ELEKTROTEHNIŠKI VESTNIK 90(3): 247–253, 2023
IZVIRNI ZNANSTVENI ČLANEK

Vrednotenje sposobnosti velikih jezikovnih modelov z nalogami strojnega učenja v času sklepanja

Klemen Grm

Univerza v Ljubljani, Fakulteta za elektrotehniko, Tržaška cesta 25, 1000 Ljubljana, Slovenija
E-pošta: klemen.grm@fe.uni-lj.si

Povzetek. Na področju strojnega učenja uporabljamo algoritme, ki se lahko učijo iz podatkov, da bi izboljšali svoje delovanje pri določeni nalogi ali naboru nalog. Tipične naloge strojnega učenja vključujejo razvrščanje, regresijo, in generativno modeliranje. Najobičajnejši sodobni primer algoritma strojnega učenja v praktični uporabi so globoka nevronska omrežja v povezavi z zunanjim optimizatorjem, kot je stohastična gradientna metoda. V zadnjem času so veliki jezikovni modeli pokazali vse večje zmožnosti metaučenja v kontekstu, ki se uporablja za izboljšanje njihove uspešnosti pri jezikovnih nalogah z dodatnimi učnimi primeri (angl. *few-shot learning*). V pričujočem članku pokažemo, da lahko vnaprej naučeni veliki jezikovni modeli delujejo kot strojni učenci glede na podatke v kontekstu, brez uporabe zunanjih optimizacijskih orodij oziroma posodobitev uteži. Z ocenjevanjem sposobnosti jezikovnih modelov opravljanja nalog strojnega učenja v času sklepanja na sintetičnih ali ustrezno preoblikovanih naborih podatkov prepričljivo pokažemo, da so sposobni modelirati zapletene odnose med podatki v vhodnem kontekstu in da je reševanje nalog strojnega učenja v času sklepanja smiselna metoda vrednotenja njihovih sposobnosti.

Ključne besede: jezikovni modeli, strojno učenje, metodologija vrednotenja

Evaluating the capabilities of large language models using machine learning tasks at inference-time

Machine learning is the domain of algorithms capable of learning from data to improve their performance on a task or set of tasks. Common machine learning tasks include classification, regression, and generative modelling. The most common modern example of machine learners in practical use is deep neural networks coupled with an extrinsic optimizer such as stochastic gradient descent. Recently, scaled-up large language models have shown increasing capabilities of in-context meta-learning, which has been used to improve their performance on language tasks through few-shot learning. In this paper, we show that pre-trained large language models can act as machine learners with regard to in-context data, without using extrinsic optimization tools or weight updates. By evaluating the language models' inference time machine learning abilities on synthetic or appropriately transformed datasets, we conclusively show that they're able to model complex relationships between data in the input context. This implies that inference-time machine learning tasks represent a meaningful capability evaluation task for large language models.

Keywords: language models, machine learning, evaluation methodology

1 UVOD

V zadnjih letih je *transformer* [27] arhitektura nevronskih omrežij omogočila skalabilno učenje velikih jezikovnih modelov, tj. modelov, parametriziranih z več milijardami prostih parametrov. Modeli, kot na primer serije GPT [21], [22], [4], [16] in Llama [25], [26] kažejo vedno boljše delovanje pri raznolikih jezikovnih nalogah, kot so prevajanje besedil, analiza sentimentov, odgovarjanje na vprašanja, ter prosta konverzacija.

Študije so pokazale, da jezikovni modeli takih velikosti pridobijo porajajoče zmožnosti, ki pri manjših modelih ne obstajajo niti v zmanjšani obliki [28]. Veliki jezikovni modeli torej predstavljajo kvalitativni preskok v sposobnostih jezikovnega modeliranja. Novejše študije najzmogljivejših jezikovnih modelov serije GPT-4 [5] kažejo na omejeno prisotnost splošne inteligence, kar med drugim dokazujejo s prisotnostjo *teorije uma* (angl. *theory of mind*), tj. sposobnosti modela opredeljevanja in razumevanja mentalnih stanj drugih agentov na podlagi daljših besedilnih scenarijev.

Vrednotenje zmožnosti oz. jezikovnega razumevanja naučenih modelov tipično poteka prek za to namenjenih podatkovnih zbirk. Pri nalogi nadaljevanja besedila se kot merilo uspeha uporabljajo npr. zbirka LAMBADA [17], ki kot nalogo predlaga napovedovanje zadnje besede paragrafa besedila glede na daljši prejšnji kontekst, ali StoryCloze [15], ki zahteva izbiro pravilne povedi glede na prejšnje. Za vrednotenje sposobnosti zdravorazumskega sklepanja se tipično uporablja zbirka vprašanj in odgovorov ARC [8]. Natančnosti (tj. deleži pravilnih odgovorov) modelov na teh in sorodnih testnih zbirkah so do nedavnega veljale kot objektivna ocena zmožnosti klasičnih pristopov k jezikovnemu modeliranju.

