UNIVERZA V LJUBLJANI | FRI Fakulteta za računalništvo in informatiko

KCUI

Kompetenčni center za umetno inteligenco

Slovenian National AI Competence Centre

Razumevanje velikih jezikovnih modelov za poslovna okolja

prof. dr. Marko Robnik-Šikonja

21. September 2026

REPUBLIKA SLOVENIJA
MINISTRSTVO ZA NOTRANJE ZADEVE IN JAVNO UPRAVO

REPUBLIKA SLOVENIJA
MINISTRSTVO ZA IZOBRAŽEVANJE, ZNANOST IN MLADINO

REPUBLIKA SLOVENIJA
MINISTRSTVO ZA GOSPODARSTVO DELO IN ŠPORT

REPUBLIKA SLOVENIJA
MINISTRSTVO ZA OBRAMBO

ARIS

Vsebina

09:00 – 10:00 Veliki jezikovni modeli (VJM): kako delujejo in kako se učijo

- Zgradba in temeljna načela delovanja VJM
- Arhitektura transformer
- Učenje VJM

10:15 – 11:15 Prilagajanje VJM in agenti za poslovna okolja

- Prilagajanje VJM konkretnim nalogam
- Oblikovanje ukaznih pozivov
- Agentni sistemi

11:30 – 12:30 Evalvacija VJM in slovenski veliki jezikovni modeli

- Pristopi k evalvaciji VJM
- VJM za slovenski jezik
- Razlaga, etika, viri

12:30 – 13:00 Diskusija, vprašanja in zaključek

Laboratorij za strojno učenje in jezikovne tehnologije UL FRI



Laboratorij za strojno učenje in jezikovne tehnologije

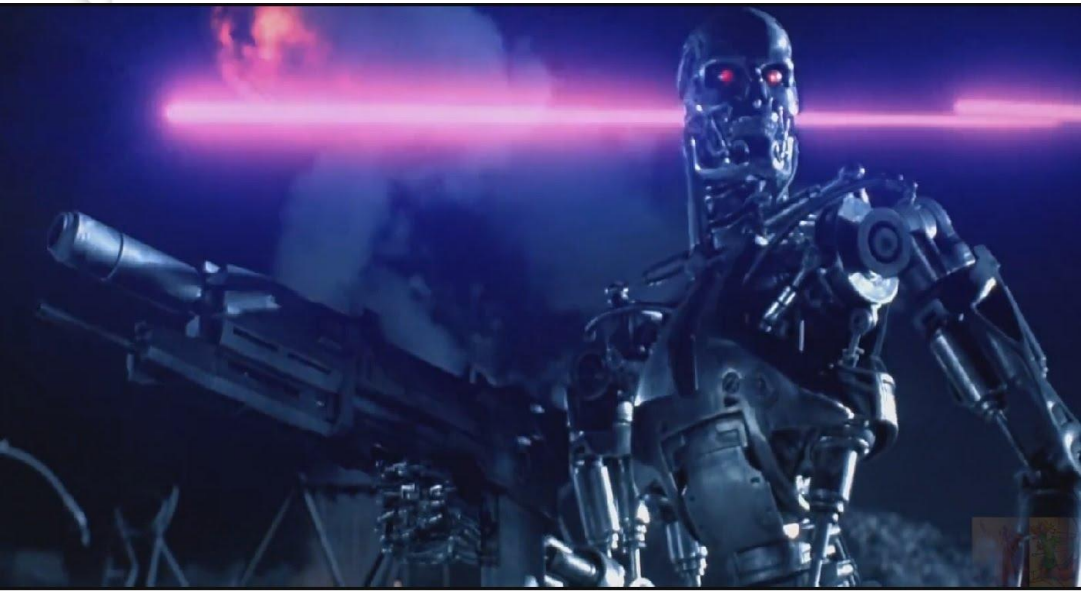
Raziskave na področju umetne inteligence in jezikovnih tehnologij

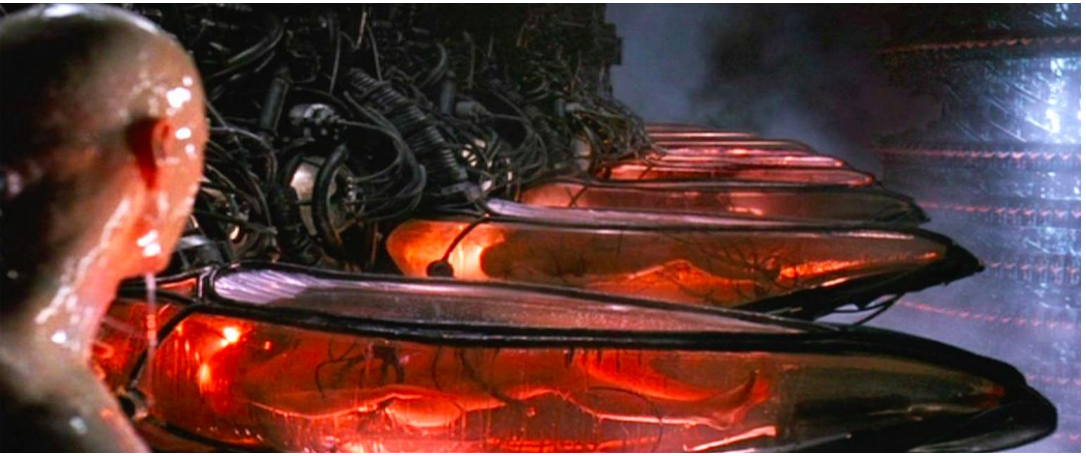
- raziskave VJM od 2018
- izgradnja številnih VJM za več jezikov z manj viri
- temeljne raziskave na področjih predstavitve podatkov, vstavljanja znanja v VJM, medjezikovnega prenosa, razložljive UI
- aplikativne raziskave v lingvistiki, berljivosti, povzemanju in poenostavljanju besedil, novičarskih medijih, parlamentarnem govoru, sentimentu, politični opredelitvi, depresiji, pristranostih in stereotipih, sovražnem govoru, lažnih novicah, slovničnih popravljalnikih, sarkazmu, parafraziranju, folkloristiki, politologiji, historiografiji, itn.

University of Ljubljana
Faculty of Computer and Information Science

Vprašanja o UI iz medijev







Nas bodo roboti uničili?

Nam bodo vzeli službe?

Bomo še potrebovali vozniško?

Bomo še potrebovali zdravnike?

Kako se bodo razvijali humanoidi roboti?

Kaj bo s kiborgi?

Kaj je splošna umetna inteligenca?

Kaj je tehnološka singularnost?

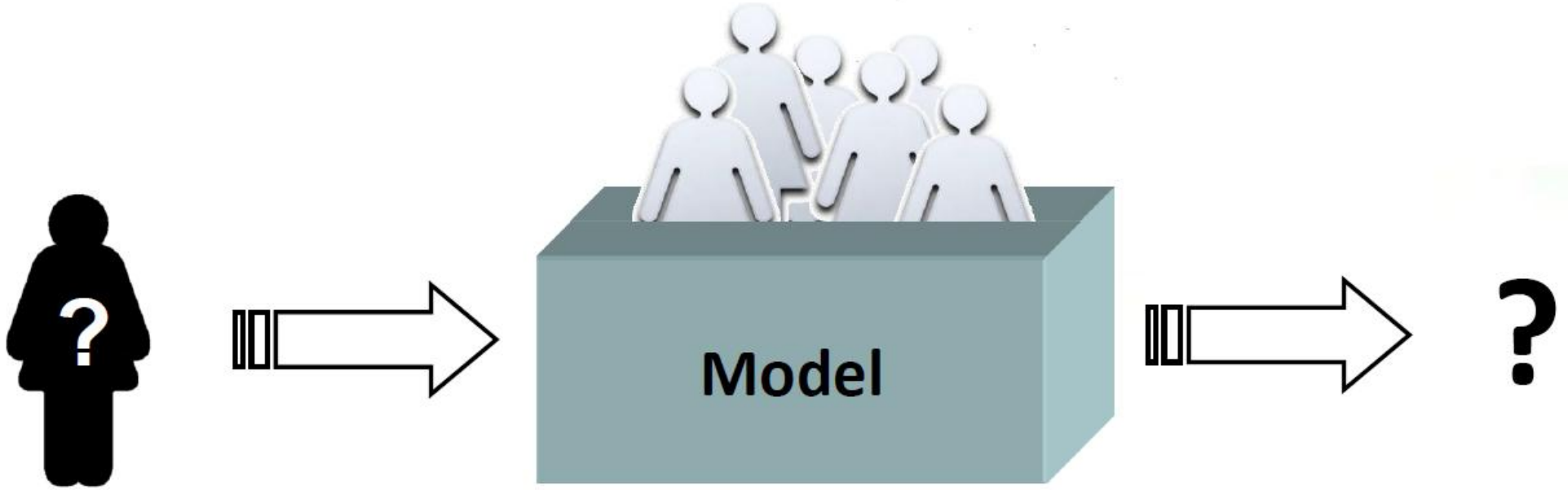




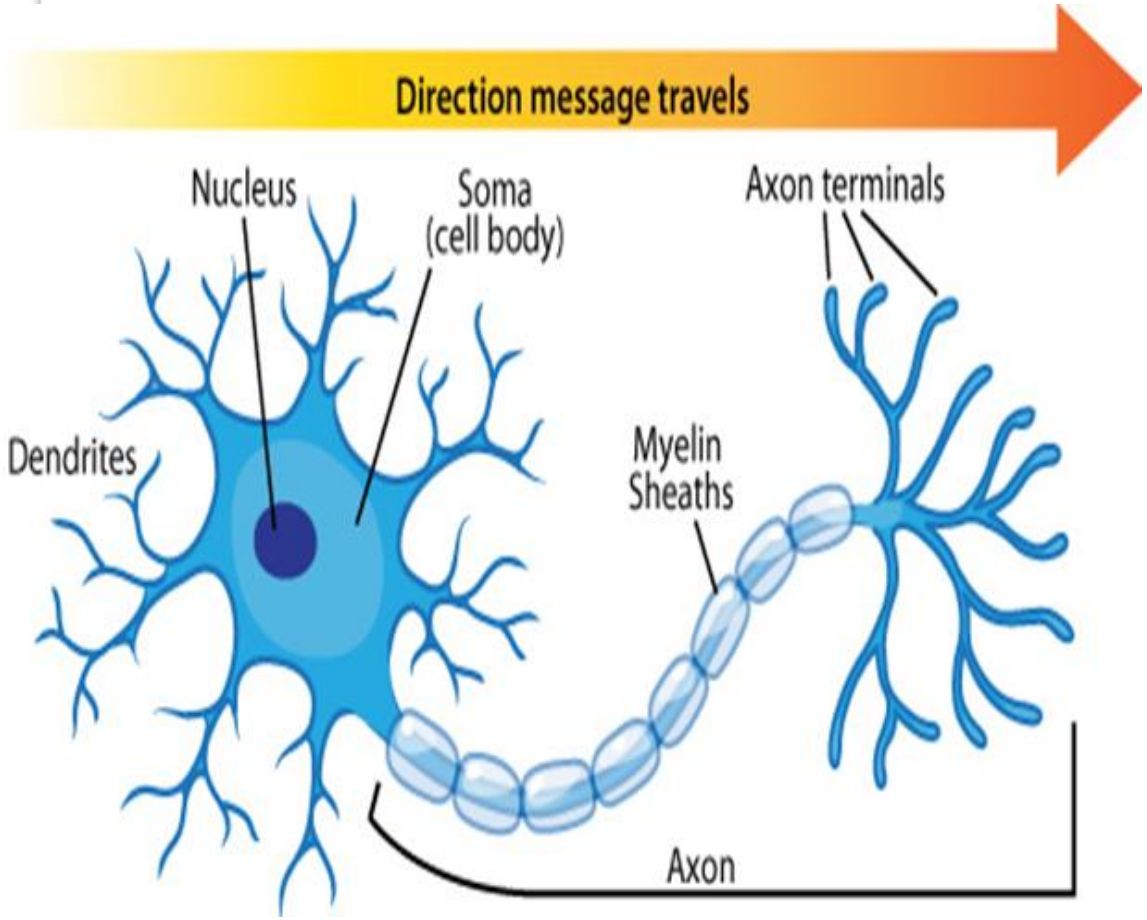
Osnovna ideja strojnega učenja

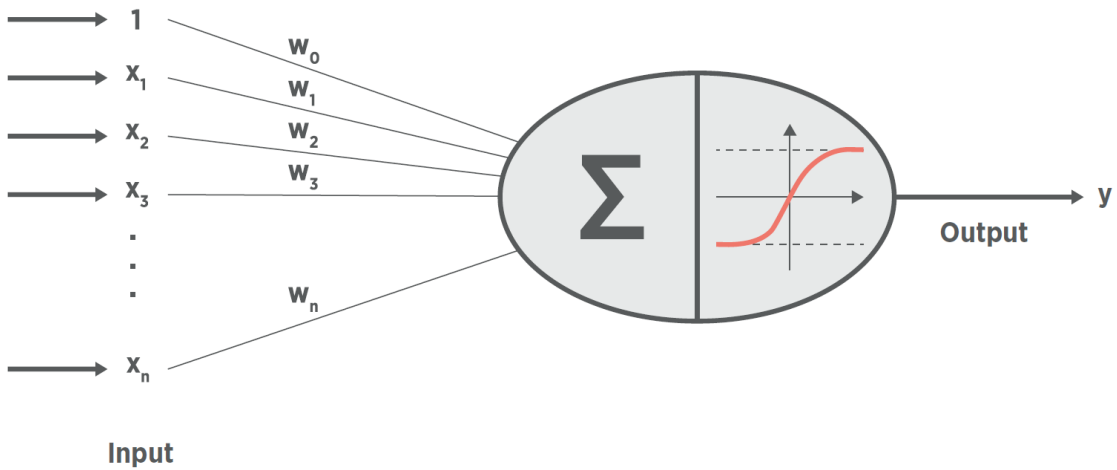
- učenje iz podatkov

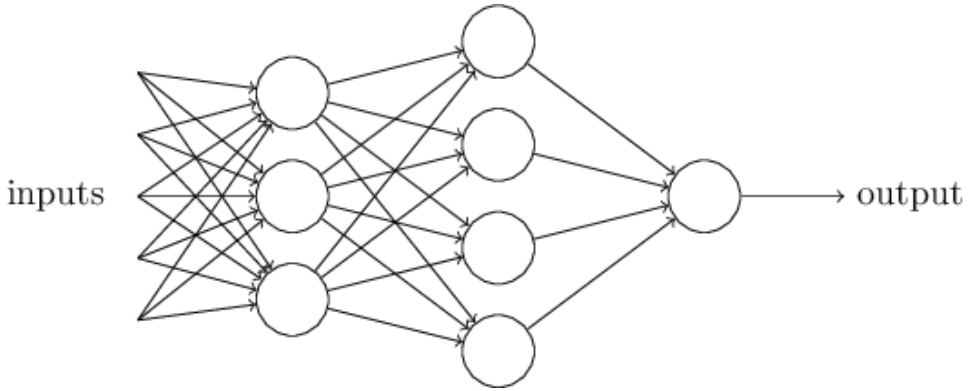
	class1	class2	menop	stage	grade	hType	PgR	inv	nLymph	cTh	hTh	famHist	LVI	ER	maxNode	posRatio	age
300	11.82	0	1	2	2	1	0	0	1	1	0	3	0	1	2	3	2
301	4.89	1	0	1	2	1	0	0	2	1	0	0	0	2	1	4	3
302	14.63	0	1	1	4	2	0	0	0	0	0	1	0	1	1	1	3
303	21.83	0	0	1	4	2	1	0	1	0	0	9	0	4	1	2	2
304	19.87	0	0	1	2	1	0	0	0	0	0	0	0	1	2	1	2
305	7.54	0	1	2	3	1	9	2	1	0	1	1	0	3	3	3	4
306	15.15	0	0	1	4	2	1	0	0	0	0	2	0	4	1	1	2
307	0.30	1	0	2	2	1	0	0	3	0	0	9	0	1	1	4	2
308	12.49	0	1	2	2	3	1	0	0	0	0	0	0	4	1	1	5
309	1.77	1	0	2	3	1	1	2	2	1	0	9	1	3	3	3	2



Umetne nevronske mreže: analogija



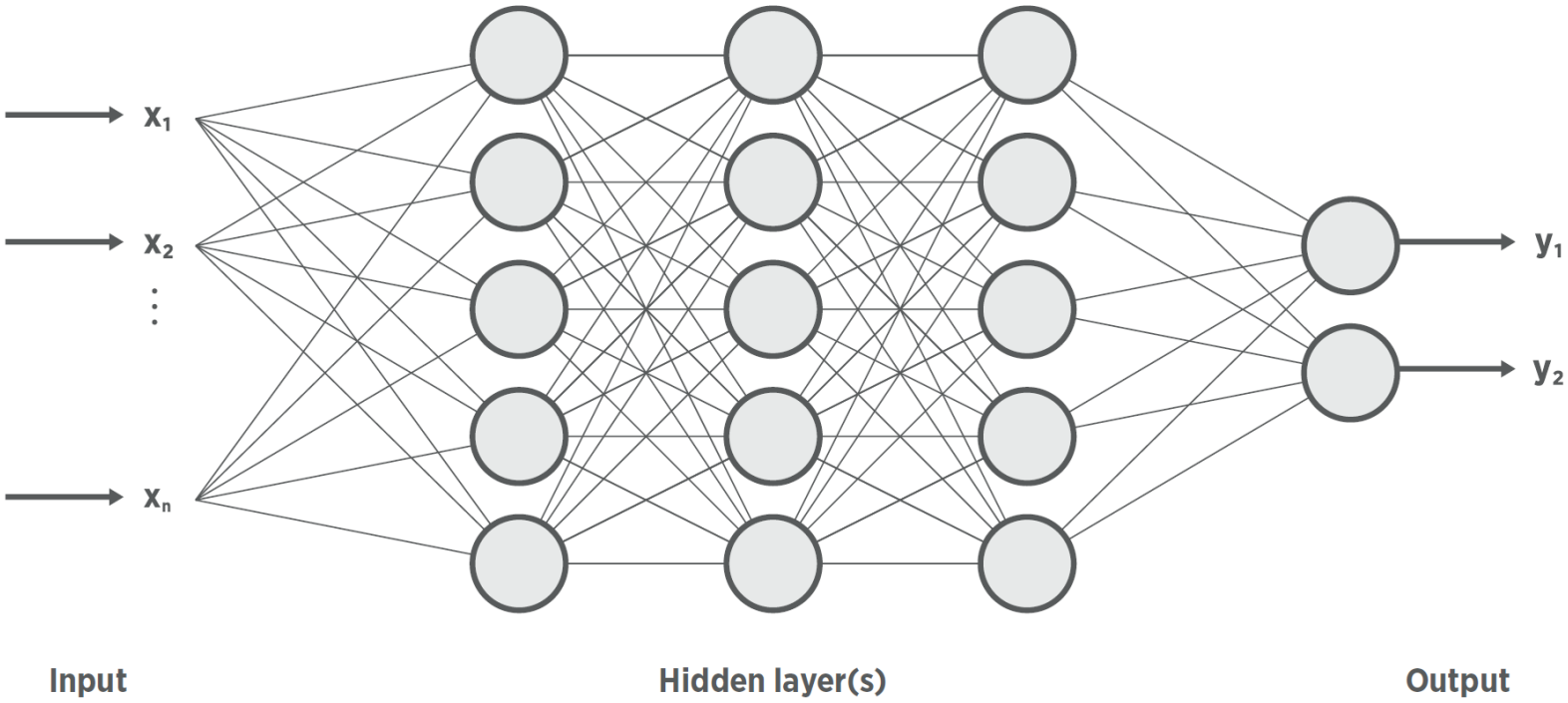




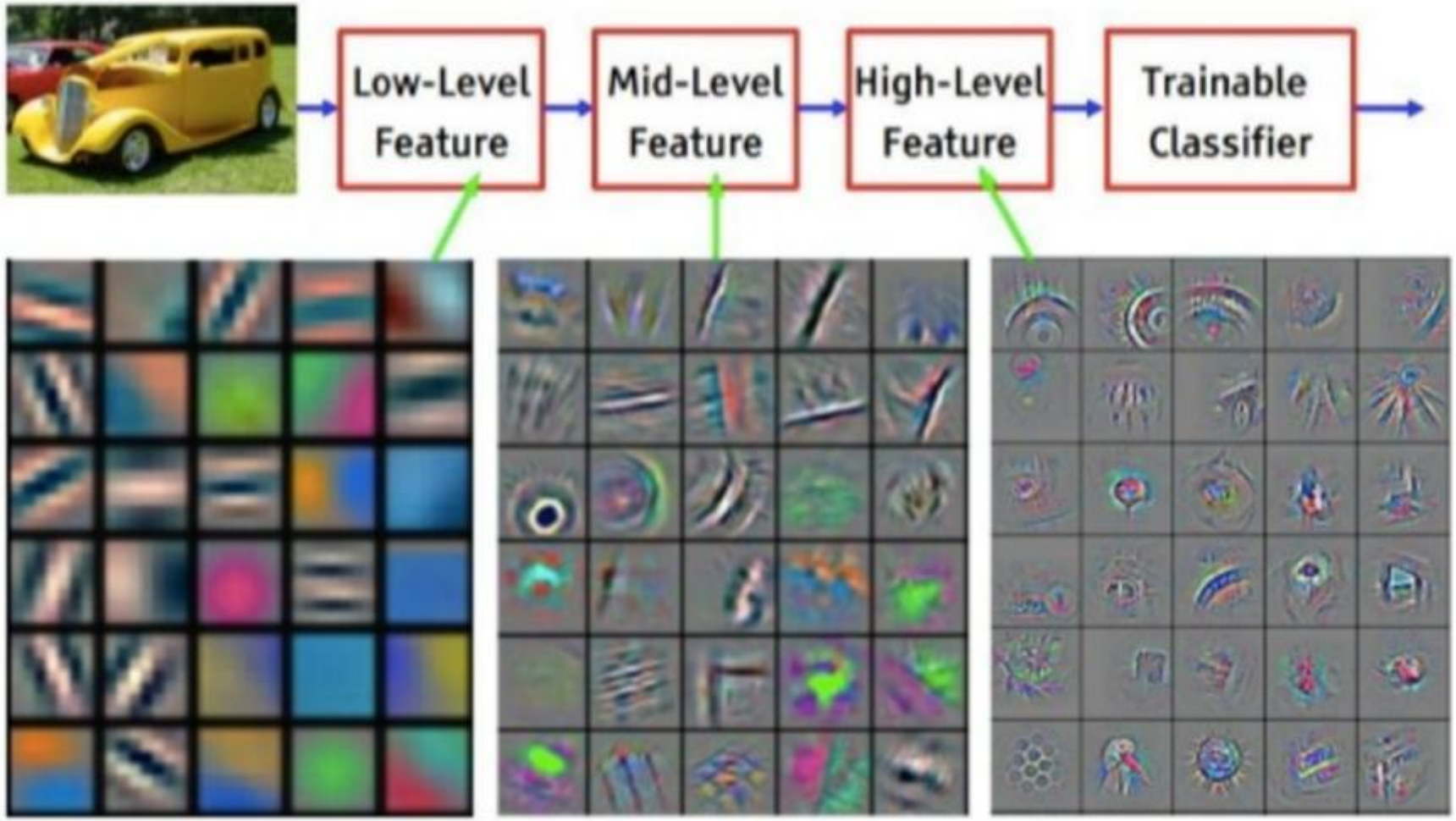
učenje: vzvratno razširjanje napake
Nobelova nagrada 2025, Hinton

Globoko učenje =

globoke nevronske mreže + velike učne množice + GPU
+mnogo novih idej + ogromna vlaganja

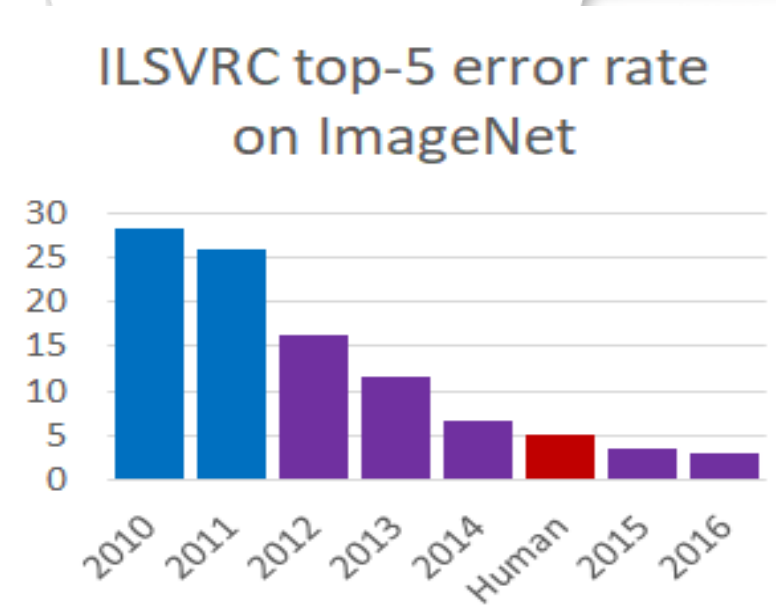


Zakaj delujejo?



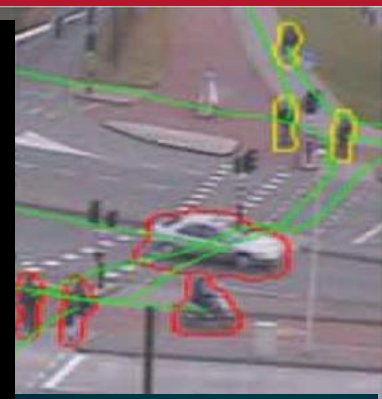
Univerza v Ljubljani
Fakulteta za računalništvo in informatiko

Nekaj uspehov globokega učenja













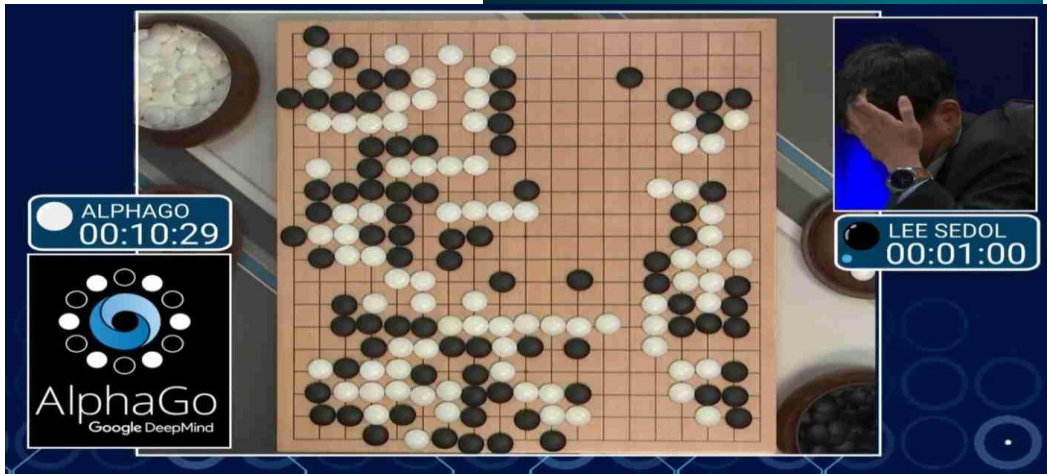




César Prado in Uncategorized · August 22, 2017 · 94 words

Microsoft claims new speech recognition record, achieving a super-human 5.1% error rate





Distribucijska semantika: razumevanje iz jezikovnih podatkov



"You shall know a word
by the company it keeps"

Firth, J. R. (1957). A synopsis of linguistic theory 1930–1955. In Studies in Linguistic Analysis, p. 11. Blackwell, Oxford.

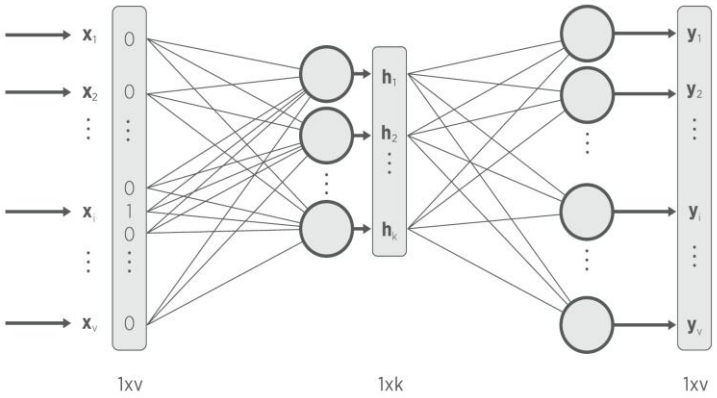


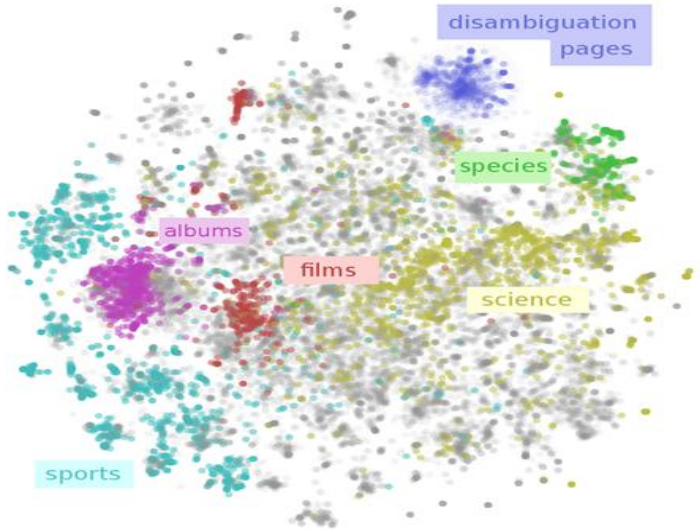
"The meaning of a word is its use in the language"

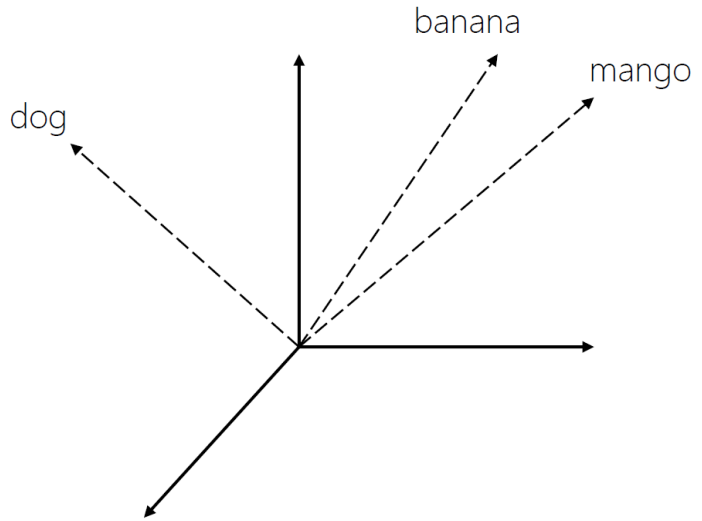
Ludwig Wittgenstein, PI #43

Goste nevronske vložitve

- nevronske besede zahtevajo številske vhode
- nevronsko mrežo uporabimo za napovedovanje konteksta besed
- vhod: beseda, izhod: kontekst sosednjih besed, metoda word2vec 2013
- semantične relacije v jeziku ustrezajo razdaljam in smerem v vložitvenem vektorskem prostoru







Relacijska podobnost



vektor(*‘kralj’*) - vektor(*‘moški’*) ≈ vektor(‘kraljica’) - vektor(*‘ženska’*)

vektor(*‘Paris’*) - vektor(*‘Francija’*) ≈ vektor(‘Rim’) - vektor(*‘Italija’*)



Ulčar, M., Vaik, K., Lindström, J., Dailidėnaitė, M. and Robnik-Šikonja, M., 2020, May. Multilingual Culture-Independent Word Analogy Datasets.
In *Proceedings of the Twelfth Language Resources and Evaluation Conference* (pp. 4074-4080).

Vložitve izražajo pristranosti

UL | FRI

Ask “Paris : France :: Tokyo : x”

→x = Japan

Ask “father : doctor :: mother : x”

→x = nurse

Ask “man : computer programmer :: woman : x”

→x = homemaker



Bolukbasi, T., Kai-Wei C., Zou, J. W., Saligrama, V., and Kalai, A. W. (2016) "Man is to computer programmer as woman is to homemaker? debiasing word embeddings." In *Advances in Neural Information Processing Systems*, pp. 4349-4357.

Ulčar, M., Supej, A., Robnik-Šikonja, M., & Pollak, S. (2021). Slovene and Croatian word embeddings in terms of gender occupational analogies. Slovenščina 2.0: empirical, applied and interdisciplinary research, 9(1), 26-59.

Verjetnostno modeliranje jezika – jezikovni modeli

- Izračunati verjetnost stavka oz. zaporedja besed:

  $P(W) = P(w_1, w_2, \ldots w_n)$

- Npr., strojno prevajanje:
  - P(**high** winds tonight) > P(**large** winds tonight)

- Izračun naslednje besede v zaporedju

  $P(w_i | w_1, w_2, \ldots w_{i-1})$
  Včeraj smo se potepali po ____

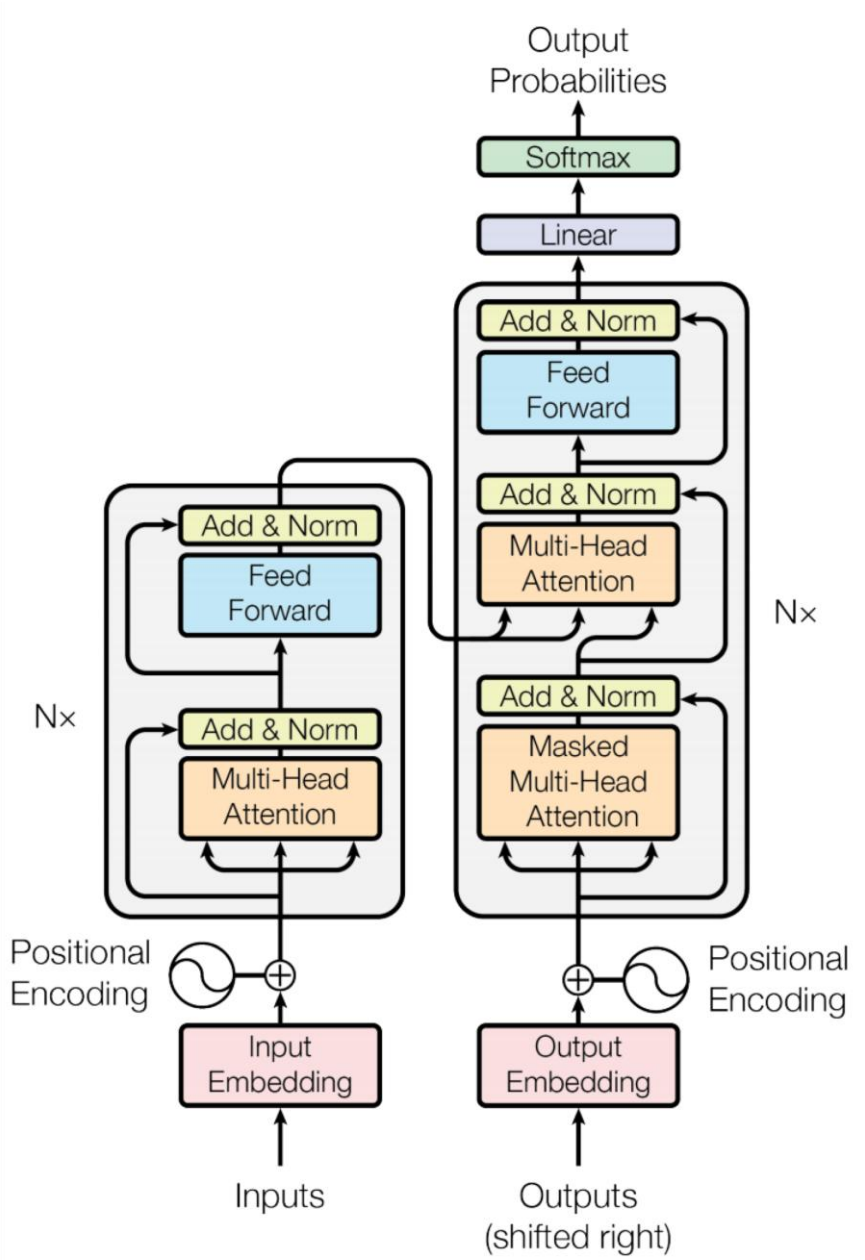
- Izračun maskirane besede v zaporedju
  $P(w_i | w_1, w_2, .. w_{i-1}, w_{i+1}, \ldots w_n)$
  Kljub _____ se nista mogla odpovedati večernemu spehodu.

- Model za izračun teh verjetnosti imenujemo jezikovni model (language model, LM)



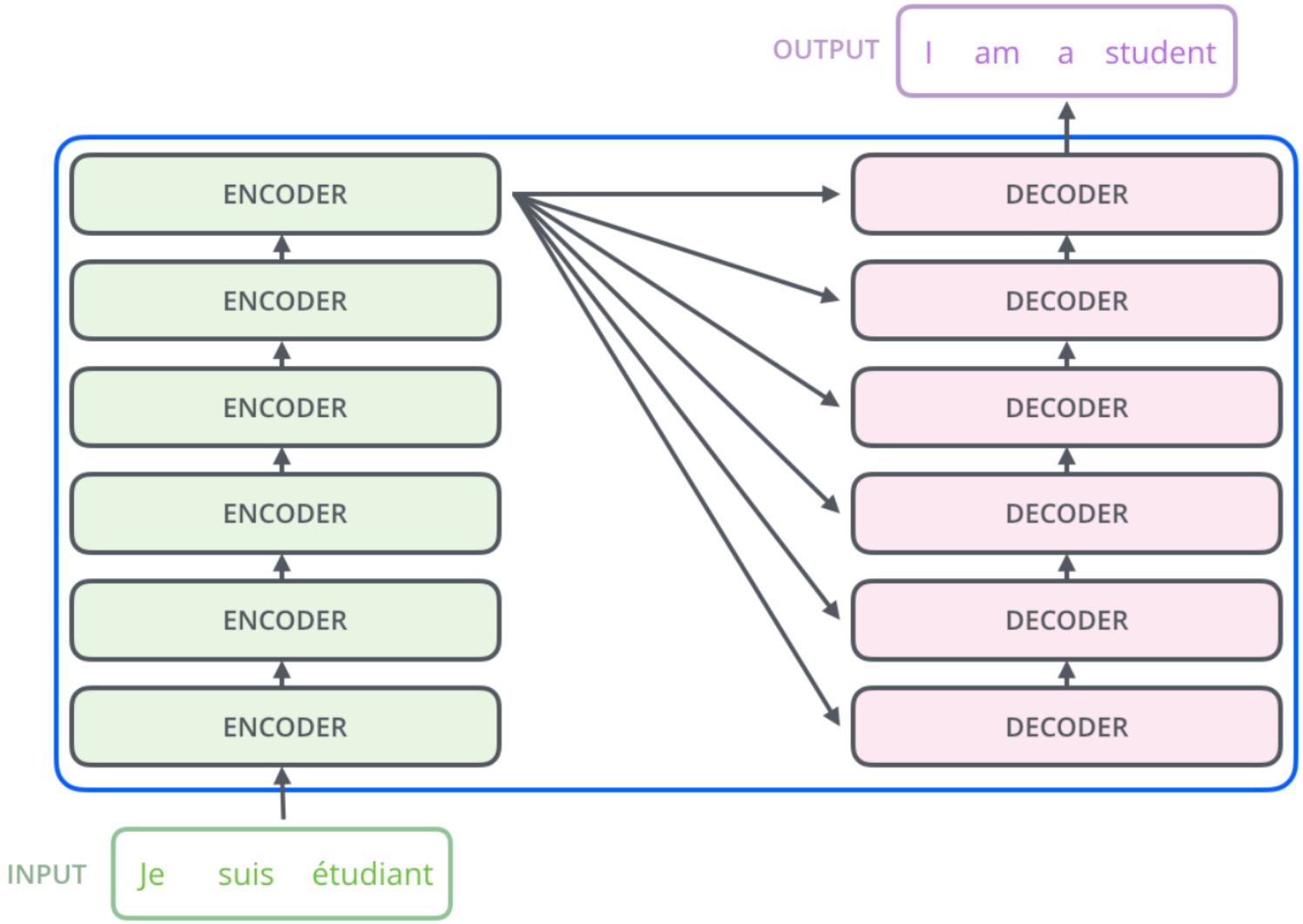
Arhitektura transformer

- trenutno najuspešnejša arhitektura globokih nevronskih mrež
- nerekurzivna, kodirnik-dekodirnik arhitektura
- fiksna vhodna sekvenca (tipično precej dolga, > 512 žetonov)
- prilagojena GPU procesiranju
- prilagojena ekstremnemu paralizmu: skalarni produkti, matrične operacije
- temelji na uporabi mehanizma pozornosti
- dobra skalabilnost
- kodirnik in dekodirnik se lahko uporabljata ločeno
- vsebuje preskočne povezave za preprečevanje pozabljanja



Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I., 2017. Attention is all you need. In *Advances in neural information processing systems* (pp. 5998-6008).

