



Konferencija Srce DEI

Kako izgraditi vlastiti veliki jezični model? Od prikupljanja podataka do treniranja na HPC infrastrukturi

dr. sc. tech. David Dukić

Sveučilište u Zagrebu

Fakultet elektrotehnike i računarstva

TakeLab



SVEUČILIŠTE U ZAGREBU
Fakultet
elektrotehnike
i računarstva



Srce DEI 2026



TakeLab



- Laboratorij za obradu prirodnog jezika @FER
- Istraživački interesi:
 - računalna društvena znanost
 - učenje reprezentacija modela
 - interpretabilnost i sigurnost modela
- Više o nama → takelab.fer.hr
- Kontaktirajte nas → david.dukic@fer.hr



Ana Barić



Marko Čuljak



Laura Majer



Iva Vukojević



David Dukić



Josip Jukić



Martin Tutek



Jan Šnajder



Kako izgraditi vlastiti veliki jezični model?

Prikupljanje i čišćenje podataka



Kako izgraditi vlastiti veliki jezični model?

Prikupljanje i čišćenje podataka

Odabir arhitekture modela i hiperparametara za učenje



Kako izgraditi vlastiti veliki jezični model?

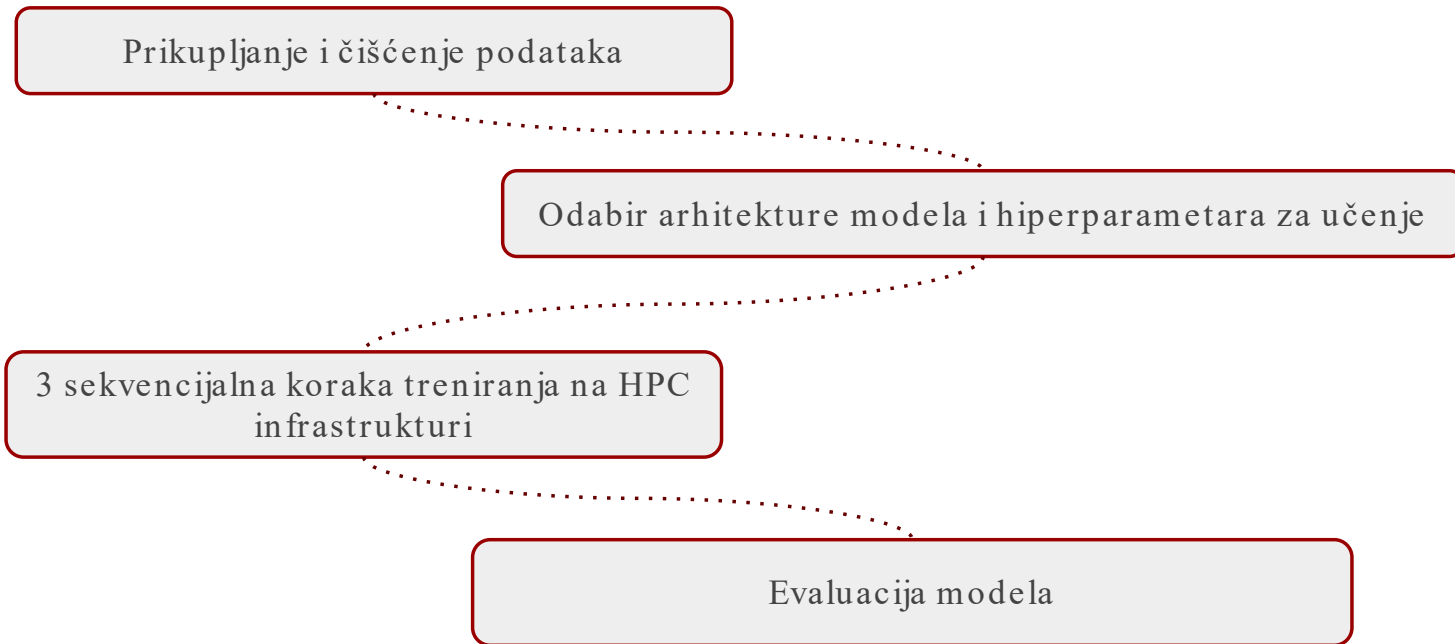
Prikupljanje i čišćenje podataka

Odabir arhitekture modela i hiperparametara za učenje

3 sekvencijalna koraka treniranja na HPC
infrastrukturi



Kako izgraditi vlastiti veliki jezični model?





Kako izgraditi vlastiti veliki jezični model?

Prikupljanje i čišćenje podataka

Odabir arhitekture modela i hiperparametara za učenje

3 sekvencijalna koraka treniranja na HPC infrastrukturi

Evaluacija modela



Back to basics

Back to basics

Large Language Models (LLMs)

Veliki jezični modeli



Vrsta umjetne inteligencije koja generira **prirodan jezik** (npr. engleski, hrvatski, kineski), ali i druge jezike poput programskih jezika

Veliki **jezični modeli**



Vrsta umjetne inteligencije koja generira **prirodan jezik** (npr. engleski, hrvatski, kineski), ali i druge jezike poput programskih jezika

Veliki jezični modeli

Trenirani na **OGROMNOM** skupu tekstnih podataka (knjige, web, novinski članci, ...čitav internet) i sadrže **VELIKI** broj parametara* (milijarde!)