Winogradove sheme [23] so do pojava velikih jezikovnih modelov veljale kot ključno merilo inteligence,

Prejet 17. oktober, 2023
Odobren 10. november, 2023

primerljive s človeško. Sheme so sestavljene iz parov povedi, kot so npr.:

1. The city councilmen refused the demonstrators a permit because they feared violence.

2. The city councilmen refused the demonstrators a permit because they advocated violence.

Pri tem je naloga dereferenciranje zaimkov, kot npr. zaimka *they* v zgornjem paru povedi. Ta se glede na spremembo konteksta povedi lahko nanaša na različne subjekte. Naloga je ljudem lahko rešljiva, za klasične pristope k jezikovnemu modeliranju in umetni inteligenci pa je bila nepremostljiva ovira.

Z uporabo velikih jezikovnih modelov je bila Winograd Schema Challenge v veliki meri rešena, podrobnejša študija [14] pa kaže njeno nezadostnost kot merilo splošne inteligence. Avtorji kot pomanjkljivosti navajajo predvsem dvoumnost velikega števila testnih primerov, nejasne kriterije vrednotenja, ter praktično nezmožnost zagotavljanja, da se primeri iz zbirke ne nahajajo v učnih korpusih velikih jezikovnih modelov.

Pogosto omenjene pomanjkljivosti velikih jezikovnih modelov [3] vključujejo njihovo omejenost na fiksne učne korpuse. Časovna in pomnilniška zahtevnost učenja velikih jezikovnih modelov v praksi pomeni, da se zaradi denarne in ekološke cene učenja njihove uteži ne posodabljajo pogosto. Posledično modeli v praktični uporabi nimajo implicitnega znanja npr. o dogodkih, novejših od časa zbiranja učnega korpusa.

Učenje velikih jezikovnih modelov je tipično sestavljeno iz dveh ločenih faz. V prvi fazi se model samonadzorovano uči jezikovnih vzorcev in pridiva splošno znanje preko z nalogami, kot je nadaljevanje besedil, prek velikih besedilnih podatkovnih zbirk [11]. V sledeči fazi se modeli prek mehanizmov, kot so npr. učenje oponašanja obstoječih modelov [20], ojačevalno učenje človeških preferenc [1], in ustavno učenje s povratno informacijo [2] učijo želenih vzorcev obnašanja, kot so sledenje človeškim navodilom v naravnem jeziku, podajanje resničnih informacij, in izogibanje škodljivemu vedenju. Primarni namen te faze učenja je ustrezno vplivati na izhodno obnašanje modelov s čim manjšim vplivom na implicitno znanje jezikovnega modela, pridobljeno v začetni fazi učenja jezikovnega modeliranja. Ker tega ni mogoče zagotoviti v celoti, pa vedno obstaja kompromis med vsiljevanjem želenih vzorcev obnašanja in ohranjanjem naučenega znanja. Pri vrednotenju znanja in sposobnosti velikih jezikovnih modelov se zato osredotočamo na modele, ki nimajo na ta način vsiljenih vzorcev obnašanja, torej na modele, ki so bili učeni samo z nalogami jezikovnega modeliranja. Na takih modelih je mogoče bolj objektivno vrednotenje naučenega znanja brez vpliva vsiljenih vzorcev obnašanja [24].

Besedilne podatkovne zbirke, uporabljene za učenje največjih jezikovnih modelov, se stalno povečujejo in zajemajo že znaten delež javno dostopnega interneta [16]. To povzroča težave pri vrednotenju sposobnosti teh jezikovnih modelov, saj je zaradi avtomatiziranih postopkov zbiranja učnih podatkov vse težje zagotoviti, da podatki za naloge, ki se uporabljajo kot standardna merila uspešnosti različnih jezikovnih sposobnosti [13], [23], [7], niso vsebovani v učnem korpusu jezikovnega modela.

Veliki jezikovni modeli s povečevanjem modelov in njihovih učnih podatkovnih zbirk pridobijo nove, prej neobstoječe sposobnosti [28], kot so modularna aritmetika, reševanje besedilnih matematičnih nalog, in aritmetične operacije nad velikimi števili, predstavljenimi z nizi znakov. Uspešnost pri matematičnih nalogah lahko torej služi kot pomembno merilo sposobnosti velikih jezikovnih modelov. Za vrednotenje jezikovnih modelov imajo matematične naloge prednost pred jezikovnimi, saj jih lahko ustvarjamo samodejno v velikih količinah, odgovore lažje samodejno in objektivno vrednotimo, zaradi njihove kombinatorične narave prek naključnega vzorčenja pa lahko z veliko gotovostjo trdimo, da nalog ni v katerem izmed velikih učnih korpusov spletnih besedil za jezikovne modele.

Ena izmed ugotovljenih prednosti povečevanja jezikovnih modelov je omogočanje njihovega meta-učenja iz vhodnega konteksta [4]. Meta-učenje se tipično izvaja tako, da jezikovnemu modelu v vhodnem kontekstu poleg želene naloge podamo več primerov rešene naloge. Pri nalogi prevajanja med dvema jezikoma se tako pred besedilom, ki naj bi ga model prevedel, v vhodni kontekst doda več parov že prevedenih povedi v želenem paru jezikov. Glavna ugotovitev tu je, da imajo večji jezikovni modeli večjo korist od količine rešenih primerov v vhodnem kontekstu.