Osnovni transformer je model tipa kodirnik-dekodirnik

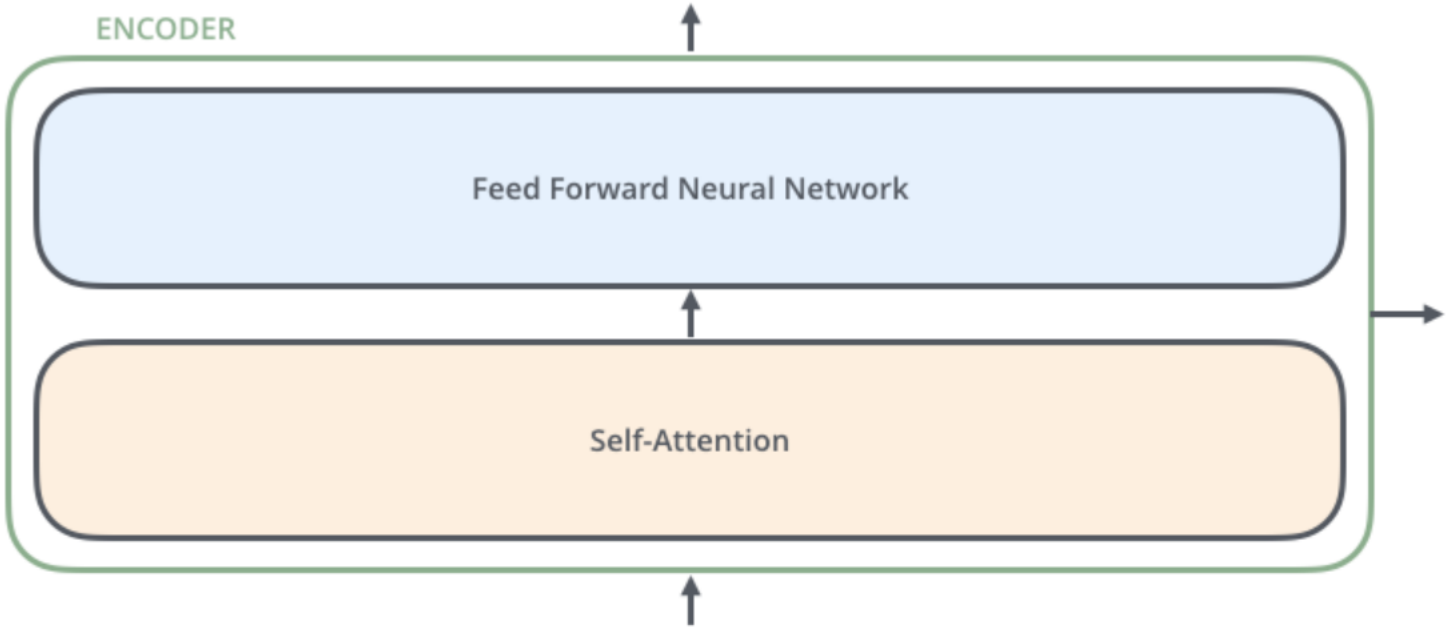


Na sliki je za ilustracijo 6 kodirnikov in 6 dekodirnikov

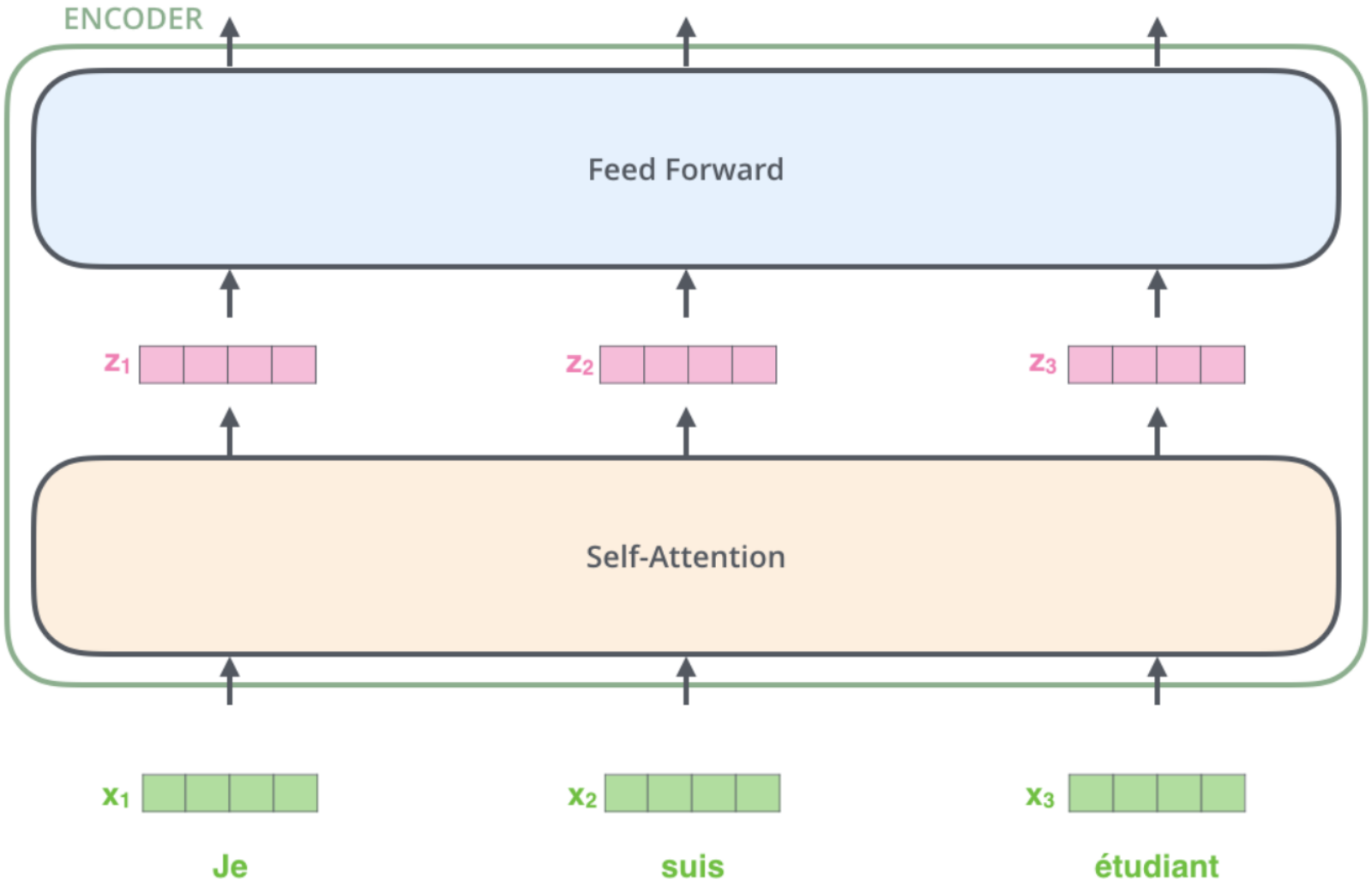
Oba dela sta lahko uporabljena tudi samostojno

Transformer: kodirnik

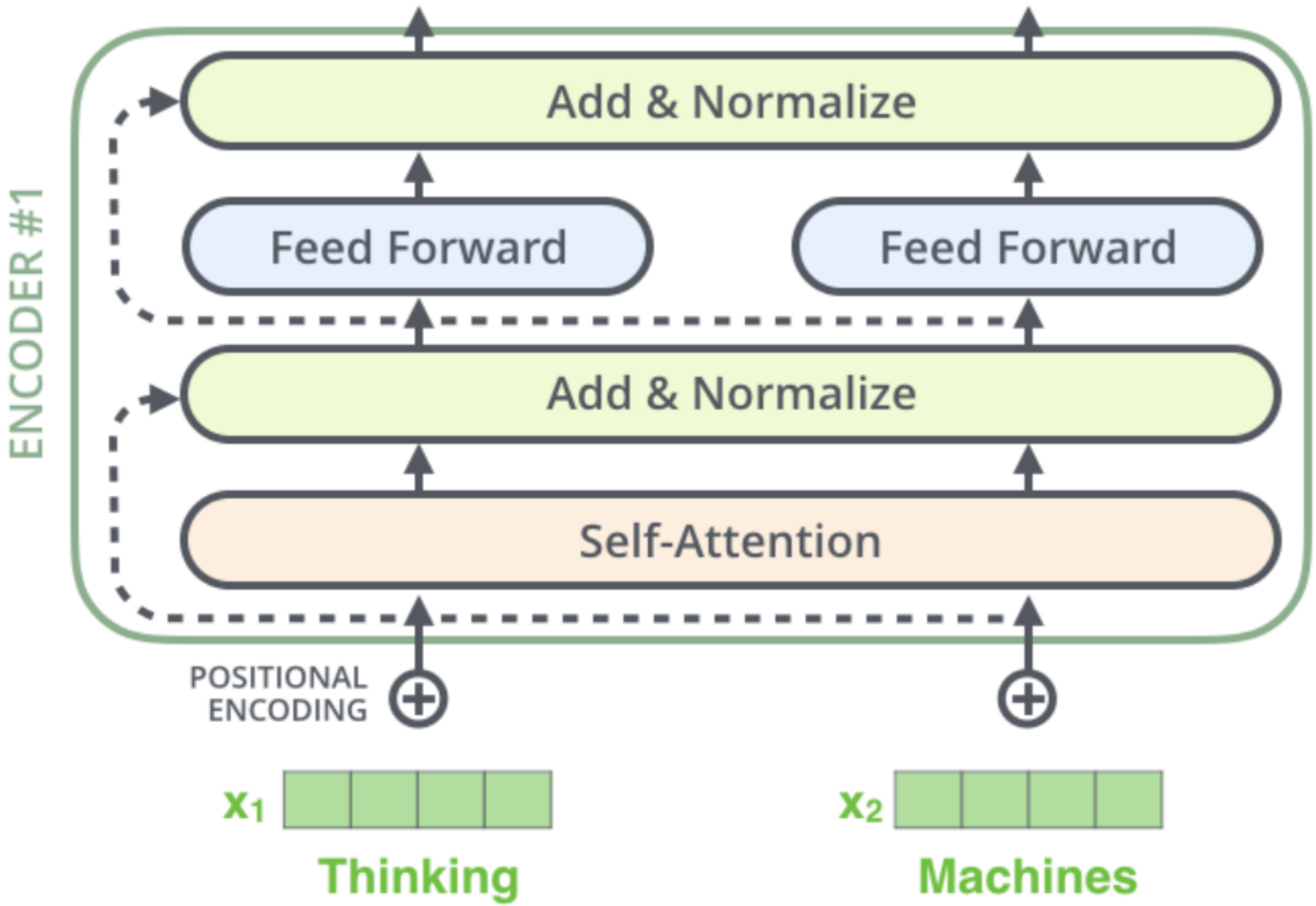
- Vsak nivo je sestavljen iz dveh delov
- Brez deljenih uteži med različnimi kodirniki
- Informacije med kodriniki se prenašajo z mehanizmom samopozornosti, ki črpa informacijo iz pomembnih delov vhoda



Normalizacija skrbi za učinkovitost računanja

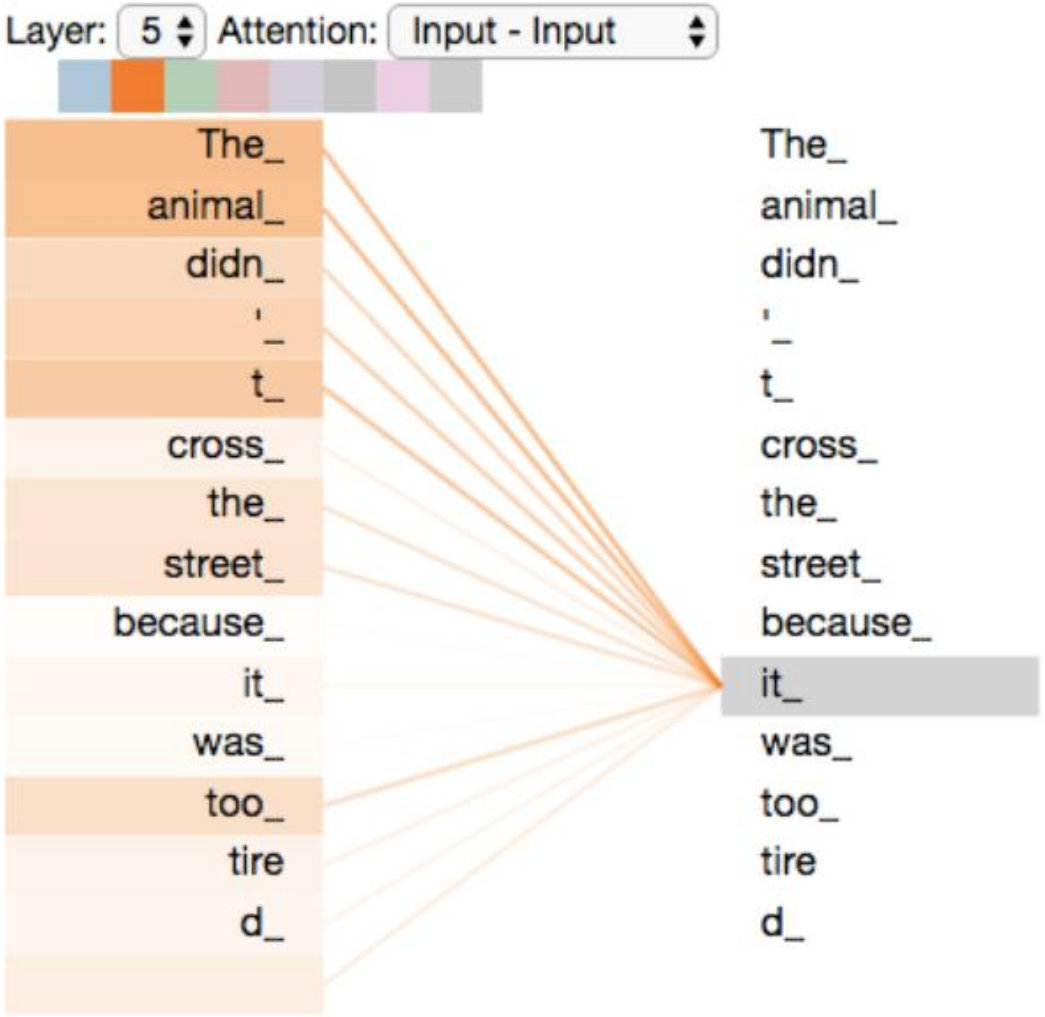


Kodirnik – celota

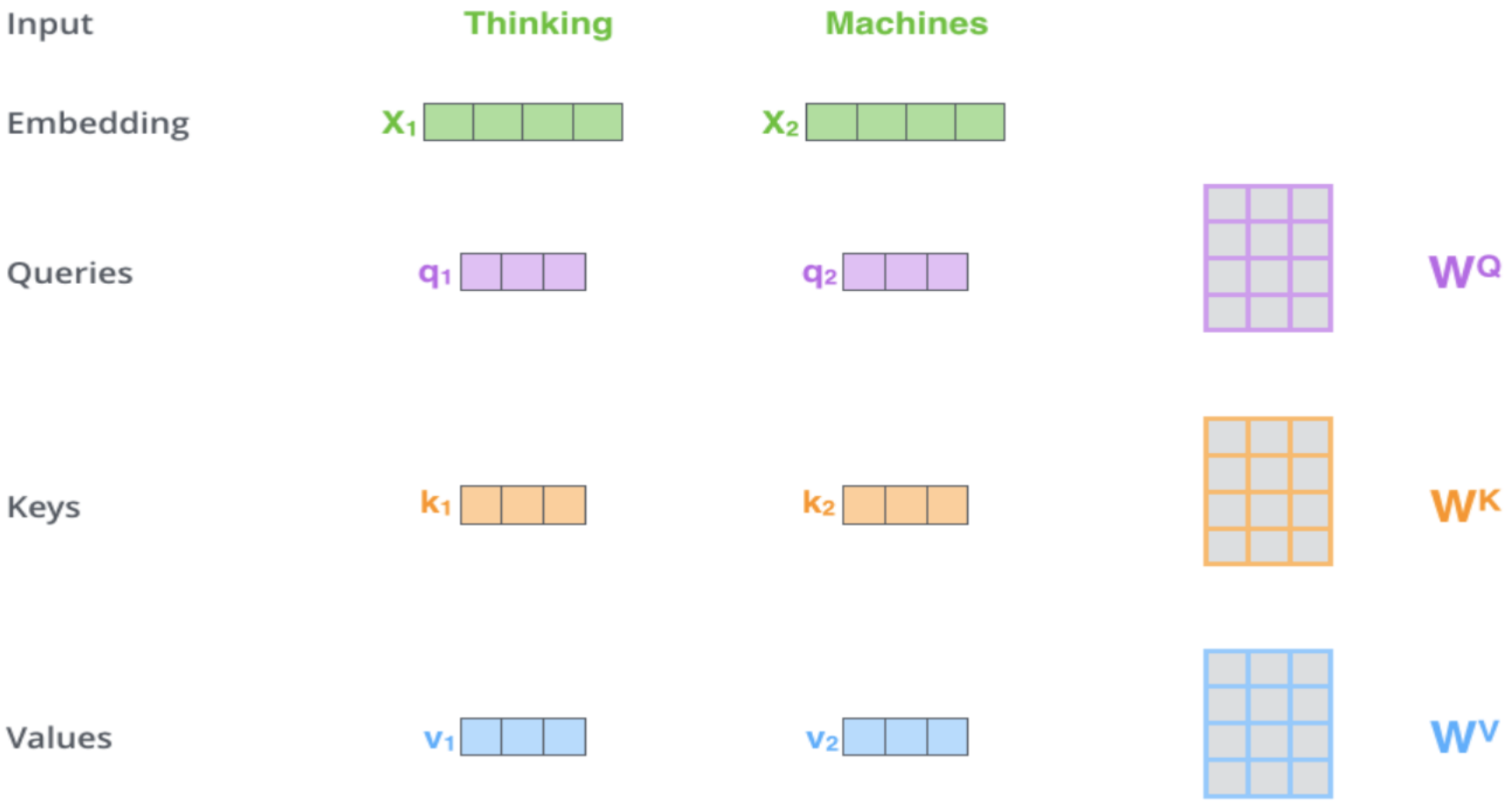


Ključna inovacija transformerjev: samopozornost

- Predstavitev besed v stavku
- *”The animal didn't cross the street because it was too tired”*
- *Se ”*it*”* nanaša na “street” ali “animal”?
- Kaj pa:
- *”The animal didn't cross the street because it was too wide”*

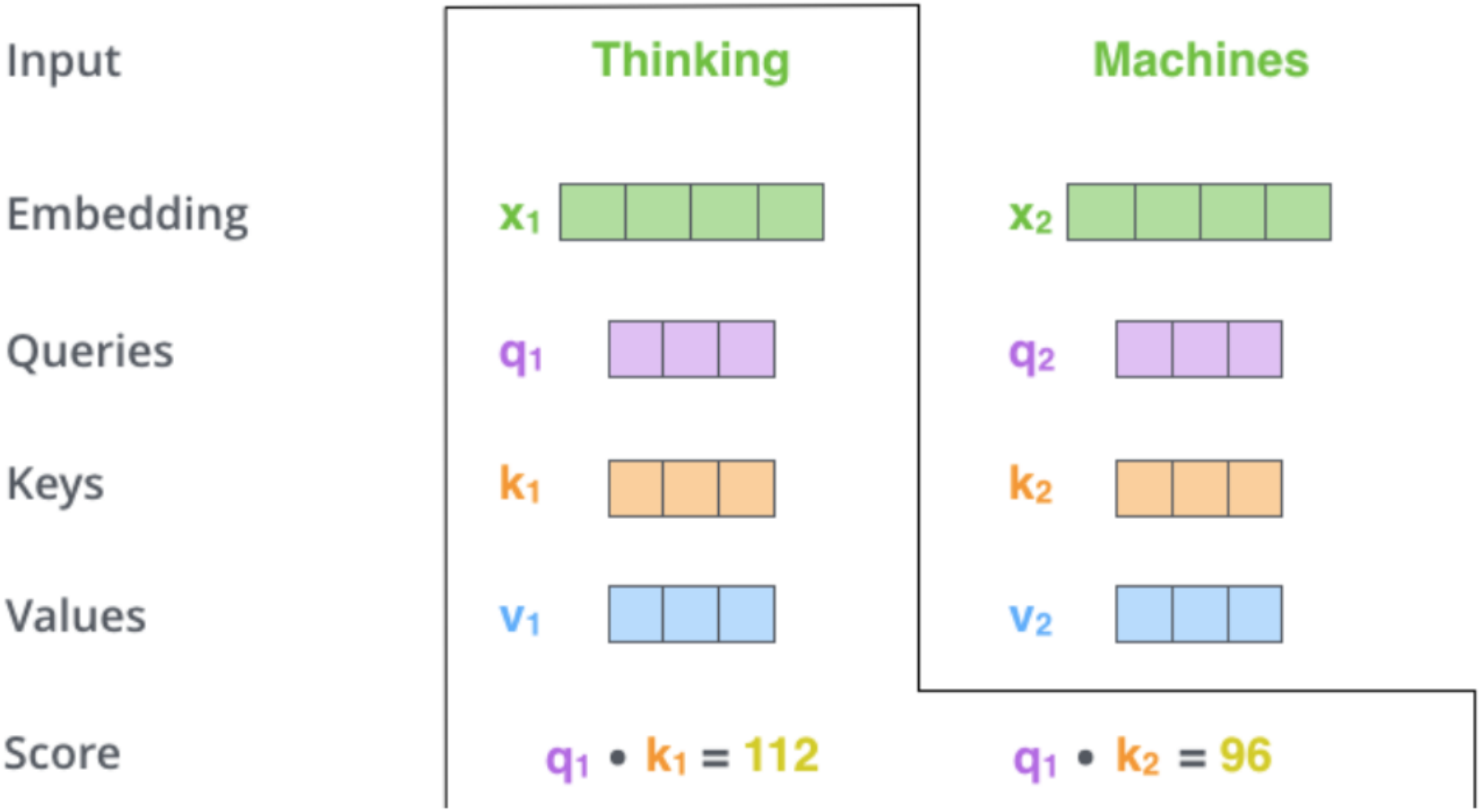


Samopozornost: natančneje 1/4

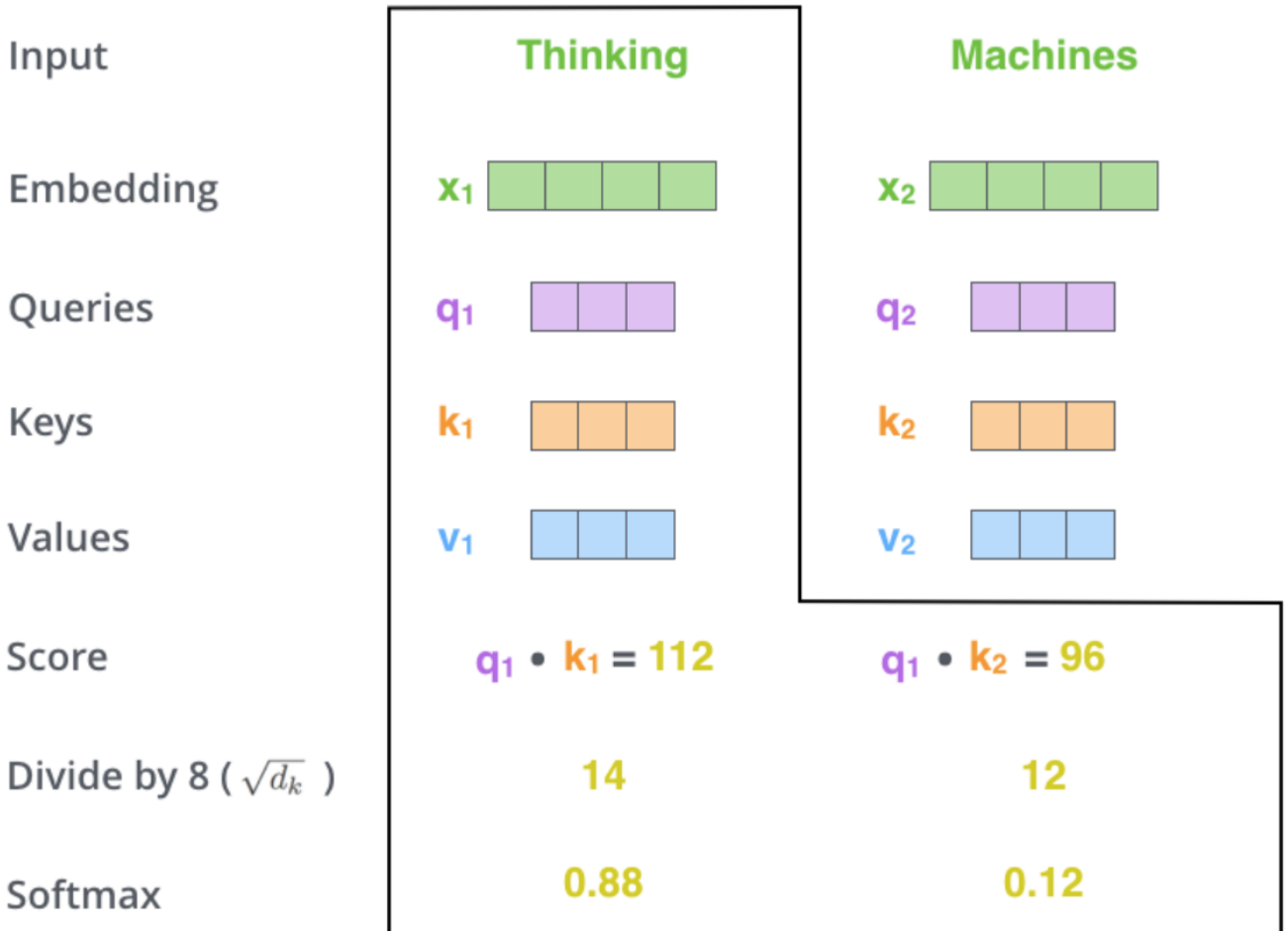


¸Iz računskih razlogov naredimo projekcije vhoda. Množenje x_1 s projekcijsko matriko W^Q ustvari vector q_1 ("query“), ki je povezan z vhodom x_1. Ustvarimo projekcije "query", "key“ in "value" za vsak žeton na vhodu.

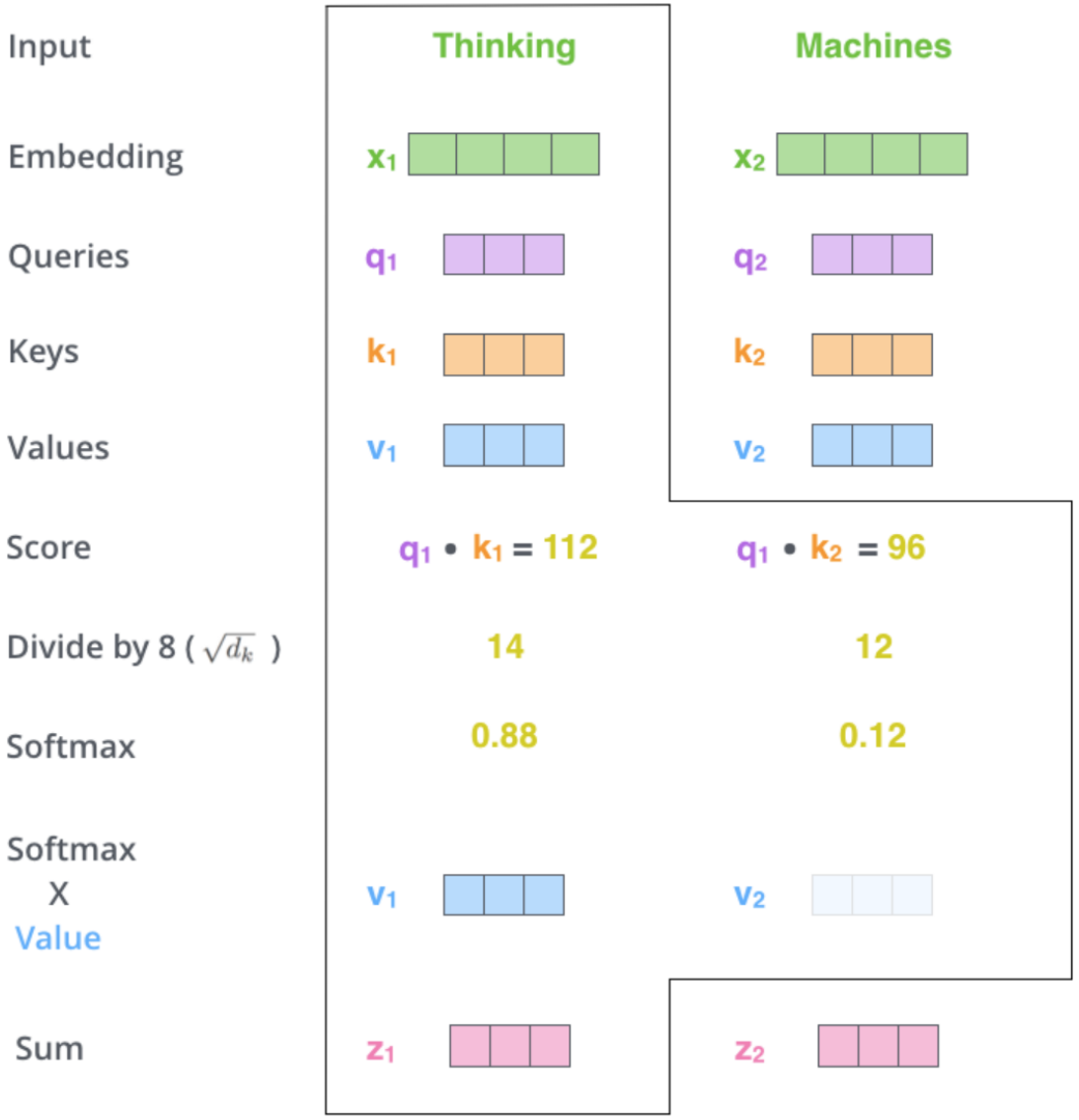
Natančneje 2/4: ocena



Natančneje 3/4: normalizacija ocen

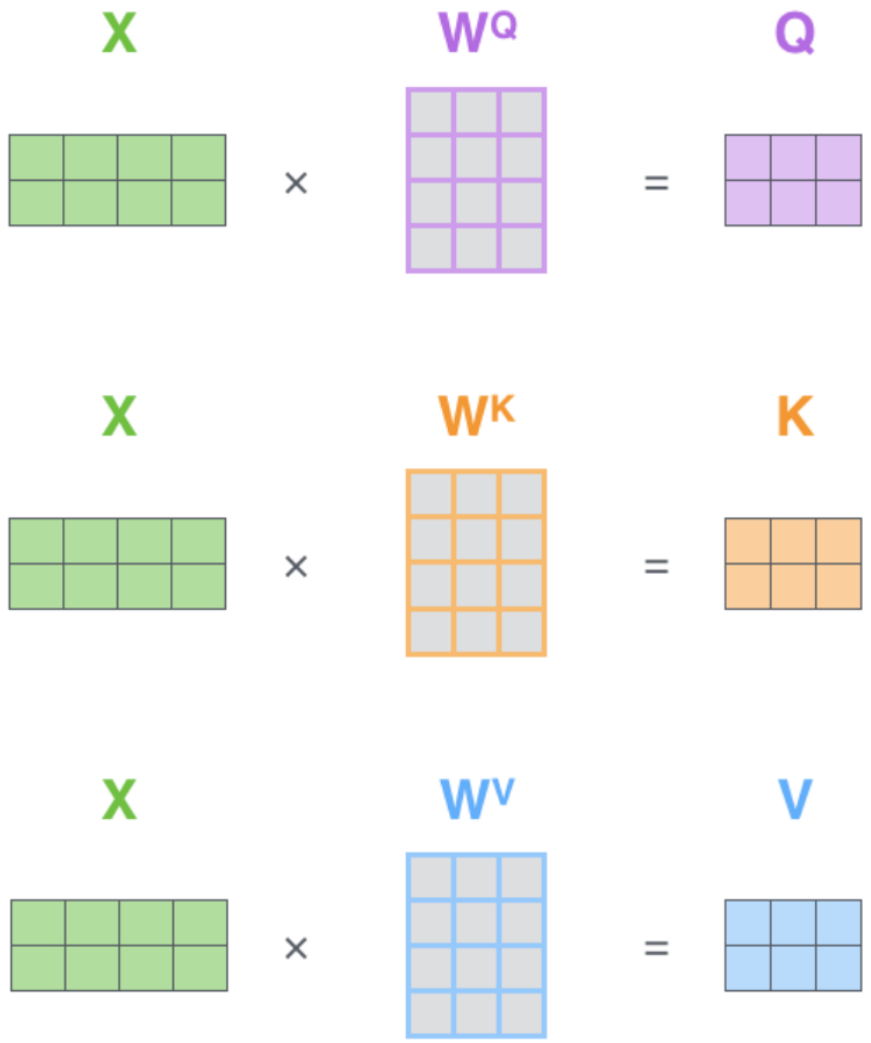


Natančneje 4/4: izhod samopozornosti



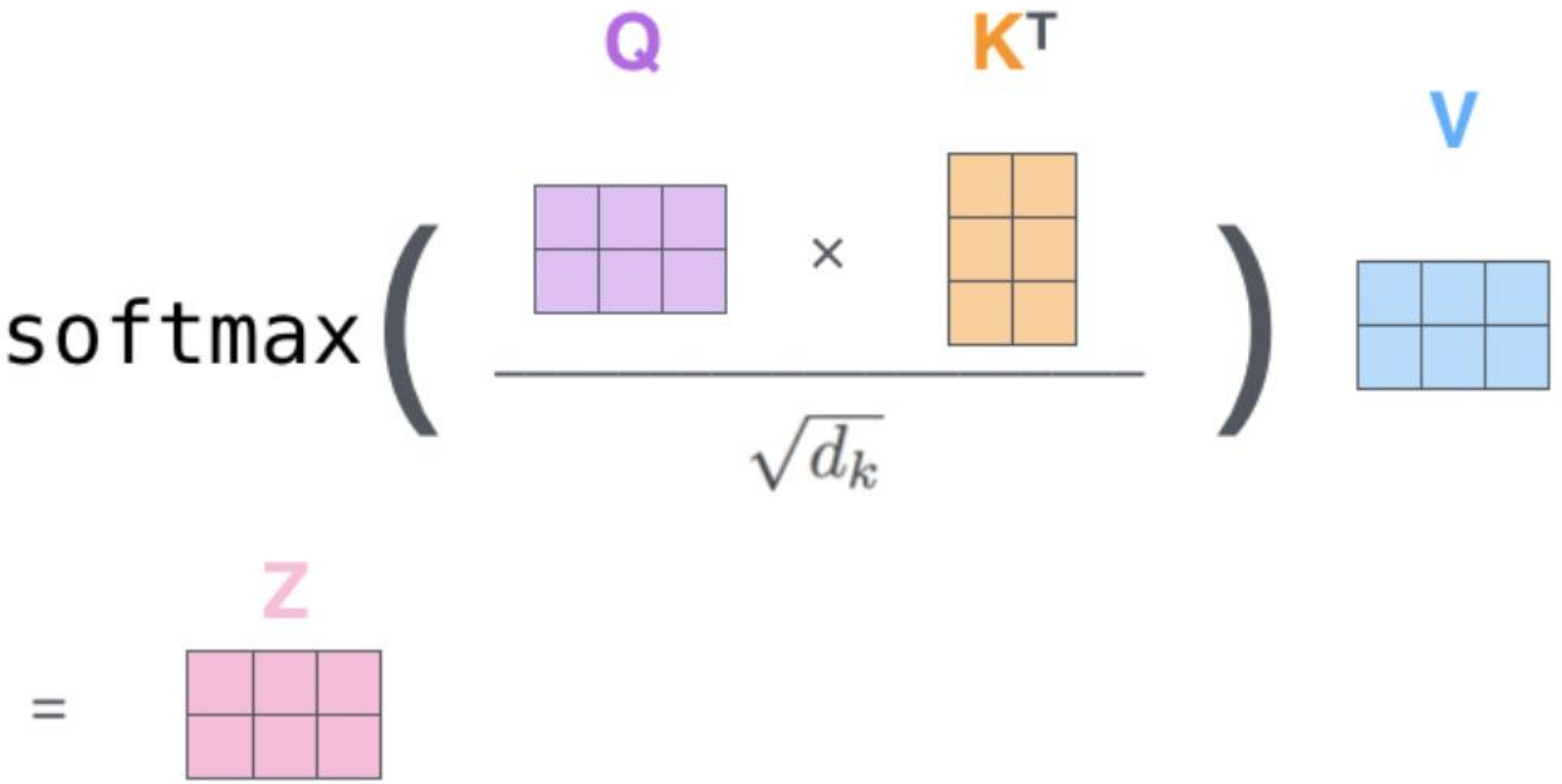
Matrični izračun samopozornosti 1/2

- Vsaka vrstica v matriki X predstavlja en žeton iz vhodnega besedila
- Vhodni vložitveni vektor x (512) je večji od projekcijskih vektorjev q/k/v (64)