parametar* = gradivna jedinica modela

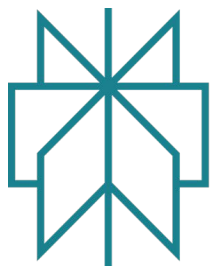


The usual suspects



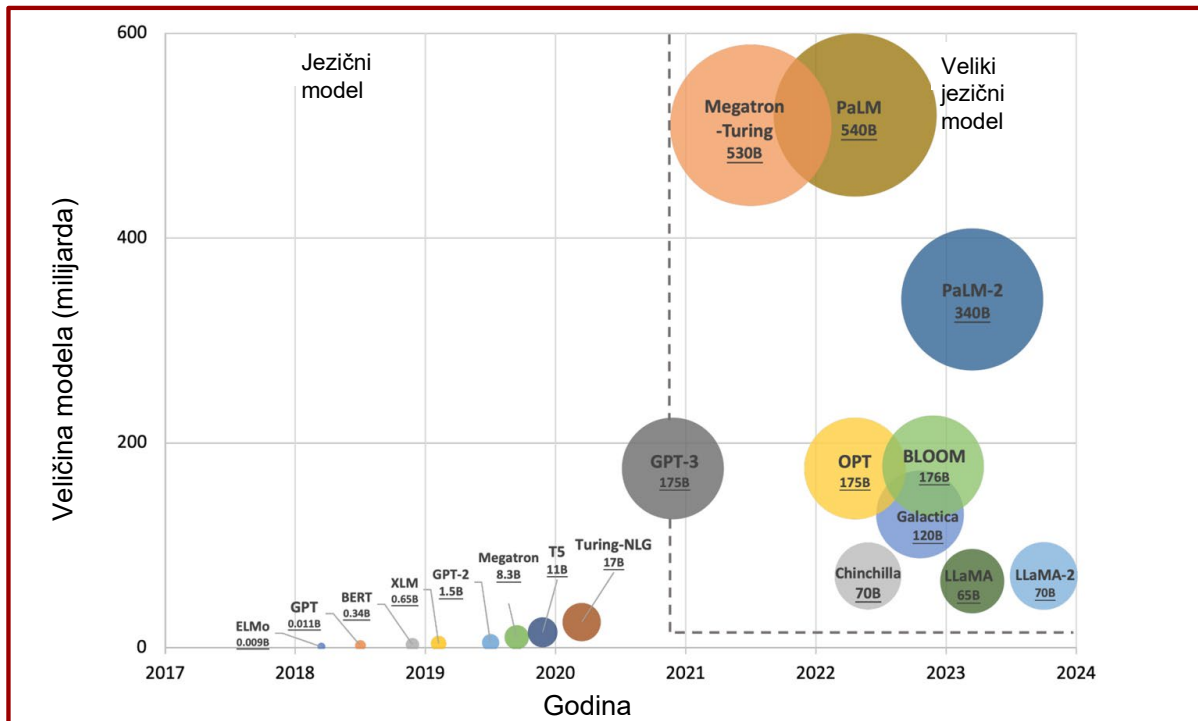
Claude

Gemini



perplexity



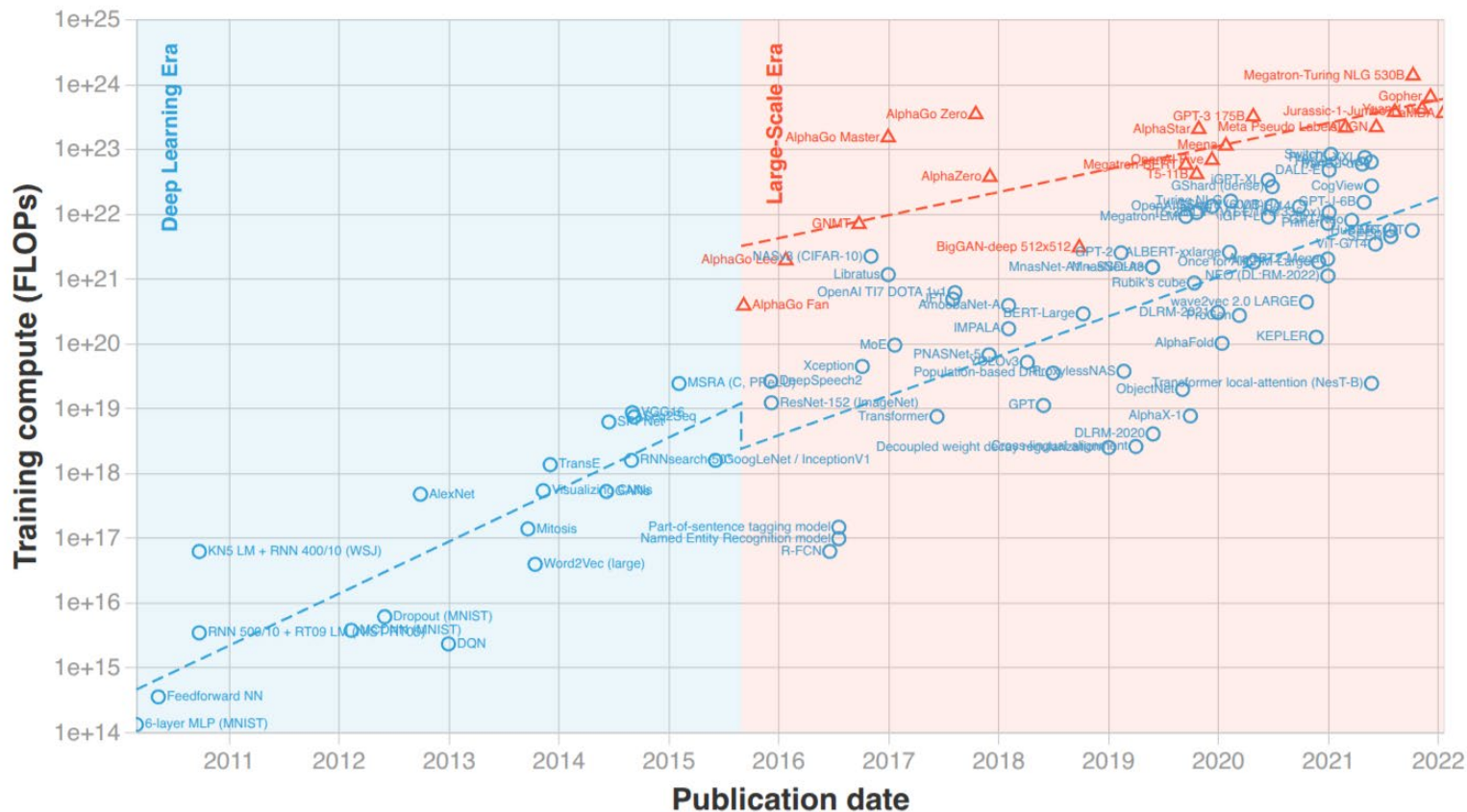


He, Kai, et al. "A survey of large language models for healthcare: from data, technology, and applications to accountability and ethics." Information Fusion 118 (2025): 102963.

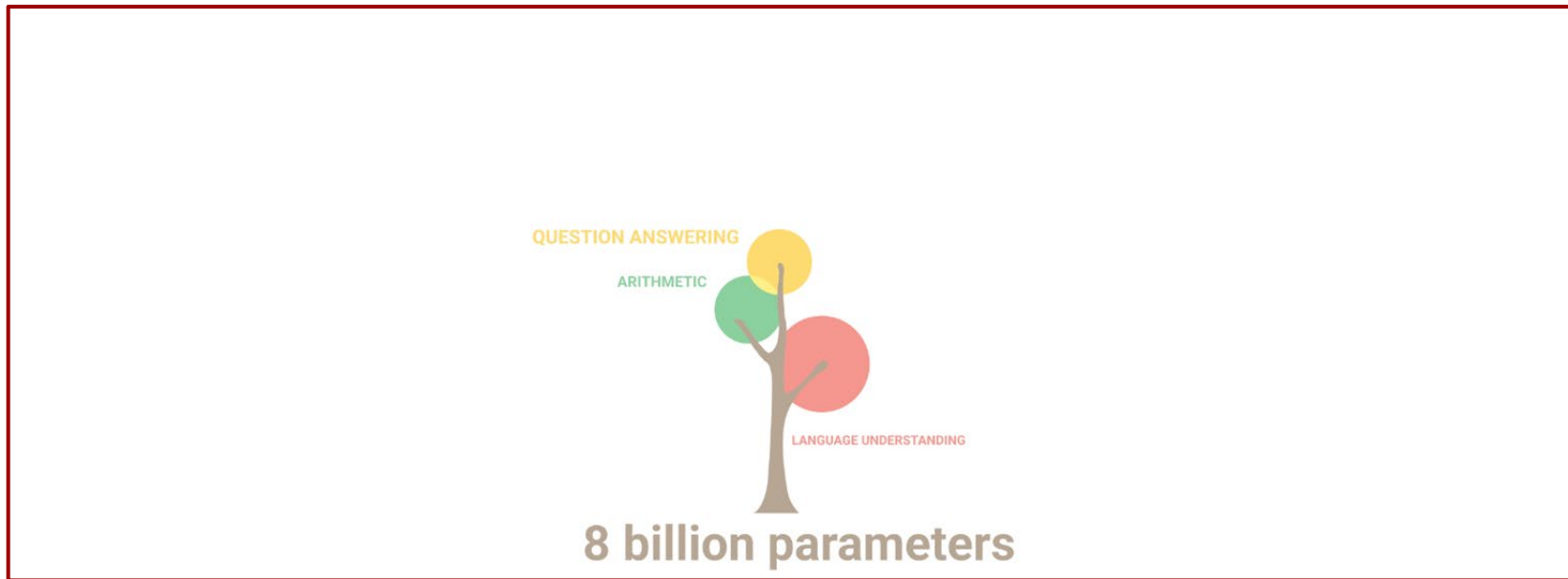
Eksplozivni porast broja parametara

Training compute (FLOPs) of milestone Machine Learning systems over time

n = 102



Sevilla, Jaime, et al. "Compute trends across three eras of machine learning." 2022 international joint conference on neural networks (IJCNN). IEEE, 2022.



