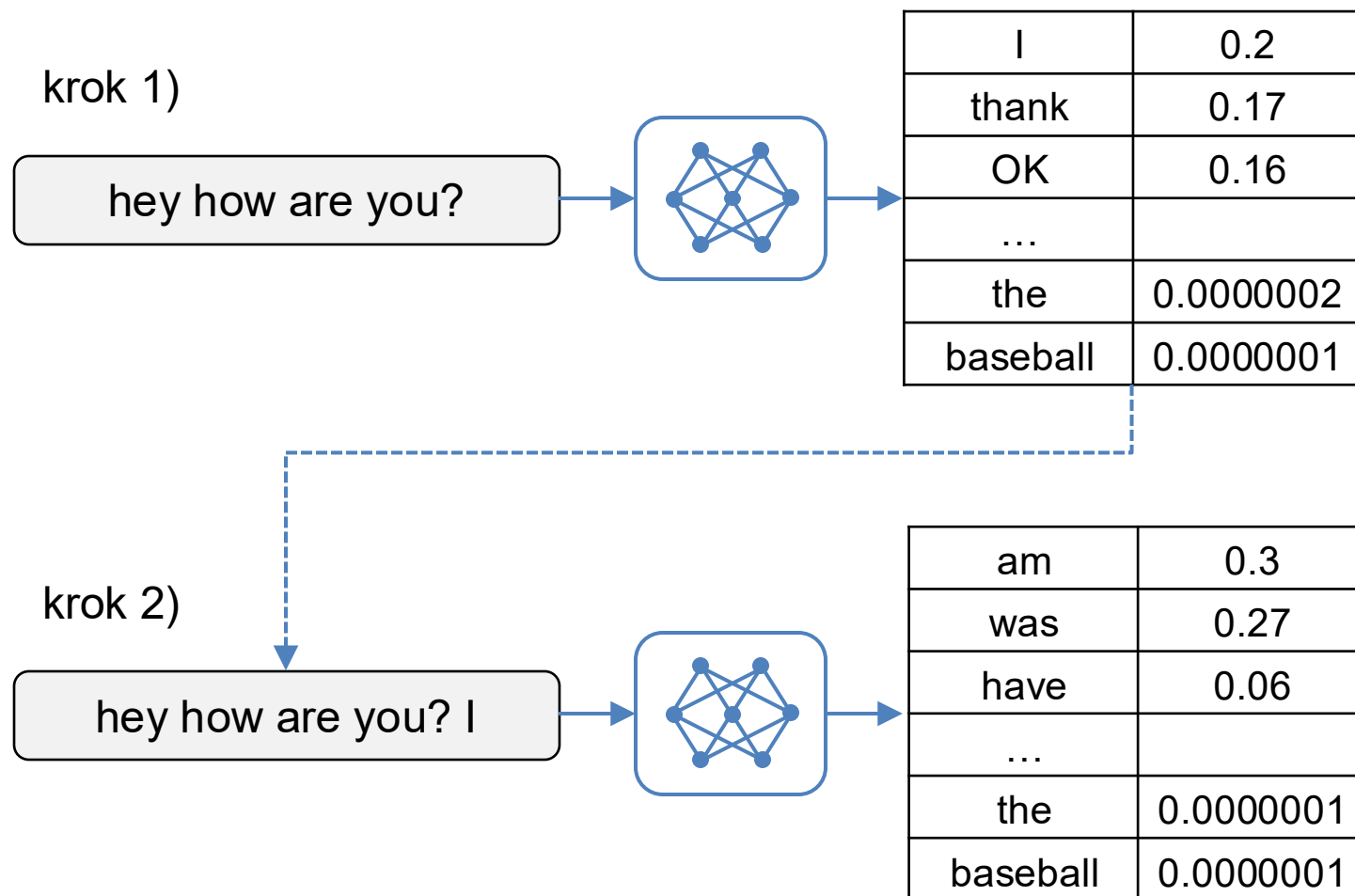


Veľké jazykové modely

Objavovanie znalostí v textoch

Peter Bednár

Jazykové modely



Aké úlohy vie „riešiť“ jazykový model?

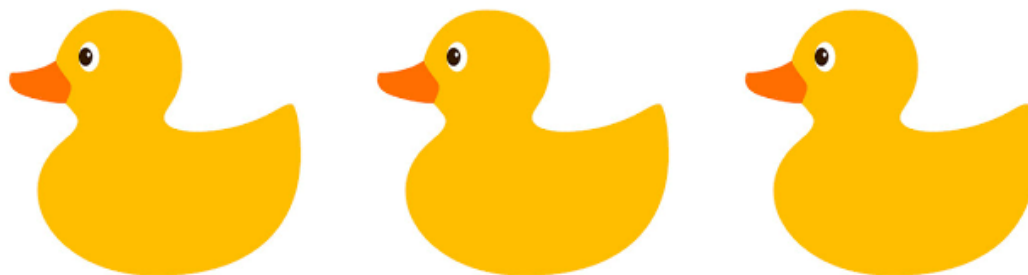
- **Vstup:** Kto bol Vincent Van Gogh?
- **Výstup:** Vincent Van Gogh bol holandský maliar 19 storočia.
- **Vstup:** Aký je sentiment vety „s ubytovaním sme boli spokojní“?
- **Výstup:** Veta má pozitívny sentiment.
- **Vstup:** Prelož vetu „kde je múzeum“ do taliančiny.
- **Výstup:** Dov'è il museo.

Aké úlohy vie „riešiť“ jazykový model?

- **Vstup:** Aké je ďalšie číslo v postupnosti 1 14 27 40?
- **Výstup:** Ďalšie číslo je 53, pretože rozdiel medzi nasledujúcimi číslami je vždy 13.
- **Vstup:** Idú dve kačiatka pred kačiatkom, dve kačiatka za kačiatkom a jedno kačiatko v strede. Koľko ide kačiatok?
- **Výstup:** ?

Aké úlohy vie „riešiť“ jazykový model?

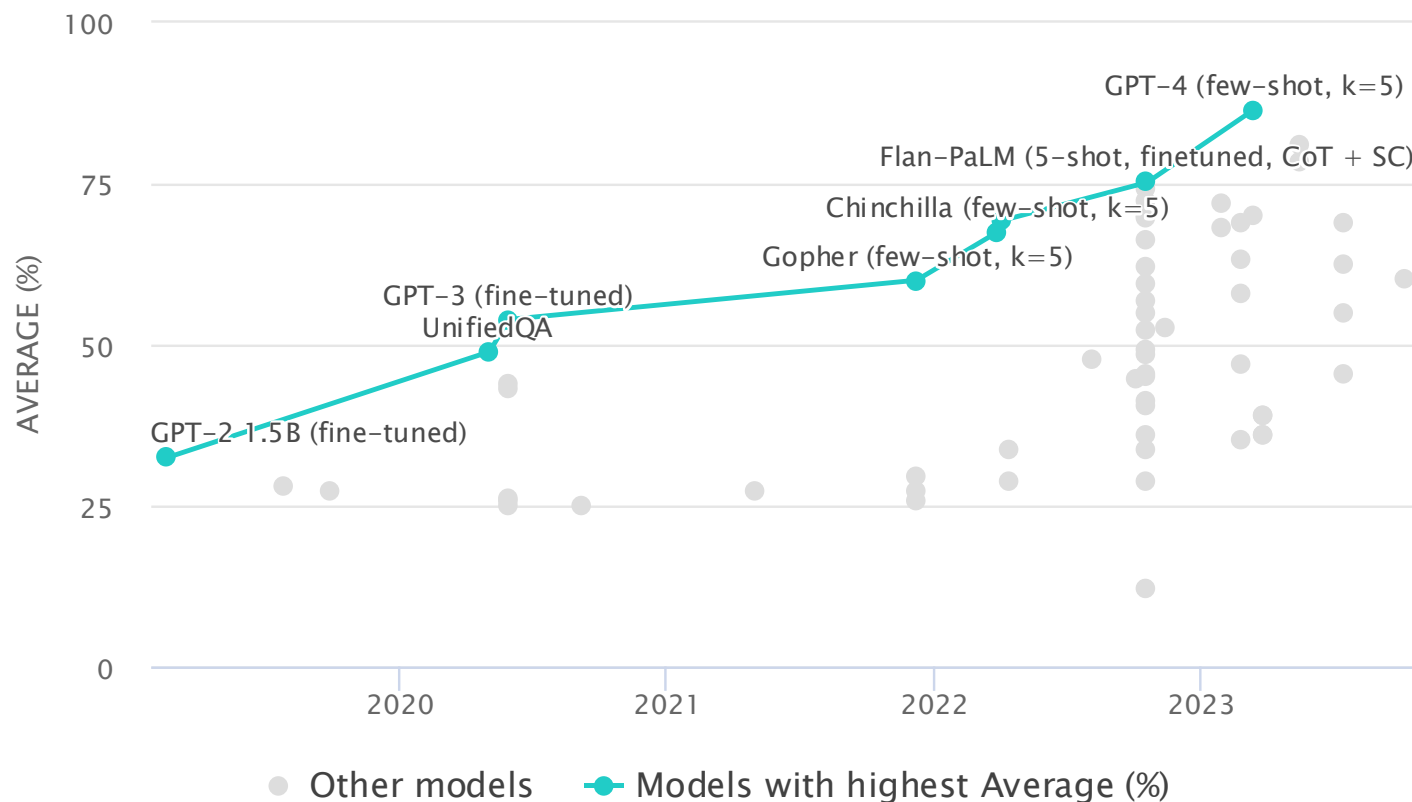
- **Vstup:** Idú dve kačiatka pred kačiatkom, dve kačiatka za kačiatkom a jedno kačiatko v strede. Koľko ide kačiatok?
- **Výstup:** Idú 3 kačiatka.



