Modely pro Seznam AI

Rok, ve kterém jsme třikrát přehazovali výhybku

Vítejte, jsem Seznam AI.

Šetřím vám čas přehlednými odpověďmi na to, co právě řešíte.
Shrnu vám aktuální dění i další praktické informace.

Zeptejte se...

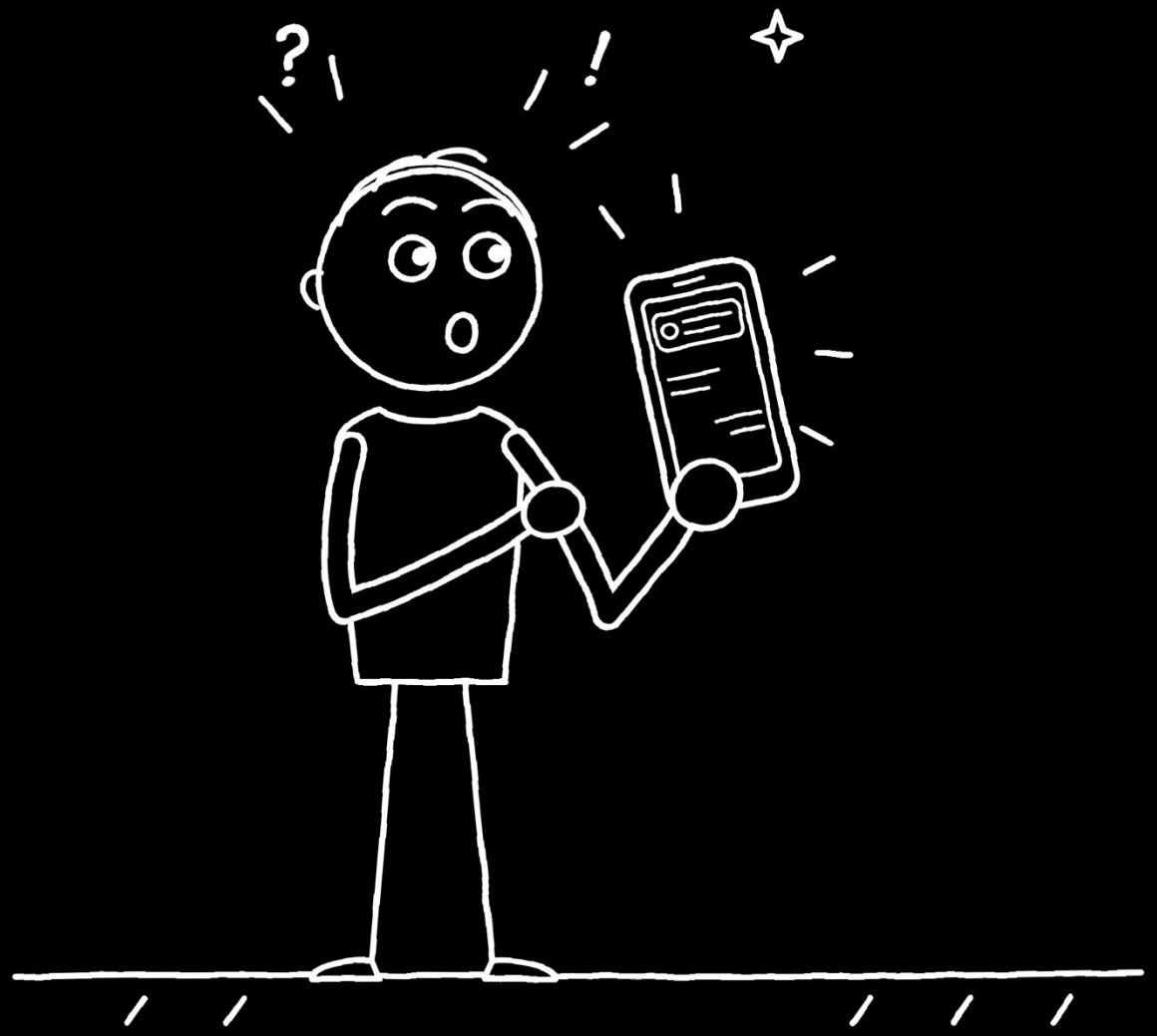
K milionům konverzací

Jak rychle naše AI roste

1. milion konverzací	**půl roku**
2. milion konverzací	**měsíc**
Současnost	**< týden**
Do konce roku	**dny**
Příští rok	**každý den**



2025: První prototyp na externích modelech





Interní infrastruktura

- Ekonomika
- Nezávislost
- Know how
- Data v ČR
- Zacílení, čeština

2025

sellma 2

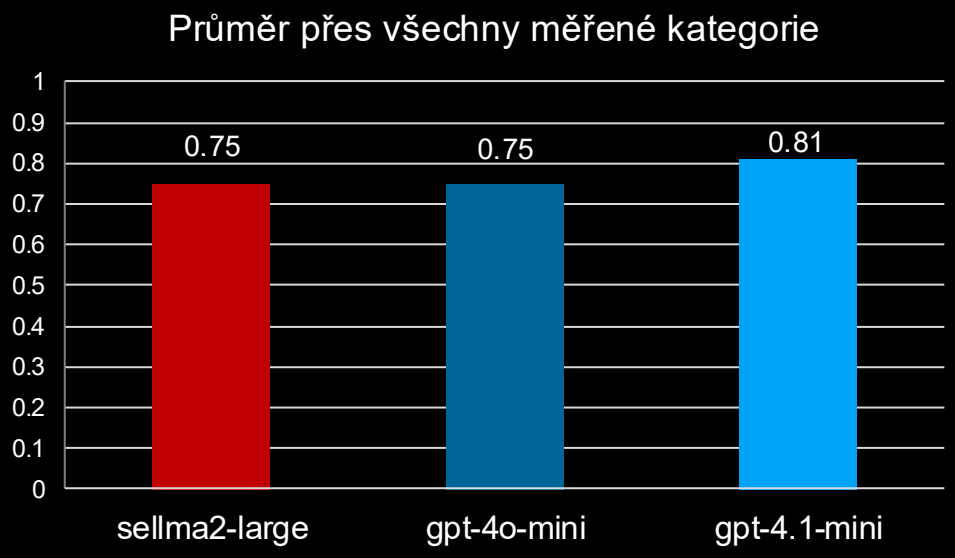
SeLLMa 2 odbavila, co jsme tehdy potřebovali

Velikost: 70B parametrů, dense, Llama3.3

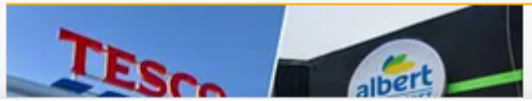
Délka kontextu: 16k tokenů

Klíčové schopnosti:

- Podpora systém promptu
- Zlepšená stabilita výsledků
- Strukturovaný výstup
- Function calling
- Bezpečnost



Novinky • Pátek 30. října. Svátek má Tadeáš.



Z Teska bude Albert

Skrýt

Shrnutí generováno Seznam AI.

Společnost Ahold Delhaize, provozující síť Albert, převezme prodejny Tesco v Česku a na Slovensku za 41,4 miliardy korun. Akvizice posílí pozici Ahold Delhaize ve střední a jihovýchodní Evropě a umožní vstup na slovenský trh. Tesco v regionu provozuje přes 360 prodejen a zaměstnává více než 14 tisíc lidí.

Celý článek

760

CN

Doručené 21. 10. 2026, 14:45

AI Policy Brief: druhé vydání

Od: CNAIP info@mail.cnaip.cz

Komu: martina.pomikalkova@firma.seznam.cz

Shrnout e-mail

Shrnutí generováno Seznam AI.

Česká vláda a CEE AI ambasador Jan Kavalírek připravili novinky o Pražské AI deklaraci, Národní AI bráně a české kandidatuře na AI Gigafactory. Evropská komise navrhla nařízení Cloud and AI Development Act (CADA), které ovlivní cloudové a AI služby v Česku. AI Act vstoupí v platnost 2. srpna 2026 a vyžaduje transparentnost u chatbotů a AI obsahu. Firmy musí dokončit inventarizaci AI systémů a compliance do prosince 2027. Evropa buduje vlastní AI infrastrukturu, včetně 19 AI Factories a prvního IPCEI pro AI, což může ovlivnit české firmy. AI mění kyberbezpečnost a firmy musí řešit nové útočné plochy spojené s AI.

Shrnutí recenzí

Vygenerováno na základě 7 recenzí pomocí Seznam AI.

Klec Ferplast Casita 120 je prostorná a esteticky povedená, s dobrým přístupem pro čištění a manipulaci. Zákazníci oceňují její odolnost a praktičnost, ale upozorňují na nekvalitní příslušenství a drobné konstrukční nedostatky. Pro většinu uživatelů však splňuje základní požadavky na klec pro malá zvířata.

