



# 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

S | Televize Seznam

SUPER.CZ

stream

SREALITY.CZ

SPORT.CZ

SMOTO.CZ

SKLIK.CZ

EMAIL

S DOVOLENÁ.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

S

# O Seznamu



# O Seznamu



# 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<sup>[1]</sup> a HuggingFace<sup>[2]</sup>

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<sup>[1]</sup> a HuggingFace<sup>[2]</sup>
  - 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<sup>[1]</sup> a HuggingFace<sup>[2]</sup>
  - 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

## Create your account

Email address

Continue

Already have an account? [Log in](#)



### Add payment method

Add your credit card details below. This card will be saved to your account and can be removed at any time.

#### Card information

#### Name on card

#### Billing address

Set as default payment method

Cancel

Add payment method



Security



Budgety pro týmy



Měření kvality



# OpenAI proxy



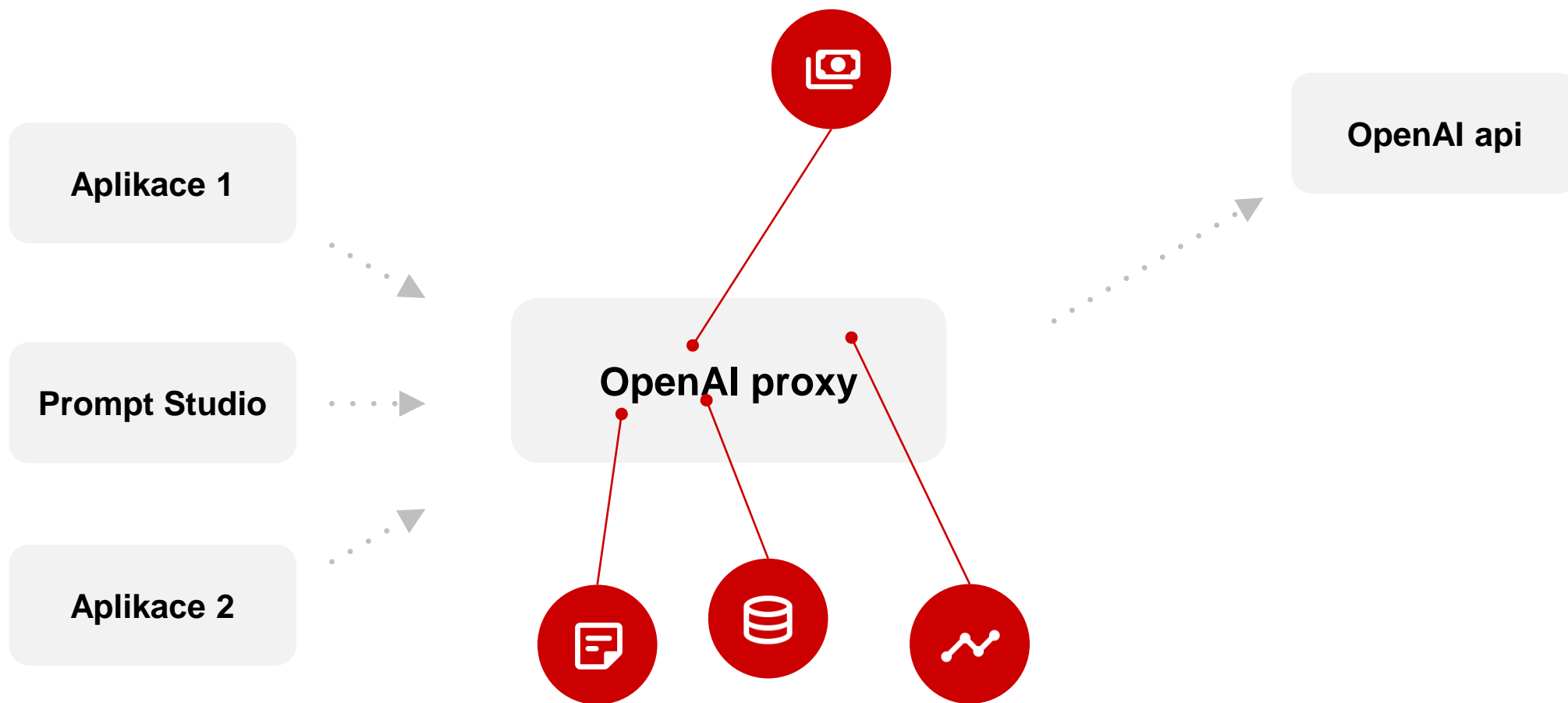
**Interní token**



**URL na naši proxy**



# OpenAI proxy



# OpenAI výpadek

OPENAI  
DEV  
DAY  
OPENAI  
DEV  
DAY



## 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.

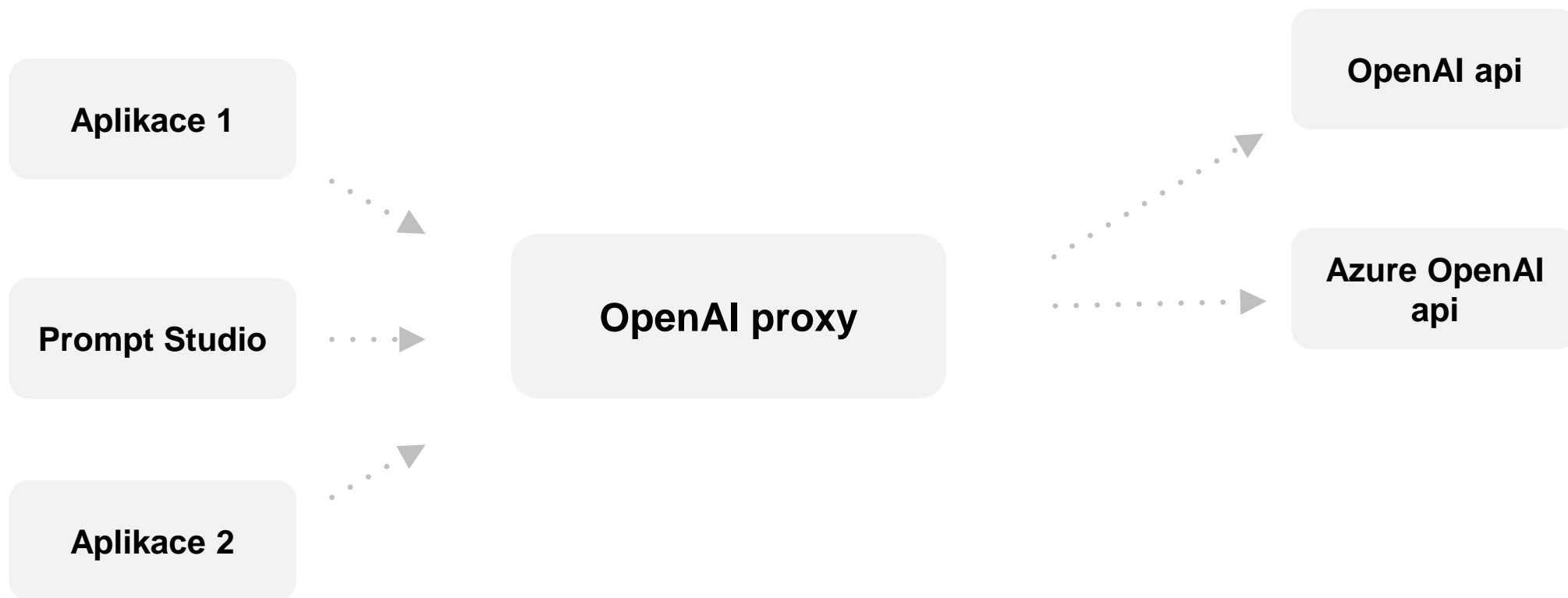
Nov 08, 2023 - 06:02 PST

**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**



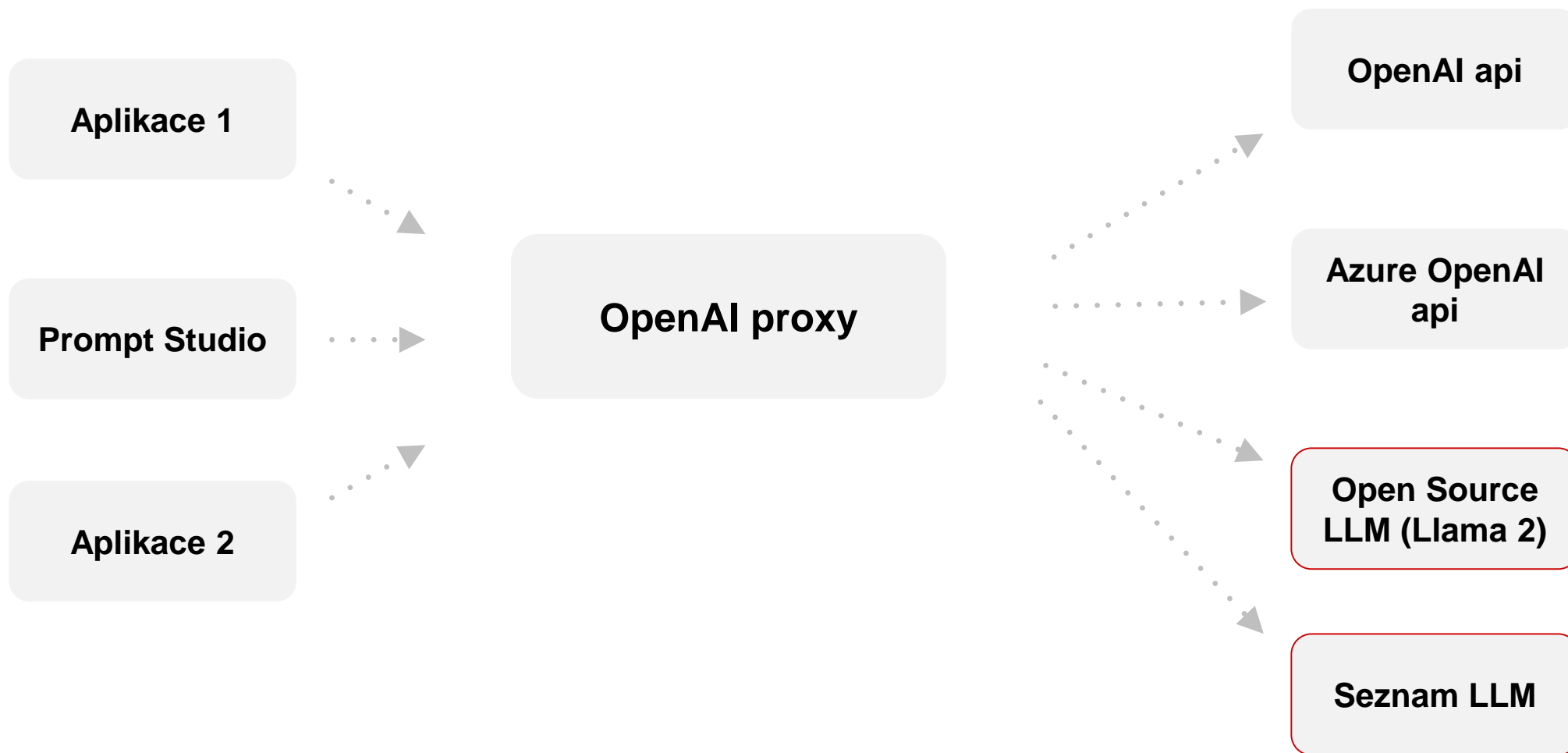
**Modely dobré na úlohy co Seznam využije**



