valira·ai

Akademija velikih jezikovnih modelov

GZS, 2025

Kdo sva?

Luka Vranješ
luka@valira.ai

Andrej Miščič
andrej@valira.ai

valira·ai

rešitve po meri na področju AI, strojnega učenja in širše podatkovne znanosti

OneQlue

AI layer unifying enterprise knowledge — from documents and ERP to 3D models and drawings

Praktični del akademije

Danes:

1. Uvod v velike jezikovne modele
2. Prompt inženiring

V nadaljevanju:

3. BERT skozi praktične primere
4. Razvoj LLM rešitev
5. Praktični trendi v svetu LLM-jev

01

Uvod v velike jezikovne modele

Agenda

1. Osnove delovanja velikih jezikovnih modelov
2. …

Velik jezikovni model

Autocomplete

“Veliki jezikovni modeli
so *autocomplete* na steroidih.”

Autocomplete

NALOGA:

- napoved naslednje besede

Autocomplete

NALOGA:

- napoved naslednje besede

MODEL:

- “stvar”, ki generira naslednjo besedo

Autocomplete

NALOGA:

- napoved naslednje besede

MODEL:

- “stvar”, ki generira naslednjo besedo

ENOSTAVEN PRIMER:

- pogledamo prejšnjo besedo, t.i. **bi-gram** model:

- (Can’t, believe) - 40%
- (Can’t, wait) - 30%
- (Can’t, remember) - 20%

Autocomplete

NALOGA:
- napoved naslednje besede

MODEL:
- "stvar", ki generira naslednjo besedo
- bi-gram model

(Can't, believe) - 40%
(Can't, wait) - 30%
(Can't, remember) - 20%

Autocomplete

NALOGA:

- napoved naslednje besede

MODEL:

- "stvar", ki generira naslednjo besedo

- bi-gram model

(Can't, believe) - 40%
(Can't, wait) - 30%
(Can't, remember) - 20%

PODATKI:

- za učenje modela

- zbirka nekaj knjig

*"As the sun dipped below the horizon, casting a golden hue over the city, Alex stood at the edge of the rooftop, his eyes wide with disbelief. "I **can't believe** it," he whispered, the words barely escaping his lips."*

Autocomplete

NALOGA:

- napoved naslednje besede

MODEL:

- "stvar", ki generira naslednjo besedo

- bi-gram model

(Can't, believe) - 40%
(Can't, wait) - 30%
(Can't, remember) - 20%

PODATKI:

- za učenje modela

- zbirka nekaj knjig

*"As the sun dipped below the horizon, casting a golden hue over the city, Alex stood at the edge of the rooftop, his eyes wide with disbelief. "I **can't believe** it," he whispered, the words barely escaping his lips."*

RAČUNSKA MOČ:

- za učenje modela

- osebni računalnik

~~Autocomplete~~ Veliki jezikovni modeli

NALOGA:

- napoved naslednje besede

MODEL:

- "stvar", ki generira naslednjo besedo

- bi-gram model

(Can't, believe) - 40%
(Can't, wait) - 30%
(Can't, remember) - 20%

PODATKI:

- za učenje modela

- zbirka nekaj knjig

*"As the sun dipped below the horizon, casting a golden hue over the city, Alex stood at the edge of the rooftop, his eyes wide with disbelief. "I **can't believe** it," he whispered, the words barely escaping his lips."*

RAČUNSKA MOČ:

- za učenje modela

- osebni računalnik

~~Autocomplete~~ Veliki jezikovni modeli

NALOGA:
- *napoved naslednje besede*
= **modeliranje jezika**
(language modeling)

MODEL:
- *"stvar", ki generira naslednjo besedo*
- bi-gram model

(Can't, believe) - 40%
(Can't, wait) - 30%
(Can't, remember) - 20%

PODATKI:
- *za učenje modela*
- zbirka nekaj knjig

*"As the sun dipped below the horizon, casting a golden hue over the city, Alex stood at the edge of the rooftop, his eyes wide with disbelief. "I **can't believe** it," he whispered, the words barely escaping his lips."*

RAČUNSKA MOČ:
- *za učenje modela*
- osebni računalnik

~~Autocomplete~~ Veliki jezikovni modeli

NALOGA:

- *napoved naslednje besede*

= **modeliranje jezika**

(language modeling)

MODEL:

- *"stvar", ki generira naslednjo besedo*

~~**bi-gram** model~~

~~(Can't, believe) - 40%~~
~~(Can't, wait) - 30%~~
~~(Can't, remember) - 20%~~

Transformer

- arhitektura, na kateri temeljijo vsi moderni LLM-ji

PODATKI:

- *za učenje modela*
- zbirka nekaj knjig

*"As the sun dipped below the horizon, casting a golden hue over the city, Alex stood at the edge of the rooftop, his eyes wide with disbelief. "I **can't believe** it," he whispered, the words barely escaping his lips."*

RAČUNSKA MOČ:

- *za učenje modela*
- osebni računalnik

~~Autocomplete~~ Veliki jezikovni modeli

NALOGA:
- *napoved naslednje besede*
= **modeliranje jezika**
(language modeling)