Co lidé chválí

- Prostorná konstrukce s velkým přístupem shora
- Boční dvířka umožňující snadný přístup
- Odolná mříž proti okusu
- Esteticky pěkné provedení
- Dobře čistitelná

Na co si dát pozor

- Plastový domeček a miska jsou nekvalitní
- Napáječka protéká
- Horní dvířka pruží a nelze plně otevřít
- Obtížnější otevírání bočních dvířek
- Návod na sestavení je hůře dostupný

halloween

AI.













Další obrázky ›

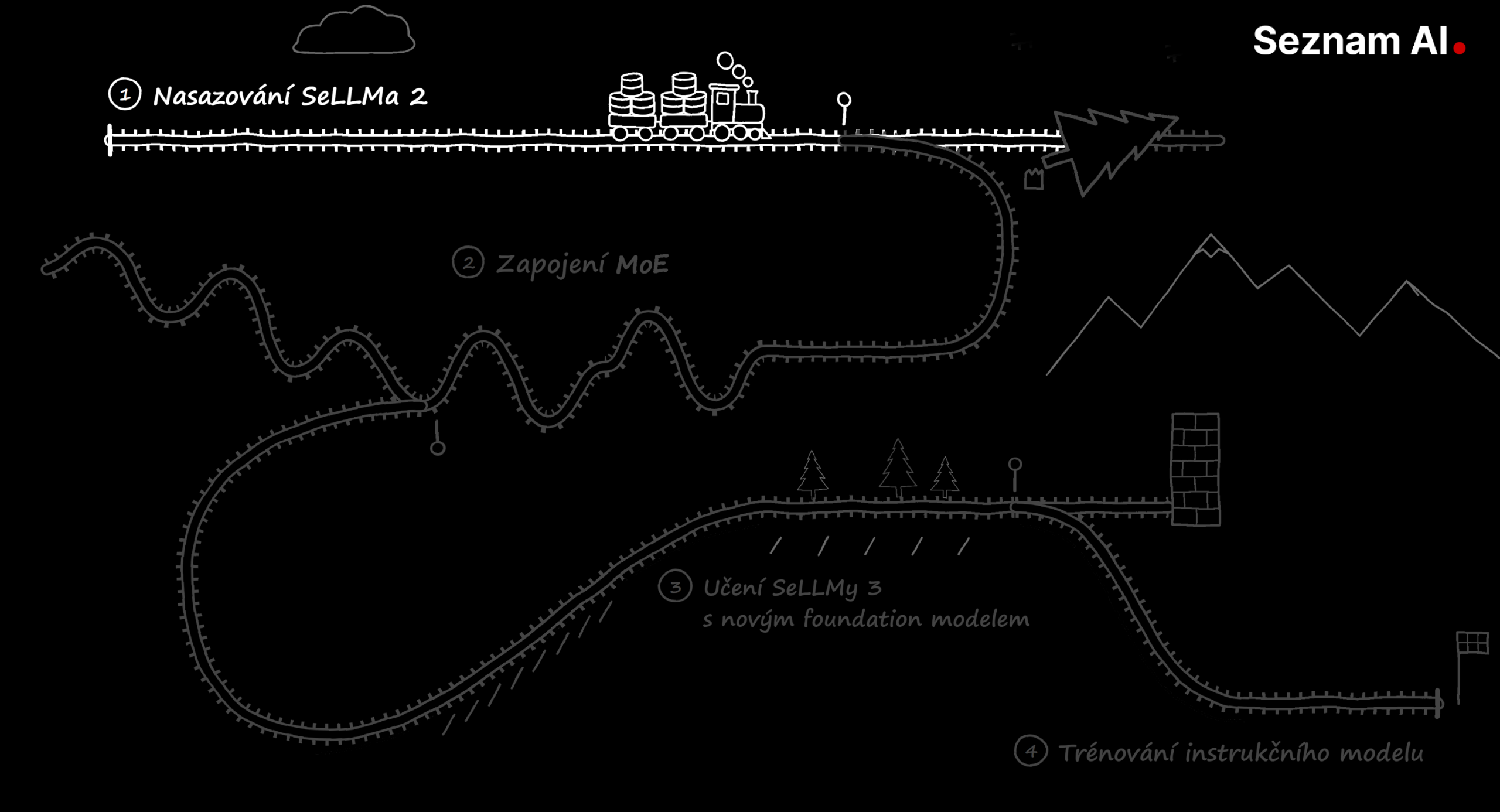
Něco se mi nezdá

Mohlo by vás zajímat

Kdy se slaví Halloween?

Jaká je historie Halloweenu?

Seznam AI.


1 Nasazování SeLLMa 2
2 Zapojení MoE
3 Učení SeLLMy 3
s novým foundation modelem
4 Trénování instrukčního modelu

Vydání modelu je okamžik, kdy dostanete nový seznam

1

Paralelní function calling

Aplikace potřebuje získat všechny v jedné odpovědi.

2

Délka kontextu 32k - 64k

Aby pojmul i dlouhé dokumenty při RAG

3

Omezení halucinací při RAG

Např. při shrnutí článků, e-mailů.

4

Robustnější strukturovaný output

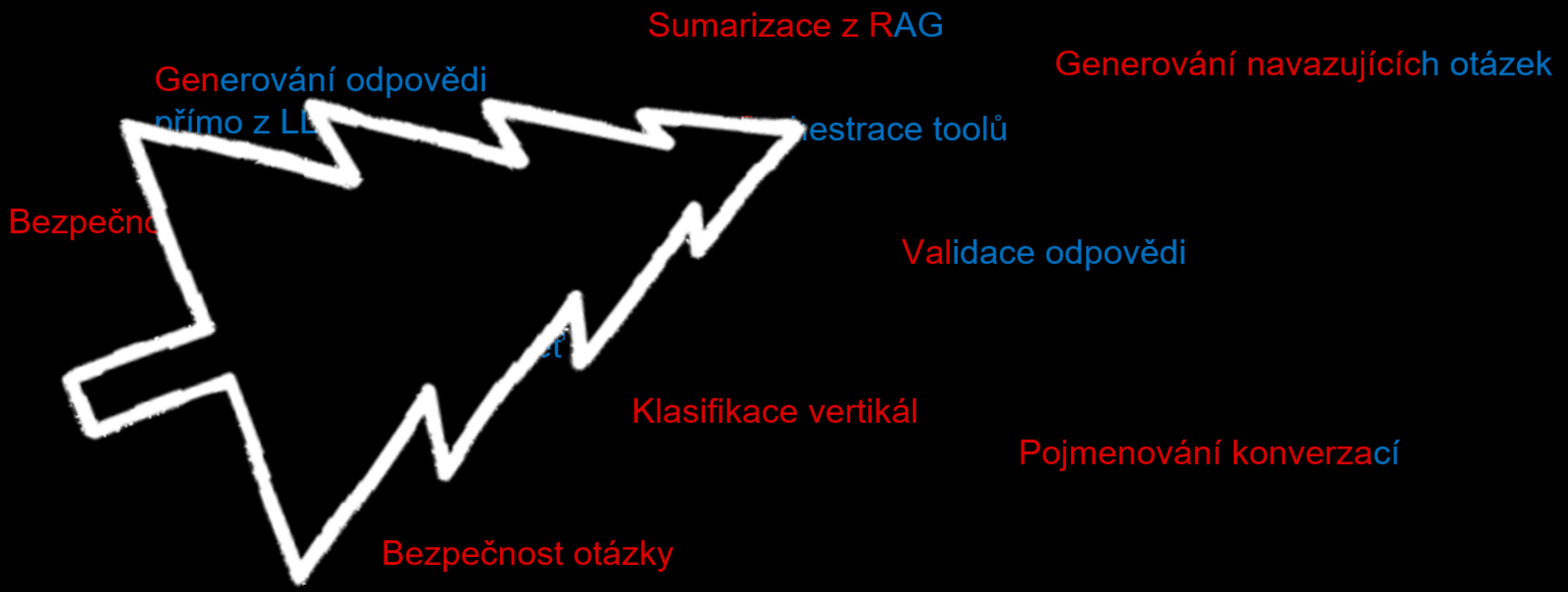
Aby model spolehlivě respektoval schéma ze system promptu..