Evolucija mogućnosti jezičnih modela

Izvor: [Google AI Blog](#)



Vrste LLM-ova



Otvoreni LLM-ovi

- Kod i težine su **javno dostupni**
- Omogućuju istraživanje

Prednosti: transparentnost, prilagodba, niži troškovi

Nedostaci: zahtijevaju tehničko znanje, manje optimizirani



Gemma

MISTRAL AI

Zatvoreni LLM-ovi

- Kod i podaci **nisu javno dostupni**

Prednosti: stabilnost, sigurnost, komercijalna podrška

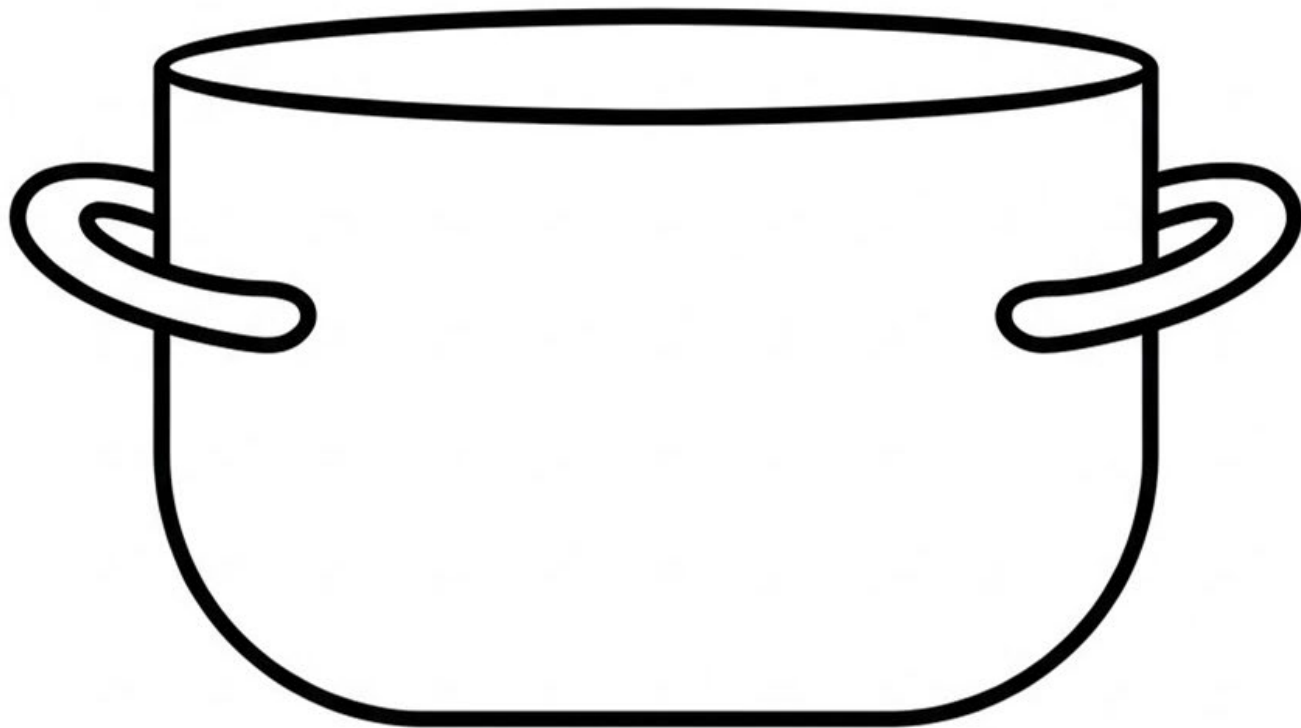
Nedostaci: ovisnost o tvrtki, nema transparentnosti