Testovanie jazykových modelov

- [HellaSwag](#)
 - Úlohy, ktoré vyžadujú jednoduché uvažovanie (*common sense*), jednoduché pre človeka ale zložité pre predchádzajúce UI systémy
- [MMLU](#)
 - 57 úloh zameraných na všeobecné znalosti z rôznych oblastí ako je napr. matematika, história, počítačové vedy, právo, náboženstvá atď.
- [TruthfulQA](#)
 - Sklony generovať nepravdy, ktoré sa často vyskytujú online
- Zjednotené metódy testovania
 - [HuggingFace Open LLM Leaderboard](#)
 - [Stanford HELM](#)

Testovanie jazykových modelov

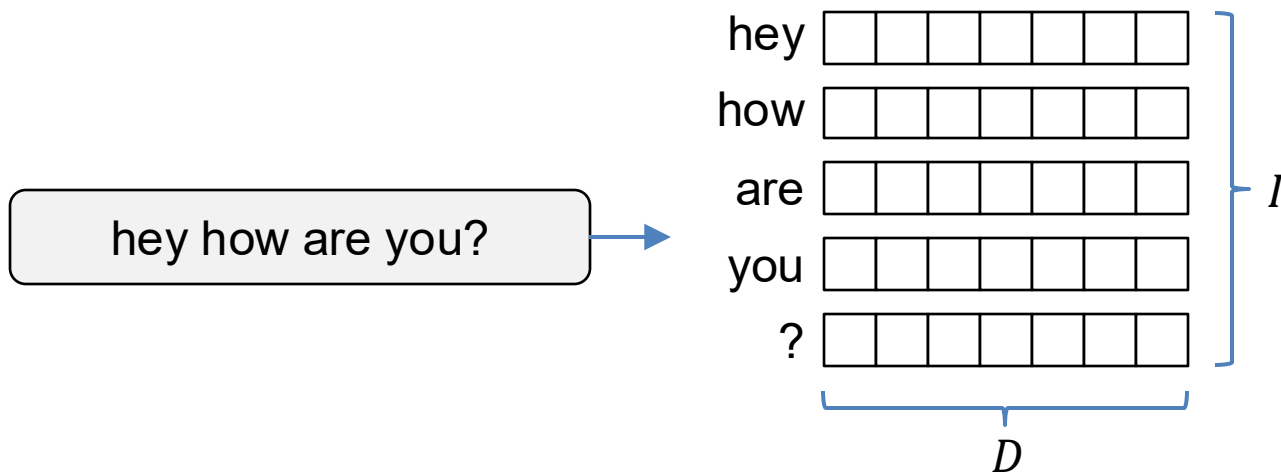


Transformer



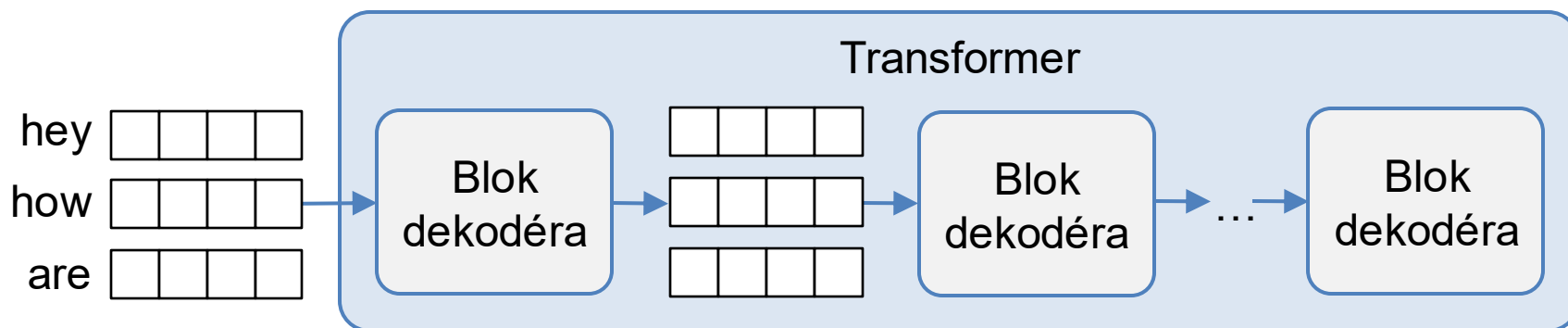
Tokenizácia a Vektorová reprezentácia slov

- Text sa rozdelí na tokeny
 - Slovník častých slov a podreťazcov
 - *Elvégezhetitek* `El` `#vé` `#ge` `#zhet` `#ite` `#k`
- Každému tokenu je priradený číselný vektor – *embedding*
 - Parametre, ktoré sa menia pri učení

