Benchmarking LLAMA-3 Large Language Model on Persian NLP

Faezeh Saghafi
Kazim Fouladi-Ghaleh

Deep Learning Research Lab
University of Tehran, Iran

11th International Conference on Web Research

Paper ID:

1568992

Peresenter:

Faezeh Saghafi

ICWR2025
April 16-17, 2025; Tehran, Iran

مقدمه

- مدل‌های زبانی بزرگ، گامی اساسی در جهت تحقق اهداف بنیان‌گذاران اولیه‌ی هوش مصنوعی همچون درک زبان، استدلال و تعامل طبیعی با ماشین‌ها هستند.
- پردازش زبان؛ گامی به سوی AGI
- مدل‌های زبانی، پایه فهم زبان
- ترانسفورمرها، افزایش توان محاسباتی و داده های بزرگ : لازمه ساخت مدل‌های زبانی بزرگ

مقدمه

- بهبود مدل‌ها، بدون Benchmarking ممکن نیست.
- مدل‌های پایه متن‌باز (open-source) در واقع ستون فقرات اکوسیستم هوش مصنوعی مدرن را تشکیل می‌دهند.
- دسترسی، شفافیت و تسهیل
- Llama, Mistral, Qwen, BERT

کارهای پیشین

پژوهشگران	مدل پایه	زبان	وظایف	سال
Baruni, Abaskohi, Masoudi, Abbasi, Babalou, Edalat, Kamahi, Mahdizadeh, Naghavian, Namazifard, Sadeghi, Yaghoobzadeh	GPT	فارسی	Sentiment Classification, Reading Comprehension, Translation, Emotion Classification, NER, Textual Entailment, Multiple-choice QA, Elementary School Questions, Mathematical Problems	۲۰۲۴
Khondaker, Abdul Waheed Nagoudi, Abdul-Mageed	GPT	عربی	Text Rewriting, Question-Gen, Summarization Paraphrase, Transliteration, MT (en→ar), MT (es→ar), MT (fr→ar), MT (ru→ar), CST (Jo-en→en)	۲۰۲۳
Khondaker, Naeem, Lyba Khan, A. Elmadany, Abdul-Mageed	Llama3-70B	عربی	Machine Translation, Code-Switching, Summarization, Title Generation, Question Answering, Question Generation, Paraphrase, Transliteration, Text Rewriting	۲۰۲۴

نوآوری

- زبان فارسی با وجود تعداد بالای گویشوران، به دلیل کمبود منابع باکیفیت و پیچیدگی‌های ساختاری و فرهنگی، چالش‌های خاصی را برای مدل‌های زبانی ایجاد می‌کند.
- چارچوب‌های ارزیابی رایج مانند HELM و BIG-Bench به‌دلیل تمرکز بر زبان‌های پرمنبع، برای ارزیابی مدل‌ها در زبان فارسی مناسب نیستند.
- بنابراین، طراحی بنچمارک‌های بومی و تخصصی برای ارزیابی دقیق‌تر عملکرد مدل‌ها در زبان فارسی ضروری است.

روش تحقیق

- مدل‌های مورد ارزیابی:

GPT 4 - GPT 3.5 turbo - Gemma 2 - Qwen 2 - Dorna – Maral – Ava - Aya - PersianMind

- وظایف مورد ارزیابی:

وظیفه	معیار	دیتاست
Sentiment Analysis	Exact Match (F1)	ParsiNLU
Emotion recognition	Exact Match (F1)	ArmanEmo
Machin translation	Bleu	mizan و ParsiNLU
Name entity recognition	Exact Match (F1)	ArmanNER
Reading comprehention	Common Tokens (F1)	ParsiNLU
Entailment	Exact Match (F1)	ParsiNLU و ConjNLI
Multiple Choice	Exact Match (Accuracy)	ParsiNLU
Elementary school	Math Equivalence (Accuracy	abbaskohi

11th International Conference on Web Research
ICWR2025
April 16-17, 2025; Tehran, Iran

خانواده مدل‌های Llama

- یک مجموعه از مدل‌های پایه هستند که توسط Meta منتشر شده است.
- متن‌باز هستند و در ۳ نسخه 8B و 70B و 405B در دسترس هستند.
- از معماری Transformer استاندارد با یک رمزگشا به‌همراه یک سری تعدیلات استفاده می‌کند.
- ۱۵ تریلیون توکن داده آموزشی