5

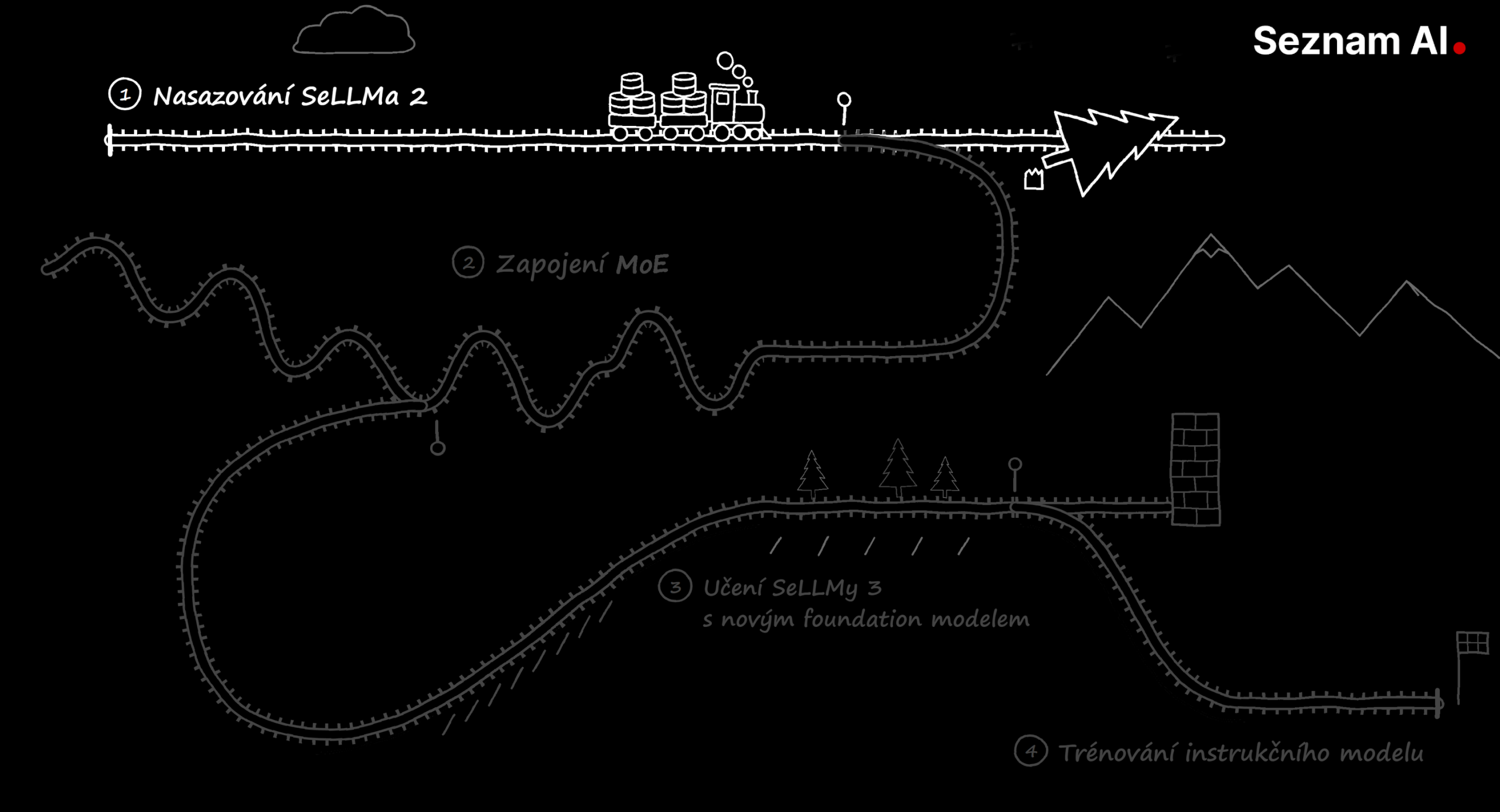
Robustnost vůči změnám v historii konverzace

Změna tématu konverzace – nové tooly – aby model nevolal ty staré.

Přepínání Seznam AI na SeLLmu 2 large



Seznam AI.


1 Nasazování SeLLMa 2
2 Zapojení MoE
3 Učení SeLLMy 3
s novým foundation modelem
4 Trénování instrukčního modelu

V čem byl problém?

Latence.
Propustnost.

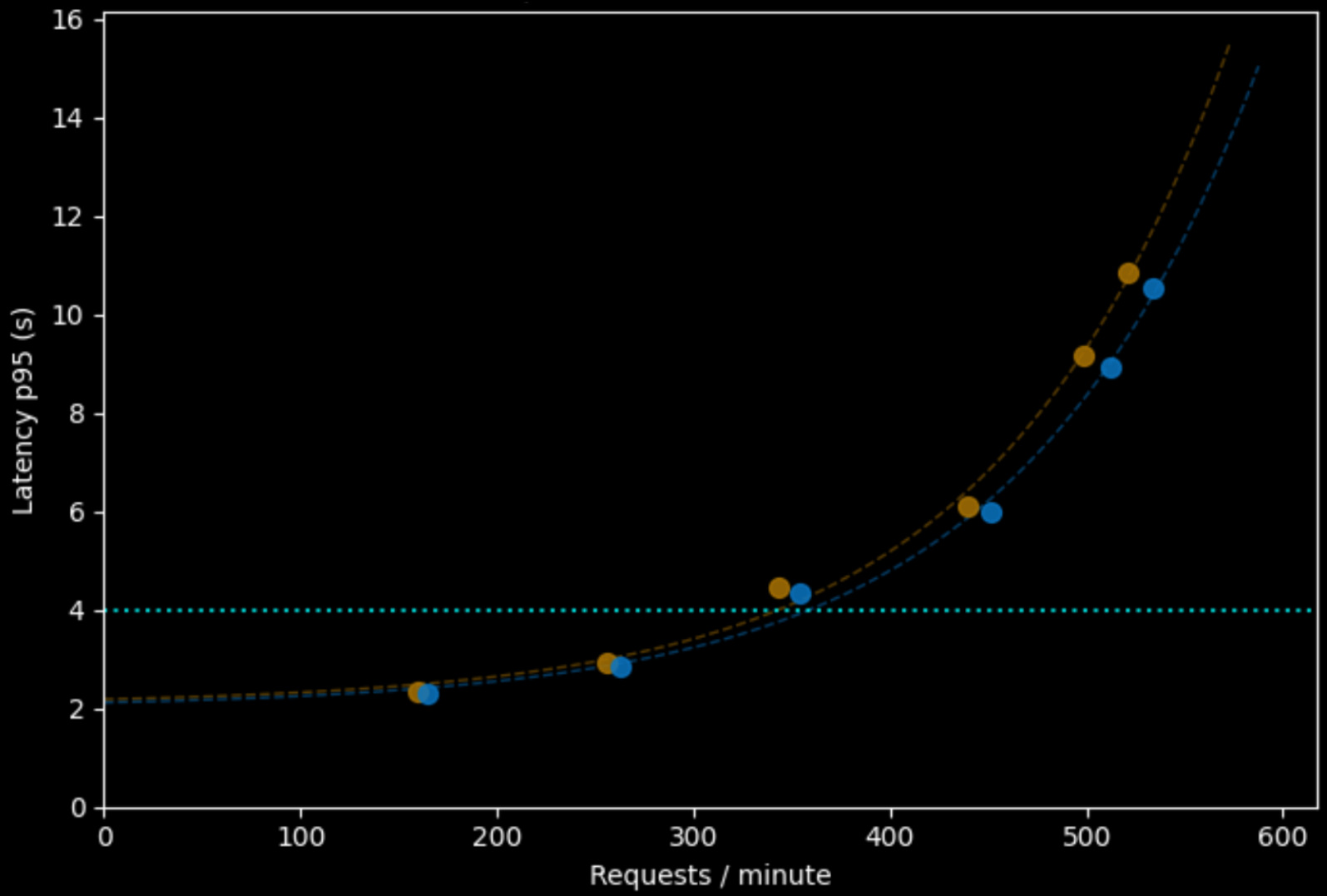


Traffic roste.
A jedna otočka Seznam AI není jedno volání modelu.



Sériová cesta určuje výsledné trvání.