MODEL:
- *"stvar", ki generira naslednjo besedo*
~~**bi-gram** model~~

~~(Can't, believe) - 40%~~
~~(Can't, wait) - 30%~~
~~(Can't, remember) - 20%~~

Transformer
- arhitektura, na kateri temeljijo vsi moderni LLM-ji

PODATKI:
- *za učenje modela*
~~- zbirka nekaj knjig~~

*"As the sun dipped below the horizon, casting a golden hue over the city, Alex stood at the edge of the rooftop, his eyes wide with disbelief. "I **can't believe** it," he whispered, the words barely escaping his lips."*

- biljoni besed, npr. crawl interneta

llama 4: 30T tokens

RAČUNSKA MOČ:
- *za učenje modela*
- osebni računalnik

~~Autocomplete~~ Veliki jezikovni modeli

NALOGA:

- napoved naslednje besede

= **modeliranje jezika**

(language modeling)

MODEL:

- "stvar", ki generira naslednjo besedo

~~**- bi-gram** model~~

~~(Can't, believe) - 40%~~
~~(Can't, wait) - 30%~~
~~(Can't, remember) - 20%~~

Transformer

- arhitektura, na kateri temeljijo vsi moderni LLM-ji

PODATKI:

- za učenje modela

~~- zbirka nekaj knjig~~

*"As the sun dipped below the horizon, casting a golden hue over the city, Alex stood at the edge of the rooftop, his eyes wide with disbelief. "I **can't believe** it," he whispered, the words barely escaping his lips."*

- biljoni besed, npr. crawl interneta

llama 4: 30T tokens

RAČUNSKA MOČ:

- za učenje modela

~~- osebni računalnik~~

- superračunalniki

llama 4: cluster of 32k H100 GPUs

“Veliki jezikovni modeli
so *autocomplete* na steroidih.”

“Veliki jezikovni modeli
so *autocomplete* na steroidih.” *

*

"People say, It's just glorified autocomplete ... Now, let's analyze that. Suppose you want to be really good at predicting the next word. If you want to be really good, you have to understand what's being said. That's the only way. So by training something to be really good at predicting the next word, you're actually forcing it to understand. Yes, it's 'autocomplete' — but you didn't think through what it means to have a really good autocomplete."

- Geoff Hinton, *"godfather of AI"*

Osnoven *autocomplete*:

- ne zahteva razumevanja jezika;
- deluje v enostavnih situacijah in ne splošno.

Perfekten *autocomplete*:

- zahteva popolno razumevanje jezika;
- neodvisno od situacije razume besedilo in pravilno napove besedo.

Osnoven *autocomplete*:

- ne zahteva razumevanja jezika;
- deluje v enostavnih situacijah in ne splošno.

Perfekten *autocomplete*:

- zahteva popolno razumevanje jezika;
- neodvisno od situacije razume besedilo in pravilno napove besedo.

AI skeptiki
“stochastic parrots”

AI doomerji
“AGI”

Osnoven *autocomplete*:

- ne zahteva razumevanja jezika;
- deluje v enostavnih situacijah in ne splošno.

Perfekten *autocomplete*:

- zahteva popolno razumevanje jezika;
- neodvisno od situacije razume besedilo in pravilno napove besedo.

Ne glede na dejansko pozicija na spektru, se izkaže:

- če učimo dovolj velike modele;
- če učimo na dovolj (kvalitetnih) podatkih;

je rezultat splošen LLM, ki ima praktično uporabnost.

Zmožnosti LLM-jev skozi prizmo *autocomplete-a*

Večjezičnost

SLOVENŠČINA:

Moj najljubši šport je …

a) nogomet ***b)*** football ***c)*** futbòl

ANGLEŠČINA:

My favourite sport is …

a) nogomet ***b)*** football ***c)*** futbòl

Večjezičnost

SLOVENŠČINA:

Moj najljubši šport je …

a)* nogomet** ***b) football ***c)*** futbòl

ANGLEŠČINA:

My favourite sport is …

a) nogomet ***b)* football** ***c)*** futbòl

Sledenje navodilom

Uporabnik: Kaj je tvoja najljubša pijača?
LLM: …

a) Coca-Cola ***b)*** Rum ***c)*** Voda

Uporabnik: Si pirat. Kaj je tvoja najljubša pijača?
LLM: …

a) Coca-cola ***b)*** Rum ***c)*** Voda

Sledenje navodilom

Uporabnik: Kaj je tvoja najljubša pijača?
LLM: …

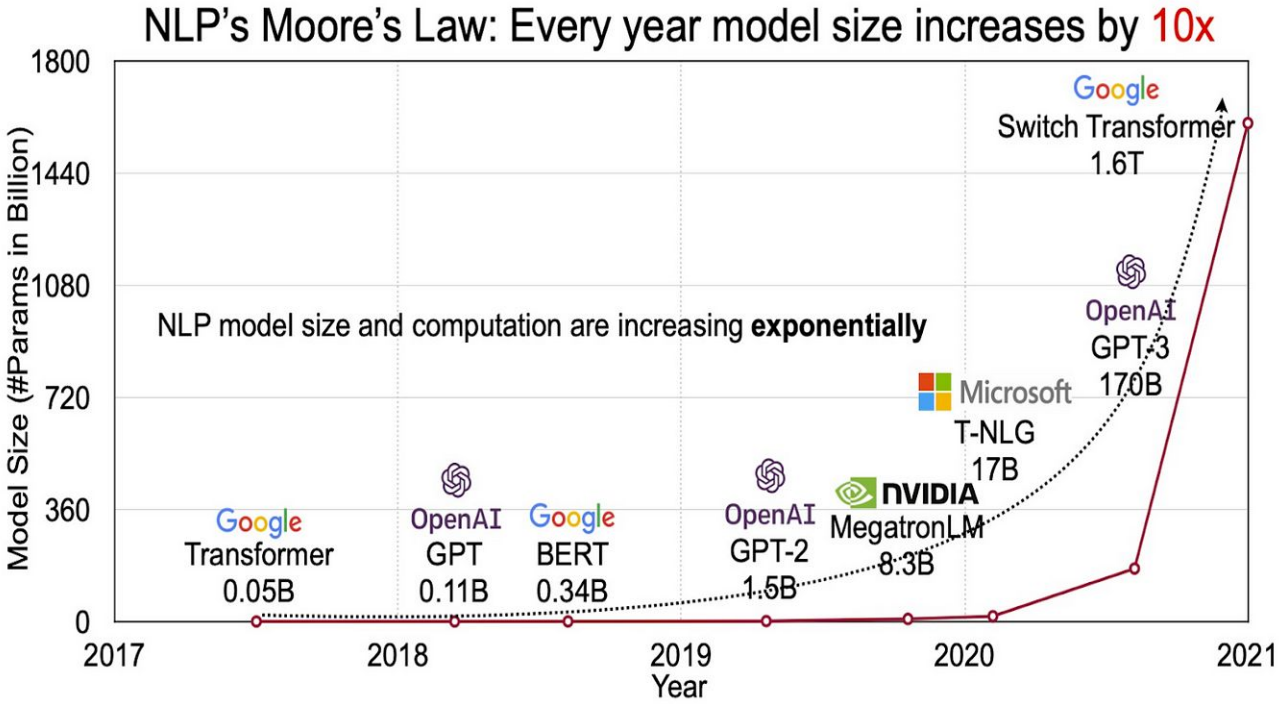
a) Coca-Cola ***b)*** Rum ***c)*** **Voda**

Uporabnik: Si pirat. Kaj je tvoja najljubša pijača?
LLM: …

a) Coca-cola ***b)*** **Rum** ***c)*** Voda

LLM-ji (podrobneje)

VELIKI jezikovni modeli



Skaliranje

Tri dimenzije:

- velikost modelov (število parametrov)
- velikost učne množice
- računska moč - kako velike modele lahko učimo na kako veliko podatkih za koliko časa

Skaliranje

Tri dimenzije:

- velikost modelov (število parametrov)
- velikost učne množice
- računska moč - kako velike modele lahko učimo na kako veliko podatkih za koliko časa

Scaling laws: za omejeno računsko moč obstaja najučinkovitejši kompromis med velikostjo modela in velikostjo učne množice

Tokeni (žetoni?)

LLMs = Large Language Models

Language modelling (modeliranje jezika) - napoved naslednje besede

Tokeni (žetoni?)

LLMs = Large Language Models

Language modelling (modeliranje jezika) - napoved naslednje besede

–

Ali LLM-ji delujejo nad besedami?

Tokeni (žetoni?)

LLMs = Large Language Models

Language modelling (modeliranje jezika) - napoved naslednje besede

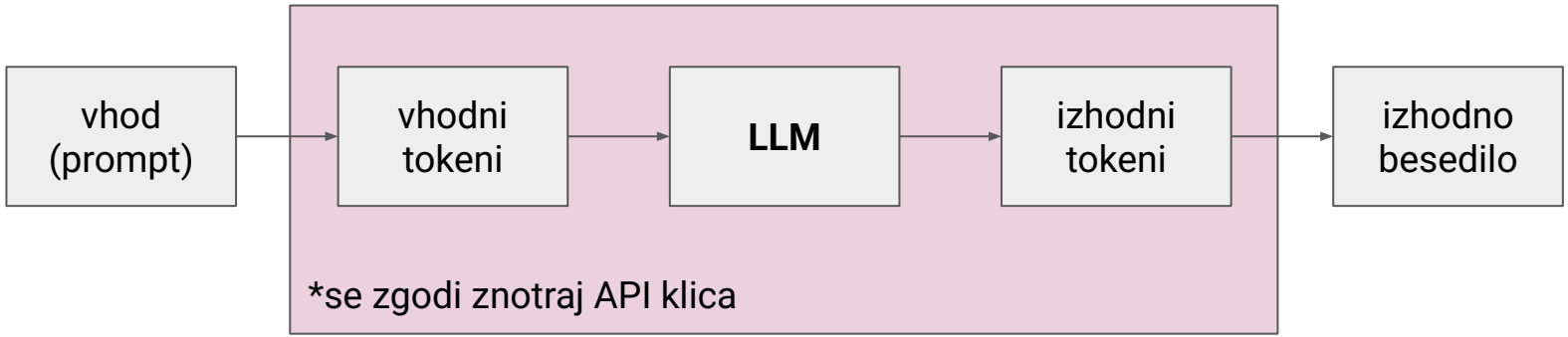
–

Ali LLM-ji delujejo nad besedami? Ne, pač pa nad tokeni.

Tokeni (žetoni?)

Token = najmanjša samostojna enota, ki jo procesirajo LLM-ji.