طراحی پرامپت

- عملکرد مدل‌های زبانی بزرگ به‌طور زیادی به طراحی دقیق و مؤثر پرامپت‌ها بستگی دارد.
- طراحی صحیح پرامپت می‌تواند کیفیت پاسخ‌ها را به میزان زیادی افزایش داده و دقت خروجی‌ها را بهبود بخشد.
- استفاده از تکنیک‌های Chain of Thought (COT) و Few-Shot Learning می‌تواند به بهینه‌سازی عملکرد مدل کمک کند.
- ترتیب چینش مثال‌ها نقش حیاتی در موفقیت مدل‌ها دارد و باید به‌صورت تدریجی از نمونه‌های ساده‌تر به پیچیده‌تر پیش برویم.
- یک پرامپت به‌خوبی طراحی‌شده می‌تواند به مدل در درک دقیق‌تر نیاز کاربر و ارائه اطلاعات مورد نظر کمک کند.

شرح وظیفه: استنتاج زبان طبیعی

جمله مقدم (Premise) و فرضیه (Hypothesis) زیر را با دقت بخوانید و رابطه بین آن‌ها را تعیین کنید.

یکی از سه دسته زیر را انتخاب کنید:

توضیحات برچسب‌ها:

Entailment (استنتاج): معنای فرضیه به‌صورت منطقی از جمله مقدم نتیجه‌گیری یا استنتاج می‌شود.

Contradiction (تناقض): معنای فرضیه با جمله مقدم در تضاد یا تناقض است.

Neutral (خنثی): هیچ رابطه منطقی واضحی بین جمله مقدم و فرضیه وجود ندارد.

توجه: جمله مقدم و فرضیه به زبان فارسی هستند.

الگوی مثال:

<جمله مقدم><جداساز><فرضیه><دسته‌بندی>

استنتاج (entailment) یا تناقض (contradiction) یا خنثی(neutral)

مثال:

<در این واکنش، سرعت هالوژناسیون مستقل از غلظت هالوژن است، اما به غلظت کتون و اسید بستگی دارد><.جداساز>

<در این واکنش، سرعت هالوژناسیون وابسته به غلظت هالوژن است، اما مستقل از غلظت کتون و اسید است>

11th International Conference on Web Research
ICWR2025
April 16-17, 2025; Tehran, Iran

یافته ها

		Gpt-3.5	Gpt-4	Qwen-2-7B	gemma-2-9b	Llama-3.2-11B	llama-3-8b	Dorna	Maral	aya	ava	PersianMind
classic	sa	0.791	0.856	0.385	0.413	0.671	0.526	0.455	0.346	0.386	0.41	0.206
	er	0.537	0.567	0.443	0.422	0.516	0.435	0.404	0.329	0.483	0.375	0.163
	ner	0.589	0.608	0.218	0.306	0.468	0.455	0.424	0.408	0.533	0.308	0.22
	Mt(en-fa)	0.343	0.377	0.229	0.304	0.301	0.282	0.265	0.366	0.318	0.26	0.296
	Mt(fa-en)	0.36	0.399	0.316	0.3	0.29	0.251	0.246	0.305	0.35	0.059	0.26
	rc	0.681	0.777	0.639	0.511	0.734	0.709	0.61	0.339	0.675	0.569	0.399
reasoning	Ent(parsnlu)	0.432	0.75	0.544	0.609	0.462	0.482	0.411	0.372	0.395	0.428	0.31
	Ent(conjnli)	0.366	0.828	0.446	0.739	0.409	0.342	0.408	0.577	0.467	0.217	0.315
	Qa(math&logic)	0.435	0.645	0.258	0.285	0.051	0.453	0.391	0.398	0.331	0.371	0
	Elem-school	0.565	0.665	0.537	0.489	0.534	0.519	0.523	0.429	0.469	0.477	0.393
knowledge	Qa(literature)	0.275	0.390	0.256	0.245	0.29	0.222	0.245	0.345	0.195	0.415	0
	Qa(common)	0.445	0.595	0.278	0.305	0.505	0.437	0.318	0.407	0.417	0.48	0

بحث و بررسی

۱. مدل‌های قدرتمندتر پاسخ های تمیزتر و واضح‌تری تولید می‌کنند، در حالی که مدل‌های ضعیف‌تر معمولاً با مشکلات شفافیت و دقت مواجه هستند.