Pri tem pa se še vedno zastavlja vprašanje, ali gre za dejansko učenje iz konteksta oziroma, ali podajanje rešenih primerov v vhodnem kontekstu služi samo namenu izzvati t. i. latentno znanje jezikovnega modela [6].

Da bi razrešili to vprašanje in obenem izboljšali standarde objektivnega vrednotenja zmožnosti velikih jezikovnih modelov, v članku preizkusimo njihovo delovanje v času sklepanja (tj. brez dodatnega učenja samih jezikovnih modelov) na klasičnih nalogah strojnega učenja, kot so razvrščanje, regresija, in generativno modeliranje. Glavni prispevki članka so torej:

1) razvoj protokola za podajanje nalog strojnega učenja v vhodnem kontekstu jezikovnih modelov;
2) vrednotenje odprtokodnih jezikovnih modelov pri nalogah strojnega učenja, in;
3) dokaz, da so veliki jezikovni modeli v času sklepanja sposobni reševanja med učenjem nevidenih kompleksnih nelinearnih problemov.

2 METODOLOGIJA

Jezikovni modeli. Za sledeče eksperimente smo uporabili družino jezikovnih modelov RWKV4 [19]. Gre za prosto dostopne jezikovne modele velikosti od 1.7×10^8 do 1.4×10^{10} parametrov, torej tipičnega razreda velikosti modernih velikih jezikovnih modelov. Modeli so na voljo z utežmi, kvantiziranimi na $32-$, $16-$, oz. $8-$bitno natančnost, kar ponuja različne nivoje kompromisa med računsko zahtevnostjo zaganjanja modelov in njihovo natančnostjo. V sledečih eksperimentih uporabljamo modele s $16-$bitnimi utežmi, ki so na uporabljeni strojni opremi približno $1.7\times$ hitrejši od $32-$bitnih, z zanemarljivim padcem natančnosti.

Modeli so bili učeni na korpusu besedil dolžine $\approx 10^{12}$ besed (0.78×10^12 žetonov) z uporabo vhodnega konteksta velikosti 4096 oz. 8192 žetonov, odvisno od modela. Velikost vhodnega konteksta je prvovrstno arhitekturna odločitev, ki vpliva na računsko težavnost učenja. Med učenjem nadaljevanja besedila optimizator kot vhodni kontekst naključno vzorči N žetonov besedila učnega korpusa, kjer je N manjši ali enak maksimalni velikosti konteksta modela. Ciljni izhod modela je nato žeton, ki sledi zadnjemu vzorčenemu.

Odprtokodna izdaja uteži omenjenih modelov izpolnjuje našo zgoraj podano zahtevo, da želimo imeti poleg modelov z vsiljenimi vzorci vedenja na voljo tudi uteži, naučene samo s posrednimi nalogami jezikovnega modeliranja - konkretni modeli so bili naučeni z nalogo nadaljevanja naključno odrezanih besedil glede na prejšnji kontekst naključno izbrane dolžine. Družina modelov RWKV4 pri tej nalogi glede na testno zbirko doseže primerljive rezultate z najboljšimi objavljenimi jezikovnimi modeli primerljivih velikosti v smislu števila parametrov in računske zahtevnosti.

Podajanje značilk in oznak vzorcev. Za preizkušanje sposobnosti jezikovnih modelov opravljati naloge strojnega učenja smo razvili protokol za podajanje značilk in oznak vzorcev. Predpostavljamo, da so vzorci v osnovi predstavljeni v numerični obliki, pri problemih razvrščanja npr. z vektorji značilk in številsko oznako razreda. V tej obliki pa jih ne moremo neposredno podati kot vhodni kontekst jezikovnim modelom, ker ti vhode sprejemajo v obliki zaporedja žetonov, ki prek kodiranja parov bajtov [10] predstavljajo vhodno besedilo.

Zahteve protokola podajanja nalog strojnega učenja v vhodnem kontekstu jezikovnih modelov so dvojne: učne primere moramo biti sposobni

1) podati na jezikovnim modelom razumljiv način, tj. na način, ki omogoča njihovo opravljanje zahtevanih nalog; ter obenem
2) na način, ki zagotovo ni vsebovan v velikih podatkovnih zbirkah besedil s spletnih strani, kot so npr. korpusi Pile [11] oz. CommonCrawl [18].

Na primeru učne zbirke vzorcev za razvrščanje naj bodo značilke vzorcev podane z vektorji $\mathbf{x} \in \mathbb{R}^d$ in njihove oznake s števili $y \in [0, N-1] \subset \mathbb{N}$, pri čemer d predstavlja razsežnost vektorjev značilk in N število razredov vzorcev. V tem primeru lahko značilke in pripadajoče oznake razredov podamo v vhodni kontekst z uporabo z vejicami ločenih decimalnih in celih števil, kot npr.:

```
x: 0.25, -0.86, -1.67, -1.21, y: 0
 x: -1.35, 1.01, -0.39, 0.21, y: 1
 x: -1.74, 0.78, -0.90, 1.18, y: 2
```