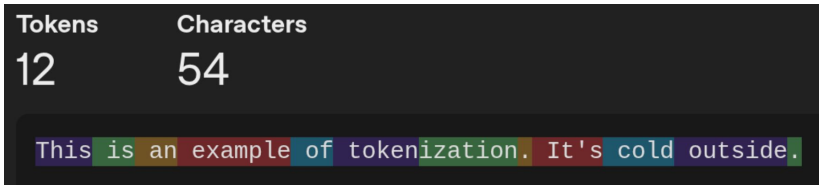
- vhod v LLM (prompt) je razbit na tokene, izhod je sestavljen iz tokenov

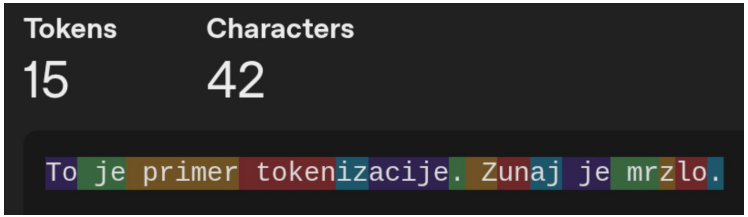


Tokeni (žetoni?)

Token = najmanjša samostojna enota, ki jo procesirajo LLM-ji.

- vsaka beseda je lahko sestavljena iz enega ali večih tokenov
- primer: OpenAI GPT-4o tokenizacija (vir)





Tokeni (motivacija)

Z omejenim naborom tokenov lahko podpremo kakršnokoli besedilo.

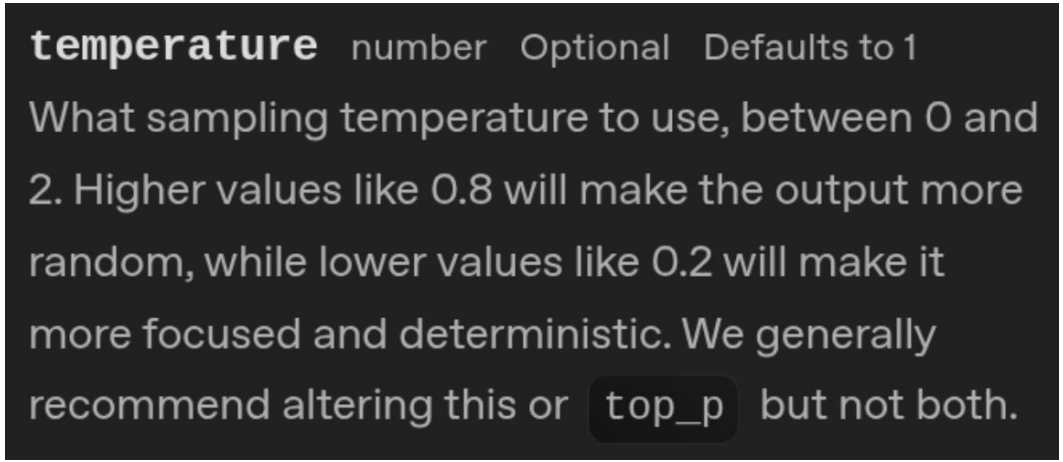
- slovenščina je primer morfološko-bogatega jezika

	ednina	dvojina	množina
imenovalnik	žeton	žetona	žetoni
rodilnik	žetona	žetonov	žetonov
dajalnik	žetonu	žetonoma	žetonom
tožilnik	žeton	žetona	žetone
mestnik	žetonu	žetonih	žetonih
orodnik	žetonom	žetonoma	žetoni

	ednina	dvojina	množina
prva oseba	kuham	kuhava	kuhamo
druga oseba	kuhaš	kuhata	kuhate
tretja oseba	kuha	kuhata	kuhajo

Temperatura

OpenAI dokumentacija:



Izhod LLM-jev

- napoved naslednje besede = *verjetnostna porazdelitev* čez vse besede

Izhod LLM-jev

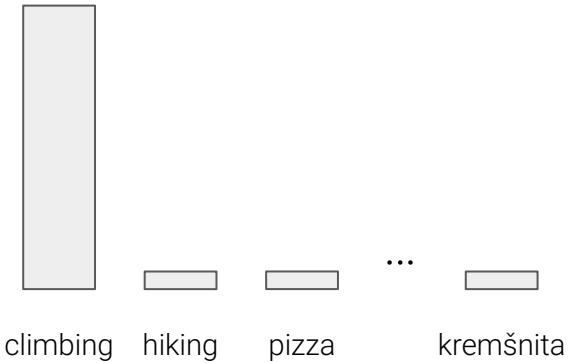
- napoved naslednje besede = *verjetnostna porazdelitev* čez vse besede

I like …

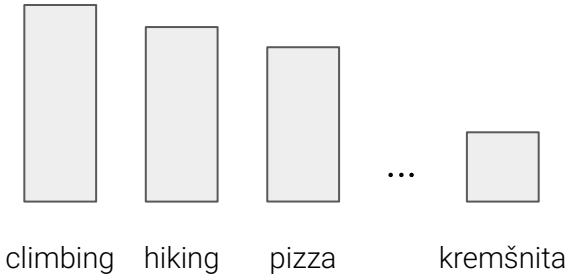
Temperatura

- nadzoruje porazdelitev

I like ...



