Priložnosti in pasti generativne umetne inteligence
(in jezikovnih modelov)

Kdo sva?

Luka Vranješ
luka@valira.ai

Andrej Miščič
andrej@valira.ai

Akademija umetne inteligence za poslovne aplikacije

Danes:

Priložnosti in pasti generativne umetne inteligence (in jezikovnih modelov)
Učinkovita komunikacija z velikimi jezikovnimi modeli (prompt engineering)

Naslednjič:

Sodobna obdelava naravnega jezika: BERT prek praktičnih primerov
Komuniciranje aplikacij in jezikovnih modelov: GPTs

Na-naslednjič:

Komuniciranje aplikacij in jezikovnih modelov: ChatGPT funkcije
Sodobna obdelava naravnega jezika: nadgradnja ChatGPT-ja s knjižnico LangChain

Agenda

00 Uvod v uvod

01 Osnove delovanja velikih jezikovnih modelov

02 Priložnosti vpeljave v podjetjih

03 Pasti vpeljave ter njihova ublažitev

04 Pregled velikih jezikovnih modelov

ODMOR

Delavnica *prompt inženiringa*

Buzzwords (*modne besede*)

Artificial Intelligence

Umetna inteligenca

Data Science

AI

Machine Learning

Strojno učenje

Podatkovna znanost

AI vs. ML vs. DS

AI vs. ML vs. DS

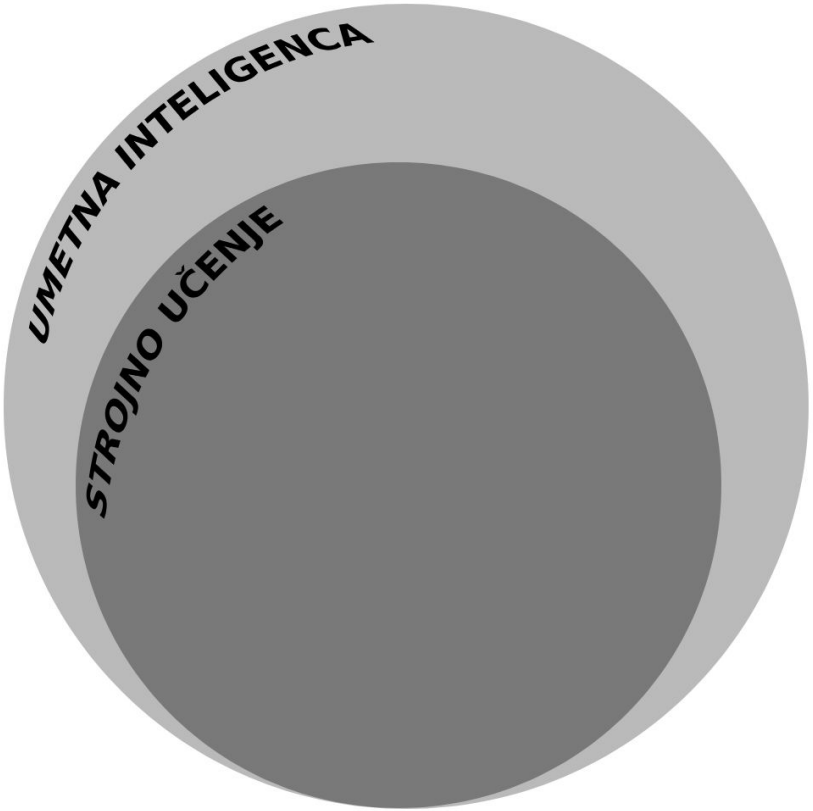
(in kako se s temi pojmi povezuje ChatGPT?)


UMETNA INTELIGENCA

AI - umetna inteligenca

CILJ:
posnemanje človekovih *inteligentnih* lastnosti

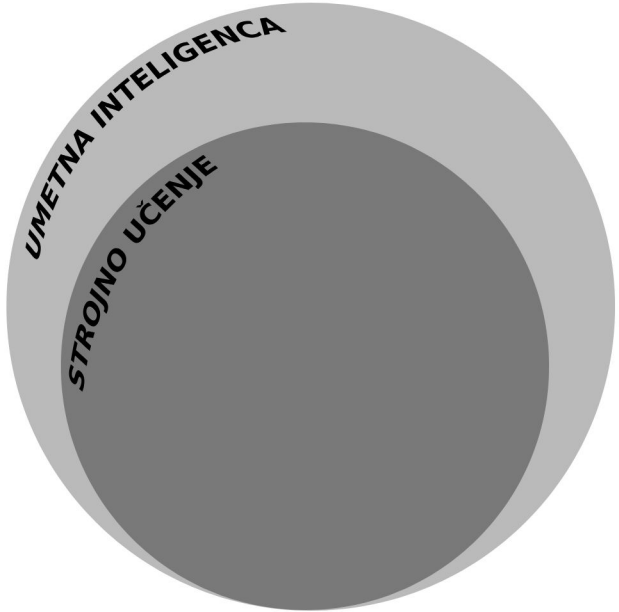
- Načrtovanje;
- Učenje;
- Sprejemanje odločitev;
- …

