

šellma

Jak vypouštíme šelmy do produkce



Diana Hlaváčová

Product Manager Senior

Lupa.cz » Seznam.cz chystá vlastní umělou inteligenci. V češtině už je o něco lepší než GPT-3.5

Seznam.cz chystá vlastní umělou inteligenci. V češtině už je o něco lepší než GPT-3.5

JAN SEDLÁK | 17. 1. 2024 | Doba čtení: 4 minuty

9 NOVÝCH NÁZ

Lupa.cz » Jazykový model Seznamu se jmenuje Šelma, firma pracuje na zapojení AI do služeb

Jazykový model Seznamu se jmenuje Šelma, firma pracuje na zapojení AI do

Zprávy » Byznys » Byznys | Rozhovory » Vyhledávání na Seznamu se dramaticky změní, říká Pav...

Vyhledávání na Seznamu se dramaticky změní, říká Pavel Zima z vedení firmy

UMĚLÁ INTELIIGENCE – 07. 3. 2024 – 4 min čtení

Nejrychlejších sedm měsíců života, říká žena, která vede vývoj generativní umělé inteligence Seznamu

Seznam.cz si jako česká internetová jednička nechce nechat příležitost v umělé inteligenci ujít. Investuje do ní desítky milionů a učí ji česky.





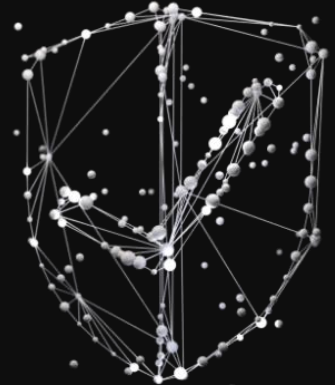
sellma



**Výborná čeština
modelu**



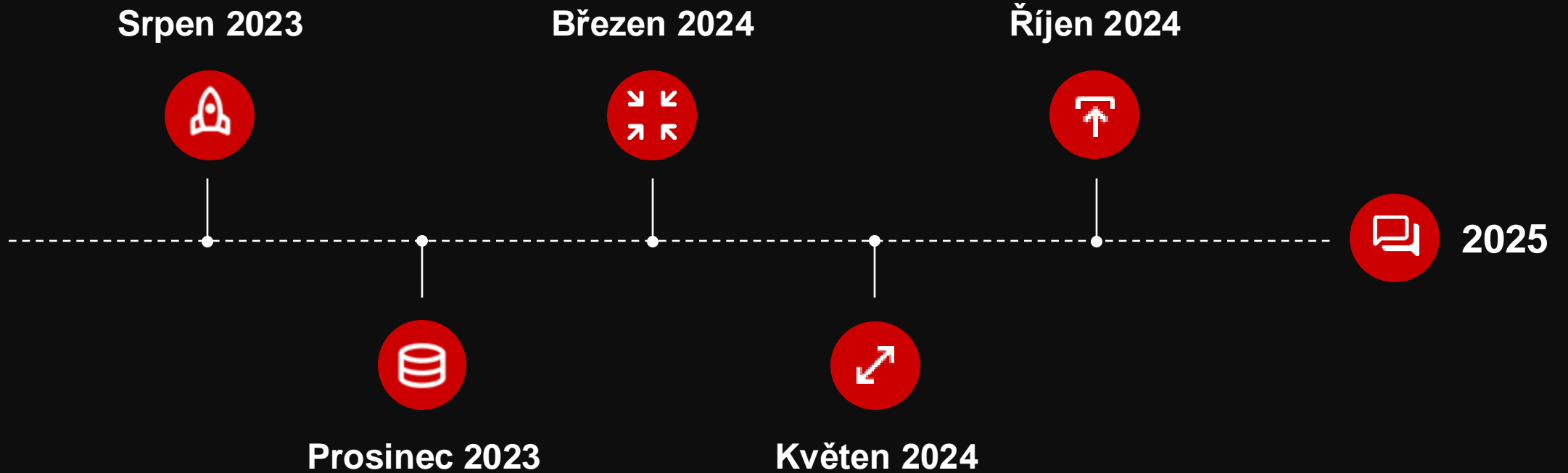
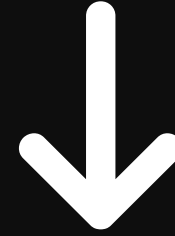
**Provoz v naší
režii**



**Bezpečnost
pro uživatele**



Jak se rodila SeLLMa?



