sellma

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 l 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 INTELIGENCE – 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.

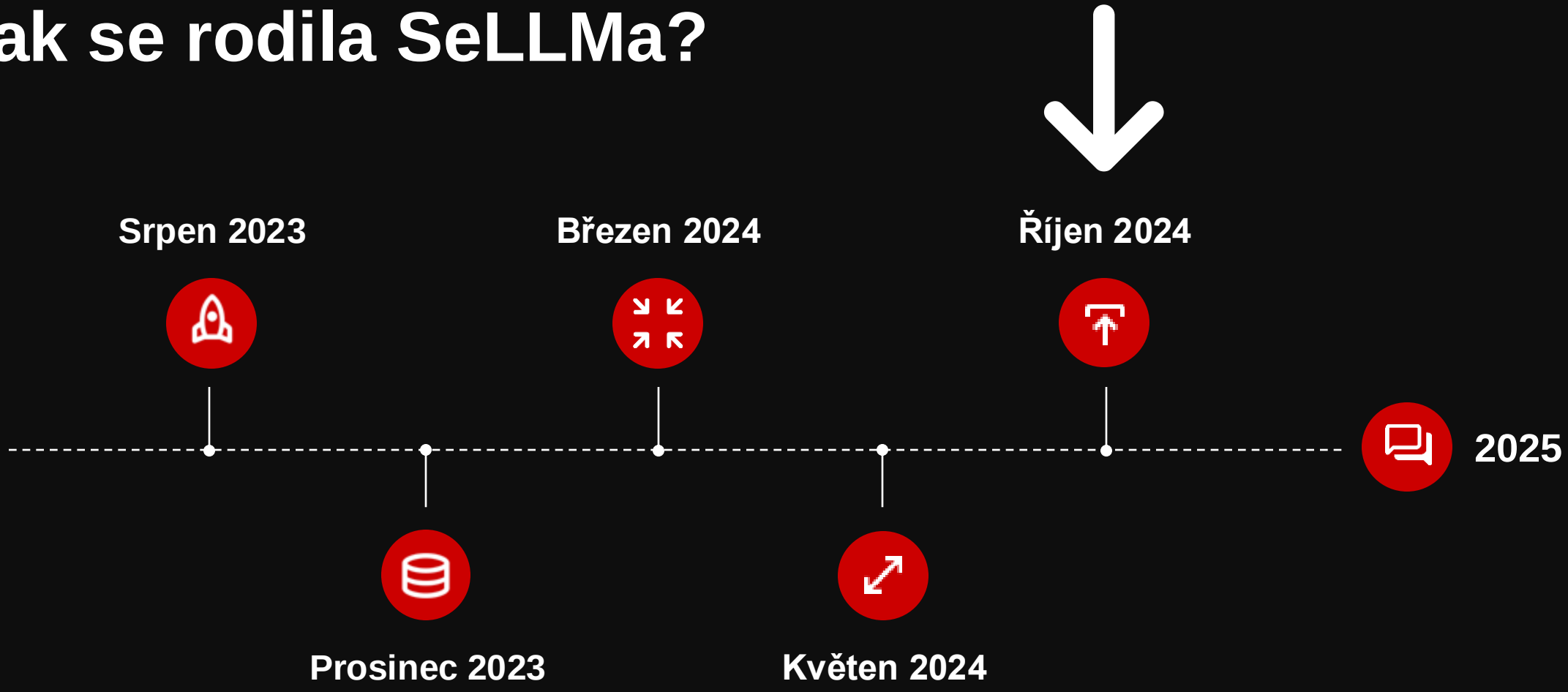

sellma

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 PCI

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

Kaufland Online marketplace

Online marketplace | Nabídka prodejny | Přihlásit se | Nákupní košík

Všechny kategorie

nvidia h100

Zpět k výsledkům vyhledávání

Grafická karta NVIDIA H100 PC bit, PCIe 5.0 x16, dvouslotová, držák ATX, podp odej

Zatím bez hodnocení

Kód výrob

Paměťová sběrnice: kód paměti: 5120 bitů

Podpora paralelního zpracování: NVLink

Paměť grafické karty: 94 GB

Grafický procesor: H100

Výrobce: PNY

Použití: Plynová karta pro grafické karty, která je součástí balení: PNY: Server/pracovní stanice