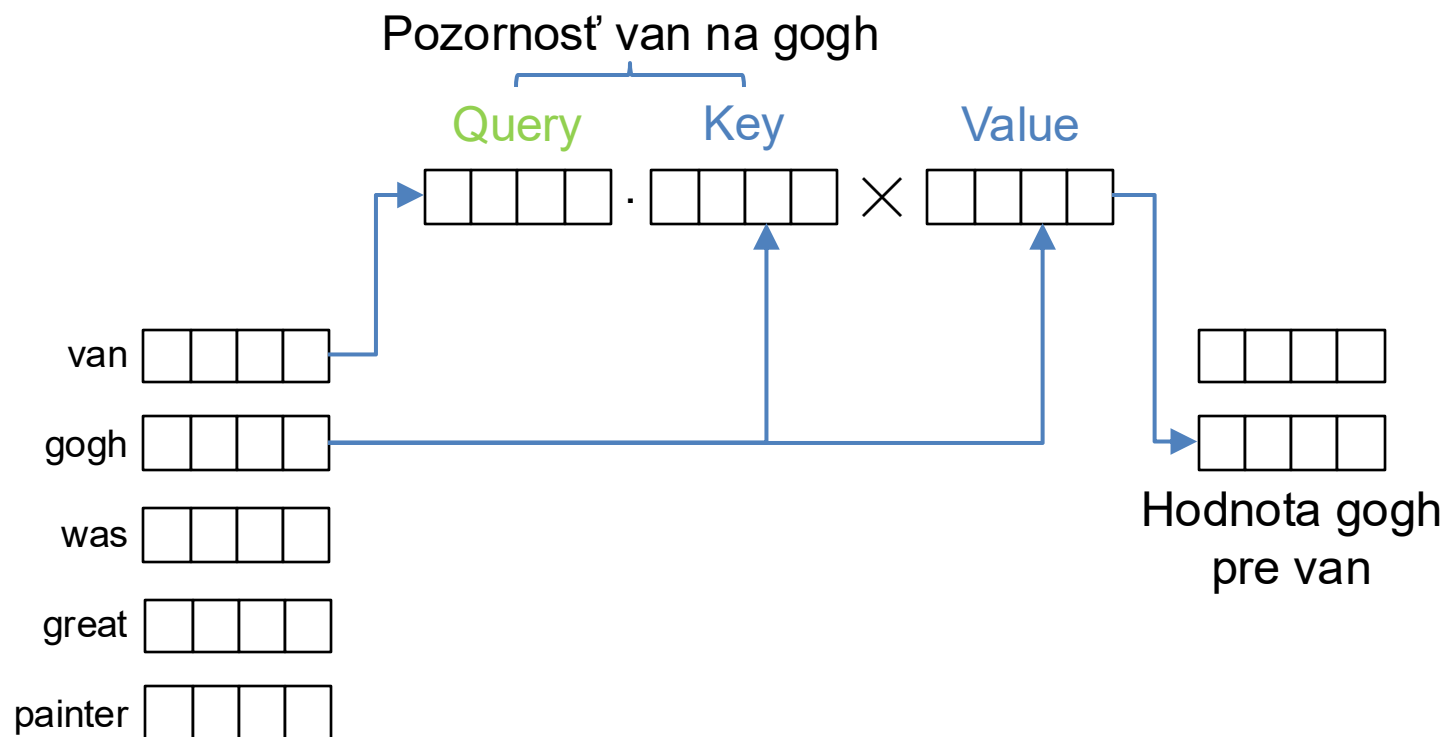


Transformer

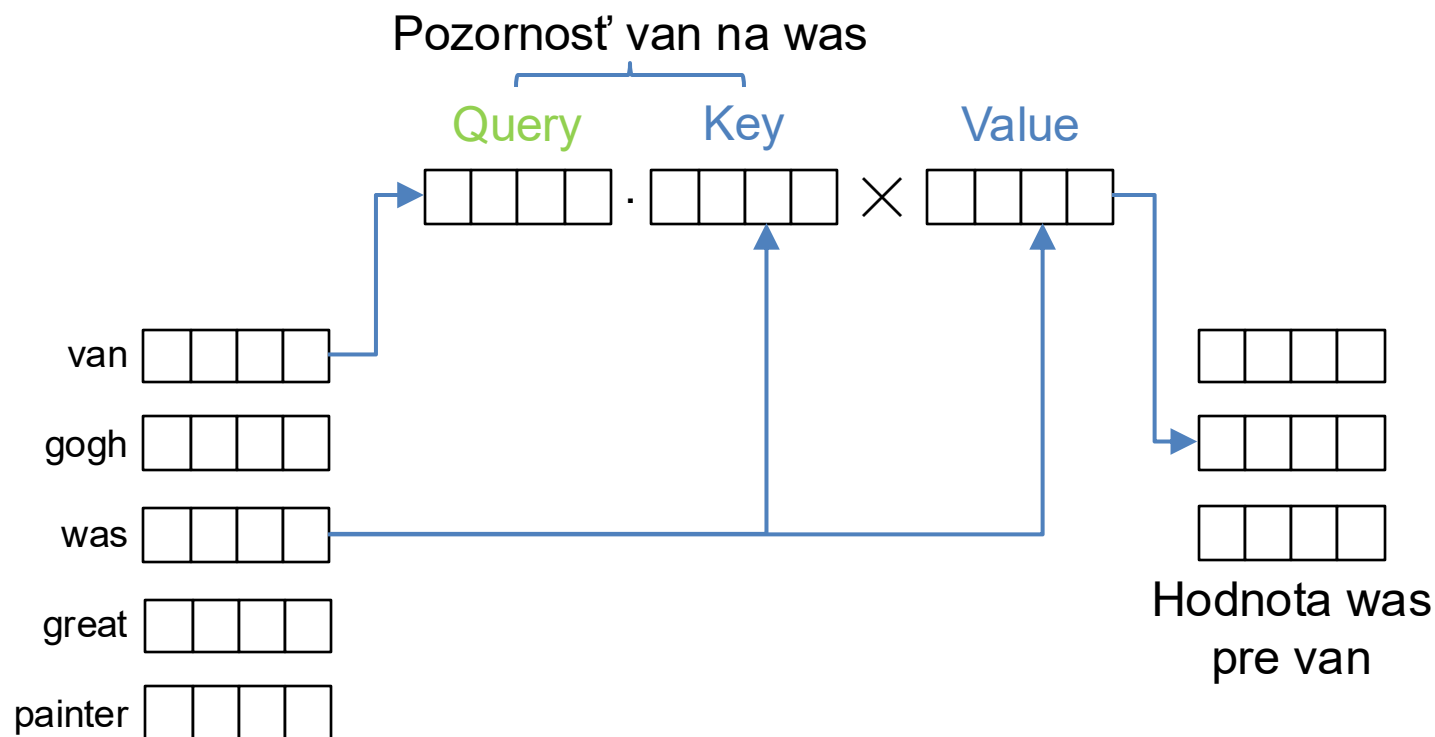
- Hlboká sieť zložená z blokov, ktoré dekodujú vstupnú maticu tokenov



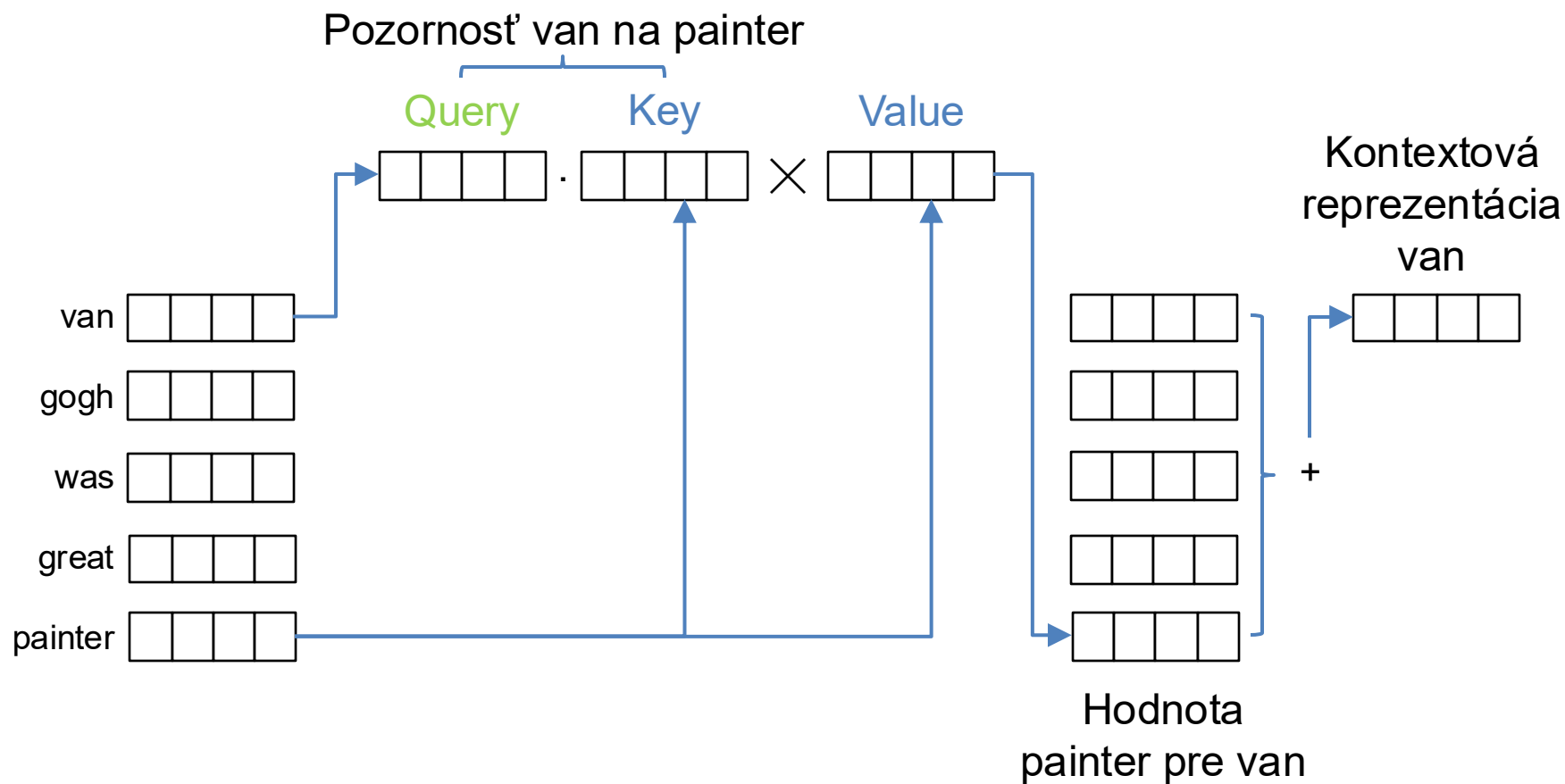
Pozornostná vrstva (2)



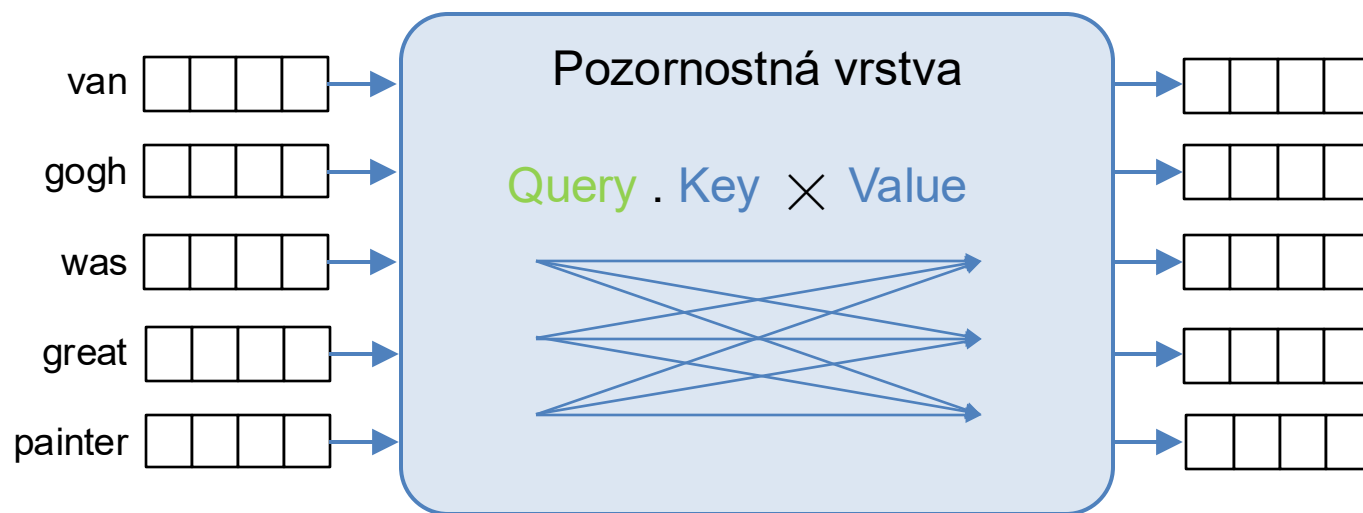
Pozornostná vrstva (3)



Pozornostná vrstva (4)



Pozornostná vrstva (5)