Požadavky na produkční řešení

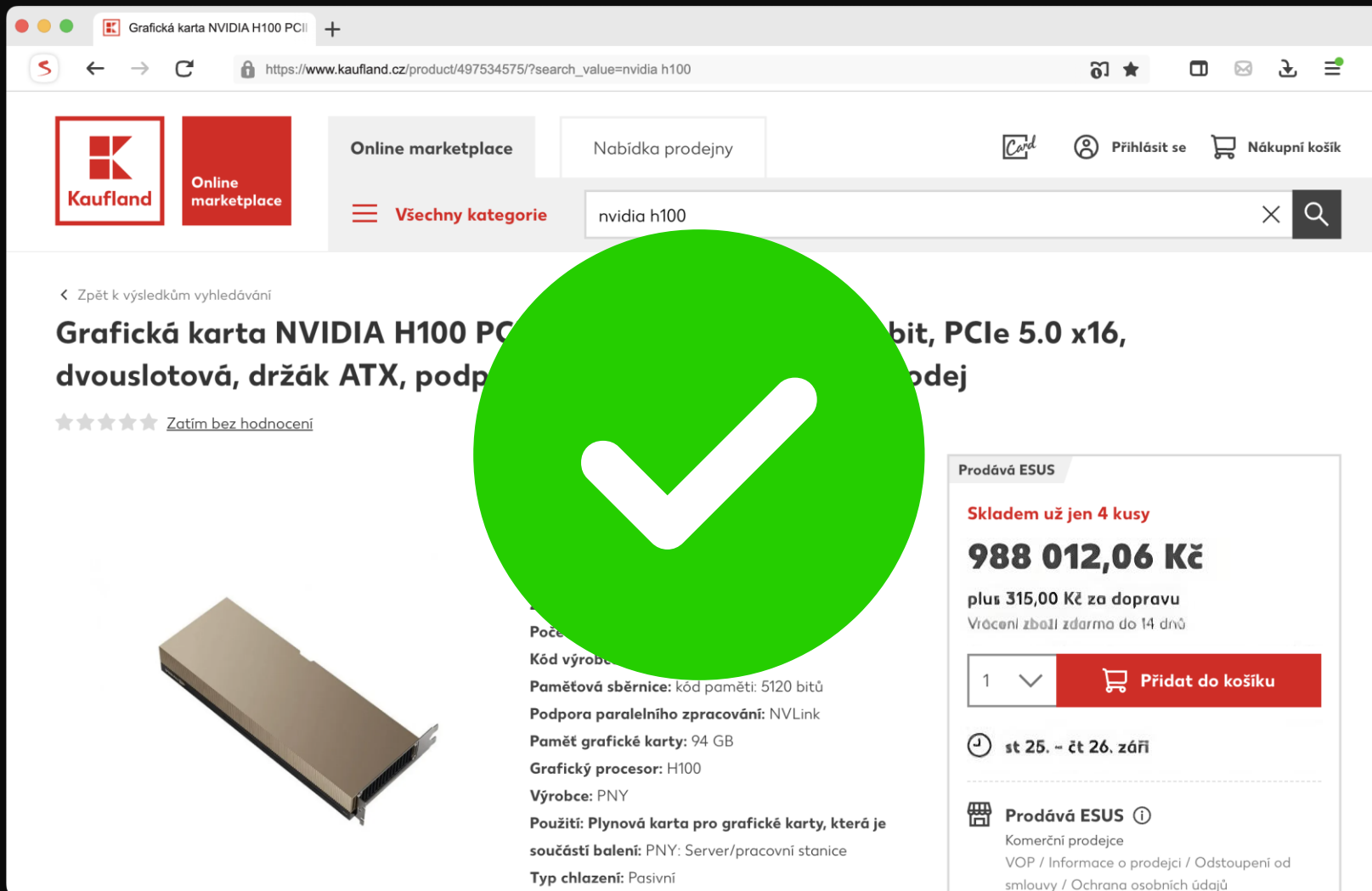
- Proměnlivý traffic během dne
- 10 až 20 modelů v provozu (aktuálně 5)
- Garance trafficu pro více uživatelů a aplikací
- Schopnost dodržovat určité latence
- Dosažení efektivního vytěžení clustru i mimo špičky