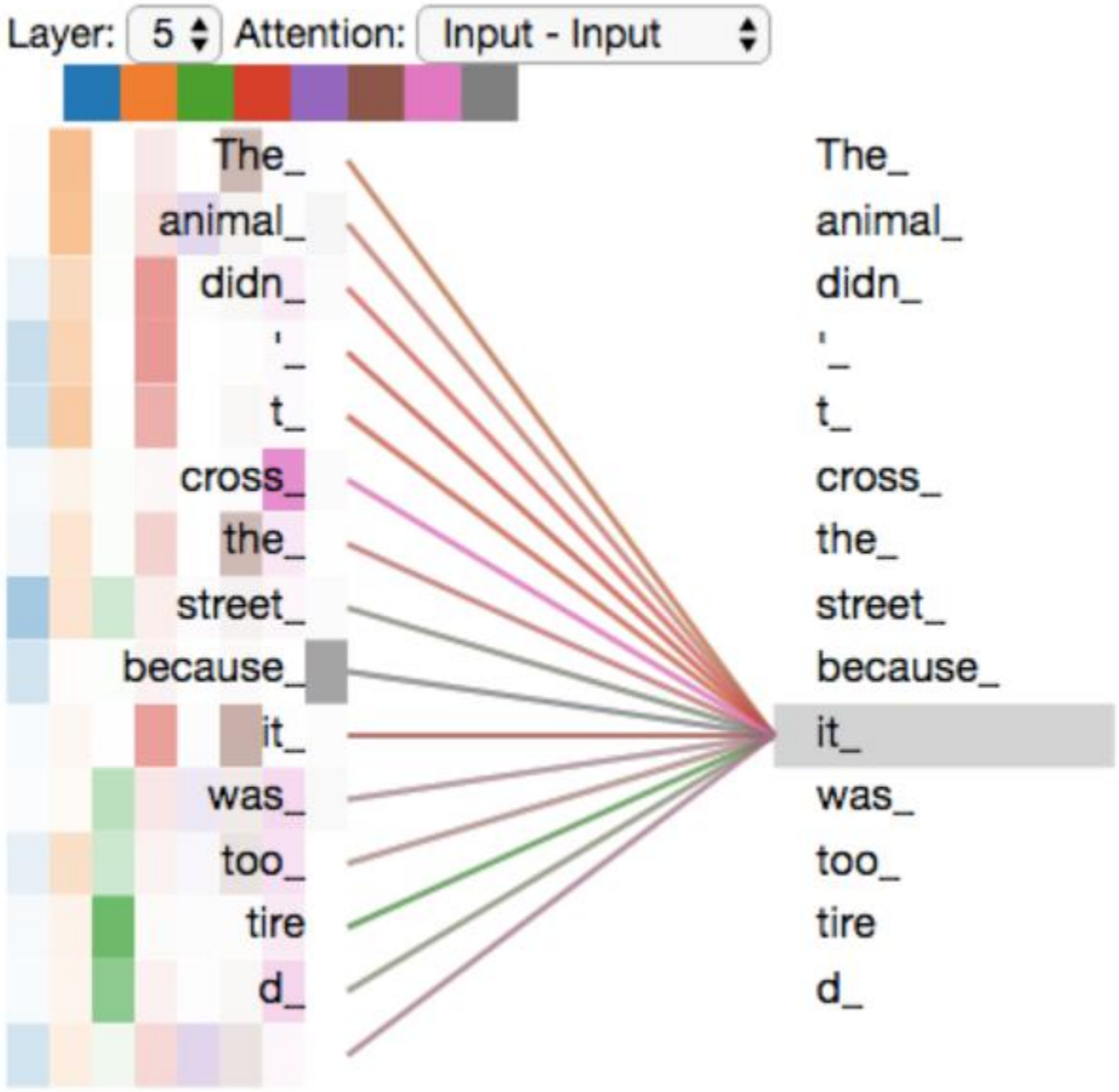
Matrični izračun samopozornosti 2/2

- Končni izračun

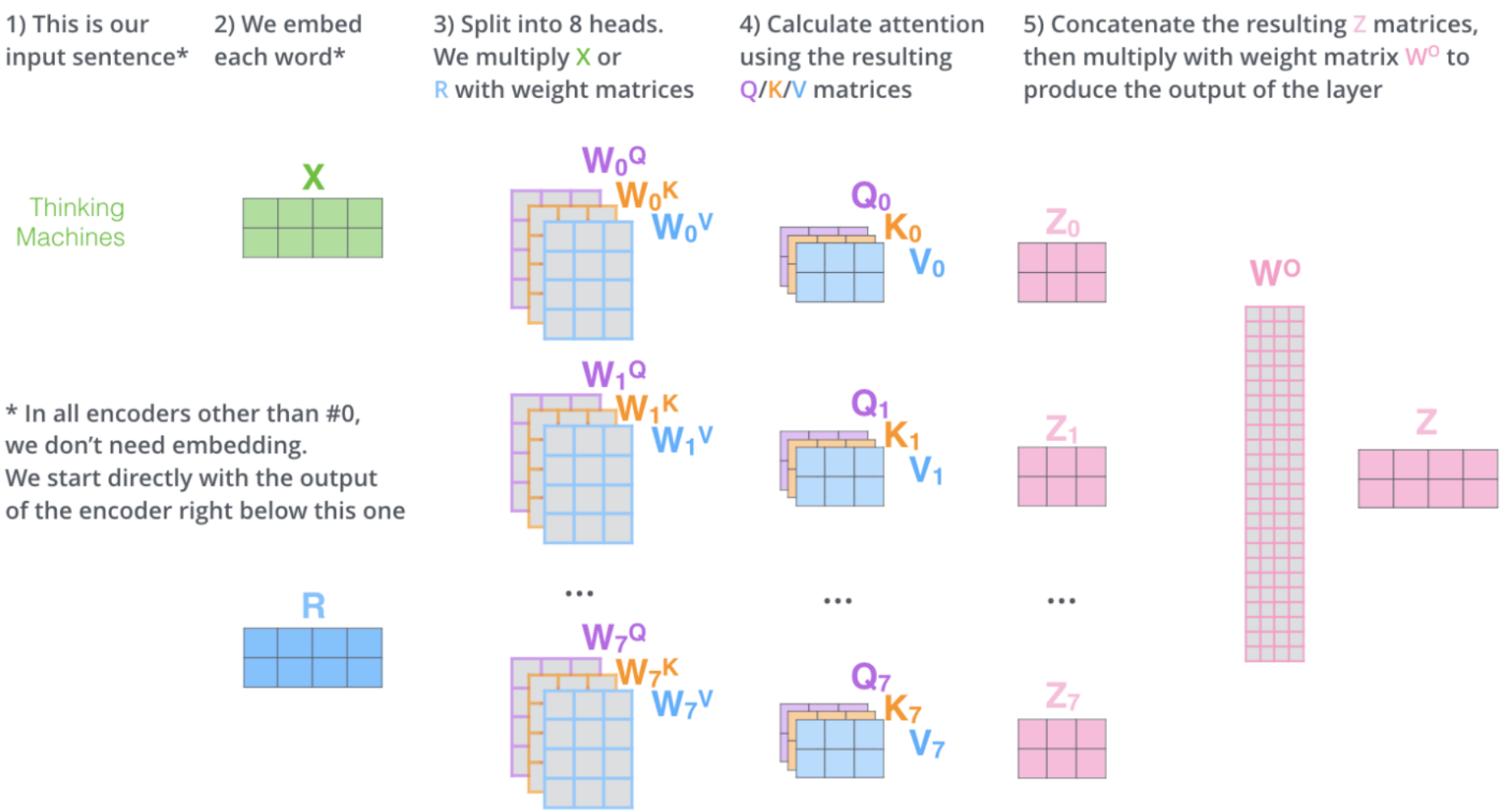


Ilustracija samopozornosti: več glav

- Vse glave je težko interpretirati

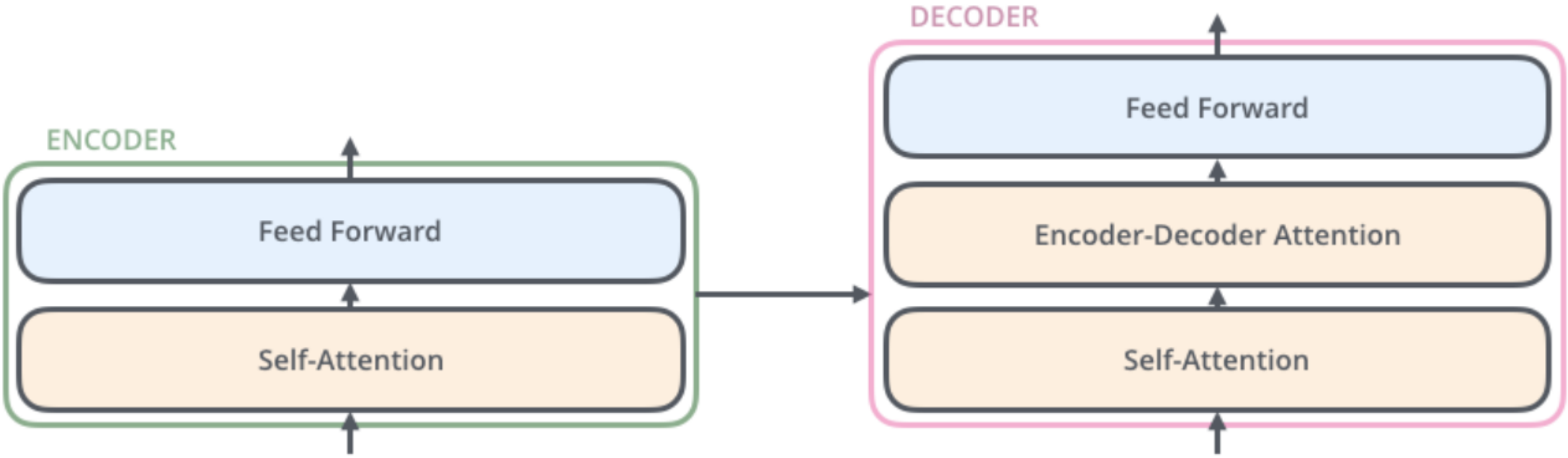


Povzetek samopozornosti

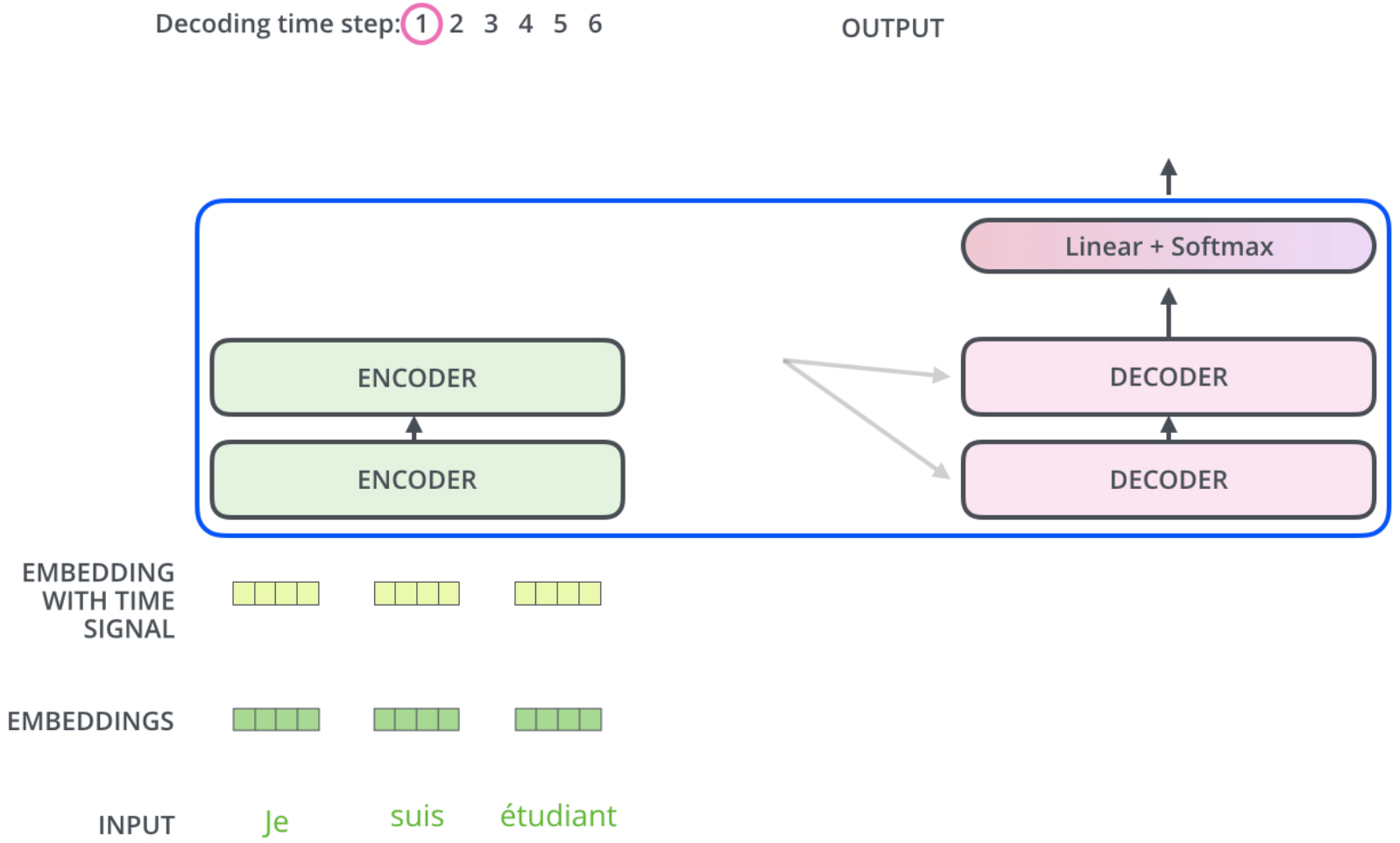


Transformer: dekodirnik

- Podoben kodirniku
- Dodaten sloj pozornosti iz kodirnika

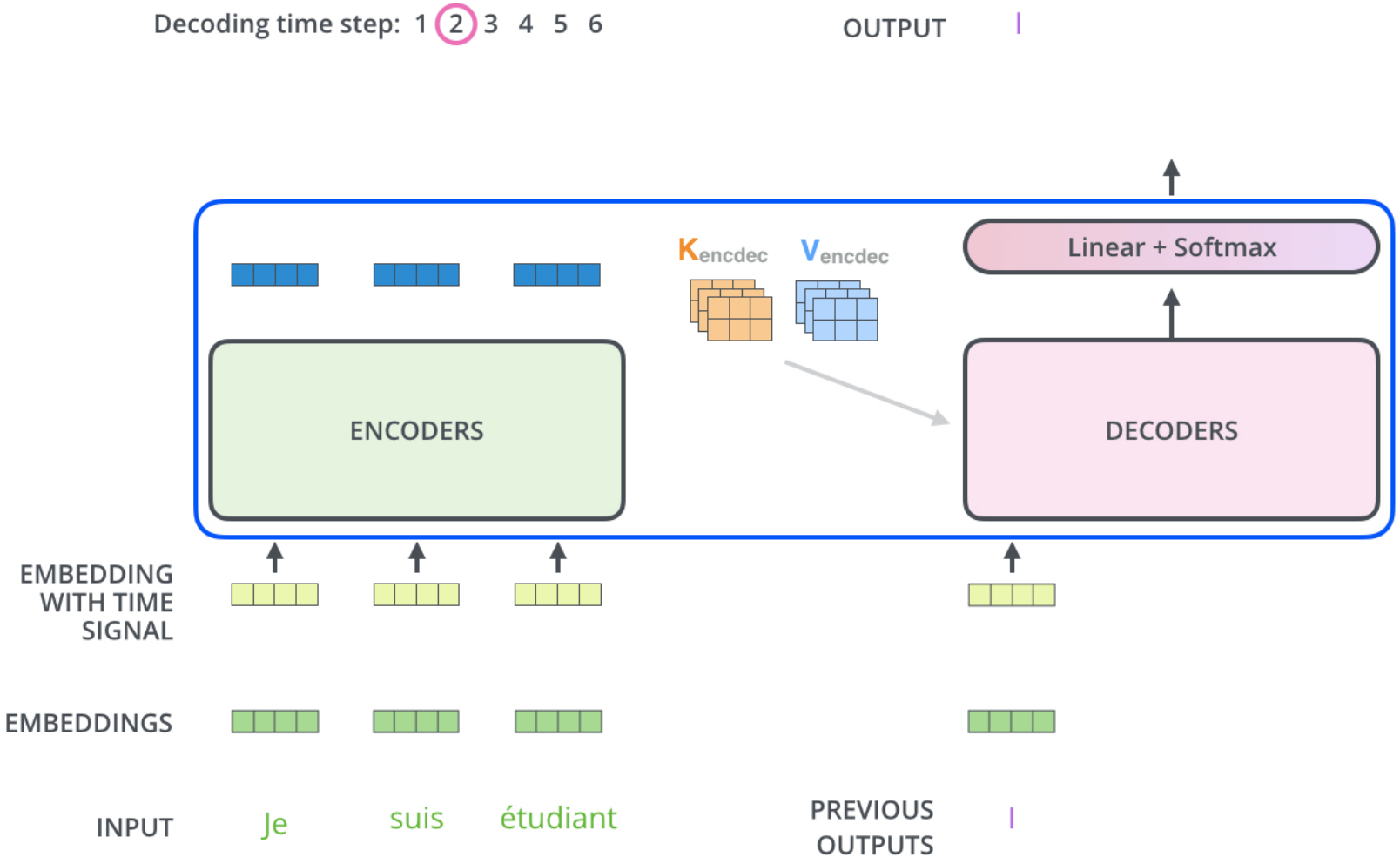


Kodirnik-dekodirnik: delovanje 1/2



Po zaključku faze kodiranja začnemo fazo dekodiranja. Vsak korak v fazi dekodiranja izpiše en žeton na izhod (v tem primeru angleški prevod).

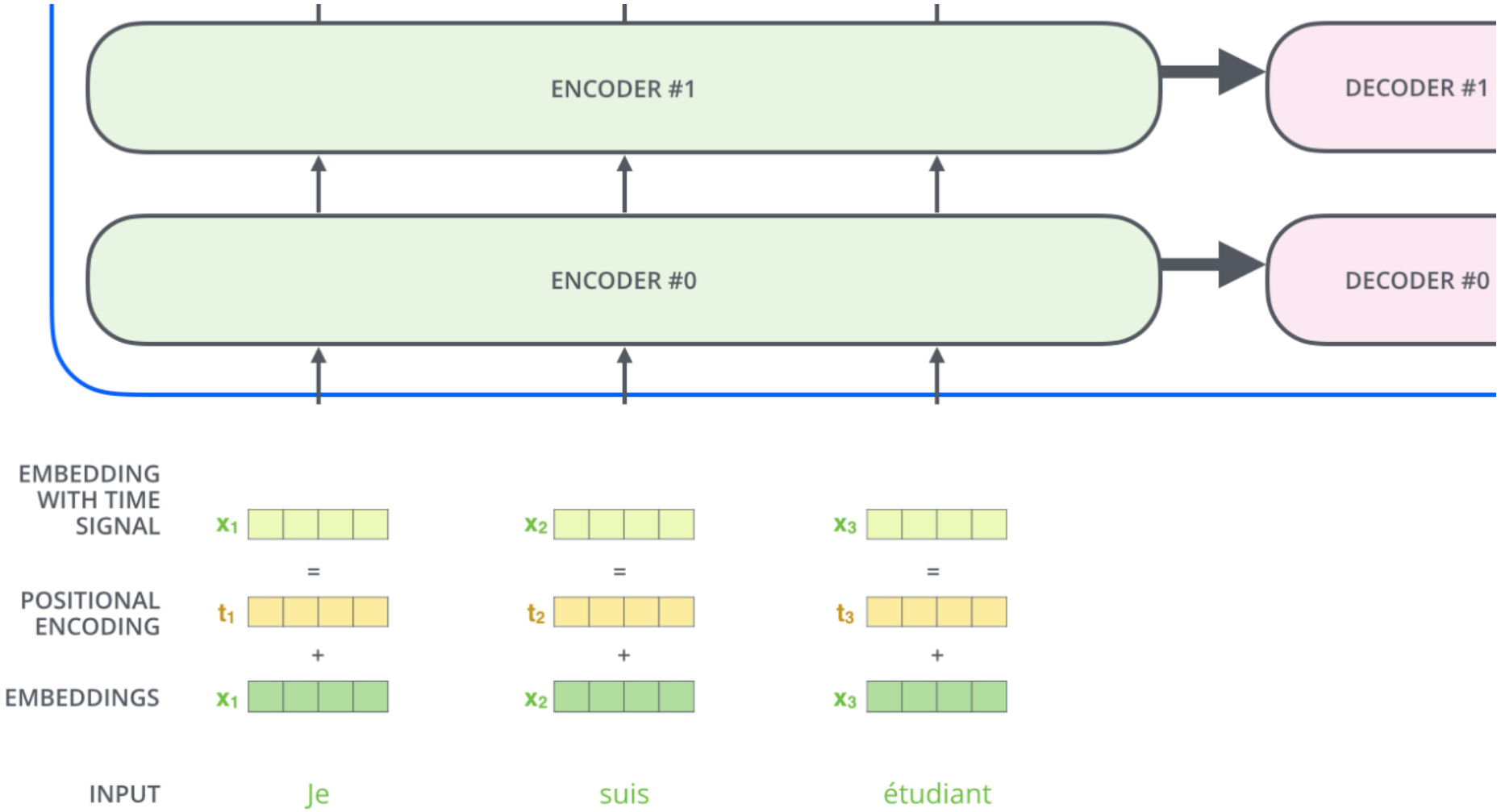
Kodirnik-dekodirnik: delovanje 2/2



Koraki se ponavljajo, dokler ni generiran poseben žeton <eos>, ki označuje konec izhoda. Izhod vsakega koraka se v naslednjem časovnem koraku posreduje na vhod dekoderja. Na vhode dekoderja dodamo tudi pozicijsko kodiranje, da označimo položaj vsakega žetona.

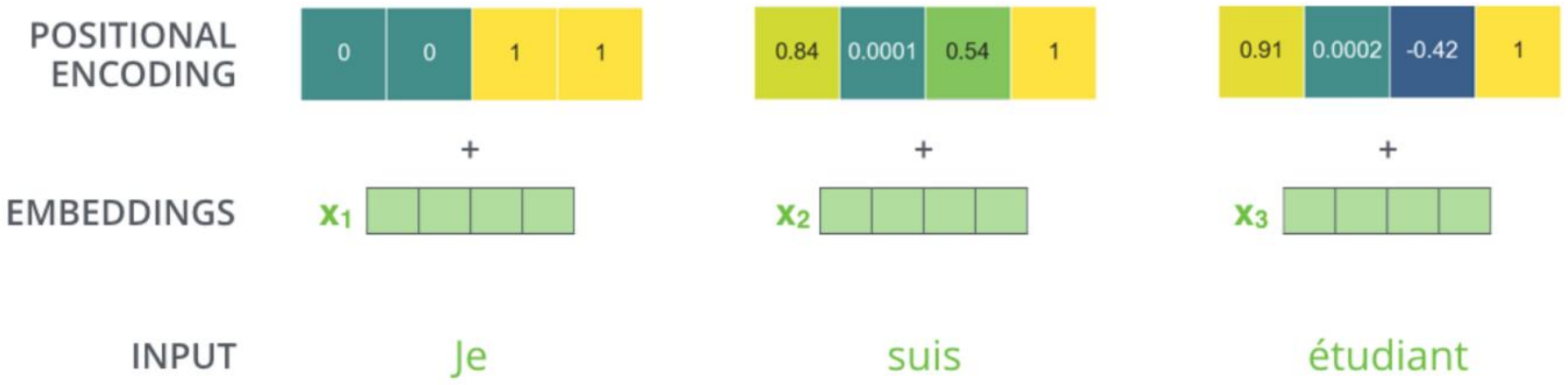
Animacija delovanja transformerja

Dodajmo pozicijsko kodiranje

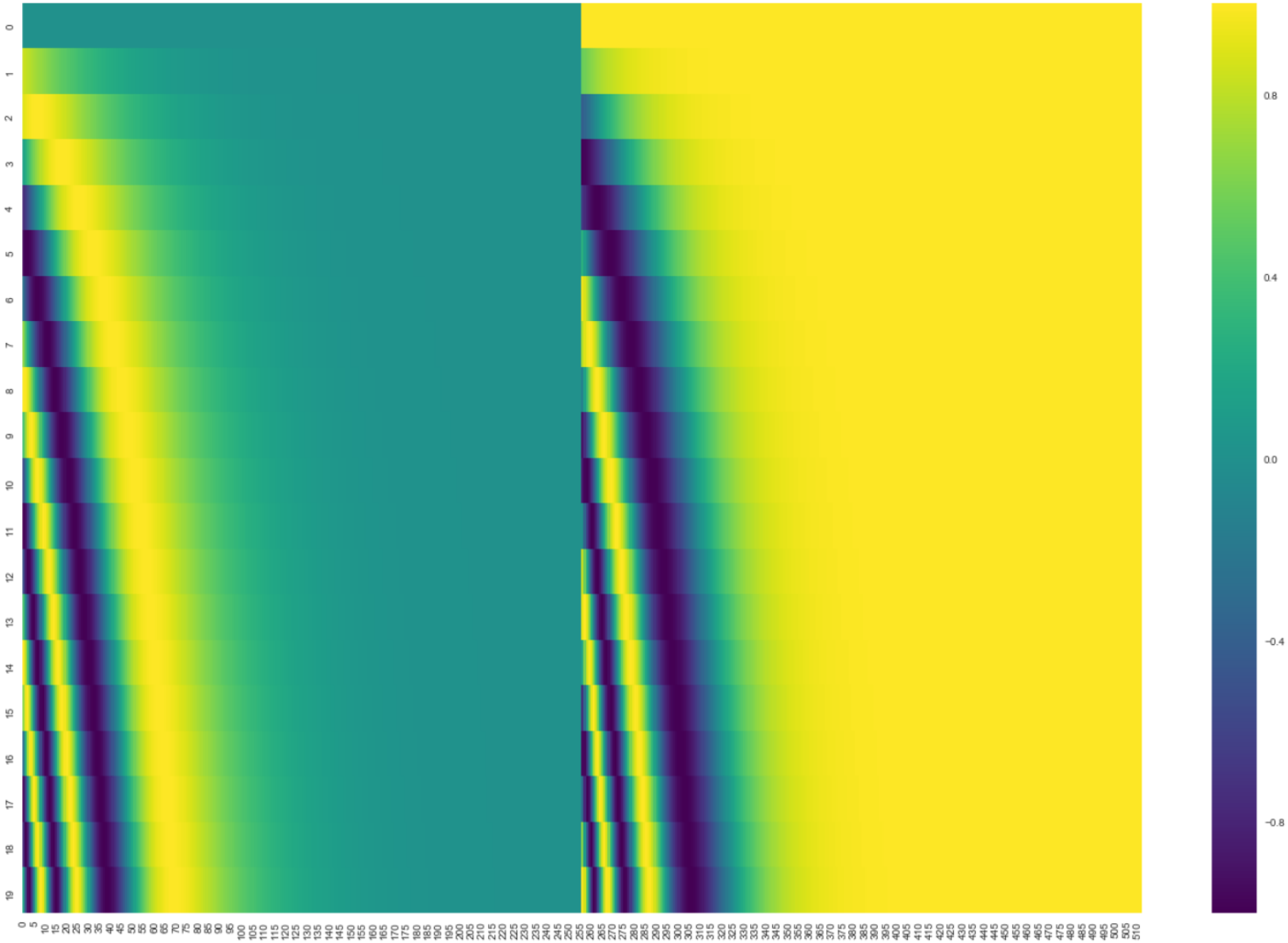


Primer kodiranja pozicije

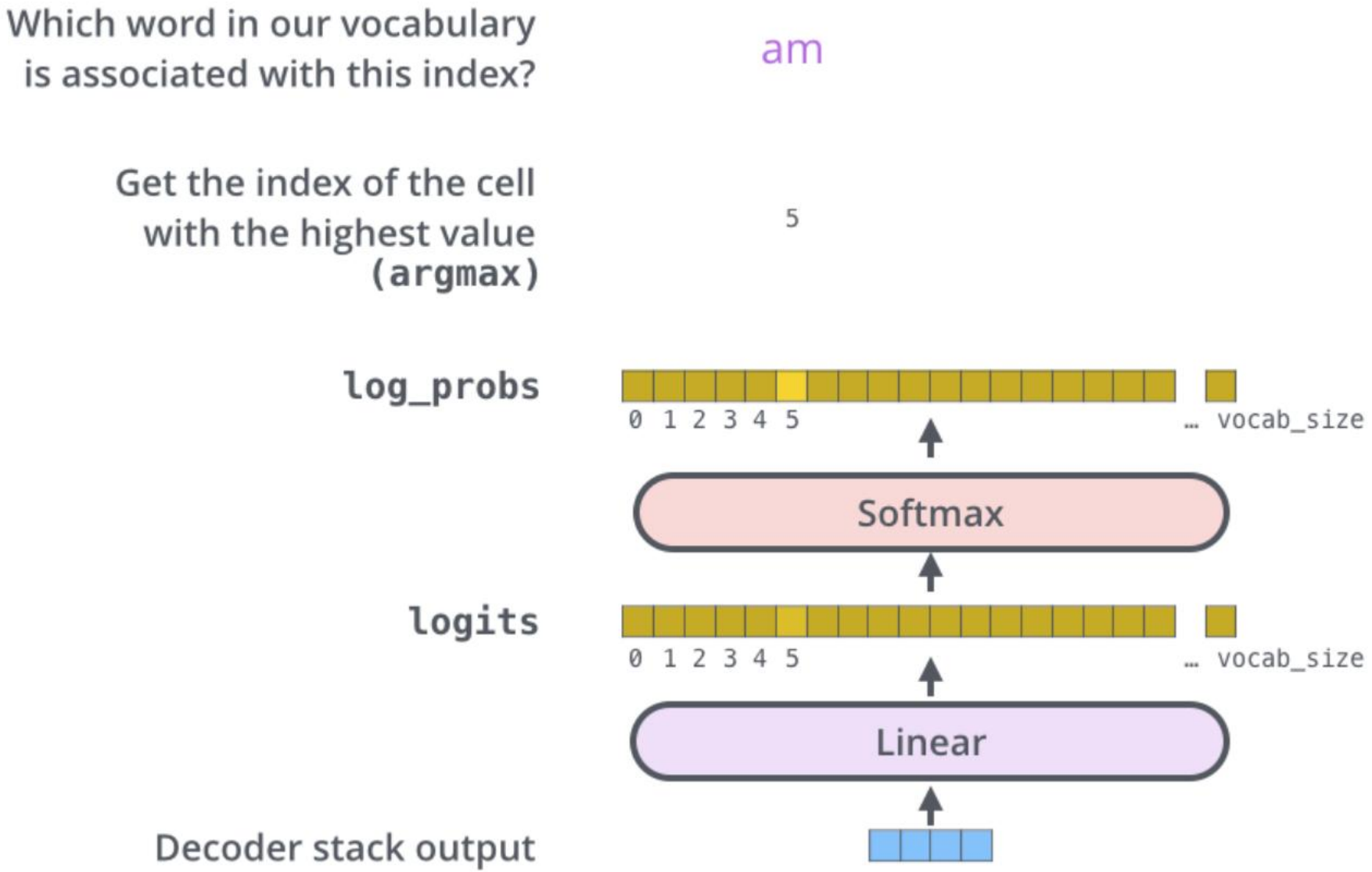
- vrednosti vektorjev za pozicijsko kodiranje sledijo določenemu vzorcu.



Primer pozicijskega kodiranja



Ustvarjanje izhodnih žetonov



Temperatura: parameter VJM

- Parameter temperatura T določa, kako naključen ali determinističen je izbor žetonov med generiranjem.
- Intuicija: nizka temperatura daje prednost žetonom z največjo verjetnostjo. Visoka temperatura povzroči bolj naključno izbiro žetonov.

Matematična definicija T

- Model začne z vrednostmi logit z_i za vsak žeton. Temperatura T jih skalira pred normalizacijo s softmaksom:

$$P(i) = \frac{e^{z_i/T}}{\sum_j e^{z_j/T}}$$

Kjer so:

- $T = 1$: originalne verjetnosti
- $T < 1$: bolj izostrena distribucija
- $T > 1$: bolj sploščena distribucija

Primer

Base probabilities, T=1

Token	Probability
cat	0.70
dog	0.20
rabbit	0.10

T=0.2 → almost always 'cat'.

Distribucija se izostri:

Token	Approx. Probability
cat	~99%
dog	~1%
rabbit	~0%

T=2.0 → more diverse outputs.

Distribucija se splošči

Token	Approx. Probability
cat	~50%
dog	~30%
rabbit	~20%

Typical settings

Temperature	Behavior	Common Use
0.0	Nearly deterministic	Code generation, data extraction
0.2–0.4	Focused and reliable	Technical writing, Q&A
0.5–0.8	Balanced	General chat assistants
1.0	Default-like sampling	Creative writing
1.2–2.0	Highly diverse	Brainstorming, fiction, experimentation

What Temperature Does Not Do

- Temperature does **not** make the model:
  - →More knowledgeable
  - →More intelligent
  - →More truthful
- It only changes how confidently the model follows its probability distribution.
- A wrong answer can still be wrong at temperature 0, and a correct answer can still be generated at temperature 1.5

Top-p (Nucleus) Sampling

UL | FRI

- Instead of considering all tokens, the model only samples from the smallest set whose cumulative probability exceeds p.

Token	Probability
cat	50%
dog	30%
rabbit	15%
elephant	5%

- With:
  - Top-p = 0.8 → {cat, dog}
  - Top-p = 0.95 → {cat, dog, rabbit}
- Typical values:
  - 0.8–0.95
- Temperature and top-p are often used together, although many practitioners tune only one

Učenje transformerja

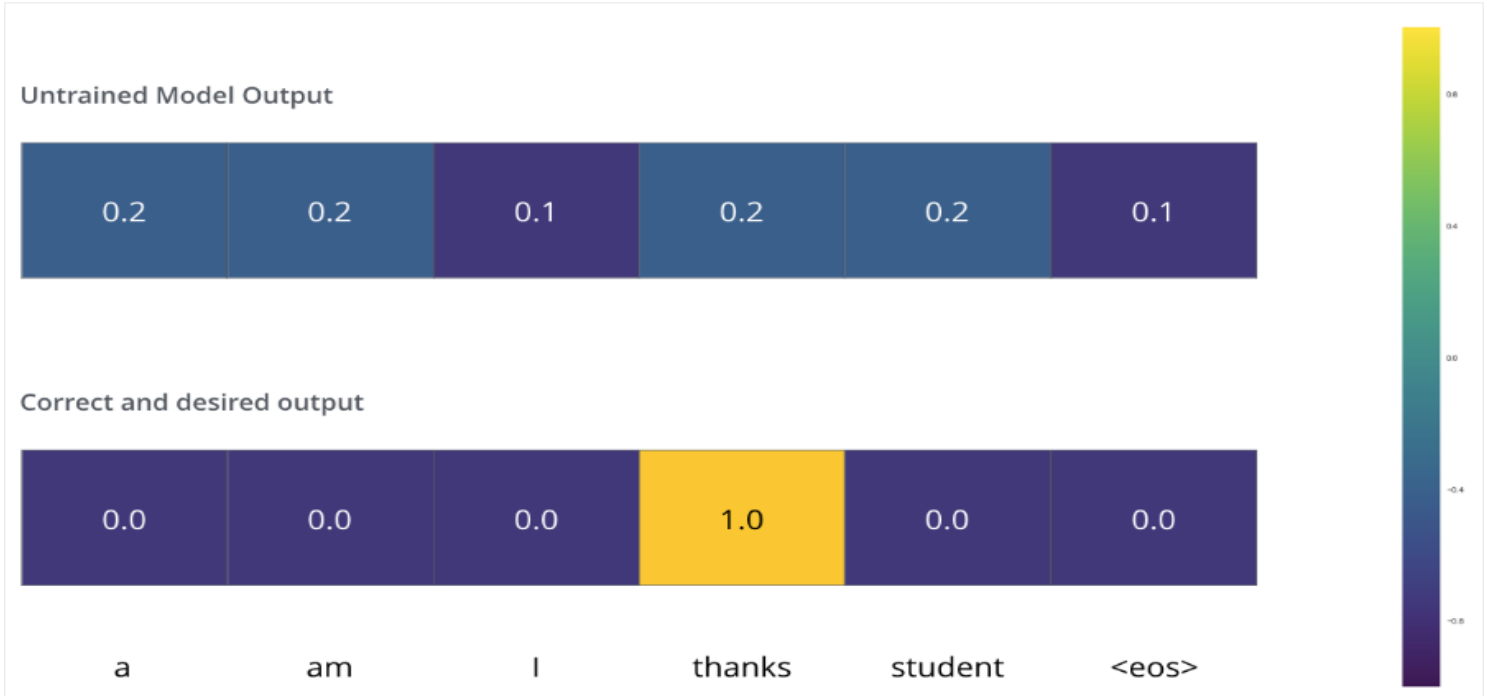
- Med učenjem (nenaučen) model opravi izračun od vhoda do izhoda. Njegov izhod primerjamo z dejanskim pravilnim izhodom.
- Za ilustracijo predpostavimo, da naš izhodni slovar vsebuje le šest besed (a, am, i, thanks, student, <eos>) (dejansko med 30.000 in 250.000)
- Vhod v model je danes običajno velikosti 10^5 žetonov (npr. 130.000)

Output Vocabulary

WORD	a	am	I	thanks	student	<eos>
INDEX	0	1	2	3	4	5

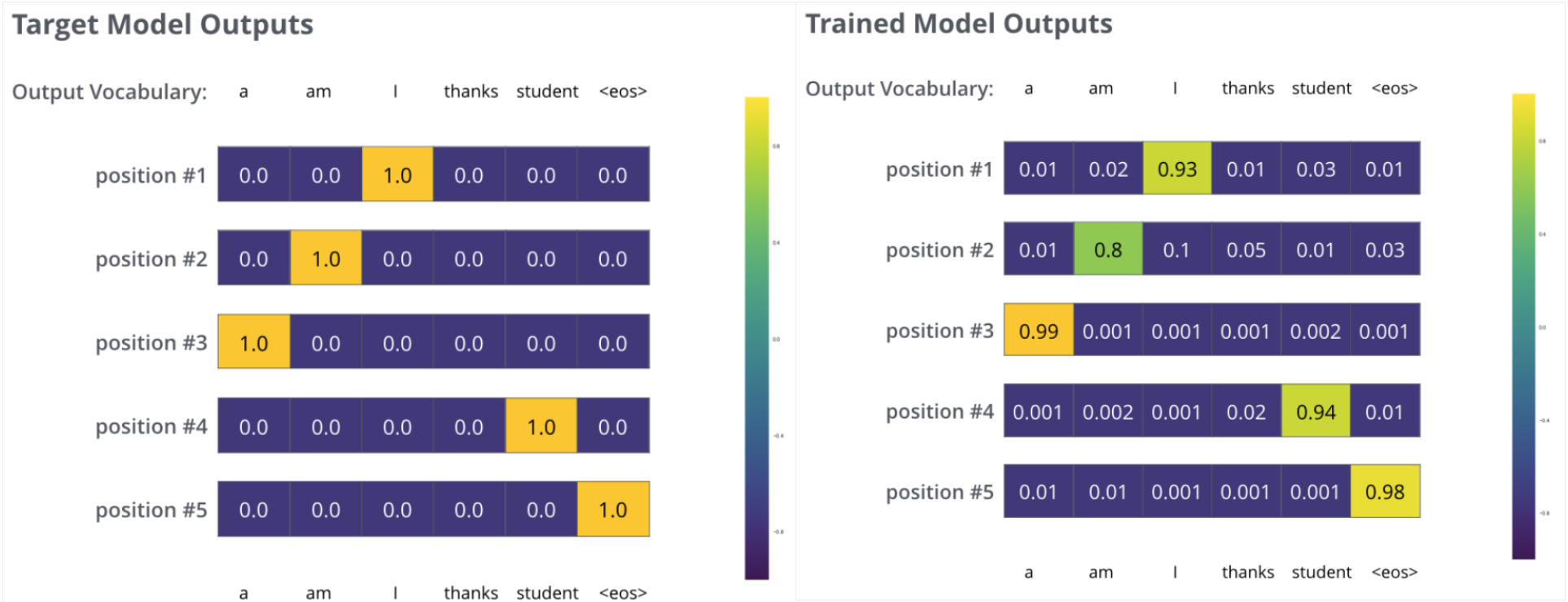
Funkcija izgube

- oceni razliko med pravim izhodom in vrnjenim izhodom
- za besedila kot funkcijo izgube običajno uporabljamo križno entropijo ali Kullback–Leiblerjevo divergenco.
- Izhod modela je verjetnostna porazdelitev, pravi izhod je eniško (1-hot) kodiran, npr.