To izpolni prvi pogoj, ne pa drugega. Zapis značilk in oznak razredov vzorcev namreč ostane nespremenjen v primerjavi z izvorno obliko zapisa podatkovnih zbirk (npr. v datotekah csv), zato obstaja možnost, da se podatkovna zbirka v taki obliki v učnem korpusu jezikovnega modela že nahaja. Da se temu izognemo značilke vzorcev transformiramo s kvantizacijo na dvomestna cela števila, tj. s preslikavo

$$\tilde{\mathbf{x}} = \lfloor a\mathbf{x} + b \rceil \,, \tag{1}$$

pri čemer $\lfloor \cdot \rceil$ predstavlja operacijo zaokroževanja na najbližje celo število, a in b pa sta skalarja, izbrana glede na definicijsko območje značilk vzorcev, tako da velja $\tilde{x}_i \in \tilde{\mathbf{x}} \in [0, 99]$. Kvantizacijo na dvomestna cela števila izberemo, ker je znan rezultat iz literature [28], da so tudi najmanjši splošni jezikovni modeli sposobni osnovnih aritmetičnih operacij z dvomestnimi decimalnimi števili, medtem, ko na daljših številih delujejo slabše.

Razvrščanje vzorcev iz testne zbirke nato rešimo tako, da v vhodni kontekst modela podamo na ta način transformirano učno zbirko in značilke enega izmed vzorcev testne zbirke, npr.:

```
x: 58, 71, 93, 58, y: 0
x: 53, 23, 81, 62, y: 1
x: 31, 46, 62, 29, y: 2
x: 35, 94, 10, 91, y:
```

Z dovolj obsežno učno zbirko vzorcev pričakujemo, da bo prvi znak, ki ga model odda na izhodu, predvidena oznaka testnega vzorca na zadnji vrstici. Metoda je nekoliko zamudna, saj je pri tem treba izvesti sklepanje za vsak testni vzorec posebej, obenem pa mora pri vsakem koraku sklepanja model kot vhodni kontekst sprejeti celotno označeno učno zbirko vzorcev.

Preizkusili smo tudi predstavitev značilk vzorcev z neštevilskimi znaki (npr. z velikimi tiskanimi črkami angleške abecede, z naborom znakov ASCII), in dobili neuspešne rezultate, prikazane v tabeli 1. Rezultat kaže, da obravnavani veliki jezikovni modeli na podlagi modeliranja jezikovnega znanja vsebujejo implicitne predstavitve številskih vrednosti, in obenem, da njihov učni korpus - glede na skoraj popolno natančnost pri razvrščanju

Tabela 1: Uspešnost razvrščanja vzorcev zbirke IRIS pri različnih metodah kvantizacije vrednosti značilk (uspešnost naključnega razvrščevalnika: 33.3 %).

Metoda	Uspešnost
Izvorne številske vrednosti	98.7 %
Kvantizacija na dvomestna cela števila	73.3 %
Kvantizacija na črke angleške abecede	32 %
Kvantizacija na 100 naključnih žetonov	34.7 %

izvornih vrednosti verjetno vsebuje podatkovno zbirko IRIS.

Za namene generativnega modeliranja obrnemo vrstni red značilk in oznak razredov vzorcev, tako, da kot vhodni kontekst jezikovnemu modelu podamo npr.:

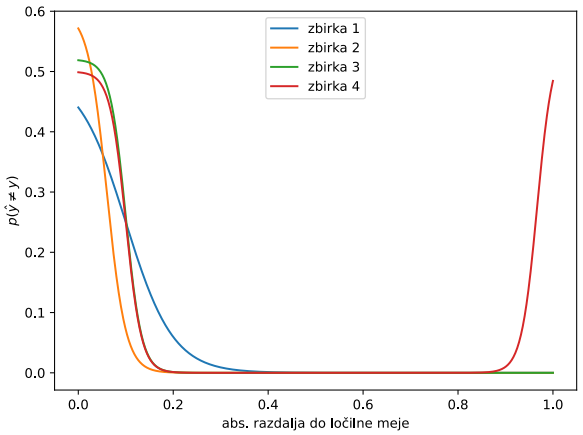
```
y: 0, x: 58, 71, 93, 58
y: 1, x: 53, 23, 81, 62
y: 2, x: 31, 46, 62, 29
y: 1, x:
```

Pri tem kot izhod modela pričakujemo porazdelitveno smiseln vektor značilk glede na podano oznako razreda, podobno, kot pri razredno pogojenih generativnih nasprotniških omrežjih [12].

3 Rezultati

Postopkovno ustvarjene podatkovne zbirke. Za preverjanje smiselnosti pristopa preizkusimo jezikovne modele na problemu dvojiškega razvrščanja postopkovno ustvarjenih podatkovnih zbirk. Učne podatkovne zbirke ustvarimo tako, da vhodne značilke podatkov vzorčimo s porazdelitve $\mathbf{x} \in \mathbb{R}^2 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, oznake razredov $y \in \{0, 1\}$ pa jim določimo glede na njihovo geometrijsko postavitev v prostoru $\mathbb{R}^2$ po pravilih, prikazanih v sliki 2 (levo). Nekatere izmed zbirk, ustvarjenih po temu postopku za uspešno razvrščanje zahtevajo tudi modeliranje nelinearnih ločilnih mej, medtem, ko zbirka 4 kot najlažji primer omogoča popolno razvrstitev z linearno ločilno mejo. V vseh primerih so vzorci enakomerno razporejeni med razredoma 0 in 1, pri naključnem ugibanju je pričakovana uspešnost razvrščanja torej 50 %.