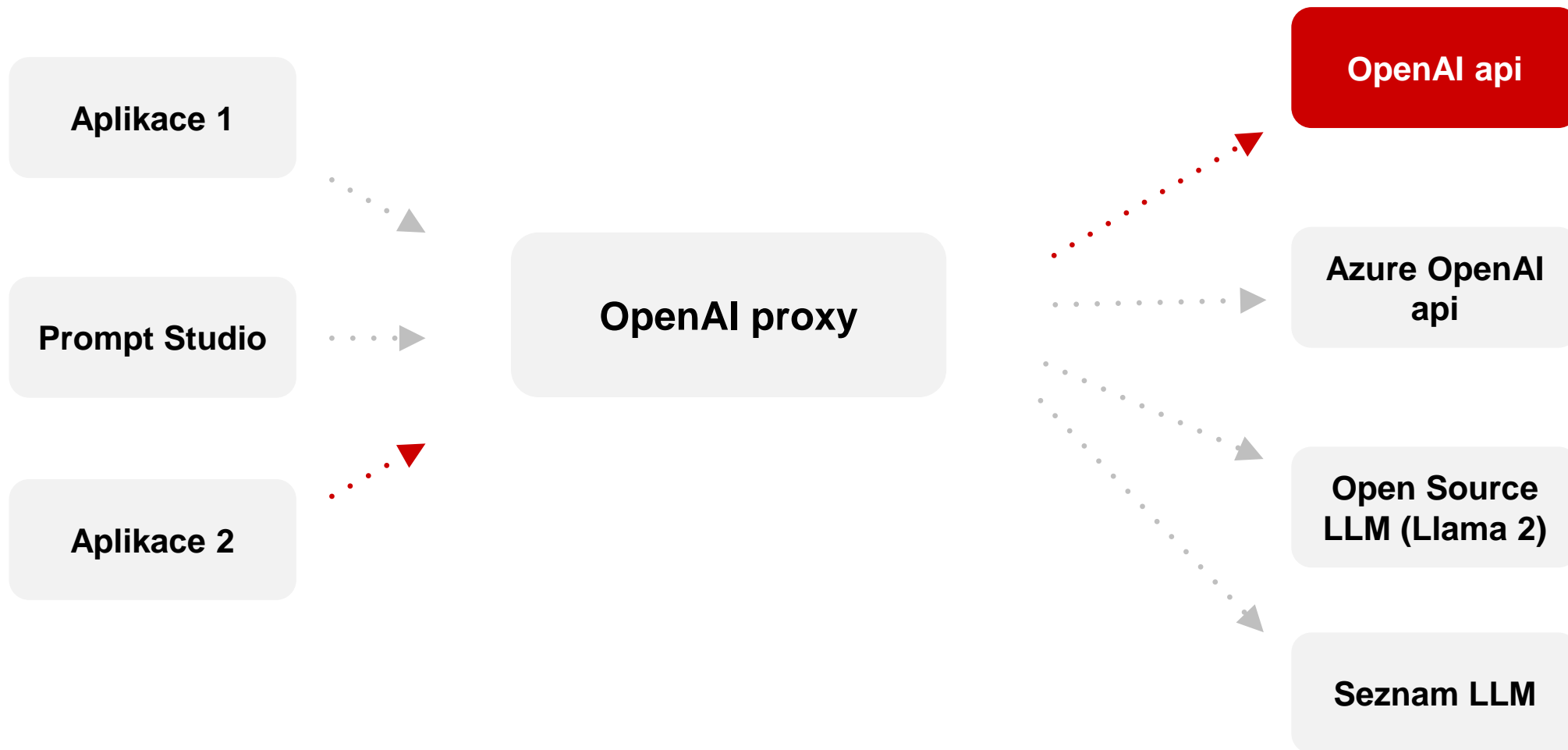
**Sbírat zpětnou vazbu od uživatelů**



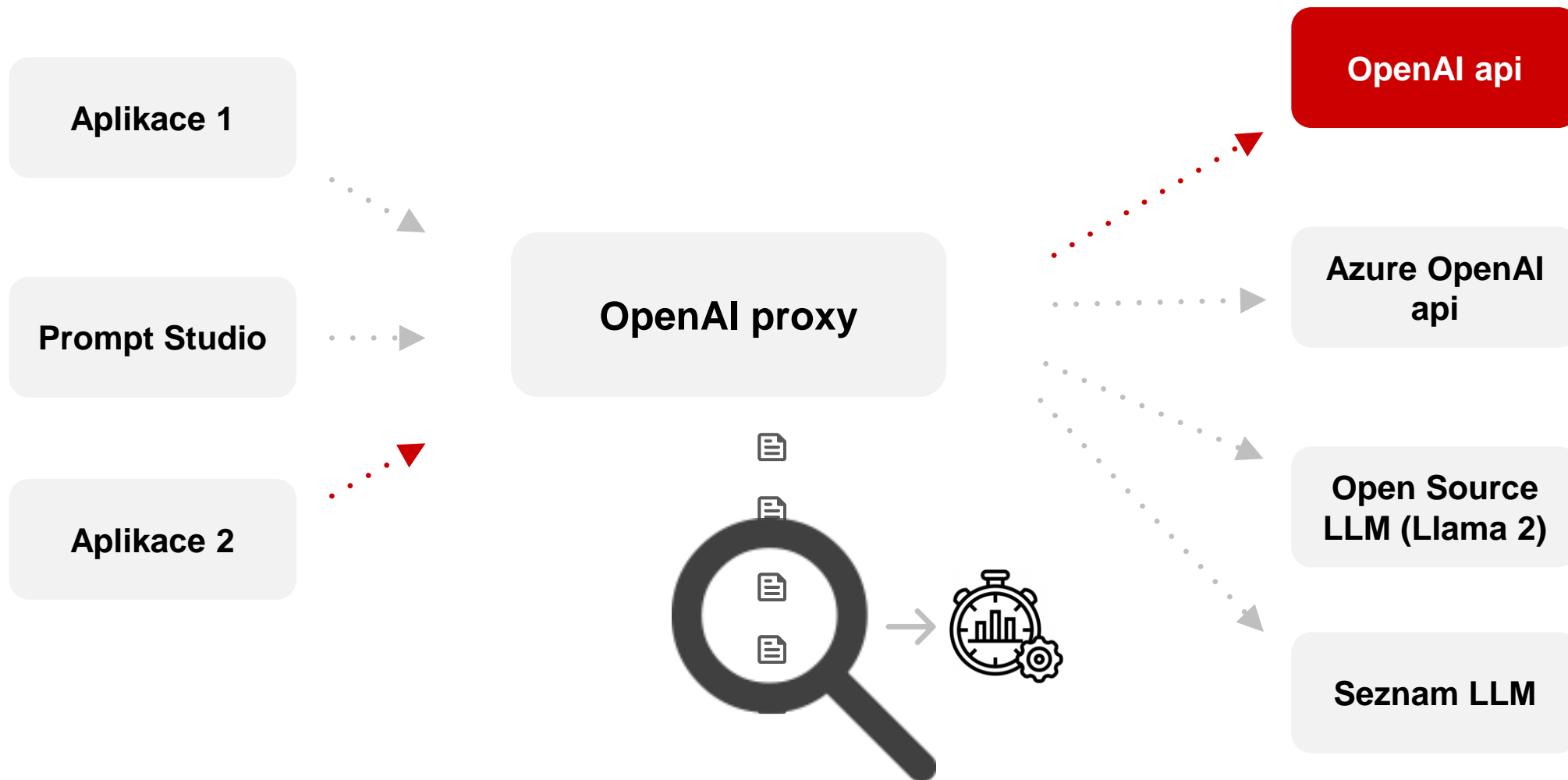
# OpenAI proxy



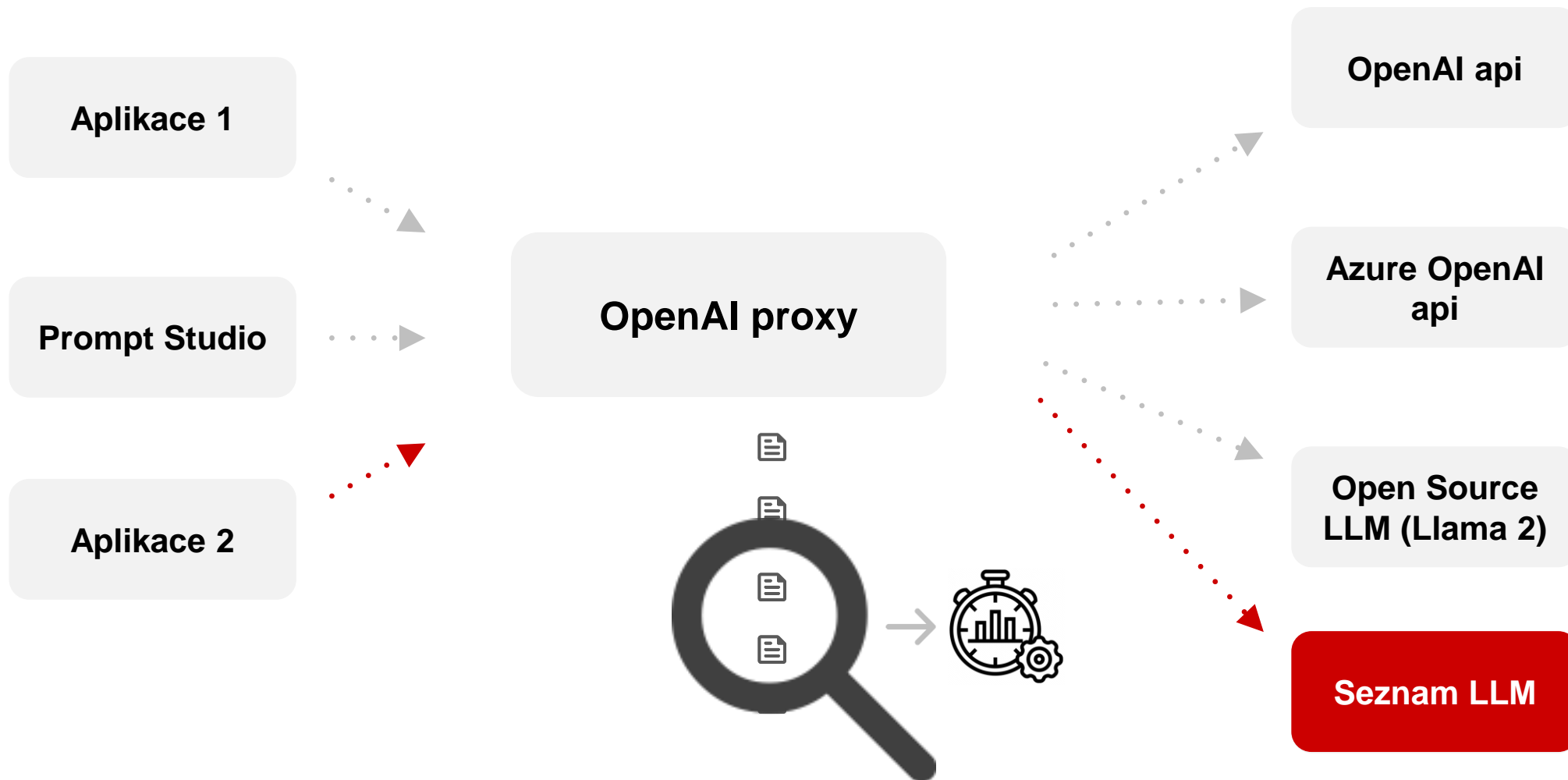
# OpenAI proxy



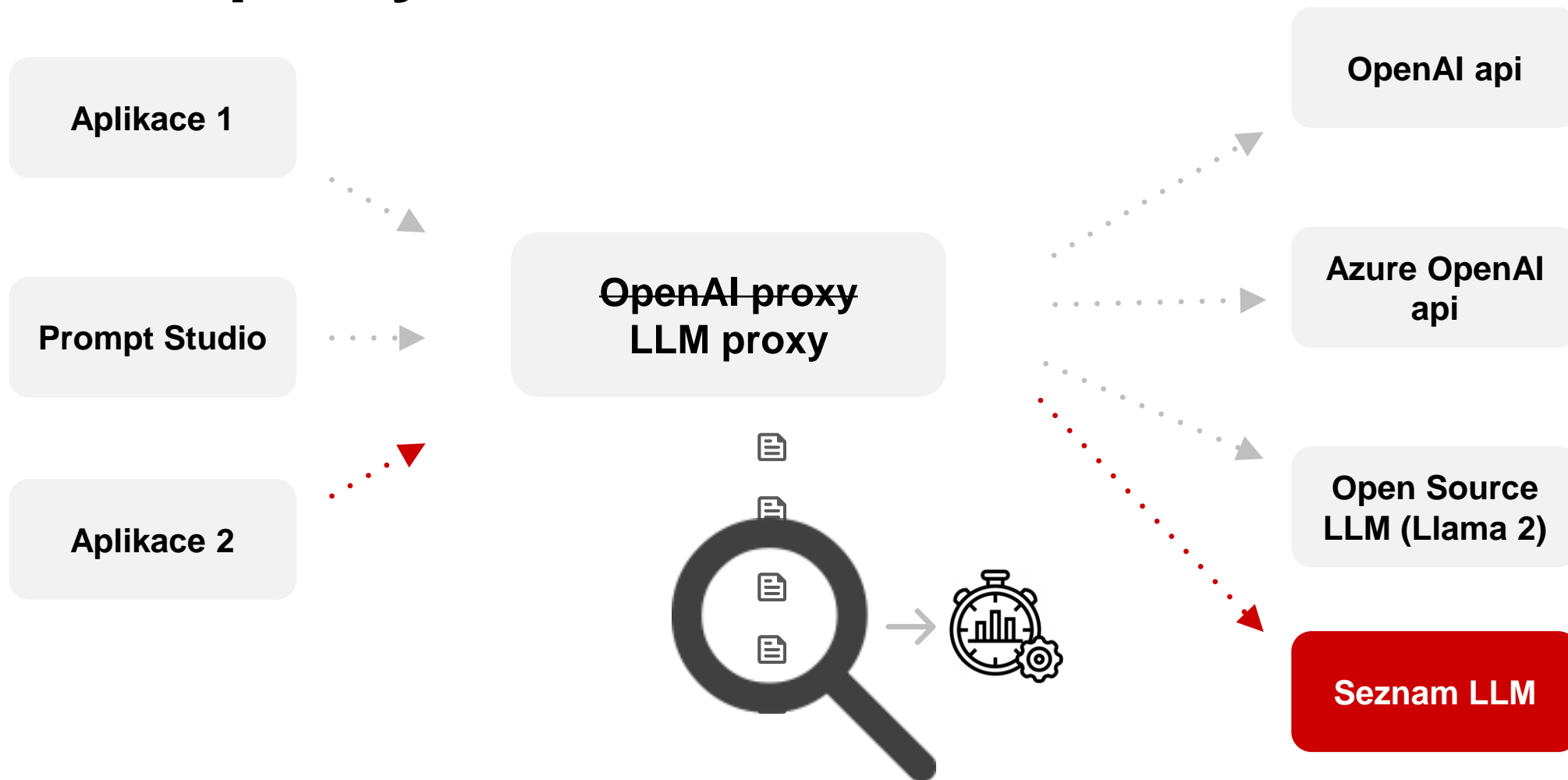
# OpenAI proxy



# OpenAI proxy



# ~~OpenAI proxy~~ LLM proxy





# Prompt Studio

Vyzkoušet

Vybrat model

Vybrat prompt

The screenshot displays the Prompt Studio interface within a browser window titled "Seznam". The main workspace contains a workflow with three nodes:

- ChatSeznamProxy**: A node with inputs for "Cache", "Model Name" (set to "gpt-3.5-turbo-1106"), "Temperature" (set to "0"), and "Additional Parameters". Its output is "ChatSeznamProxy".
- Prompt Template**: A node with an input for "Template" containing the text "Vyber mi z následujících dat jména a příjmení ---- {data}". It has a "Format Prompt Values" button and an output labeled "PromptTemplate".
- LLM Chain**: A node with inputs for "Language Model" and "Prompt", and an "Output Parser". Its "Chain Name" is "Name Your Chain" and its output is "LLM Chain".

Arrows indicate the flow of data from the Prompt Template node to the LLM Chain node, and from the ChatSeznamProxy node to the LLM Chain node. A red arrow points from the "gpt-3.5-turbo-1106" dropdown to the "Vybrat model" label. Another red arrow points from the "Vyber mi z následujících dat jména a příjmení" text to the "Vybrat prompt" label.