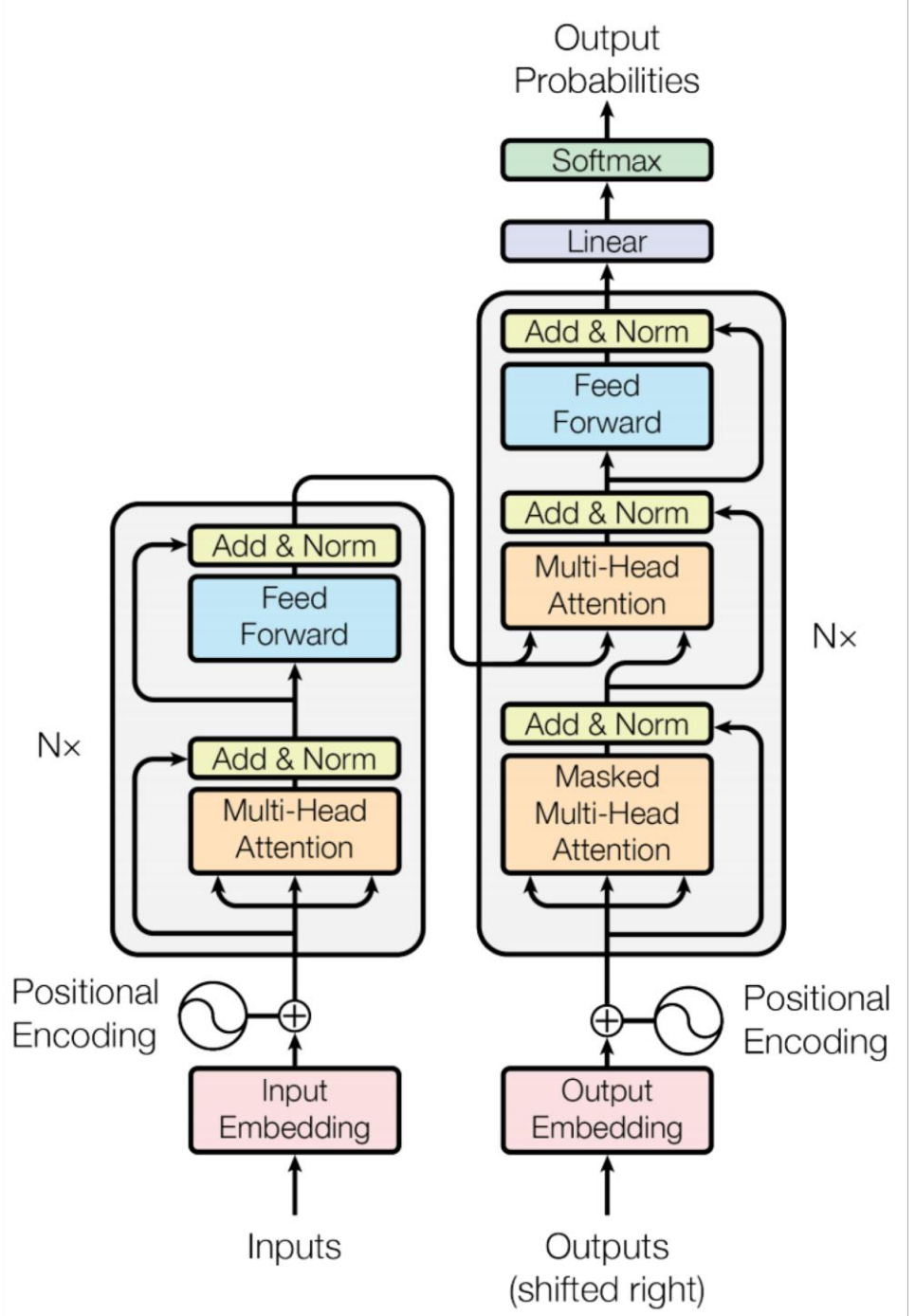
Funkcija izgube za zaporedja

- Funkcijo izgube moramo oceniti za celoten stavek, ne le za en žeton
- Uporabljamo požrešno dekodiranje ali iskanje v snopu



Arhhitektura transformer

- Običajno vhod najprej tokeniziramo s kodiranjem podbesed (subword tokenization)



BPE token learner

An example corpus :(

low low low low low lowest lowest newer newer newer newer newer newer wider wider wider new new

Add end-of-word tokens and segment:

corpus	vocabulary
5 l o w _	_, d, e, i, l, n, o, r, s, t, w
2 l o w e s t _	
6 n e w e r _	
3 w i d e r _	
2 n e w _	

BPE token learner

corpus	vocabulary
5 l o w _	_, d, e, i, l, n, o, r, s, t, w
2 l o w e s t _	
6 n e w e r _	
3 w i d e r _	
2 n e w _	

Merge e r to er

corpus	vocabulary
5 l o w _	_, d, e, i, l, n, o, r, s, t, w, er
2 l o w e s t _	
6 n e w er _	
3 w i d er _	
2 n e w _	

BPE

corpus

5 l o w _
2 l o w e s t _
6 n e w er _
3 w i d er _
2 n e w _

vocabulary

_, d, e, i, l, n, o, r, s, t, w, er

Merge er _ to er_

corpus

5 l o w _
2 l o w e s t _
6 n e w er_
3 w i d er_
2 n e w _

vocabulary

_, d, e, i, l, n, o, r, s, t, w, er, er_

BPE

corpus

5 l o w _
2 l o w e s t _
6 n e w er_
3 w i d er_
2 n e w _

vocabulary

_, d, e, i, l, n, o, r, s, t, w, er, er_

Merge n e to ne

corpus

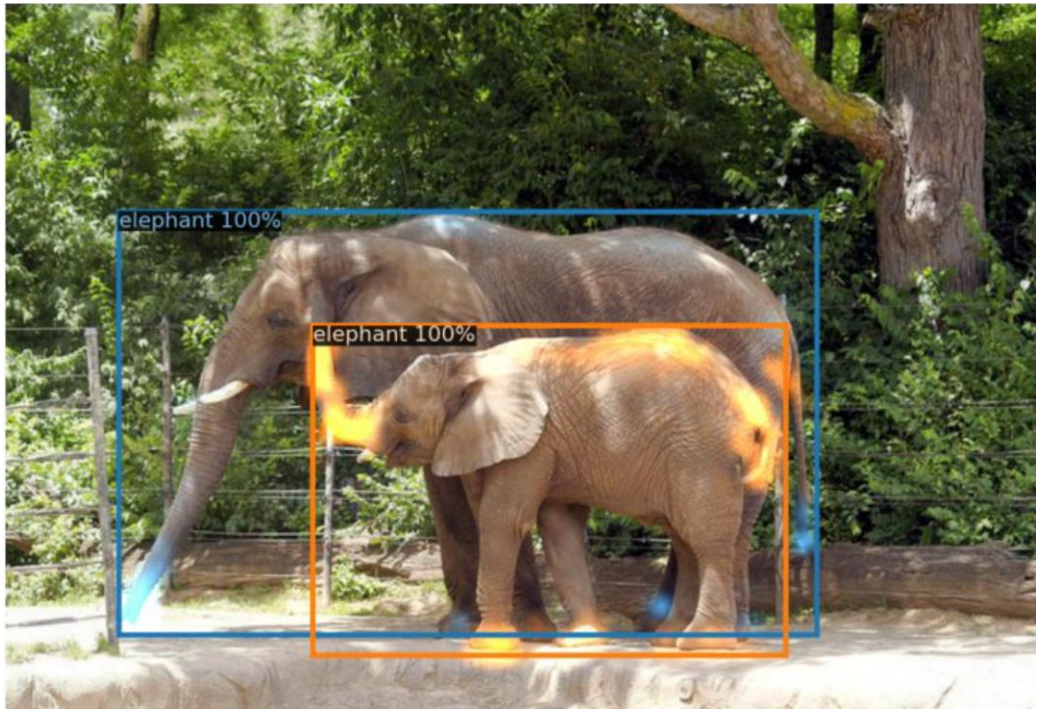
5 l o w _
2 l o w e s t _
6 ne w er_
3 w i d er_
2 ne w _

vocabulary

_, d, e, i, l, n, o, r, s, t, w, er, er_, ne

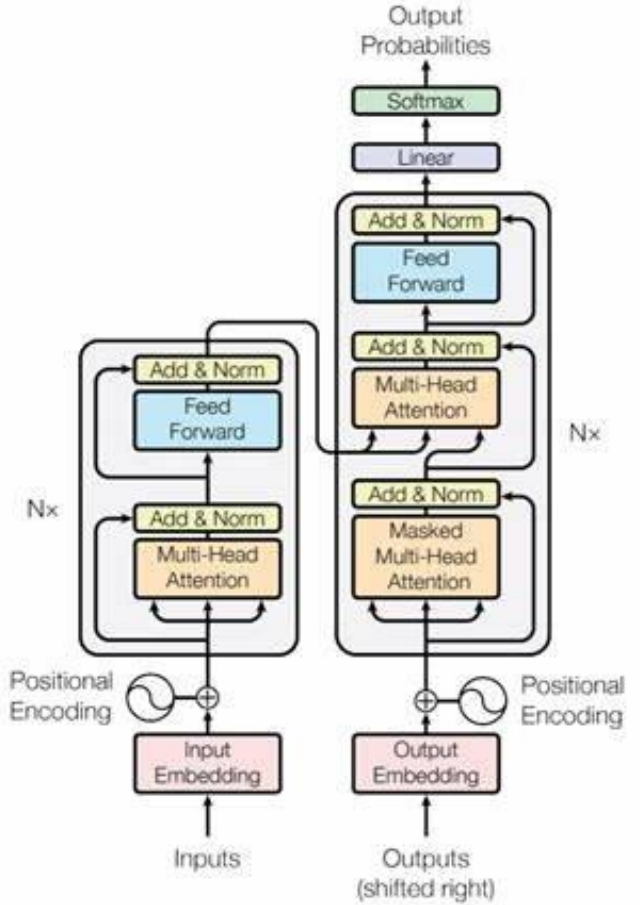
Transformerji so danes povsod

- glasba: slovar sestavljajo MIDI višine, pavze, hitrost
- Zaznavanje objektov: pozornost na predmete, dele slike
- časovne vrste, kjer so signali diskretizirani, vrednosti pa obravnavane kot črke



Današnji veliki jezikovni modeli

- Arhitektura nevronskih mrež transformer
- Učenje v dveh fazah
- 1. faza: prednaučeni modeli
  - napoved maskirane besede, naslednje besede,
  - ogromno besedila
  - zelo dolgo učenje
- 2. faza: model doučimo za specifično nalogo
  - veliko hitreje kot 1. faza
  - prenos splošnega znanja o jeziku iz 1. faze
- Na voljo so številni prostodostopni VJM, tudi večjezični in za slovenščino



Tri različice transformerjev

- Samo kodirnik (BERT)
- kodirnik-dekodirnik (npr. modeli T5 in strojno prevajanje)
- Samo dekodirnik (GPT)

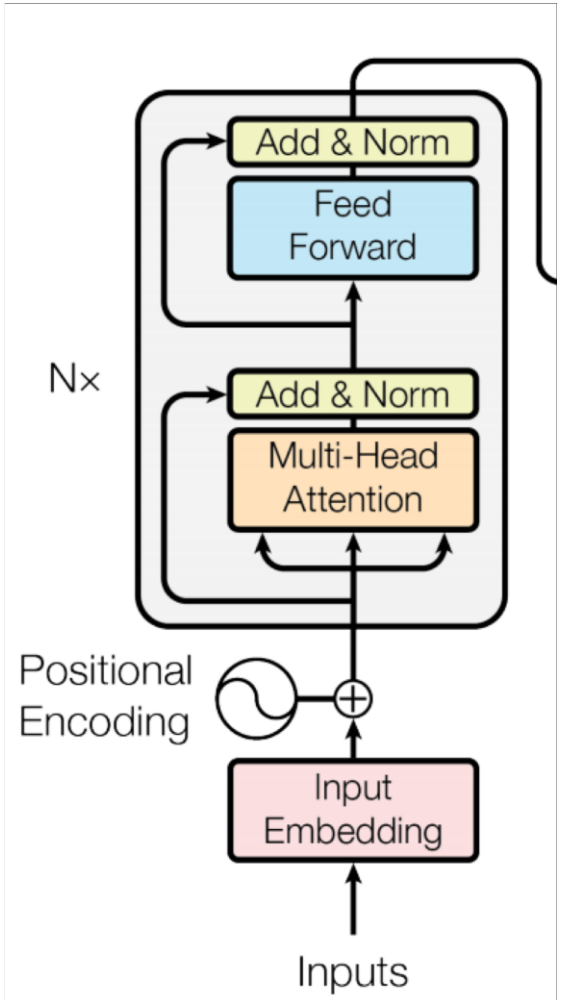
Prilagajanje VJM-jev

Trije glavni načini uporabe in prilagajanja VJM-jev:

- Z ukaznimi pozivi
- Prilagajanje za specifično nalogo
- Prilagajanje sledenja navodilom za prilagoditev domeni

BERT – predstavitveni model

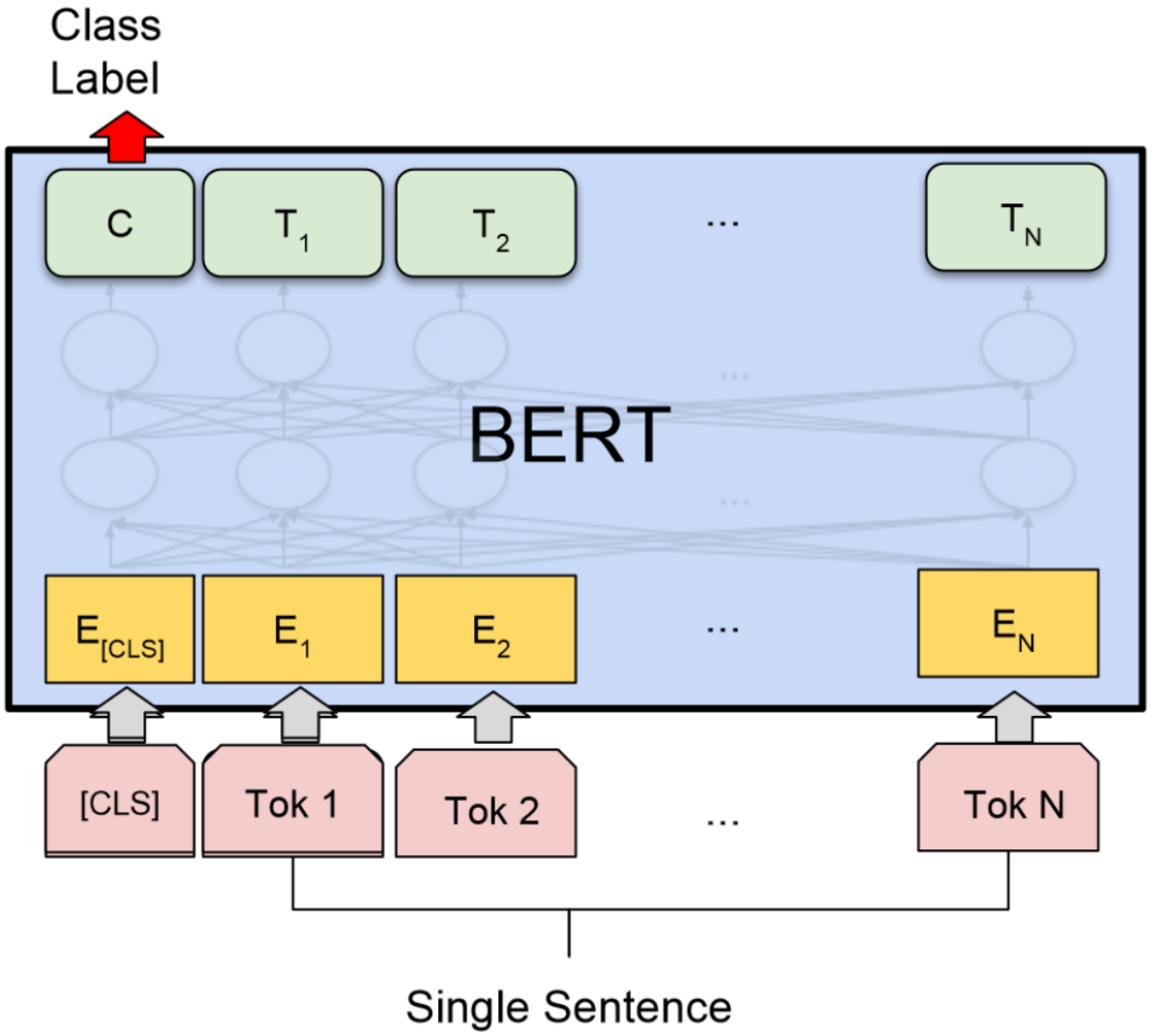
- Med učenjem združuje dve nalogi:
  1. napoveduje zamaskirane besede v stavku
  2. napoveduje vrstni red stavkov: ali je stavku A sledi stavek B ali ne
- 14, 28 ali več plasti samo kodirnika
- prilagajamo za različne **klasifikacijske** naloge
- večjezikovna različica podpira 104 jezike z učenjem na Wikipediji
- mnogo javno dostopnih modelov



- Devlin, J., Chang, M.W., Lee, K. and Toutanova, K., 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Volume 1, pp. 4171-4186.
- Ulčar, M. and Robnik-Šikonja, M., 2021. SloBERTa: Slovene monolingual large pretrained masked language model. *Proceedings of SI-KDD within the Information Society 2021*, pp.17-20.
- Ljubešić, N., & Lauc, D. (2021). BERTić-The Transformer Language Model for Bosnian, Croatian, Montenegrin and Serbian. In Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing (pp. 37-42).
- Ulčar, M., & Robnik-Šikonja, M. (2020). FinEst BERT and CroSloEngual BERT. In *International Conference on Text, Speech, and Dialogue* (pp. 104-111).
- Ulčar, M. & Robnik-Šikonja, M. (2023) Sequence-to-sequence pretraining for a less-resourced Slovenian language. Frontiers in Artificial Intelligence, Section on Language and Computation, Volume 6 – 2023, https://doi.org/10.3389/frai.2023.932519

Prilagajanje modela BERT

- Klasifikacija stavkov z uporabo modela BERT – sentiment, slovnična pravilnost

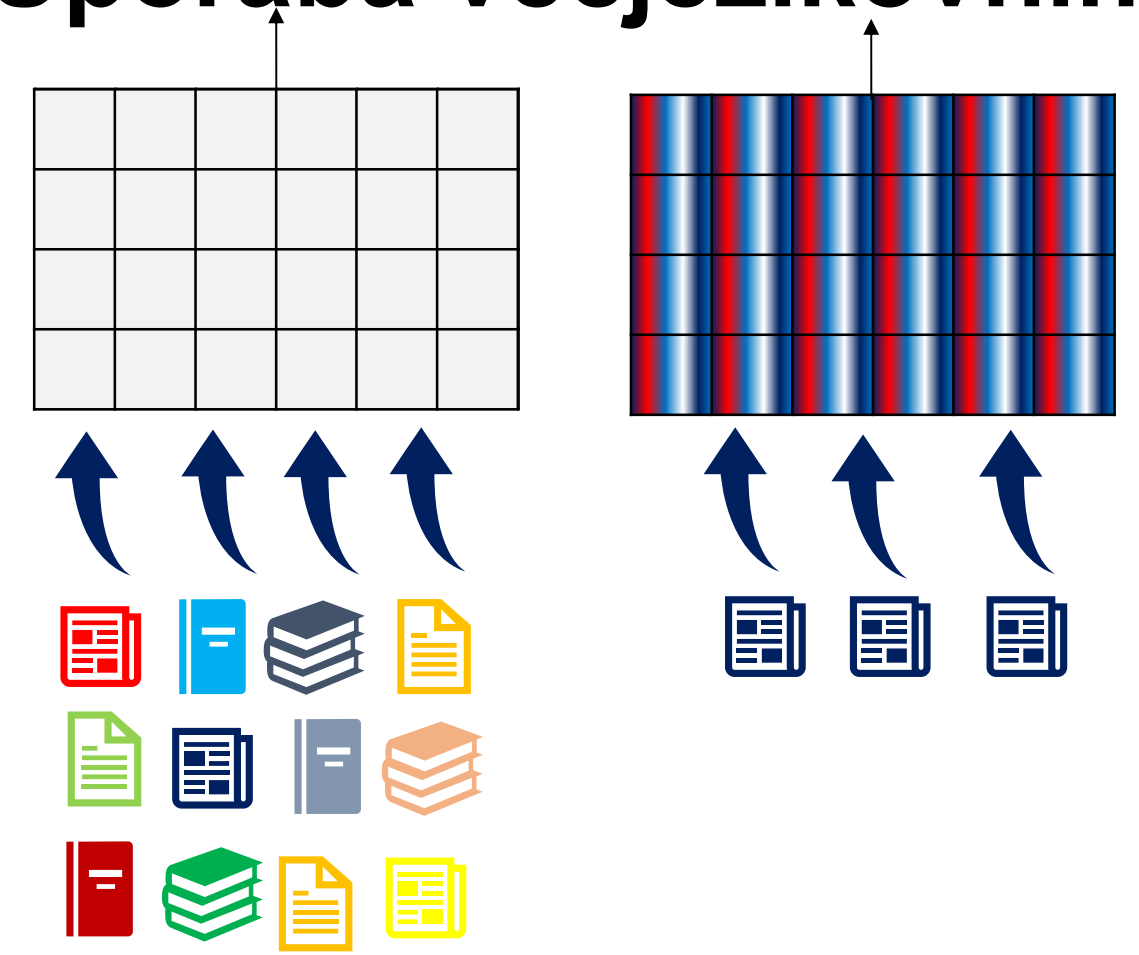


Večjezikovni VJM



- Naučeni na več jezikih hkrati
- Npr. večjezikovni model BERT podpira 104 jezike
- Omogočajo medjezikovni prenos
- Delno rešujejo problem nezadostnih virov za jezike z manj viri

Uporaba večjezikovnih BERT modelov



predsednik je danes najavil ...

Vnaprejšnje učenje

Prilagajanje

Klasifikacija

Prenos brez in z nekaj dodatnimi primeri

- Ulčar, M., & Robnik-Šikonja, M. (2020). FinEst BERT and CroSloEngual BERT. In International Conference on Text, Speech, and Dialogue (pp. 104-111).

University of Ljubljana
Faculty of Computer and Information Science

Zakaj ne samo večjezikovni modeli?

- Uspešnost pri številnih nalogah upade z večjezikovnimi modeli
- rezultati za nekaj nalog v slovenščini (prepoznavanje imenskih entiet – NER, oblikoskladenjsko označevanje – POS, skladenjsko razčlenjevanje – DP, analiza sentimenta – SA, besedne analogije – WA)

Model	NER	POS	DP	SA	WA
mBERT	0.885	0.984	0.681	0.576	0.061
XLM-R	0.912	0.988	0.793	0.604	0.146
SloBERTa	**0.933**	**0.991**	**0.844**	**0.623**	**0.405**

Ulčar, M., Žagar, A., Armendariz, C. S., Repar, A., Pollak, S., Purver, M., & Robnik-Šikonja, M. (2021). Evaluation of contextual embeddings on less-resourced languages. *arXiv preprint arXiv:2107.10614*.

Česa se naučijo veliki večjezikovni modeli?

- We would like to travel to [MASK], ki je najlepši otok v Mediteranu.

SloBERTa: ..., Slovenija, I, Koper, Slovenia

CSE-BERT: Hvar, Rab, Cres, Malta, Brač

XLM-R: Mallorca, Tenerife, otok, Ibiza, Zadar

mBERT: Ibiza, Gibraltar, Tenerife, Mediterranean, Madeira

BERT (en): Belgrade, Italy, Serbia, Prague, Sarajevo

University of Ljubljana
Faculty of Computer and Information Science

Slovarji VJM

- Tokenizacija je odvisna od slovarja VJM
- Slovar VJM je sestavljen statistično(npr. Byte-Pair Encoding)

- Poved: “Letenje je bilo predmet precej starodavnih zgodb.”
- SloBERTa (10 žetonov):

'▁Le', 'tenje', '▁je', '▁bilo', '▁predmet', '▁precej', '▁staroda', 'vnih', '▁zgodb', '.’

- mBERT (17 žetonov):

'Let', '##en', '##je', 'je', 'bilo', 'pred', '##met', 'pre', '##cej', 'star', '##oda', '##vnih', 'z', '##go', '##d', '##b', '.'

Ulčar, M., & Robnik-Šikonja, M. (2021) Training dataset and dictionary sizes matter in BERT models: The case of Baltic languages. In *International Conference on Analysis of Images, Social Networks and Texts* (pp. 162-172)

Prilagajanje: LoRA –Low Rank Approximation za VJM

- Glejte Substack objavo
- vstavimo AB v vsako plast
- r je lahko zelo majhen, npr. 1 % originalne dimenzije d

Finetuned Weights — Weight Update

$$W_{\text{ft}} = W_{\text{pt}} + \Delta W$$

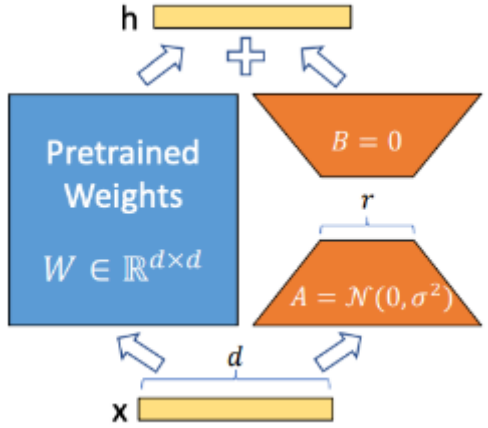
Pretrained Weights

Rank Decomposition Matrix

$$W_{\text{ft}} = W_{\text{pt}} + \underbrace{\Delta W}_{\text{Approximation}} = W_{\text{pt}} + AB$$

$$\text{where} \quad W_{\text{ft}}, W_{\text{pt}}, \Delta W, AB \in \mathbb{R}^{d \times d}$$

$$\text{and} \quad \underbrace{A \in \mathbb{R}^{d \times r}, B \in \mathbb{R}^{r \times d}}_{\text{Low Rank}}$$



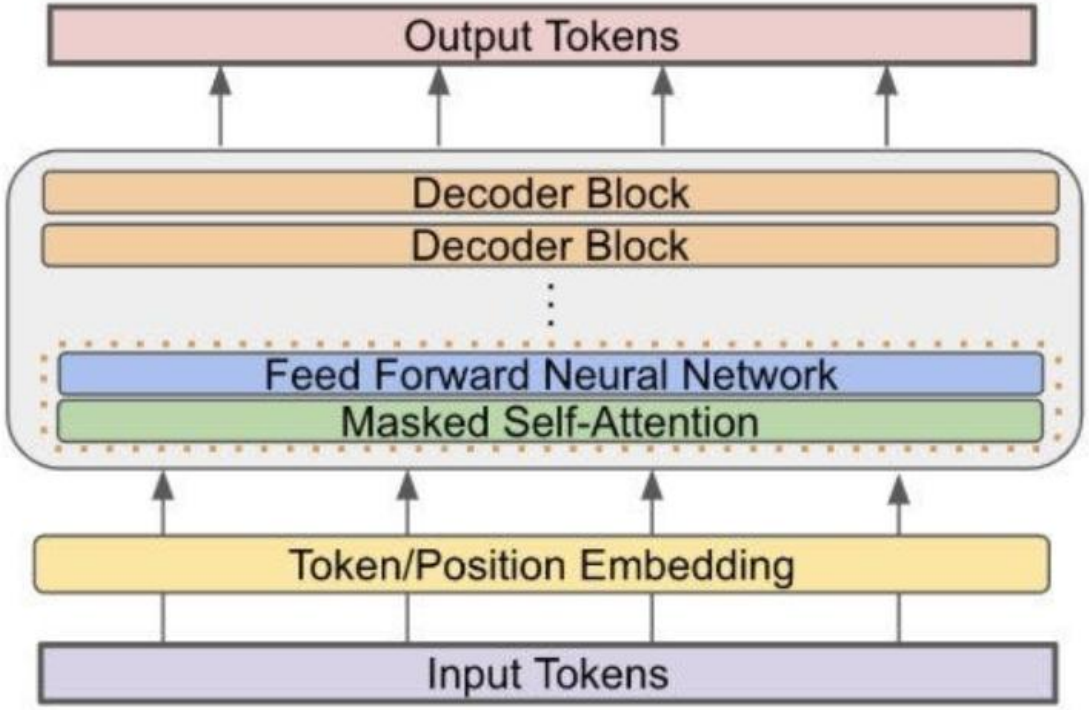
Scaling Factor

$$W_{\text{ft}} = W_{\text{pt}} + \frac{\alpha}{r} AB$$

Rank Decomposition Matrix

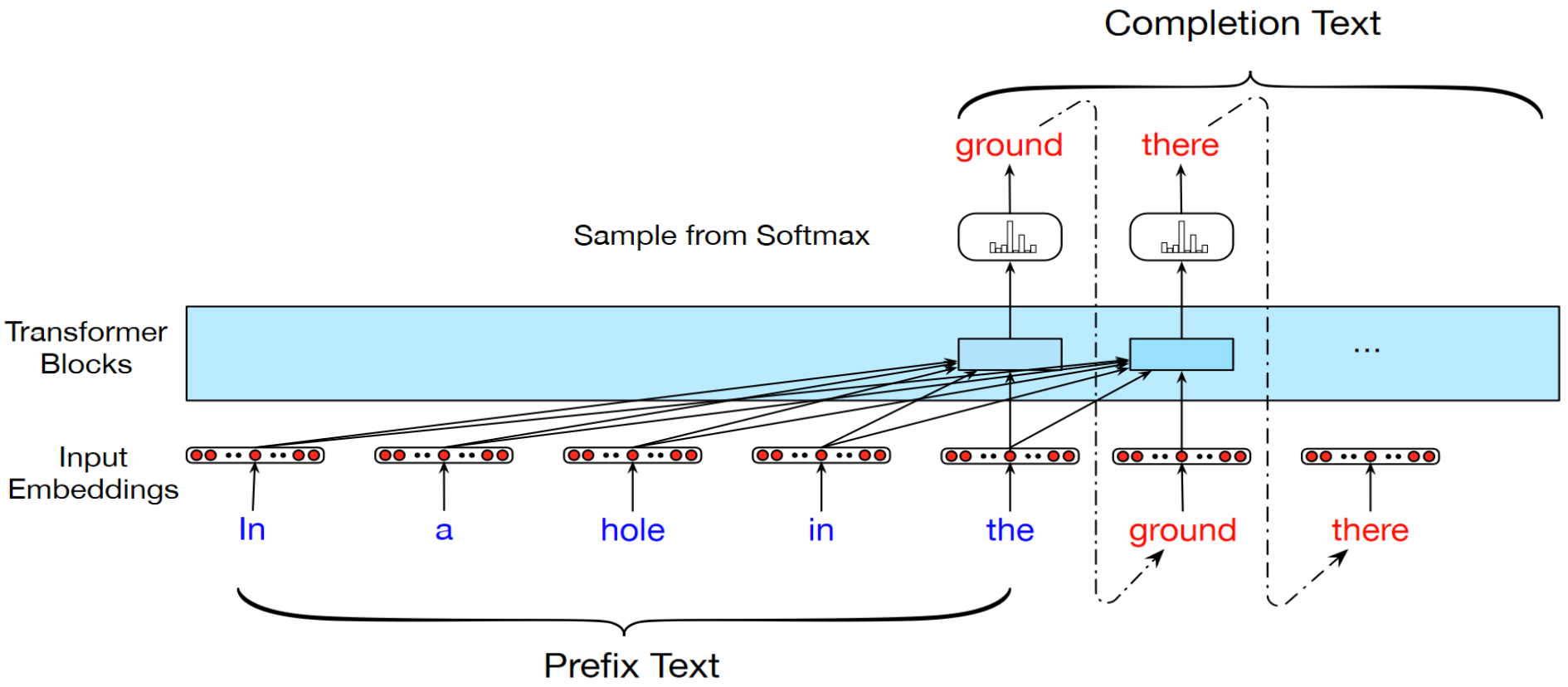
Generativni modeli

- arhitektura transformer: samo dekoder
- GPT, GPT-2, GPT-3, ChatGPT, GPT-4, 5, 6
- Gemini, Gemma
- OPT, LLaMA, LLaMA-2, 3, 4
- MPT, Falcon
- Mistral, Mixtral
- EuroLLM, Open EuroLLM
- OLMo
- DeepSeek, Kimi, Qwen
- GaMS



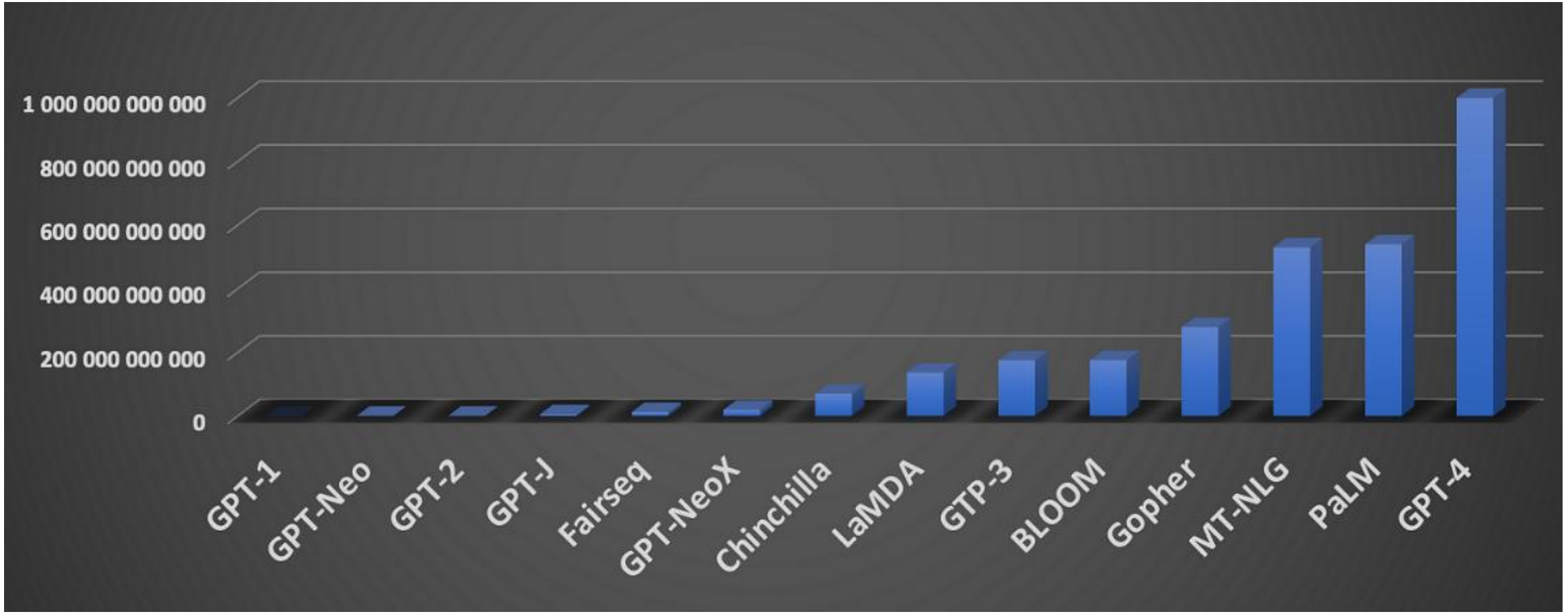
Avtoregresivni generatorji

- generator usmerimo z vhodnim kontekstom, npr. vprašanjem



- uporaba za različne naloge sledenja navodilom: odgovarjanje na vprašanja, povzemanje in druge generativne naloge

Res veliki jezikovni modeli



Veliki generativni jezikovni modeli



- OpenAI: ChatGPT november 2022; temelječ na GPT-3.5, nevronski mreža s 175 milijardami parametrov
- dostop na https://chatgpt.com/
- na podlagi GPT-3.5 z dodatnim učenjem za dialog in sledenje ukazom
- Uporabljal je RLHF (spodbujevano učenje s človeškimi povratnimi informacijami)
- Ideja: dodaten napovedni model, ki se nauči rangirati kakovost odgovorov. Ta model dodatno nadzoruje učenje VJM pri dialogih
- Danes mešanica ekspertov (Mixture of Experts, MoE), npr.
  8 x 220 milijard parametrov in agentni sistemi
- velik vpliv na javnost,
- potencialno velike spremembe za vse poklice, ki uporabljajo jezik, tudi za učenje, študij, poučevanje, znanstveno pisanje, intelektualne poklice
- LLaMa, Alpaca, Koala, LiMa, Gemma, Qwen, Kimi, EuroLLM, GaMS: manjši, odprti in zelo konkurenčni modeli

GPT-2 and GPT-3

- Layer norm applied to input of each subblock
- GPT-3 also uses some sparse attention layers
- More data, larger batch sizes (GPT-3 uses batch size of 3.2M)
- The models are scaled:

GPT-2:
48 layers, 25 heads
d_m = 1600, d = 64
context size = 1024
~ 1.5B parameters

GPT-3:
96 layers, 96 heads
d_m = 12288, d = 128
context size = 2048
~ 175B parameters

[1] Radford et al.: Language Models are Unsupervised Multitask Learners, 2019.
[2] Brown et al.: Language Models are Few-Shot Learners, 2020.

- see demos at https://transformer.huggingface.co/

Oblikovanje ukaznih pozivov

- prompt engineering (PE)
- relativno novo področje za razvoj in optimizacijo ukaznih pozivov za učinkovito uporabo VJM
- oblikovanje robustnih in učinkovitih ukaznih pozivov za VJM in ostale pristope
- v bistvu nabor mehkih vodil kako načrtovati ukazne pozive
- vse bolj avtomatizirani pristopi
- glej https://www.promptingguide.ai/

Osnove oblikovanja pozivov

- Oblikovanje ukaznih pozivov temelji na mehanizmu pozornosti in razmeroma dolgem vhodu (kontekstu) VJM-jev
- Ljudje običajno ustvarjamo razmeroma kratke pozive
- Dober poziv mora zagotoviti dovolj konteksta
- Majhna varianca v pozivu lahko povzroči veliko varianco v rezultatih
- Oblikovanje ukaznih pozivov je nekje med inženirstvom in ugibanjem, kaj bi lahko delovalo
- Vendar pa obstajajo nekateri dobro uveljavljeni pristopi, kot je s pridobivanje obogateno pridobivanje (Retrieval Augmented Generation, RAG)

Osnove ukaznih pozivov

- Kaj jenarobe s tem pozivom?
- Prompt

The sky is

- Output:

blue.

- Namen, kontekst?

Osnove ukaznih pozivov

- Boljše?

Complete the sentence:

The sky is

- Output:

blue during the day and dark at night.

- Še vedno brez namena in primerov

Tipični elementi pozivov

- **Ukaz –** določena naloga ali ukaz, ki naj jo model izvaja
- **Kontekst –** zunanje informacije ali dodatni kontekst, ki lahko usmeri model k boljšim odgovorom
- **Vhodni podatki –** vhod ali vprašanje, na katerega želimo najti odgovor
- **Indikator izhoda –** vrsta ali format izhoda.

Primer

Classify the text into neutral, negative, or positive

Text: I think the food was okay.

Sentiment:

- Tukaj ni konteksta; kaj bi bilo primerno?

Splošna navodila za ročno oblikovanje pozivov

- Oblikovanje pozivov je iterativno, za dobre rezultate zahteva precej eksperimentiranja
- Začnimo z enostavnimi pozivi in postopoma dodajmo več elementov ter konteksta
- V mnogih primerih bodo specifičnost, preprostost in natančnost prinesli boljše rezultate.
- Za večje naloge z več različnimi podnalogami poskusimo nalogo razdeliti na enostavnejše podnaloge in jih postopoma povečevati, ko dobivamo boljše rezultate.

Specifičnost

- Prompt:

Extract the name of places in the following text.

Desired format:

Place: <comma_separated_list_of_places>

Input: "Although these developments are encouraging to researchers, much is still a mystery. “We often have a black box between the brain and the effect we see in the periphery,” says Henrique Veiga-Fernandes, a neuroimmunologist at the Champalimaud Centre for the Unknown in Lisbon. “If we want to use it in the therapeutic context, we actually need to understand the mechanism.“"

- Output:

Place: Champalimaud Centre for the Unknown, Lisbon

- Naj bodo ukazni pozivi dovolj opisni in podrobni
- V pozivu navedimo primere

Povej, kaj naj naredi, izogibaj se negacijam

- VJM so semantično šibki pri negacijah
- Prompt:

The following is an agent that recommends movies to a customer. DO NOT ASK FOR INTERESTS. DO NOT ASK FOR PERSONAL INFORMATION.

Customer: Please recommend a movie based on my interests.
Agent:

- Output:

Sure, I can recommend a movie based on your interests. What kind of movie would you like to watch? Do you prefer action, comedy, romance, or something else?

- A better prompt:

The following is an agent that recommends movies to a customer. The agent is responsible to recommend a movie from the top global trending movies. It should refrain from asking users for their preferences and avoid asking for personal information. If the agent doesn't have a movie to recommend, it should respond "Sorry, couldn't find a movie to recommend today.".

Customer: Please recommend a movie based on my interests.

Agent:

- Output:

Sorry, I don't have any information about your interests. However, here's a list of the top global trending movies right now: [list of movies]. I hope you find something you like!

Pozivi z nekaj primeri

- Primeri podajo kontekst za usmerjanje modela k boljšemu delovanju.
- Poziv:

A "whatpu" is a small, furry animal native to Tanzania. An example of a sentence that uses the word whatpu is:

We were traveling in Africa and we saw these very cute whatpus.

To do a "farduddle" means to jump up and down really fast. An example of a sentence that uses the word farduddle is:

- Izhod:

When we won the game, we all started to farduddle in celebration.

- Različno število primerov: 1, 3, 10 (1-shot, 3-shot, 10-shot)

Nasveti za dodajanje primerov v poziv

- Pomembna sta tako prostor label (oznak) in razporeditev vhodnega besedila (ne glede na to, ali so oznake primerov pravilne)
- Pomemben je format primerov
- Boljše so naključne oznake, kot jih ni
- Distribucija v primerih uporabljenih oznak naj sledi dejanski distribucij.

Min, S., Lyu, X., Holtzman, A., Artetxe, M., Lewis, M., Hajishirzi, H. and Zettlemoyer, L., 2022, December. Rethinking the Role of Demonstrations: What Makes In-Context Learning Work?. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (pp. 11048-11064).

Primer

- Naključne oznak še vedno delujejo
- Prompt

This is awesome! // Negative
This is bad! // Positive
Wow that movie was rad! // Positive
What a horrible show! //

- Output:

Negative

Omejitve dodatnih primerov v pozivih

- Dodatni primeri pomagajo le pri nekaterih nalogah
- Če pozivi brez primerov ali z nekaj primeri ne zadoščajo, razmislimo o prilagajanju modela in naprednejših tehnikah pozivanja, kot je sklepanje z verigo misli, agenti, itd.

Primer: poziv brez primerov

- Prompt

The odd numbers in this group add up to an even number: 15, 32, 5, 13, 82, 7, 1.

Answer:

- Output:

Yes, the odd numbers in this group add up to 107, which is an even number.

Primer: nekaj primerov

- Prompt:

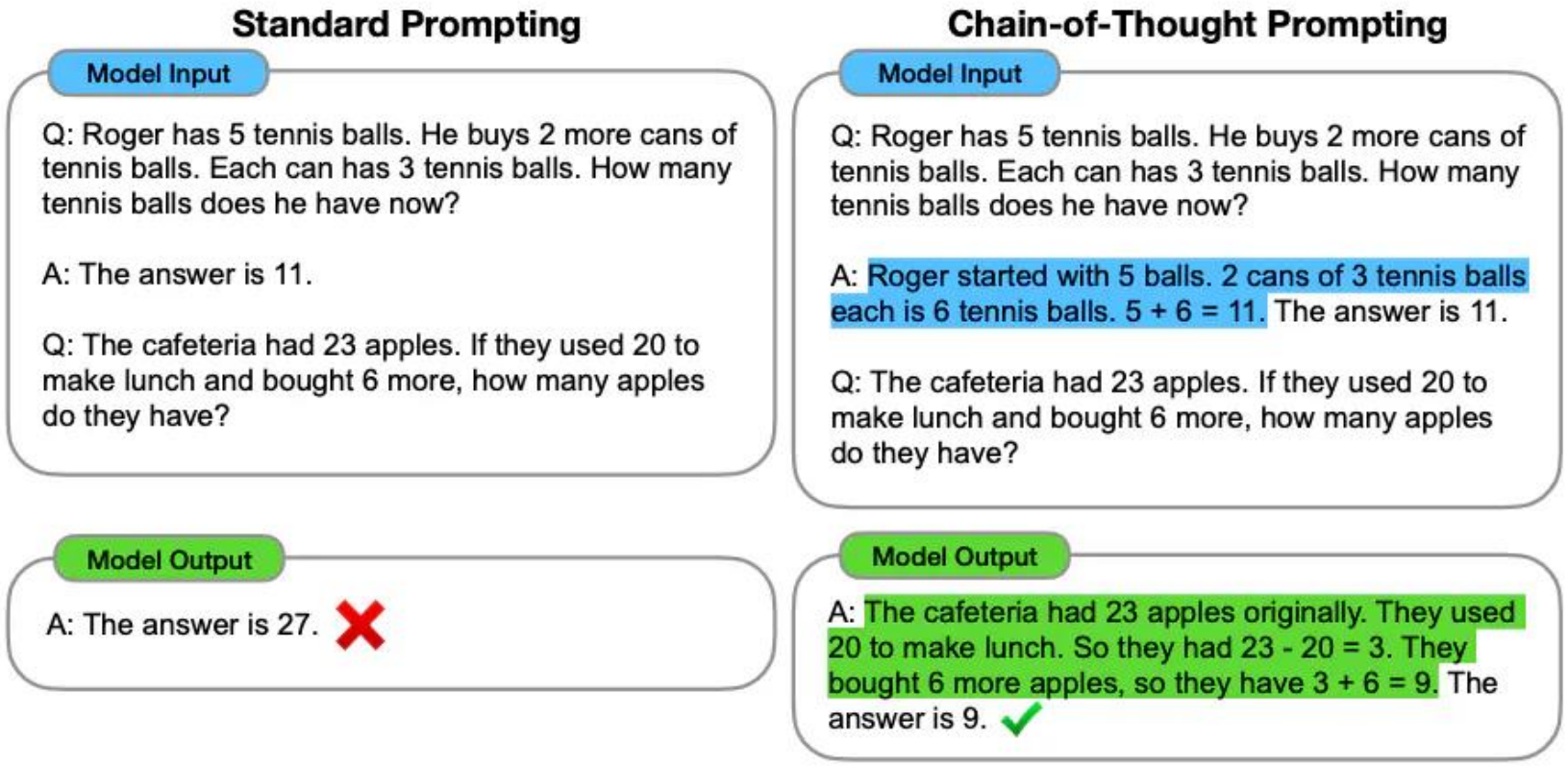
The odd numbers in this group add up to an even number: 4, 8, 9, 15, 12, 2, 1.
A: The answer is False.
The odd numbers in this group add up to an even number: 17, 10, 19, 4, 8, 12, 24.
A: The answer is True.
The odd numbers in this group add up to an even number: 16, 11, 14, 4, 8, 13, 24.
A: The answer is True.
The odd numbers in this group add up to an even number: 17, 9, 10, 12, 13, 4, 2.
A: The answer is False.
The odd numbers in this group add up to an even number: 15, 32, 5, 13, 82, 7, 1.
A:

- Output:

The answer is True.