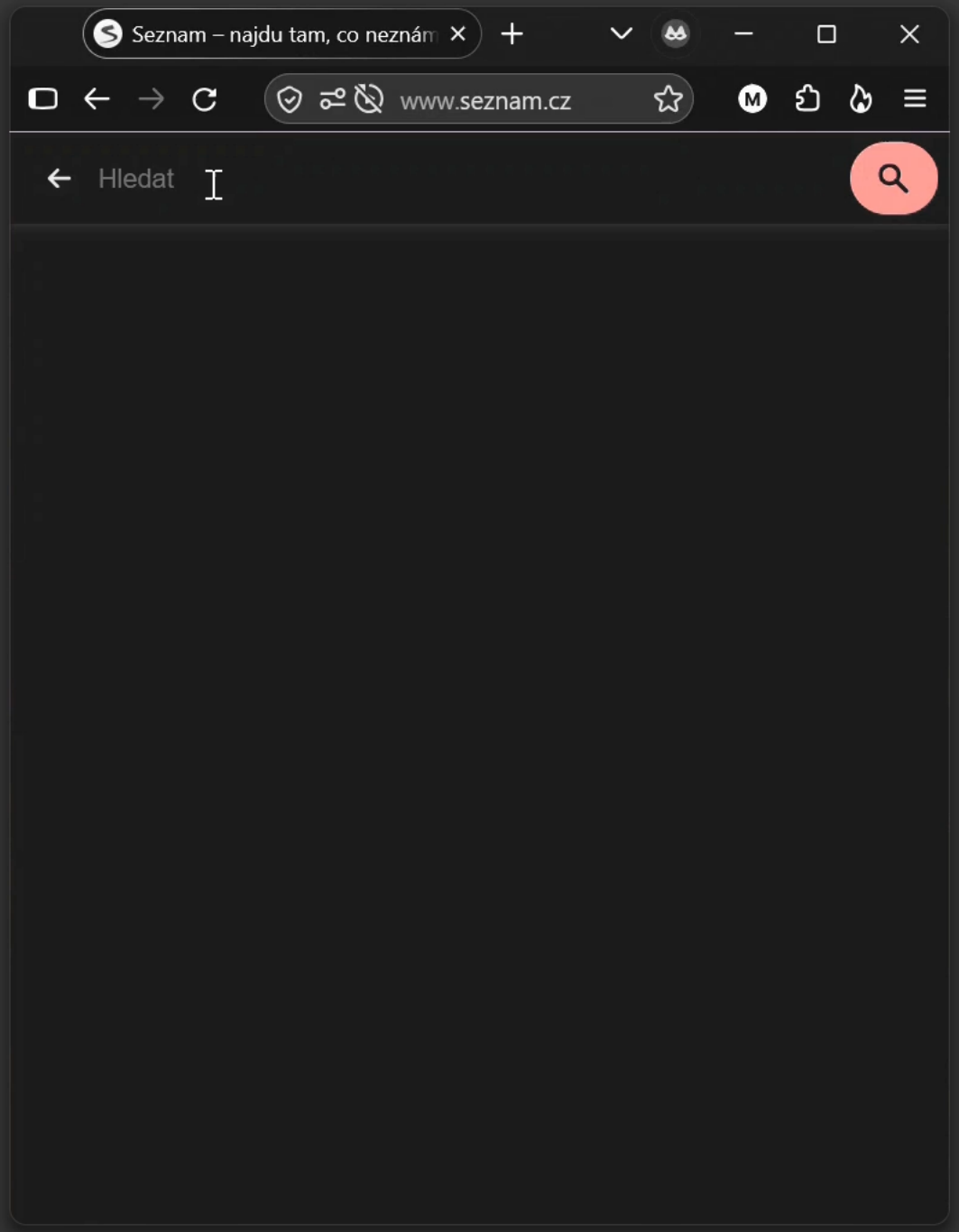
E2E latence rostou s počtem requestů



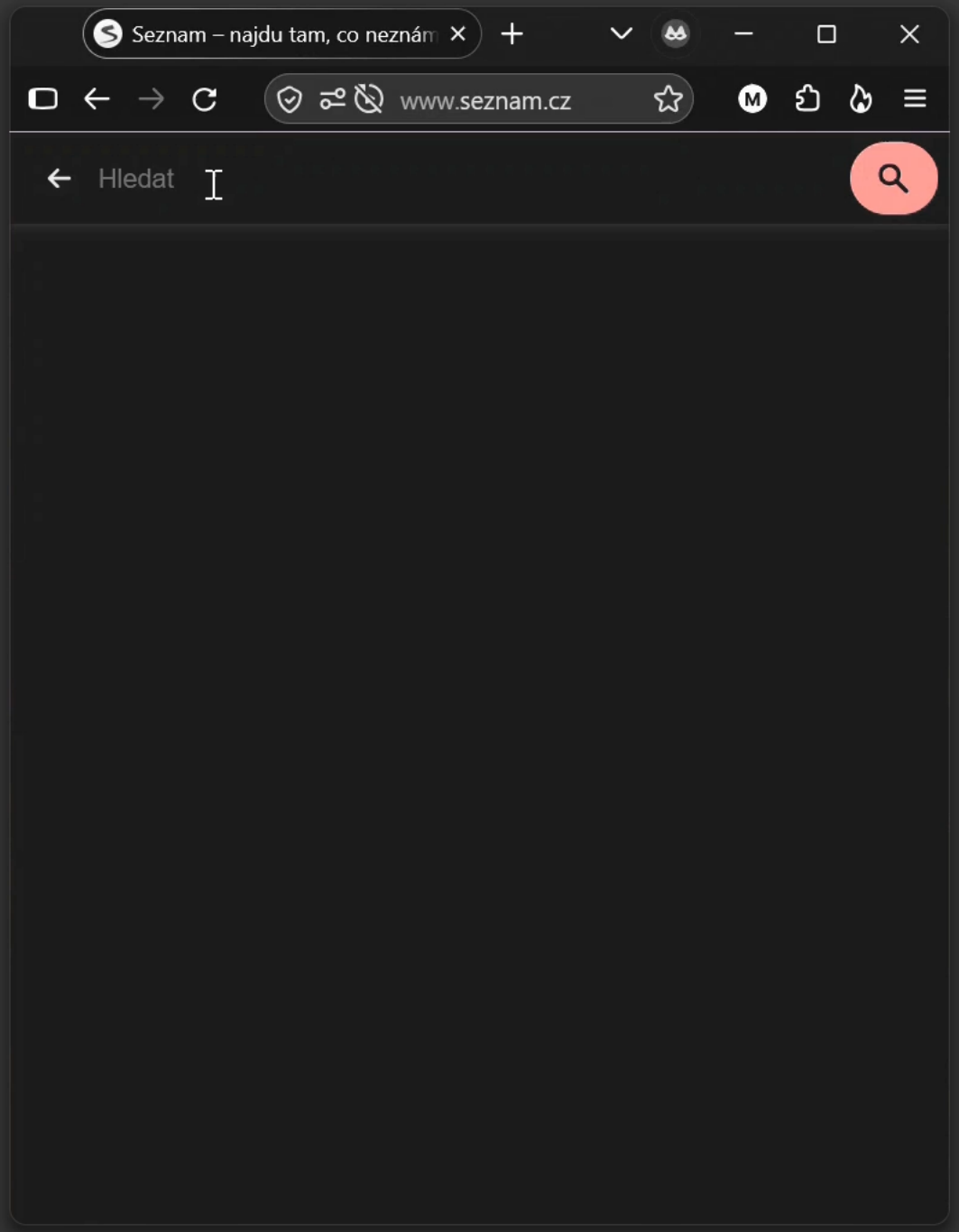
Proč jsou latence tak důležité?