Předpoklady pro produkci



1



Grafická karta NVIDIA H100 PCII

https://www.kaufland.cz/product/497534575/?search_value=nvidia h100

Kaufland Online marketplace


Online marketplace Nabídka prodejny

Všechny kategorie nvidia h100

Zpět k výsledkům vyhledávání

Grafická karta NVIDIA H100 PCII 100W, 100 GB, 100 bit, PCIe 5.0 x16, dvouslotová, držák ATX, podpora NVLink

★★★★★ [Zatím bez hodnocení](#)



Proává ESUS

Skladem už jen 4 kusy

988 012,06 Kč

plus 315,00 Kč za dopravu
Vrácení zboží zdarma do 14 dnů

1

st 25. - čt 26. září

Proává ESUS ⓘ
Komerční prodejce
VOP / Informace o prodeji / Odstoupení od smlouvy / Ochrana osobních údajů



2

LLM proxy



Aplikace 1

Prompt Studio

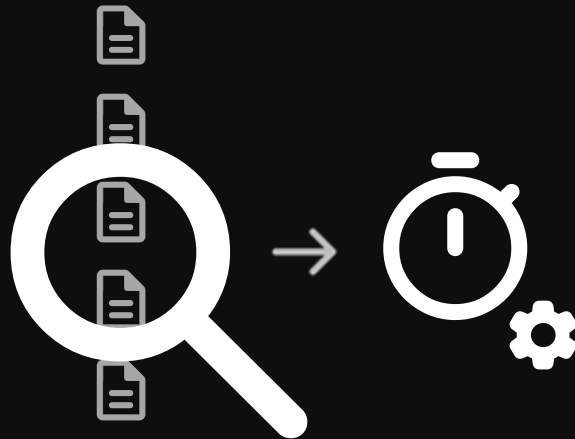
Aplikace 2

LLM proxy

Komerční api

Open Source LLM

Seznam LLM



3

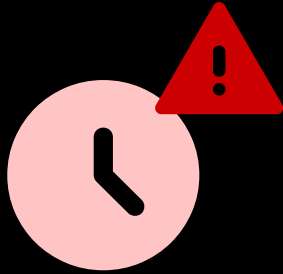
Víme, jaké aplikace nasazujeme



- Jaký model je využíváný
- Kolik tokenů je v průměru na vstupu a výstupu
 - Do velikosti vstupu se započítává např. i historie chatu atd.
- Počet requestů ve špičce
- Jaké jsou očekávané latence
- Další 15 detailů



Aplikace z pohledu rychlosti



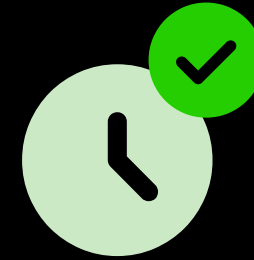
Nečeká/Live chat

Chci odpověď co nejdříve.
Maximálně 10 sekund + streaming.



