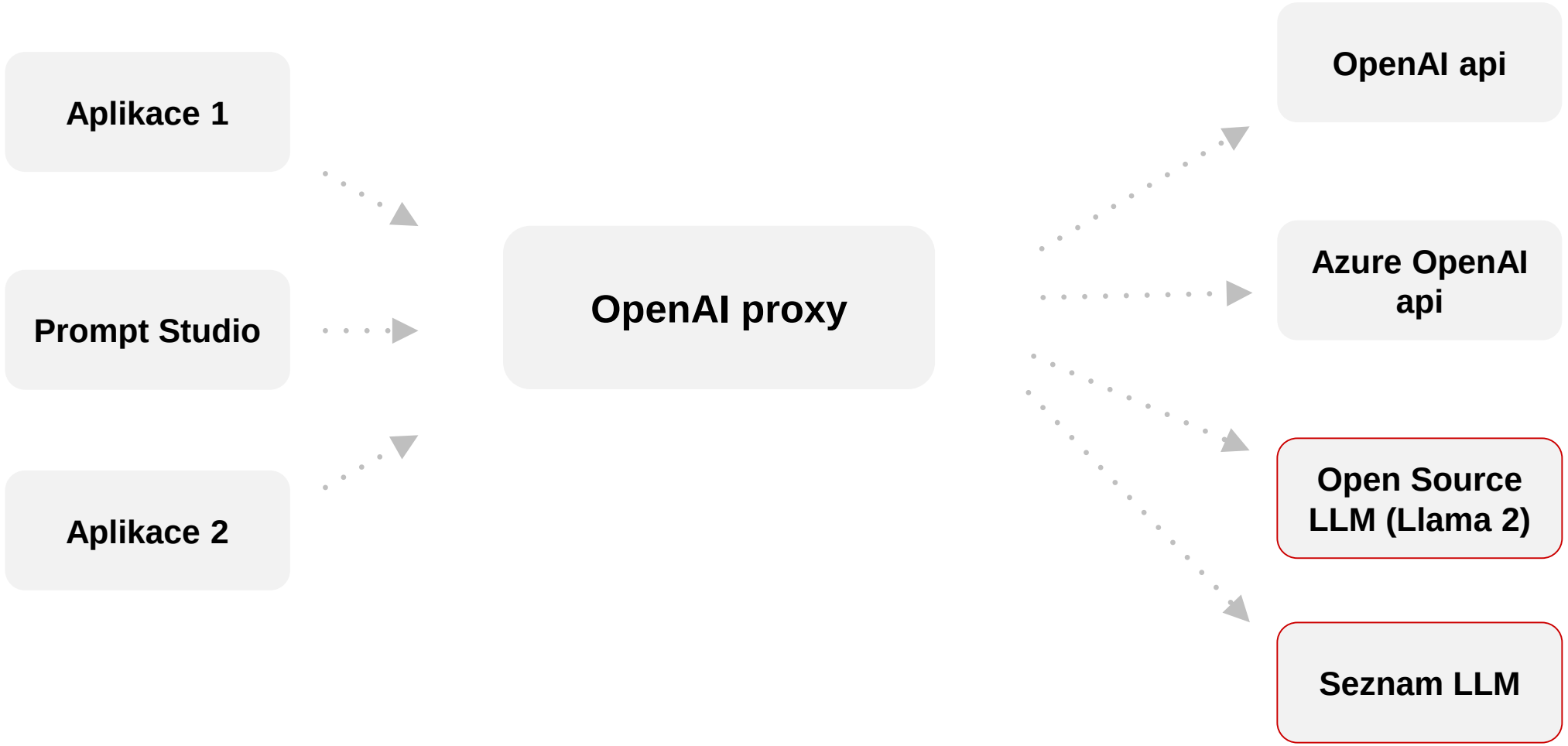
Modely dobré na češtinu

Nečekat a ukázat MVP

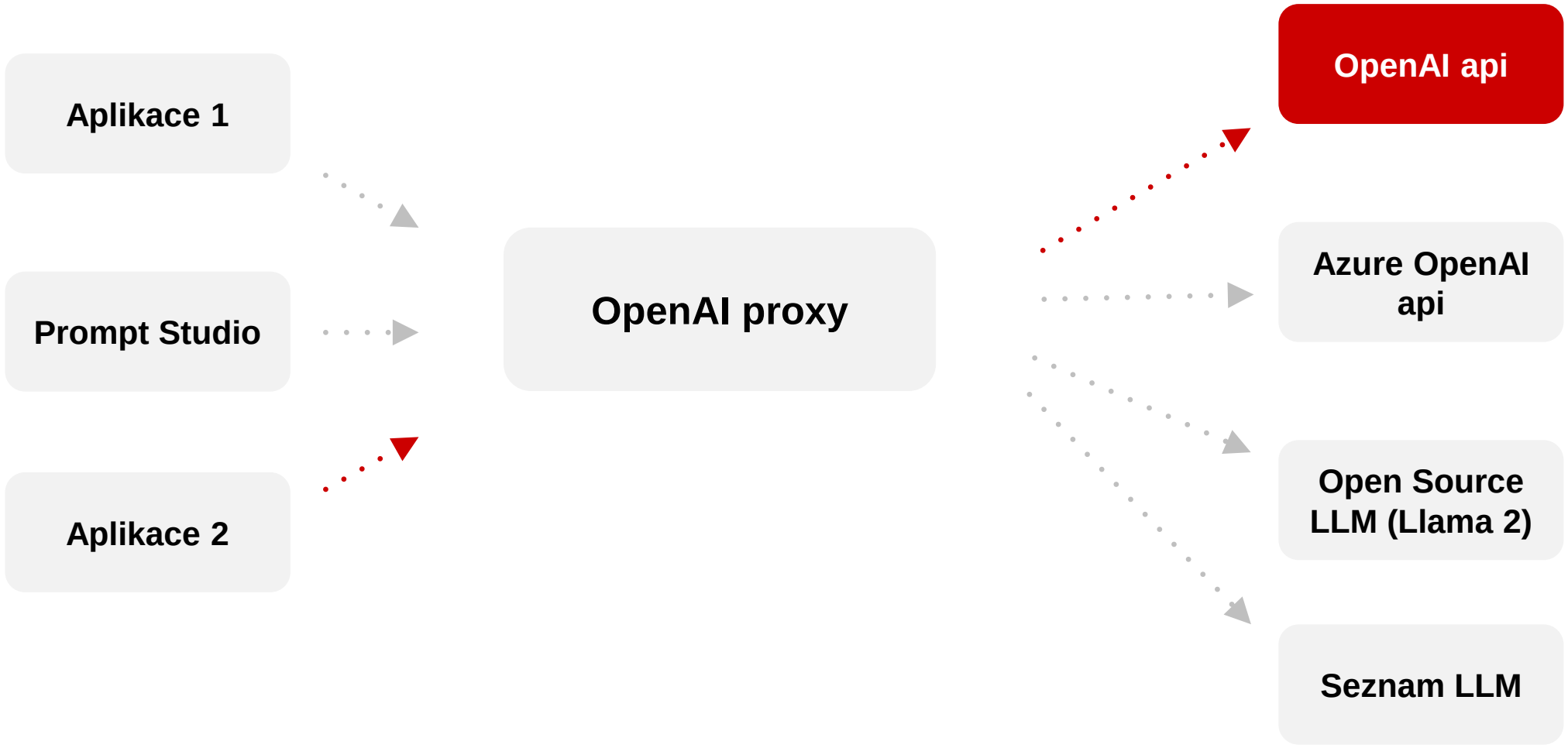
Modely dobré na úlohy co Seznam využije

Sbírat zpětnou vazbu od uživatelů

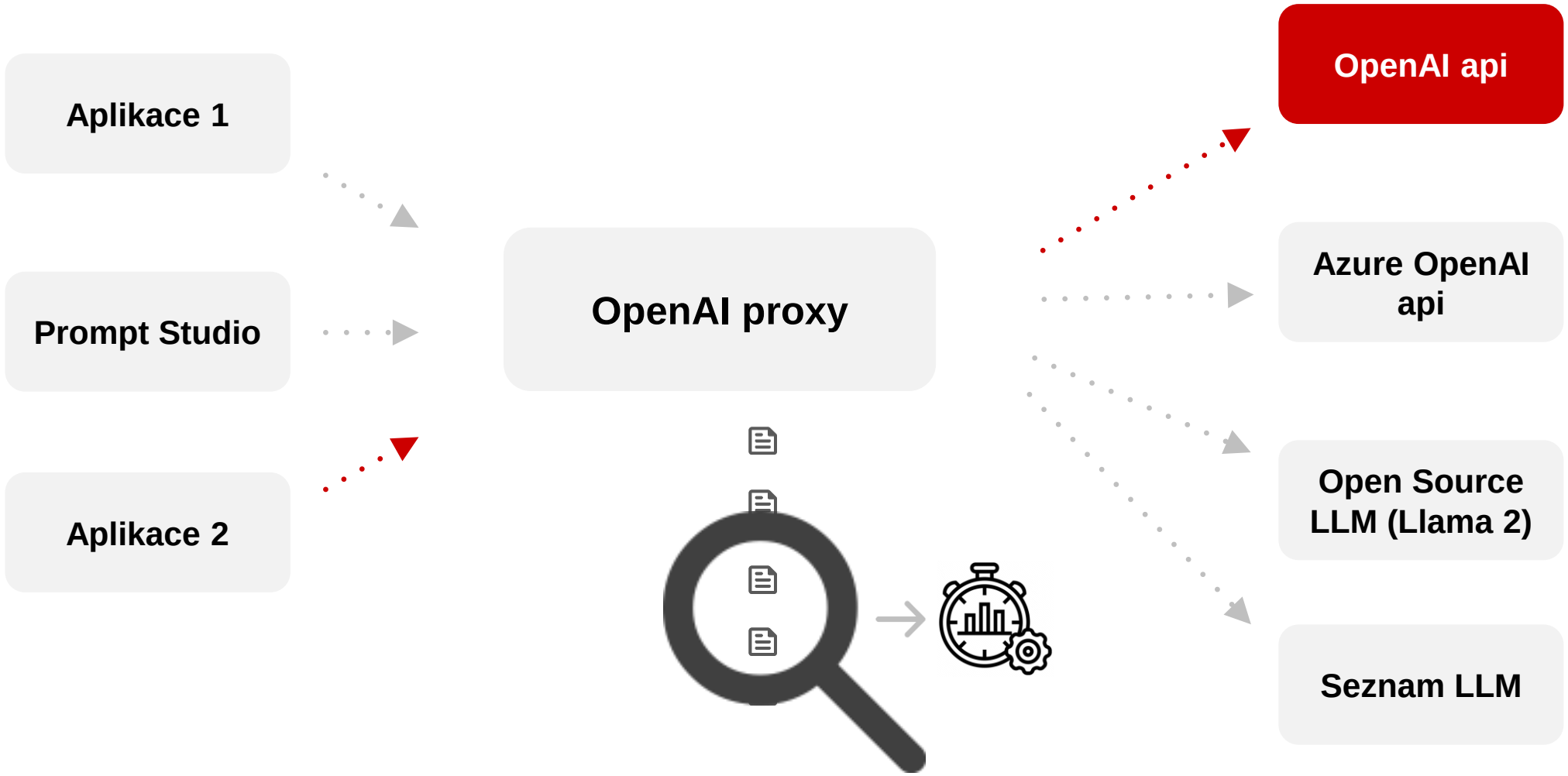
OpenAI proxy



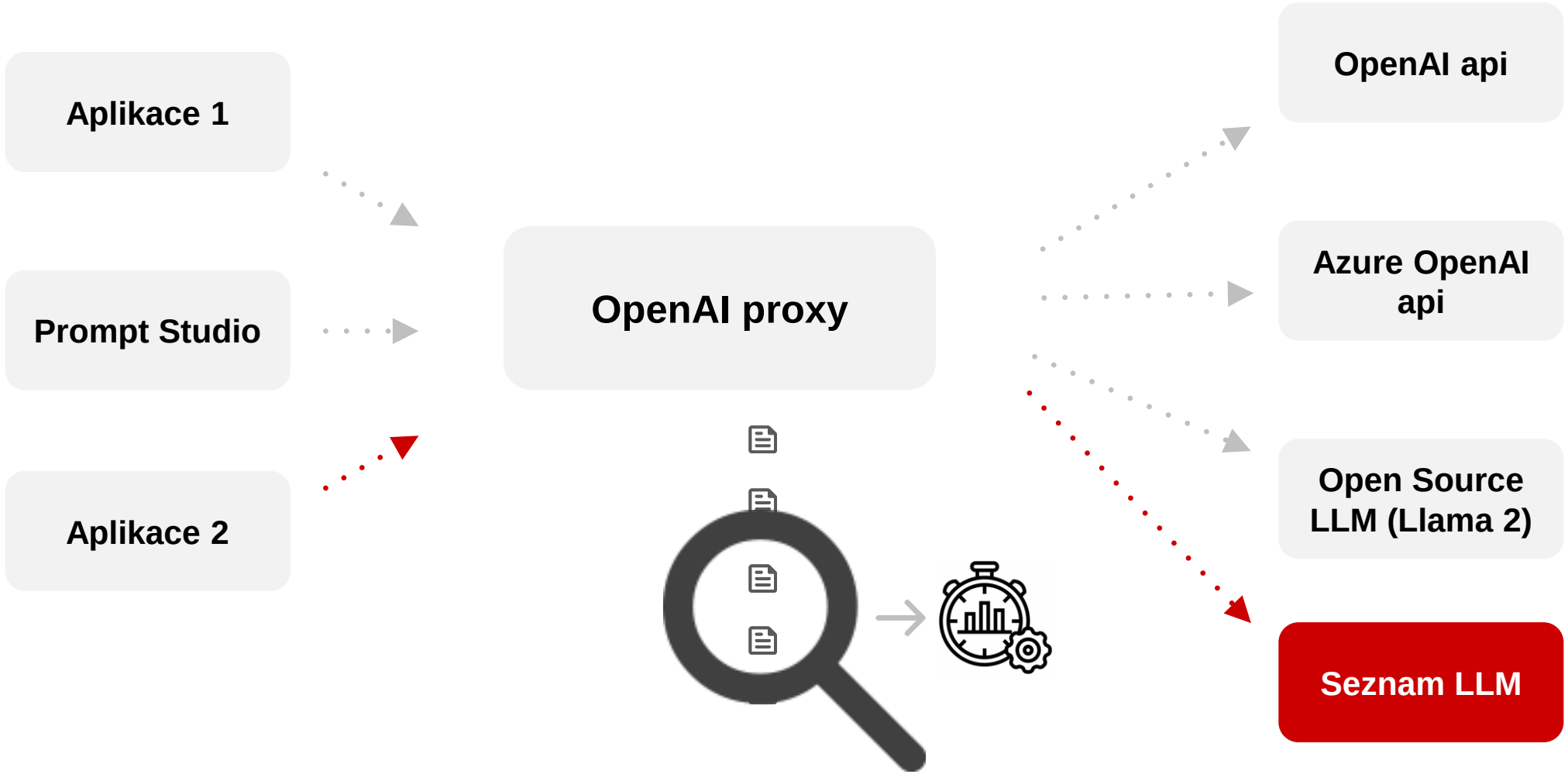
OpenAI proxy



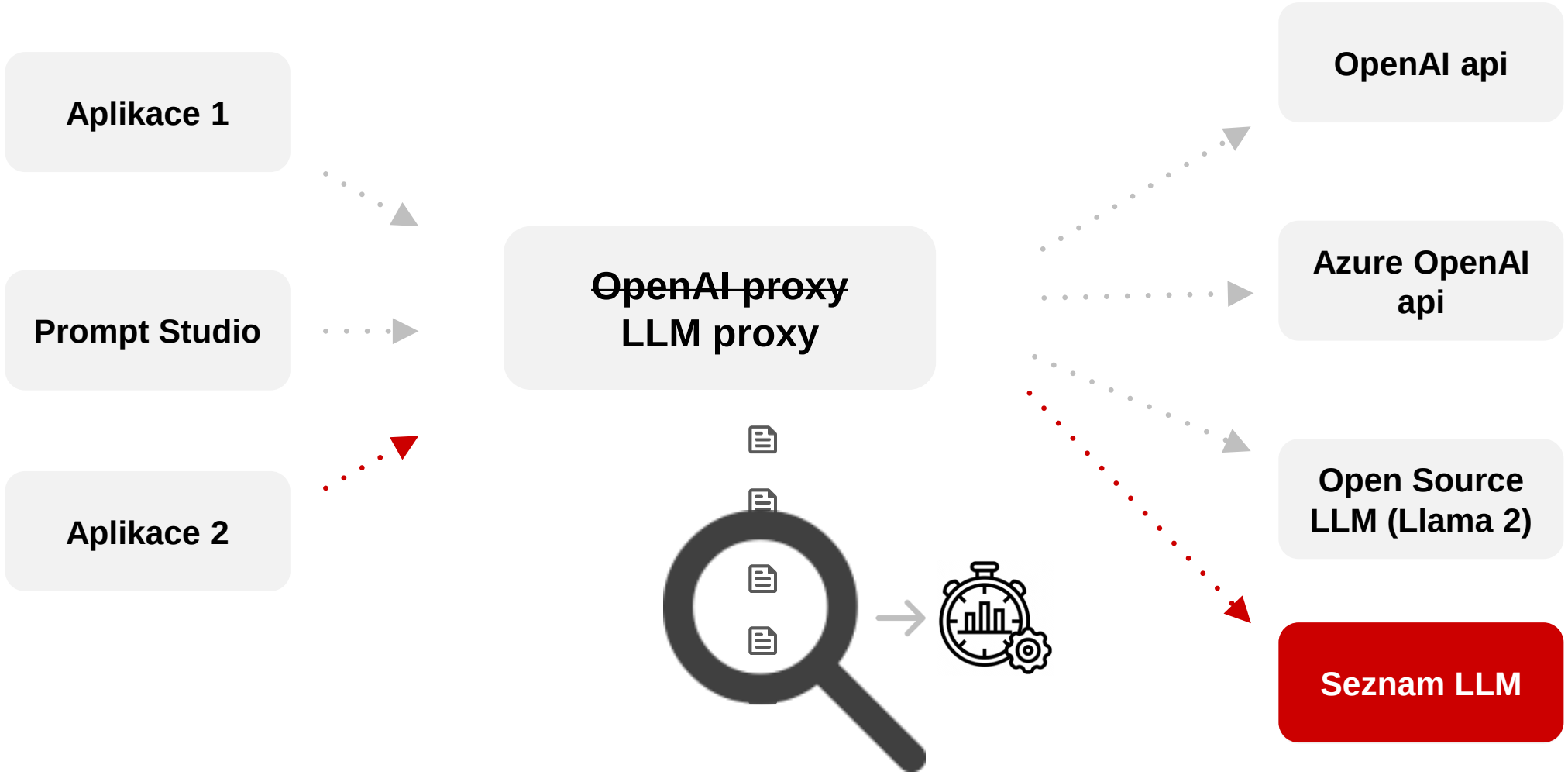