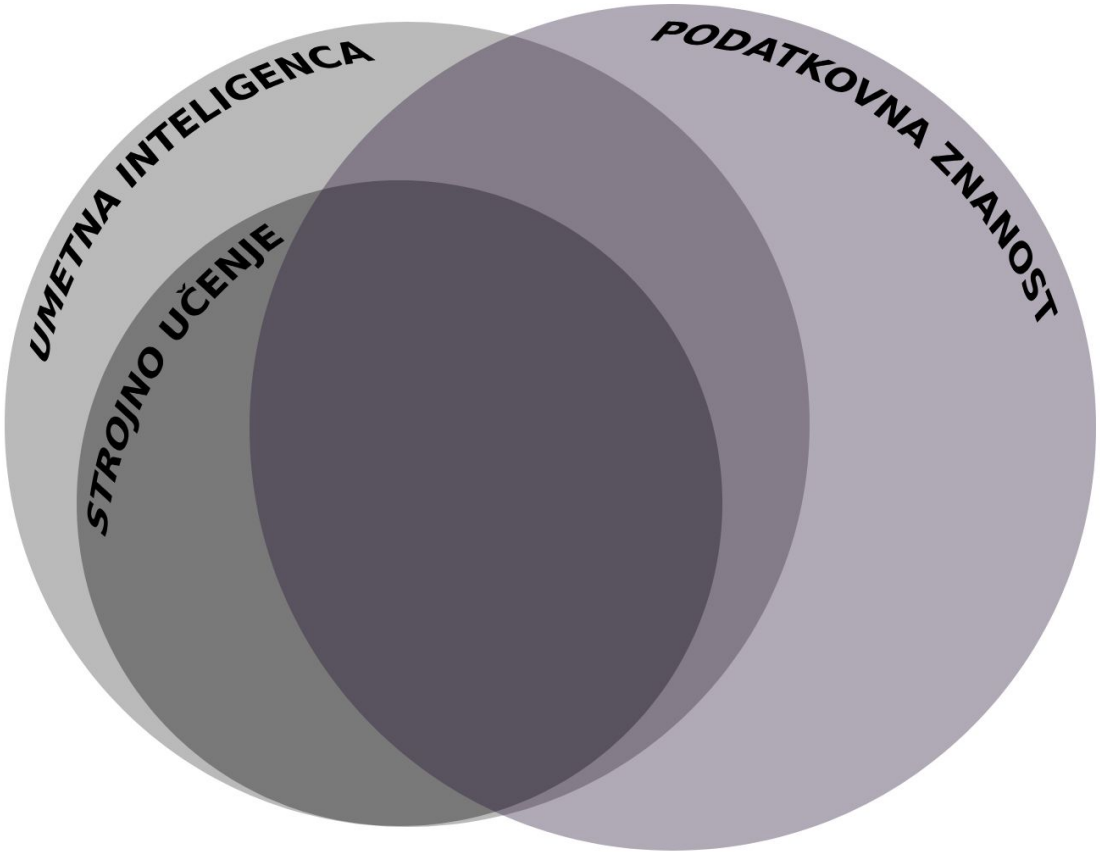

UMETNA INTELIGENCA
STROJNO UČENJE

ML - strojno učenje

Področje **umetne inteligence**
CILJ:
Stroj posnema človeško učenje

1. Avtomatsko pridobivanja znanja iz podatkov (tabularičnih, besedil, slik, …)
2. Določanje lastnosti novih podatkov (glede na pridobljeno znanje, izkušnje)