Může chvíli čekat

Chci odpověď co nejdříve,
pravidelně přepočítávám, latence
do 1–5 min nevadí.



Čas není problém

Mám hodně requestů,
ale odpověď počká.



Jak je (pro teď) vnímaná naše produkce?



Online

Balancování mezi dobou odpovědí
a celkovou propustností systému.



Offline

Nejjednodušší: zaměření na
maximální propustnost.



Co musíme sledovat?



Online

Balancování mezi dobou odpovědí a celkovou propustností systému.



Offline (Batch)

Nejjednodušší: zaměření na maximální propustnost.

Streaming

metriky:

- Time to first token
- Inter token latency

Sequential

metriky:

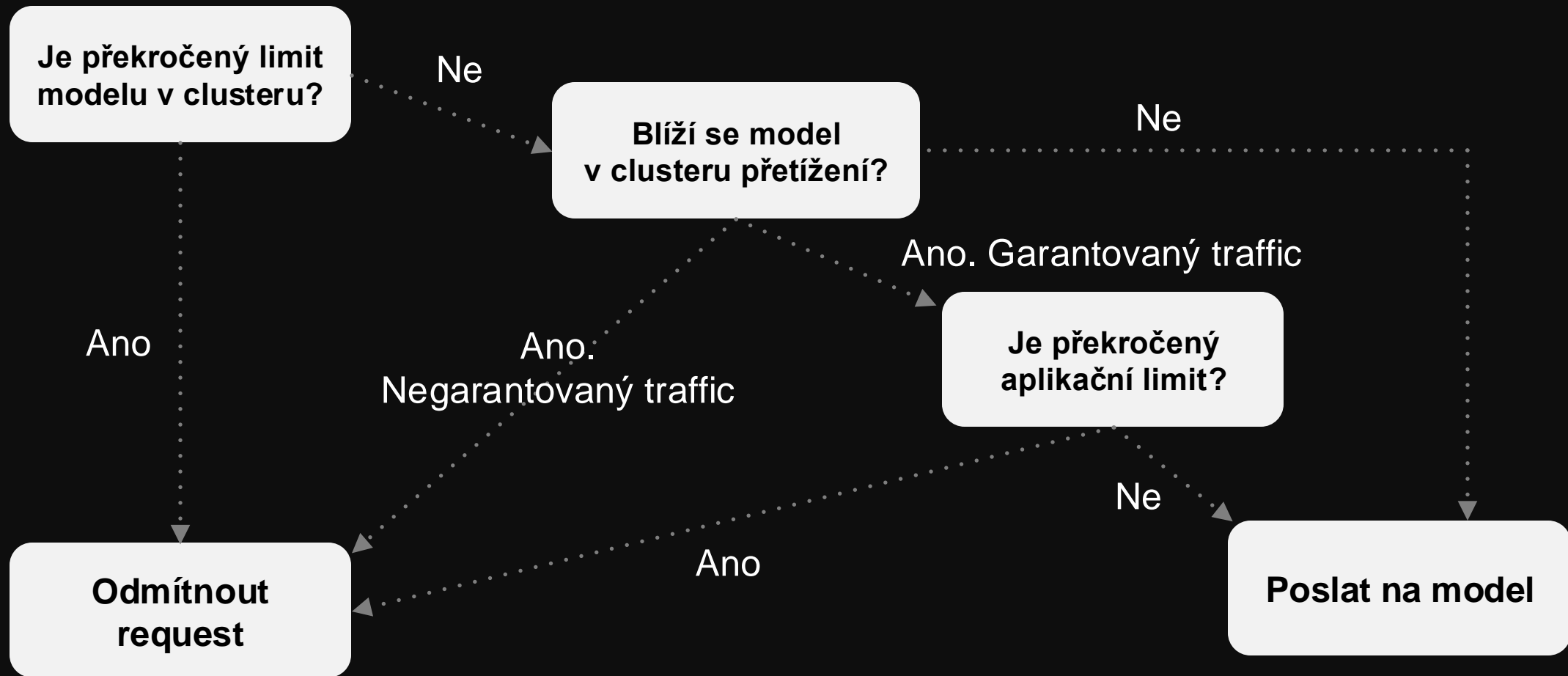
- End to end latency (request latency)

metriky:

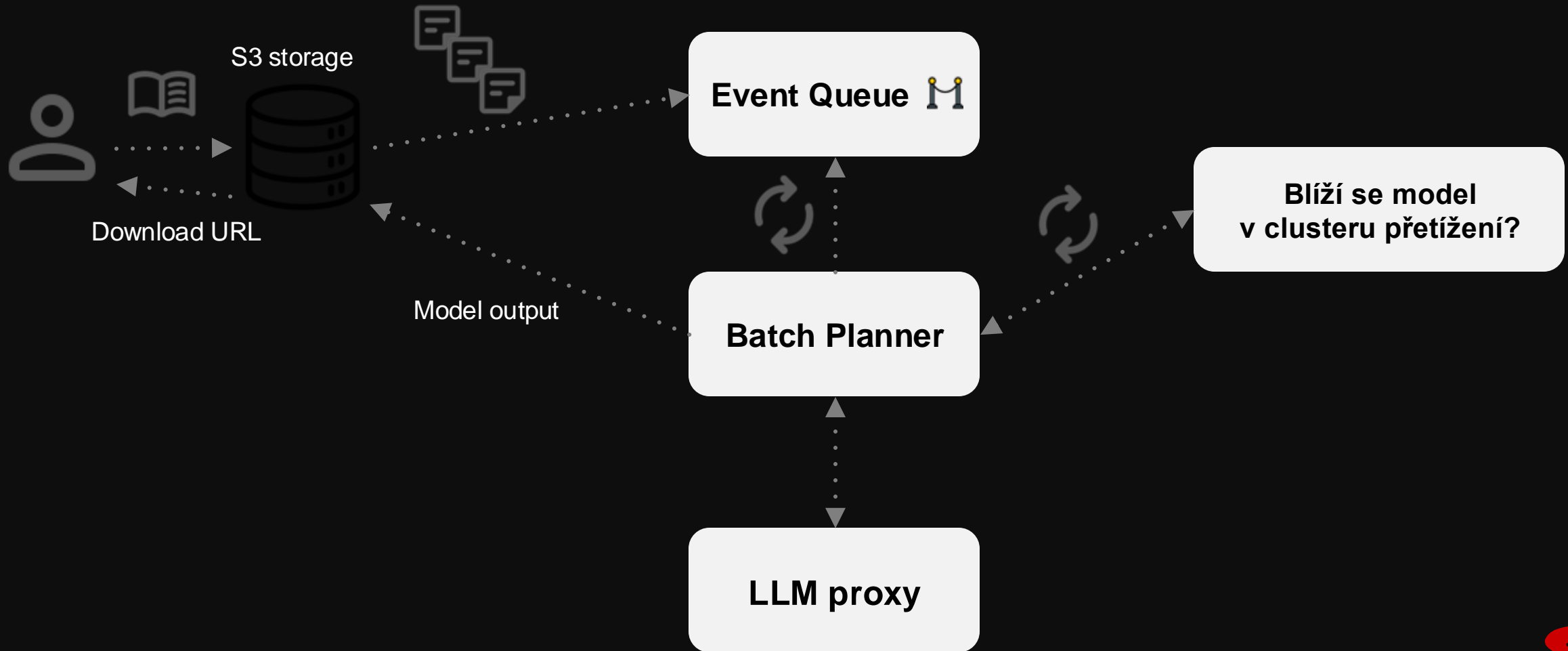
- request_throughput
- output_token_throughput



Garance pro provozní aplikace



Offline (batch) inference



Výběr SeLLMa modelu

The screenshot shows the 'Upload' section of the application. At the top, there are three tabs: 'Upload' (active), 'In progress', and 'Results'. Below the tabs, there is a 'Select a Model' dropdown menu with 'sellma-70b-240413-preview' selected. Underneath, there is an 'Upload a file' section with a 'Drag and drop file here' area (limit 200MB per file, JSONL) and a 'Browse files' button. A 'Submit' button is located at the bottom left.