Typ chlazení: Pasivní

Prodává ESUS

Skladem už jen 4 kusy

988 012,06 Kč

plus 315,00 Kč za dopravu

Vrácení zboží zdarma do 14 dnů

1 | Přidat do košíku

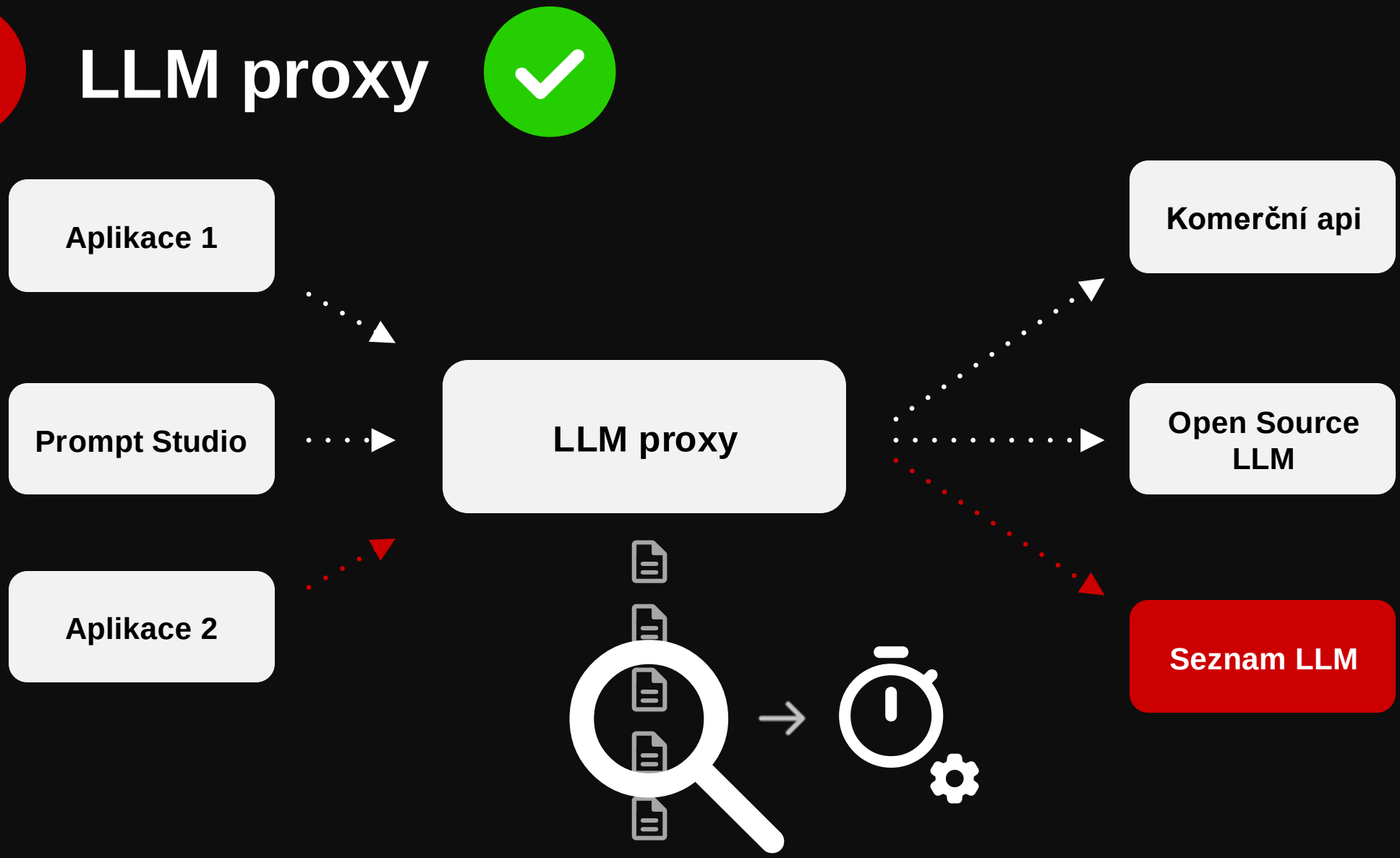
st 25. – čt 26. září

Prodává ESUS

Komerční prodejce

VOP / Informace o prodejci / Odstoupení od smlouvy / Ochrana osobních údajů

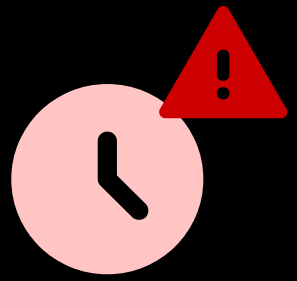
2 LLM proxy


Aplikace 1
Prompt Studio
Aplikace 2
LLM proxy
Komerční api
Open Source LLM
Seznam LLM

3 Víme, jaké aplikace nasazujeme

- Jaký model je využívaný
- 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
- Dalších 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.