- Naloga zahteva več korakov sklepanja

Sklepanje z verigo misli (Chain-of-Thought prompting, CoT)



Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V. and Zhou, D., 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, *35*, pp.24824-24837.

Primer verige misli

- Prompt:

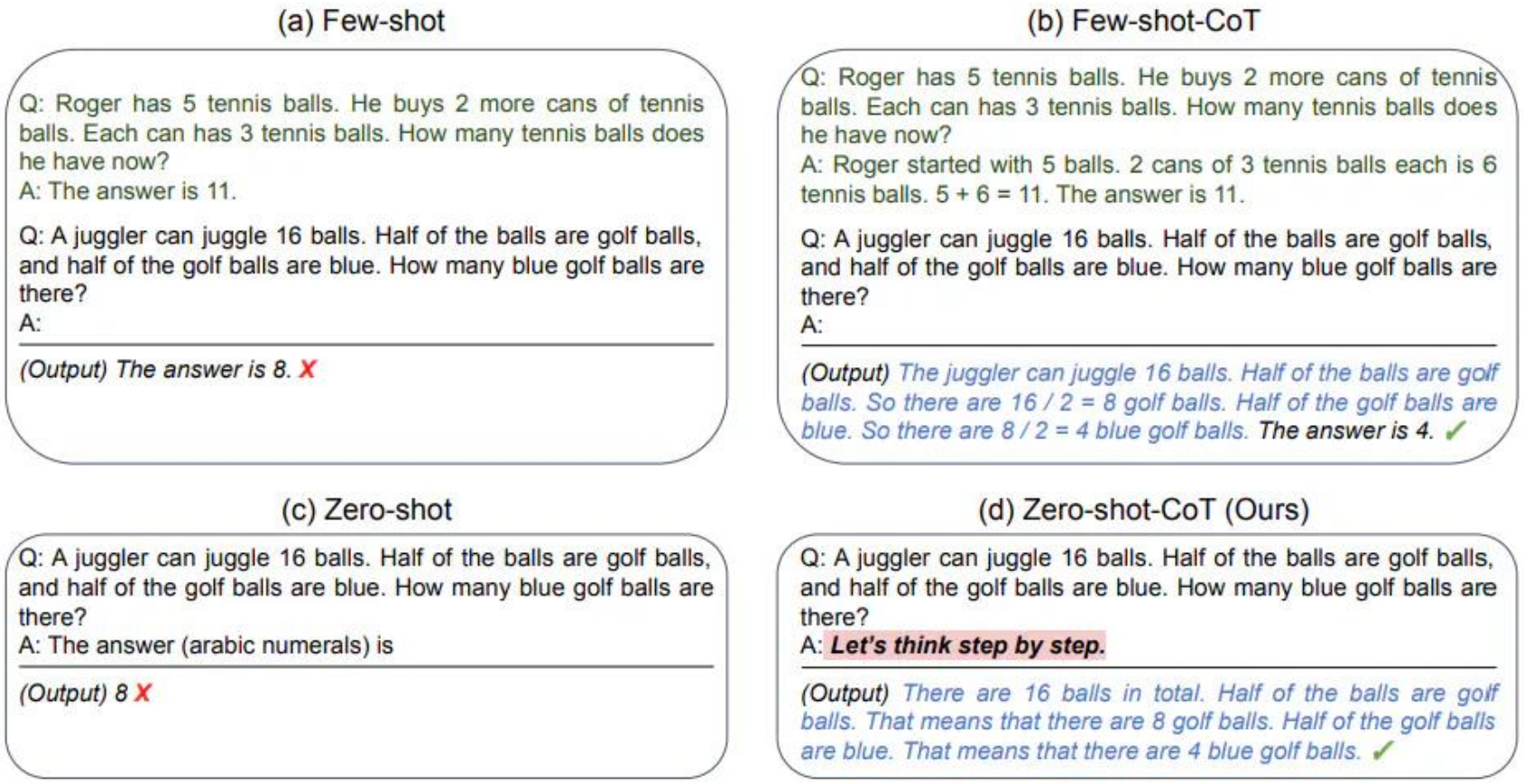
The odd numbers in this group add up to an even number: 4, 8, 9, 15, 12, 2, 1.
A: Adding all the odd numbers (9, 15, 1) gives 25. The answer is False.
The odd numbers in this group add up to an even number: 17, 10, 19, 4, 8, 12, 24.
A: Adding all the odd numbers (17, 19) gives 36. The answer is True.
The odd numbers in this group add up to an even number: 16, 11, 14, 4, 8, 13, 24.
A: Adding all the odd numbers (11, 13) gives 24. The answer is True.
The odd numbers in this group add up to an even number: 17, 9, 10, 12, 13, 4, 2.
A: Adding all the odd numbers (17, 9, 13) gives 39. The answer is False.
The odd numbers in this group add up to an even number: 15, 32, 5, 13, 82, 7, 1.
A:

- Output:

Adding all the odd numbers (15, 5, 13, 7, 1) gives 41. The answer is False.

- Deluje tudi z enim razširjenim primerom

COT brez primerov



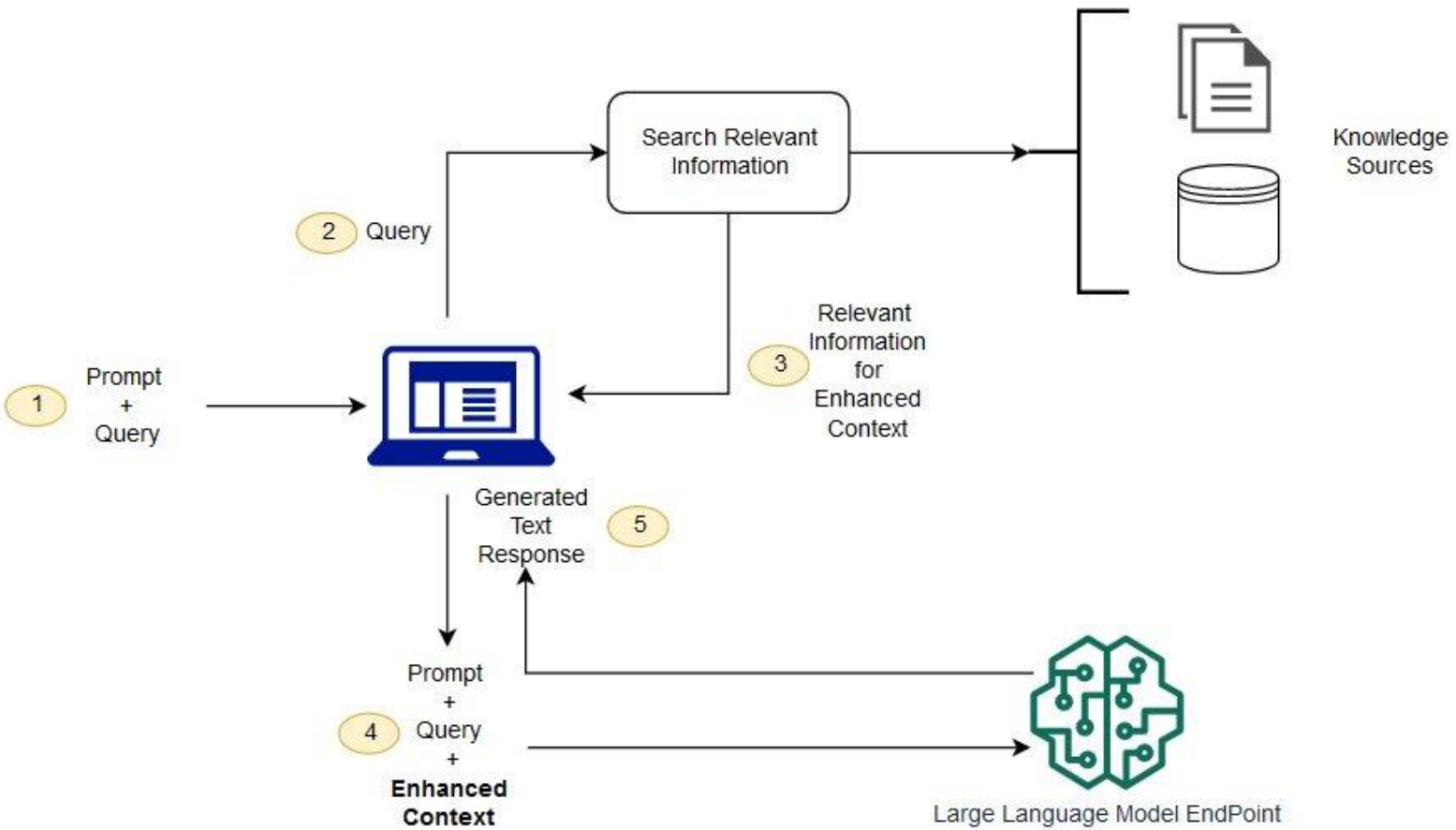
- Tak poziv deluje tudi v mnogih drugih primerih

Kojima, T., Gu, S.S., Reid, M., Matsuo, Y. and Iwasawa, Y., 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, *35*, pp.22199-22213.

S pridobivanjem obogateno generiranje

- Retrieval Augmented Generation (RAG)
- Za zahtevne naloge in naloge, ki zahtevajo dostop do zunanjih virov znanja
- Izboljšuje pravilnost, konsistentnost in zanesljivost generiranih odgovorov, zmanjšuje halucinacije
- RAG sprejme vhodne podatke in pridobi nabor relevantnih/podpornih dokumentov glede na vir (npr. Wikipedia). Primerni koščki dokumenta se združijo v kontekst in se skupaj z izvornim vhodnim pozivom uporabijo kot vhod v VJM
- RAG se lahko prilagaja dinamičnim situacijam (dejstva se lahko sčasoma spreminjajo)
- uspešen pri mnogih nalogah odgovarjanja na vprašanja (question answering, QA)

RAG shema

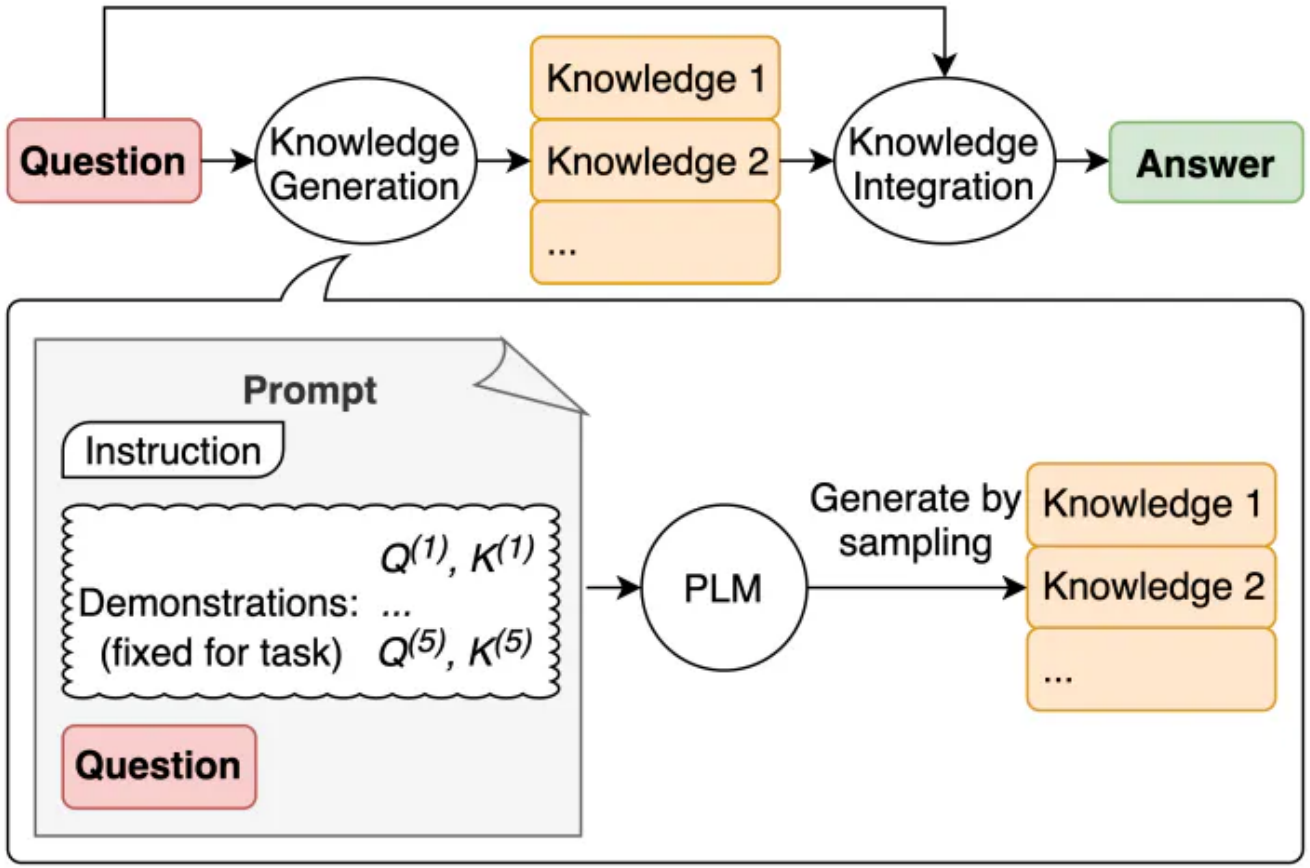

Search Relevant Information
Knowledge Sources
2 Query
Relevant Information for Enhanced Context
3
1 Prompt + Query
Generated Text Response
5
Prompt + Query + Enhanced Context
4
Large Language Model EndPoint

Pozivanje z generiranjem znanja

- RAG skuša v poziv vključiti znanje ali informacije iz baze podatkov, da bi modelu pomagal pri boljših napovedih.
- Če baze podatkov nimamo, podobna ideja uporablja VJM za generiranje znanja pred napovedjo.
- ustvarjeno besedilo se uporabi kot del ukaznega poziva

Liu, J., Liu, A., Lu, X., Welleck, S., West, P., Le Bras, R., Choi, Y. and Hajishirzi, H., 2022, May. Generated Knowledge Prompting for Commonsense Reasoning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 3154-3169).

Pozivanje z generiranjem znanja



Primer

- Prompt:

Part of golf is trying to get a higher point total than others. Yes or No?

- Output:

Yes.

- Primer izboljšamo z ustvarjanjem znanja.
- Najprej ustvarimo nekaj "znanja".

Primer: poziv za generiranje znanja

- Prompt:

Input: Greece is larger than Mexico.

Knowledge: Greece is approximately 131,957 sq km, while Mexico is approximately 1,964,375 sq km, making Mexico 1,389% larger than Greece.

Input: Glasses always fog up.

Knowledge: Condensation occurs on eyeglass lenses when water vapor from your sweat, breath, and ambient humidity lands on a cold surface, cools, and then changes into tiny drops of liquid, forming a film that you see as fog. Your lenses will be relatively cool compared to your breath, especially when the outside air is cold.

Input: A fish is capable of thinking.

Knowledge: Fish are more intelligent than they appear. In many areas, such as memory, their cognitive powers match or exceed those of 'higher' vertebrates including non-human primates. Fish's long-term memories help them keep track of complex social relationships.

Input: A common effect of smoking lots of cigarettes in one's lifetime is a higher than normal chance of getting lung cancer.

Knowledge: Those who consistently averaged less than one cigarette per day over their lifetime had nine times the risk of dying from lung cancer than never smokers. Among people who smoked between one and 10 cigarettes per day, the risk of dying from lung cancer was nearly 12 times higher than that of never smokers.

Input: A rock is the same size as a pebble.

Knowledge: A pebble is a clast of rock with a particle size of 4 to 64 millimetres based on the Udden-Wentworth scale of sedimentology. Pebbles are generally considered larger than granules (2 to 4 millimetres diameter) and smaller than cobbles (64 to 256 millimetres diameter).

Input: Part of golf is trying to get a higher point total than others.

Knowledge:

Primer: generiranje znanja

- Znanje 1:

The objective of golf is to play a set of holes in the least number of strokes. A round of golf typically consists of 18 holes. Each hole is played once in the round on a standard golf course. Each stroke is counted as one point, and the total number of strokes is used to determine the winner of the game.

- Znanje 2:

Golf is a precision club-and-ball sport in which competing players (or golfers) use many types of clubs to hit balls into a series of holes on a course using the fewest number of strokes. The goal is to complete the course with the lowest score, which is calculated by adding up the total number of strokes taken on each hole. The player with the lowest score wins the game.

Primer: integrirano znanje in generiranje

- Prompt:

Question: Part of golf is trying to get a higher point total than others. Yes or No?

Knowledge: The objective of golf is to play a set of holes in the least number of strokes. A round of golf typically consists of 18 holes. Each hole is played once in the round on a standard golf course. Each stroke is counted as one point, and the total number of strokes is used to determine the winner of the game.

Explain and Answer:

- Answer 1 (confidence very high):

No, the objective of golf is not to get a higher point total than others. Rather, the objective is to play a set of holes in the least number of strokes. The total number of strokes is used to determine the winner of the game, not the total number of points.

Question: Part of golf is trying to get a higher point total than others. Yes or No?

Knowledge: Golf is a precision club-and-ball sport in which competing players (or golfers) use many types of clubs to hit balls into a series of holes on a course using the fewest number of strokes. The goal is to complete the course with the lowest score, which is calculated by adding up the total number of strokes taken on each hole. The player with the lowest score wins the game.

Explain and Answer:

- Answer 2 (confidence is a lot lower):

Yes, part of golf is trying to get a higher point total than others. Each player tries to complete the course with the lowest score, which is calculated by adding up the total number of strokes taken on each hole. The player with the lowest score wins the game

- A few more details to consider before arriving at the final answer

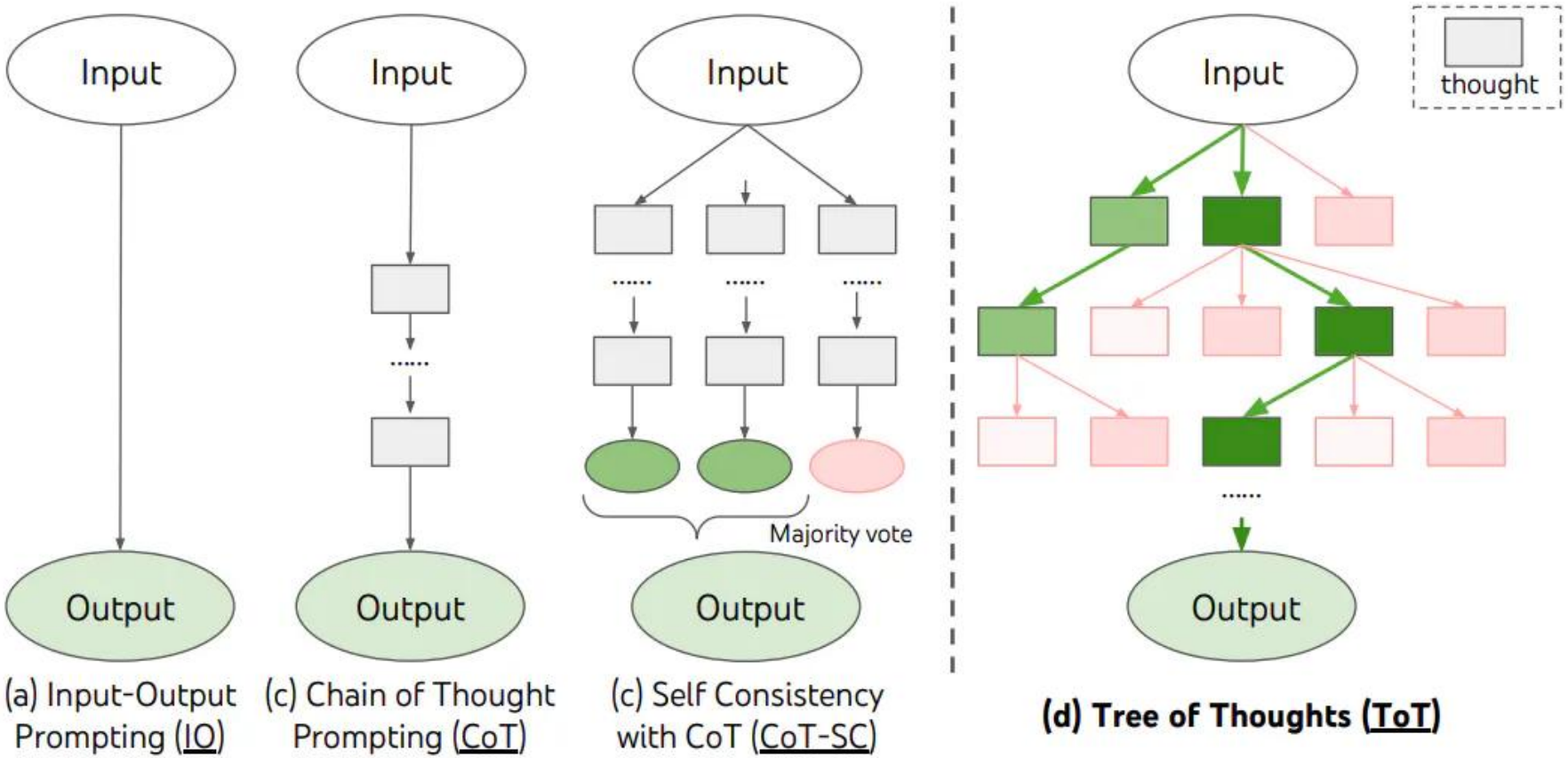
Veriženje pozivov

- Za izboljšanje zanesljivosti in zmogljivosti LLM-jev pri kompleksnih nalogah razdelimo nalogo na podnaloge.
- identificirajmo podnaloge, pozivamo VJM s podnalogo in uporabimo odgovor kot vhod za drugo podnalogo
- veriga pozivov ustvari verigo operacij
- lahko poveča preglednost aplikacije, poveča nadzor in zanesljivost.
- lažje odpravljanje napak z vmesnimi odzivi modela
- lažje je analizirati in izboljšati uspešnost v različnih fazah, ki jih je treba izboljšati.

Drevo misli (Tree of Thoughts, ToT)

- Drevo misli: razbitje problema na vmesne korake ali alternativne korake, ki jih rešuje VJM
- Misli: koherentno besedilo, ki koristi pri reševanju problema
- VJM ovrednoti napredek in misli skozi proces sklepanja
- VJM preišče drevo (v širino, v globino, vračanje nazaj, regeneriranje)
- VJM ovrednoti napredek skozi vmesne korake, ki jih naredimo pri reševanju problema skozi proces sklepanja

Primerjava ToT



Alternativna formulacija ToT

- Ukažimo VJM, da oceni vmesne rezultate

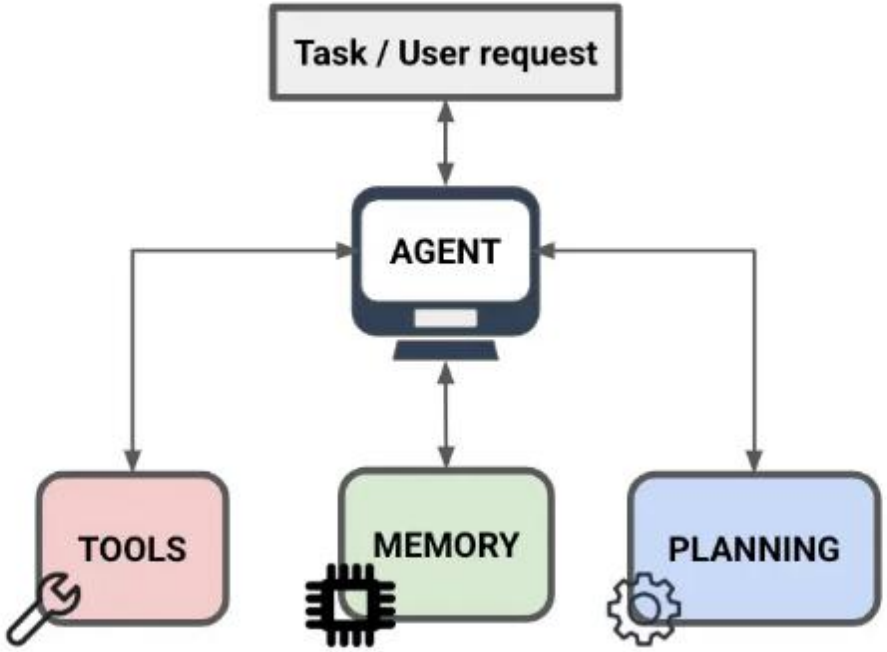
Prompt:
Imagine three different experts are answering this question.
All experts will write down 1 step of their thinking, then share it with the group.
Then all experts will go on to the next step, etc.
If any expert realises they're wrong at any point then they leave.
The question is:
Bob is in the living room.
He walks to the kitchen, carrying a cup.
He puts a ball in the cup and carries the cup to the bedroom.
He turns the cup upside down, then walks to the garden.
He puts the cup down in the garden, then walks to the garage.
Where is the ball?

- Boljši rezultati kot brez navodil za razčlenitev

Dave Hulbert (2023). Using Tree-of-Thought Prompting to boost ChatGPT's reasoning. https://github.com/dave1010/tree-of-thought-prompting

VJM agenti

- Uporaba VJM za različne zahtevne naloge



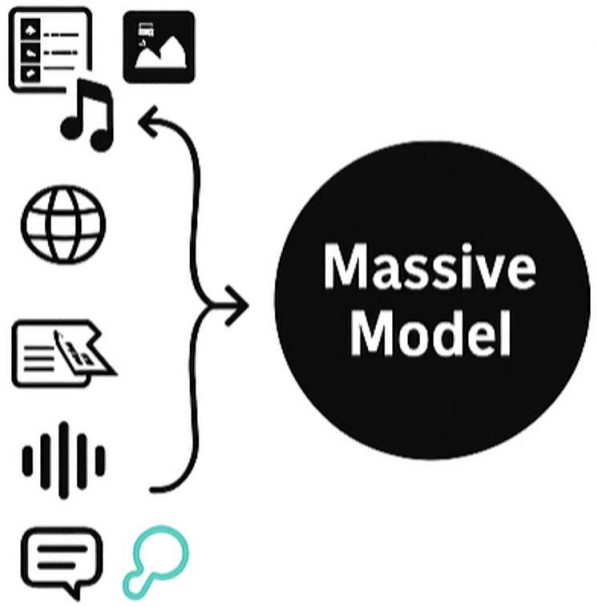
From ML to agents

Traditional ML



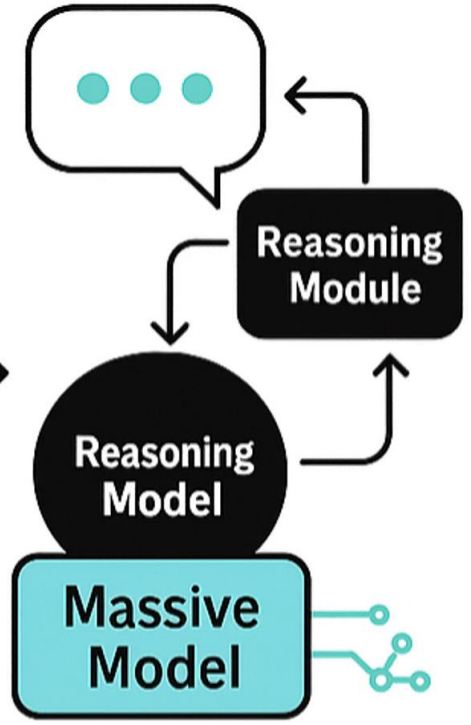
- Individual siloed models
- Require task-specific training
- Lots of human supervised training

Foundation models or LLMs



- Massive multi-modal model
- Adapted with minimal training
- Pre-trained unsupervised learning

Reasoning models



- Deliberative multi-step logic
- Self-consistency checks
- Slower but more accurate inference

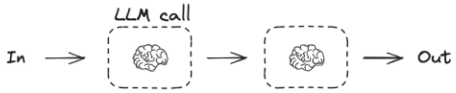
Agent architectures

- **Workflows:** Systems where LLMs and tools are orchestrated through predefined code paths.
- **Agents: S**ystems where LLMs dynamically direct their own processes and tool usage, maintaining control over how they accomplish tasks.

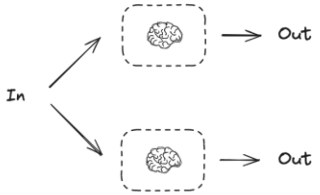
Workflows and agents

Workflows

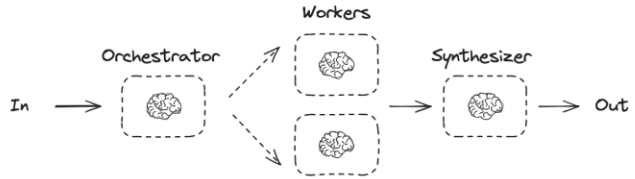
Prompt Chaining



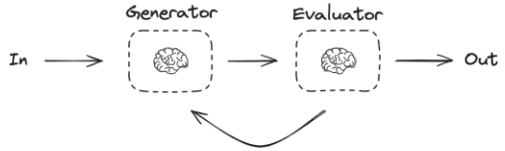
Parallelization



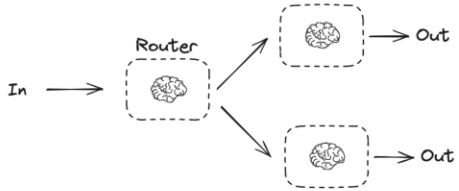
Orchestrator-Worker



Evaluator-optimizer



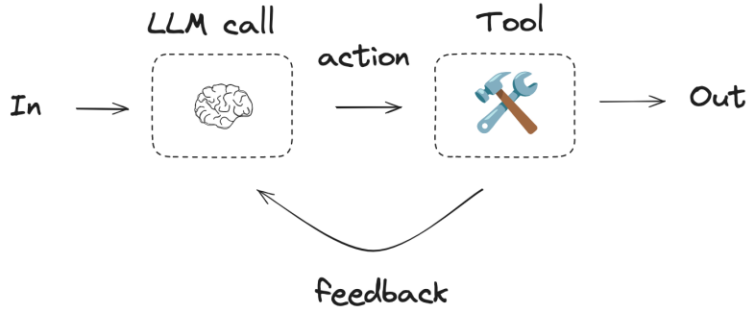
Routing



LLM is embedded in predefined code paths

LLM directs control flow through predefined code paths

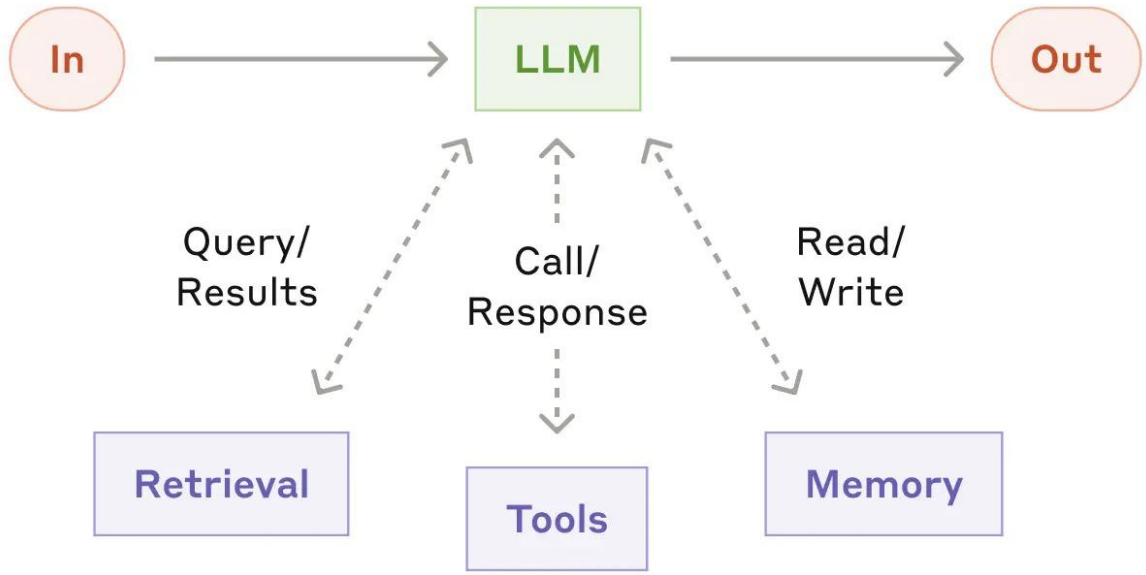
Agent



LLM directs its own actions based on environmental feedback

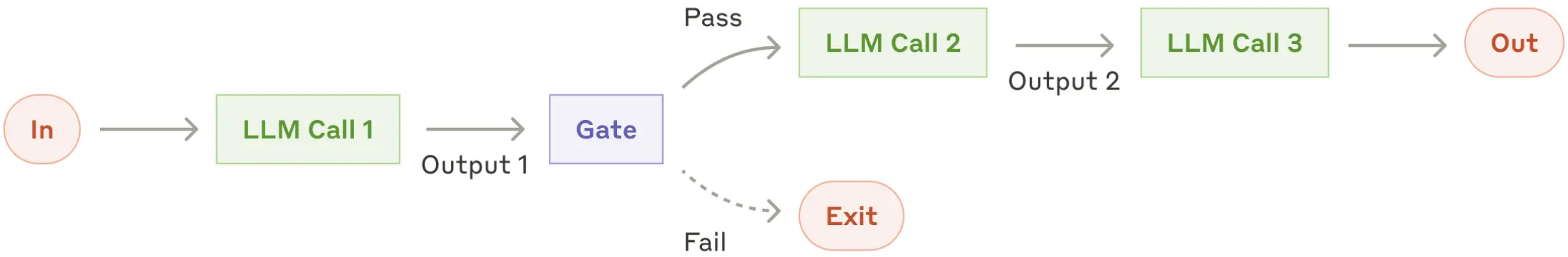
Workflows

- Building block: The augmented LLM



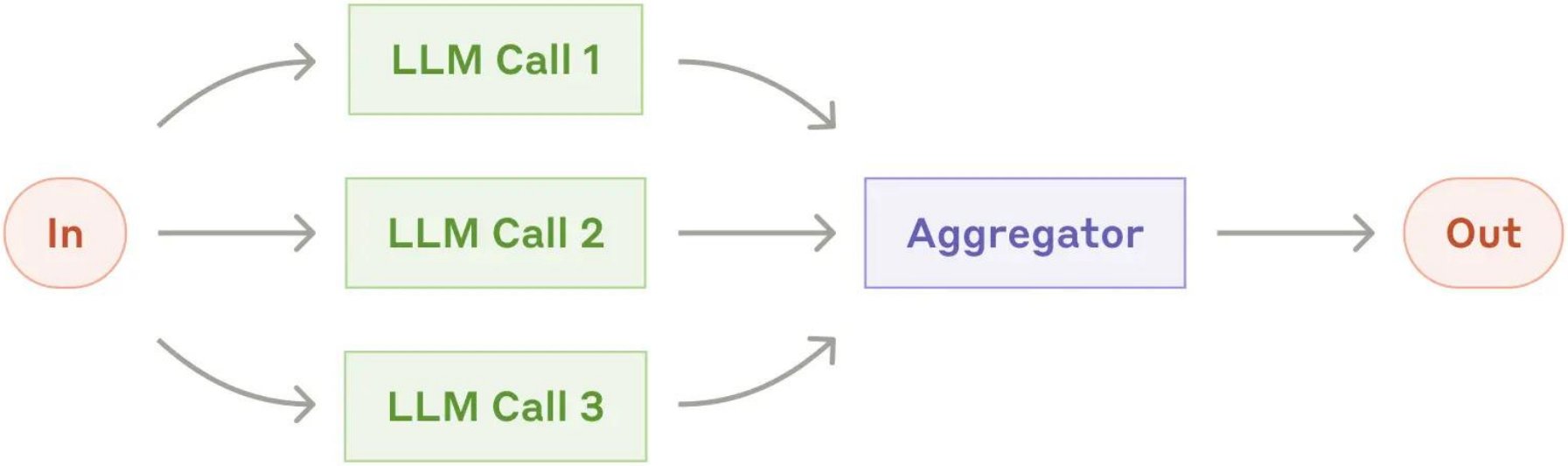
Prompt chaining workflows

- Decompose a task into a sequence of steps, where each LLM call processes the output of the previous one.
- You can add programmatic checks ("gate" in the diagram) on any intermediate steps to ensure that the process is still on track.



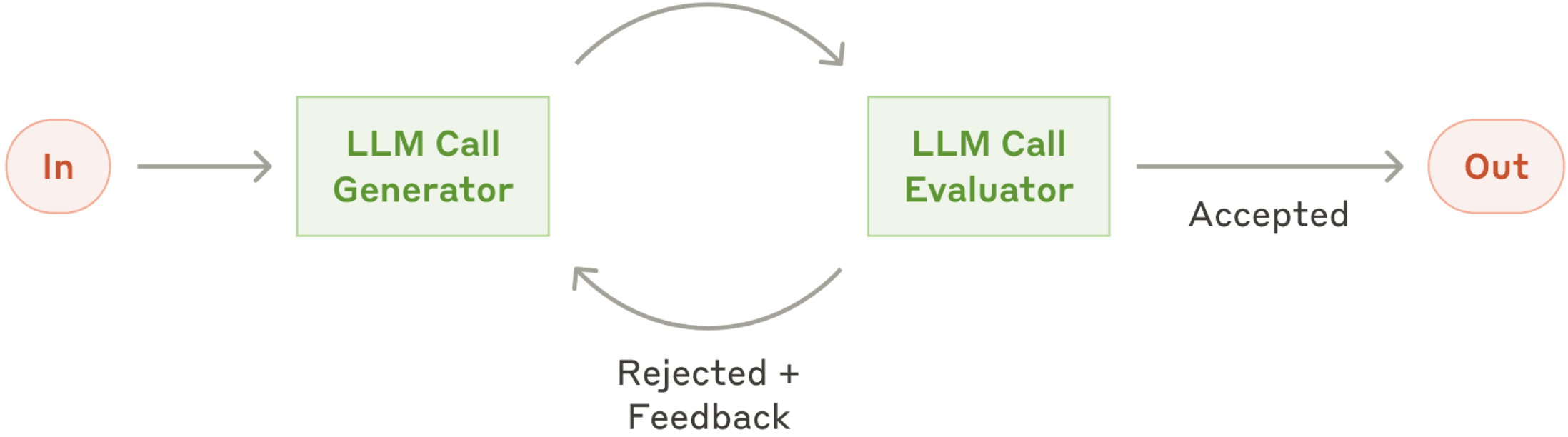
Routing workflows

- Routing classifies an input and directs it to a specialized followup task. Allows for separation of concerns, and building more specialized prompts. Otherwise, optimizing for one kind of input can hurt performance on others.