OpenAI proxy



OpenAI proxy



~~OpenAI proxy~~
LLM proxy

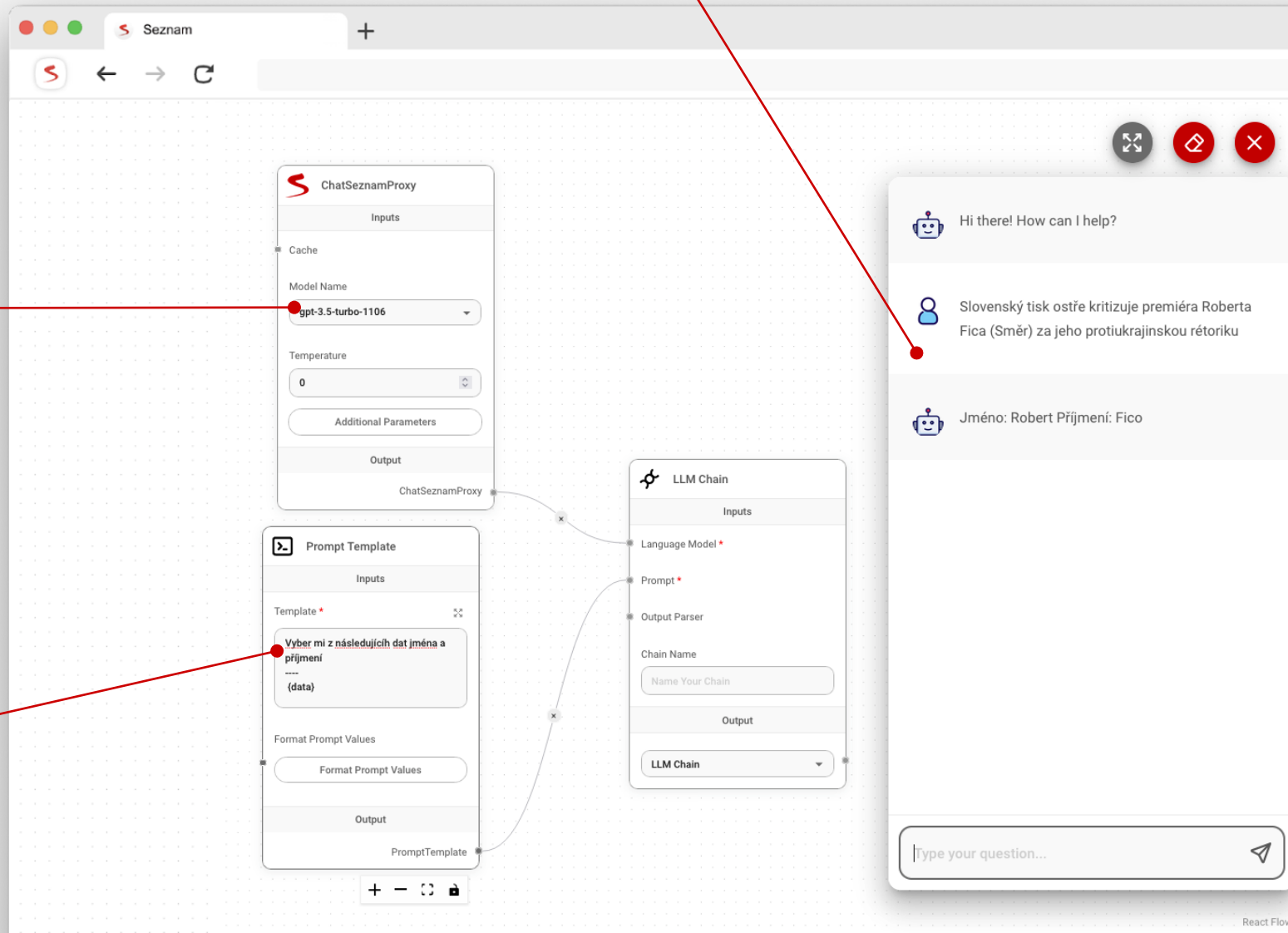


Prompt Studio

Vyzkoušet

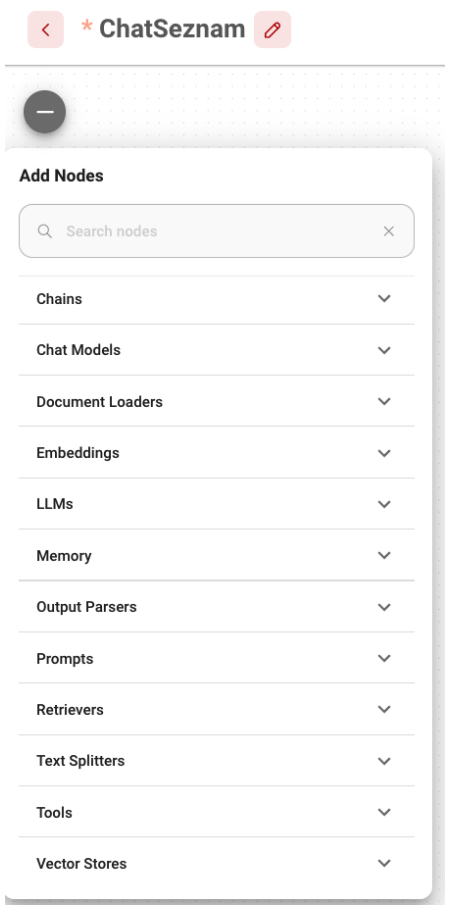
Vybrat model

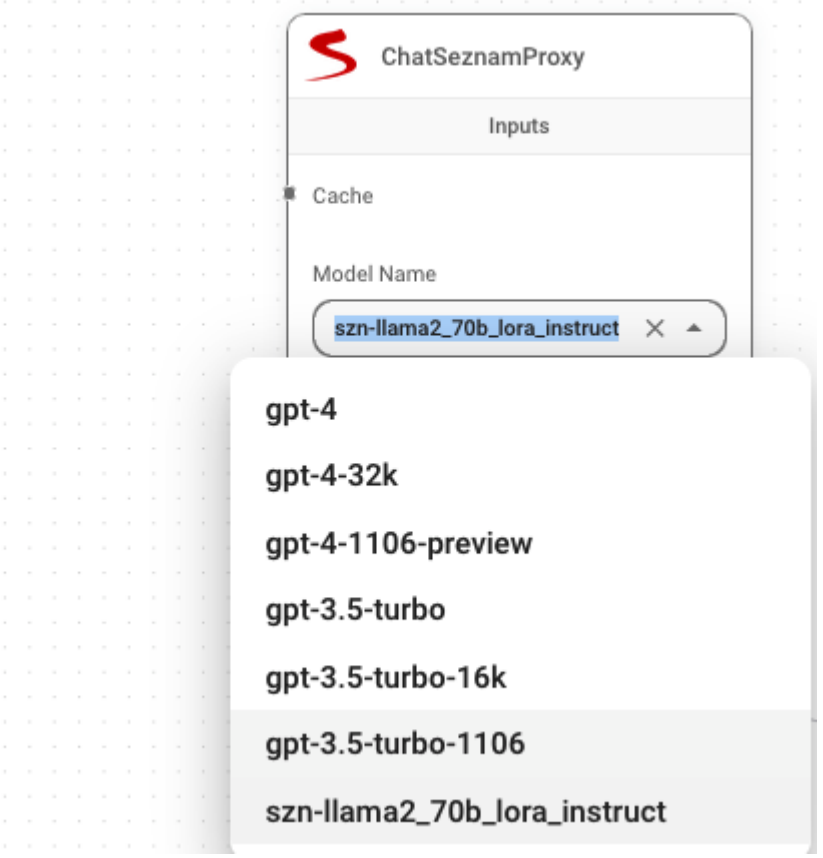
Vybrat prompt



Prompt Studio

- https://flowiseai.com/
- napojení interních API





Benchmark inferencí

	Naivní		Triton (s TensorRT-LLM)	
