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

- مدل‌های زبانی بزرگ، گامی اساسی در جهت تحقق اهداف بنیان‌گذاران اولیه‌ی هوش مصنوعی همچون درک زبان، استدلال و تعامل طبیعی با ماشین‌ها هستند.
- بهبود مدل‌ها، بدون ارزیابی ممکن نیست.
- مدل های پایه منبع-باز در واقع ستون فقرات اکوسیستم هوش مصنوعی مدرن را تشکیل می‌دهند.
- مدل زبانی llama از مهم‌ترین مدل‌های پایه است که به‌صورت منبع-باز منتشر شده است.
- زبان فارسی با وجود تعداد بالای گویشوران، به‌دلیل کمبود منابع باکیفیت و پیچیدگی‌های ساختاری و فرهنگی، چالش‌های خاصی را برای مدل‌های زبانی ایجاد می کند.

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

- ساختار پرامپت:

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

یافته‌ها

		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

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

بحث و بررسی

۱. مدل‌های قدرتمندتر پاسخ‌های تمیزتر، واضح‌تر و مرتبط‌تر تولید می‌کنند، و عملکرد یک‌دست و متعادلی در تمامی وظایف دارند.

۲. زمان اجرای مدل ها بستگی زیادی به قدرت آنها دارد.

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

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

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

جمع بندی و پیشنهادها

- 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.