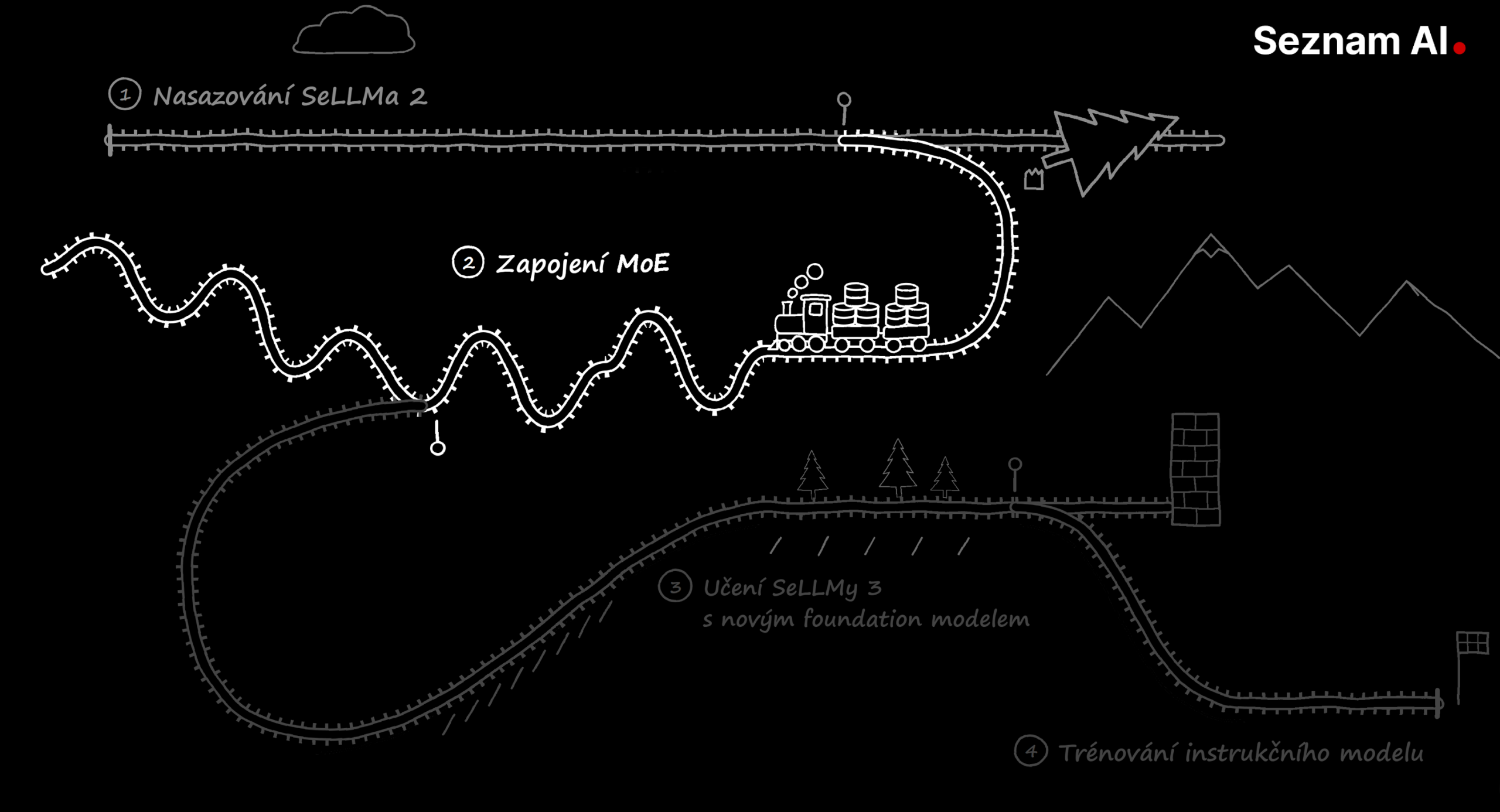

Seznam AI
asistent.seznam.cz
9+
Vítejte, jsem Seznam AI.
Šetřím vám čas přehlednými odpověďmi na to, co právě řešíte.
Shrnu vám aktuální dění i další praktické informace.
Jakým onemocněním trpí papež Lev?
Jaká je situace na akciových trzích?
V jaké oblasti získala Anne Carsonová letošní Nobelovu cenu?
Jaké deskové hry zkusit s přáteli?
Co je nového v ČR?
Jaká místa navštívit na Jižní Moravě?
Zeptejte se...
Odpovědi si ověřujte a nesdílejte citlivá data. Soukromí a nápověda.


Seznam – najdu tam, co neznám
www.seznam.cz
Hledat


Seznam AI
asistent.seznam.cz
9+
Vítejte, jsem Seznam AI.
Šetřím vám čas přehlednými odpověďmi na to, co právě řešíte.
Shrnu vám aktuální dění i další praktické informace.
Jakým onemocněním trpí papež Lev?
Jaká je situace na akciových trzích?
V jaké oblasti získala Anne Carsonová letošní Nobelovu cenu?
Jaké deskové hry zkusit s přáteli?
Co je nového v ČR?
Jaká místa navštívit na Jižní Moravě?
Zeptejte se...
Odpovědi si ověřujte a nesdílejte citlivá data. Soukromí a nápověda.


Seznam – najdu tam, co neznám
www.seznam.cz
Hledat

Seznam AI.


1 Nasazování SeLLMa 2
2 Zapojení MoE
3 Učení SeLLMy 3
s novým foundation modelem
4 Trénování instrukčního modelu

SeLLMa + GPT → SeLLMa + Gemma = rychlost vyřešena

Proč MoE

Gemma 4 26B A4B (MoE)

Aktivní parametry výrazně nižší než celkové - kvalita zůstává, latence klesá.

Co jsme získali

Propustnost ~6-12x vyšší ve srovnání se SeLLMou 2 large.

Možnost využít i menší stroje.

Za jak dlouho

Všechny testy byly připravené. Po měsíci jsme začali přepínat.



V čem byl problém?

Čeština.
České znalosti.



„Jak chutná španělský ptáček?“

MoE

V španělské kuchyni se připravuje mnoho druhů ptáků (např. bažant nebo přepelici), jejichž chuť závisí na způsobu úpravy – bývá to jemné, ale výrazné maso.

Nicméně, pokud je to otázka typu „vtip“, odpověď by mohla být: **„Jak chutná španělský ptáček? Jako španělský ptáček!“** (což nedává smysl, ale je to typický typ suchého humoru).

S

Španělský ptáček je tradiční české jídlo, které se skládá z hovězího masa, které je zavinuté do ruličky a plněné různými ingrediencemi, jako je slanina, cibule, okurka a vajíčko. Chuť španělského ptáčka je bohatá a vydatná, s výraznou chutí hovězího masa a slaniny, doplněná o kyselou a slanou chuť okurky a cibule. Vajíčko dodává jemnou a krémovou chuť. Celkově je chuť španělského ptáčka komplexní a vyvážená, s kombinací slaných, kyselých a jemných chutí.

„Vytvoř mi krátký kvíz o filmech“

MoE

Kolik Oscarů získala filmová série Vládce prstenů: Návrat krále?

A) 5
B) 7
C) 11

Není to náhoda, má to logické důvody

1

Zastoupení v korpusu

Čeština tvoří v MoE ~0,3 % pretrain dat*

* Odhad založený na:
- podpoře 140 jazyků,
- dominanci angličtiny,
- deklarovaném teplotním samplování ve prospěch malých jazyků

2

Tokenizace

EN: ~1,3 tokenu na slovo
CZ: ~1,8 tokenu na slovo

Víc tokenů na slovo znamená kratší efektivní kontext.

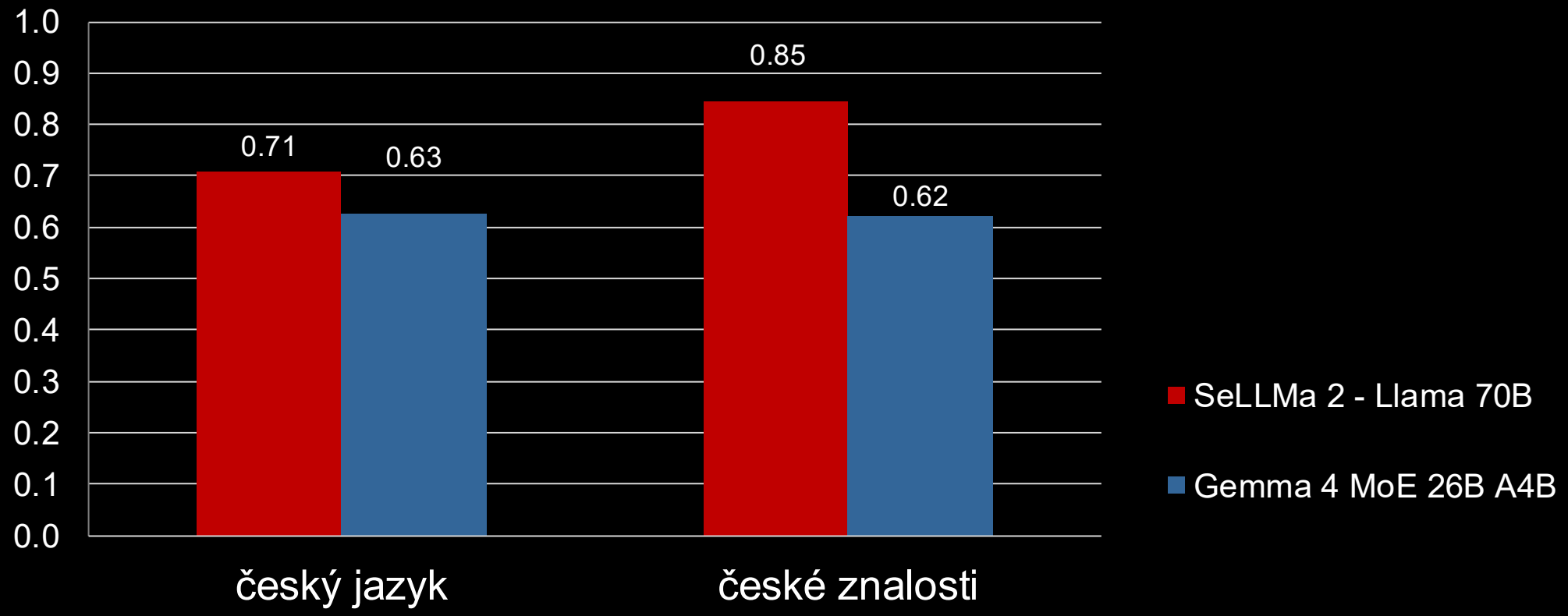
3

Morfologie

Bohaté skloňování tvoří chyby, které anglická evaluace nezachytí.