Model	token/s	H100s	token/s	H100s
Falcon 180B	3.1	6		
Falcon 40B_fp8	15.2	1		

Benchmark inferencí

	Naivní		Triton (s TensorRT-LLM)	
Model	token/s	H100s	token/s	H100s
Falcon 180B	3.1	6		
Falcon 40B_fp8	15.2	1		

Rychlost lidského čtení 8.3 token/s až 33.3 token/s

Benchmark inferencí

	Naivní		Triton (s TensorRT-LLM)	
Model	token/s	H100s	token/s	H100s
Falcon 180B	3.1	6	**492**	8
Falcon 40B_fp8	15.2	1	**700**	1

Rychlost lidského čtení 8.3 token/s až 33.3 token/s

Tensor RT LLM

- Quantization (BF16, FP8, GPTQ, AWQ)
- In-flight Batching - gpu utilization
- Page Attention - reduce memory wastes
- Graph Rewriting - optimize the underlying graph of neural network

Tensor RT LLM

- rychlejší

- složitější nasazení
- menší podpora modelů

https://github.com/NVIDIA/TensorRT-LLM

Tensor RT LLM

- rychlejší

- složitější nasazení
- menší podpora modelů

https://github.com/NVIDIA/TensorRT-LLM

vLLM

- snadné nasazení
- podpora OpenAI formátu
- rychlejší podpora nových modelů

- pomalejší

https://github.com/vllm-project/vllm

Benchmark inferencí Llama 7B

Translation task 1024 (max input and output)

GPU	token/s	zrychlení
H100s 80GB	1490	99.3x
L4 24 GB	25	1.66x
A4000 16GB	15	1x

Lessons learned

- těžší vytáhnout lidi ze zajetých kolejí
- context modelu je nový parametr
- inspiruj se (německý LLM https://laion.ai/blog/leo-lm/, mistral etc.)
- udělejte si vlastni evaluaci pro sve vlastni vyuziti

Shrnutí

- měříme naše způsoby použití
- stavíme ekosystém pro snadné použití pro naše zaměstnance
- nespoléháme na closed LLM
  - míříme na skvělou češtinu česky
  - zrychlujeme inferenci, zlevňujeme provoz

Shrnutí

- měříme naše způsoby použití
- stavíme ekosystém pro snadné použití pro naše zaměstnance
- nespoléháme na closed LLM
  - míříme na skvělou češtinu česky
  - zrychlujeme inferenci, zlevňujeme provoz