Streaming

metriky:
- Time to first token
- Inter token latency

Sequential

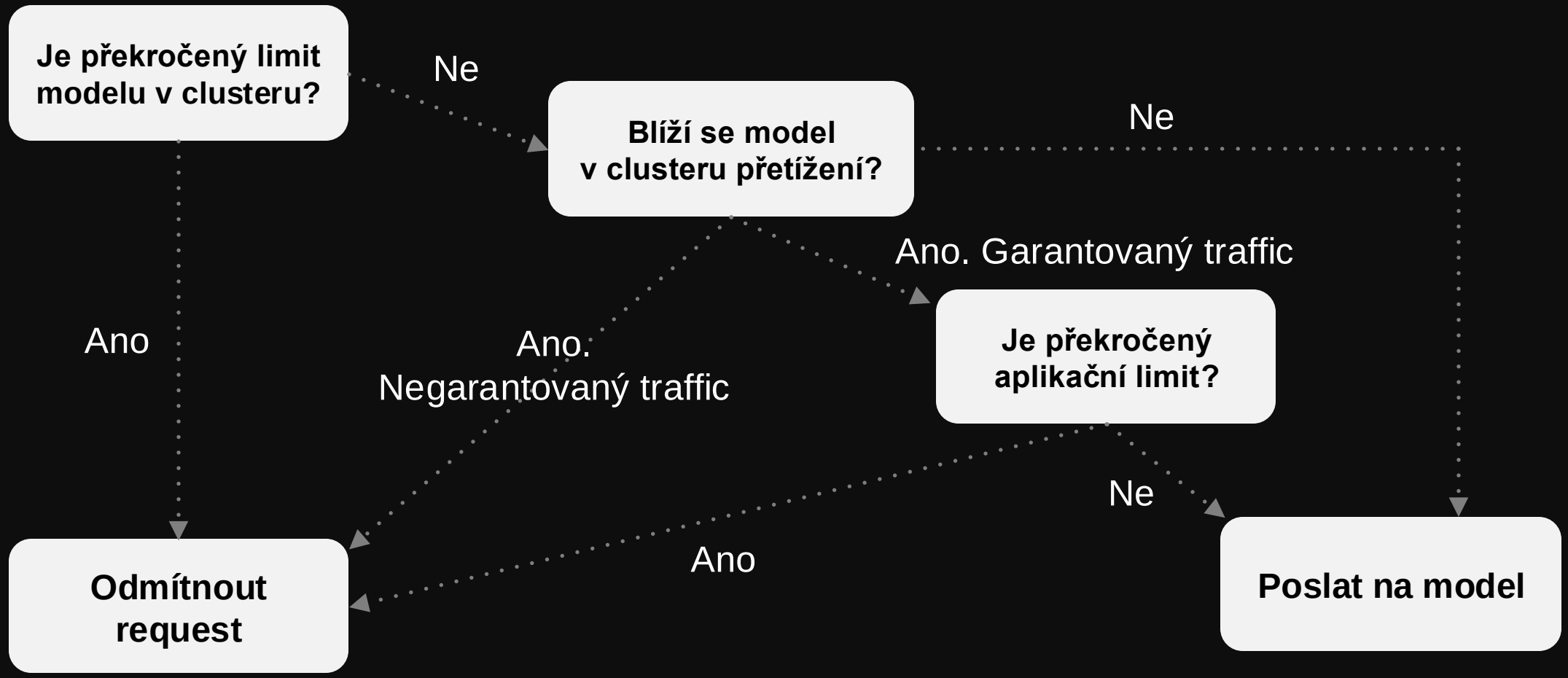
metriky:
- End to end latency (request latency)

Offline (Batch)

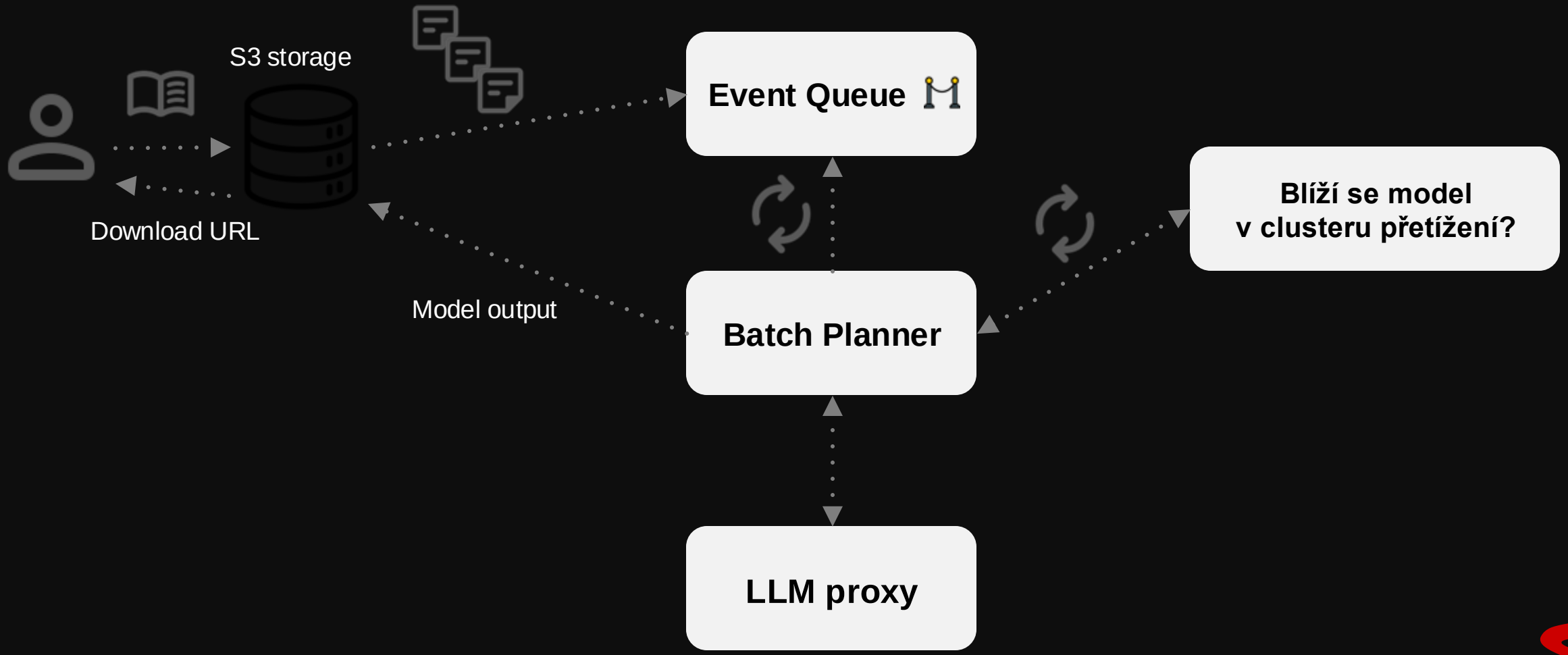
Nejjednodušší: zaměření na maximální propustnost.