Vložit soubor s requesty

The screenshot shows the 'Results' section of the application. At the top, there are three tabs: 'Upload', 'In progress', and 'Results' (active). Below the tabs, there is a 'Files in S3 Bucket' section with a 'Prefix filter path' input field. Underneath, there is a 'Results' section with a list of files and their corresponding 'Download' buttons. The files listed are: 'sellma-70b-240413-preview/example-file.jsonl-output', 'sellma-mistral-7b-240117/example-file.jsonl-output', 'sellma-mistral-7b-240328-dws/example-file.jsonl-output', 'sellma-mistral-7b-240328-dws/test-document-output', and 'sellma-mistral-7b-240328-dws/test2-output'. Below this, there is an 'Uploaded input request files' section with a list of files and their corresponding 'Download' buttons. The files listed are: 'sellma-70b-240413-preview/example-file.jsonl', 'sellma-mistral-7b-240117/example-file.jsonl', 'sellma-mistral-7b-240328-dws/example-file.jsonl', and 'sellma-mistral-7b-240328-dws/test-document'.

Po dokončení možné stáhnout výsledky

The screenshot shows the 'Status of uploaded batches' section. At the top, there are three tabs: 'Upload', 'In progress' (active), and 'Results'. Below the tabs, there is a table with the following data:

| batch_id | created_at | updated_at | status | status_message | valid_lines |
|----------|------------|---------------------|---------------------|----------------|-------------|
| 0 | 86 | 2024-09-18T16:53:48 | 2024-09-18T16:54:02 | COMPLETED | 10 |

Po nahrání souboru můžu sledovat stav

Vstupní nahrané soubory

Powered by
sellma

proč polníček

Internet Obrázky Zboží Mapy Video Zprávy Firmy Slovník

- Polníček je **neskutečně výživný** a chutí byste si ho snadno spletli s hlávkovým salátem.¹
- Jelikož ho lze **snadno pěstovat** a má významné nutriční vlastnosti, rozhodně byste ho neměli přehlížet, protože může ozdobit i ochutit vaše pokrmy.¹
- Polníček obsahuje **vitamin C, E, B, provitamin A, kyselinu listovou, hořčík a železo**, což z něj činí cenný zdroj živin.²
- V období vegetačního klidu nám může **poskytovat tolik důležité vitamíny**, což je jeho výjimečnost.²
- Polníček je **jednoduchý důkaz** toho, že i v zimě si můžeme doma nachystat čerstvý a velice chutný salát.²

Polníček: Jak jej pěstovat a čím je...
kupi.cz 1

Polníček – Sezónní potraviny
sezonka.cz 2

Polníček: Jak jej pěstovat
kupi.cz

Před 14 dny · Věc polní? Kde se vzal? Článek!



Pastrňák svatba

Internet Obrázky Zboží Mapy Video Zprávy Firmy Slovník

Zprávy na téma: Pastrňák svatba

Český hokejový útočník David Pastrňák se oženil se švédskou partnerkou Rebeccou Rohlssonovou na chorvatském ostrově Hvar, kde vychovávají roční dceru Freyu Ivy.^{1, 2, 3} Pastrňák se na Instagramu pochlubil fotkou ze svatby, která byla třídenním luxusním veselím, na němž nešetřil a které stálo miliony.^{4, 5}

Tento text je vygenerovaným shrnutím tématu

Adele

Adele – Wikipedie
cs.wikipedia.org/wiki/adele

Adele Laurie Blue Adkins, MBE (* 5. května 1988 Londýn, Spojené království) je anglická zpěvačka a skladatelka známá jen jako Adele.

