Veliki jezikovni modeli so strojni učenci v času sklepanja

Klemen Grm

Univerza v Ljubljani, Fakulteta za Elektrotehniko
E-pošta: klemen.grm@fe.uni-lj.si

Large language models are inference-time machine learners

A machine learner is an algorithm capable of learning from data to improve its 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, I 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, I conclusively show they're able to model complex relationships between data in the input context.

1 Uvod

Učenje velikih jezikovnih modelov tipično sestoji iz dveh ločenih faz: V prvi fazi se model uči jezikovnih vzorcev in splošnega znanja preko samonadzorovanega učenja na nalogah, kot so nadaljevanje besedil, preko velikih besedilnih podatkovnih zbirk [7]. V sledeči fazi se model preko mehanizmov, kot npr. ojačevalno učenje človeških preferenc [1] ali učenje oponašanja obstoječih modelov [13] učijo želenih vzorcev obnašanja, kot so na primer 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 brez, da bi bistveno spreminjali njihovo implicitno znanje, pridobljeno v začetni fazi učenja jezikovnega modeliranja. 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 možno bolj objektivno vrednotenje naučenega znanja brez vpliva vsiljenih vzorcev obnašanja [15].

Besedilne podatkovne zbirke, uporabljene za učenje največjih jezikovnih modelov, se stalno povečujejo in zajemajo že znaten delež javno dostopnega interneta [10]. To predstavlja problem pri vrednotenju sposobnosti teh jezikovnih modelov, ker je zaradi avtomatiziranih postopkov zbiranja učnih podatkov vedno težje zagotoviti, da podatki za naloge, ki se uporabljajo kot standardna merila uspešnosti različnih jezikovnih sposobnosti [9, 14, 5] niso vsebovani v učnem korpusu jezikovnega modela.

Veliki jezikovni modeli s povečevanjem modelov in njihovih učnih podatkovnih zbirk pridobijo nove, prej neobstoječe sposobnosti [16], kot npr. modularna aritmetika, reševanje besedilnih matematičnih nalog, in aritmetične operacije nad velikimi števili, predstavljenimi z nizi znakov. Uspešnost na matematičnih nalogah more torej služiti kot pomembno merilo sposobnosti velikih jezikovnih modelov. Za vrednotenje jezikovnih modelov imajo matematične naloge te prednosti pred jezikovnimi, da jih moremo ustvarjati samodejno v velikih količinah, da moremo odgovore lažje samodejno in objektivno vrednotiti, in da moremo vsled njihove kombinatorične narave preko naključnega vzorčenja z visoko ravnjo gotovosti trditi, da se naloge ne nahajajo 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 [3]. Glavna ugotovitev je, da imajo večji modeli pri nalogah, kot je strojno prevajanje, veliko korist od tega, da v vhodnem kontekstu poleg besedila, ki ga želimo prevesti, podamo več rešenih primerov v želenem paru jezikov. Pri tem pa še vedno obstaja vprašanje, če gre za dejansko učenje iz konteksta, oziroma, če primeri v vhodnem kontekstu služijo samo namenu izzvati latentno znanje jezikovnega modela [4].

Da bi razrešil to vprašanje in obenem izboljšal standarde objektivnega vrednotenja zmožnosti velikih jezikovnih modelov, v pričujočem prispevku preizkusim njihovo delovanje na klasičnih nalogah strojnega učenja, kot so razvrščanje, regresija in generativno modeliranje. Glavni doprinosi prispevka so torej:

1. Razvoj protokola za podajanje nalog strojnega učenja v vhodnem kontekstu jezikovnih modelov,
2. vrednotenje odprtokodnih jezikovnih modelov na 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 namene sledečih eksperimentov uporabim družino jezikovnih modelov RWKV4 [12]. Gre za prosto dostopne jezikovne modele velikosti od 1.7×10^8 do 1.4×10^{10} parametrov, učene na korpusu besedil dolžine $\approx 10^{12}$ besed, z vhodnim kontekstom velikosti 4096 oz. 8192 žetonov, odvisno od modela.

Podajanje značilk in oznak vzorcev. Zahteve protokola podajanja nalog strojnega učenja v vhodnem kontekstu jezikovnih modelov so dvojne: učne primere podati na jezikovnim modelom razumljiv način, ter obenem na način, ki zagotovo ni vsebovan v velikih podatkovnih zbirkah besedil s spletnih strani, kot so npr. korpusi Pile [7] oz. CommonCrawl [11].