Claude 

perplexity

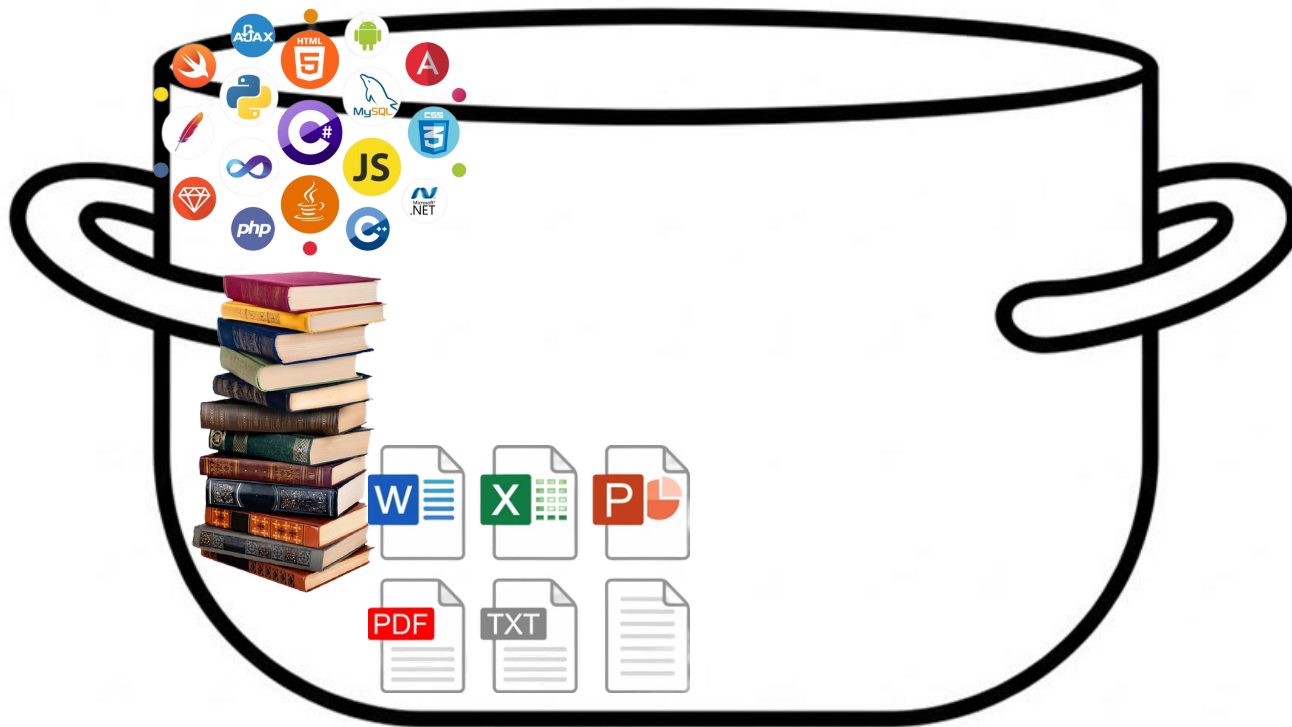
Gemini

Sastojci za izgradnju LLM-a



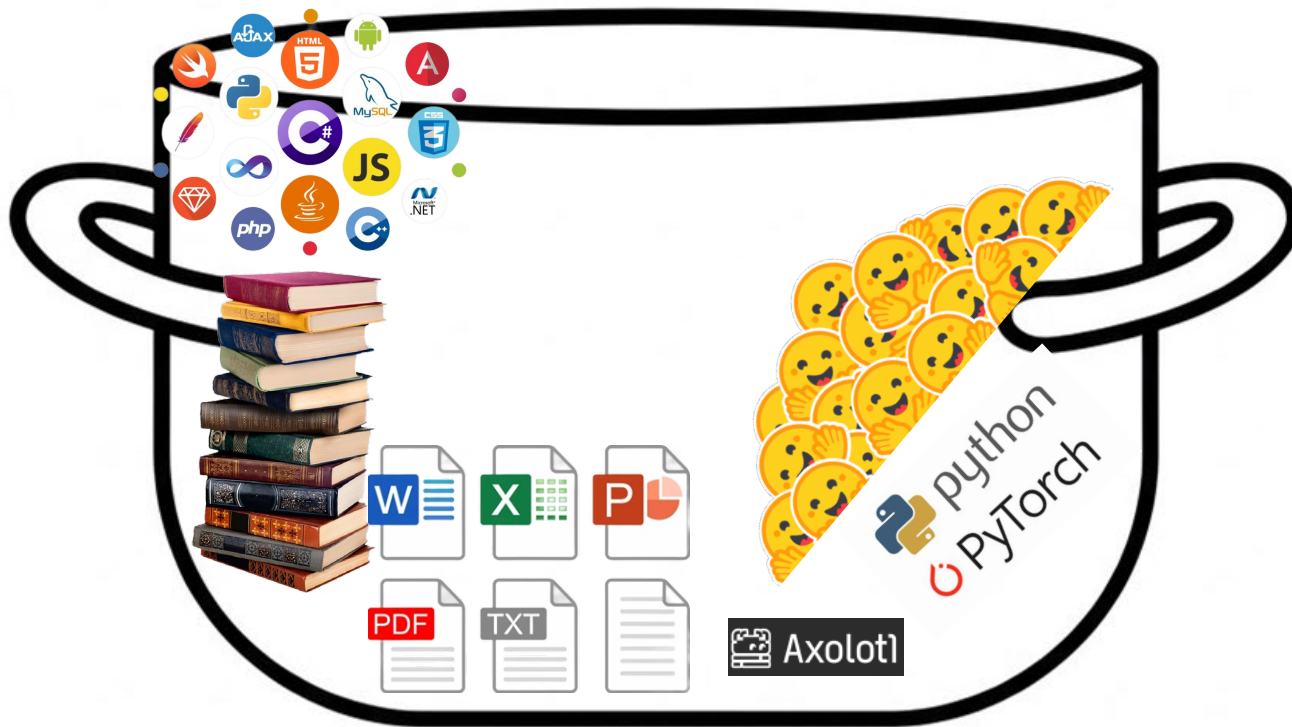


Sastojci za izgradnju LLM-a



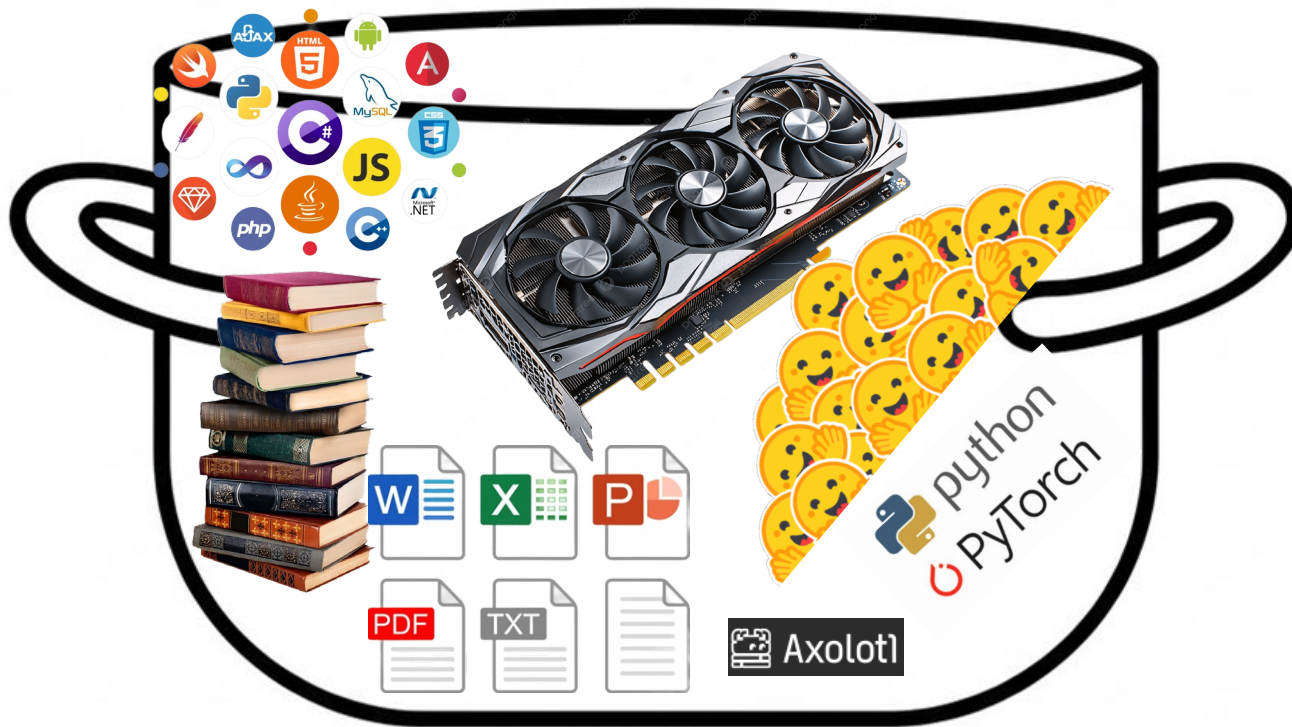


Sastojci za izgradnju LLM-a

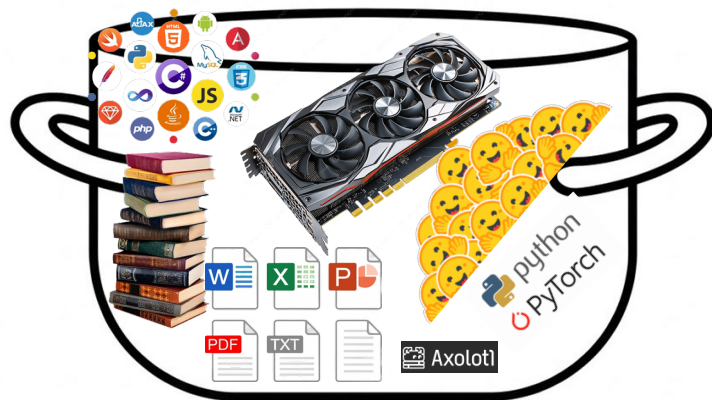


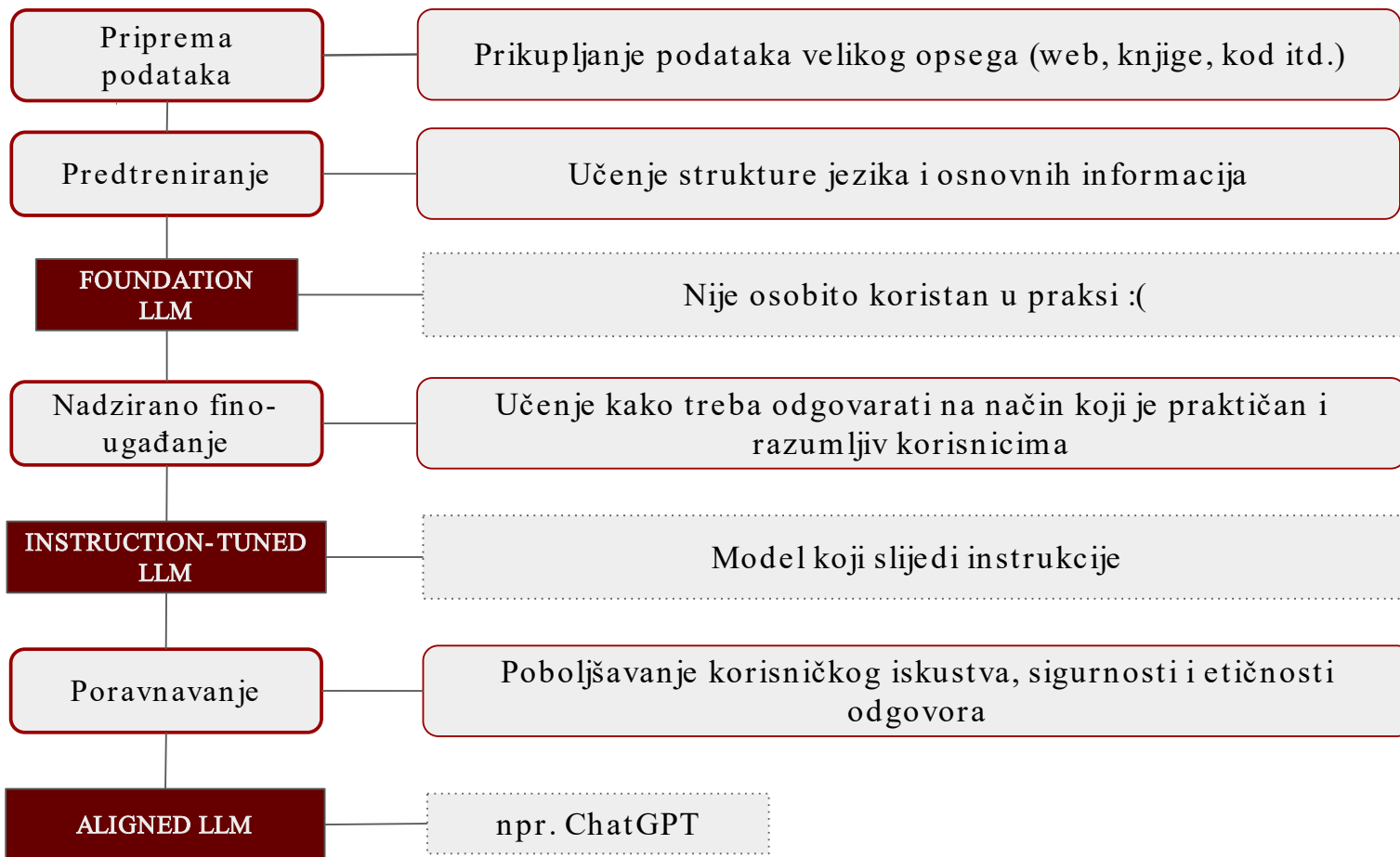


Sastojci za izgradnju LLM-a



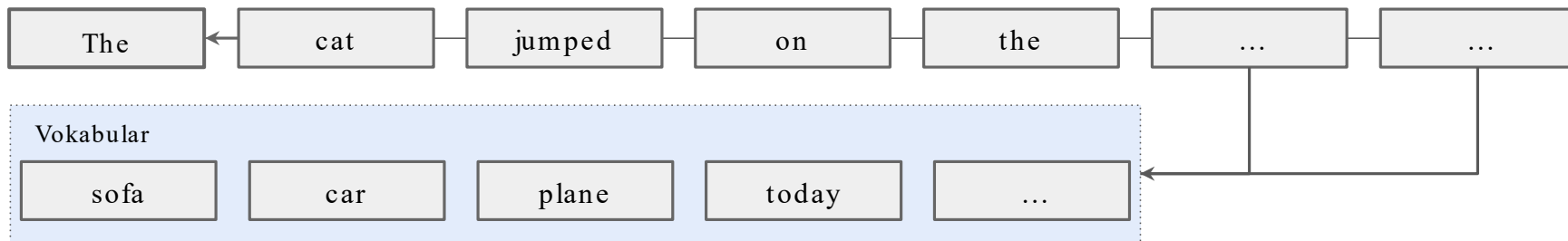
Kuhanje na HPC-u







Kako napraviti temeljni model

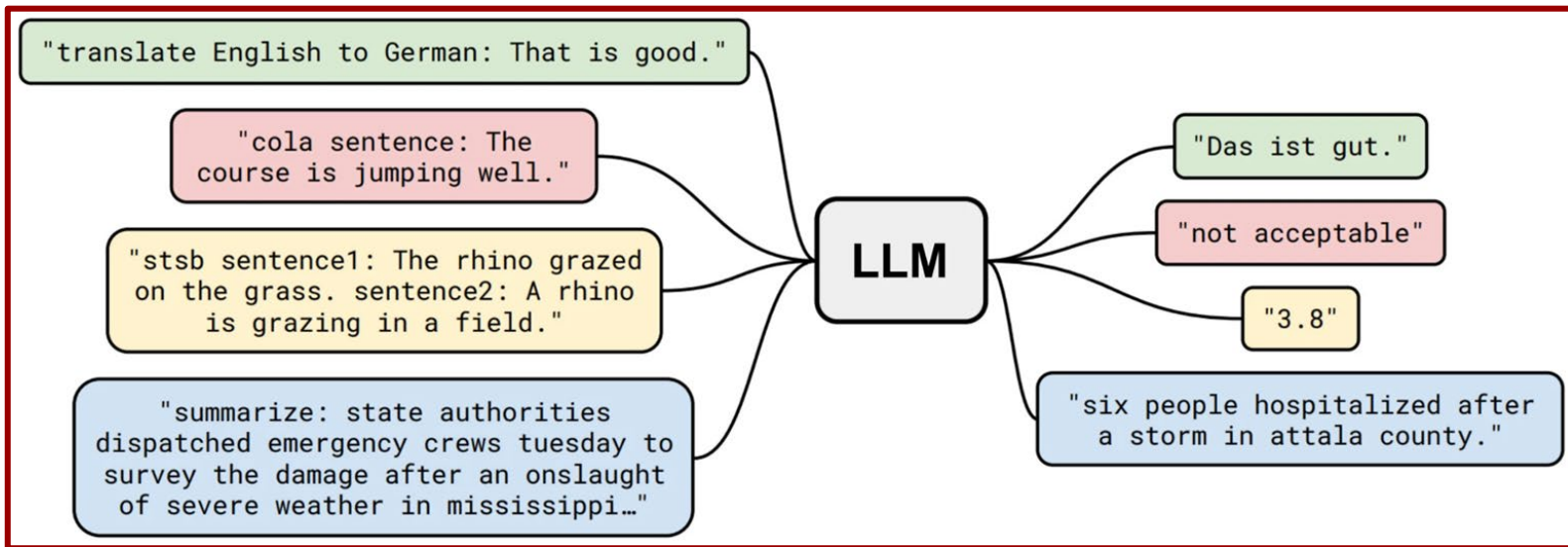


LLM-ovi su trenirani tzv. *samonadziranim* učenjem

Predviđanje iduće najvjerojatnije riječi



Kako napraviti model koji prati instrukcije



Raffel, Colin, et al. "Exploring the limits of transfer learning with a unified text-to-text transformer." Journal of machine learning research 21.140 (2020): 1-67.

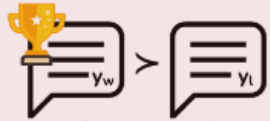