Kot osnovo za primerjavo uspešnosti razvrščanja velikih jezikovnih modelov uporabimo razvrščanje z dvema klasičnima pristopoma razpoznavanja vzorcev, tj. z metodo podpornih vektorjev (angl. SVM), in metodo prileganja najbližjih sosedov (angl. kNN). Rezultati so prikazani v tabeli 2. Iz rezultatov je razvidno, da se največji obravnavani jezikovni model približa uspešnosti klasičnih pristopov strojnega učenja, vsi jezikovni modeli pa delujejo bistveno bolje od razvrščanja z naključnim ugibanjem. To pomeni, da so modeli sposobni



Slika 1: Odvisnost med razdaljo vzorca do ločilne meje in verjetnostjo njegove napačne razvrstitve.

Tabela 2: Uspešnost na postopkovno ustvarjenih podatkovnih zbirkah (uspešnost naključnega razvrščevalnika: 50 %).

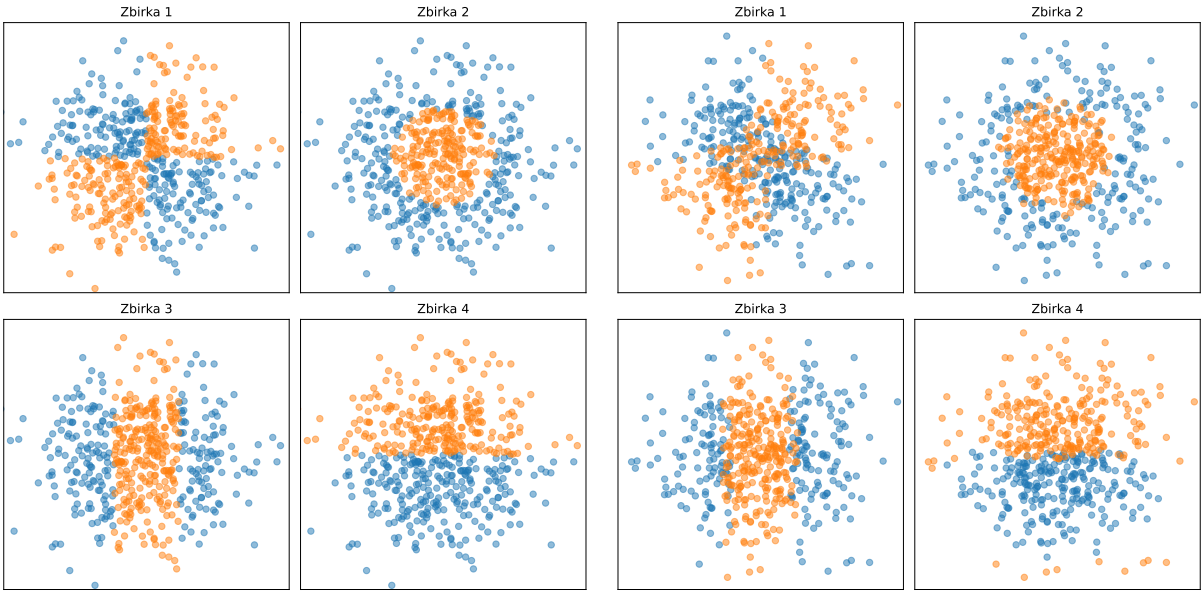
Metoda Zbirka	1	2	3	4
SVM	92 %	99 %	97 %	98 %
kNN	90 %	99 %	96 %	97 %
RWKV4-1.5B	61 %	69 %	73 %	78 %
RWKV4-3B	58 %	72 %	78 %	75 %
RWKV4-7B	78 %	84 %	88 %	91 %
RWKV4-14B	89 %	96 %	92 %	94 %

modeliranja odvisnosti med značilkami vzorcev in dodeljenimi oznakami in, da je izbrani pristop vrednotenja smiseln.

Iz grafičnih rezultatov v sliki 2 (desno) je razvidno, da se model RWKV4-14B nauči linearnih in nelinearnih geometrijskih vzorcev v dvorazsežnem prostoru vhodnih značilk, do napak pride večinoma blizu ločilne meje, kjer model kaže majhno gotovost. Slika 1 prikazuje odvisnost med oddaljenostjo vzorca od ločilne meje in verjetnostjo njegove napačne razvrstitve modela RWKV4-14B. Ta pri prvih treh zbirkah strogo pada z razdaljo od ločilne meje, pri zbirki 4 pa se model nauči vse modele daleč od koordinatnega izhodišča razvrstiti v enega izmed razredov ne glede na njihove oznake, kot je prikazano na sliki 2 (desno, zbirka 4).

Poleg tega naj omenimo še, da vsi jezikovni modeli v vseh testnih primerih vrnejo smiselne izhode, torej znak `0` oz. `1`. Rezultati kažejo, da so se veliki jezikovni modeli sposobni naučiti tako linearnih kot nelinearnih vzorcev odvisnosti med značilkami in razrednimi oznakami v vhodnem kontekstu.