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 temu primeru moremo značilke in pripadajoče oznake razredov podati 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
```

kar 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), in obstaja možnost, da se podatkovna zbirka v taki obliki že nahaja v učnem korpusu jezikovnega modela. V izogib temu značilke vzorcev transformiram 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 izberem, ker je znan rezultat iz literature [16], da so tudi najmanjši splošni jezikovni modeli sposobni osnovnih aritmetičnih operacij na dvomestnih decimalnih številih, medtem, ko na daljših številih delujejo slabše. Razvrščanje vzorcev iz testne zbirke nato rešim tako, da v vhodni kontekst modela podam 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:
```

Tabela 1: 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%

Z dovolj obsežno učno zbirko vzorcev pričakujem, da bo prvi znak, ki ga model odda na izhodu, predvidena oznaka testnega vzorca na zadnji vrstici. Za namene generativnega modeliranja obrnem vrstni red značilk in oznak razredov vzorcev, tako, da kot vhodni kontekst jezikovnemu modelu podam npr.:

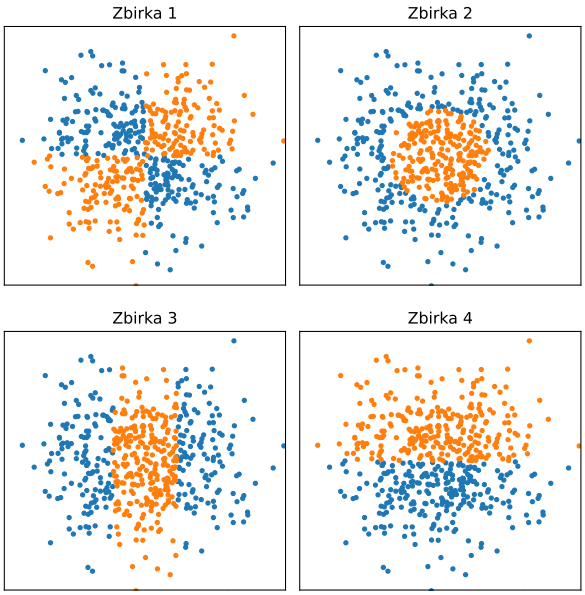
```
y:  0, x:  58, 71, 93, 58
y:  1, x:  53, 23, 81, 62
y:  2, x:  31, 46, 62, 29
y:  1, x:
```

Pri čemer kot izhod modela pričakujem porazdelitveno smiseln vektor značilk glede na podano oznako razreda, podobno, kot pri razredno pogojenih generativnih nasprotniških omrežjih [8].

3 Rezultati

Postopkovno ustvarjene podatkovne zbirke. Za namene preverjanja smiselnosti pristopa preizkusim jezikovne modele na problemu dvojiškega razvrščanja postopkovno ustvarjenih podatkovnih zbirk. Učne podatkovne zbirke ustvarim tako, da značilke vzorcev vzorčim s porazdelitve $\mathbf{x} \in \mathbb{R}^2 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, oznake razredov $y \in \{0, 1\}$ pa jim določim glede na vzorce, prikazane v sliki 1. Kot osnovo za primerjavo uspešnosti razvrščanja jezikovnih modelov uporabim razvrščanje z metodama podpornih vektorjev in prileganja najbližjih sosedov. Rezultati so prikazani v tabeli 1. Iz rezultatov je razvidno, da se največji obravnavan jezikovni model približa uspešnosti klasičnih pristopov strojnega učenja, vsi jezikovni modeli pa delujejo bistveno bolje od razvrščanja z naključnim ugibanjem. Poleg tega omenim še, da vsi jezikovni modeli vrnejo v vseh testnih primerih smiselne izhode, torej znak `0` oz. `1`. Rezultati kažejo, da so se veliki jezikovni modeli sposobni naučiti vzorcev odvisnosti med značilkami in razrednimi oznakami v vhodnem kontekstu.

Podatkovna zbirka IRIS [6]. Po postopku, opisanem v sekciji 2, preizkusimo razvrščanje in generativno modeliranje na podatkovni zbirki IRIS, ki sestoji iz 150 4-razsežnih vzorcev, enakomerno porazdeljenih med 3 razrede. Za vrednotenje generativnega modeliranja jezikovne modele primerjamo z modelom Gaussovih mešanic,



Slika 1: Postopkovno ustvarjene podatkovne zbirke vzorcev za binarno razvrščanje, prikazane v prostoru $\mathbb{R}^2$. Samo zbirka 4 omogoča razvrščanje z linearno ločilno mejo.

Tabela 2: Rezultati na podatkovni zbirki IRIS (uspešnost naključnega razvrščevalnika: 33.3%; razdalja χ^2 do enorazredne diagonalne 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

Tabela 3: 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

ter z razredno pogojenim generativnim nasprotniškim omrežjem. Z vsakim izmed generativnih modelov generiramo množico vzorcev enake velikosti, kot 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. Rezultati razvrščanja v tabeli 2 kažejo na to, da so jezikovni modeli sposobni učenja iz vhodnega konteksta tudi na temu realnem problemu, kljub temu, da 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.