Routing workflow: Evaluator-optimizer

- One LLM call generates a response while another provides evaluation and feedback in a loop



The agent loop

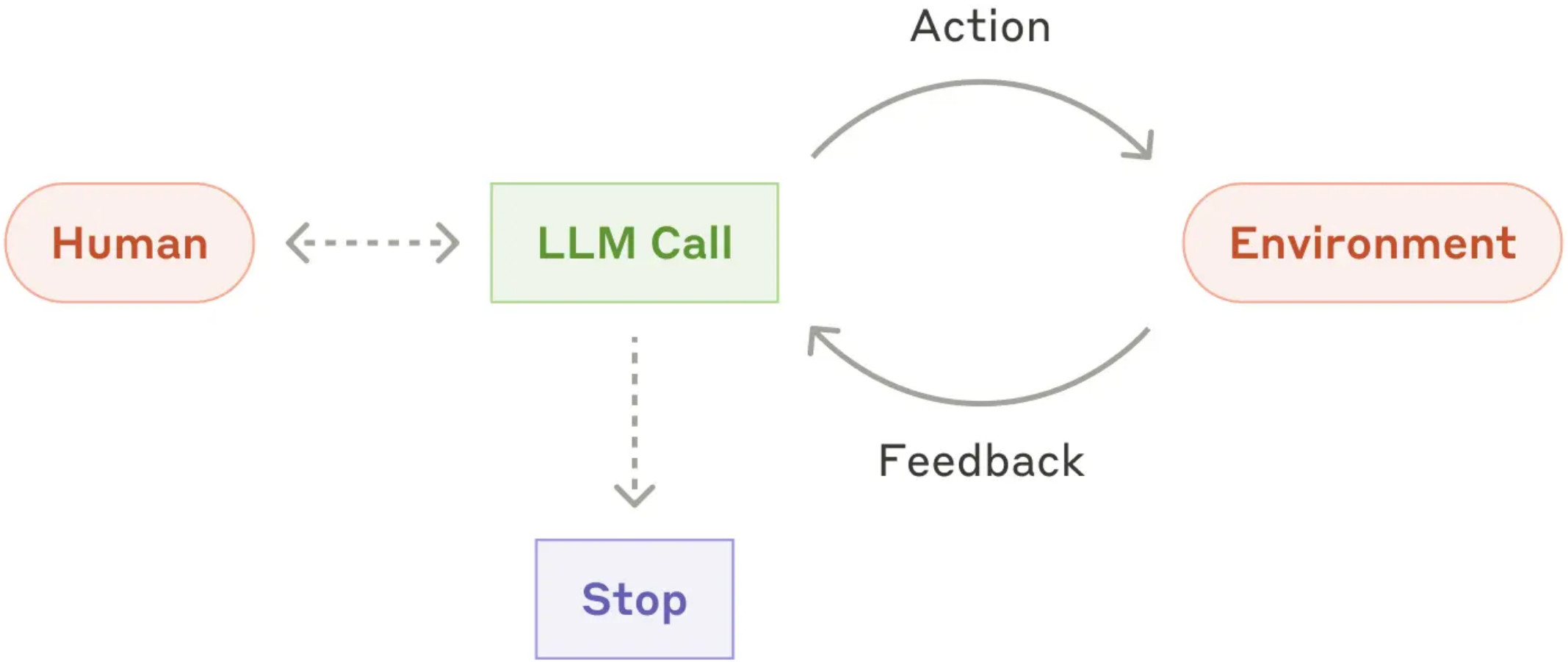
- The “sense-plan-act-learn” loop:
- Initialise state and goals
- Repeat until termination:
  - **Sense**: Gather observations from the environment
  - **Plan**: Evaluate goals and state, plan/select next action
  - **Act**: Execute chosen action on the environment
  - **Learn**: Update internal state, memory, or model based on outcomes

- **In sequence:** call tool, invoke LLM, test for termination; repeat

LLM agents

- Agents begin their work with either a command from, or interactive discussion with, the human user.
- Once the task is clear, agents plan and operate independently, potentially returning to the human for further information or judgement.
- During execution, agents gain “ground truth” from the environment at each step (such as tool call results or code execution) to assess progress.
- Agents can pause for human feedback at checkpoints or when encountering blockers.

LLM agents



Agenti za poslovne probleme

- Veliko kompleksnih poslovnih procesov je mogoče avtomatizirati z uporabo več povezanih VJM-jev, prilagojenih za specifične naloge
- Primer:agent za obdelavo računov v oddelku financ (avtomatizacija finančnega procesa)

Motivacija

- **Velik in neenakomeren obseg naloge**
  Vsak mesec prispe stotine računov v različnih oblikah (PDF, skenirani dokumenti, e-pošta), ki jih je treba terminsko uskladiti z drugimi naloga v računovodstvu
- **Drago ročno preverjanje**
  Ročno preverjanje vsakega računa glede na naročilnico in dobavnico je počasno, neskladja pa pod časovnim pritiskom pogosto spregledamo.
- **Neenotna revizijska sled**
  Kadar kdo kasneje podvomi v odločitev, pogosto ni jasno, kdo je kaj preveril, na podlagi katerih dokazov in zakaj je bil vnos odobren.

Agentni delotok

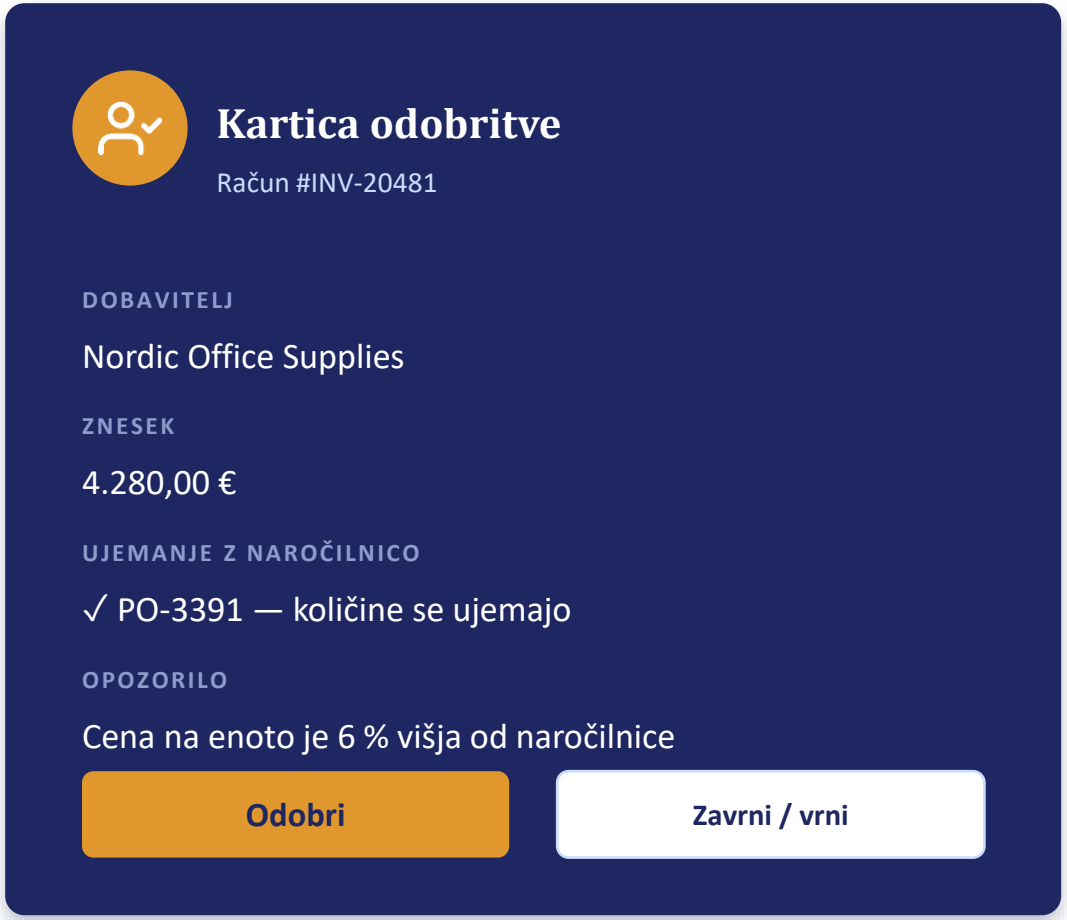
- En račun, šest korakov agenta, ena človeška kontrolna točka



- Koraki 1–4 potekajo samodejno. Pri koraku 5 nadzor prevzame človek, preden se karkoli knjiži v poslovne knjige.

Znotraj kontrolne točke: pogled odobritelja

- Agent nikoli ne knjiži neposredno v poslovne knjige — človeku preda informacije, ki jih potrebuje za odločitev
- Kaj agent prikaže
  - Izvirni račun, naročilnico in dobavnico drug ob drugem
  - Usklajena polja in tista, ki se ne ujemajo
  - Razumljivo pojasnilo za vsako označeno neskladje
  - Zahtevo za potrditev, ne tiho samodejno odobritev
- Ob odobritvi agent vnos knjiži in shrani celotno dokazno sled. Ob zavrnitvi se ne zapiše nič - primer se vrne s pripombo odobritelja



Zakaj pristop deluje

- Isti trije mehanizmi dobro delujejo tudi za druge poslovne procese



Človek v zanki

Agent predlaga; osnutek v verodostojen zapis spremeni šele odobritev osebe.



Izvor in sledljivost

Vsak rezultat vključuje izvorne dokumente in obrazložitev, zato ga je mogoče preveriti, ne le zaupati vanj.



Opredeljene meje odločanja

Kaj agent sme brati, predlagati ali o čemer nikoli ne sme odločati sam, je določeno vnaprej ; agent ne improvizira med delovanjem.

Enak pristop tudi za druge procese

- Prejem → osnutek → človeška kontrolna točka → knjiženje



Kadrovsko predizbiranje

Prijave razvrsti glede na opis delovnega mesta in pripravi ožji izbor, ki ga odobri kadrovnik.



Podpora uporabnikom

Iz baze znanja pripravi osnutke odgovorov na zahtevke; vse, kar zadeva vračila ali račune uporabnikov, pregleda človek.



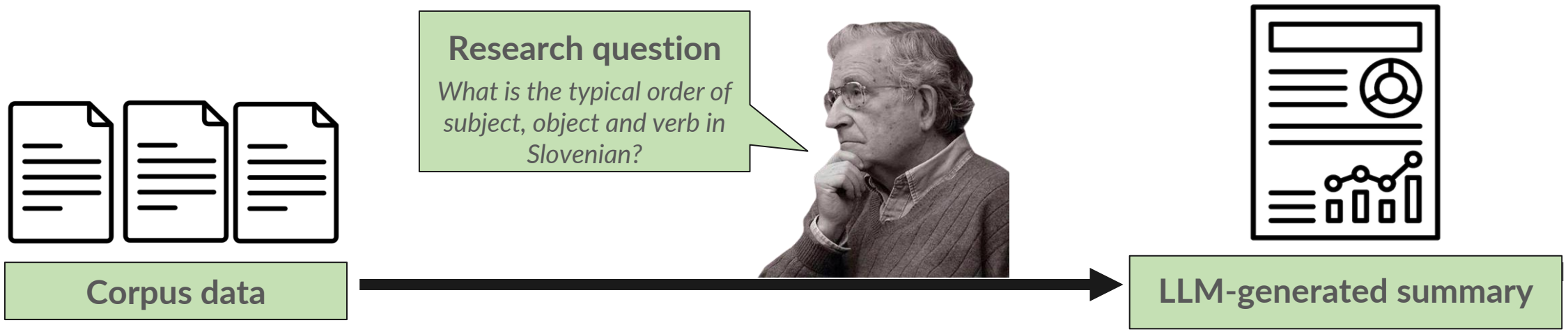
Pregled pogodb

Glede na priročnik označi nestandardne klavzule in izjeme posreduje pravni službi v potrditev.

- Avtomatiziramo ponavljajoči se osrednji del — zajem podatkov, usklajevanje, pripravo osnutkov
- Odobritev, knjiženje in izjeme naj ostanejo v rokah človeka (vsaj dokler ni process dobro upeljan in brez izjem)
- Beležimo dokaze in izvor, da je mogoče vsako odločitev pregledati.

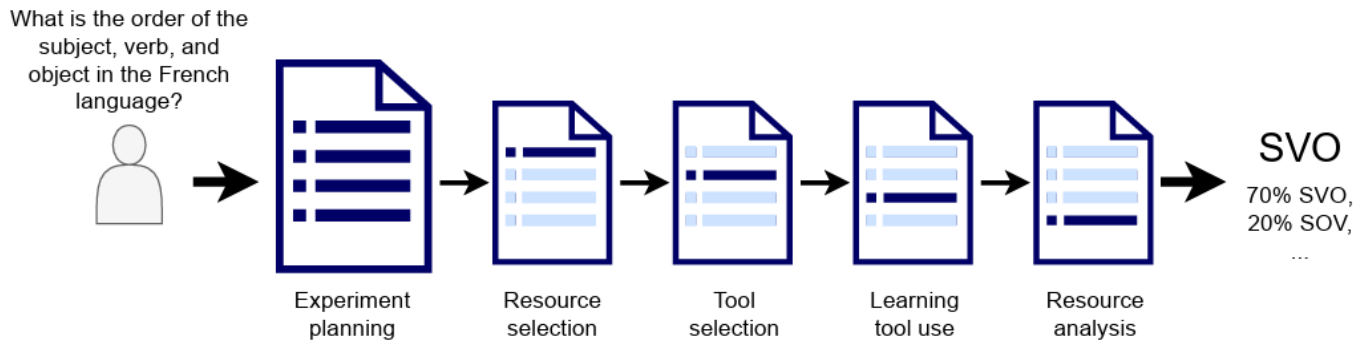
Example: Agents for Linguistics

- Aim
  - Explore LLM-based methods for discovering, summarizing and generating **grammatical knowledge** from multilingual **corpus data**.
- Two guiding questions
  - What do LLMs already know about grammatical structure across languages?
  - Can this knowledge be improved by grounding LLM reasoning in annotated corpus data?



From Implicit Knowledge to Data-Grounded Analysis

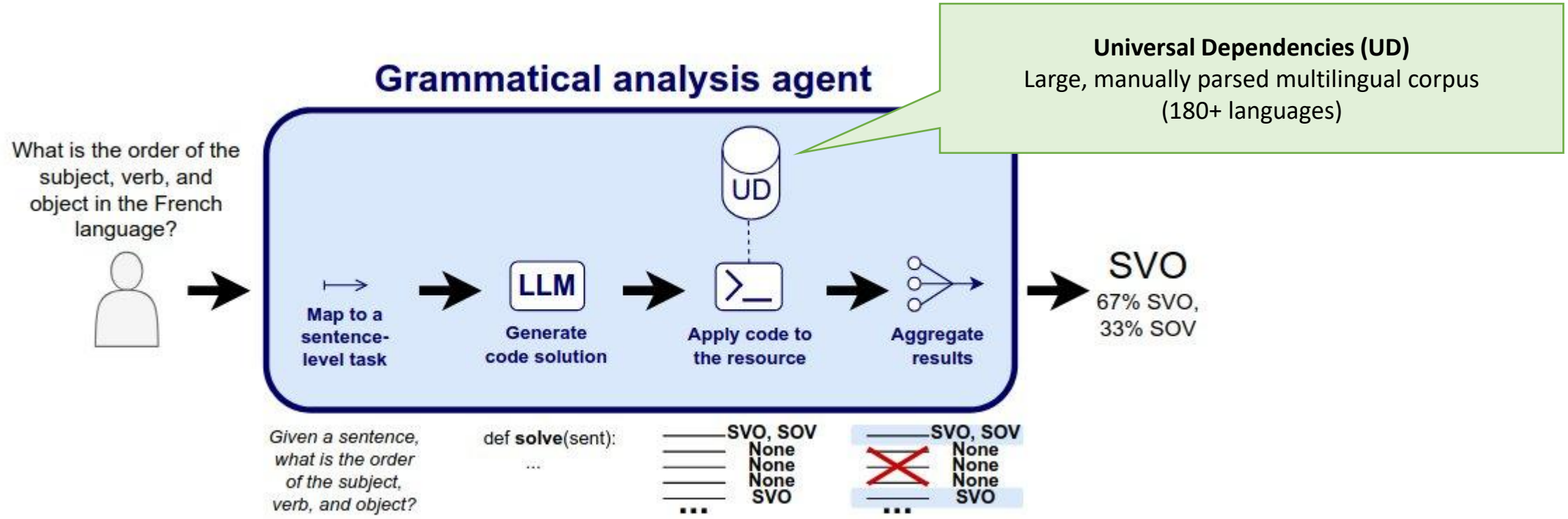
Traditional approach:



Our proposal:

Matej Klemen, Tjaša Arčon, Luka Terčon, Marko Robnik-Šikonja, Kaja Dobrovoljc (2026) *Towards Corpus-Grounded Agentic LLMs for Multilingual Grammatical Analysis*. https://arxiv.org/abs/2512.00214

Agentic LLM System (proof of concept)



Evaluation Results

- **Setup:** 13 WALS-inspired word-order features on UD corpora (170+ languages)
- **Ground truth:** distributions extracted with custom scripts

	Dominant order prediction			Identification of all orders			Order distribution prediction		
	(Accuracy ↑)			(Coverage F_1 ↑)			(Hellinger dist. ↓)		
Feature	Majority	GPT-5	UDagent	Majority	GPT-5	UDagent	Majority	GPT-5	UDagent
81A									
82A									
83A									
84A									
85A									
86A									
87A									
88A									
89A									
90A									
94A									
144A									
144B									

RESULTS

- **Strong gains:** Best results in most tasks
- **High reliability:** Scores mostly > 0.9, distributions closely match ground truth
- **Cross-lingual:** Perfect accuracy in 112/159 languages (excl. 144A/B).
- **Errors:** code generation failures, but also valid linguistic alternatives.

Strojno učenje se spreminja

- Transformerji za tabelarične podatke in časovne vrste:
  - Vsako vhodno značilko obravnavamo kot "žeton"
  - Uporabimo pozornost za samodejno učenje interakcij
-
- Temeljni modeli (foundation models) za tabelarične podatke in časovne vrste
  - Veliki vnaprej naučeni modeli na osnovi transformerjev, ki se učijo splošnih predstavitev strukturiranih podatkov na tisočih učnih množic (sintetičnih in realnih)
  - sledi prilagoditev na druge tabelarične naloge



Zakaj potrebujemo odprt velik jezikovni model za slovenščino?

Mistral-NeMo-Minitron-8B-Instruct:
Srečno, sem rad, da bom vam pomagal pri pripravi recepta za špagete bolognese. Ena od najpopuljarnejših italijanskih jame je relativno enostavna, vendar pa je tudi zelo omembljiva. Ena od mojih omembljivih različic je naslednja:
Špagete bolognese
Čas priprave: 45 minut
Dosežek: 4-6 portions
Oreški:

- 500g gibanice
- 2 stikle majoneze

Source: https://okusno.je/recept/prekmurska-gibanica

Razvoj velikega jezikovnega modela GaMS - pomen

GaMS: “

- GaMS (generativni model slovenščine)
- Pomen velikih jezikovnih modelov (VJM)
  - Osnovna infrastruktura umetne inteligence
  - Z agentnimi sistemi, vključevanjem zunanjih virov, sklepanjem in klicanjem zunanjih servisov postajajo bistveno močnejši
  - Postajajo del splošne znanstvene in poslovne metodologije
  - Osnova za tehnološko suverenost
- Gradnja VJM zahteva
  - ogromno besedil (zbiralna akcija, veliki besedijodajalci) in
  - računskih virov (Vega, Leonardo, Nvidia, FRIDA)



Razvoj velikega jezikovnega modela GaMS - pristop

GaMS: “

- Osnova sta odprtodostopna modela Gemma 2 in Gemma 3
- Gradnja VJM poteka v treh fazah
  1. Nadaljevanje vnaprejšnjega učenja:
     - ogromno slovenskih besedil in računske moči
  2. Učenje sledenja navodilom in odgovarjanja na vprašanja
     - velike slovenske učne množice
     - kulturno prilagajanje
     - klici zunanjih servisov, sklepanje, slike
  3. Prilagajanje človeškim preferencam
     - Značilnosti odgovorov, varnost
- Razvoj omogočitvenih tehnologij
  - strojno prevajanje, optično prepoznavanje dokumentov, čiščenje učnih množic, evalvacija, zbiranje preferenc – arena



Razvoj velikega jezikovnega modela GaMS - stanje

- Eden najuspešnejših odprtodostopnih modelov velikosti okoli 10 milijard parametrov
- Ne le eden, cela čreda GaMSov: 2B, 9B, 12B, 27B
- V naslednjih verzijah še več poudarka na klicih zunanjih servisov, sklepanju, podpori agentom in slikam
- Financiranje še ni zagotovljeno



GaMS 2

PoVeJMo
SLOVENSKI
VELIKI JEZIKOVNI MODEL

GaMS:
Generativni model za slovenščino
Generative Model for Slovene

ODPRTODOSTOPEN

- Vaši podatki ostanejo pri vas

VEČ RAZLIČIC

- 2B
- 9B
- 27B

- Učenje verzije 9B: 3 dni na 32 GPU vozliščih (128 GPU) na Leonardo Booster HPC
- Sledenje navodilom: 25.000 primerov, prilagoditev parametrov, 1 dan, 1 GPU vozlišče

DODATNO UČEN NA SLOVENSKIH BESEDILIH

- Osnova: Gemma 2
- 8 milijard slovenskih “žetonov” (6 milijard besed)

NAJBOLJŠI ODPRT MODEL ZA SLOVENŠČINO

- SloBench
- GaMS 3 - še boljši

Od kod besedila v slovenščini?

PoVeJMo
SLOVENSKI
VELIKI JEZIKOVNI MODEL

Obstoječa besedila

- MetaFida
  - 4,66 mrd (34,18 %)
- Korpus akademske slovenščine
  - 2,77 mrd (20,34 %)
- Wikipedia
  - Wikipedia-Sl
    - 0,16 mrd (1,19 %)
  - Wikipedia-En
    - 5,45 mrd (39,99 %)
  - Wikipedia-Sr-Latin
    - 0,36 mrd (2,68 %)
  - Wikipedia-Hr
    - 0,15 mrd (1,13 %)
  - Wikipedia-Bs
    - 0,07 mrd (0,50 %)

Nova besedila

- Narodna in univerzitetna knjižnica
- Radiotelevizija Slovenija
- Slovenska tiskovna agencija
- Arhiv RS
- Druge javne ustanove
- Založbe
- Mediji

Zbiranje gradiv v slovenščini

PoVeJMo
SLOVENSKI
VELIKI JEZIKOVNI MODEL

- Portal: zbiranje.povejmo.si
- “Dogovor o uporabi avtorskih del v zvezi z zbiranjem besedil”
- Direktiva 2019/790 o avtorski in sorodnih pravicah na enotnem digitalnem trgu
- Zakon o avtorski in sorodnih pravicah (ZASP)
- Akt o umetni inteligenci



GAMS 2 :” 9B

UL | FRI

- temelji na Gemma 2 9B.
- nadaljno učenje na 15 milijardah besed (9 milijard slovenskih)
- Učenje: 3 dni na 32 GPU vozliščih (128 GPU) na Leonardo Booster superračunalniku
- sledenje navodilom: 25.000 primerov, prilagoditev vseh parametrov, 1 dan, 1 GPU vozlišče
- najuspešnejši VJM za slovenščino
- evalvacija:
  - SloBENCH : SuperGLUE, strojno prevajanje
  - SlovenianEval

Rank	Title	Authors and Affiliations	Average
1	GaMS-27B	Domen Vreš, FRI	0.7601
2	PrešernGPT 0.1	Žan Knafelc	0.7568
3	Gemma 2 27B	Google, Domen ...	0.7546
4	GaMS-9B	Domen Vreš, FRI	0.7309
5	GaMS-9B-Instru...	Domen Vreš, FRI	0.6997
6	Gemma 2 9B	Google, Domen ...	0.6980
7	SlovenianGPT-C...	Domen Vreš, FRI	0.6436
8	CroSloEngual BE...	Aleš Žagar, FRI ...	0.6078
9	GaMS-1B-Chat-f...	Domen Vreš, FRI	0.5806
10	OPT_GaMS-1B-...	Domen Vreš, FRI	0.5645

https://povejmo.si

Gemma 2 -> GaMS :”

1. faza: poravnava jezikov
   → učenje na poravnanih angleško-slovenskih prevodih
2. faz:. kakovostna slovenska besedila
   - Wikipedia + MetaFida + KAS
   - brez spletnih besedil in tvitov
   - novice urejene časovno naraščajoče
3. faza: učenje sledenja navodilom

GaMS: “ 3

UL FRI

Tehnične podrobnosti:

- Osnovan na Googlovi Gemmi 3
- 12 milijard parametrov
- Učen na več kot 100 miljardah besed
- Približno 30-40 milijard slovenskih besed
- Učen na 3 različnih superračunalnikih:
  - LEONARDO - 150 tisoč GPU ur
  - FRIDA - tisoč GPU ur
  - NVIDIA DGX Lepton - 40 tisoč GPU ur

Nadgradnje prejšnje generacije:

- Kontekstno okno 131 tisoč žetonov
  - Omogoča procesiranje daljših dokumentov
- Boljši rezultati na klasifikacijskih testih
- Potencial za podporo več modalnosti
  - Gemma 3 omogoča procesiranje slik

Nadaljnji razvoj:

- RAG po slovenski wikipediji
- Klicanje funkcij/orodij → agentni način
- Sklepanje/razmišljanje

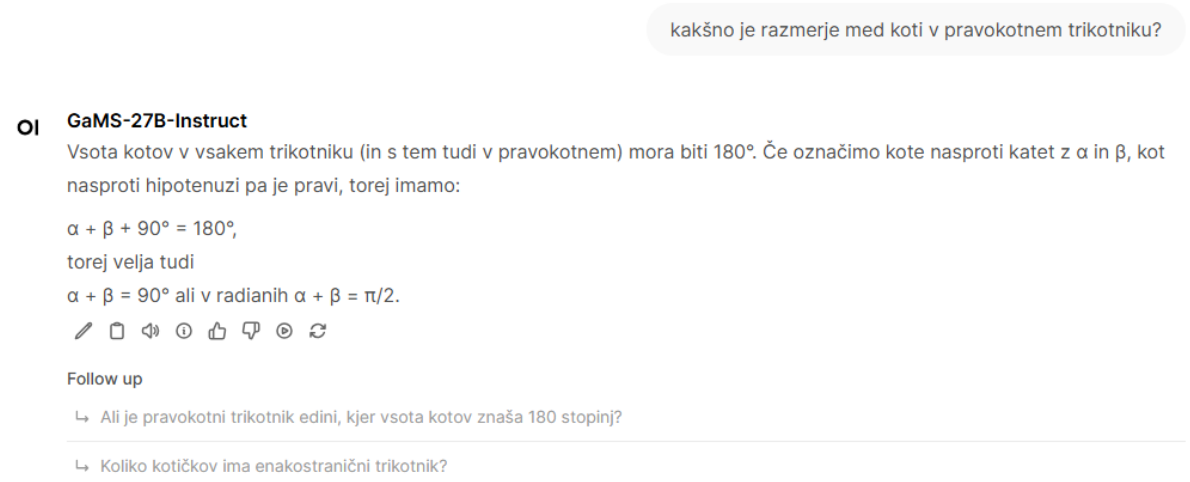
Financira Evropska unija NextGenerationEU
NAČRT ZA OKREVANJE IN ODPORNOST
ARIS Javna agencija za znanstvenoraziskovalno in inovacijsko dejavnost Republike Slovenije

Pogovor z GaMSom

https://povejmo.si/klepet/



http://llm.ijs.si/

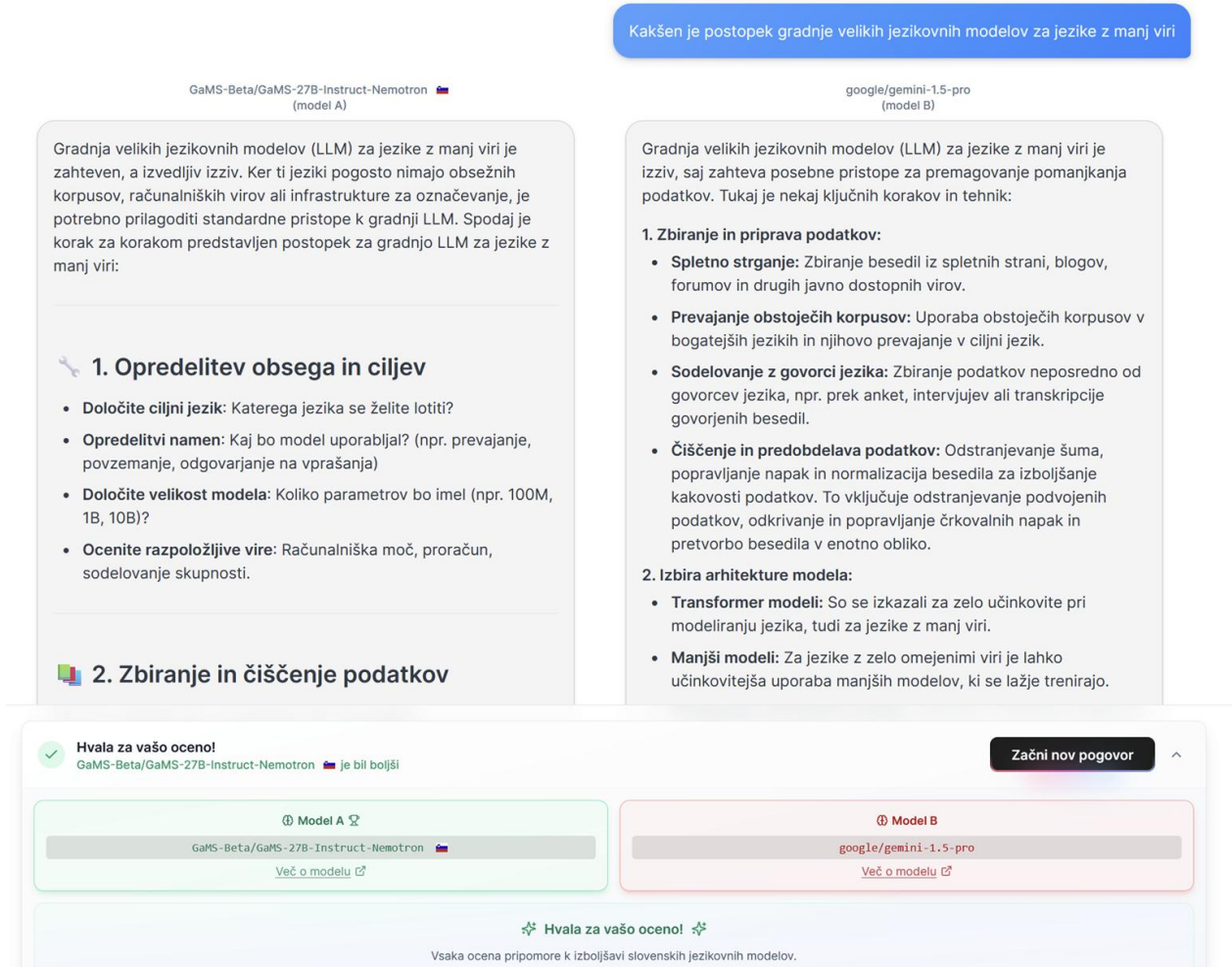


LLM Arena - zbiranje podatkov o uporabnosti modelov

PoVeJMo
SLOVENSKI
VELIKI JEZIKOVNI MODEL

- Na vprašanje uporabnika istočasno odgovorita dva modela
- Uporabnik izbere boljši odgovor
- Izdelana slovenska LLM arena, kjer odgovarjajo najuspešnejši modeli za sloveščino
- Poteka zbiranje in analiza uporabniških odgovorov

- Prvi izsledki: GaMS le malo zaostaja za 1000x večjim Gemini 2.5 Pro in je boljši od ostalih modelov
- https://arena.cjvt.si/sl



Evalvacija: SloBench - SuperGLUE

PoVeJMo
SLOVENSKI
VELIKI JEZIKOVNI MODEL

Leaderboard

Slovene SuperGLUE 1.0

Description

This leaderboard is a Slovene (translated) version of an existing English version. The task initially consists of multiple tasks, which test a system's ability to understand and reason about texts in Slovene. All the tasks should be solvable by most college-educated Slovene speakers.

The original SuperGLUE benchmark consists of 8 tasks, while 6 tasks are included in the Slovene evaluation. These tasks are BoolQ, CB, COPA, MultiRC, RTE and WSC.

Leaderboard Editors

Aleš Žagar, ales.zagar@fri.uni-lj.si
Slavko Žitnik, slavko@zitnik.si

Related Leaderboards

Slovene SuperGLUE (v1)

https://slobench.cjvt.si/

Leaderboard Submissions

Rank	Title	Authors and Affiliations	Average	BoolQ Accuracy	CB Accuracy	CB F1 Score	CB Average	COPA Accuracy	MultiRC EM	MultiRC F1a Score	MultiRC Average	RTE Accuracy	WSC Accuracy	Tag
1	GaMS-27B	Domen Vreš, FRI	0.7601	0.8333	0.6440	0.5864	0.6152	0.9540	0.3904	0.7504	0.5704	0.7931	0.7945	
2	PrešernGPT 0.1	Žan Knafelc	0.7568	0.8333	0.8520	0.5868	0.7194	0.9740	0.4985	0.8061	0.6523	0.8276	0.5342	
3	Gemma 2 27B	Google, Domen ...	0.7546	0.8333	0.6680	0.5972	0.6326	0.9140	0.4174	0.7295	0.5735	0.8276	0.7466	
4	GaMS-9B	Domen Vreš, FRI	0.7309	0.7000	0.8400	0.7955	0.8178	0.9000	0.3243	0.6551	0.4897	0.7931	0.6849	
5	GaMS-27B-Instr...	Domen Vreš, FRI	0.7038	0.8333	0.7200	0.6322	0.6761	0.9400	0.0511	0.5813	0.3162	0.7931	0.6644	
6	GaMS-9B-Instru...	Domen Vreš, FRI	0.6997	0.8000	0.7960	0.7128	0.7544	0.8140	0.0721	0.6174	0.3447	0.7931	0.6918	
7	Gemma 2 9B	Google, Domen ...	0.6980	0.8333	0.8240	0.5683	0.6962	0.8700	0.2282	0.5310	0.3796	0.7241	0.6849	

GaMS3: najboljši na INCLUDE-base-44 benchmarku

- 20 evropskih jezikov

Rank	Model	Parameters	Average Score	Polish	Albanian	Belarusian	Bulgarian	Croatian	Dutch	Estonian	Finnish	French	German	Greek	Hungarian	Italian	Lithuanian	Portuguese	Russian	Serbian	Spanish	Turkish	Ukrainian
1	GaMS3-12B-Instruct,bos	12	0.6579	0.6259	0.7768	0.3400	0.7782	0.7582	0.7350	0.7098	0.5554	0.6444	0.5683	0.6033	0.5618	0.7664	0.7416	0.6570	0.6286	0.7018	0.6818	0.6332	0.6909
2	Qwen3-14B	14	0.6507	0.5967	0.7550	0.3582	0.7273	0.7691	0.7241	0.5402	0.5445	0.7017	0.5971	0.6069	0.5673	0.7828	0.6479	0.6878	0.6268	0.7382	0.7236	0.6259	0.6927
3	Bielik-11B-v3.0-Instruct	11	0.6492	0.6880	0.7949	0.3545	0.7327	0.7545	0.7477	0.7232	0.5499	0.6396	0.5108	0.5870	0.5545	0.7372	0.7341	0.5935	0.5815	0.7291	0.6873	0.6150	0.6691
4	cogito-v1-preview-qwen-14B	14	0.6436	0.6277	0.6733	0.3455	0.7018	0.7527	0.7550	0.5714	0.5318	0.7017	0.5899	0.5471	0.5127	0.7901	0.6273	0.7187	0.6504	0.7636	0.7364	0.6022	0.6727
5	gemma-3-12b-it,bos	12	0.6393	0.6296	0.7677	0.3509	0.7291	0.7564	0.7132	0.6652	0.5336	0.6539	0.5396	0.5598	0.5345	0.7500	0.7172	0.6370	0.5906	0.7218	0.6800	0.5949	0.6618
6	Qwen2.5-14B-Instruct	14	0.6166	0.5894	0.6552	0.3491	0.6945	0.7218	0.7423	0.5312	0.4809	0.6945	0.5755	0.5525	0.4855	0.7646	0.5487	0.6515	0.6178	0.6964	0.7364	0.5912	0.6527
7	Qwen2.5-14B	14	0.6109	0.6131	0.6552	0.3182	0.6891	0.6945	0.7368	0.5268	0.5045	0.7160	0.5540	0.4511	0.4564	0.7701	0.5730	0.6661	0.6467	0.6909	0.7418	0.5821	0.6309
8	Qwen3-8B	8	0.6105	0.5949	0.6461	0.3491	0.6836	0.7036	0.7024	0.5179	0.4737	0.6587	0.5899	0.5942	0.5182	0.7646	0.5936	0.6388	0.5924	0.6582	0.6927	0.5748	0.6618
9	phi-4	15	0.5882	0.4964	0.6770	0.2873	0.6655	0.7509	0.7205	0.4375	0.4501	0.6874	0.5468	0.4257	0.4818	0.7628	0.5094	0.6588	0.5815	0.6327	0.7273	0.6058	0.6582
10	Apertus-8B-Instruct-2509	8	0.5789	0.4964	0.7078	0.2964	0.6582	0.7073	0.6570	0.6250	0.4719	0.5919	0.5036	0.5145	0.4745	0.6697	0.6030	0.5445	0.5344	0.6673	0.6636	0.5766	0.6145
11	Llama-3.1-8B-Instruct	8	0.5534	0.5383	0.6134	0.3073	0.6491	0.6582	0.6116	0.4777	0.4356	0.5823	0.4892	0.4402	0.5145	0.7007	0.4925	0.5844	0.5163	0.6000	0.6509	0.6423	0.5636
12	EuroLLM-9B-Instruct	9	0.5510	0.5201	0.4446	0.2891	0.6618	0.6873	0.6007	0.6562	0.4319	0.5943	0.4676	0.4873	0.5164	0.6478	0.6105	0.5209	0.5399	0.5782	0.6509	0.5493	0.5655

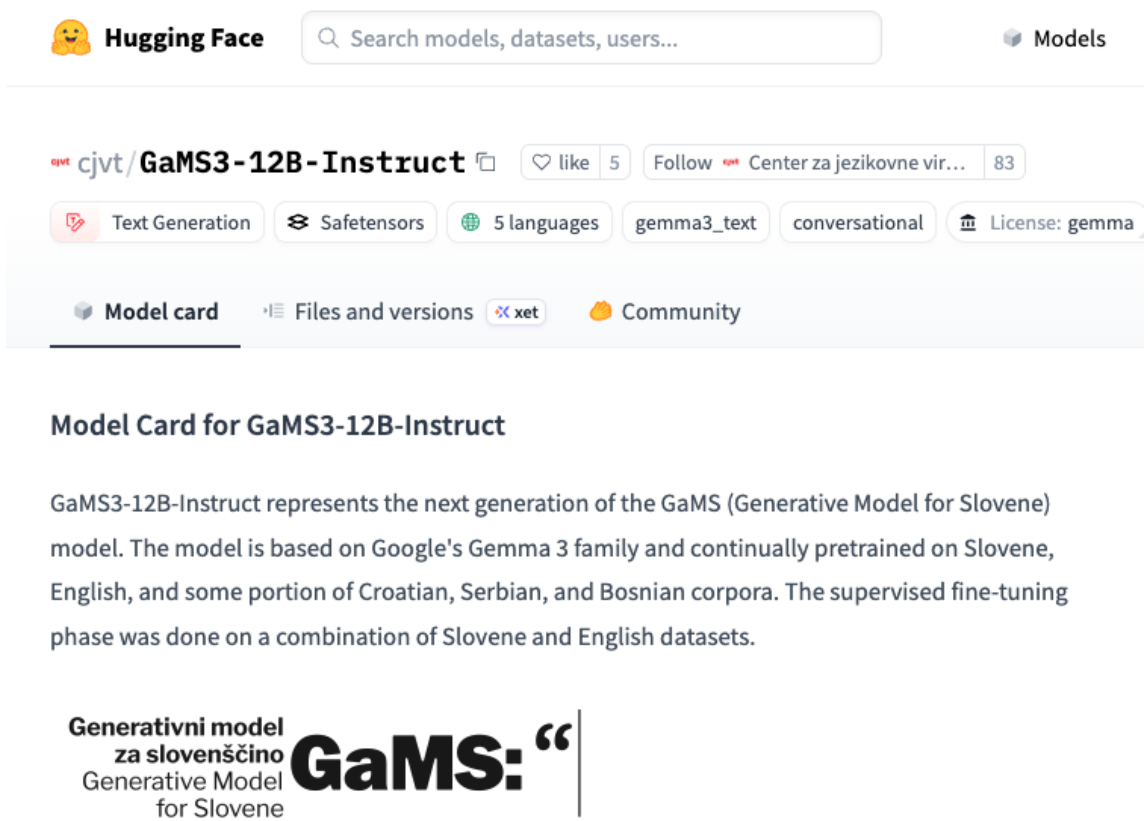
UL | FRI

Preizkusimo GaMS-a ☺

Klepetalnik

Pogovorna arena

Prenos GaMS-a v odprtem dostopu



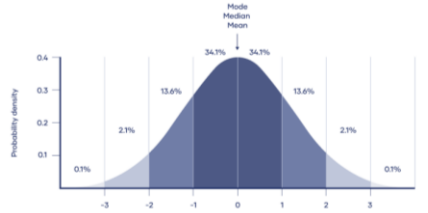
gams.povejmo.si/odprtidostop

Vstavljanja dodatnega znanja v velike generativne jezikovne modele

PoVeJMo
SLOVENSKI
VELIKI JEZIKOVNI MODEL

- Pripravljena slovenska učna množica za zdravorazumsko sklepanje KE-WSC
- Pripravljena slovenska učna množica za sklepanje v naravnem jeziku SI-NLI
- Pripravljene učne množice za logično sklepanje in matematiko
- Pripravljena varnostna učna množica z etičnimi normami
- Razvoj metod za prenos jezikoslovnega in leksikografskega znanja v VJM
- Vključeno v naslednjo verzijo modelov GaMS
- Nove aplikacije na podlagi slovenskih modelov GaMS, npr. jezikovni popravljalniki, prevajalniki, povzemanje in poenostavljanje besedil
- GaMS 3.0 je že na voljo, a se še izpopolnjuje

Statistical evaluation of LLMs' results



- LLMs can answer any question, even questions about complex phenomena

You are a senior data analyst. You have access to a SQL database containing 3 years of customer transaction data across 5 regions. Your task is to determine whether our new loyalty program — launched 6 months ago — has statistically significantly increased average order value (AOV) and purchase frequency.
Perform the following steps iteratively:
1. Query the database to split customers into treatment (enrolled in loyalty program) and control (not enrolled) cohorts, stratified by region and customer tenure.
2. For each cohort, pull AOV and purchase frequency for the 6 months pre- and post-launch.
…

- We have no guarantee that the answers are correct
- LLMs do not abstain; we almost always get an answer
- To statistically analyze phenomena, the prompts (or fine-tuned LLMs) are typically applied to many texts such as news, articles, folktales, etc.
- To trust the results and use them (in science), LLMs' results have to be verified first

Verification of LLMs outputs

UL | FRI

- Using LLMs for complex questions:
  - Ask LLM
  - Verify the answers:
  - If the accuracy is high, use LLM; otherwise, this is senseless, revise
- Three verification paths:
  1. Comparison with human gold-standard answers
  2. Human verification of LLM's answers
  3. LLM-as-a-judge



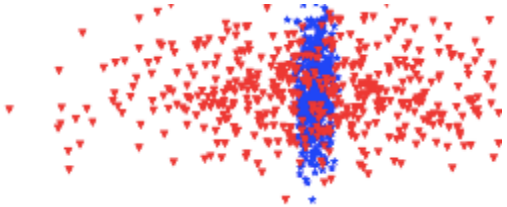
Verification 1: Comparison with human gold-standard answers

- It might be time-consuming or expensive to get the gold-standard answers
- For classification, use metrics such as accuracy, F_1, precision, recall, etc.
- In contrast to classification, the comparison is more difficult for generated texts, but methods exist
  - n-gram based matching (BLEU, ROUGE, etc),
  - sentence encoders and cosine distance
  - LLM as a judge (stronger LLM, has to be verified first, potentially circular)

Verification 2:
Human verification of LLM's answers

- Humans check the answers of LLMs (correctness, style, format, consistency)
- Note: Checking the answers is a weaker form of verification than producing the same or similar answers (the priming effect).