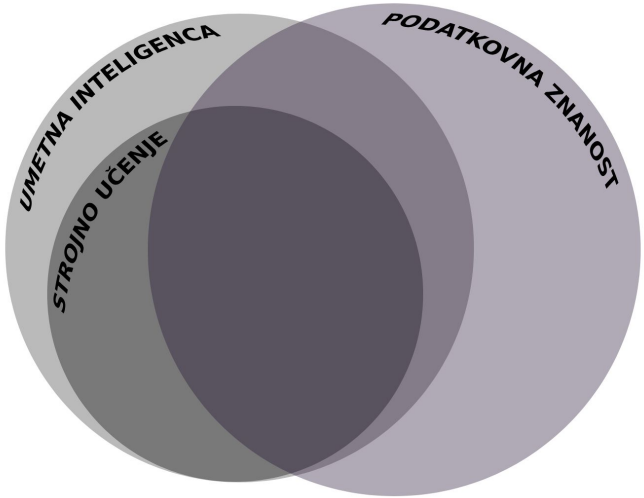


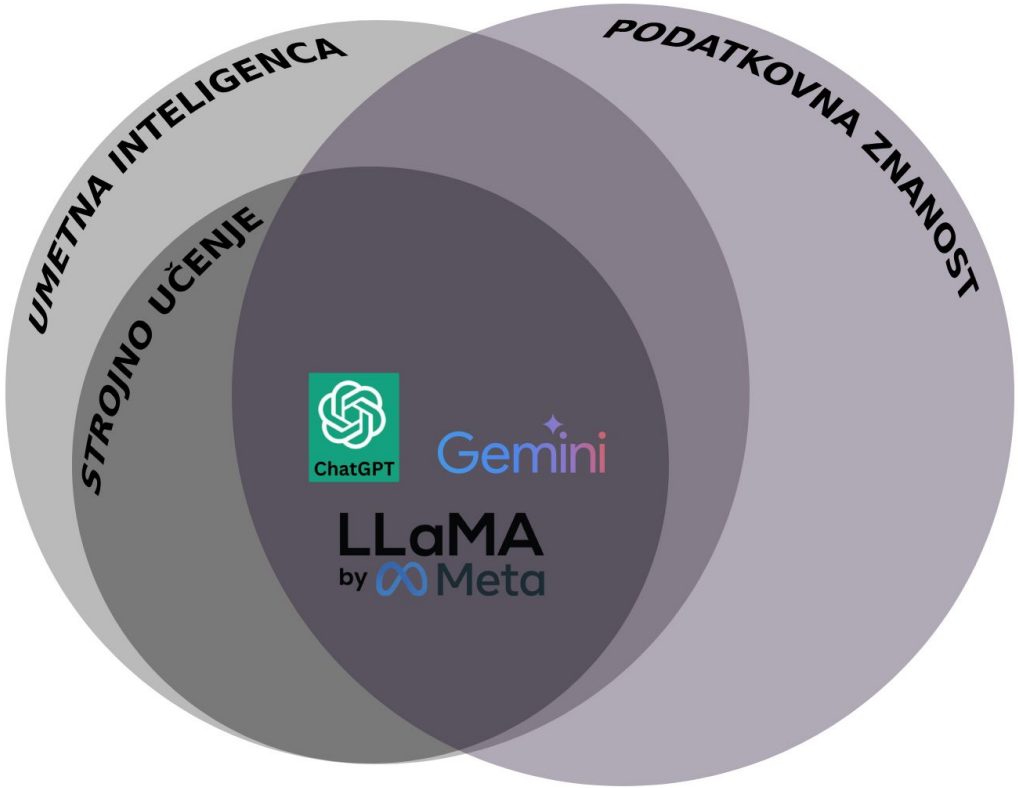

UMETNA INTELIGENCA
STROJNO UČENJE
PODATKOVNA ZNANOST

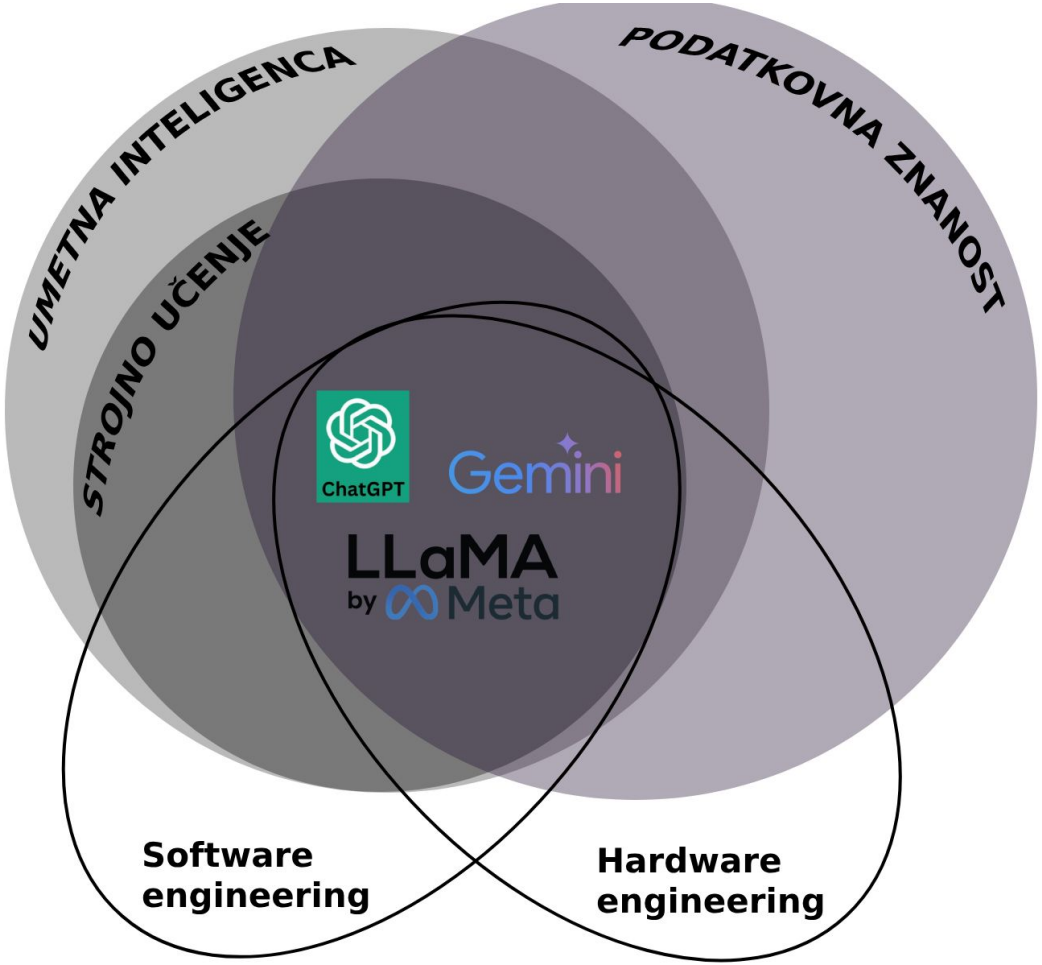
DS - podatkovna znanost

Sistematičen način pridobivanja informacij iz podatkov

- od poslovne analitike…
  - Analiza podatkov;
  - Statistika;
  - Vizualizacija podatkov;
- do prediktivne analitike…
  - Procesiranje podatkov;
  - AI in ML;
- in vse okrog.




UMETNA INTELIGENCA
STROJNO UČENJE
PODATKOVNA ZNANOST
ChatGPT
Gemini
LLaMA
by Meta


UMETNA INTELIGENCA
STROJNO UČENJE
PODATKOVNA ZNANOST
ChatGPT
Gemini
LLaMA
by Meta
Software
engineering
Hardware
engineering

Osnove velikih jezikovnih modelov

Autocomplete

(predlogi naslednje besede)

“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."*

RAČUNSKA MOČ:

- za učenje modela

- osebni računalnik

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**, e.g.:
  - OpenAI ChatGPT,
  - Meta LLama,
  - vsi moderni LLM-ji.

PODATKI:

- za učenje modela

- zbirka nekaj knjig

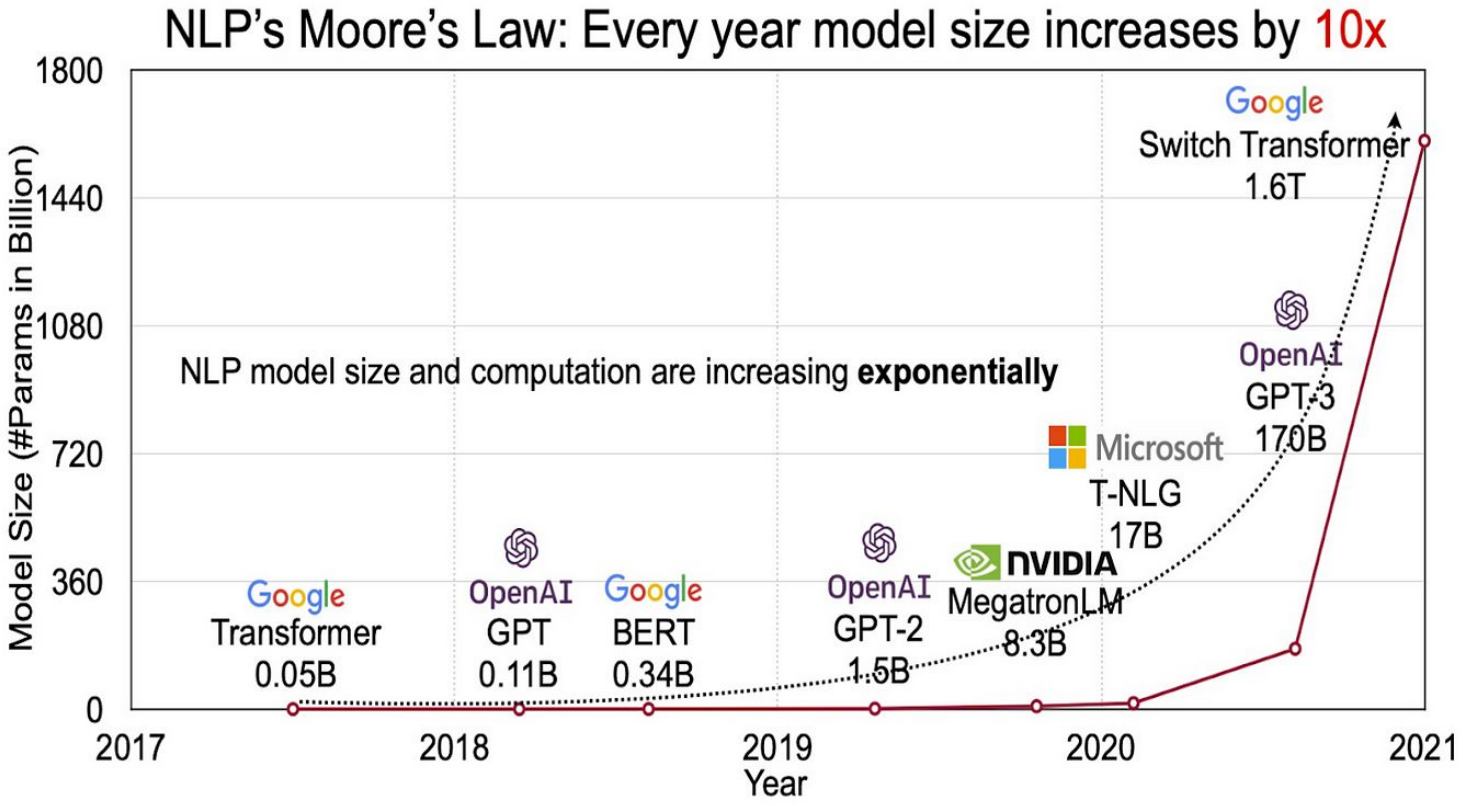
*"As the sun dipped below the horizon, casting a golden hue over the city, Alex stood at the edge of the rooftop, his eyes wide with disbelief. "I **can't believe** it," he whispered, the words barely escaping his lips."*

RAČUNSKA MOČ:

- za učenje modela

- osebni računalnik

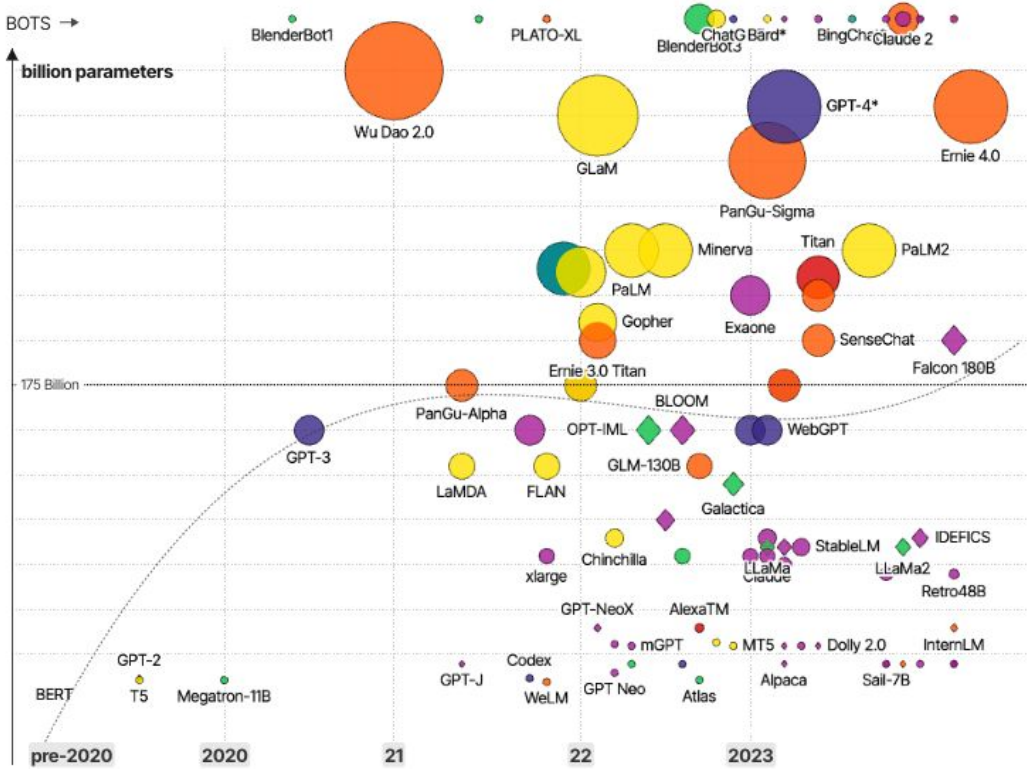
Intermezzo: VELIKI jezikovni modeli



The Rise and Rise of A.I.
Large Language Models (LLMs) & their associated bots like ChatGPT

size = no. of parameters open-access

Amazon-owned Chinese Google Meta / Facebook Microsoft OpenAI Other



David McCandless, Tom Evans, Paul Barton
Information is Beautiful // UPDATED 2nd Nov 23

source: news reports, LifeArchitect.ai
* = parameters undisclosed // see the data

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**, e.g.:
  - OpenAI ChatGPT,
  - Meta LLama,
  - 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 - več 100 TB podatkov:
  - množica knjig,
  - Internet,
  - ...

RAČUNSKA MOČ:

- za učenje modela

- osebni računalnik

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**, e.g.:
  - OpenAI ChatGPT,
  - Meta LLama,
  - 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 - več 100 TB podatkov:
  - množica knjig,
  - Internet,
  - ...

RAČUNSKA MOČ:

- za učenje modela

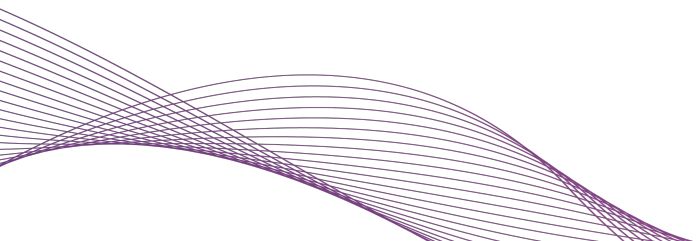
- ~~osebni računalnik~~

- superračunalniki s kopico GPU-jev

“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/aplikacijah in ne splošno.

Perfekten *autocomplete*:

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

Veliki jezikovni modeli

Osnoven *autocomplete*:

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

Perfekten *autocomplete*:

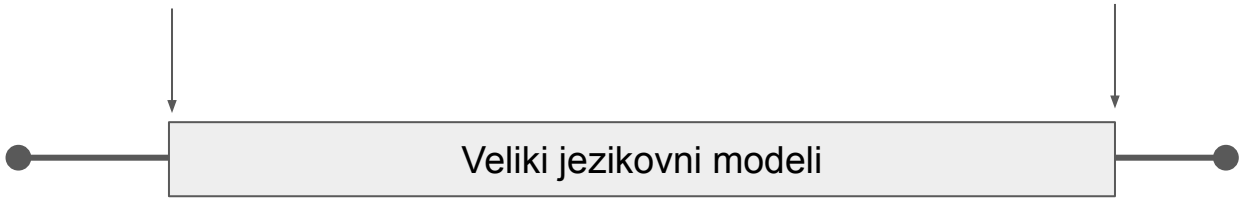
- zahteva popolno razumevanje jezika;
- neodvisno od situacije, razume jezik in pravilno napove besedo.

AI skeptiki

- To so zgolj papige!

AI evangelisti

- To je splošna umetna inteligenca!



Osnoven *autocomplete*:

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

Perfekten *autocomplete*:

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

Ne glede na pozicijo LLM-jev na prejšnji skali, ne moremo zanikati njihove praktične uporabnosti!

Izkaže se, da:

- če učimo dovolj velike modele (več miljard učljivih parametrov) in
- če učimo na dovolj veliko podatkov (več miljard do biljonov besed)

je rezultat splošno uporaben model.

Ne glede na pozicijo LLM-jev na prejšnji skali, ne moremo zanikati njihove praktične uporabnosti!

Izkaže se, da:

- če učimo dovolj velike modele (več miljard učljivih parametrov) in
- če učimo na dovolj veliko podatkov (več miljard do biljonov besed)

je rezultat splošno uporaben model.

Emergent abilities *(spontane sposobnosti)*

V STARIH ČASIH (beri: 3 leta nazaj):

- Specializiran model za različne naloge:
  - Odgovarjanje na vprašanja;
  - Povzemanje besedil;
  - Klasifikacija besedil;
  - Tvorjenje pesmi;
  - Matematika;
  - Prevajanje;
  - ...

DANES:

- En velik jezikovni model, ki je zmožen opraviti vse te naloge.

Kaj so posledice dejstva, da so LLM-ji izboljšan *autocomplete?*

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

Učenje velikih jezikovnih modelov

1. Predučenje *(pretraining)*
2. Učenje za dialoge
3. Učenje človeških preferenc

Učenje velikih jezikovnih modelov

1. **Predučenje *(pretraining)***
2. Učenje za dialoge
3. Učenje človeških preferenc

PREDUČENJE:

Cilji:

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

Podatki:

- splošen jezik (Internet, knjige, …)

Modeli:

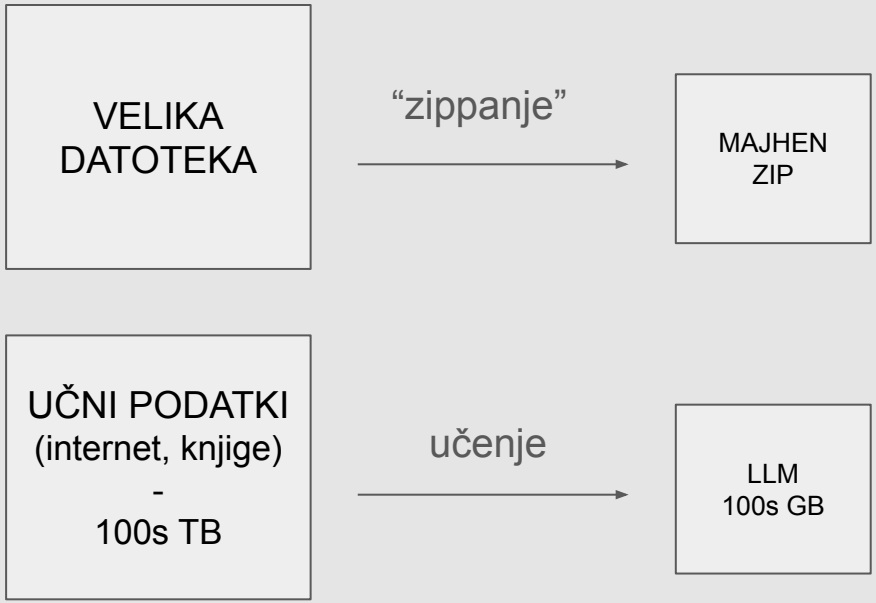
- GPT 1, 2, 3, 3.5

Učenje velikih jezikovnih modelov

1. **Predučenje *(pretraining)***
2. Učenje za dialoge
3. Učenje človeških preferenc