Regresija. Za preizkus učenja regresije se omejim na napovedovanje vrednosti funkcij tipa $f : \mathbb{R} \mapsto \mathbb{R}$, pri čemer me zanima sposobnost velikih jezikovnih modelov glede na vhodno-izhodne vzorce f v vhodnem kontekstu

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

Za primerjavo z uspešnostjo velikih jezikovnih modelov tu uporabljam metodo najmanjših kvadratov s polinomskim modelom ustrezne stopnje. Za linearno regresijo preizkusim delovanje na funkciji

$$f_1(x) = -238x + 1475 + \epsilon, \tag{2}$$

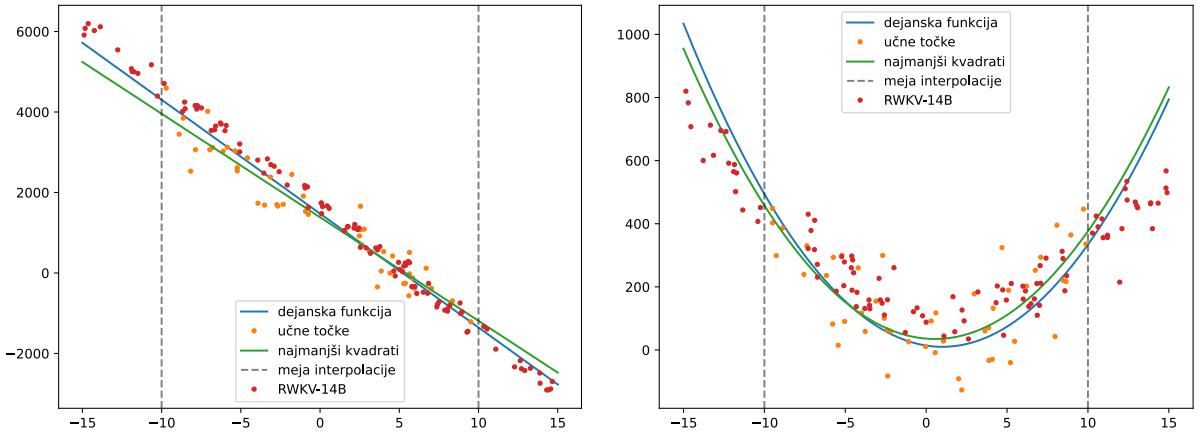
za nelinearno regresijo pa na funkciji

$$f_2(x) = 4x^2 - 8x + 14 + \epsilon, \tag{3}$$

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 preizkusim interpolacijo na definicijskem območju $x_{in} \in [-10, 10]$ ter ekstrapolacijo na $x_{ex} \in [-15, 15]$. Rezultati poizkusov so prikazani na sliki 2. Kvaliteto regresije kvantitativno ocenimo preko korena srednje kvadratne napake med napovedmi modelov in vrednostmi dejanske funkcije, ki so podane v tabeli 3. Jezikovni model RWKV4-14B doseže bistveno slabše prileganje pravi funkciji od metode najmanjših kvadratov, je pa iz slik razvidno, da kljub temu uspe modelirati tako linearno, kot nelinearno odvisnost med vhodno in izhodno spremenljivko. Kljub večjemu odstopanju pri ekstrapolaciji je razvidno, da se ekstrapolirane vrednosti smiselno navezujejo na vrednosti z definicijskega območja učnih točk, in niso npr. izbrane naključno.

4 Pomen odprtokodnih jezikovnih modelov

Najnovejši in najzmogljivejši veliki jezikovni modeli, kot na primer GPT-4 [10] niso odprtokodni, ampak so končnim uporabnikom dostopni bodisi preko aplikacijskega programskega vmesnika, bodisi preko aplikacij tipa ChatGPT. V slednjih so tipično na voljo samo modeli z vsiljenimi vzorci obnašanja preko ojačevalnega učenja človeških preferenc, ki pa za eksperimente s prejšnje sekcije niso primerni. Glede na objavljene cene API dostopa podjetja OpenAI v juliju 2023 ceno ponovitve eksperimentov iz prejšnje sekcije samo na modelu GPT-4 ocenjujem na



Slika 2: Rezultati regresije na linearnem (levo) oz. nelinearnem problemu (desno).

$7,000$€. V primerjavi s tem moremo uporabljene odprtokodne in prosto dostopne jezikovne modele družine RWKV4 zaganjati na potrošniški delovni postaji, kupljeni l. 2017, vsi eksperimenti pa so bili izvedeni v okviru 110 GPU-ur, kar predstavlja zanemarljive dodatne stroške. To kaže na izrazit pomen odprtokodnih in prosto dostopnih jezikovnih modelov, saj širši akademski skupnosti omogočajo obsežnejše eksperimentiranje, kar vodi k boljšemu razumevanju delovanja in hitrejšemu razvoju sposobnejših jezikovnih modelov.

5 Zaključki

Sposobnosti velikih jezikovnih modelov sem ovrednotil z nalogami strojnega učenja, za katere obstajajo močna zagotovila, da se ne nahajajo v nobenem izmed večjih učnih korpusov besedil. Rezultati kažejo na to, da imajo veliki jezikovni modeli sposobnost učiti se kompleksnih nelinearnih modelov podatkov, podanih v vhodnih kontekstih, in ne delujejo zgolj kot "stohastični papagaji" [2]. Naloge strojnega učenja in sorodne naloge matematičnega modeliranja imajo močen potencial kot metoda 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] 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.

[3] 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.

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

[5] 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.

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

[7] 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.

[8] 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.

[9] 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.

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

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

[12] 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.

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

[14] 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.

[15] 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.

[16] 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.