Kariéra Soukromý život Diskografie Turné Odkazy

Mohlo by vás zajímat

Jaké je celé jméno zpěvačky Adele?

Jaké je de...

Proč se hledá Nehoda na Rakovnicku?

...ošlo u Lubné na Rakovnicku k tragické nehodě, při které řidička sjela ze silnice a převrátila auto na střechu.^{1, 2, 3} Řidička na místě... dvě děti, které s ní cestovaly, byly zraněny a převezeny do...^{4, 5}

2. idnes.cz 3. extra.cz 4. iprima.cz 5. blesk.cz

erovaným shrnutím tématu



Shrnutí aktuálního dění pro nejhledanější dotazy

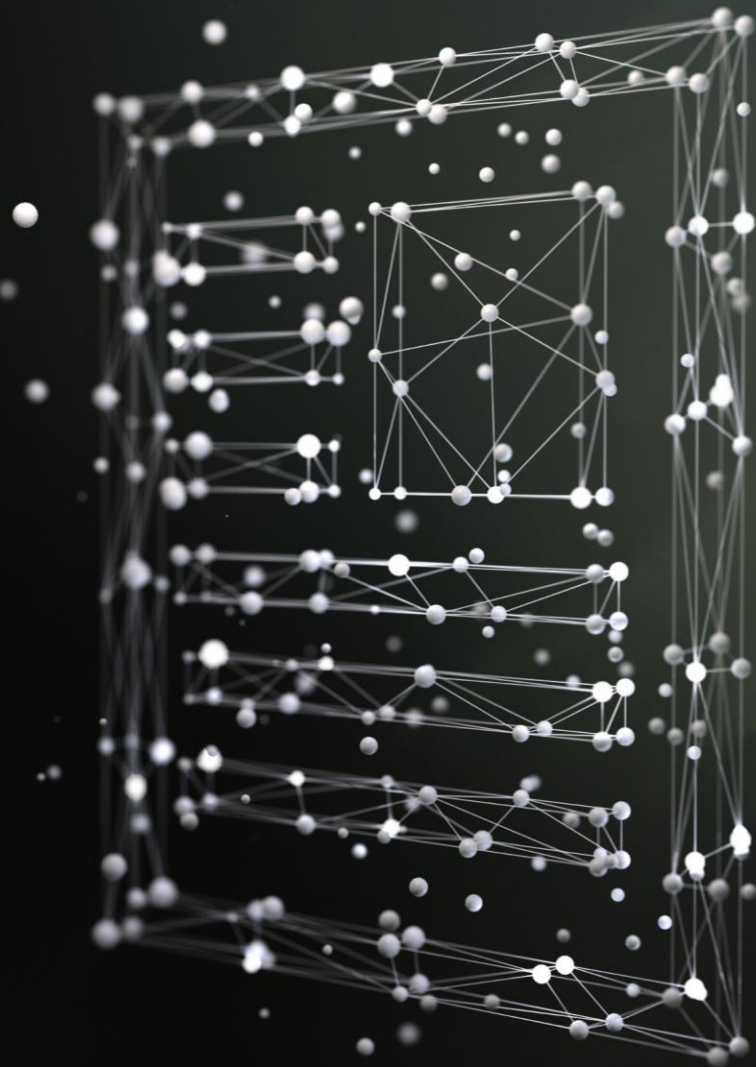
The screenshot shows the Seznam.cz homepage. At the top, there is a search bar with the text "...najdu tam, co neznám" and a "Vyhledat" button. Below the search bar, there are several trending topics: "Ptáček psycholog", "Bomba ve školách", "Princezna Kate", and "Bouřky". The main content area features a news article titled "Ukrajinská vláda se otfásá. Odejde nejméně polovina členů, včetně Kuleby" with a sub-header "Shrnutí". A red box highlights the "Shrnutí" link. To the right of the article, there is a "Krátké shrnutí" section with the text "Generováno AI" and a "Celý článek" button. Below this, there are "Podobné články" with thumbnails and titles like "České noviny Lipavský i Fiala po útoku v Poltavě obhajují dodávky pomoci Ukrajině", "Médium Tři scénáře k porážce Ruska", and "Hlídací pes Slovensko chce svůj díl z čínského koláče: „Copak Evropa neporoučuje lidská práva?“". On the right side of the page, there is an "Email" section with "Firmní e-mail zdarma" and a list of "Doručené (16)" emails. At the bottom, there is a "Služby" section with icons for "Mapy", "Bazar", "Reality", "Autá", "Slovník", and "TV program".