Analýza sentimentu

Vážený predajcovia <oddelenie predaja>

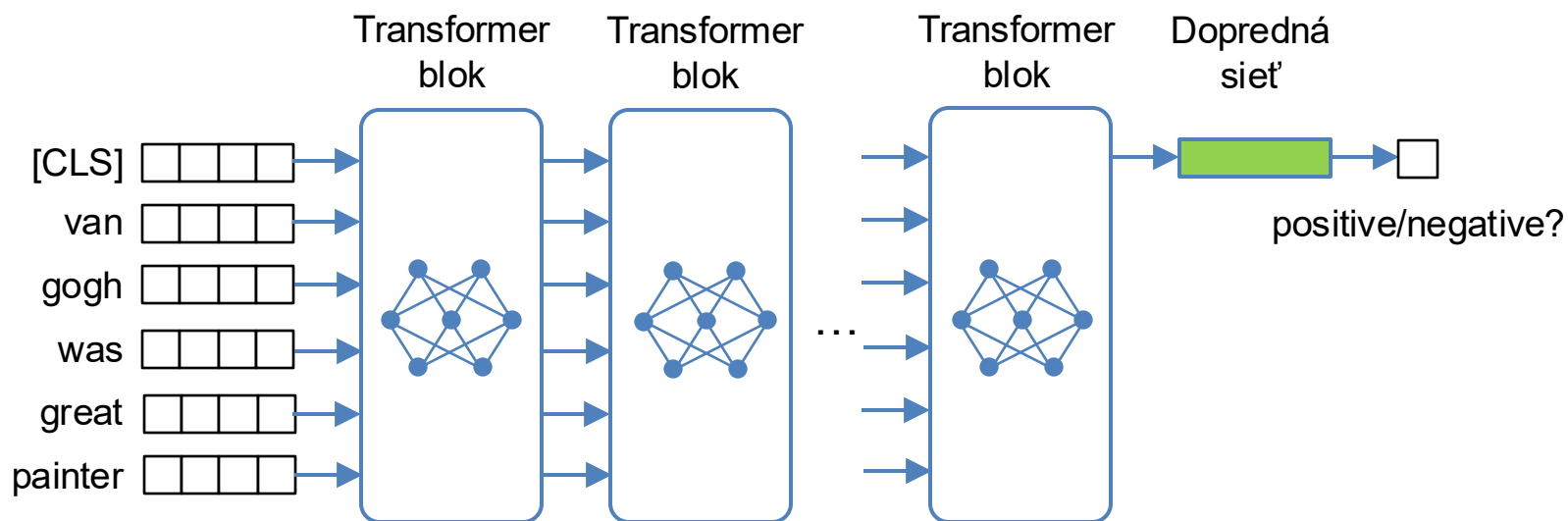
Včera som navštívil <konkurenčnú predajňu>. Majú skvelý výber, najlepšie ceny, a nápomocný personál.

Vaša firma je nanič.

S pozdravom ...

- Správne klasifikovanie sentimentu vyžaduje vo všeobecnosti rozšírený kontext a zachovanie syntaktických väzieb

Klasifikácia/analýza sentimentu



Extrahovanie informácií

- Vety z medicínskych článkov

Stress is associated with migraines

Stress can lead to loss of magnesium

Calcium channel blockers prevent some migraines

Magnesium is a natural calcium channel blocker

Spreading cortical depression (SCD) is implicated in some migraines

High levels of magnesium inhibit SCD

Migraine patients have high platelet aggregability

Magnesium can suppress platelet aggregability

Extrahovanie informácií

- Extrahovanie entít

Stress is associated with migraines

Stress can lead to loss of magnesium

Calcium channel blockers prevent some migraines

Magnesium is a natural calcium channel blocker

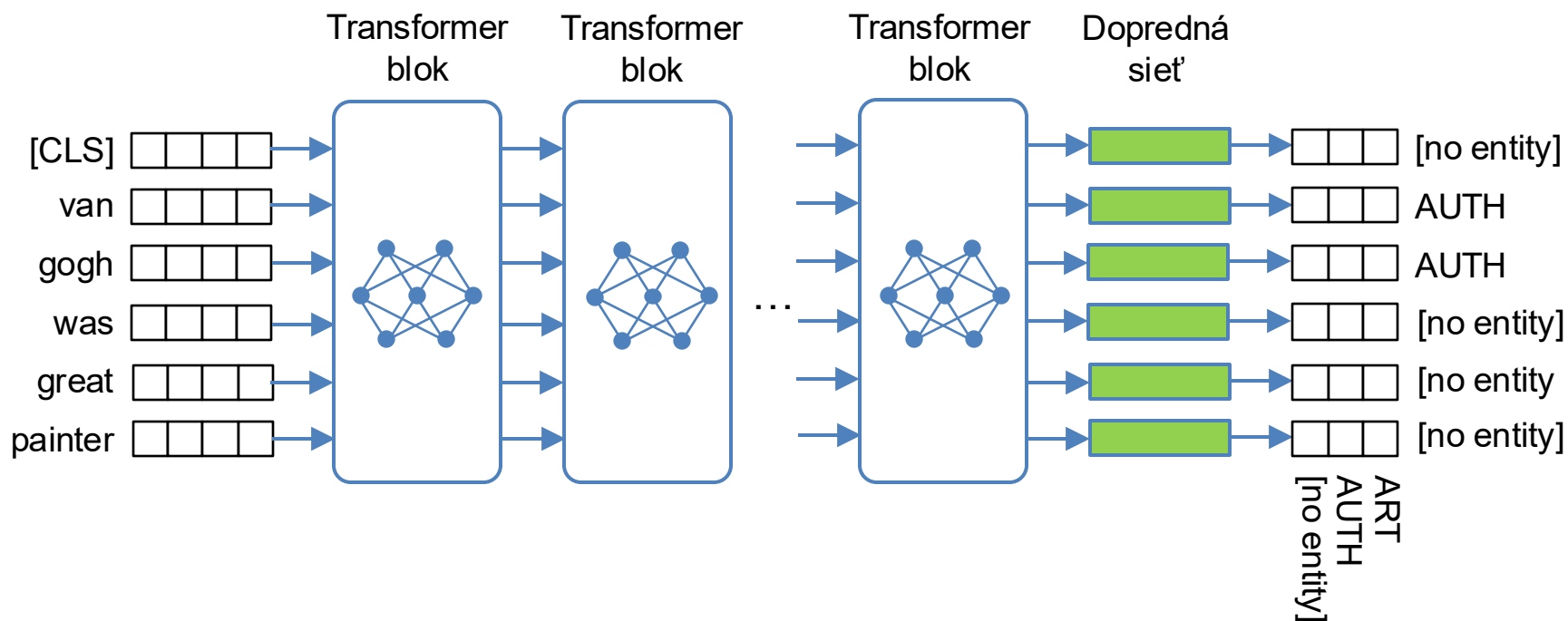
Spreading cortical depression (SCD) is implicated in some migraines

High levels of magnesium inhibit SCD

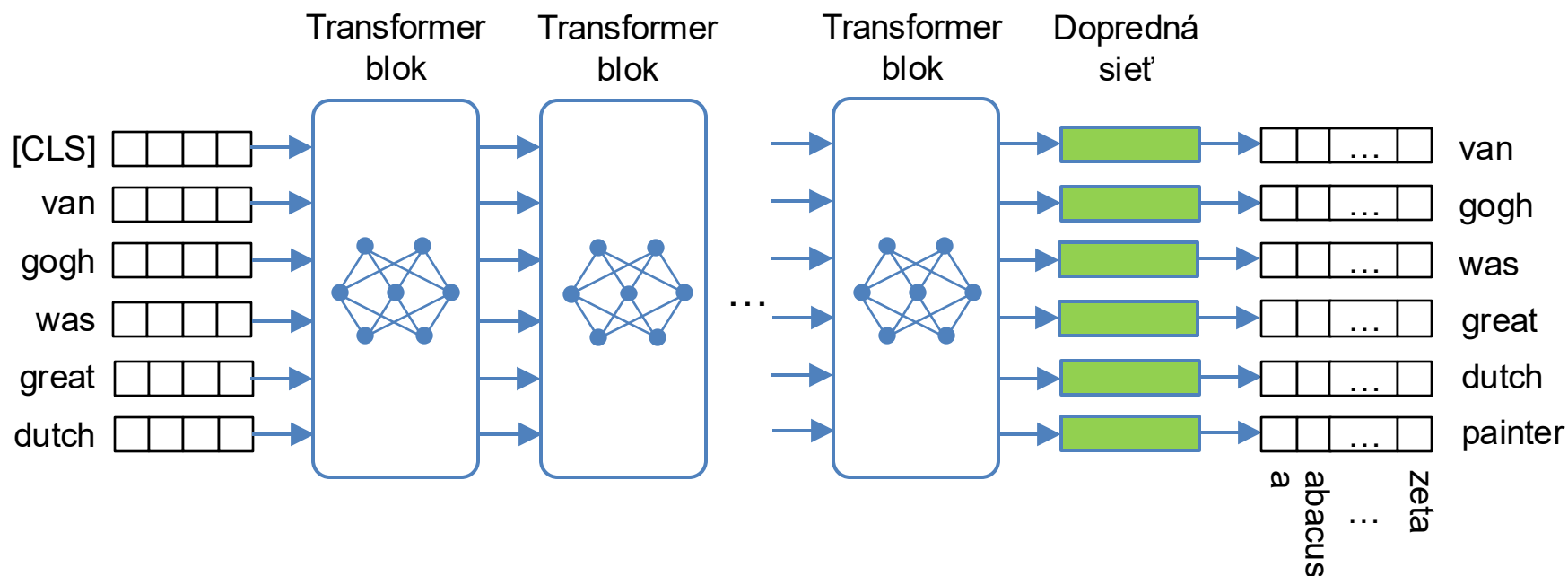
Migraine patients have high platelet aggregability