Podatkovna zbirka IRIS [9]. Po postopku, opisanem v sekciji 2, preizkusimo razvrščanje in generativno modeliranje na podatkovni zbirki IRIS - klasični testni zbirki za preizkušanje pristopov strojnega učenja. Gre za zbirko meritev dolžin in širin cvetnih listov



Slika 2: Levo: Postopkovno ustvarjene podatkovne zbirke vzorcev za binarno razvrščanje, prikazane v prostoru $\mathbb{R}^2$. Desno: rezultati razvrščanja z modelom RWKV4-14B. Samo zbirka 4 omogoča razvrščanje z linearno ločilno mejo.

treh različnih vrst perunik. Podatkovna zbirka je sestavljena iz 150 vzorcev, ki so predstavljeni vsak s štirimi meritvami - realnimi vrednostmi. Vzorci zbirke so enakomerno porazdeljeni med 3 razrede, pričakovana uspešnost razvrščanja pri naključnem ugibanju je torej 33.3 %. Podatkovno zbirko lahko torej enako kot v prejšnjem primeru predstavimo z vzorci $\mathbf{x} \in \mathbb{R}^d$, pri čemer je $d = 4$, in oznakami $y \in [0, N-1] \subset \mathbb{N}$, pri čemer je $N = 3$. Pri eksperimentu polovico vzorcev vsakega razreda uporabimo kot učne vzorce, preostale vzorce pa uporabimo za preizkušanje uspešnosti.

Za vrednotenje sposobnosti razvrščanja velike jezikovne modele primerjamo z metodo prileganja najbližjih sosedov, ter metodo podpornih vektorjev. Za vrednotenje sposobnosti generativnega modeliranja jezikovne modele primerjamo z modelom Gaussovih mešanic (angl. GMM), ter z razredno pogojenim generativnim nasprotniškim omrežjem (angl. GAN). Z vsakim izmed generativnih modelov generiramo množico vzorcev enake velikosti, kot je originalna zbirka podatkov. Nato izračunamo njen histogram v prostoru $\mathbb{R}^4$ ter uspešnost generativnega modeliranja vrednotimo kot razdaljo χ^2 med histogramom izvornih oz. generiranih vzorcev danega razreda,

$$\chi^2(X, Y) = \sum_{i=1}^{N} \frac{(x_i - y_i)^2}{x_i + y_i}, \tag{2}$$

pri čemer je X histogram porazdelitve izvornih vzorcev, Y histogram porazdelitve generiranih vzorcev, in i predstavlja indeks bin-ov v histogram, ter x_i in y_i pa število vnosov v i-ti bin. Pri tem boljše ujemanje porazdelitve generiranih vzorcev z dejanskimi pomeni manjšo razdaljo χ^2. Rezultati razvrščanja v tabeli 3 kažejo, da so jezikovni modeli sposobni učenja iz vhodnega konteksta tudi na tem, realnem problemu, čeprav se po uspešnosti ne približajo klasičnim metodam strojnega učenja. Podobno kot v prejšnjem primeru tudi pri večrazrednem razvrščanju vsi modeli v 100 % testnih primerov vrnejo smiselne odgovore, torej cela števila z intervala $[0, N-1]$. Tudi rezultati generativnega modeliranja so pozitivni, saj kažejo, da jezikovni modeli na problemu učenja porazdelitve vzorcev iz konteksta delujejo bolje, kot če vzorcem vseh razredov priredimo skupno večrazsežno normalno porazdelitev z diagonalno kovariančno matriko, in da se obenem z velikostjo jezikovnih modelov njihova sposobnost generativnega modeliranja izboljšuje.

Tabela 3: Rezultati na podatkovni zbirki IRIS (uspešnost naključnega razvrščevalnika: 33.3 %; razdalja χ^2 do enorazredne diagonalne normalne porazdelitve: 0.0997).

Metoda	Uspešnost razvrščanja	generativno modeliranje $[\chi^2]$
SVM	93.3 %	-
kNN	92 %	-
GMM	-	0.0420
GAN	-	0.0181
RWKV4-1.5B	32 %	0.1873
RWKV4-3B	41.3 %	0.1455
RWKV4-7B	65.3 %	0.0812
RWKV4-14B	73.3 %	0.0661

Regresija. Pri problemu regresije za razliko od razvrščanja na podlagi učnih vzorcev napovedujemo

zvezne številske vrednosti. Tipična predpostavka je, da realne številske oznake učnih vzorcev predstavljajo funkcijsko vrednost, pokvarjeno z virom šuma. Za preizkus sposobnosti velikih jezikovnih modelov učenja regresije se omejimo na napovedovanje vrednosti skalarnih funkcij realnih vrednosti, tj. funkcij tipa $f : \mathbb{R} \mapsto \mathbb{R}$, pri čemer nas zanima sposobnost velikih jezikovnih modelov glede na vhodno-izhodne vzorce $(x, f(x))$ v vhodnem kontekstu

1) interpolirati med učnimi točkami; in
2) ekstrapolirati na širše definicijsko območje.