PREDUČENJE:

Analogija:



Učenje velikih jezikovnih modelov

1. Predučenje *(pretraining)*
2. **Učenje za dialoge**
3. Učenje človeških preferenc

UČENJE ZA DIALOGE:

Cilji:

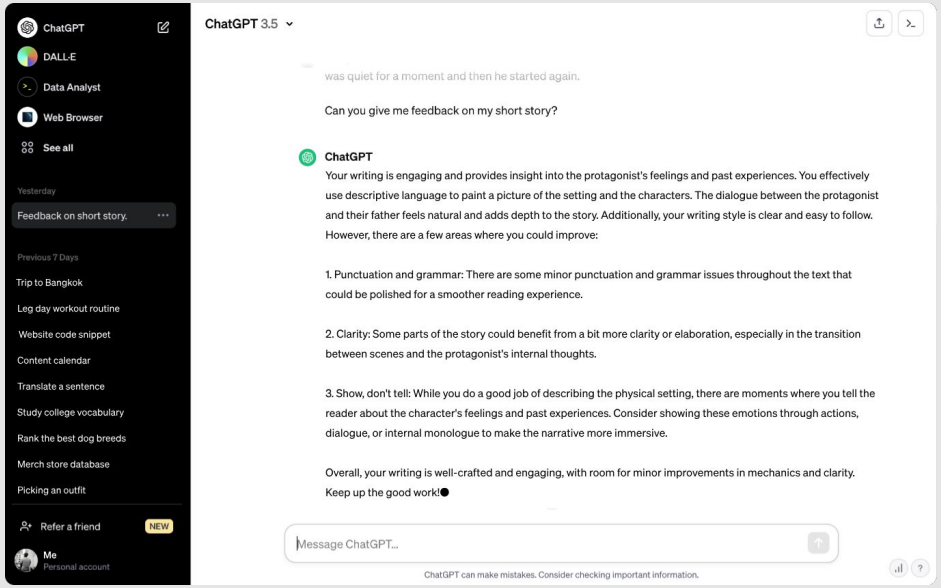
- LLM-ji razumejo interakcije v obliki dialogov;

Podatki:

- kurirani dialogi;
- učna naloge še zmeraj napoved naslednje besede.

Učenje velikih jezikovnih modelov

1. Predučenje *(pretraining)*
2. **Učenje za dialoge**
3. Učenje človeških preferenc



Učenje velikih jezikovnih modelov

1. Predučenje *(pretraining)*
2. Učenje za dialoge
3. **Učenje človeških preferenc**

The New York Times

Artificial Intelligence › | A.I. Faces Quiz | How the A.I. Race Began | Key Figures in the Field | One Year of ChatGPT

Microsoft Created a Twitter Bot to Learn From Users. It Quickly Became a Racist Jerk.

MICROSOFT / WEB / TL;DR

Twitter taught Microsoft's AI chatbot to be a racist asshole in less than a day

Učenje velikih jezikovnih modelov

1. Predučenje *(pretraining)*
2. Učenje za dialoge
3. **Učenje človeških preferenc**

UČENJE ČLOVEŠKIH PREFERENC:

Cilji:

- varnostni mehanizem;
- mitigacija pristranskosti, toksičnosti in druge škodljive vsebine.

Podatki:

- označene človeške preference.

Priložnosti vpeljave v podjetjih

Dimenzije uporabe

- priložnosti
- končni uporabniki
- nivo zahtevnosti
- uporaba

Dimenzije uporabe

- **priložnosti**
- končni uporabniki
- nivo zahtevnosti
- uporaba

produktivnost/stroški	**prihodki**
- večanje produkivnosti zaposlenih: izboljšanje sposobnosti ustvarjanja, urejanja besedil, slik, … (povzemanje, prevajanje besedil, preverjanje kode); - izboljšanje procesov: (pol)avtomatizacija rutinskih nalog.	- razvoj novih produktov (od raziskav trga do pisanja kode in pitchev); - nadgradnje obstoječih produktov/storitev.

Dimenzije uporabe

- priložnosti
- **končni uporabniki**
- nivo zahtevnosti
- uporaba

za interno uporabo	**za zunanje stranke**
- osebni asistent (pisanje emailov, pitchev, poročil, …); - avtomatizacija rutinskih opravil; - za izboljšavo delovnih procesov;	- nov uporabniški vmesnik za interakcijo s produktom; - nadgradnja uporabniške izkušnje; - …

Dimenzije uporabe

- priložnosti
- končni uporabniki
- **nivo zahtevnosti**
- uporaba

"off-the-shelf"	**nadgrajeno AI orodje**
- ChatGPT aplikacija; - GitHub copilot; - …	- LLM-je lahko nadgradimo z internimi podatki; - LLM-ji lahko komunicirajo z našimi aplikacijami (preko API vmesnikov); - …

Dimenzije uporabe

- priložnosti
- končni uporabniki
- nivo zahtevnosti
- **uporaba**

Ustvarjanje besedil	**Transformacija besedil**
- Pisanje mailov, - poročil, - dokumentacij, - ...	- povzemanje (dolg dokument -> povzetek); - klasifikacija (besedilo -> kategorija); - prevajanje; - ekstrakcija (besedilo -> strukturirane informacije)

Koraki uvajanja

1. Ovrednotenje notranjih zmogljivosti in zrelosti UI
2. Razumevanje UI, določanje poslovnih potreb in prioritizacija primerov uporabe UI
3. Razvijanje spretnosti ter uporaba znanja in storitev, ki so na voljo prek omrežij in poslovnih združenj
4. Priprava podatkov za podporo rešitvam UI
5. Dokvalifikacija lastne ekipe ali vzpostavitev sodelovanja s partnerji s področja UI in podatkov
6. Obvladovanje stroškov s pomočjo partnerstev

Koraki uvajanja

1. Ovrednotenje notranjih zmogljivosti in zrelosti UI

2. Razumevanje UI, določanje poslovnih potreb in prioritizacija primerov uporabe UI

Začnite interno:

“off-the-shelf” orodje za interno uporabo

Primera:

- *ChatGPT aplikacija*: asistent pri delu - pisanje emailov, tvorjenje poročil, …
- *GitHub Copilot*: asistent za pisanje programske kode.