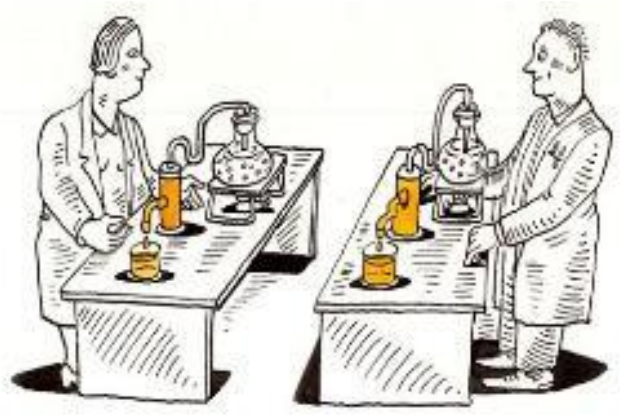
Reliability



UL | FRI

- **Variance** in answers (humans' or LLMs') is always present
- Humans: different background, education, political orientation, sense of humour, mood, etc.
- LLMs: different prompts, different LLMs, temperature parameters, etc.
- If possible, compute
  - Inter-annotator agreement
  - Fleiss' kappa,
    Landis and Koch scale: $\kappa > 0.80$ (almost perfect), $0.60 < \kappa \leq 0.80$ (substantial), etc.
  - Annotator's self-agreement
- To improve reliability: repeat and average, show variance

Reproducibility



UL | FRI

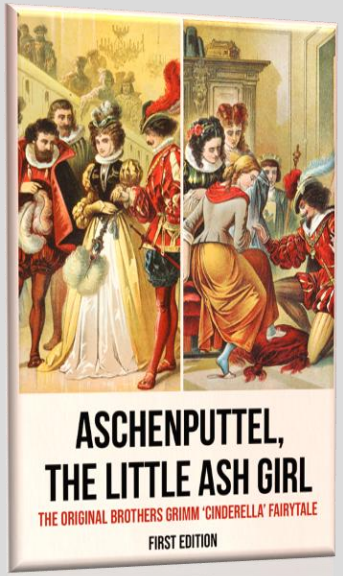
- **Reproducibility** concerns:
  - The research in LLMs is very fast
  - Commercial LLMs have a life span of around a year
  - However, there are many (less capable) open-source models

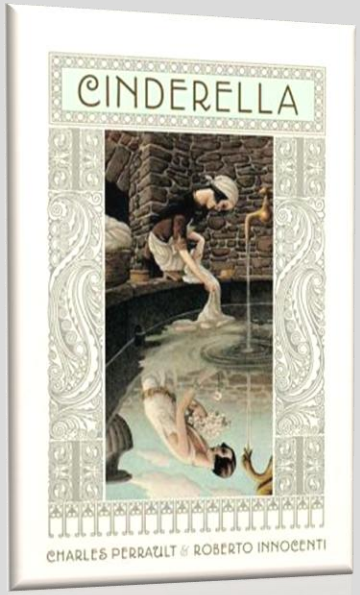
Use case: Complex questions in folkloristics

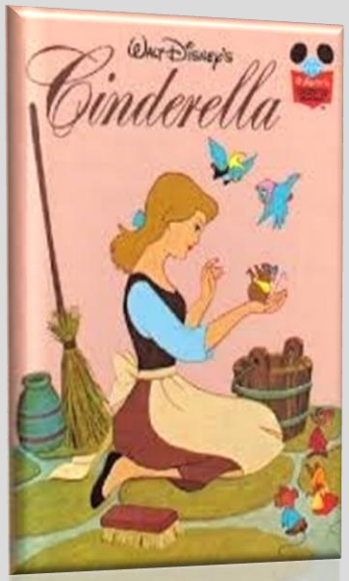
- An example of evaluated LLM for motif detection in folkloristics



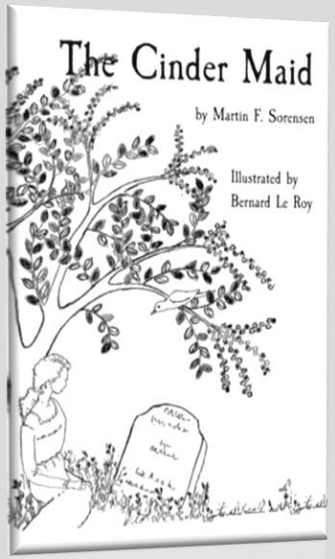
Tjaša Arčon, Marko Robnik-Šikonja, Polona Tratnik (2025): Motif detection using large language models: The Cinderella case study. https://doi.org/10.48550/arXiv.2510.18561

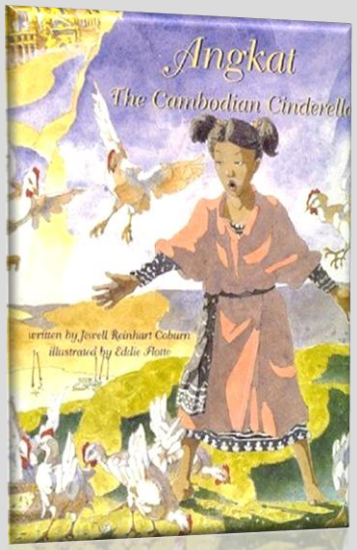

ASCHENPUTTEL,
THE LITTLE ASH GIRL
THE ORIGINAL BROTHERS GRIMM 'CINDERELLA' FAIRYTALE
FIRST EDITION

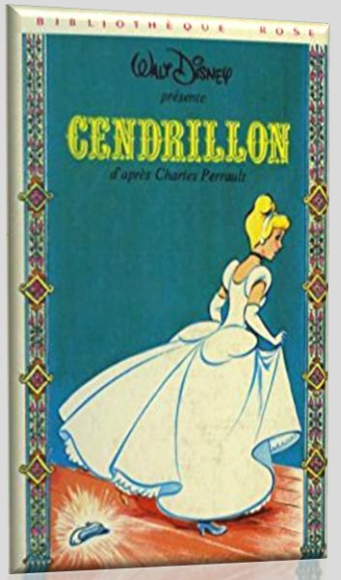

CINDERELLA
CHARLES PERRAULT & ROBERTO INNOCENTI

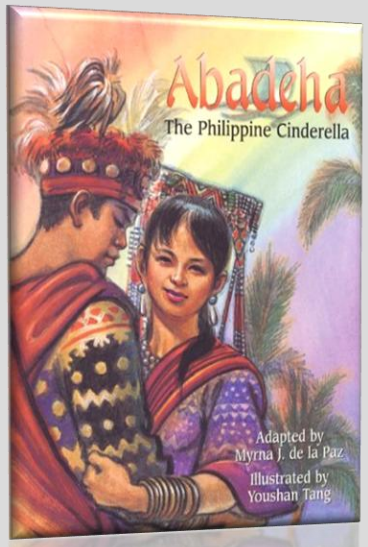

Walt Disney's
Cinderella

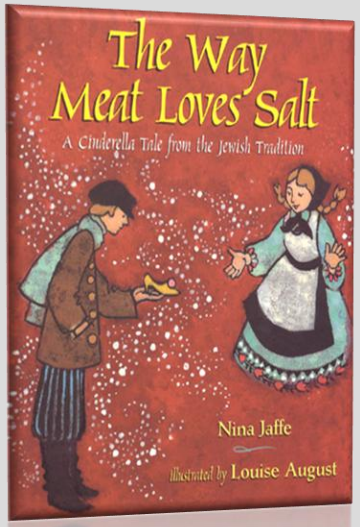

PEPELKA

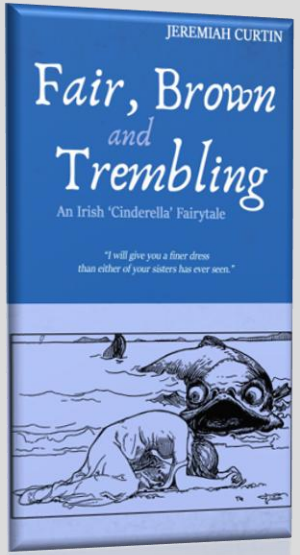

The Cinder Maid
by Martin F. Sorensen
Illustrated by
Bernard Le Roy

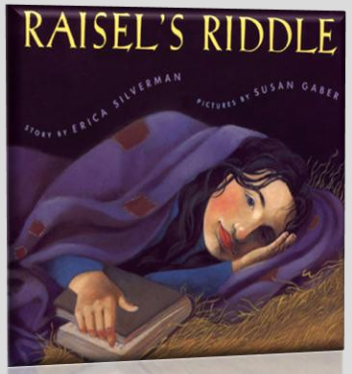

Angkat
The Cambodian Cinderella
written by Jewell Reinhart Coburn
illustrated by Eddie Flott

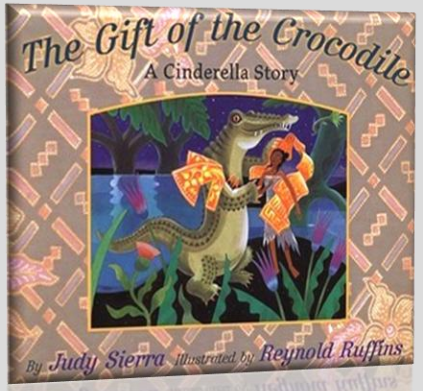

BIBLIOTHÈQUE ROSE
Walt Disney
présente
CENDRILLON
d'après Charles Perrault

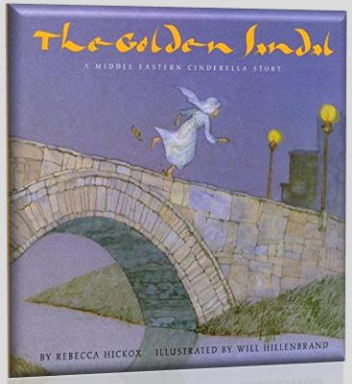

Abadeha
The Philippine Cinderella
Adapted by
Myrna J. de la Paz
Illustrated by
Youshan Tang

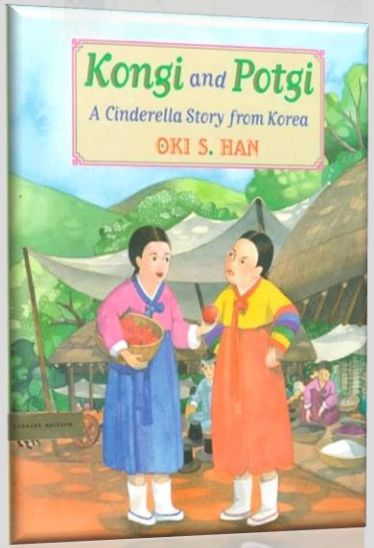

The Way
Meat Loves Salt
A Cinderella Tale from the Jewish Tradition
Nina Jaffe
Illustrated by Louise August


JEREMIAH CURTIN
Fair, Brown
and
Trembling
An Irish 'Cinderella' Fairytale
"I will give you a finer dress
than either of your sisters has ever seen."


RAISEL'S RIDDLE
STORY BY ERICA SILVERMAN
PICTURES BY SUSAN GABER


The Gift of the Crocodile
A Cinderella Story
By Judy Sierra Illustrated by Reynold Ruffins


The Golden Sandal
A MIDDLE EASTERN CINDERELLA STORY
BY REBECCA HICKOX ILLUSTRATED BY WILL HILLENBRAND


Kongi and Potgi
A Cinderella Story from Korea
OKI S. HAN

AIM

Can we develop a methodology for **automatic motif detection** in text, based on LLMs?

a motif: a repeated element in a story

cruel stepmother
glass shoes
helpful birds etc.

Do LLMs recognize motifs in texts?

- 13 Cinderella variants from different parts of the world (text in English)
- motifs from ATU 510A tale type – manually labelled: 15 motifs
- accuracy – the proportion of correct answers (true positives, true negatives)
- **98 %** motifs – correctly labelled by GPT-4.5-Preview

Which tales were included in the study?

- we extended the study to more Cinderella tales:
  - **First**: **77 variants** from five different continents

    (44 from Europe, 18 from Asia, 11 from the Americas, 3 from Africa,

    1 from Antarctica) - in English
  - Then: **33 Slovene** variants - in Slovene

What LLM prompt did we use?

- GPT-4.5-Preview (API)
- „zero-shot“ prompts – no examples

System message: "You are a helpful assistant, skilled in finding motifs in fairy tales. Follow the instructions carefully."

User message: "Find motifs in the following story. For each motif, return 'yes' if the motif is present in the story, return 'no' if the motif is absent. If the motif is modified, add a very short comment next to yes or no. Do not write the motif. Choose among the following motifs: 1. Cruel stepmother, 2. Stepdaughter heroine, 3. Helpful birds, 4. Clothes produced by magic, 5. Fairy as helper, 6. Magic object received from mother, 7. Magic object found on mother‘s grave, 8. Dead mother returns to aid persecuted children, 9. Prince sees heroine at ball and is enamored, 10. Prince sees maiden at church and is enamored, 11. Tabu: staying too long at ball. Must leave before certain hour, 12. Heroine’s three-fold flight from ball, 13. Glass shoes/slippers, 14. Shoe/slipper test. Identification by fitting of shoe/slipper, 15. False bride‘s mutilated feet. In order to wear the shoes/slippers with which the husband is testing the identity of his bride, the false bride cuts her feet. She is detected.“

Do LLMs recognize motifs in texts?

- yes/no answers – transformed into an *m* x *n matrix:*

  0 – marks motif absence, *1* – marks motif presence

	A	B	C	D	E	F	G	H	I	J
1	Title	1. Cruel stepmother	. Stepdaughter heroin	3. Helpful birds	lothes produced by m	5. Fairy as helper	c object received from	bject found on moth	r returns to aid perse	heroine at ball an
2	Abadeha-the Philippine Cinderell	1	1	0	0	0	0	1	0	
3	Angkat-The Cambodian Cinderell	1	1	1	0	0	0	0	1	
4	Anklet for a Princess-India-Cinde	1	1	0	1	0	0	0	0	
5	Ashey-Pelt-Cinderella-Ireland.txt	1	1	0	1	0	0	0	0	
6	Ashpitel-Scottish-Cinderella.txt	1	1	1	0	0	0	0	0	
7	Benizara and Kakezara-Japan-Cin	1	1	0	1	0	0	0	0	
8	Cendrillon-A Carribean Cinderell	1	1	0	1	0	0	0	0	
9	Cenerentola-Italian-Cat Cinderell	1	1	1	1	1	0	0	0	
10	Cinderella and the Golden Sanda	1	1	0	1	1	0	0	0	
11	Cinderella-Bulgarian.txt	1	1	1	1	0	0	1	1	
12	Cinderella-Conkiajgharuna, the L	1	1	0	1	0	0	1	0	
13	Cinderella-Dona Labismina-Brasil	0	0	0	1	0	0	0	0	
14	cinderella-holland.txt	1	1	0	0	0	0	0	0	
15	Cinderella-Indian-Canada.txt	0	0	0	0	0	0	0	0	
16	Cinderella-Indian-Canada-2.txt	0	0	0	0	0	0	0	0	
17	Cinderella-Islamic.txt	1	1	0	0	0	0	0	0	
18	Cinderella-Italian.txt	0	0	1	1	0	0	0	0	
19	Cinderella-Little Burnt Face-A Mi	0	0	0	0	0	0	0	0	
20	Cinderella-Massignon.txt	1	1	0	0	1	0	0	0	
21	Cinderella-Norway-Kraake-Lange	1	1	0	1	0	0	1	0	
22	Cinderella-or-Aschenputtel-Germ	1	1	1	1	0	0	1	0	
23	Cinderella-Pepelyouga-A Serbian	1	1	1	1	0	0	1	1	
24	Cinderella-Rhodopis and Her Litt	0	0	0	0	0	0	0	0	
25	Cinderella-Rushen Coatie-An Eng	1	1	0	1	0	0	0	0	
26	Cinderella-Rushen Coatie-English	1	1	0	1	0	0	0	0	
27	Cinderella-Sodewa Bai-A Tale fro	0	0	0	0	0	0	0	0	
28	Cinderella-Swedish-Askungen_Ci	1	1	0	1	1	0	0	0	
29	Cinderella-Swedish-The-Little-Go	1	1	1	1	0	0	0	0	

What motifs did we use?

- ATU = Aarne–Thompson–Uther **classification** system for organizing folktales, fairytales, and other narratives
- tales are classified on the basis of typical **motifs** that each type contains
- Thompson – a **motif** index
- ATU classification – handy tool, but problematic (overlapping types, structural and geographical bias, etc.)

ATU 510A Uther's motifs, according to Thompson	Added ATU 510A motifs	Supermotifs, according to Thompson
S31. **Cruel stepmother**	S31. **Cruel stepmother**	†S0--S99. **Cruel relatives.**
	A1273.1. Incestuous first parents	
L55. **Stepdaughter heroine**	**Stepdaughter heroine**	†L0--†L99. **Victorious youngest child.**
	Stepson hero	
B450. **Helpful birds**	**Helpful domestic animals**	†B400--†B499. **Kinds of helpful animals.**
	Helpful wild animals	
D1050.1 **Clothes produced by magic**		D1050. **Magic clothes.**
N815. **Fairy as helper**		N810. **Supernatural helpers.**
D815.1. **Magic object received from mother**		D810. **Magic object a gift.**
D842.1. **Magic object found on mother's grave**		D840. **Magic object found.**
E323.2. **Dead mother returns to aid persecuted children**		E320. **Dead relative's friendly return.**
N711.6. **Prince sees heroine at ball and is enamored**		N710. **Accidental meeting of hero and heroine.**
N711.4. **Prince sees maiden at church and is enamored**		
C761.3. **Tabu: staying too long at ball. Must leave before certain hour**		C750. **Time tabus.**
R221. **Heroine's three-fold flight from ball**		R220. **Flights (as hurried escapes).**
F823.2. **Glass shoes**		F820. **Extraordinary clothing and ornaments.**
H36.1. **Slipper test. Identification by fitting of slipper**		†H0--†H199. **Identity tests: Recognition.**
K1911.3.3.1. **False bride's mutilated feet.** In order to wear the shoes with which the husband is testing the identity of his bride, the false bride cuts her feet. She is detected		K1910. **Marital impostors.**

AIM

Can we find groups of similar folktale variants?

- different types of clustering:
  - **Kmeans**, Kmedoids, aglommerative, DBSCAN
- **Silhouette coefficient** – measures clustering quality
  - value [-1, 1]

Can we find groups of similar folktale variants?

Results – Kmeans clustering:

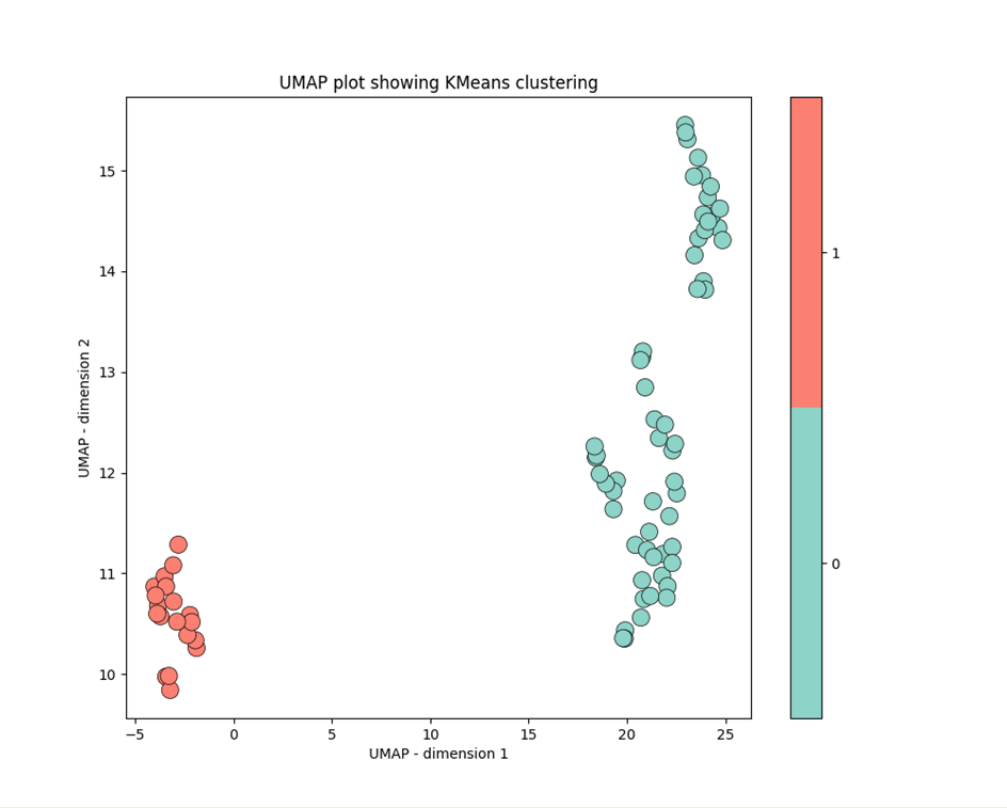
15 original specific motifs:
2 clusters
s = 0.9

18 specific motifs:
2 clusters
s = 0.81

14 supermotifs:
4 clusters
s = 0.6

How do we recognize similar tales?

- reduce dimensionality with **UMAP** algorithm – to remove uninformative motifs and to perform visualization



15 specific motifs:
2 clusters

1. Results – foreign vs. Slovene Cinderella variants: 15 original specific motifs

motif occurs in at least 66% of the folktales

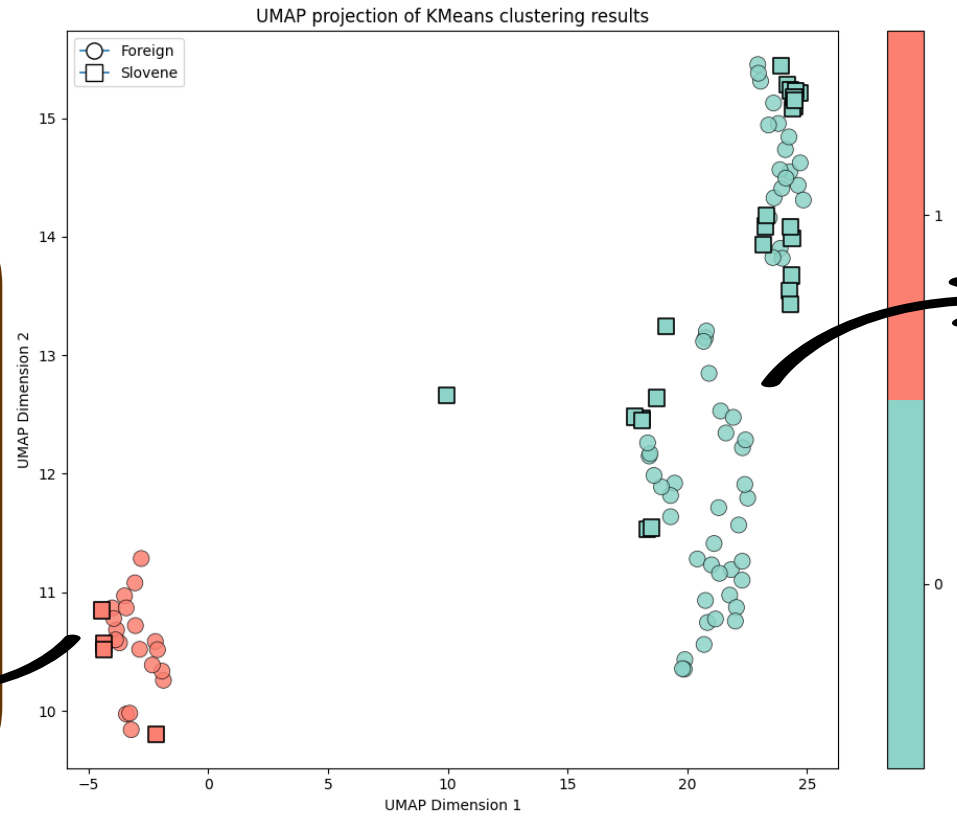
motif occurs in at least 33% of the folktales

+ 4 Slovene

clothes produced by magic

shoe/slipper test

19 folktales



cruel stepmother

stepdaughter heroine

clothes produced by magic

shoe/slipper test

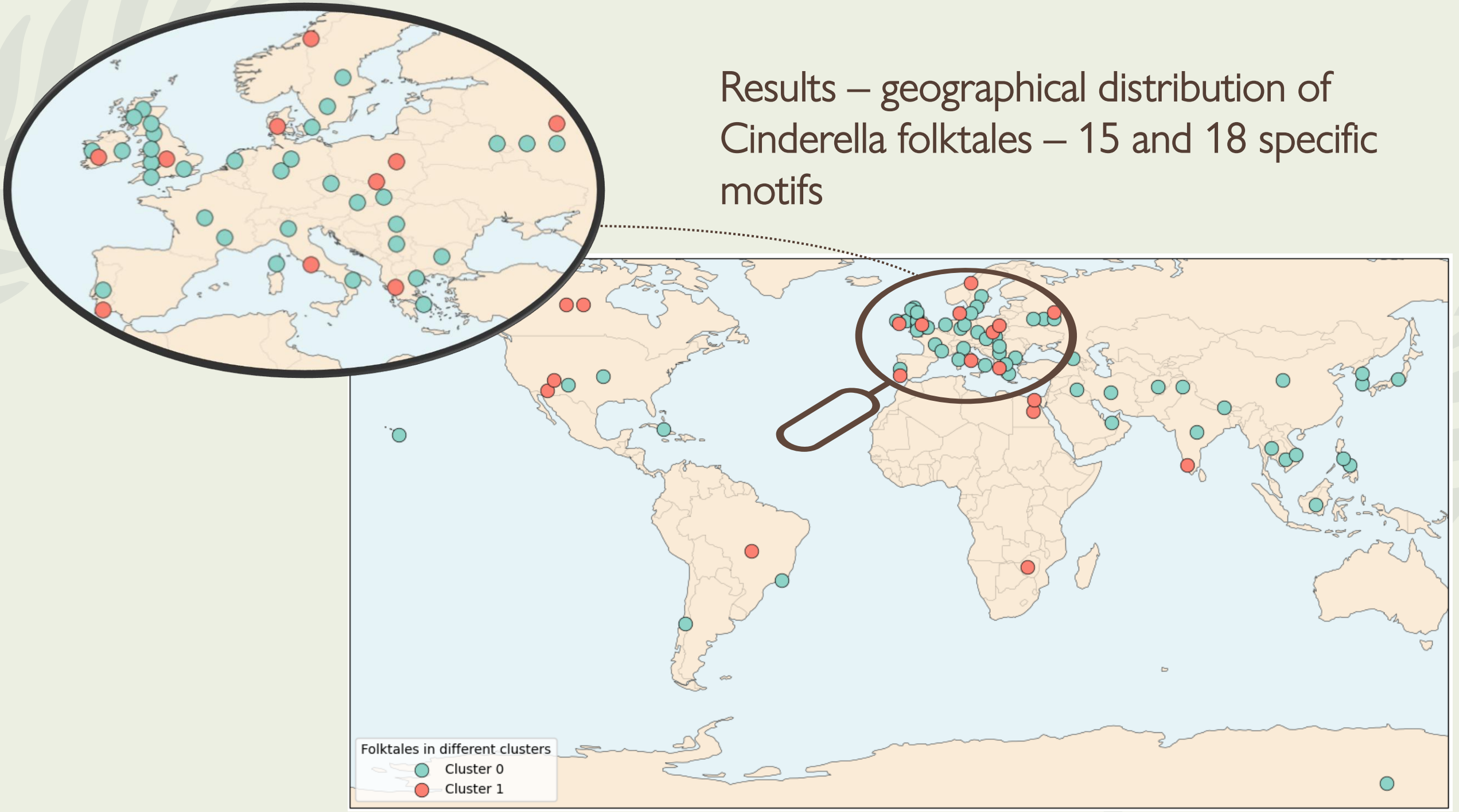
helpful birds

58 folktales

+ 29 Slovene

Source: 77 Cinderella folktales from around the world, 33 Slovene folktales

Results – geographical distribution of Cinderella folktales – 15 and 18 specific motifs



- Conclusion:

  • LLMs can **effectively** detect motifs in tales – reduces the need for time-consuming manual annotation

  • LLMs capture also **variations** and **deviations** from established motif patterns

  • presence/absence of motifs in a folktale can be the basis for finding **similar** tales (types) and their further analysis and visualization of their relations

  • LLM revealed **limitations** in the current motif typology:
  need for typology that is **better aligned with motif patterns** across diverse tales

  • new possibilities for **large-scale** narrative analyses – foundation for more systematic and computationally grounded frameworks



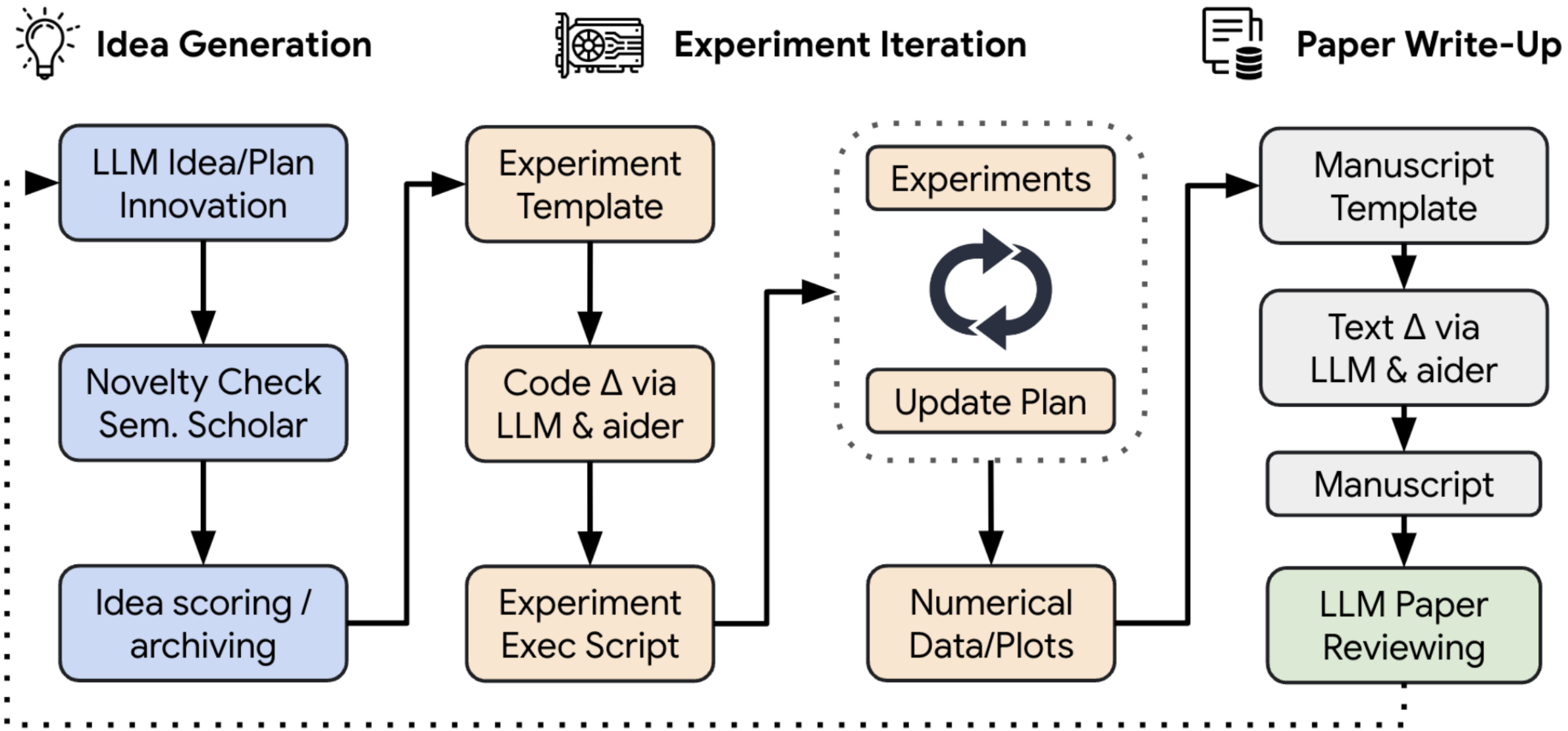
The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery

Chris Lu[1,2,*], **Cong Lu**[3,4,*], **Robert Tjarko Lange**[1,*], **Jakob Foerster**[2,†], **Jeff Clune**[3,4,5,†] **and David Ha**[1,†]

[*]Equal Contribution, [1]Sakana AI, [2]FLAIR, University of Oxford, [3]University of British Columbia, [4]Vector Institute, [5]Canada CIFAR AI Chair, [†]Equal Advising

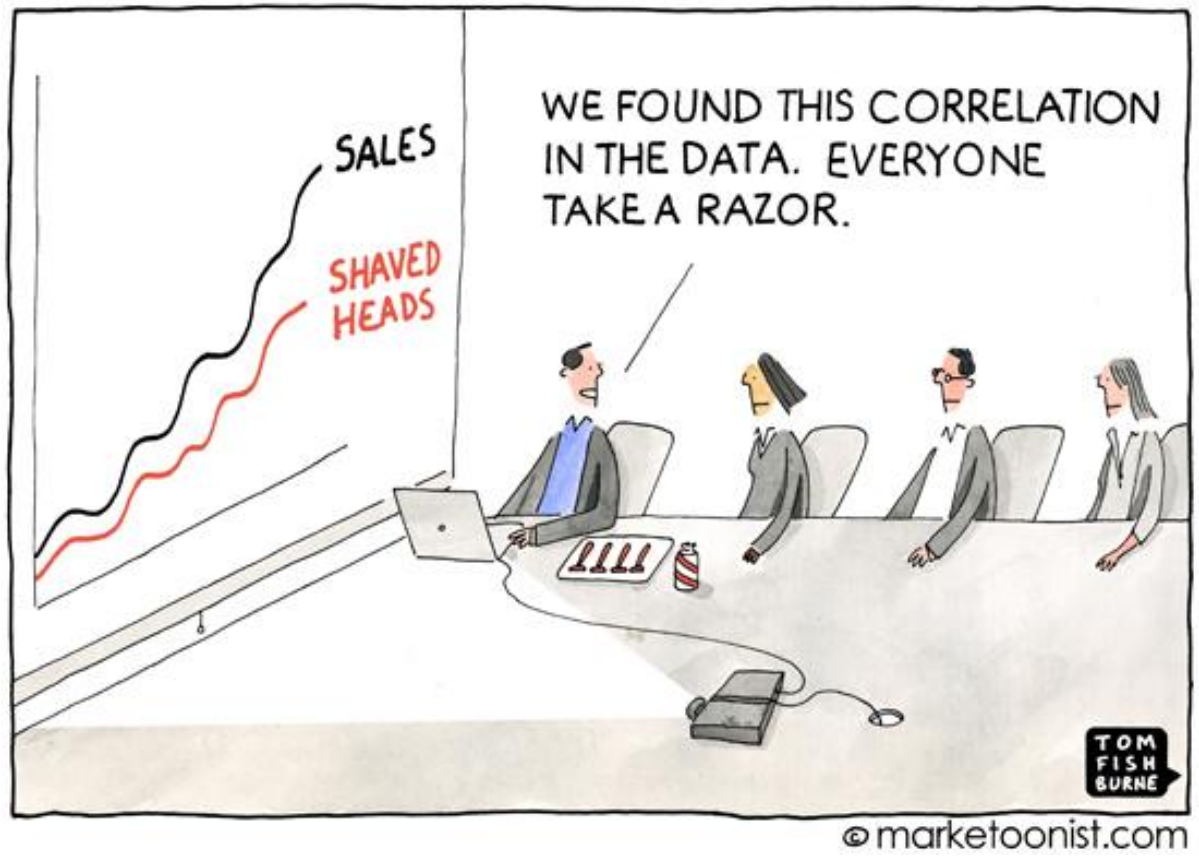
arXiv:2408.06292v2 [cs.AI] 15 Aug 2024

One of the grand challenges of artificial general intelligence is developing agents capable of conducting scientific research and discovering new knowledge. While frontier models have already been used as aides to human scientists, e.g. for brainstorming ideas, writing code, or prediction tasks, they still conduct only a small part of the scientific process. This paper presents the first comprehensive framework for fully *automatic scientific discovery*, enabling frontier large language models (LLMs) to perform research independently and communicate their findings. We introduce The AI Scientist, which generates novel research ideas, writes code, executes experiments, visualizes results, describes its findings by writing a full scientific paper, and then runs a simulated review process for evaluation. In principle, this process can be repeated to iteratively develop ideas in an open-ended fashion and add them to a growing archive of knowledge, acting like the human scientific community. We demonstrate the versatility of this approach by applying it to three distinct subfields of machine learning: diffusion modeling, transformer-based language modeling, and learning dynamics. Each idea is implemented and developed into a full paper at a meager cost of less than $15 per paper, illustrating the potential for our framework to democratize research and significantly accelerate scientific progress. To evaluate the generated papers, we design and validate an automated reviewer, which we show achieves near-human performance in evaluating paper scores. The AI Scientist can produce papers that exceed the acceptance threshold at a top machine learning conference as judged by our automated reviewer. This approach signifies the beginning of a new era in scientific discovery in machine learning: bringing the transformative benefits of AI agents to the *entire* research process of AI itself, and taking us closer to a world where *endless affordable creativity and innovation* can be unleashed on the world's most challenging problems. Our code is open-sourced at https://github.com/SakanaAI/AI-Scientist.