Nadzirano fine-ugađanje Supervised (instruction) fine-tuning



Kako napraviti model koji je poravnat (s našim vrijednostima, očekivanjima, namjerama, stavovima, ...)

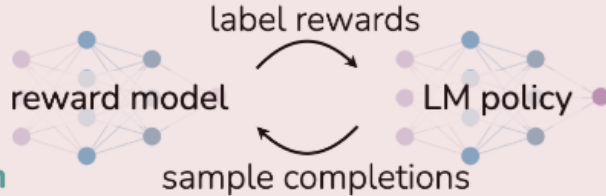
Reinforcement Learning from Human Feedback (RLHF)

x: "write me a poem about
the history of jazz"



preference data

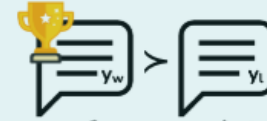
maximum
likelihood



reinforcement learning

Direct Preference Optimization (DPO)

x: "write me a poem about
the history of jazz"



preference data

maximum
likelihood



Rafailov, Rafael, et al. "Direct preference optimization: Your language model is secretly a reward model." Advances in neural information processing systems 36 (2023)

Bez poravnavnja, model može davati **netočne, pristrane ili opasne odgovore**. Cilj je stvoriti model koji je **koristan, vjerodostojan i siguran**.