Magnesium can suppress platelet aggregability

Extrahovanie informácií



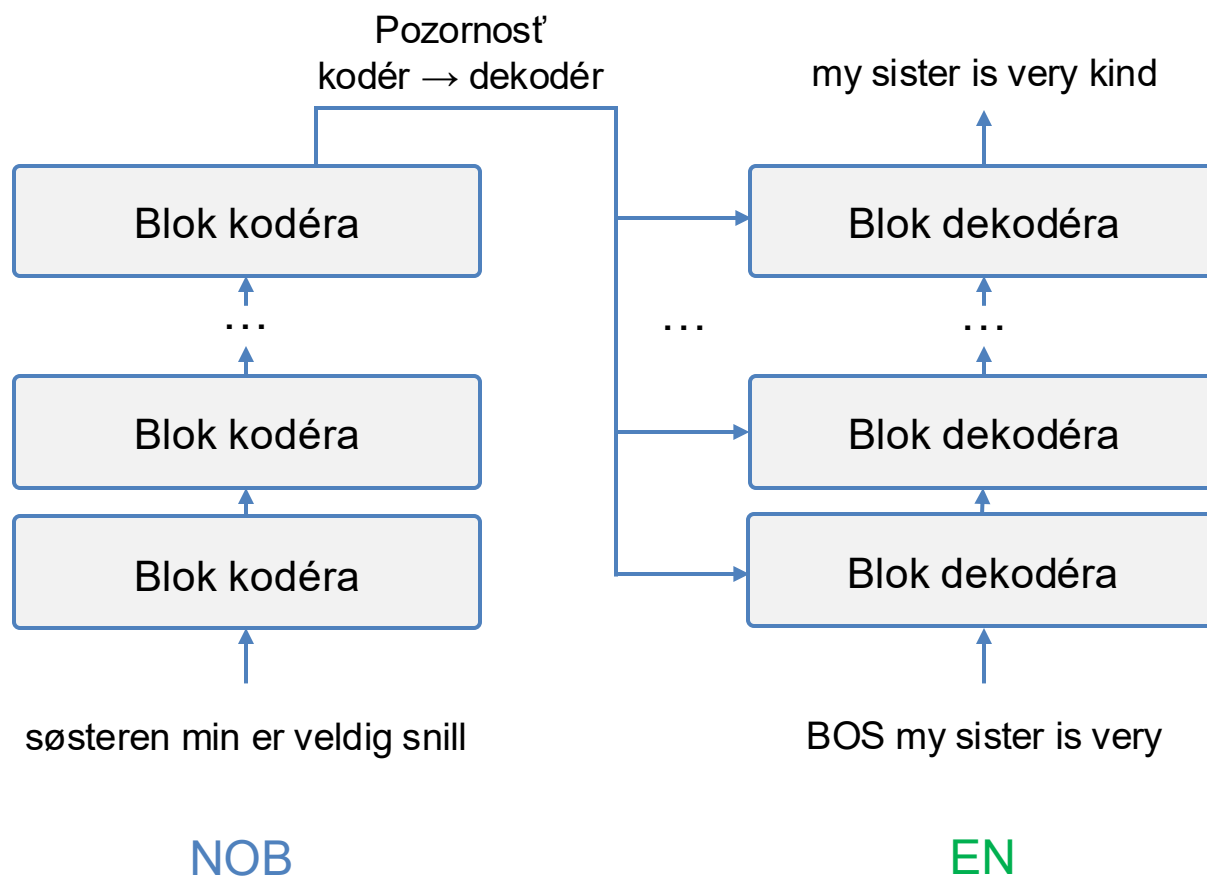
Jazykový model



Automatický preklad

Kodér zdrojového jazyka

Dekodér cieľového jazyka



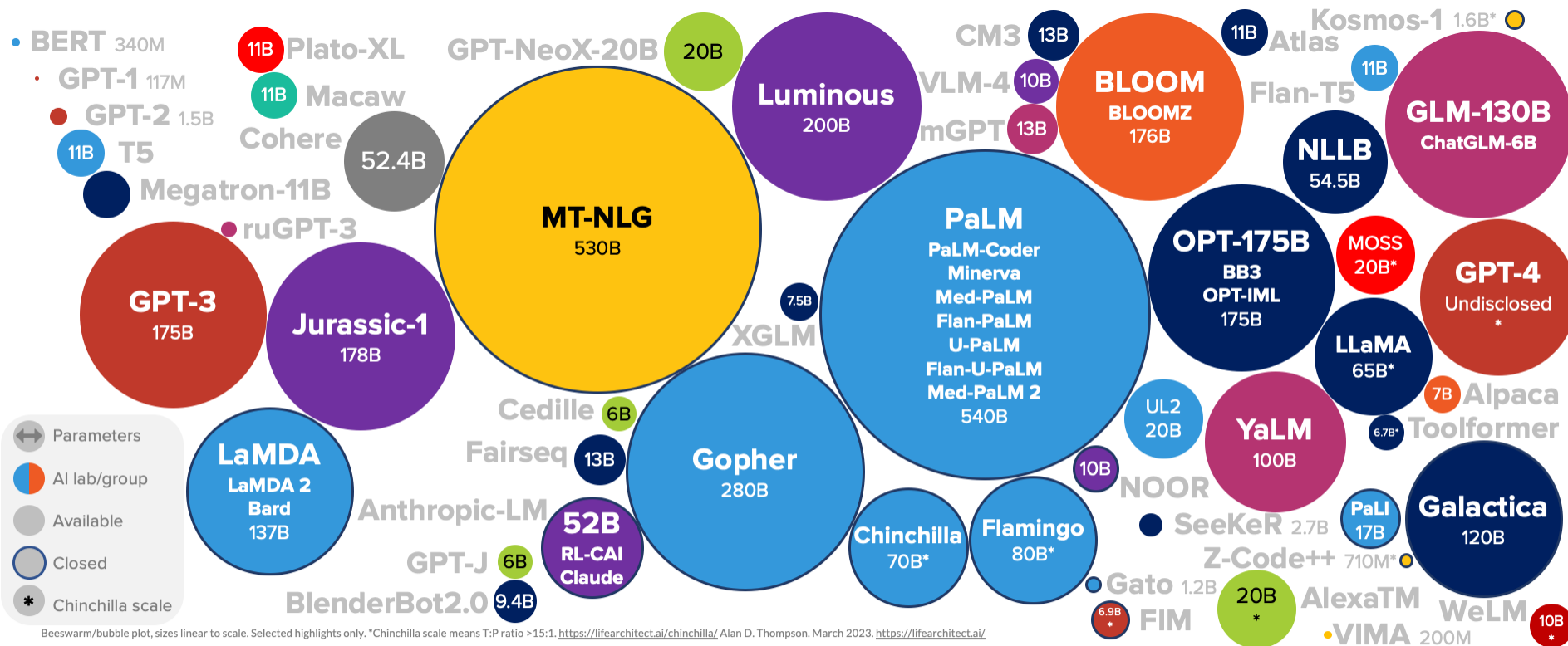
BERT

- Pôvodný model architektúry pre Transformer siete
- Slovník 30 000 tokenov, 1024 rozmerné vektory (*embeddings*)
- 24 blokov, 16 hláv (tzn. matice $\Phi_{q,h}$, $\Phi_{k,h}$ a $\Phi_{v,h}$ majú rozmer 1024×64), skrytá vrstva doprednej siete má 4096 neurónov
- Celkový počet parametrov cca 340 mil.