Měření českosti modelu

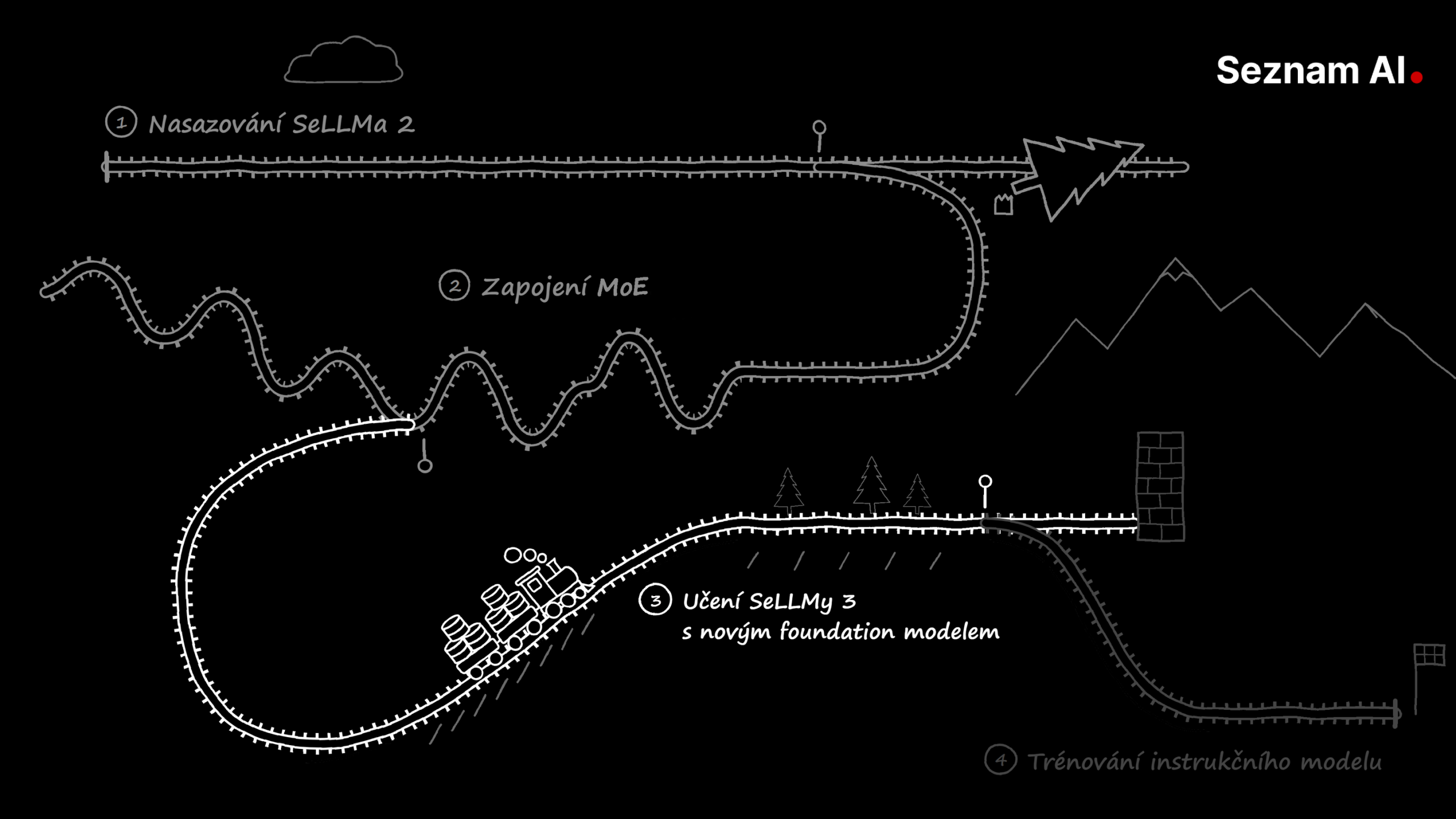


Příští SeLLMa musí být rychlá.

Žádný dílčí update poptávaných vlastností. Hledáme nový foundation model.



Seznam AI.


1 Nasazování SeLLMa 2
2 Zapojení MoE
3 Učení SeLLMy 3
s novým foundation modelem
4 Trénování instrukčního modelu

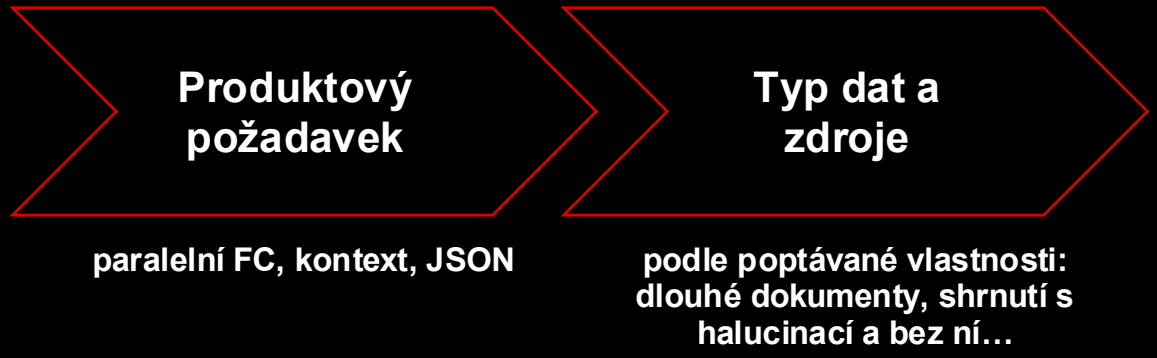
Každá nová schopnost je především datový problém



paralelní FC, kontext, JSON

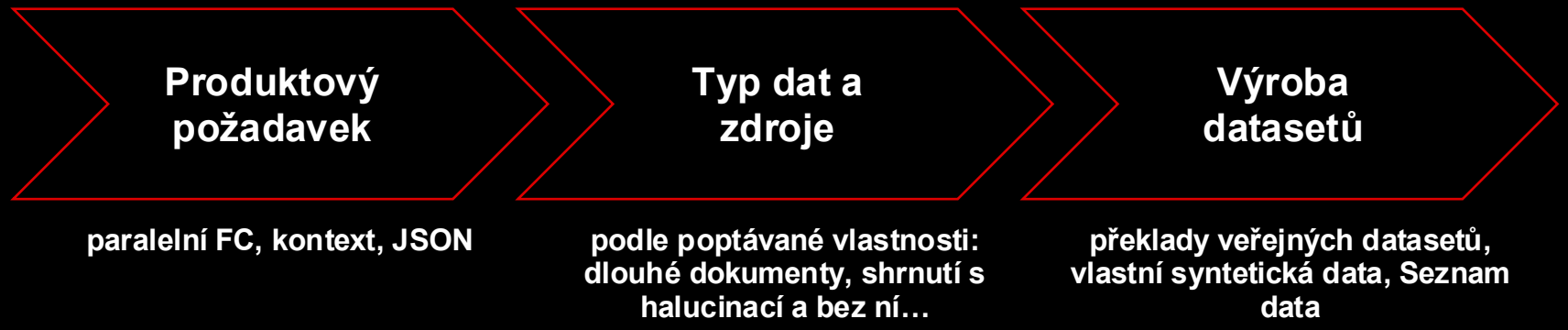
Mezi „aplikace vyžaduje paralelní function calling“ a hotovým modelem leží hlavně rozhodnutí o datech a jejich příprava.

Každá nová schopnost je především datový problém



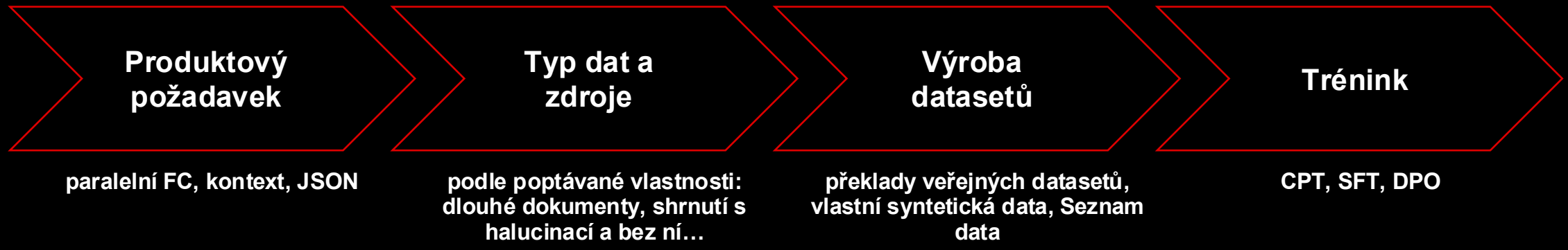
Mezi „aplikace vyžaduje paralelní function calling“ a hotovým modelem leží hlavně rozhodnutí o datech a jejich příprava.

Každá nová schopnost je především datový problém



Mezi „aplikace vyžaduje paralelní function calling“ a hotovým modelem leží hlavně rozhodnutí o datech a jejich příprava.

Každá nová schopnost je především datový problém



Mezi „aplikace vyžaduje paralelní function calling“ a hotovým modelem leží hlavně rozhodnutí o datech a jejich příprava.