۲. زمان اجرای مدل‌ها بستگی زیادی به قدرت آنها دارد.

۳. تولید پاسخ‌های بی‌ربط در مدلهای ضعیف‌تر

۴. تغییر زبان در طول پاسخ دادن به پرسش‌ها به‌طور مکرر اتفاق می‌افتاد.

۵. توجه به نوع وظیفه‌ای که مدل برا ی آن تنظیم شده، در ارزیابی عملکرد آن بسیار اهمیت دارد.

۶. مدل‌هایی که بر روی متون شبکه‌های اجتماعی آموزش دیده‌اند، در انجام وظایف کلاسیک مانند تحلیل احساسات و شناسایی موجودیت‌ها عملکرد بهتری دارند.

بحث و بررسی

۱. مدل LLaMA 3.2 نسبت به نسخه قبلی خود، قابلیت چندزبانه بودن بهتری دارد، اما هنوز از زبان فارسی به‌طور کامل پشتیبانی نمی‌کند.

۲. مدل‌های فارسی که با داده‌های فارسی فاین‌تیون شده‌اند، هنوز عملکردی پایین‌تر از مدل‌های پیشرفته‌تر دارند.

۳. مدل‌های عمومی با امتیازهای یکدست قدرت بالاتری در وظایف مختلف دارند، در حالی که مدل‌های تخصصی با امتیاز بالاتر در یک وظیفه خاص نشان‌دهنده بهینه‌سازی برای آن حوزه هستند.

۴. بررسی‌ها نشان داد که بهترین روش پرامپت‌دهی،
ترکیب تکنیک‌های few-shot و chain of thought (CoT) است.

نتیجه‌گیری

- در این پژوهش، عملکرد مدل‌های زبانی بزرگ در وظایف مختلف پردازش زبان طبیعی فارسی مقایسه شد. یافته‌ها نشان دادند که مدل‌های قوی‌تر مانند GPT-4 و LLaMA-3.2-11B در وظایف پیچیده و چندزبانه عملکرد برتری دارند. در مقابل، مدل‌های فاین‌تیون‌شده فارسی، به‌ویژه در وظایف مرتبط با فرهنگ و زبان فارسی، نتایج قابل قبولی ارائه دادند. محدودیت‌های سخت‌افزاری به‌عنوان مانعی جدی در گستردگی آزمایش‌ها شناسایی شدند. برای تحقیقات آینده، ارتقاء زیرساخت‌ها، طراحی بنچمارک‌های خاص برای زبان فارسی، و بهره‌گیری از داده‌های متنوع‌تر توصیه می‌شود. این رویکردها می‌توانند به توسعه مدل‌های زبانی مؤثرتر برای زبان فارسی کمک کنند.

پیشنهادها

۱. جمع‌آوری و تهیه داده‌های **با کیفیت** و **متنوع**

۲. تمرکز هر چه بیشتر سایر مدل‌های متن‌باز بر روی زبان‌های کم‌منبع مانند فارسی

۳. اهتمام هر چه بیشتر به‌منظور تهیه مستندات مدل

مراجع

1. Naveed, Humza, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. "A comprehensive overview of large language models." *arXiv preprint arXiv:2307.06435* (2023).
2. Chu, Zhibo, Shiwen Ni, Zichong Wang, Xi Feng, Chengming Li, Xiping Hu, Ruifeng Xu, Min Yang, and Wenbin Zhang. "History, Development, and Principles of Large Language Models-An Introductory Survey." *arXiv preprint arXiv:2402.06853* (2024).
3. Abaskohi, Amirhossein, Sara Baruni, Mostafa Masoudi, Nesa Abbasi, Mohammad Hadi Babalou, Ali Edalat, Sepehr Kamahi et al. "Benchmarking Large Language Models for Persian: A Preliminary Study Focusing on ChatGPT." *arXiv preprint arXiv:2404.02403* (2024).