Veľké jazykové modely

M – 1 000 000, B – 1 000 000 000

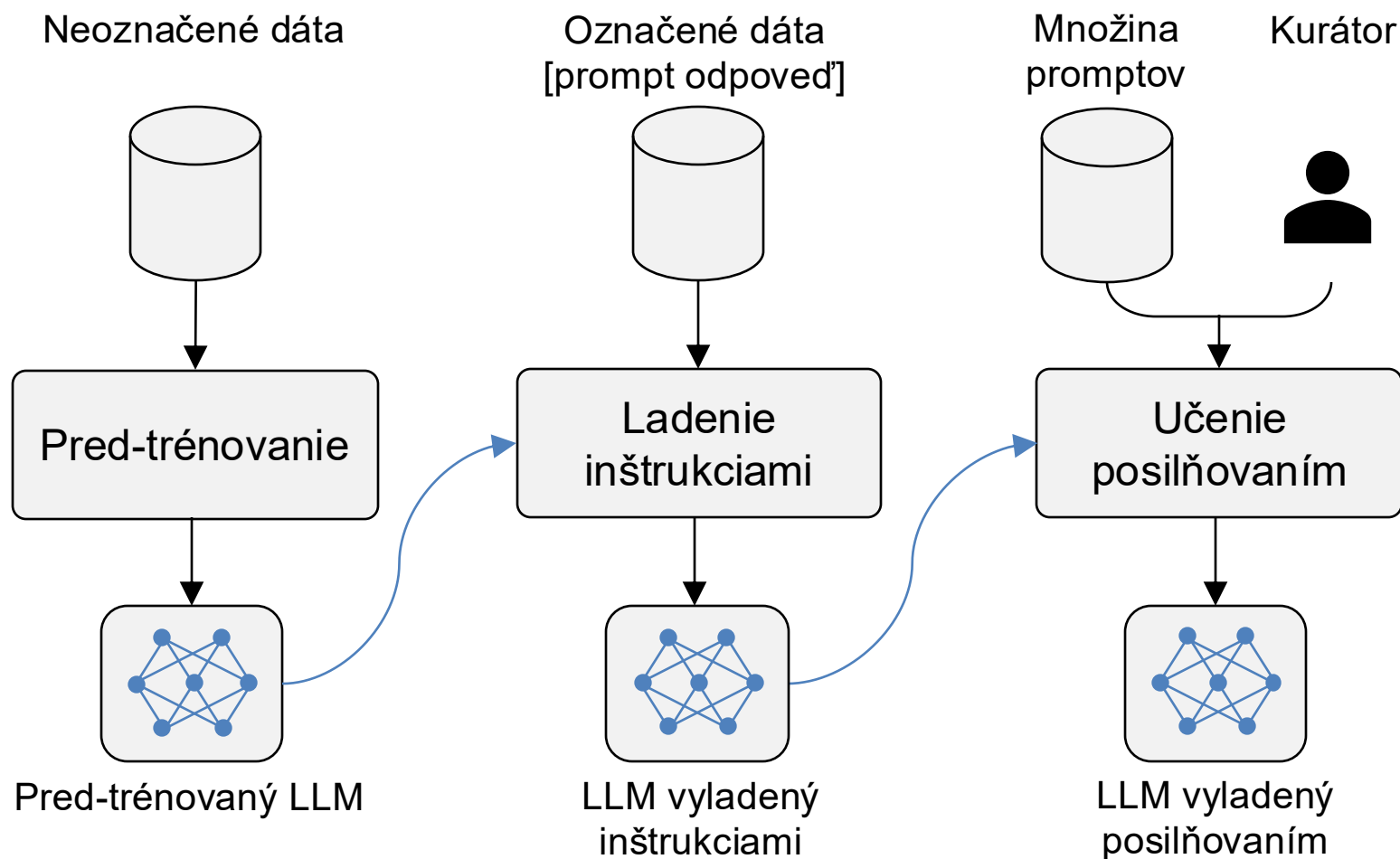


LifeArchitect.ai/models

Trénovanie veľkých jazykových modelov

- Trénovanie modelu LLaMA – 65 miliárd ($\approx 65\,000\,000\,000$) parametrov vyžaduje 21 dní, 2048 A100 grafických kariet (každá má 80 GB pamäte)
- Odhadované náklady pre GPT3 boli \$4.6M
- Open source modely
 - Falcon
 - OpenLLaMA / LLaMA 2
 - Alpaca
 - MPT
 - FastChat-T5
 - Bloom

Trénovanie veľkých jazykových modelov

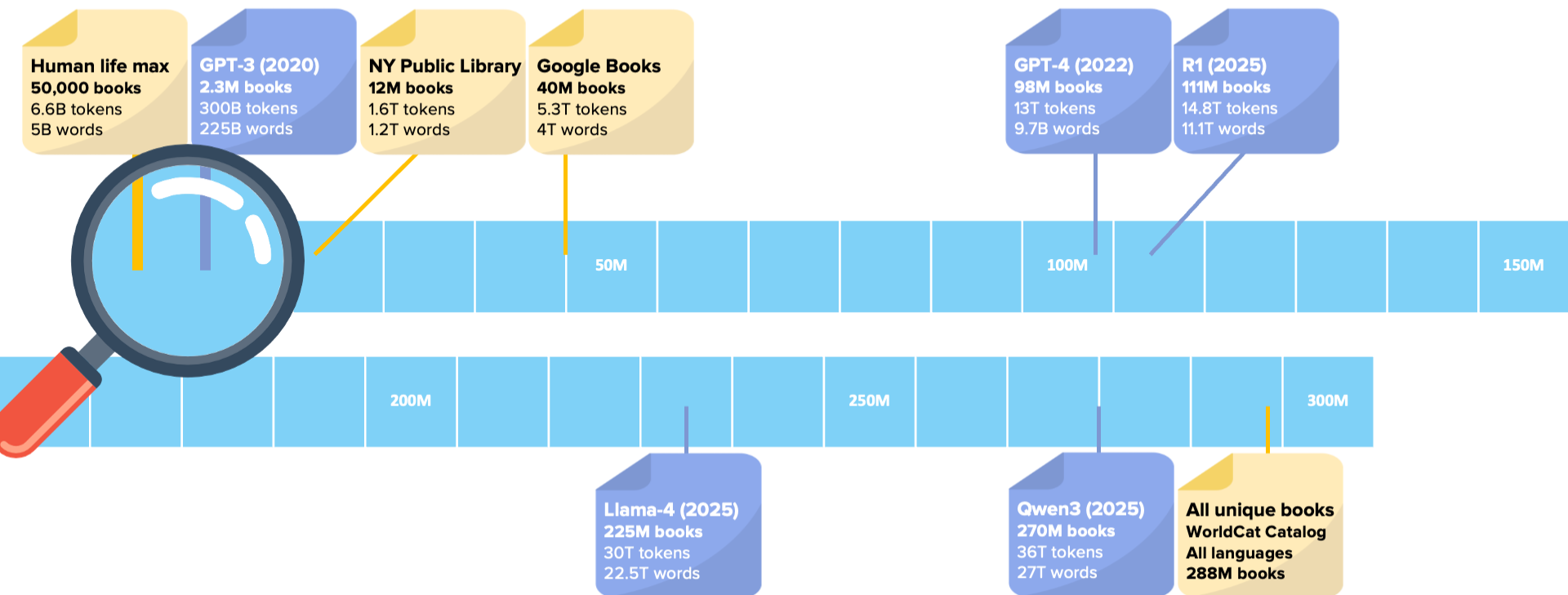


Pred-trénovanie

- Samo-kontrolované učenie
 - Predikcia nasledujúceho tokenu
 - Cieľom je hlavne zachytiť jazykové vlastnosti
- Rozsiahle textové dáta
 - Web, Wikipédia, knihy, ...
- Rozsah dát
 - PALM – 780 miliárd tokenov
 - RedPajama – 1.2 biliónov tokenov, 2.67 TB dát

Veľkosť tréningových dát

AI TRAINING DATA COMPARED TO TOTAL BOOKS (2025)



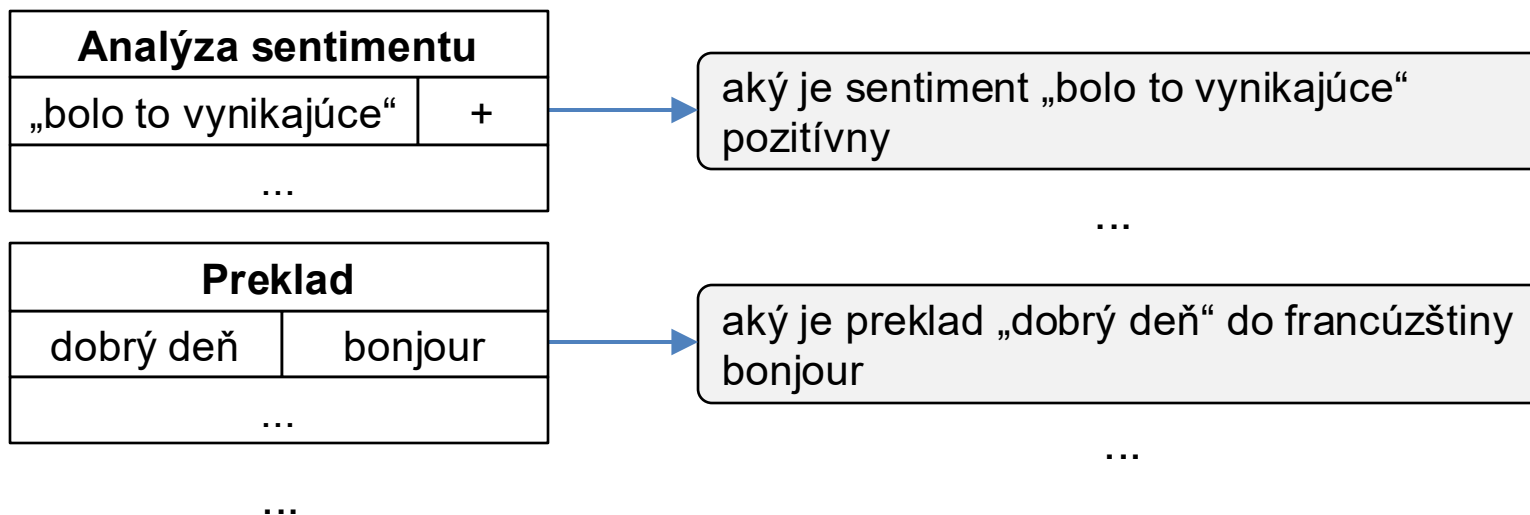
Training data sizes converted from tokens. 133K tokens = 100K words = 1 book equiv. Text datasets only. For simplicity, books are shown here in place of web data, code data, other data. Alan D. Thompson. 2025.