Jak naučit open-weight model česky

0 Data

Příprava dat

- velké množství českých dat
- datasety na míru vylepšovaným funkcím
- čištění, deduplikace, filtr kvality
- replay: angličtina, kód

1 Čeština do base modelu

Continued pretraining

- base model na mixu češtiny a replaye
- nižší learning rate
- cooldown na nejkvalitnějších datech

SFT

- české instrukce, multi-turn
- function calling, JSON
- dlouhý kontext

DPO

- kvalita češtiny, RAG halucinace, délka výstupu, cyklení
- RL s ověřitelnou odměnou
- on-policy DPO

2 Instrukční vrstva

Merge s instruct

- přiblížit váhy instrukčnímu modelu
- např. přičíst chat vector nebo zprůměroval
- vrací instrukční schopnosti

3 Nasazení

Provoz

- kvantizace FP8
- spekulátor pro rychlejší dekódování



Base



IT



Base



IT



Base +
post-training



Base



IT



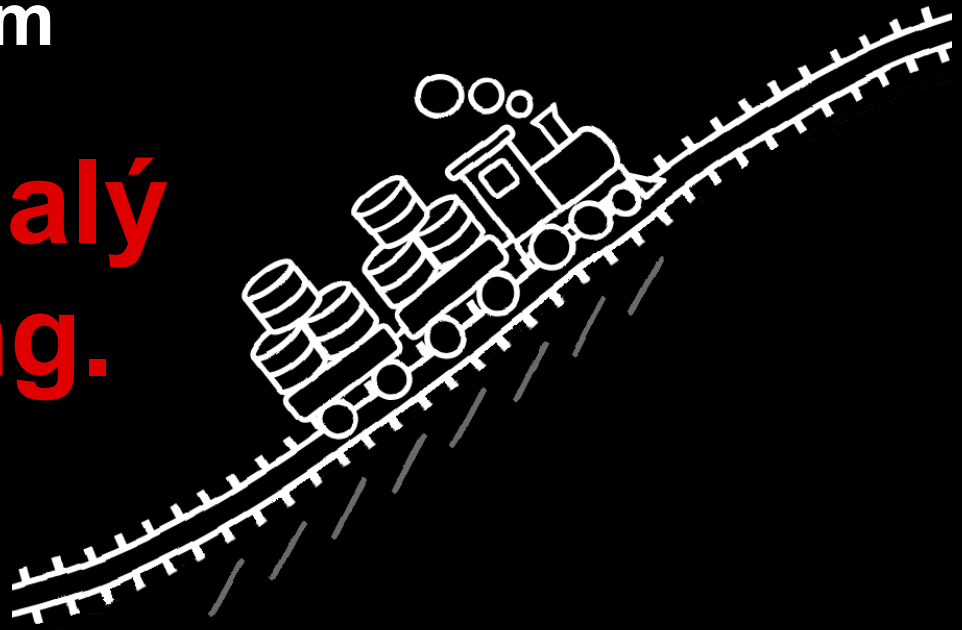
Base +
post-training





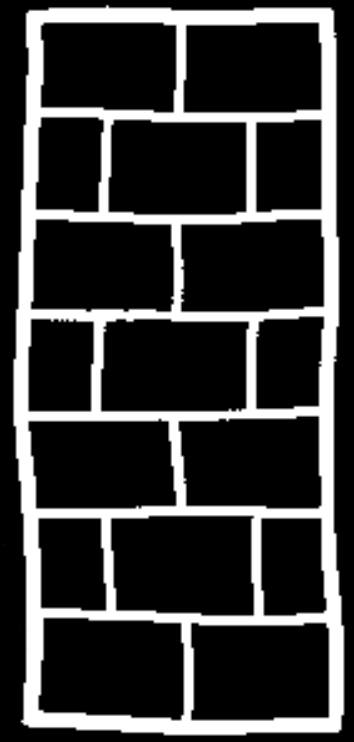
Menší problém

Velmi pomalý pretraining.



Zásadní problém

Doučený base model nelze sloučit s instrukčním.



Base



IT



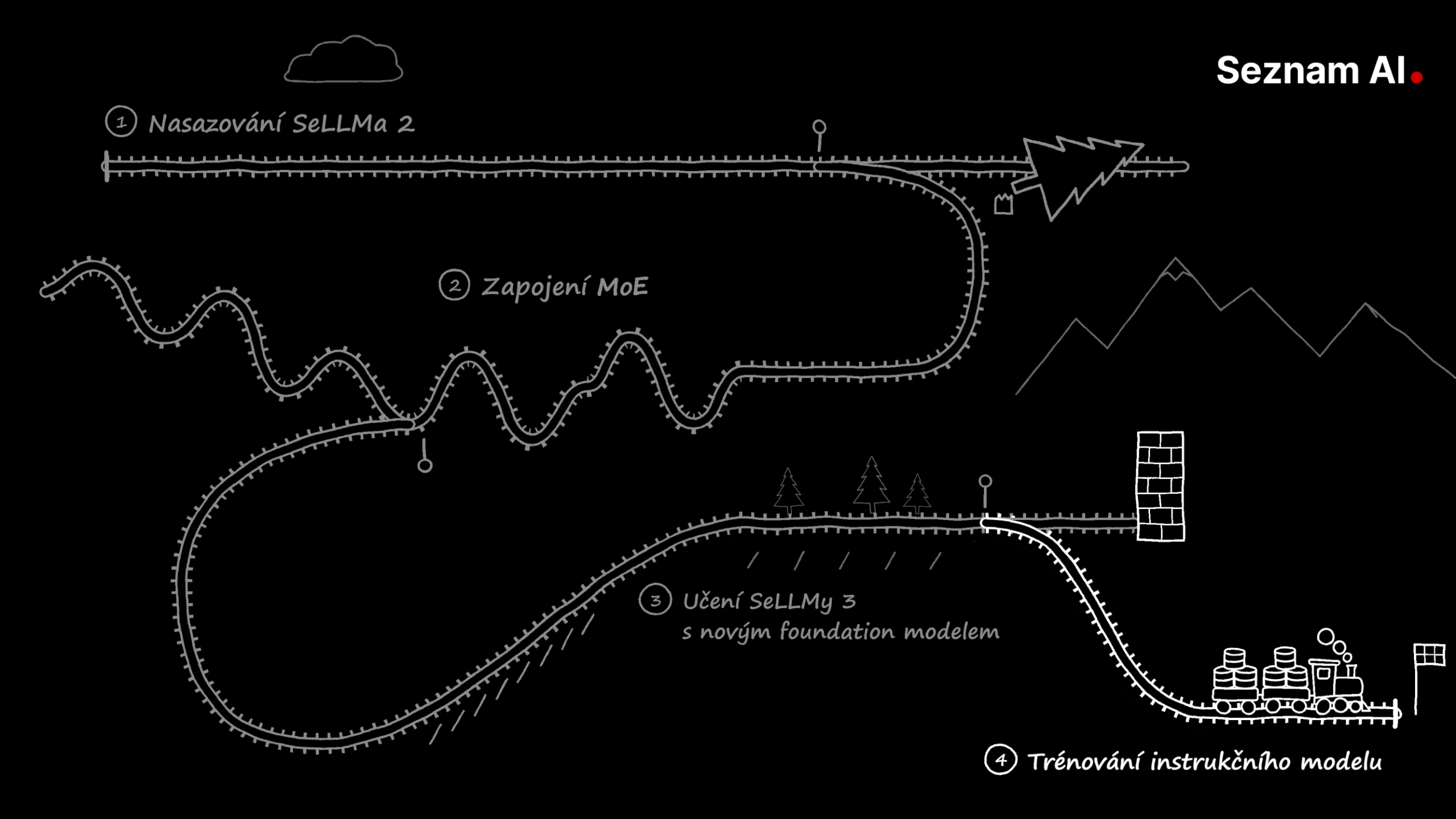
Base +
post-training







Seznam AI.


1 Nasazování SeLLMa 2
2 Zapojení MoE
3 Učení SeLLMy 3
s novým foundation modelem
4 Trénování instrukčního modelu

1

Čeština v cooldownu

Nerozbilo se to!

2

Víc češtiny

Vylepšujeme jazykové schopnosti.

