بررسی Self-Explanation در مدل‌های زبانی بزرگ (LLM) در تسک Question-Answering

۱ مقدمه

هدف اصلی این پروژه بررسی کیفیت توضیحاتی است که مدل‌های زبانی بزرگ (LLM) برای پاسخ‌های تولیدی خود ارائه می‌دهند. به طور خاص، دو معیار مهم را بررسی می‌کنیم:

- **Faithfulness** — آیا توضیحات مدل واقعاً منطق درونی آن را منعکس می‌کنند؟ آیا این توضیحات مستقیماً از فرآیند تصمیم‌گیری مدل نشأت می‌گیرند یا صرفاً تولیداتی سطحی و غیرمرتبط هستند؟
- **Plausibility** — آیا توضیحات مدل از دیدگاه انسانی قابل فهم، منطقی و متقاعدکننده هستند؟

۲ تسک مورد بررسی و روش تحقیق

ما این مطالعه را در چارچوب پاسخ به سؤالات چندگزینه‌ای (Question-Answering) انجام می‌دهیم. در این تسک، هر نمونه شامل یک **متن زمینه (Context)**، یک **سؤال چهارگزینه‌ای** و **یک پاسخ صحیح** است. انتظار داریم که مدل:

۱. براساس **متن زمینه**، پاسخ صحیح را انتخاب کند.

۲. توضیحی ارائه دهد که نشان دهد چرا این پاسخ را انتخاب کرده است.

۳. کیفیت و صحت این توضیحات را ارزیابی کنیم.

برای این هدف، از **دیتاست Belebele** استفاده می‌کنیم که شامل سؤالات چندزبانه، از جمله فارسی، است.

۳ بهبود عملکرد مدل با استفاده از توضیحات مدل‌های بزرگ‌تر

در گام نخست، رویکردی را آزمودیم که در آن از **مدل‌های زبانی بزرگ‌تر** برای تولید **توضیحات** و هایلایت بخش‌های کلیدی متن استفاده کردیم. سپس این اطلاعات را به عنوان ورودی به یک **مدل کوچک‌تر** (مانند Llama-۸B) ارائه دادیم. نتایج نشان دادند که دقت مدل کوچک‌تر با این روش حدود **۱۰ درصد بهبود یافت** که نشان‌دهنده کارآمد بودن توضیحات و بخش‌های برجسته‌شده توسط مدل‌های بزرگ‌تر (مانند Command-R-Plus Cohere) است.

۴ استفاده از Log-Probability در ارزیابی Faithfulness

در ادامه، قصد داریم از قابلیت Probability Log در مدل‌های زبانی استفاده کنیم. این ویژگی به ما امکان می‌دهد که برای هر **توکن تولیدی مدل**، میزان **احتمال انتخاب آن توکن** را مشاهده کنیم. بررسی نحوه استفاده از این قابلیت برای ارزیابی Faithfulness یکی از چالش‌های کلیدی ما در این پروژه است.

۵ چالش‌های پیش رو

۱. **داده‌های محدود برای زبان فارسی** — هدف اصلی ما بررسی **مدل‌ها در زبان فارسی** است، اما ممکن است در تهیه داده‌های کافی با چالش مواجه شویم. اگرچه دیتاست **Belebele** گزینه مناسبی برای فارسی محسوب می‌شود، اما تنها **۹۰۰ نمونه** دارد. در صورت نیاز به افزایش داده‌ها، گزینه‌هایی مانند **OpenBookQA** یا **ScienceQA** را بررسی خواهیم کرد.

۲. **تعریف متریک‌های کارآمد** — یکی از چالش‌های مهم این است که **متریک‌هایی برای سنجش کیفیت توضیحات** طراحی کنیم که واقعاً قابل‌اعتماد و معتبر باشند.

۳. **نحوه استفاده از Log-Probability برای سنجش Faithfulness** — روش استفاده از این قابلیت هنوز به طور دقیق مشخص نشده است و نیاز به بررسی‌های بیشتری دارد.