Koraki uvajanja

1. Ovrednotenje notranjih zmogljivosti in zrelosti UI

2. Razumevanje UI, določanje poslovnih potreb in prioritizacija primerov uporabe UI

Začnite interno:

Določite smernice uporabe / interni pravilnik:

- obseg uporabe (katera orodja in za kaj);

- ne vnašajte:
    - osebnih podatkov;
    - poslovnih skrivnosti;
    - intelektualne lastnine;
    - kakršnihkoli občutljivih informacij.

Koraki uvajanja

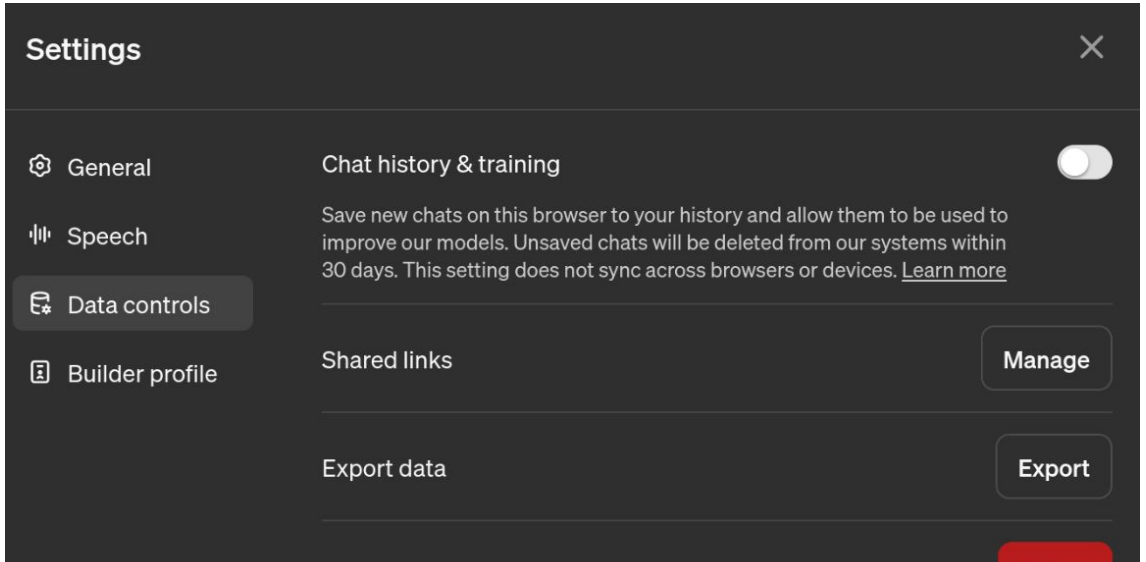
1. Ovrednotenje notranjih zmogljivosti in zrelosti UI

2. Razumevanje UI, določanje poslovnih potreb in prioritizacija primerov uporabe UI

Začnite interno:

Določite smernice uporabe / interni pravilnik:

- izklopite zgodovino;



Koraki uvajanja

1. Ovrednotenje notranjih zmogljivosti in zrelosti UI

2. Razumevanje UI, določanje poslovnih potreb in prioritizacija primerov uporabe UI

Začnite interno:

Določite smernice uporabe / interni pravilnik:

- odgovore/besedila obravnavajte kot izhodišče:
    - Proofread;
    - Fact-check;
    - ...

Koraki uvajanja

4. Priprava podatkov za podporo rešitvam UI

Ko ste na točki, ko:

- razumete delovanje in potencial LLM-jev;
- imate nadaljne primere uporabe;

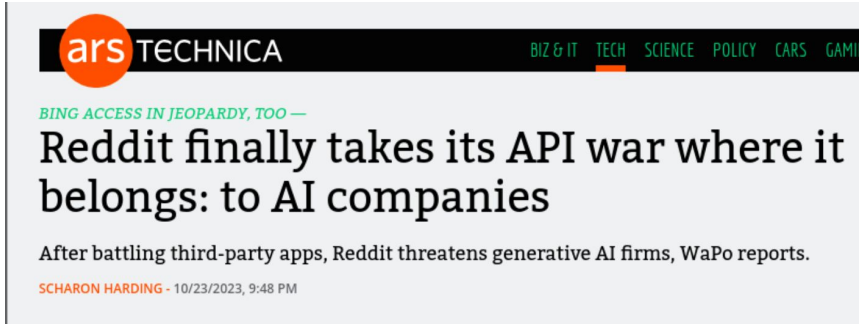
lahko začnete razmišljat o **nadgrajevanju** LLM-jev.

S čim lahko obogatite osnoven ChatGPT?

Koraki uvajanja

4. Priprava podatkov za podporo rešitvam UI

Podatki so zlato!



Koraki uvajanja

4. Priprava podatkov za podporo rešitvam UI

Uporaba vaših podatkov (internih, kuriranih):

- za interno uporabo:
  - chatbot, ki pomaga novim zaposlenim pri procesu uvajanja (podatki: interni pravilniki, dokumentacija …)
- za zunanje stranke:
  - chatbot, ki kupcem odgovarja na vprašanja o izdelku na spletni trgovini (dokumentacija izdelka, podobni izdelki …)

Koraki uvajanja

3. Razvijanje spretnosti ter uporaba znanja in storitev, ki so na voljo prek omrežij in poslovnih združenj

5. Dokvalifikacija lastne ekipe ali vzpostavitev sodelovanja s partnerji s področja UI in podatkov

- Delavnice;
- Povezovanje s partnerji s področja UI.

Pasti vpeljave (in ublažitev)

1. Statičnost

= LLM-ji ne morejo odgovarjati na vprašanja, katerih odgovorov niso videli.

Težava so učni podatki!

Posledice statičnosti

1. Nedavni dogodki
2. Specifične domene
3. Privatni podatki
4. Halucinacije

Posledice statičnosti

1. **Nedavni dogodki**
2. Specifične domene
3. Privatni podatki
4. Halucinacije

LLM-ji so naučeni na podatkih zbranih do neke točke v času.

Primer:

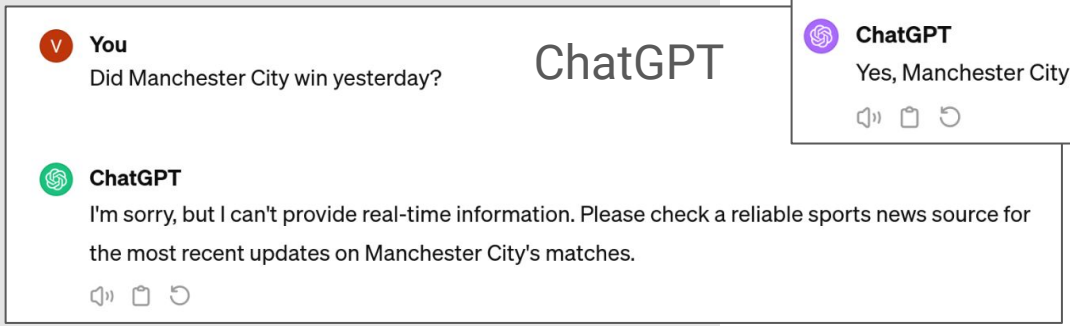
ChatGPT: *I am sorry I can't answer you question, my training data cutoff point was September 2021.*