metriky:
- request_throughput
- output_token_throughput

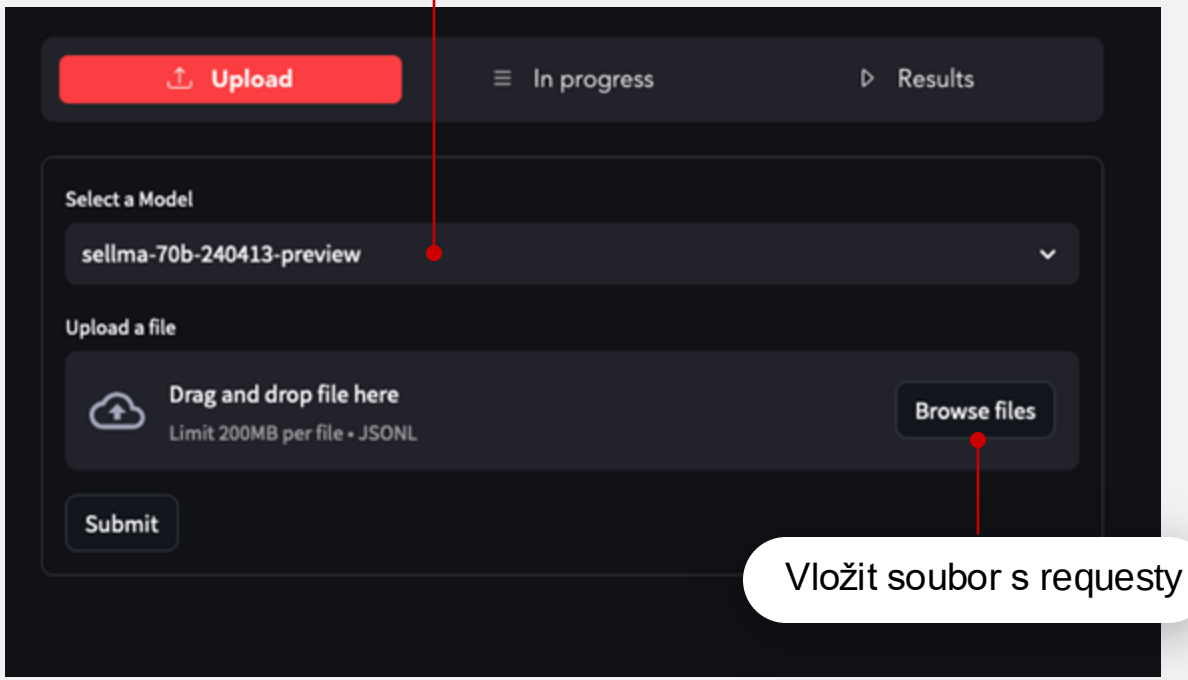
Garance pro provozní aplikace



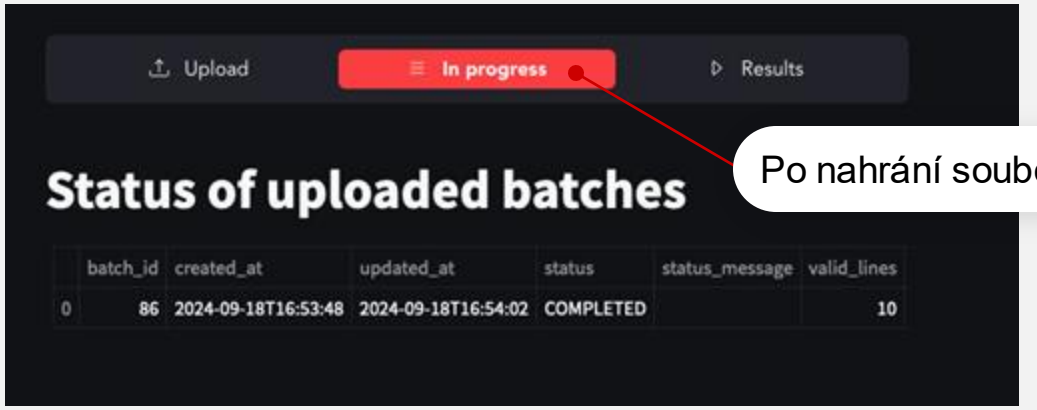
Offline (batch) inference



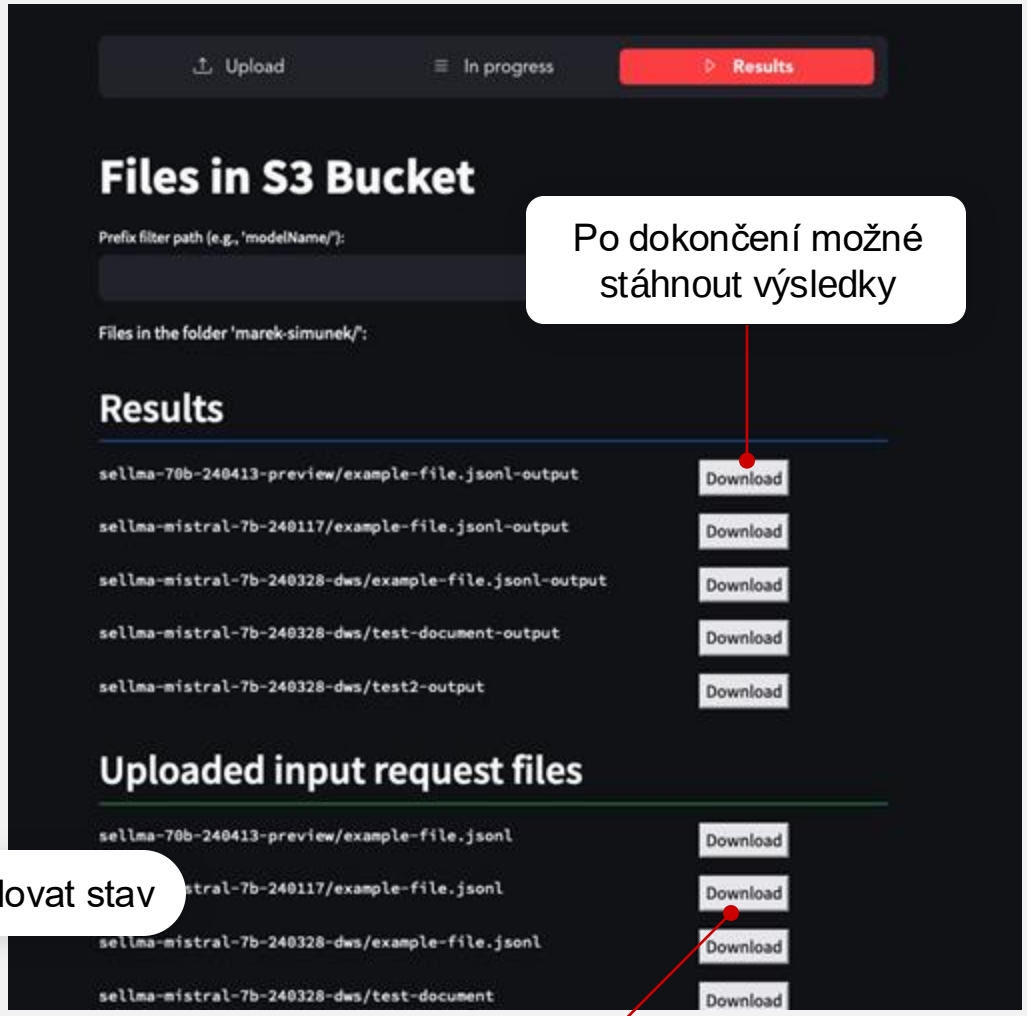
Výběr SeLLMa modelu



Vložit soubor s requesty



Po nahrání souboru můžu sledovat stav



Po dokončení možné stáhnout výsledky

Vstupní nahrané soubory


Powered by
sellma

proč polníček

Internet Obrázky Zboží Mapy Videa 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 **vitamín C, E, B, provitamín 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 · Vě
polní? Kde se vz
článku!