Ladenie inštrukciami (1)

1. Využitie existujúcich označených dátových množín

- Rôzne typy úloh (klasifikácia, preklad, zodpovedanie otázok, ...)
- Použitím šablón sa prepíšu do tvaru [prompt - očakávaný výstup]
- Učenie prebieha jednotne, predikovaním nasledujúceho tokenu



Ladenie inštrukciami (2)

2. Ručným zadaním požadovaného výstupu pre predchádzajúce prompty
 - InstructGPT/OpenAI – 13 000 promptov, 40 anotátorov
3. Využitie LLM na generovanie trénovacích dát
 - Anotátor zadá iba jeden prompt pre danú úlohu ako príklad, pred-trénovaný LLM vygeneruje množinu promptov a odpovedí, ktoré sa použijú na ladenie

Ladenie inštrukciami – príklady modelov

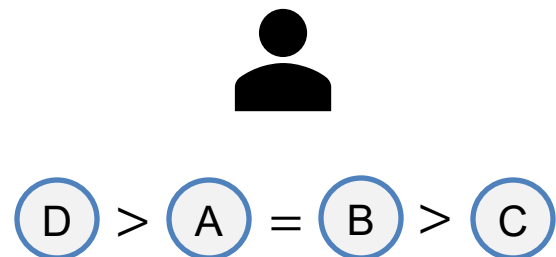
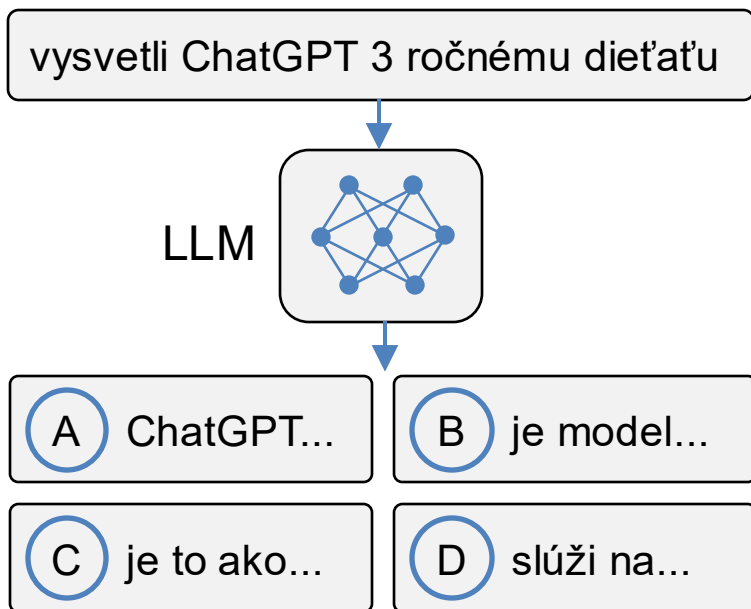
- Alpaca – 52K vygenerovaných inštrukcií na LLaMA modely, podobný výkon ako GPT-3, oveľa menší model (7B oproti 176B), ľahko reprodukovateľný (600\$)
- Flan-PaLM – 1.8K inštrukcií, lepší než PaLM (v priemere o 9.4%), ale vyžaduje na tréning iba 0.2% času

Učenie posilňovaním

- Snažíme sa zarovnať model (*alignment*) tak aby produkoval správny výstup
- Anotátor môže ohodnotiť celkový výstup, ktorý sa však generuje po tokenoch – ako propagovať chybu výstupu a určiť gradient pre jednotlivé tokeny?
- **Metódy učenia posilňovaním** (*reinforcement learning*)
 - Algoritmus vygeneruje postupnosť akcií (pre LLM postupnosť tokenov)
 - Určí sa ohodnotenie celej postupnosti
 - Metóda učenia posilňovaním určí ako sa má zmeniť prediktívna funkcia

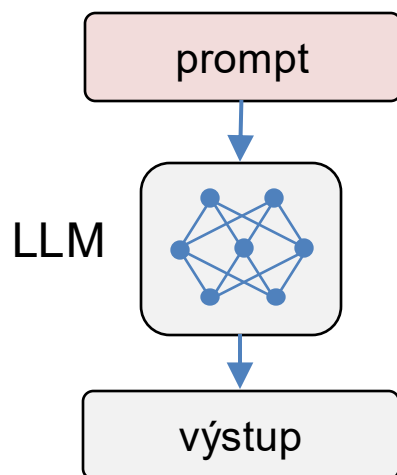
Učenie ohodnotenia (1)

1. Pre zvolené prompty sa nasampluje viacero odpovedí
2. Anotátor určí ich poradie od najlepšej po najhoršiu

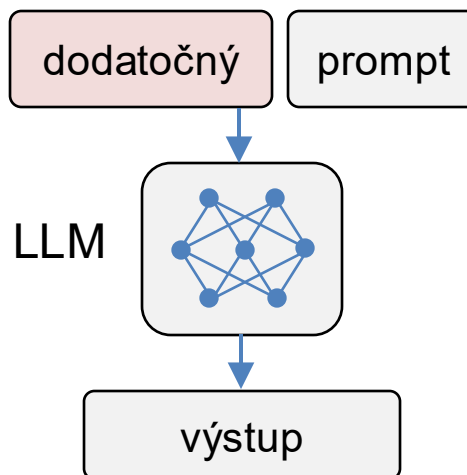


Použitie veľkých jazykových modelov

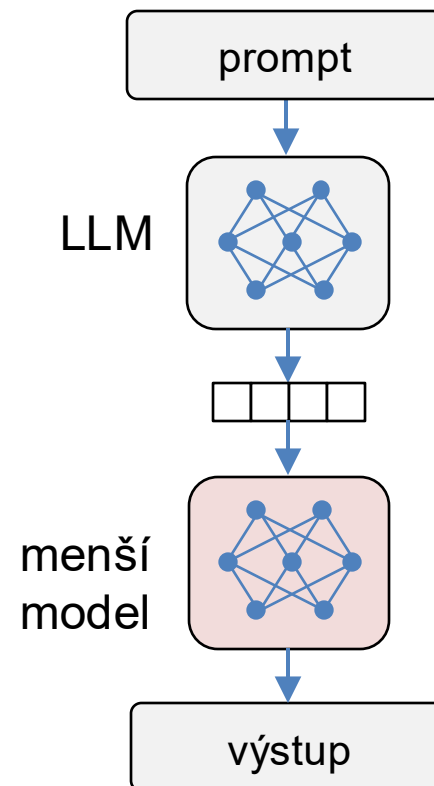
1 Vytváranie promptov



2 Ladenie promptov



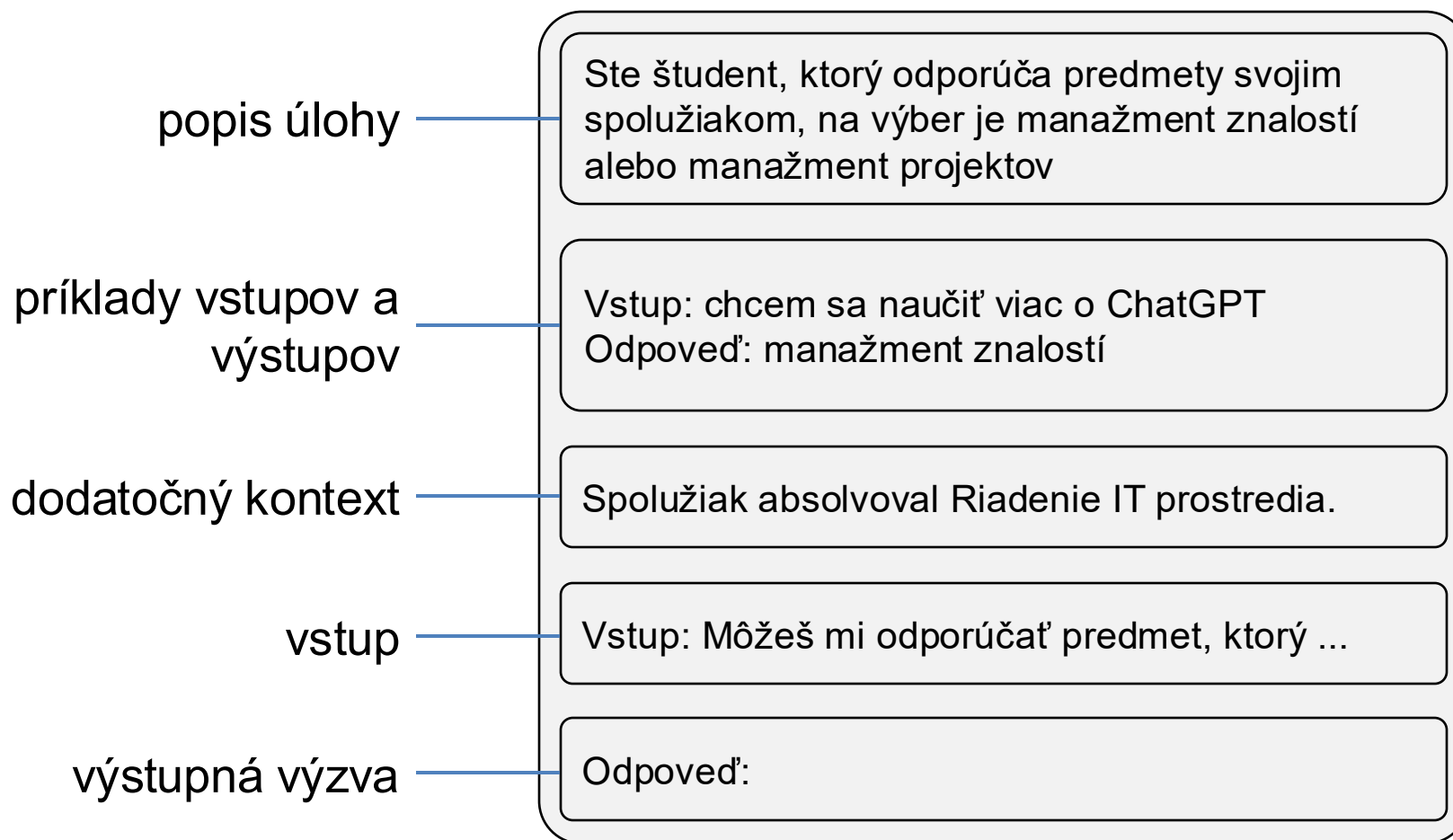
3 Využitie reprezentácie



Vytváranie/ladenie promptov

- *Few-shot* – uvidíme pár príkladov aj s odpoveďami a cieľovú otázku
- Generovaná znalosť – najprv sa opýtame všeobecnú otázku z danej domény a vygenerovanú odpoveď použijeme ako kontext pre špecifickú otázku
- Samplovanie – vygenerujeme viac odpovedí a použijeme majoritné hlasovanie pre získanie finálnej odpovede
- *Chain-of-thoughts* – v prompte uvidíme viac krokov ako zodpovedať otázku