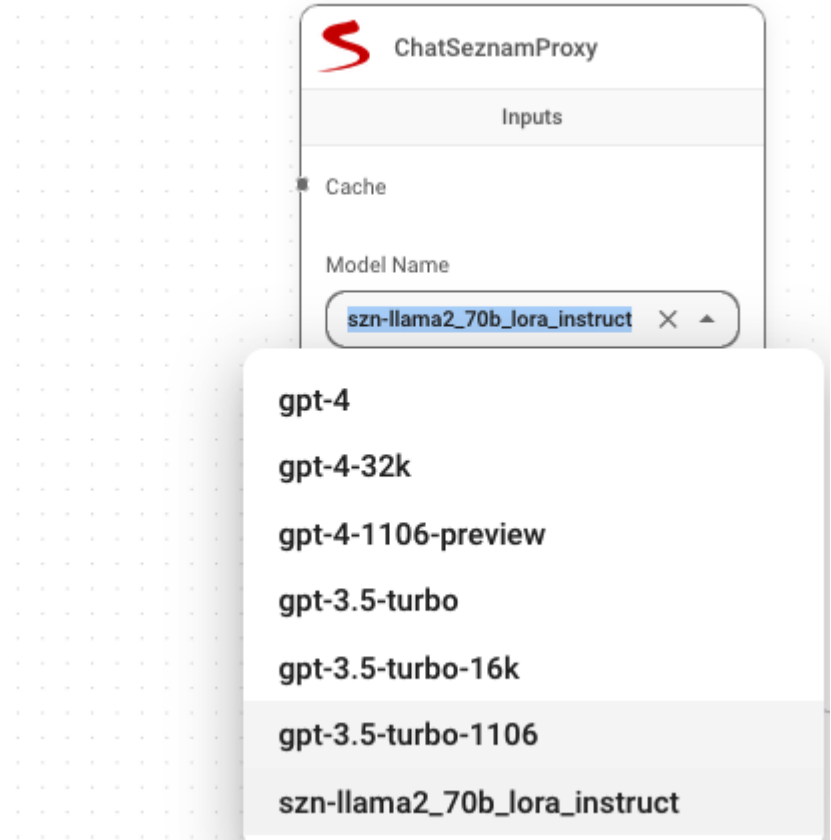
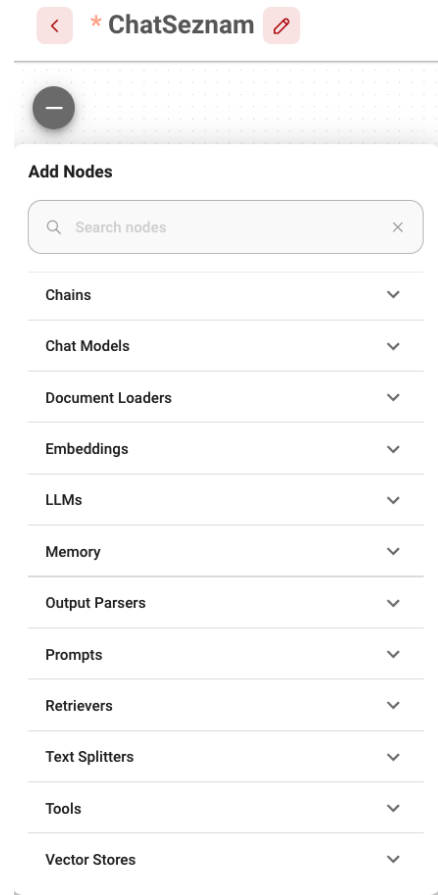
On the right side, there is a chat window with three messages:

- Robot icon: "Hi there! How can I help?"
- User icon: "Slovenský tisk ostře kritizuje premiéra Roberta Fica (Směr) za jeho protiukrajinskou rétoriku"
- Robot icon: "Jméno: Robert Příjmení: Fico"

At the bottom right, there is a text input field with the placeholder "Type your question..." and a send button.

# Prompt Studio

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



# Benchmark inferenci

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



# Benchmark inferencí

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

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



# Tensor RT LLM

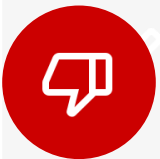
- Quantization (BF16, FP8, GPTQ, AWQ)
- In-flight Batching - gpu utilization
- Page Attention - reduce memory wastes
- Graph Rewriting - optimize the underlying graph of neural network



# Tensor RT LLM



- rychlejší



- složitější nasazení
- menší podpora modelů

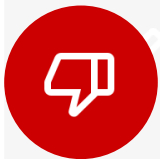
<https://github.com/NVIDIA/TensorRT-LLM>



# 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

GPU	Translation task 1024 (max input and output) 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 vlastní evaluaci pro své vlastní využití



# Shrnutí

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