Chinchilla zakon skaliranja



Za fiksni budžet (compute), najbolji model dobijemo **uravnotežujući veličinu modela** (u parametrima P) i **broj podataka za učenje** (u tokenima T)

Empirijski optimum otprilike prati formulu: $T \approx 20 * P$

Za model veličine $P=70B$ (milijardi) trebamo otprilike $T=1.4T$ (bilijuna) tokena.

Hoffmann, Jordan, et al. "Training compute-optimal large language models." *arXiv preprint arXiv:2203.15556* 10 (2022).



Primjer izgradnje hrvatskog LLM-a



Primjer izgradnje hrvatskog LLM- a

Koliko čega trebamo? Podaci, parametri, hiperparametri, grafičke kartice?

Podaci

Što više podataka to bolje. Potrebno je pronaći podatke iz različitih domena na hrvatskom jeziku (novinarstvo, medicina, pravo, znanost, književnost, wikipedija, ..). Potrebno je ubaciti i određen postotak podataka na engleskom jeziku te određen postotak podataka na programskim jezicima. Potrebno je provesti deduplikaciju podataka i očistiti tekstove od netočnog i uvredljivog sadržaja...

Podaci

Što više podataka to bolje. Potrebno je pronaći podatke iz različitih domena na hrvatskom jeziku (novinarstvo, medicina, pravo, znanost, književnost, wikipedija, ..). Potrebno je ubaciti i određen postotak podataka na engleskom jeziku te određen postotak podataka na programskim jezicima. Potrebno je provesti deduplikaciju podataka i očistiti tekstove od netočnog i uvredljivog sadržaja...

Podaci

Što više podataka to bolje. Potrebno je pronaći podatke iz različitih domena na hrvatskom jeziku (novinarstvo, medicina, pravo, znanost, književnost, wikipedija, ...). Potrebno je ubaciti i određen postotak podataka na engleskom jeziku te određen postotak podataka na programskim jezicima. Potrebno je provesti deduplikaciju podataka i očistiti tekstove od netočnog i uvredljivog sadržaja...

Podaci

Što više podataka to bolje. Potrebno je pronaći podatke iz različitih domena na hrvatskom jeziku (novinarstvo, medicina, pravo, znanost, književnost, wikipedija, ..). Potrebno je ubaciti i određen postotak podataka na engleskom jeziku te određen postotak podataka na programskim jezicima. Potrebno je provesti deduplikaciju podataka i očistiti tekstove od netočnog i uvredljivog sadržaja...



Sakupljene podatke je potrebno deduplicirati i tokenizirati



Sakupljene podatke je potrebno deduplicirati i tokenizirati

**Deduplikacija i tokenizacija zahtijevaju više od 512GB RAM-a
za nekoliko desetaka milijardi tokena**



Sakupljene podatke je potrebno deduplicirati i tokenizirati

**Deduplikacija i tokenizacija zahtijevaju više od 512GB RAM-a
za nekoliko desetaka milijardi tokena**

Učimo s 50 milijardi dedupliciranih tokena (90% hrv/ 10% eng + kod)!

Parametri i hiperparametri



Za $T=50B$ tokena idealan je model veličine $P=2.5B$ parametara kako bismo izbjegli podnaučenost.

Parametri i hiperparametri



Za $T=50B$ tokena idealan je model veličine $P=2.5B$ parametara kako bismo izbjegli podnaučenost.



Gemma4-E2B, **2.3B** efektivnih parametara, kontinuirano predtreniranje (engl. continual pretraining), mijenjamo sve parametre modela.

Parametri i hiperparametri



Za $T=50B$ tokena idealan je model veličine $P=2.5B$ parametara kako bismo izbjegli podnaučenost.



Gemma4-E2B, **2.3B** efektivnih parametara, kontinuirano predtreniranje (engl. continual pretraining), mijenjamo sve parametre modela.



```
num_epochs: 1 sequence_len: 4096 per_device_micro_batch_size: 4
gradient_accumulation_steps: 2 learning_rate: 2e-4
lr_scheduler: cosine warmup_ratio: 0.01 weight_decay: 0.1
optimizer: adamw_bnb_8bit bf16: true flash_attention: true
```

HPC resursi

Prema Kaplan et al., 2020, FLOPs procjena iznosi $\approx 6 * P * T \approx 7 \times 10^{20}$

Kaplan, Jared, et al. "Scaling laws for neural language models." *arXiv preprint arXiv:2001.08361* (2020).

HPC resursi

Prema Kaplan et al., 2020, FLOPs procjena iznosi $\approx 6 * P * T \approx 7 \times 10^{20}$

Kaplan, Jared, et al. "Scaling laws for neural language models." *arXiv preprint arXiv:2001.08361* (2020).

Kada bismo koristili cijeli Supek (80 A100 kartica s 40 GB VRAM-a):

1 A100 $\approx 156 \times 10^{12}$ model FLOPs utilizacija (MFU)

Chowdhery, Aakanksha, et al. "Palm: Scaling language modeling with pathways." *Journal of machine learning research* 24.240 (2023): 1-113.

80 A100 $\approx 1.25 \times 10^{16}$

Treniranje modela trajalo bi 56000 sekundi \approx 16 sati

HPC resursi

Prema Kaplan et al., 2020, FLOPs procjena iznosi $\approx 6 * P * T \approx 7 \times 10^{20}$

Kaplan, Jared, et al. "Scaling laws for neural language models." *arXiv preprint arXiv:2001.08361* (2020).

Kada bismo koristili cijeli Supek (80 A100 kartica s 40 GB VRAM-a):

1 A100 $\approx 156 \times 10^{12}$ model FLOPs utilizacija (MFU)

Chowdhery, Aakanksha, et al. "Palm: Scaling language modeling with pathways." *Journal of machine learning research* 24.240 (2023): 1-113.