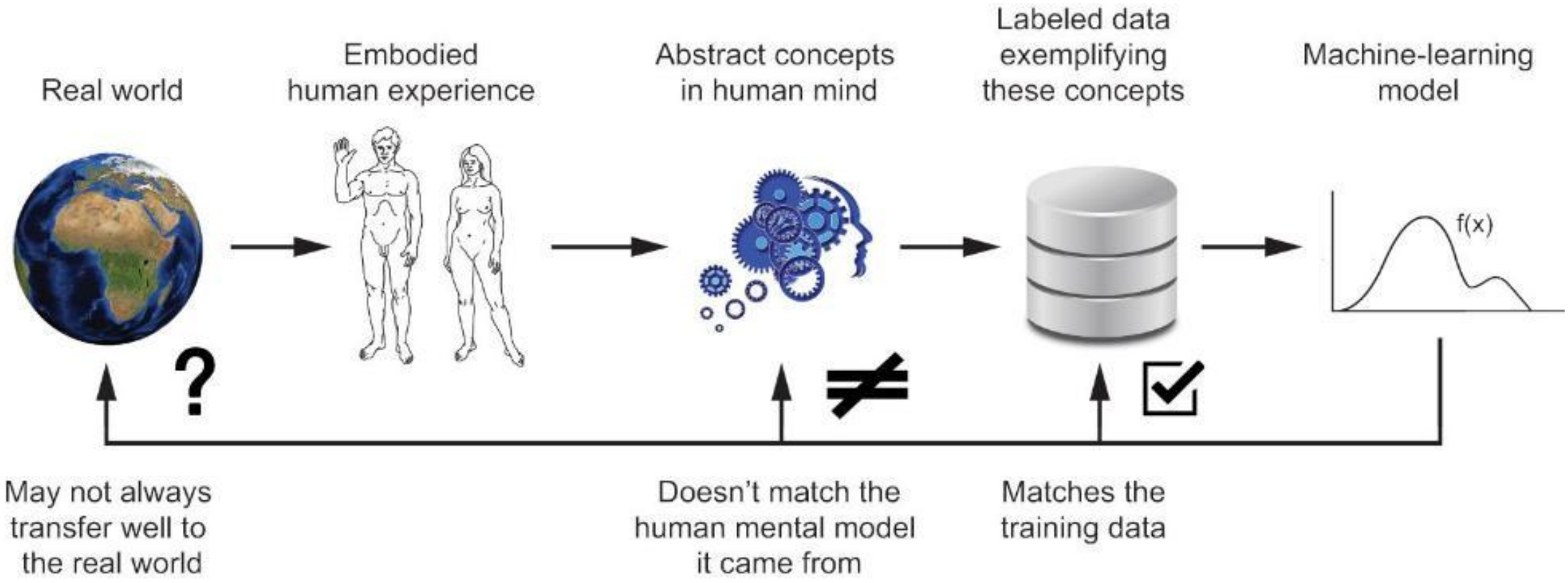

Idea Generation
Experiment Iteration
Paper Write-Up
LLM Idea/Plan Innovation
Novelty Check Sem. Scholar
Idea scoring / archiving
Experiment Template
Code Δ via LLM & aider
Experiment Exec Script
Experiments
Update Plan
Numerical Data/Plots
Manuscript Template
Text Δ via LLM & aider
Manuscript
LLM Paper Reviewing

Razložljiva umetna intelligence (Explainable AI, XAI)

- Modeli strojnega učenja so pogosto črne škatle
- Na veliko področjih je transparentnost modelov zelo pomembna: medicina, pravo, svetovanje, javne storitve itd.
- Pravna obveznost po GDPR (Splošna uredba o varstvu podatkov) in Zakonu o umetni inteligenci (AI Act)

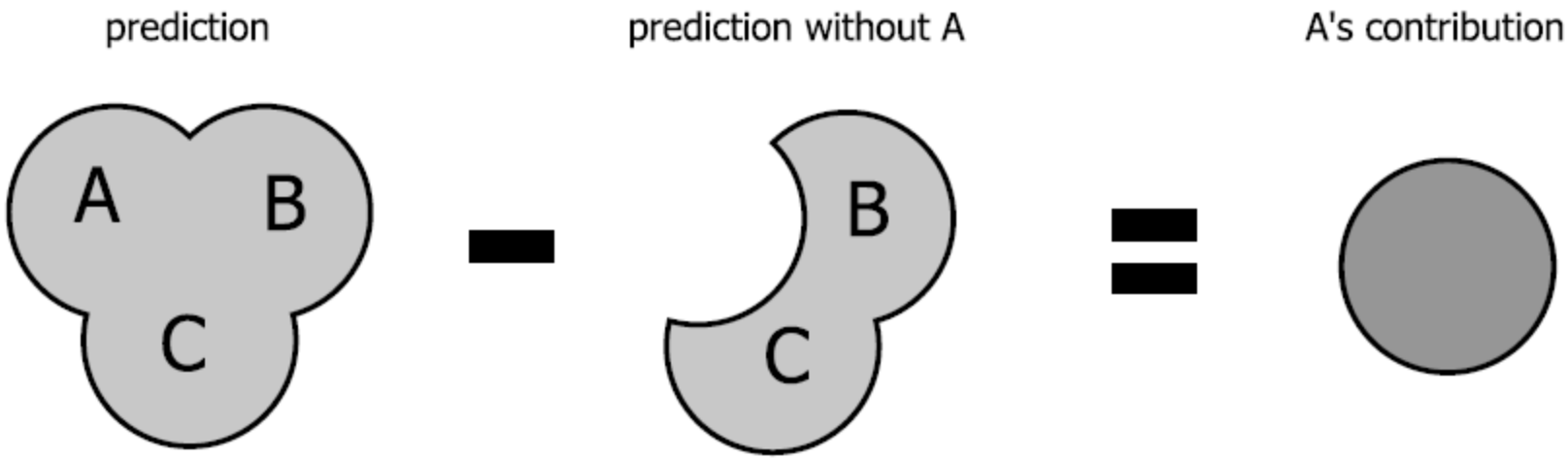


Učenje in strojno učenje

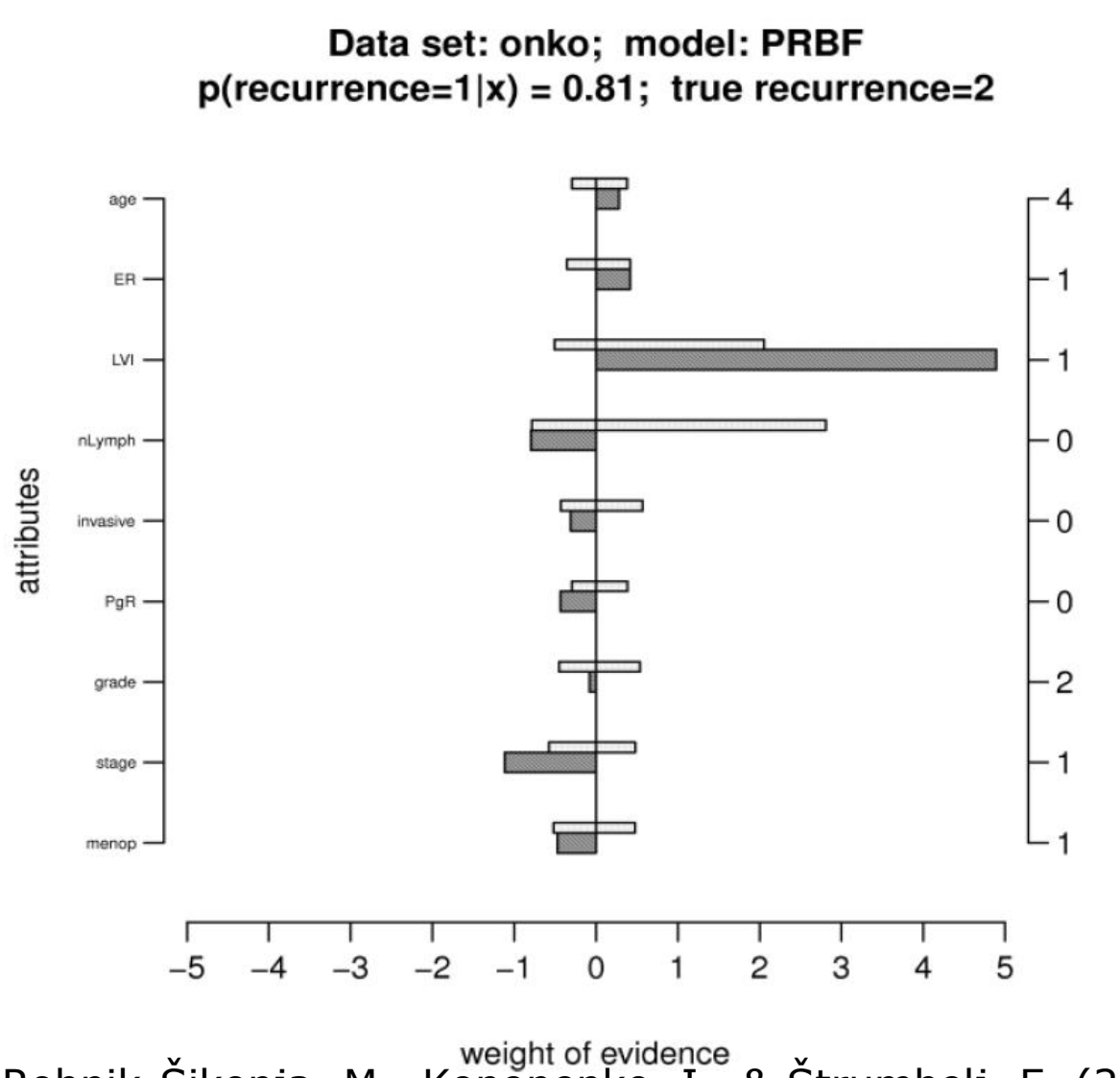


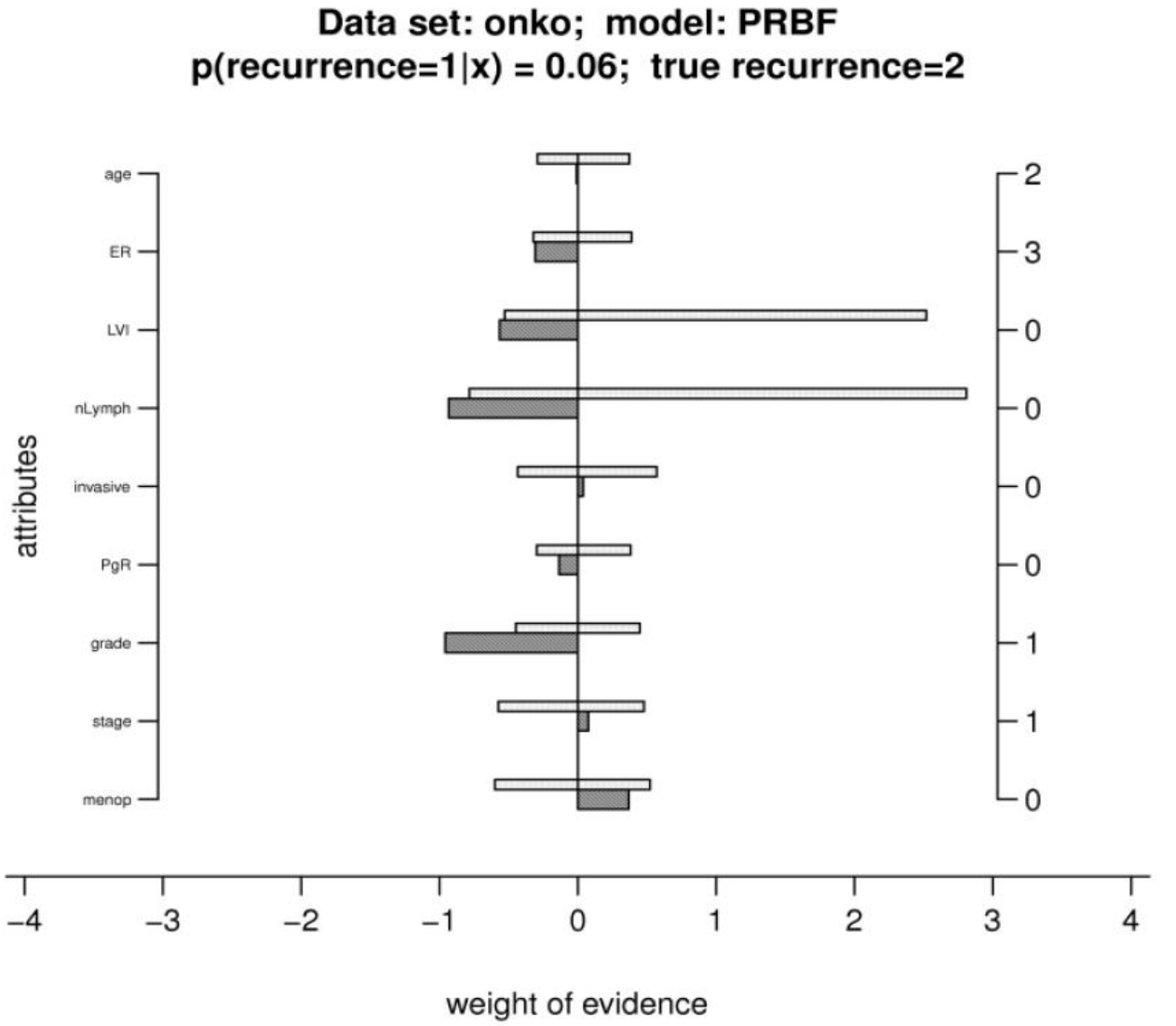
Razlage, ki temeljijo na perturbacijah

- Pomembnost značilke ali skupine značilkv v nekem modelu lahko ocenimo s simulacijo skrivanja informacij o vrednostih te značilke



Primer: ponovitev raka na prsih





Robnik-Šikonja, M., Kononenko, I., & Štrumbelj, E. (2012). Quality of classification explanations with PRBF. *Neurocomputing*, *96*, 37-46.

Primer razlage v VJM-jih: zdravorazumsko sklepanje

Winograd Schema Challenge (WSC, 2012): določanje pomena zaimkov, ki zahteva znanje o svetu in zdravorazumsko sklepanje

Površinski vzorci niso zadostni

Obstoječe razširitve: WinoGrande, WinoWhy, WinoGender itd.

The trophy doesn't fit in the brown suitcase because ***it*** *is too big.* What is too big?
A. **The trophy** B. The suitcase

Research questions



Question: Models achieve high accuracy, but can they also provide the right explanation?

Shortcomings of current LLMs

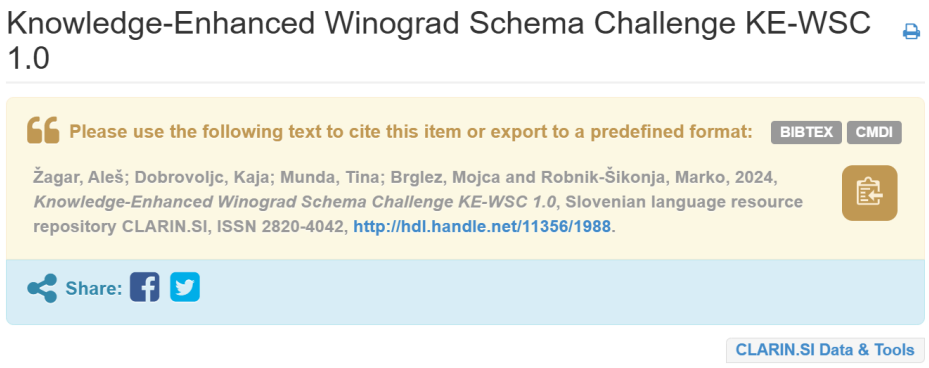
- →**Explanation gap**: Can often guess the right answer, but explanations are weak or missing
- →**Dataset gap**: Lack of structured explanations and reasoning-aware evaluation

Dataset

Extend the original dataset -> **KE-WSC: Knowledge-Enhanced Winograd Schema Challenge**

Primary objectives and research questions:

- →Enrich WSC examples with **human explanations** (create ontologies on different levels, generated explanations, etc.)
- →Study impact of explanations on model reasoning
- →Analyze model-generated explanations for quality and consistency



CLARIN.SI

UL | FRI

KE-WSC dataset construction steps

1. Translate English samples to Slovene.
2. Create new Slovene samples by modifying existing English (semantic & syntactic).
3. Developed a two-level ontology of knowledge.
4. Create human explanations.
5. Human explanations were translated to English.
6. 1001 examples (Slovene + English).

Experiments

Do human explanations improve commonsense reasoning?

3 Evaluation Settings (few-shot scenario):

- →Query → Label
- →Query + Explanation → Label
- →Query → Explanation + Label

LLMs tested (DeepSeek, Gemma, OpenAI o-series, GaMS)

- →Reasoning vs. non-reasoning models
- →Languages: English and Slovene

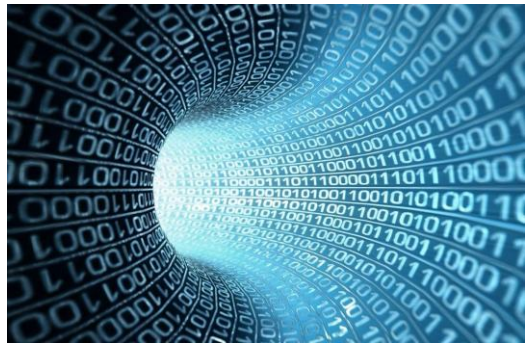
UL | FRI

KE-WSC Conclusions

- KE-WSC introduces structured explanations and multilingual data
- Adding explanations can improve results
- Surprising findings on automatic explanation quality assessment (LLM-as-a-judge):
  →Some low-accuracy models generate good explanations
  →Some high-accuracy models fail to explain well
- Models trained with reasoning benefit most from explanations
- Future work:
  →Human evaluation of explanation quality
  →Fine-grained error analysis by knowledge type
  →Cross-lingual generalization experiments

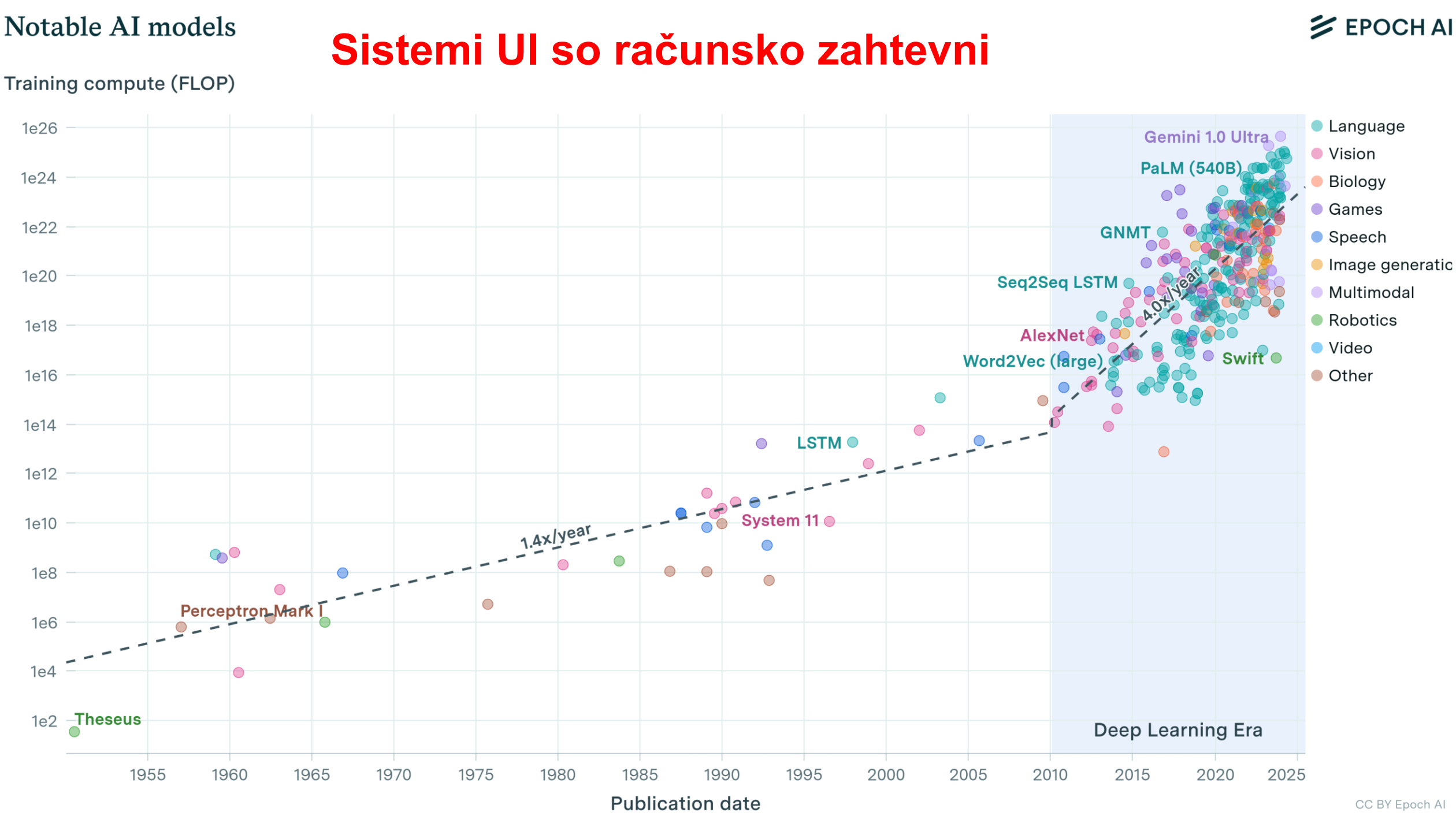
Aleš Žagar, Marko Robnik-Šikonja (2026): Human Explanations for Improving Commonsense Reasoning (submitted)

Etične dileme uporabe velikih jezikovnih modelov



- pristranost in poštenost modelov
  družbene in podatkovne pristranosti
- kakovost učnih podatkov
- privatnost in varnost podatkov
  modeli so zgrajeni na velikih zbirkah: zaščita, dostop
- odgovornost razvijalcev/lastnikov modelov
- transparentnost in razložljivost modelov
  veliki modeli so črne škatle
- zanesljivost in robustnost modelov
- obnavljanje modelov (spremenjene okoliščine, podatki, predpisi)

Sistemi UI so računsko zahtevni



Učenje VJM je podatkovno izjemno požrešno

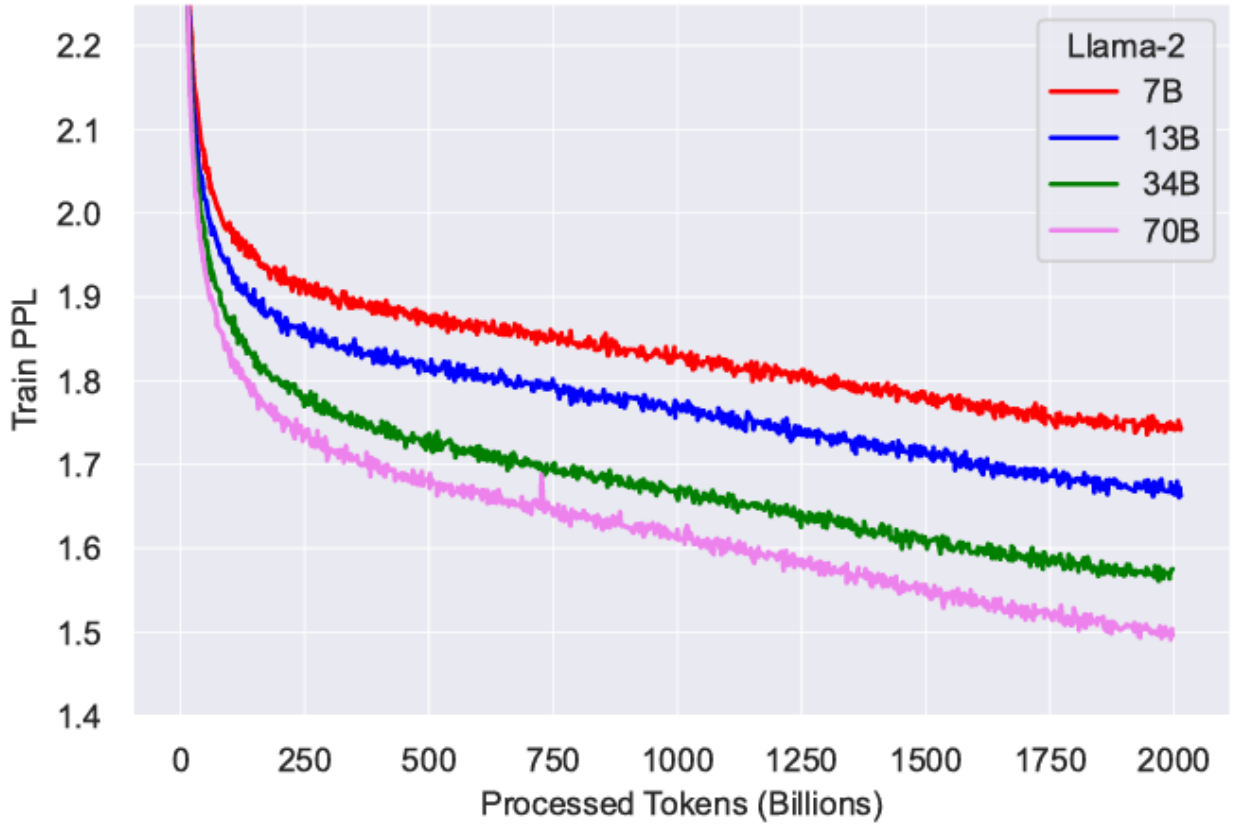


Figure 5: Training Loss for Llama 2 models. We compare the training loss of the Llama 2 family of models. We observe that after pretraining on 2T Tokens, the models still did not show any sign of saturation.

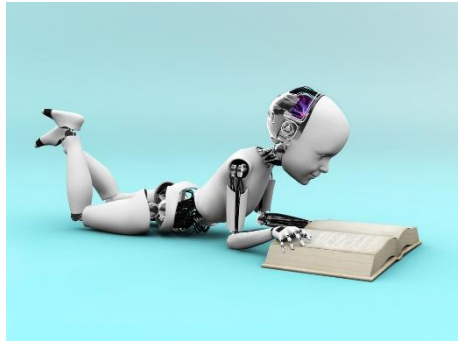
HPC is an essential infrastructure for LLMs

- Building LLM from scratch is computationally extremely costly
  - LLaMa-2 7B > 360,000 GPU hours,
  - SLING tender: Vega GPU offers 80,000 node hours = 320.000 GPU hours annually
- Building LLM by further training existing LLM:
  - Data quantity and quality issues
  - Platform issues
  - Model parallelization (within 1 GPU node, i.e. 4 or 8 A/H100),
  - Data parallelization (across nodes), reduces inter-node communication
  - Building LLM requires lots of HPC expertise
- Using platforms like NVidia NeMo framework:
  - +speed
  - +easier parallelization
  - -less models available
  - -needs model and code tweaking
- In deployment: squeeze everything on a single computational node (if possible)

Univerza v Ljubljani
Fakulteta za računalništvo in informatiko

Kam gredo veliki jezikovni modeli?

- na tisoče novih aplikacij z VJM
- multimodalni modeli: mešanje besedila, slik, zvoka in videa
- vključevanje obstoječega znanja v nevronske mreže, npr. omrežja z jezikovnimi, pravnimi ali biološkimi relacijami, kot sta Legal-RDF in BioRDF
- novi semantični viri in tehnologije
- modeli, ki kličejo zunanje vtičnike, t.i. function calling
- Sklepanje in agenti
- razlage in vključevanje etičnega vidika



EU Uredba o UI (AI Act)

UL FRI

- Nekomercialno raziskovanje je izvzeto iz regulacije
- Za raziskave, ki lahko vodijo do tržnih produktov so sistemi UI razvrščeni v 4 kategorije:
  - **Nesprejemljivo tveganje**: UI aplikacije, ki predstavljajo resno grožnjo varnosti ali temeljnim pravicam, so prepovedane.
  - **Visoko tveganje**: Sistemi umetne inteligence, ki lahko pomembno vplivajo na zdravje, varnost ali temeljne pravice, morajo izpolnjevati stroge zahteve, vključno s kakovostjo, preglednostjo, človeškim nadzorom in varnostnimi standardi.
  - **Omejeno tveganje**: Ti sistemi imajo določene zahteve glede preglednosti, tako da uporabniki vedo, da komunicirajo z umetno inteligenco.
  - **Minimalno tveganje**: Večina aplikacij umetne inteligence spada v to kategorijo in se sooča z minimalnimi regulativnimi zahtevami.
- Raziskave umetne inteligence, ki se usmerjajo v **visoko tvegane aplikacije**, morajo upoštevati te zahteve že v zgodnji fazi razvoja.
- Akt vključuje tudi določila za **splošno uporabne sisteme umetne inteligence**, kot so **temeljni modeli** (npr. ChatGPT). Ti sistemi so predmet zahtev glede **preglednosti**, tisti z večjim tveganjem pa morajo prestati podrobnejše ocene. Raziskovalci, ki razvijajo takšne modele, morajo biti pozorni na te obveznosti, zlasti če bi njihov sistem lahko imel velik vpliv v različnih sektorjih.
- Raziskovalci se morajo zavedati klasifikacij tveganja in predvideti, katere zahteve bo treba izpolniti, če bodo želeli svoj AI sistem **komercializirati**.
- Akt poudarja pomen **etičnega razvoja umetne inteligence**, zato morajo raziskovalci upoštevati etične vidike že med razvojem.

Projekt MzK: Razvoj slovenščine v digitalnem okolju, RSDO, 2021-2023

- nadgradnja slovenske besedilne korpusov in leksikona besednih oblik
- učne množice in modeli za strojno označevanje slovenščine
- govorna baza 1000 ur govora
- splošni razpoznavalnik govora
- digitalna slovarska baza z različnimi tipi jezikovnih podatkov o slovenščini
- semantična mreža za zdravorazumsko sklepanje
- viri in orodja za razdvoumljanje pomenov in prepoznavanje semantičnih premikov
- avtomatsko povzemanje in odgovarjanje na vprašanja
- korpus prevodov in strojni prevajalnik za par angleščina-slovenščina
- terminološki portal z iskalnikom
- orodja za luščenje in urejanje terminologije iz korpusov
- svetovalnica za reševanje terminoloških vprašanj
- infrastruktura CLARIN.SI za distribucijo jezikovnih virov in orodij

- www.slovenscina.eu

www.slovenščina.eu

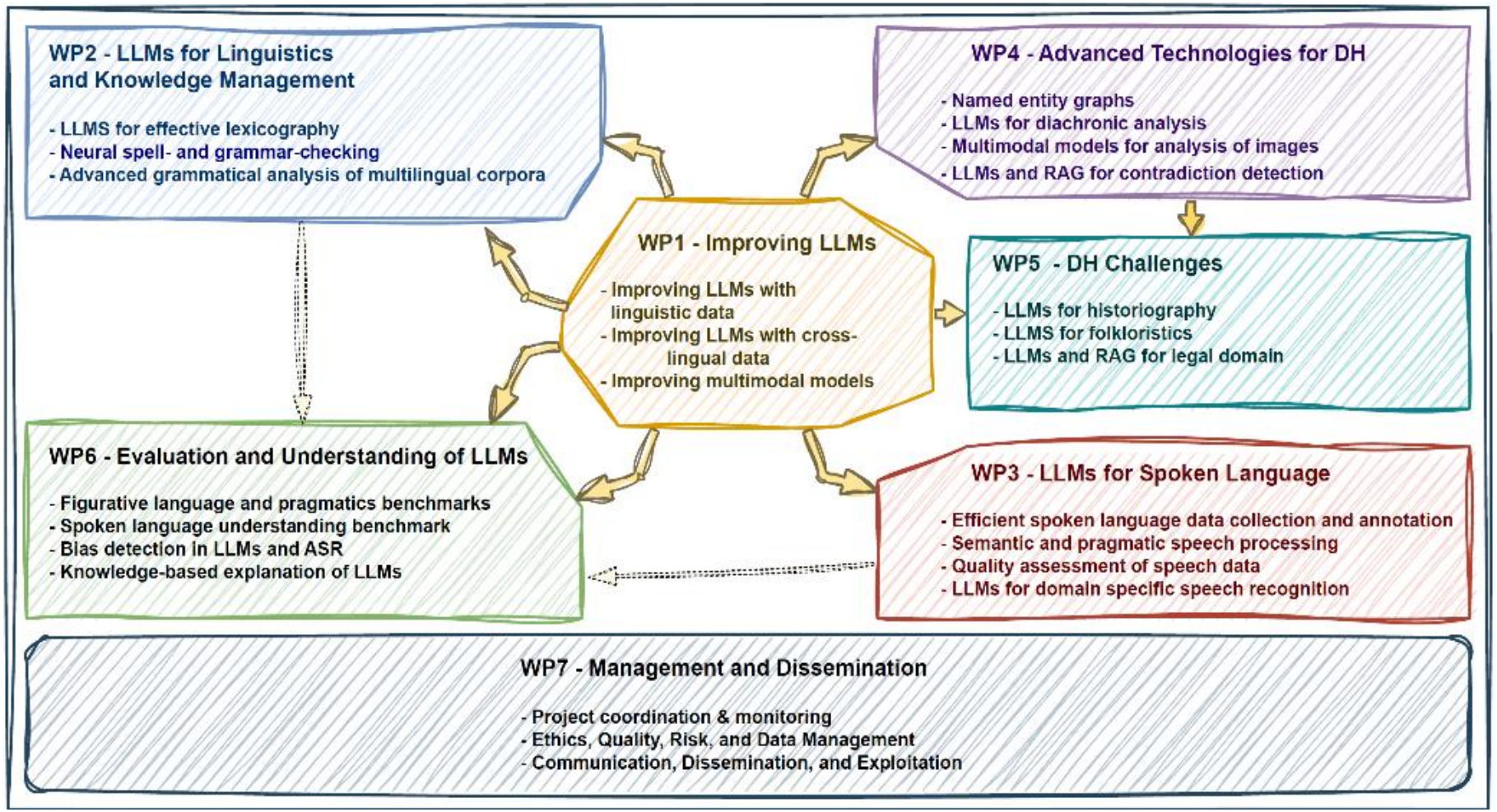
REPUBLIKA SLOVENIJA
MINISTRSTVO ZA KULTURO

EVROPSKA UNIJA
EVROPSKI STRUKTURNI IN INVESTICIJSKI SKLADI
NALOŽBA V VAŠO PRIHODNOST

UNIVERSITY OF LJUBLJANA | FRI Faculty of Computer and Information Science

Veliki jezikovni modeli za digitalno humanistiko, 2024-2027





Slovenski nacionalni ARIS projekt za povezovanje raziskav umetne inteligence in digitalne humanistike

Sodelovanje 5 institucij, 2,25 milijjona €

Glavne naloge:

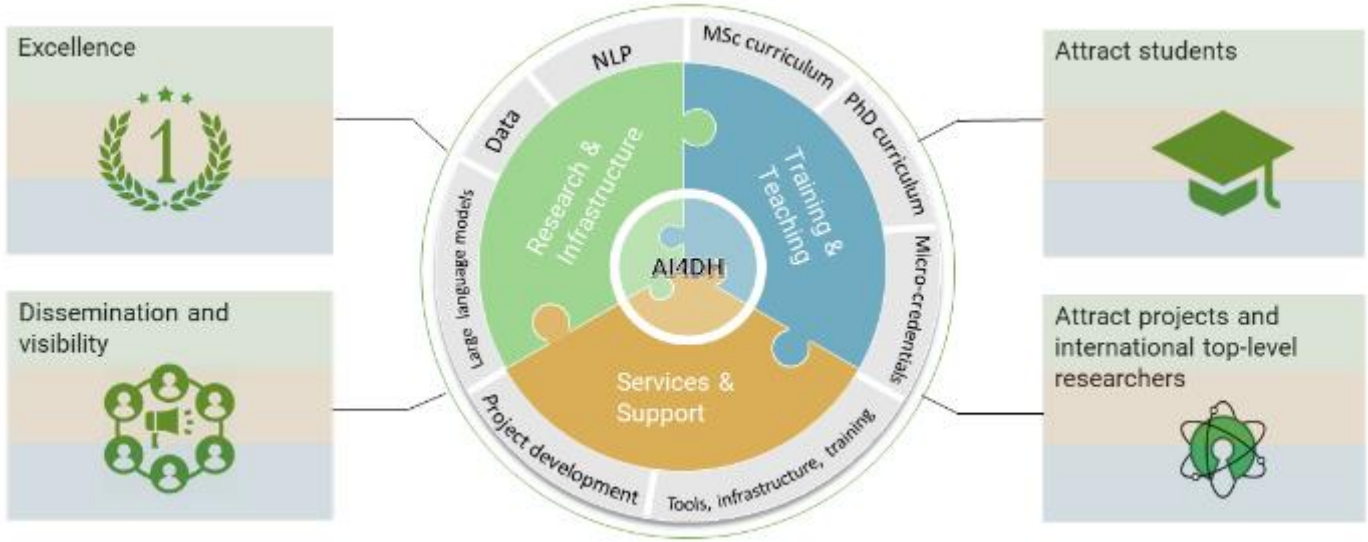
- VJM in MVM
- jezikoslovje
- historigrafija
- govorjeni jezik
- folkloristika
- pravo
- razlage itd.

Cilj: vrhunske raziskave

Centre of Excellence in Artificial Intelligence for Digital Humanities

Center odličnosti umetne intelligence v digitalni humanistiki

- EU ERA Chair projekt, 2025 - 2030, 2.5 milijona €
- Project coordinator: Marko Robnik-Šikonja
- Vzpostavitev Centra odličnosti na UL
- Glavne dejavnosti:
  - nuditi raziskovalno podporo na področju umetne inteligence
  - gostiti raziskovalce, ki pripravljajo predloge projektov
  - prijavljati in gostiti projekte na področju umetne inteligence za digitalno humanistiko
  - izdelava orodij in odprtih zbirk podatkov
  - graditi in prilagajati VJM
  - računalniška podpora raziskovalcem na področju DH
  - izobraževanje na področju UI in DH



AI4DH

Center za jezikovne vire in tehnologije Univerze v Ljubljani - www.cjvt.si (FF, PEF, FDV, FE, FU, FRI)

- Besedilni korpusi
  - Referenčni korpus pisne standardne slovenščine
  - Korpus govorjene slovenščine Gos
  - Korpus šolskih pisnih izdelkov Šolar
  - povzemalnik korpusnih podatkov Korpusnik
  - Jezikovni sledilnik
- Digitalna slovarska baza za slovenščino
  - slovar sopomenk
  - slovensko-madžarski slovar
  - kolokacijski slovar



Center za jezikovne vire in tehnologije Univerze v Ljubljani - www.cjvt.si (FF, PEF, FDV, FE, FU, FRI)

- Strojna obdelava besedil
  - jezikoslovno označevanje in razčlenjevanje
  - poenostavljanje besedil
  - določanje berljivosti in kompleksnosti besedil
  - orodja za jezikovne popravke (npr. Vejice)
- Govorne tehnologije
  - sprotno strojno tolmačenje predavanj (ON)
- Umetna inteligenca
  - veliki jezikovni modeli



Slovenska tovarna umetne intelligence
Slovene AI factory - SLAIF



- namen: razširiti ekosistem umetne inteligence v Sloveniji
- povezava s superračunalniškim omrežjem Euro HPC
- strojna oprema v Sloveniji: superračunalnik Vega 2
- podporni projekt Slo/EU za prenos znanja v znanost in industrijo v okviru programa Obzorje 2020, 2025–2028
- razvoj talentov in znanja
- izobraževalni programi
- dostop do podatkov in upravljanje podatkov
- podpiranje znanosti in podjetij pri uporabi umetne inteligence
- UL FRI je vodilni partner na področju HPC, LLM in računalniškega vida

SLAIF – sodelovanje s podjetji

UL FRI



Učenje GaMS-a na infrastrukturi SLAIF (Vega)

Postavitev lokalnega strežnika GaMS

Kompetenčni center UI

KCUI

- KCUI služi kot vstopna točka za podjetja, raziskovalce, javne institucije in druge uporabnike, ki želijo uporabljati ali razvijati rešitve umetne inteligence.
- Namen ukrepa je spodbuditi prenos vrhunskih kompetenc in tehnologij v prakso, omogočiti dostop do strokovnega znanja, okrepiti uporabo UI v gospodarstvu ter ustvariti podporno okolje za razvoj, testiranje in uporabo naprednih rešitev.

Infrastruktura za slovenščino => evropska infrastruktura

PoVeJMo
SLOVENSKI VELIKI JEZIKOVNI MODEL

ALT-EDIC

- Alliance for Language Technology – European Digital Infrastructure Consortium
- Član iz Slovenije: MDP, zdajMNZJU (Ministrstvo za notranje zadeve in javno upravo)
- 100M€ - trije projekti
- Odprt jezikovni model za vse evropske jezike
- Podatkovna infrastruktura (Language Data Space)

Projekti

- ALT-EDIC4EU (infrastruktura)
- OpenEuroLLM (model)
- LLMs4EU (podatki, evalvacija)

UNIVERSITY OF LJUBLJANA | FRI Faculty of Computer and Information Science

Institut "Jožef Stefan" Ljubljana, Slovenija

eventregistry

TelekomSlovenije

XLAB

ARCTUR

Zaključki

- Veliki jezikovni modeli so osrednji pristop k sodobni umetni inteligenci
- Demokratizacija uporabe UI
- Trije glavni načini uporabe in prilagajanja:
- ukazni pozivi
- prilagajanje za dano nalogo
- prilagajanje za sledenje navodilom za dano domeno
- HPC je bistvena infrastruktura za VJM-je
- Naprednejša uporaba zahteva dostop do API-jev, prilagajanje, RAG, dodatno vnaprejšnje učenje, sklepanje, klicanje funkcij, agente.



Stik

UL | FRI

AI4DH

GaMS: “

- marko.robnik@fri.uni-lj.si

- Splet:
  - https://povejmo.si/
  - https://ai4dh.eu
  - https://www.cjvt.si/
  - https://www.cjvt.si/llm4dh/

Funded by the European Union

NOO | NAČRT ZA OKREVANJE IN ODPORNOST

aris

UNIVERZA V LJUBLJANI | FRI Fakulteta za računalništvo in informatiko

KCUI

Kompetenčni center za umetno inteligenco

Slovenian National AI Competence Centre

Hvala za udeležbo

REPUBLIKA SLOVENIJA
MINISTRSTVO ZA NOTRANJE ZADEVE IN JAVNO UPRAVO

REPUBLIKA SLOVENIJA
MINISTRSTVO ZA IZOBRAŽEVANJE, ZNANOST IN MLADINO

REPUBLIKA SLOVENIJA
MINISTRSTVO ZA GOSPODARSTVO DELO IN ŠPORT

REPUBLIKA SLOVENIJA
MINISTRSTVO ZA OBRAMBO

ARNES