Za primerjavo z uspešnostjo velikih jezikovnih modelov tu uporabljamo metodo najmanjših kvadratov s polinomskim modelom iste stopnje, kot je dejanska polinomska funkcija, ki jo modeliramo. Za linearno regresijo preizkusimo delovanje na funkciji

$$f_1(x) = -238x + 1475 + \epsilon, \tag{3}$$

za nelinearno regresijo pa na funkciji

$$f_2(x) = 4x^2 - 8x + 14 + \epsilon, \tag{4}$$

pri čemer so bili koeficienti izbrani naključno in ima šum $\epsilon \sim \mathcal{N}\left(0, \sigma^2\right)$ varianco, sorazmerno s koeficientom najvišje stopnje funkcij f_1 oz. f_2.

Za obe funkciji preizkusimo interpolacijo na definicijskem območju $x_{in} \in [-10, 10]$, na katero so omejeni učni pari vzorcev ter pripadajočih funkcijskih vrednosti. Zunaj definicijskega območja učnih vzorcev preizkusimo ekstrapolacijo na območju $x_{ex} \in [-15, 15]$. Rezultati poskusov so prikazani na sliki 3. Kakovost regresije kvantitativno ocenimo prek korena srednje kvadratne napake med napovedmi modelov in vrednostmi dejanskih funkcij (3) oz. (4). Rezultati mere RMSE so podani v tabeli 4. Jezikovni model RWKV4-14B doseže bistveno slabše prileganje pravi funkciji od metode najmanjših kvadratov, iz slik pa je razvidno, da jezikovni kljub temu uspe modelirati tako linearno kot nelinearno odvisnost med vhodno in izhodno spremenljivko. Razvidno je, da imajo napovedi modela za razliko od metode najmanjših kvadratov raztros, primerljiv z varianco šuma učnih primerov. Kljub raztrosu in večjemu odstopanju pri ekstrapolaciji pa je razvidno tudi, da se ekstrapolirane vrednosti smiselno navezujejo na vrednosti z definicijskega območja učnih točk in niso npr. izbrane naključno. Omenimo naj še, da manjši jezikovni modeli RWKV4 funkcij f_1 oz. f_2 niso sposobni modelirati niti v definicijskem območju učnih vzorcev (tj. v interpolacijskem režimu), kar kaže, da je regresija realnih skalarnih funkcij porajajoča zmožnost velikih jezikovnih modelov, sorodna tistim, odkritim v študiji [28].

Tabela 4: Kvantitativni rezultati poizkusov regresije.

Metoda in funkcija	RMSE, x_{in}	RMSE, x_{ex}
Najmanjši kvadrati, f_1	412.3	472.6
RWKV4-14B, f_1	425.0	613.5
Najmanjši kvadrati, f_2	109.7	122.5
RWKV4-14B, f_2	179.1	208.5

4 Pomen odprtokodnih jezikovnih modelov

Najnovejši in najzmogljivejši veliki jezikovni modeli, kot je GPT-4 [16] niso odprtokodni, ampak so končnim uporabnikom dostopni bodisi prek aplikacijskega programskega vmesnika, bodisi prek aplikacij tipa ChatGPT. V slednjih so tipično na voljo samo modeli z vsiljenimi vzorci obnašanja prek ojačevalnega učenja človeških preferenc, ki pa za eksperimente iz prejšnje sekcije niso primerni. Glede na objavljene cene API-dostopa podjetja OpenAI v juliju 2023, tj. $\$0.06/1000$ vhodnih žetonov in $\$0.12/1000$ izhodnih žetonov, ceno ponovitve eksperimentov iz prejšnje sekcije samo na modelu GPT-4 ocenjujemo na 7200 €. V primerjavi s tem moremo uporabljene odprtokodne in prosto dostopne jezikovne modele družine RWKV4 zaganjati na potrošniški delovni postaji, kupljeni leta 2017, vsi eksperimenti pa so bili izvedeni v okviru 110 GPU-ur, kar prinaša zanemarljive dodatne stroške. To kaže na izrazit pomen odprtokodnih in prostodostopnih jezikovnih modelov, saj širši akademski skupnosti omogočajo obsežnejše eksperimentiranje, to pa vodi k boljšemu razumevanju delovanja in hitrejšemu razvoju sposobnejših jezikovnih modelov.