nižji T
bolj deterministično



višji T
bolj stohastično (kreativno)

Učenje velikih jezikovnih modelov

Učenje velikih jezikovnih modelov

1. Predučenje *(pretraining)*
2. Učenje za dialoge
3. Učenje človeških preferenc *(alignment)*
4. (razmišljujoči modeli)

Učenje velikih jezikovnih modelov

1. **Predučenje *(pretraining)***
2. Učenje za dialoge
3. Učenje človeških preferenc *(alignment)*
4. (razmišljujoči modeli)

CILJ:

- razumevanje jezika (slovnica, besedišče, ločila, pojmi …)
- pridobivanje znanja (informacije o svetu …)

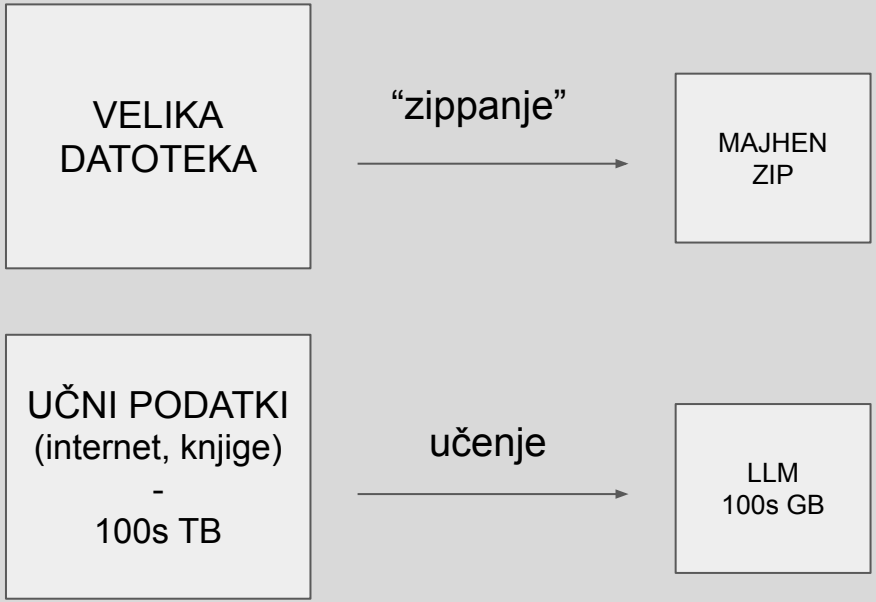
PODATKI:

- splošen jezik (Internet, knjige, …)

Učenje velikih jezikovnih modelov

1. **Predučenje *(pretraining)***
2. Učenje za dialoge
3. Učenje človeških preferenc *(alignment)*
4. (razmišljujoči modeli)

Analogija:



Učenje velikih jezikovnih modelov

1. Predučenje *(pretraining)*
2. **Učenje za dialoge**
3. Učenje človeških preferenc *(alignment)*
4. (razmišljujoči modeli)

CILJ:

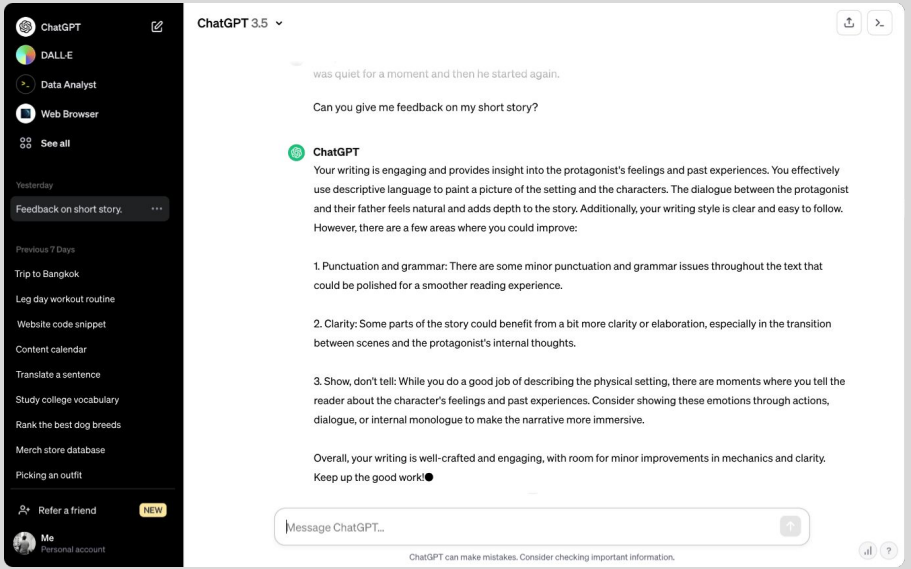
- LLM-ji razumejo interakcije v obliki dialogov (npr.: vprašanje - odgovor)

PODATKI:

- kurirani dialogi
- učna naloga še zmeraj ostaja napoved naslednje besede

Učenje velikih jezikovnih modelov

1. Predučenje *(pretraining)*
2. **Učenje za dialoge**
3. Učenje človeških preferenc *(alignment)*
4. (razmišljujoči modeli)



Učenje velikih jezikovnih modelov

1. Predučenje *(pretraining)*
2. Učenje za dialoge
3. **Učenje človeških preferenc *(alignment)***
4. (razmišljujoči modeli)