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 deb

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

Pastrňák svatba

Internet Obrázky Zboží Mapy Videa Zprávy Firmy Slovník

Zprávy na téma: Pastrňák svatba

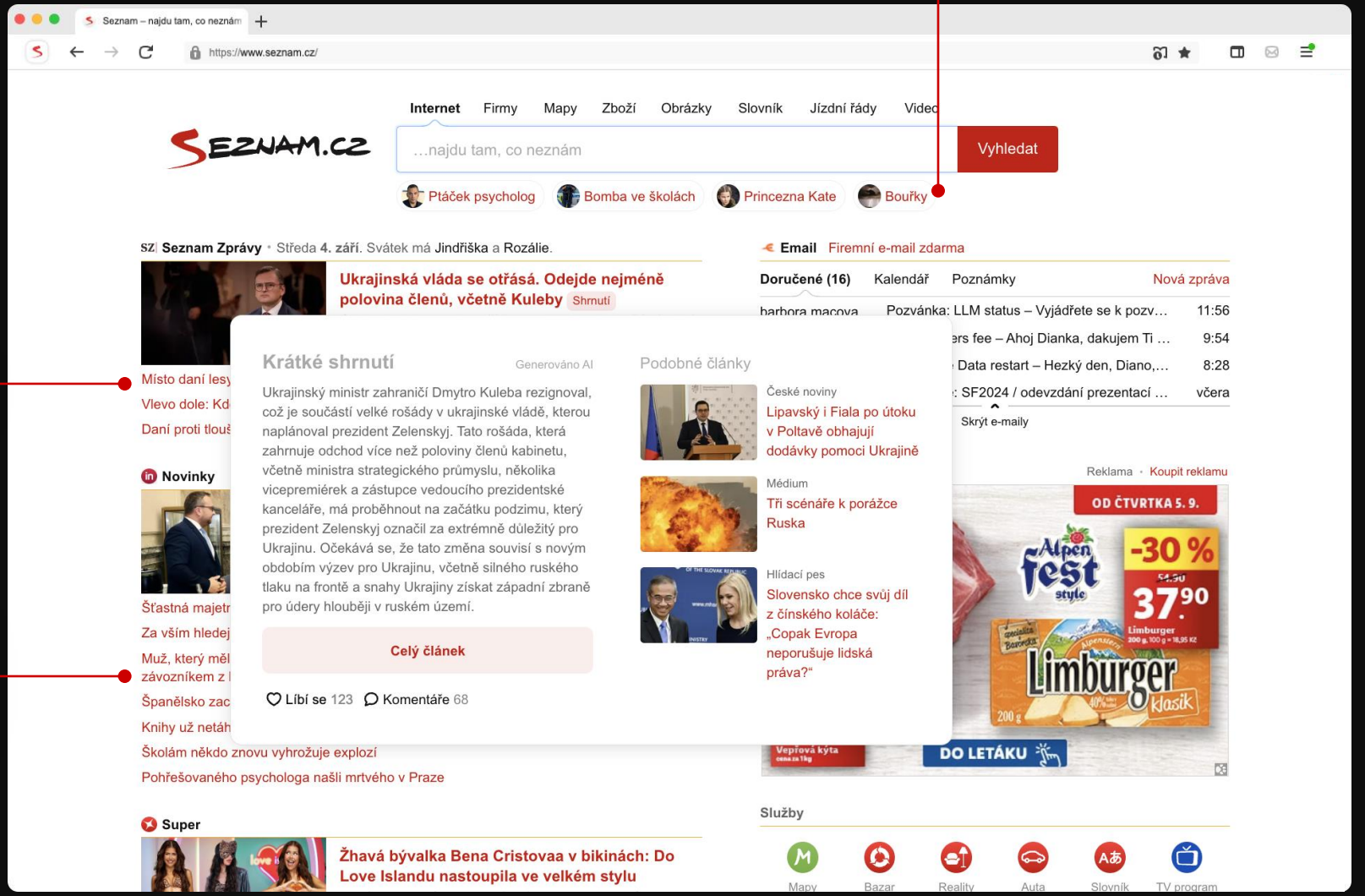
Český hokejový útočník David Pastrňák se oženil se švédskou partnerkou Rebeccou Rohlssonovou na chorvatském ostrově Hvar, kde vychovávají roční dceru Freyu Ivy.[1, 2, 3] Pastrňák se na Instagramu pochlubil fotkou ze svatby, která byla třídenním luxusním veselím, na němž nešetřil a které stálo miliony.[4,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