Príklad promptu



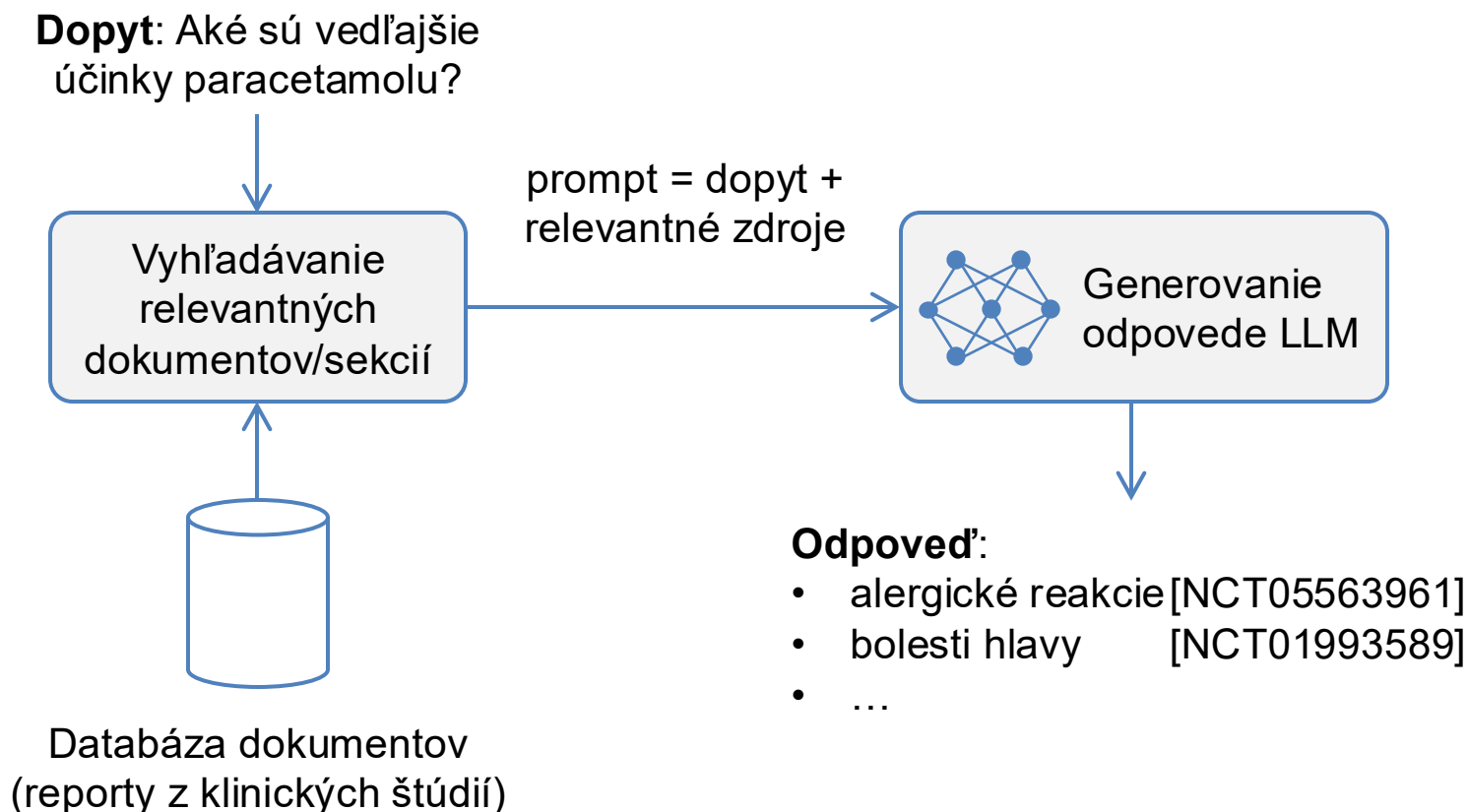
Aké otázky vie LLM zodpovedať?

- Nedostatky LLM pri zodpovedaní otázok
 - LLM je naučený na statickej dátovej množiny a nemá prístup k aktuálnym informáciám
 - Halucinovanie (zvlášť pre informácie chýbajúce v trénovacích dátach)
 - Nemá prístup k proprietárnym dátam
 - Nevie spoľahlivo poskytnúť zdroj dát z ktorých odvodil odpoveď

RAG - Retrieval-Augmented Generation (1)

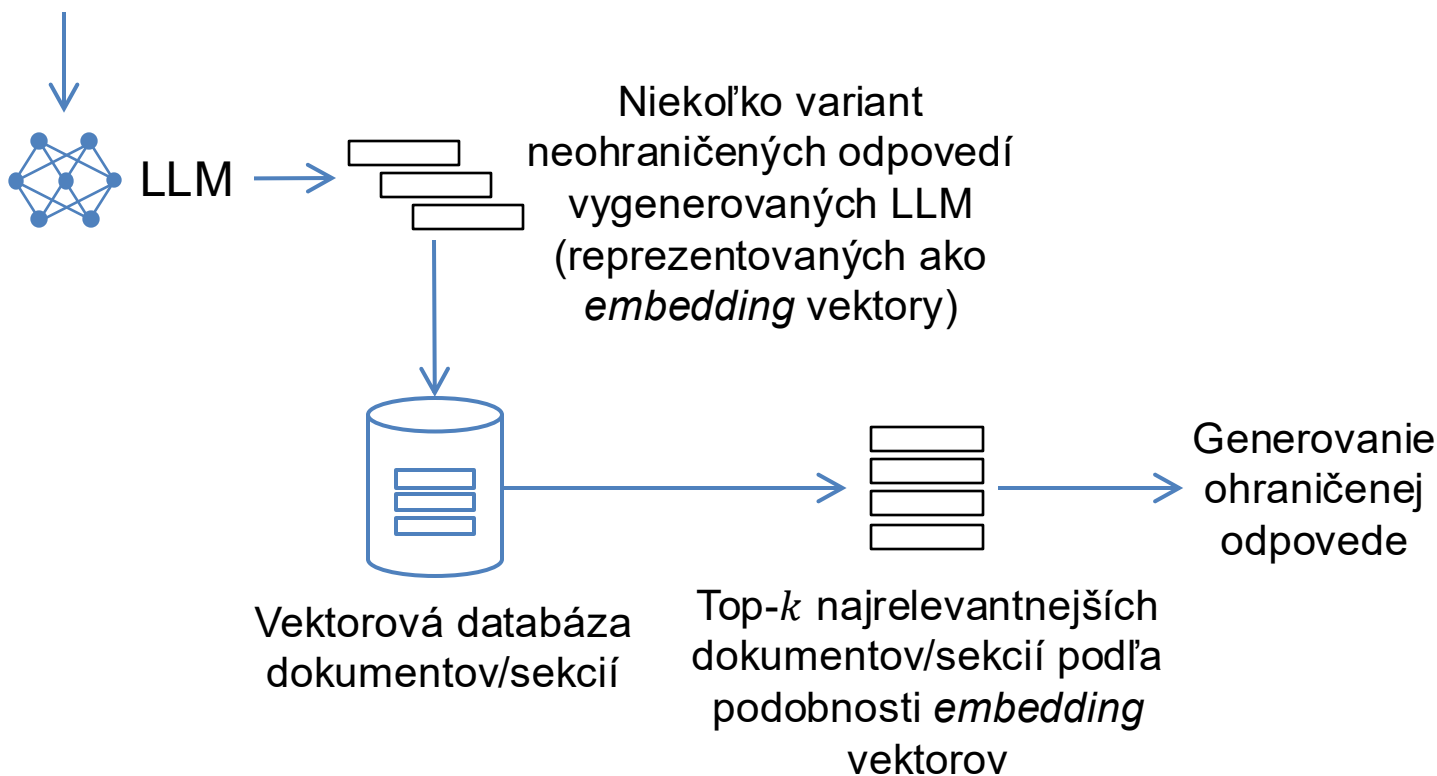
- Prepojíme databázu dokumentov/znalostnú bázu s LLM
- Vo vstupnom prompte ohraničíme generovanie odpovede v LLM iba na informácie z relevantných dokumentov vyhládaných v báze znalostí
- Stačí aktualizovať databázu dokumentov pre pridanie nových informácií (nie je potrebné doučiť LLM)
- Databáza môže obsahovať aj privátne dokumenty
- V odpovedi môžeme zahrnúť aj odkazy na zdrojové dokumenty

RAG - Retrieval-Augmented Generation (2)



RAG - Retrieval-Augmented Generation (3)

Dopyt: Aké sú vedľajšie účinky paracetamolu?



Otvorené otázky a výskum jazykových modelov

- Kvalitné trénovacie dáta a menšie modely
- Dostupnejšie metódy pre prispôsobenie modelov
- Dlhší kontext
- Dôveryhodnosť
- Ochrana súkromia a bezpečnosť
- Ochrana autorských práv