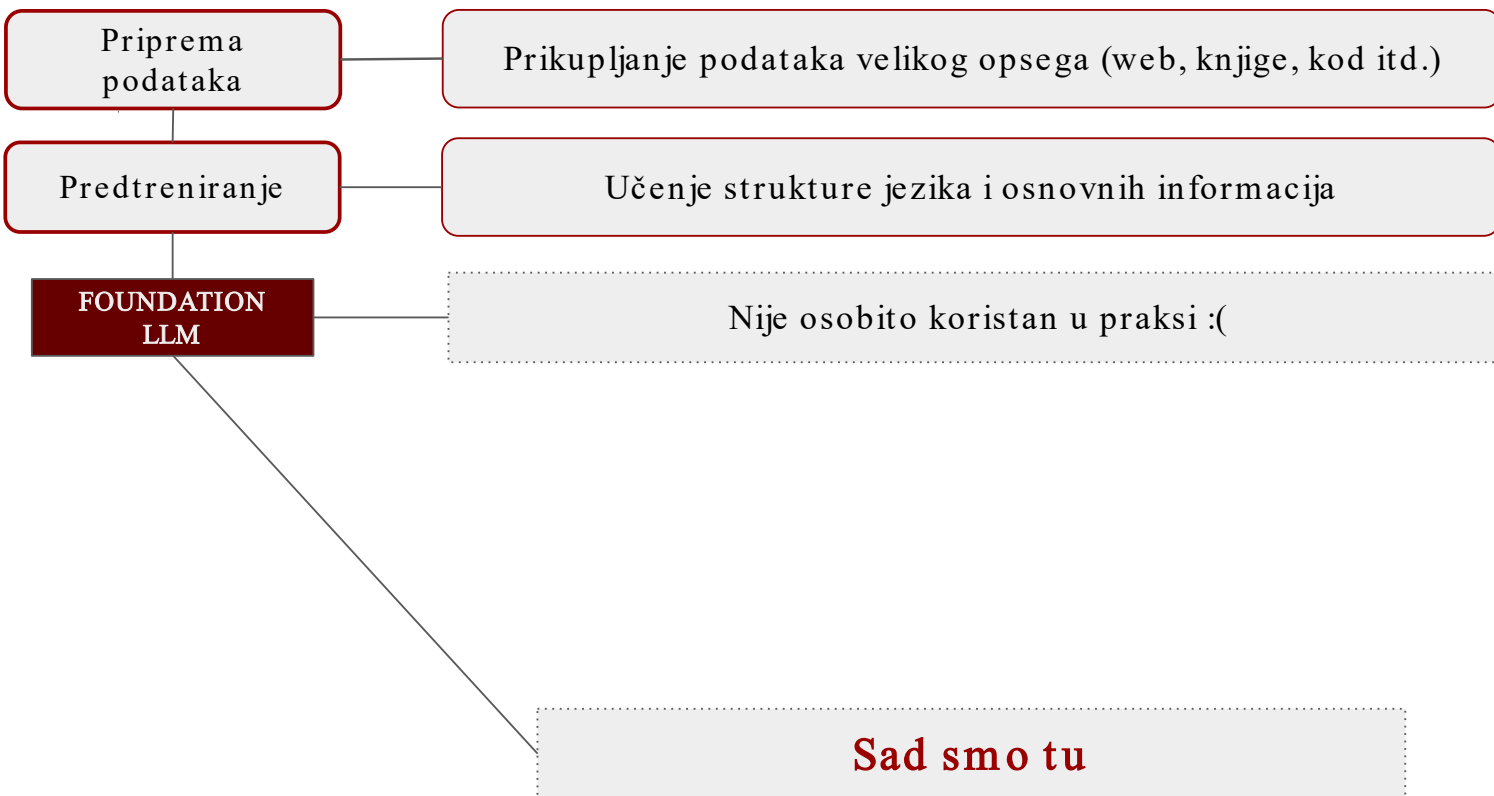
80 A100 $\approx 1.25 \times 10^{16}$

Treniranje modela trajalo bi 56000 sekundi \approx 16 sati

Za usporedbu: treniranje modela na 1 čvoru (4 A100) trajalo bi 13 dana



Kuharica





Supervised fine-tuning (SFT)

Je li moguće stvoriti
tamu energiju?

PROMPT

Tama energija je hipotetski oblik energije za koji se teoretizira da je odgovoran za ubrzano širenje svemira. Iako je njezino postojanje ključno za moderno razumijevanje kozmosa, ona se i dalje nalazi isključivo na razini teorije jer nikada nije izravno opažena niti stvorena u laboratorijskim uvjetima...

ODGOVOR

Datasets: lmsys/lmsys-chat-1m like 874 Follow Large Model Systems O... 889

Modalities: Text Formats: parquet Size: 1M-10M ArXiv: arxiv:2309.11998 Libraries: Datasets Dask

Dataset card Data Studio Files and versions xet Community 23

Dataset Viewer Auto-converted to Parquet API Duplicate Data Studio

Split (1)
train · 1M rows

Search is not available for this dataset

model	conversation	turn	language
wizardlm-13b	[{"content": "I have 1000 documents to download from a website. So as not to overloa..."}]	1	English
vicuna-13b	[{"content": "summarise below transcript \\\nStudent: Chat for help with Field Experienc..."}]	1	English
llama-2-13b-chat	[{"content": "\n\u041e\u043f\u0440\u0435\u0434\u0438\u0442\u0435 \u0432\u0430\u0436\u043d\u0435\u0439\u0448\u0438\u0435 \u0441\u043c\u044b\u0441\u043b\u044b \u0432 \u0442\u0435\u043a\u0441\u0442\u0435 \u043d\u0438\u0436\u0435. \u041a\u0430\u0436\u0434\u044b\u0439 \u0441\u043c\u044b\u0441\u043b \u043e\u043f\u0438\u0448\u0438 \u043e\u0434\u043d\u0438\u043c..."}]	3	unknown
vicuna-13b	[{"content": "Buenas noches!", "role": "user"}, {"content": "Buenas noches! \u00bfEn qu..."}]	8	Spanish
vicuna-13b	[{"content": "hola puedes hablar espa\u00f1ol de argentina?", "role": "user"}, {"content": "..."}]	5	Spanish
vicuna-13b	[{"content": "Please focus on preparing for the college entrance examination again after..."}]	1	English

< Previous 1 2 3 ... 10,000 Next >



Trošak SFT-a

Ponovno učimo svih 2.3B parametara modela, ovaj put s `learning_rate: 1e-5`

Pretpostavimo 100 milijuna tokena

Za $P=2.3B$, $T=100M$, FLOPs procjena iznosi $\approx 6 * P * T \approx 1.38 \times 10^{18}$



Trošak SFT-a

Ponovno učimo svih 2.3B parametara modela, ovaj put s `learning_rate: 1e-5`

Pretpostavimo 100 milijuna tokena

Za $P=2.3B$, $T=100M$, FLOPs procjena iznosi $\approx 6 * P * T \approx 1.38 \times 10^{18}$

Treniranje na jednom čvoru Supeka: 2.5 sati



Trošak SFT-a

Ponovno učimo svih 2.3B parametara modela, ovaj put s `learning_rate: 1e-5`

Pretpostavimo 100 milijuna tokena

Za $P=2.3B$, $T=100M$, FLOPs procjena iznosi $\approx 6 * P * T \approx 1.38 \times 10^{18}$

Treniranje na jednom čvoru Supeka: 2.5 sati

Treniranje na cijelom Supeku: 110 sekundi



Kuharica



Direct preference optimization (DPO)

Je li moguće stvoriti
tamnu energiju?

PROMPT



Tamna energija je hipotetski oblik energije za koji se teoretizira da je odgovoran za ubrzano širenje svemira. Iako je njezino postojanje ključno za moderno razumijevanje kozmosa, ona se i dalje nalazi isključivo na razini teorije jer nikada nije izravno opažena niti stvorena u laboratorijskim uvjetima...

PREFERIRANI
ODGOVOR



Tamna energija je nešto što znanstvenici misle da postoji jer se cijeli svemir širi sve brže i brže. Nitko je još nije stvarno vidio niti je nitko uspio napraviti u laboratoriju, pa je ona zapravo jedna velika zagonetka...

ODBAČEN
I
ODGOVOR



Datasets: HuggingFaceH4/**ultrafeedback_binarized** like 329 Following Hugging Face

Tasks: Text Generation Modalities: Tabular Text Formats: parquet Languages: English Size: 100K

Libraries: Datasets pandas Croissant +1 License: mit

[Dataset card](#) [Data Studio](#) [Files and versions](#) xet [Community](#)

Dataset Viewer Auto-converted to Parquet [API](#) [Embed](#) [Duplicate](#) [Data Studio](#)

Split (6)
train_prefs · 61.1k rows

Search this dataset

	chosen list · lengths	rejected list · lengths	messages list · lengths	score_chose float64
5	[{ "content": "how can i develop a habit of...	[{ "content": "how can i develop a habit of drawing...	[{ "content": "how can i develop a habit of...	
b	[{ "content": "how can I transform the...	[{ "content": "how can I transform the getPosition...	[{ "content": "how can I transform the...	
8	[{ "content": "Given a sentence in French,...	[{ "content": "Given a sentence in French, provide...	[{ "content": "Given a sentence in French,...	
a	[{ "content": "Which animal has two hands, a...	[{ "content": "Which animal has two hands, a hyrax or a...	[{ "content": "Which animal has two hands, a...	
a	[{ "content": "Can you explain more about how...	[{ "content": "Can you explain more about how...	[{ "content": "Can you explain more about how...	
d	[{ "content": "Write a eulogy for a public...	[{ "content": "Write a eulogy for a public figure...	[{ "content": "Write a eulogy for a public...	