Sumarizace článků

Nadpisy zahraničních článků v češtině



Co ještě chystáme?



Nový zážitek z inzerce

SREALITY.CZ Seznam.cz

AGENTURA ZVONEK
sreality-test-client-admin 340 inzerátů

Nový inzerát

| | | | |
|------------------------------|--------------------------|-----------------------------------|---------------------------------|
| 1 topovaný inzerát | 0 tipů makléře | 6 běžících tipů regionu | 23 končících inzerátů |
|------------------------------|--------------------------|-----------------------------------|---------------------------------|

Statistiky za posledních 14 dní

| | | | |
|-----------------------------------|--|------------------------------------|-----------------------------------|
| 17593 zobrazení detailů | 2 přidání do uložených hledání | 10 přidání do oblíbených | 10 odpovědí na inzeráty |
|-----------------------------------|--|------------------------------------|-----------------------------------|

Kredit

| | |
|------------------------------------|-----------------------------|
| 98 881,71 Kč v peněžence | Kč utraceno včera |
|------------------------------------|-----------------------------|

Doporučení na topování inzerátu
Vybíráme inzeráty, které nebyly za posledních 7 dní topované

| | | | | |
|--|---|--|--|---|
| <p>Prodej rodinného domu 456 m², pozemek 452 m² Sázava, Svážná 460 000 Kč</p> | <p>Prodej rodinného domu 150 m², pozemek 327 m² Racková 3 150 000 Kč</p> | <p>Prodej rodinného domu 50 m², pozemek 977 m² Bohuslavice u Zlína 2 640 000 Kč</p> | <p>Pronájem rodinného domu 120 m², pozemek 566 m² P 26 900 Kč</p> | <p>Pronájem bytu 2+kk 123 m² Sázava, Svážná</p> |
|--|---|--|--|---|

[Kontaktujte nás](#) » Jsme offline

https://master.klient.sreality.test.dszn.cz



... i z nákupů

SEZNAM.CZ VP ⁴

...najdu tam, co neznám **Vyhledat**

Právě se hledá Bomba v Litvínově Počasí radar B

Pátek 30. srpna. Svátek má Vladěna. **Meteoradar**

29°C Dnes **27°C** Sobota **26°C** Neděle

Na později TV program Mapy **Jízdní řády**

Email Firemní e-mail zdarma

Doručené (973) Kalendář Poznámky **+**

| | | |
|--------------------|--------------------------|-------|
| neodpovidat@s... | Upozornění: Produkt... | 13:45 |
| peter.pekarovic... | Změna výskytu: Prod... | 12:21 |
| Mikeš, Radoslav | [YouTrack, Updated] I... | 12:14 |
| peter.pekarovic... | Spolecna priprava na ... | 11:55 |

Skrýt e-maily

Reklama · **Koupit reklamu**

SAMSUNG

Galaxy Flip6 | Watch7





Diana Hlaváčová

Product Manager Senior

diana.hlavacova@firma.seznam.cz

Copyright © 1996–2024 Seznam.cz, a. s.