Učenje velikih jezikovnih modelov

1. Predučenje *(pretraining)*
2. Učenje za dialoge
3. **Učenje človeških preferenc *(alignment)***
4. (razmišljujoči modeli)

CILJ:
- varnostni mehanizem
- mitigacija pristranskosti, toksičnosti, drugih škodljivih vsebin

PODATKI:
- rangirani odgovori modelov (izražena preferenca)

Učenje velikih jezikovnih modelov

1. Predučenje *(pretraining)*
2. Učenje za dialoge
3. Učenje človeških preferenc *(alignment)*
4. **(razmišljujoči modeli)**

CILJ:

- izboljšanje rezultatov LLM-jev s tem, da jim pustimo *razmišljati*

Nadgradnja LLM-jev

Nadgradnja LLM-jev - kdaj?

- Ko imate jasne in močne primere uporabe, ki jih off-the-shelf orodja ne rešujejo dovolj dobro.
- Ko želite nadgraditi oziroma obogatiti obstoječe modele z vašimi podatki in aplikacijami.
- Zahteva po zrelosti procesov povezanih z vpeljavo umetne inteligence (npr. LLMOps).

l’AI pour l’AI (AI for AI’s sake)

- strah pred zaostankom za konkurenco
- želja po pridobitvi konkurečne prednosti

l'AI pour l'AI (AI for AI's sake)

MIT report: 95% of generative AI pilots at companies are failing

BY SHERYL ESTRADA
SENIOR WRITER AND AUTHOR OF CFO DAILY

August 18, 2025 at 6:54 AM EDT

- "tools like ChatGPT excel for individuals but stall in enterprise as companies don't adapt to workflows"

- "more than half of budget devoted to sales and marketing tools (op.a. external use), where biggest ROI in back-office automation"

Nadgradnja LLM-jev

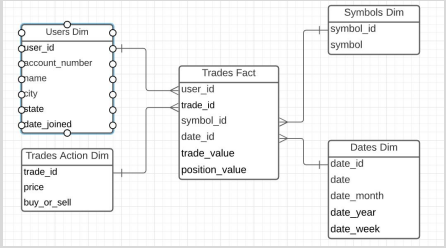
1. Vaši podatki
2. Vaše aplikacije

Nadgradnja LLM-jev

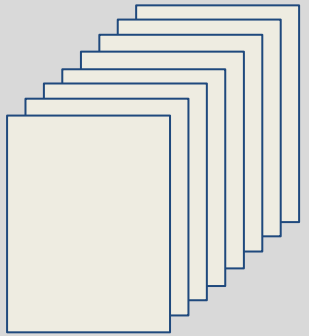
1. **Vaši podatki**
2. Vaše aplikacije

V dobi umetne inteligence so podatki zlato:

- strukturirani



- nestrukturirani



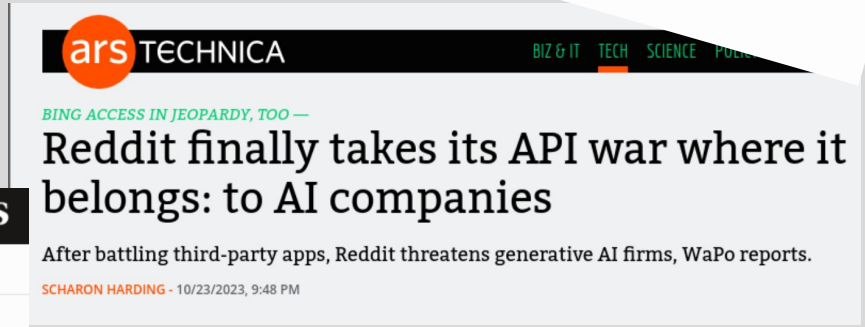
Nadgradnja LLM-jev

1. **Vaši podatki**
2. Vaše aplikacije


Apps

Reddit will begin charging for access to its API

Kyle Wiggers @kyle_l_wiggers / 9:08 PM GMT+2 • April 18, 2023


ars TECHNICA BIZ & IT TECH SCIENCE

BING ACCESS IN JEOPARDY, TOO —

Reddit finally takes its API war where it belongs: to AI companies

After battling third-party apps, Reddit threatens generative AI firms, WaPo reports.

SCHARON HARDING - 10/23/2023, 9:48 PM


Forbes

FORBES > BUSINESS

BREAKING

Twitter Ends Its Free API: Here's Who Will Be Affected

Jenae Barnes Former Staff

Forbes Staff

Nadgradnja LLM-jev

1. **Vaši podatki**
2. Vaše aplikacije

Česa LLM-ji niso videli med učenjem?

- vaših internih podatkov? Dokumentacija, pravilniki, CRM, itd.
- ...