< Previous **1** 2 3 ... 612 Next >

Trošak DPO-a

Ponovno učimo svih 2.3B parametara modela, ovaj put s `learning_rate: 1e-7`

Trošak DPO-a

Ponovno učimo svih 2.3B parametara modela, ovaj put s `learning_rate: 1e-7`

Pretpostavimo

100 milijuna tokena \approx 50 tisuća parova (preferirani_odgovor, odbačeni_odgovor)

Za $P=2.3B$, $T=100M$, FLOPs procjena iznosi $\approx 6 * P * T \approx 1.38 \times 10^{18}$

Trošak DPO-a

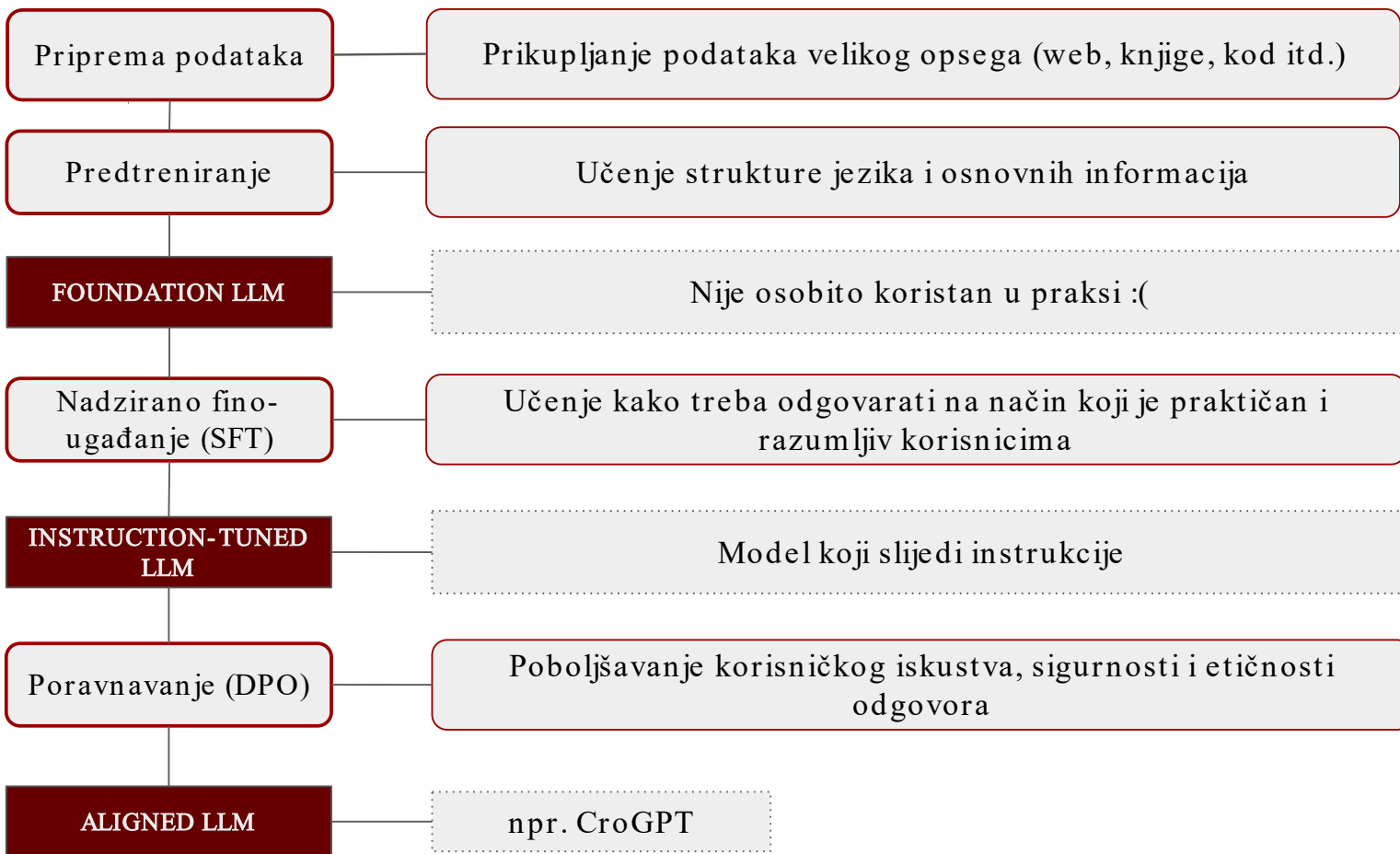
Ponovno učimo svih 2.3B parametara modela, ovaj put s `learning_rate: 1e-7`

Pretpostavimo

100 milijuna tokena \approx 50 tisuća parova (preferirani_ odgovor, odbačeni_ odgovor)

Za $P=2.3B$, $T=100M$, FLOPs procjena iznosi $\approx 6 * P * T \approx 1.38 \times 10^{18}$

Treniranje na jednom čvoru Supeka: **2.5 sati**





O čemu nismo pričali?

- Najteži dio je sakupljanje, čišćenje i filtriranje podataka (micanje uvredljivog sadržaja, netočnih informacija, ...)

O čemu nismo pričali?

- Najteži dio je sakupljanje, čišćenje i filtriranje podataka (micanje uvredljivog sadržaja, netočnih informacija, ...)
- Ne postoje kvalitetni skupovi podataka za SFT i DPO faze na hrvatskom jeziku → potrebno prevesti $\approx 200\text{M}$ tokena



O čemu nismo pričali?

- Najteži dio je sakupljanje, čišćenje i filtriranje podataka (micanje uvredljivog sadržaja, netočnih informacija, ...)
- Ne postoje kvalitetni skupovi podataka za SFT i DPO faze na hrvatskom jeziku → potrebno prevesti $\approx 200\text{M}$ tokena
- Potrebno je provesti detaljnu evaluaciju na standardnim skupovima za evaluaciju performansi modela u svakom koraku učenja (ne postoje hrvatske verzije, potreban prijevod)



O čemu nismo pričali?

- Najteži dio je sakupljanje, čišćenje i filtriranje podataka (micanje uvredljivog sadržaja, netočnih informacija, ...)
- Ne postoje kvalitetni skupovi podataka za SFT i DPO faze na hrvatskom jeziku → potrebno prevesti $\approx 200\text{M}$ tokena
- Potrebno je provesti detaljnu evaluaciju na standardnim skupovima za evaluaciju performansi modela u svakom koraku učenja (ne postoje hrvatske verzije, potreban prijevod)
- Potrebno je isprobati različite hiperparametre, veličine i verzije modela uz evaluaciju na izdvojenom skupu za vrijeme učenja



LinkedIn

david.dukic@fer.hr



Ovo djelo je dano na korištenje pod licencom Creative Commons *Imenovanje* 4.0 međunarodna.

Srce politikom otvorenog pristupa široj javnosti osigurava dostupnost i korištenje svih rezultata rada Srca, a prvenstveno obrazovnih i stručnih informacija i sadržaja nastalih djelovanjem i radom Srca.

www.srce.unizg.hr

creativecommons.org/licenses/by/4.0/deed

www.srce.unizg.hr/otvoreni-pristup