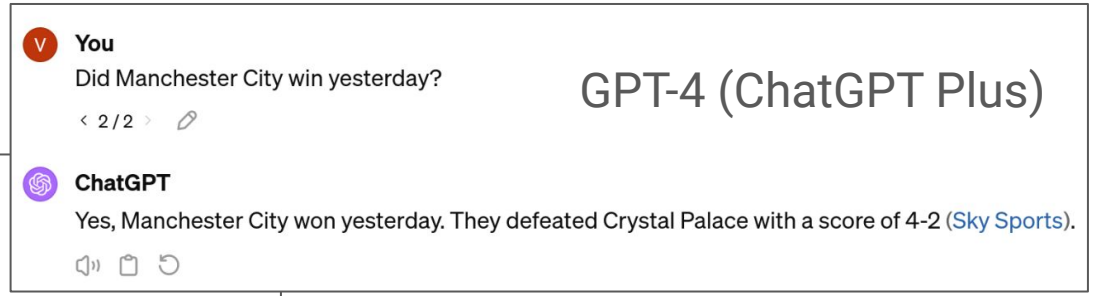
Posledice statičnosti

1. **Nedavni dogodki**
2. Specifične domene
3. Privatni podatki
4. Halucinacije

Rešitev:

- Omogočiti LLM-jem dostop do podatkov v realnem času *(real-time data)*





Posledice statičnosti

1. Nedavni dogodki
2. **Specifične domene**
3. Privatni podatki
4. Halucinacije

Bolj kot je neka domena nišna, manj je informacij v učnih podatkih.

Primer:

- Pravo
  - Intelektualna lastnina
    - Patenti
      - Biotehnološki patenti
        - Pravni spor glede patenta za CRISPR

Posledice statičnosti

1. Nedavni dogodki
2. **Specifične domene**
3. Privatni podatki
4. Halucinacije

REŠITVE:

- Dodatno učenje LLM-jev na domensko-specifičnih podatkih *(finetuning)*
- Boljša navodila*:*
  - Dodajanje informacij v navodila *(RAG)*
  - Boljše pozivanje *(prompting)*

↓

Enostaven primer:

“You are an IT specialist”

Posledice statičnosti

1. Nedavni dogodki
2. Specifične domene
3. **Privatni podatki**
4. Halucinacije

LLM-ji so naučeni na javno dostopnih podatkih.

PRIMER:

Če želimo, da LLM-ji odgovarjajo na vprašanja o našem podjetju.

Posledice statičnosti

1. Nedavni dogodki
2. Specifične domene
3. **Privatni podatki**
4. Halucinacije

REŠITEV:

- Boljša navodila:
  - Boljše pozivanje *(prompting)*
  - Dodajanje informacij v navodila *(RAG)*

Andrej: Kako mi je ime?

ChatGPT: Kako naj bi to vedel?

Andrej: Ime mi je Andrej. Kako mi je ime?

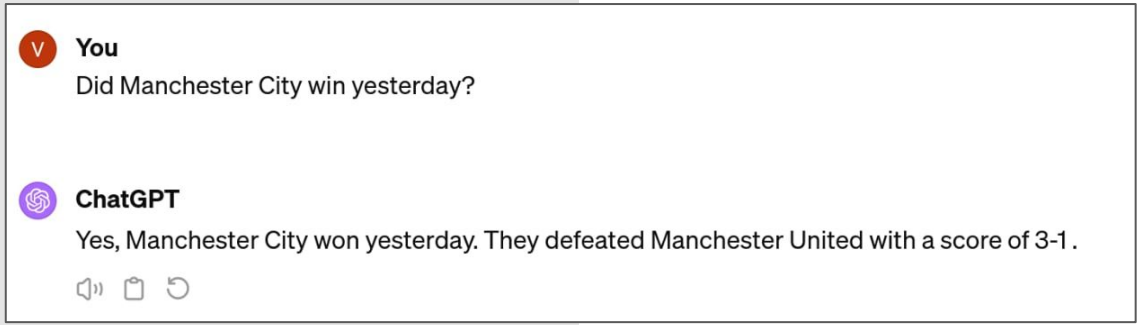
ChatGPT: Andrej.

Posledice statičnosti

1. Nedavni dogodki
2. Specifične domene
3. Privatni podatki
4. **Halucinacije**

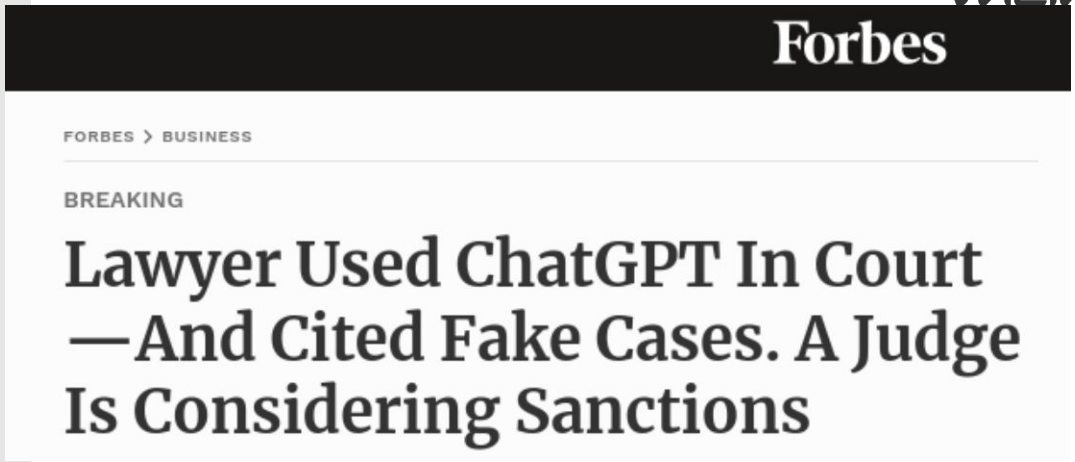
Halucinacije v LLM-jih so samozavestno generiranje napačnih informacij, ki se na prvi pogled zdijo verjetne.

- Poleg nevidenih informacij (1., 2., 3. na levi), se v učnih podatkih pojavlja tudi ogromno napak.



Posledice statičnosti

1. Nedavni dogodki
2. Specifične domene
3. Privatni podatki
4. **Halucinacije**



Posledice statičnosti

1. Nedavni dogodki
2. Specifične domene
3. Privatni podatki
4. **Halucinacije**

REŠITVE:

- Bolj deskriptivna navodila *(prompting)*
- Dodajanje informacij v navodila - prizemljitev *(RAG)*

Veliko izboljšav lahko dosežemo z boljšim promptingom

Izoblikovala se je vloga *prompt inženirja (prompt engineer)*.

Prompt = poziv, ukaz, navodilo?
Engineer = inženir
Prompt engineer = inženir za pozive, inženir za navodila?