Tento text je vygenerovaným shrnutím tématu

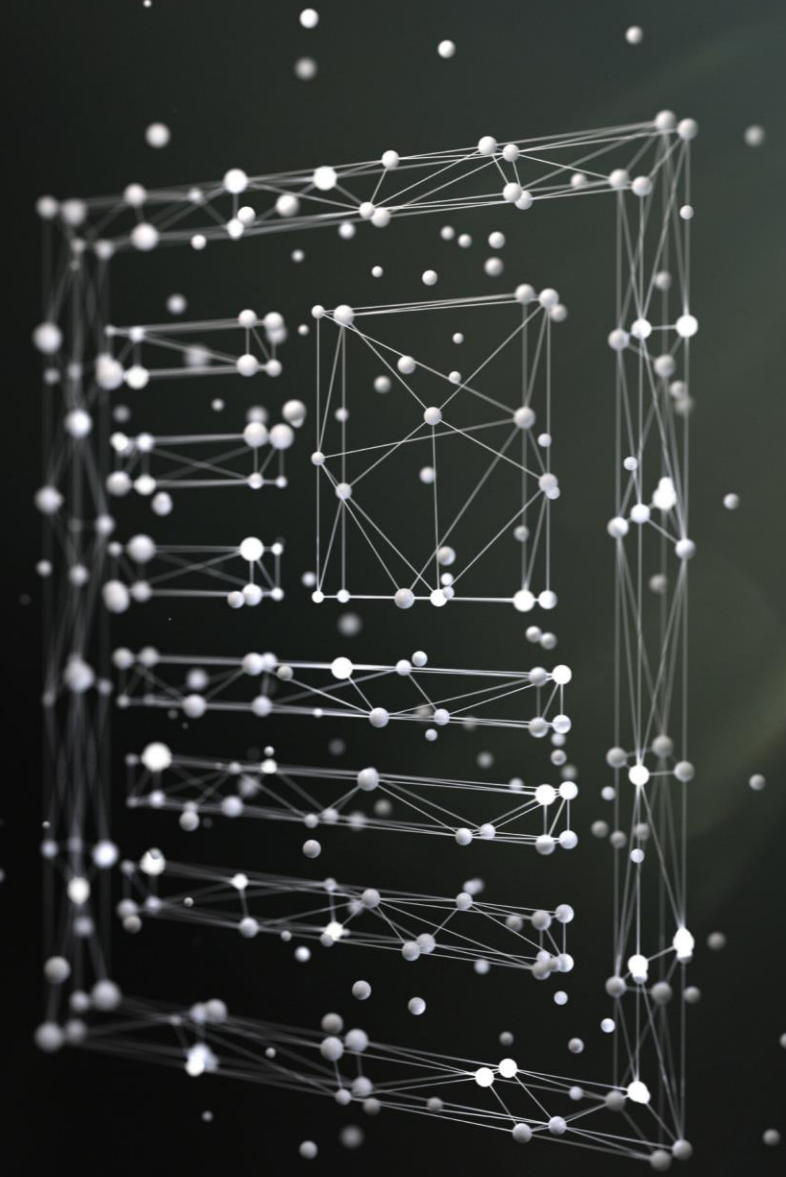
Shrnutí aktuálního dění pro nejhledanější dotazy