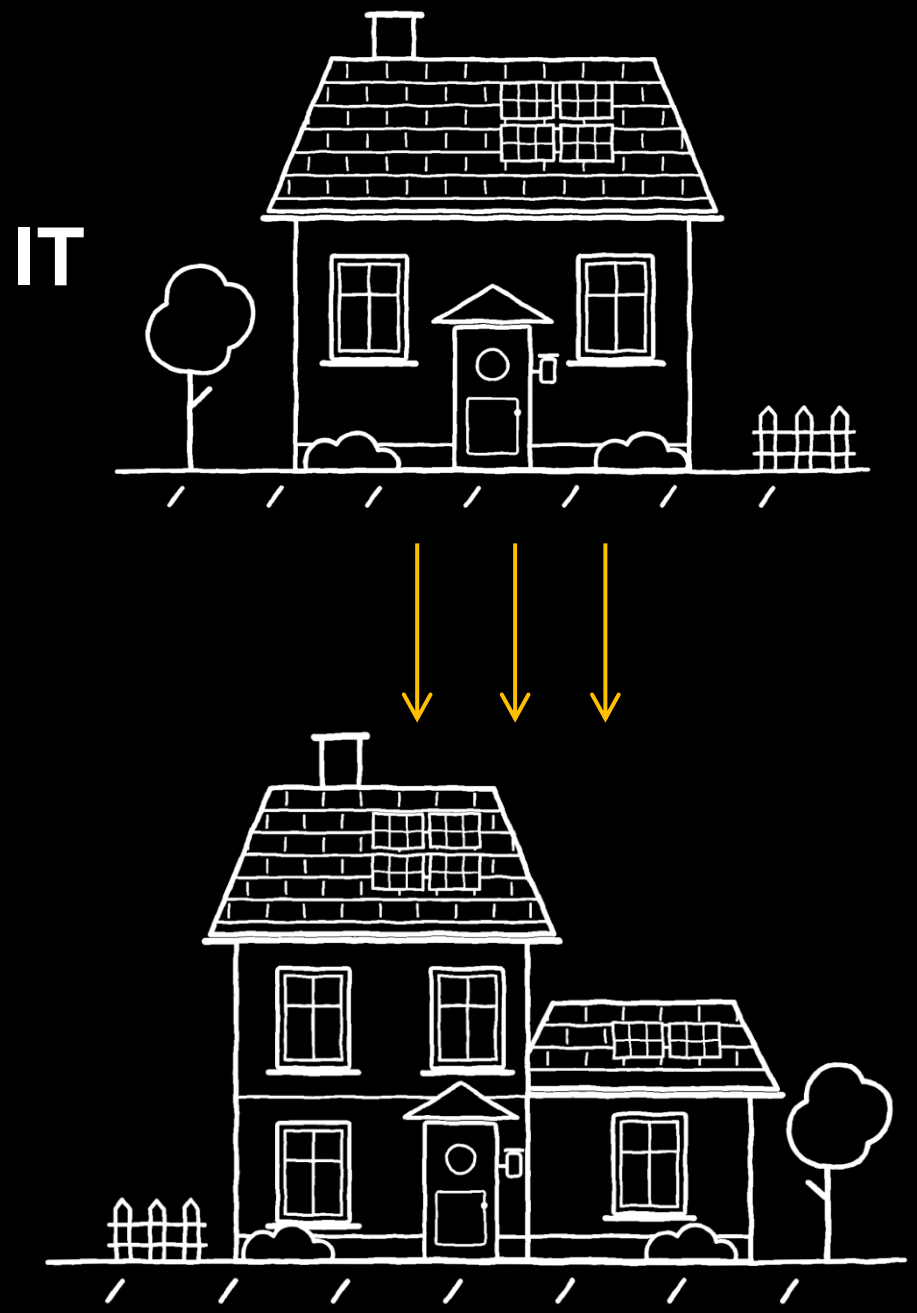
3

Kompletní post training instruct modelu

Průběžné mergování s původním instrukčním modelem.

Skvělá čeština i české znalosti.

Vylepšené vlastnosti.



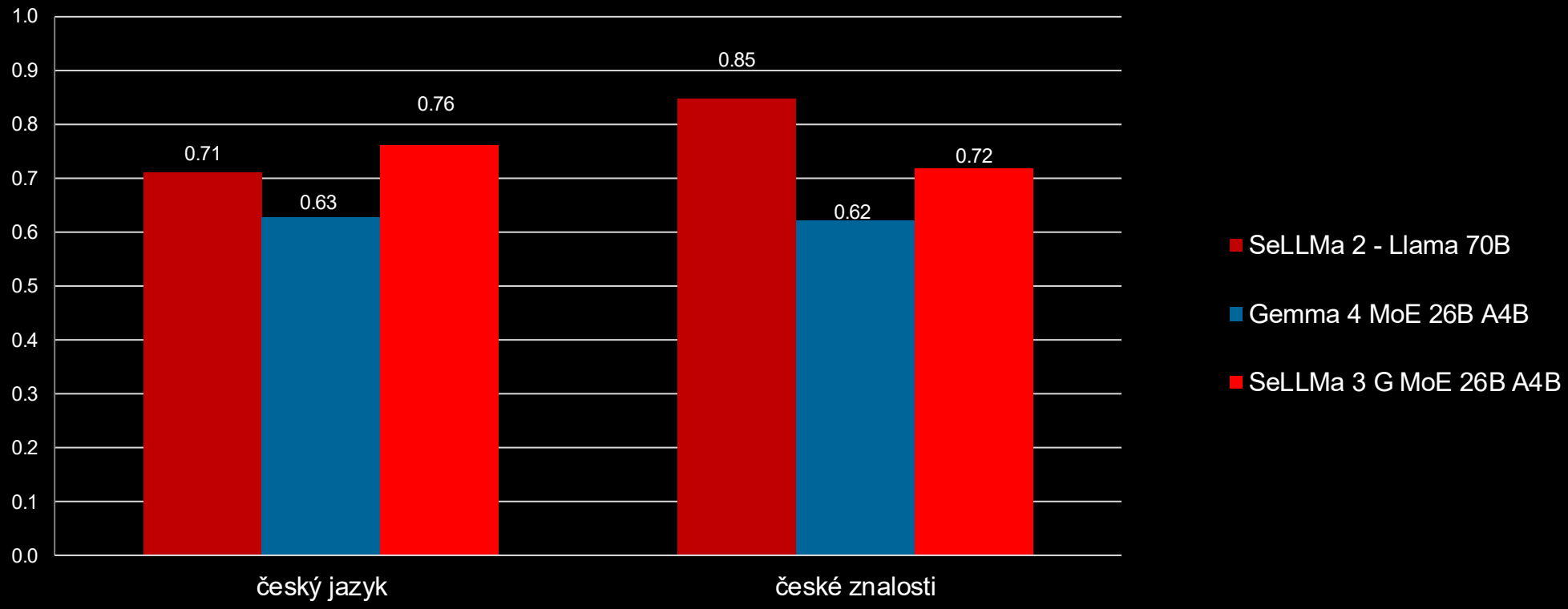
sellma 3

SeLLMa 3 odbavuje celý seznam požadavků



A má rychlost MoE Gemmy.

Měření českosti modelu



Použít, nebo stavět s námi

Použít

SeLLMa, TTS, ASR i Search API - odladěné na našem vlastním provozu.

solutions@firma.seznam.cz

Stavět s námi

Pojďte naučit Seznam AI něco ještě lepšího.

https://o-seznam.cz/kariera/seznam-ai/

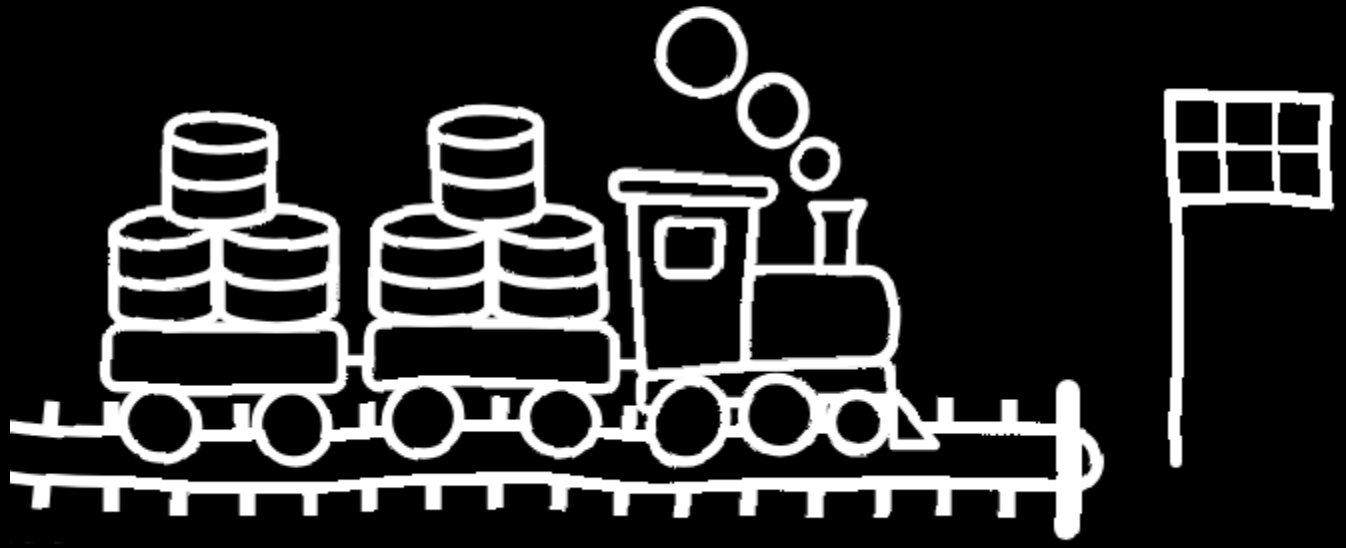


Pro příště

01 Provést realistický zátěžový test včas.

02 Chci-li opakovat minulé řešení, mám pro něj stejný materiál?

03 Nebát se přehazovat výhybky.



SEZNAM
MEETUP 26