Intermezzo: halucinacije

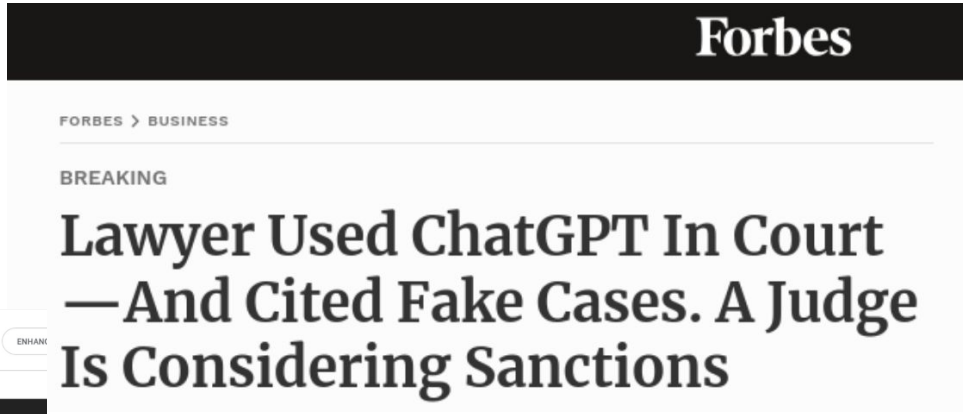
Samozavestno generiranje napačnih informacij, ki se na prvi pogled zdijo verjetne.

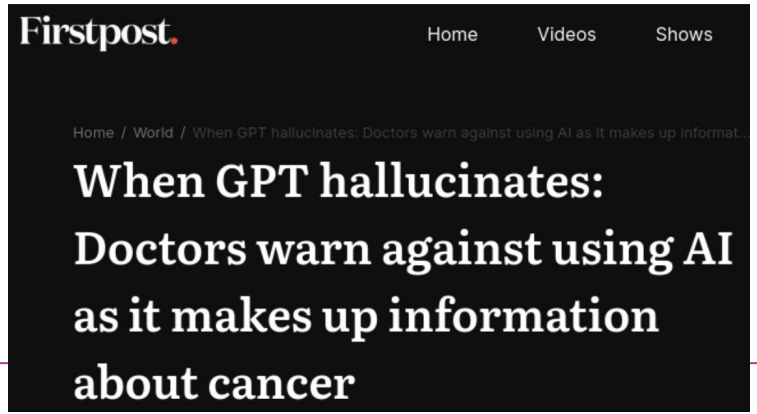
Intermezzo: halucinacije

Samozavestno generiranje napačnih informacij, ki se na prvi pogled zdijo verjetne.

- LLM-ji nimajo direktnega dostopa do *resnice* (baze dejstev).
- Zanašajo se na vzorce naučene v procesu učenja (slaba generalizacija?)

Intermezzo: halucinacije



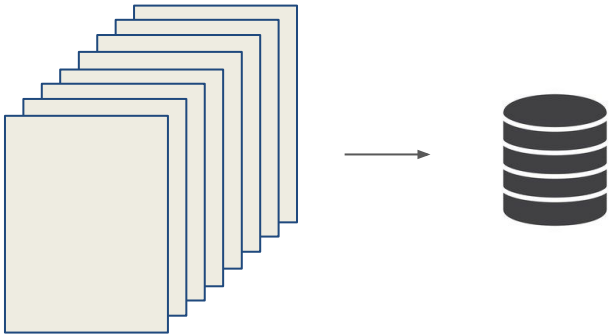


RAG (Retrieval-augmented generation)

(“odgovarjanje na vprašanja s pridobivanjem informacij”)

RAG (Retrieval-augmented generation)

(“odgovarjanje na vprašanja s pridobivanjem informacij”)