Sumarizace článků

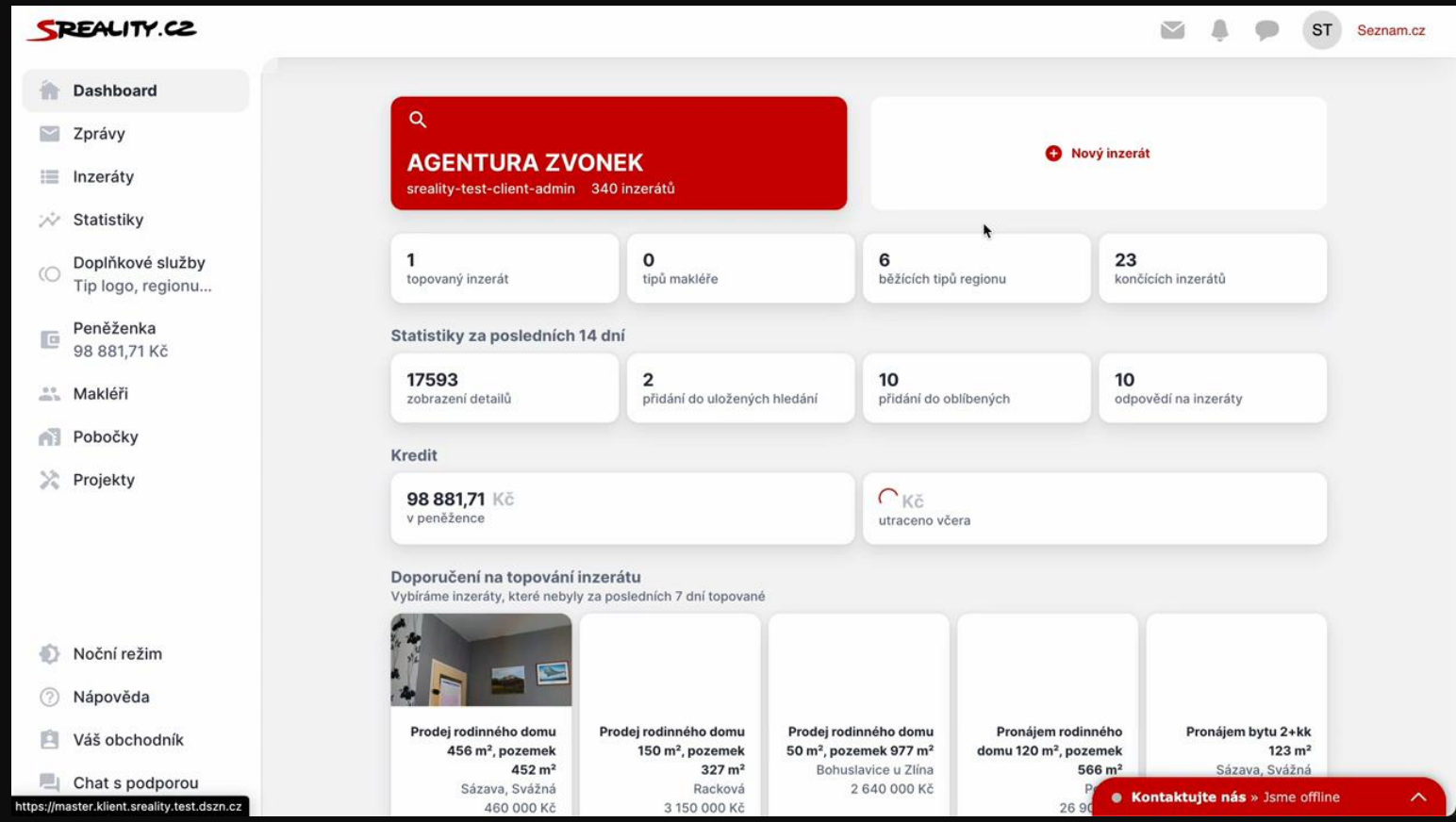
Nadpisy zahraničních článků v češtině



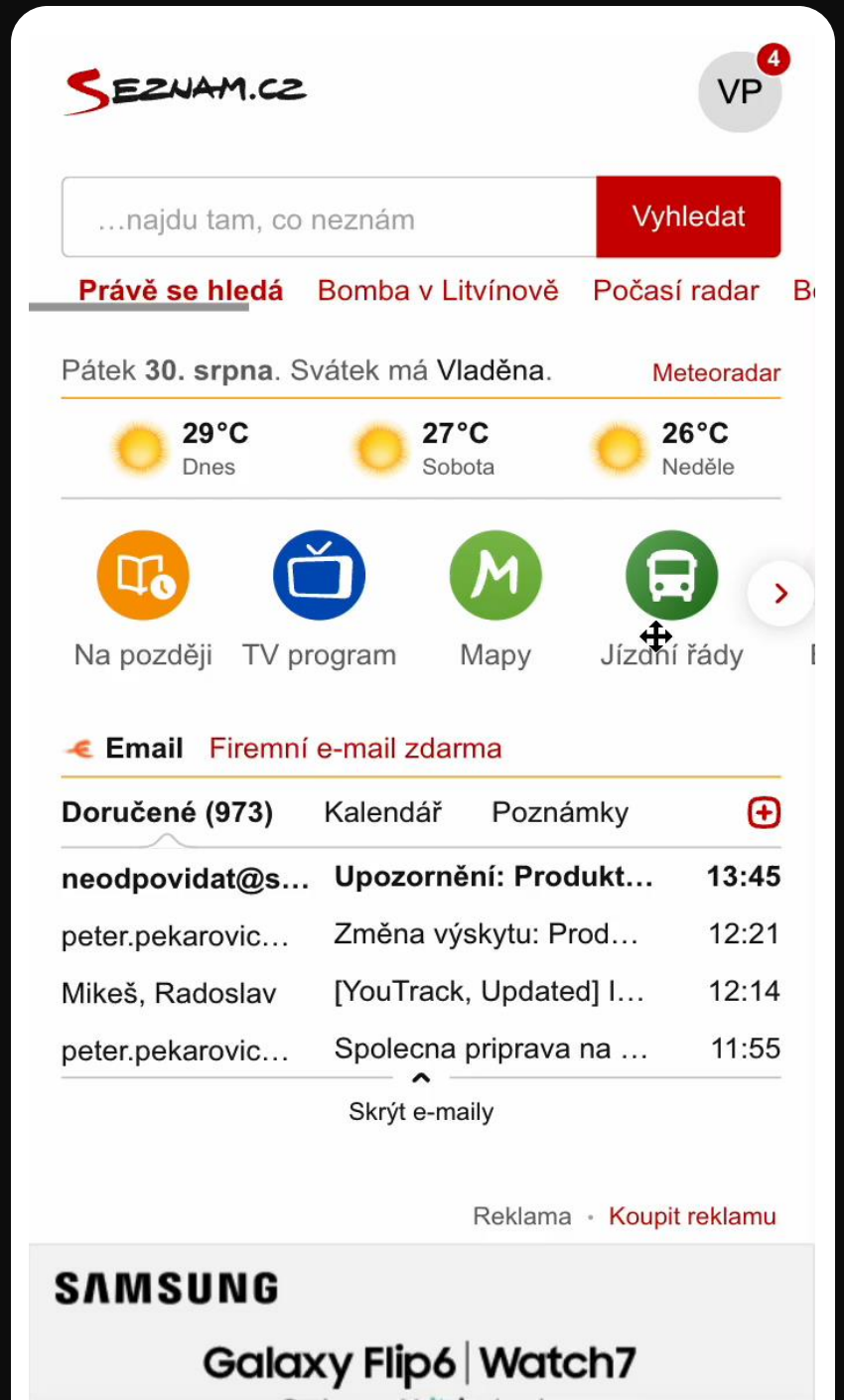
Co ještě chystáme?

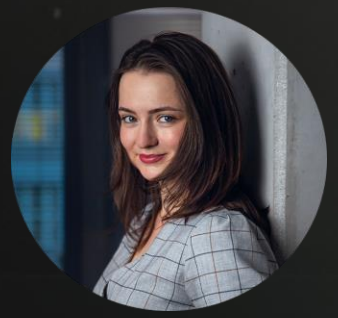


Nový zážitek z inzerce



… i z nákupů


SEZNAM.CZ
4
VP
…najdu tam, co neznám
Vyhledat
Právě se hledá
Bomba v Litvínově
Počasí radar
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