5 Zaključki

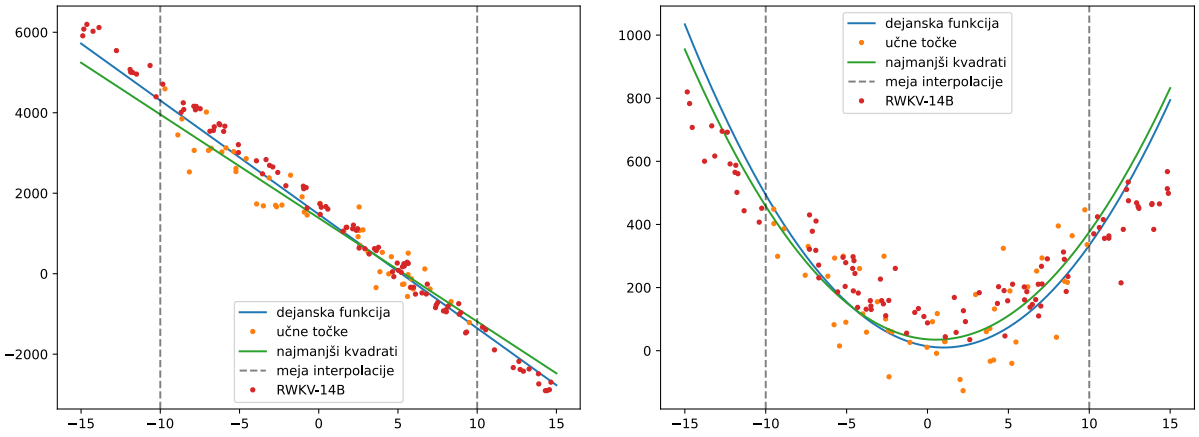
Sposobnosti velikih jezikovnih modelov smo ovrednotili z nalogami strojnega učenja, za katere obstajajo močna zagotovila, da niso prisotne v nobenem izmed večjih učnih korpusov besedil. Rezultati kažejo, da imajo veliki jezikovni modeli sposobnost učiti se kompleksnih nelinearnih modelov podatkov, podanih prek vhodnih kontekstov, brez posodobitev parametrov modela, in ne delujejo zgolj kot "stohastični papagaji" [3]. Posledično ocenjujemo, da imajo naloge strojnega učenja in sorodne naloge matematičnega modeliranja močen potencial kot metode objektivnega vrednotenja sposobnosti velikih jezikovnih modelov.

Literatura

[1] Y. Bai, A. Jones, K. Ndousse, A. Askell, A. Chen, N. DasSarma, D. Drain, S. Fort, D. Ganguli, T. Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

[2] Y. Bai, S. Kadavath, S. Kundu, A. Askell, J. Kernion, A. Jones, A. Chen, A. Goldie, A. Mirhoseini, C. McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.

[3] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623, 2021.



Slika 3: Rezultati regresije na linearnem (levo) oz. nelinearnem problemu (desno).

[4] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[5] S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.

[6] C. Burns, H. Ye, D. Klein, and J. Steinhardt. Discovering latent knowledge in language models without supervision. *arXiv preprint arXiv:2212.03827*, 2022.

[7] M. Chen, J. Tworek, H. Jun, Q. Yuan, H. P. d. O. Pinto, J. Kaplan, H. Edwards, Y. Burda, N. Joseph, G. Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

[8] P. Clark, I. Cowhey, O. Etzioni, T. Khot, A. Sabharwal, C. Schoenick, and O. Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

[9] R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of eugenics*, 7(2):179–188, 1936.

[10] P. Gage. A new algorithm for data compression. *C Users Journal*, 12(2):23–38, 1994.

[11] L. Gao, S. Biderman, S. Black, L. Golding, T. Hoppe, C. Foster, J. Phang, H. He, A. Thite, N. Nabeshima, et al. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.

[12] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[13] D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

[14] V. Kocijan, E. Davis, T. Lukasiewicz, G. Marcus, and L. Morgenstern. The defeat of the winograd schema challenge. *Artificial Intelligence*, page 103971, 2023.

[15] N. Mostafazadeh, N. Chambers, X. He, D. Parikh, D. Batra, L. Vanderwende, P. Kohli, and J. Allen. A corpus and evaluation framework for deeper understanding of commonsense stories. *arXiv preprint arXiv:1604.01696*, 2016.

[16] OpenAI. Gpt-4 technical report, 2023.

[17] D. Paperno, G. Kruszewski, A. Lazaridou, Q. N. Pham, R. Bernardi, S. Pezzelle, M. Baroni, G. Boleda, and R. Fernández. The lambada dataset: Word prediction requiring a broad discourse context. *arXiv preprint arXiv:1606.06031*, 2016.

[18] J. M. Patel. *Introduction to Common Crawl Datasets*, pages 277–324. Apress, Berkeley, CA, 2020.

[19] B. Peng, E. Alcaide, Q. Anthony, A. Albalak, S. Arcadinho, H. Cao, X. Cheng, M. Chung, M. Grella, K. K. GV, et al. Rwkv: Reinventing rnns for the transformer era. *arXiv preprint arXiv:2305.13048*, 2023.

[20] B. Peng, C. Li, P. He, M. Galley, and J. Gao. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*, 2023.

[21] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever. Improving language understanding with unsupervised learning. *Technical report, OpenAI*, 2018.

[22] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[23] K. Sakaguchi, R. L. Bras, C. Bhagavatula, and Y. Choi. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106, 2021.

[24] K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, S. Pfohl, et al. Large language models encode clinical knowledge. *arXiv preprint arXiv:2212.13138*, 2022.

[25] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[26] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[27] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[28] J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, E. H. Chi, T. Hashimoto, O. Vinyals, P. Liang, J. Dean, and W. Fedus. Emergent abilities of large language models. *Transactions on Machine Learning Research*, 2022. Survey Certification.

Klemen Grm je leta 2020 doktoriral s področja elektrotehnike na Fakulteti za elektrotehniko Univerze v Ljubljani. Je asistent v Laboratoriju za strojno inteligenco na Fakulteti za elektrotehniko. Njegovo področje raziskav obsega strojno učenje, biometrijo, in obdelavo slik.