Veliko izboljšav lahko dosežemo z boljšim promptingom

Izoblikovala se je vloga *prompt inženirja (prompt engineer)*

Prompt = poziv, ukaz, navodilo?
Engineer = inženir
Prompt engineer = ~~inženir za pozive, inženir za navodila?~~

Terapevt za jezikovne modele

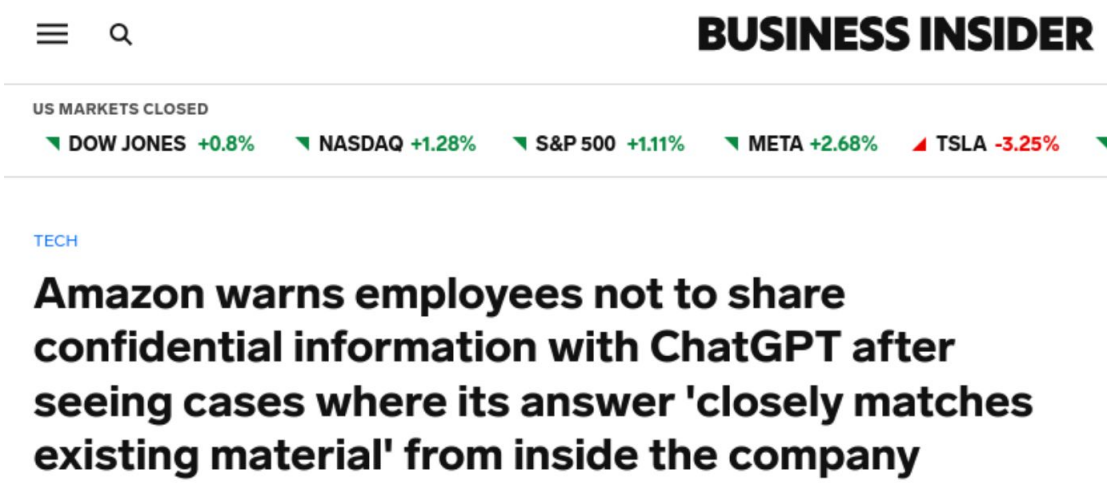
Prompt inženiring

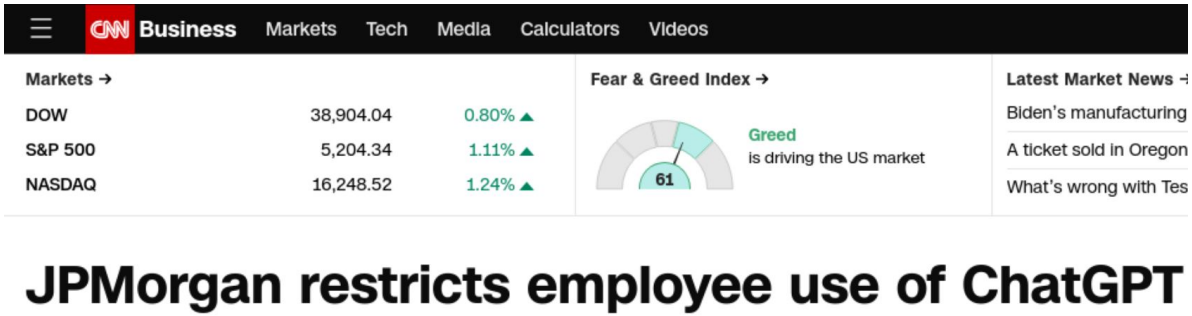
= učinkovita komunikacija z velikimi jezikovnimi modeli

2. Privatnost

“If you are not paying for the product, you are the product.”

2. Privatnost





2. Privatnost

REŠITVE:

- Interna politika uporabe AI orodij:
  - Deljenje zaupnih podatkov, poslovnih skrivnosti, osebnih podatkov…
  - Izklopljena zgodovina in izklopljena možnost učenja na vnešenih podatkih.

- Azure OpenAI services (Microsoft Azure politika)
- Lastna postavitev odprtokodnih LLM-jev *(self-hosted)*

3. Stroški (uporabe)

- število uporabnikov;
- naloga LLM-ja.

MITIGACIJA:

- boljša navodila - manj uporabljenih besed *(prompting)*
- uporaba manjših modelov (e.g. ChatGPT-3.5 namesto GPT-4)
- lastna postavitev odprtokodnih LLM-jev *(self-hosted)*

Pregled velikih jezikovnih modelov


ChatGPT


Gemini


LLaMA
by Meta


Claude

Primerjava

- Zmogljivost
- Odprtokodni vs. lastniški
- Cena
- Število besed

Primerjava

- **Zmogljivost**
- Odprtokodni vs. lastniški
- Cena
- Število besed

LMSYS ChatBot arena

Rank	Model	Arena Elo	95% CI	Votes	Organization	License	Knowledge Cutoff
1	Claude 3 Opus	1255	+3/-4	37663	Anthropic	Proprietary	2023/8
1	GPT-4-1106-preview	1252	+3/-3	56936	OpenAI	Proprietary	2023/4
1	GPT-4-0125-preview	1249	+3/-4	38105	OpenAI	Proprietary	2023/12
4	Bard (Gemini Pro)	1204	+5/-5	12468	Google	Proprietary	Online
4	Claude 3 Sonnet	1200	+3/-4	40389	Anthropic	Proprietary	2023/8
6	GPT-4-0314	1185	+4/-4	35803	OpenAI	Proprietary	2021/9
7	Claude 3 Haiku	1177	+3/-4	26773	Anthropic	Proprietary	2023/8
8	GPT-4-0613	1160	+3/-5	54509	OpenAI	Proprietary	2021/9
8	Mistral-Large-2402	1157	+5/-4	28356	Mistral	Proprietary	Unknown
9	Qwen1.5-72B-Chat	1149	+4/-5	21981	Alibaba	Qianwen LICENSE	2024/2
10	Claude-1	1146	+4/-5	21868	Anthropic	Proprietary	Unknown
10	Mistral Medium	1146	+4/-5	27059	Mistral	Proprietary	Unknown
10	Command R	1146	+5/-6	12739	Cohere	CC-BY-NC-4.0	2024/3
14	Gemini Pro (Dev API)	1127	+4/-4	16041	Google	Proprietary	2023/4
14	Claude-2.0	1127	+5/-5	13484	Anthropic	Proprietary	Unknown

Primerjava

- **Zmogljivost**
- Odprtokodni vs. lastniški
- Cena
- Število besed

Kako zmogljiv model zares potrebujemo?

- za splošno uporabo (asistenti, chatboti)
- za specializirano uporabo (transformacija besedil)

Primerjava

- Zmogljivost
- **Odprtokodni vs. lastniški**
- Cena
- Število besed

Lastniški jezikovni modeli:

- OpenAI: ChatGPT (3.5), GPT-4
- Google: Gemini (Pro, Ultra)
- Antropic: Claude (1, 2.1, 3: Opus, Sonnet, Haiku)

Odprtokodni:

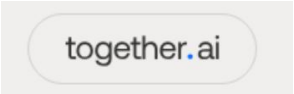
- Meta: LLama 2 (7B, 13B, 34B, 70B)
- Mistral: Mistral-7B 2

Primerjava

- Zmogljivost
- Odprtokodni vs. lastniški
- **Cena**
- Število besed

- Lastniški jezikovni modeli:
  - Plačilo za API klice;
  - na 1k oz. 1M tokenov *(žetonov?)*.

- Odprtokodni jezikovni modeli:
  - Lastna postavitev *(on-premise, cloud)*;
  - Postavljena rešitev (together.ai, …)



Primerjava

- Zmogljivost
- Odprtokodni vs. lastniški
- Cena
- **Število besed**

Število besed, ki jih jezikovni model lahko obdeluje.

Dolžina konteksta = dolžina vhoda + dolžina izhoda

Vprašanja?