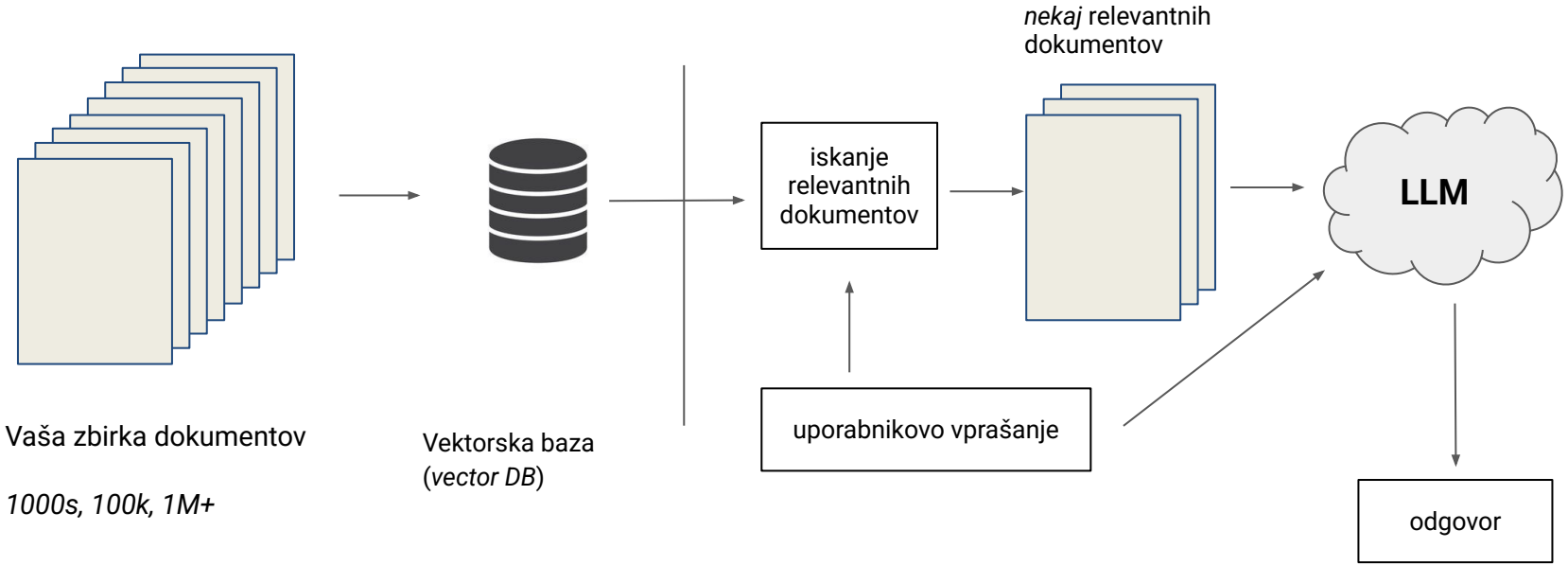
Vaša zbirka dokumentov

1000s, 100k, 1M+

Vektorska baza
(*vector DB*)

RAG (Retrieval-augmented generation)

(“odgovarjanje na vprašanja s pridobivanjem informacij”)



Nadgradnja LLM-jev

1. Vaši podatki
2. **Vaše aplikacije**

Dostop prek programskih vmesnikov:

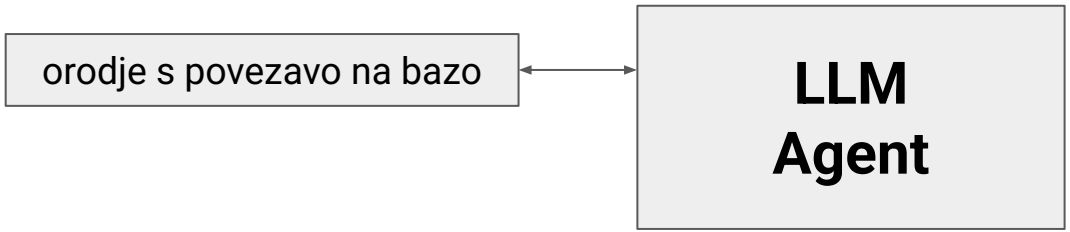
- pridobivanje podatkov v realnem času;
- izvajanje akcij.

Agent

LLM, ki komunicira z zunanjim svetom z izvajanjem *orodij*

Agent

LLM, ki komunicira z zunanjim svetom z izvajanjem *orodij*



User: V katerih trgovinah so na voljo ti čevlji?

LLM: SQL orodje, query: “SELECT …”

SQL rezultat: *empty*

LLM: Žal jih ni na zalogi.

Agent

LLM, ki komunicira z zunanjim svetom z izvajanjem *orodij*

User: Kakšno je vreme v Kranju?

LLM: ARSO API, {“location”: “Kranj”}

ARSO API rezultat: oblačno, 10°C

LLM: Mrzlo.

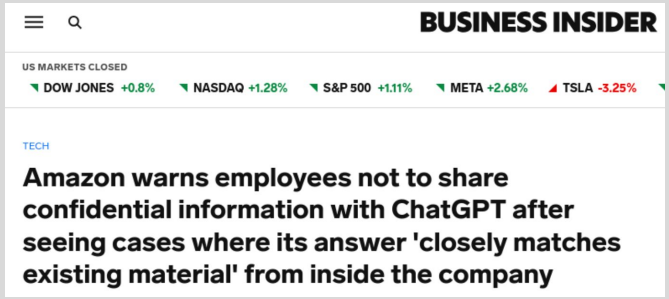
Pasti vpeljave

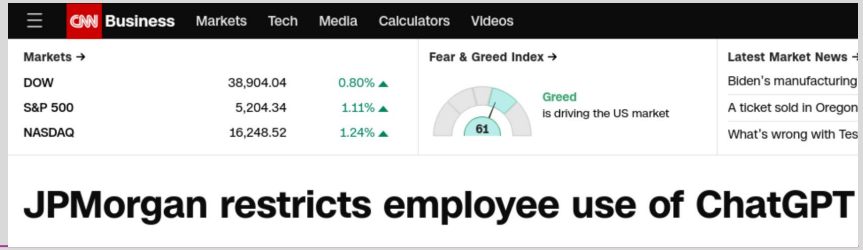
1. Privatnost
2. LLMOps
3. Dolžina konteksta

Pasti vpeljave

1. **Privatnost**
2. LLMOps
3. Dolžina konteksta

"If you are not paying for the product, you are the product."





Pasti vpeljave

1. **Privatnost**
2. LLMOps
3. Dolžina konteksta

- lastne postavitve (on-prem/cloud)
- API providerji, hranijo podatke, a ne učijo na njih

Pasti vpeljave

1. Privatnost
2. **LLMOps**
3. Dolžina konteksta

- evalvacija sistemov
- monitoring, observability, tracing:
  - skladnost rezultatov
  - pregled stroškov
  - debugging
  - latenca
- upravljanje s podatki: RAG
- pravice, dostopi, ...

Pasti vpeljave

1. Privatnost
2. LLMOps
3. **Dolžina konteksta**

Število tokenov, ki jih LLM lahko obdela:
št. vhodnih + št izhodnih tokenov

Moderni LLM-ji imajo ogromne kontekste - potrebno paziti na eksplozijo stroškov.