| موضوع: **بررسی توهمات در مدل‌های زبانی** | | |
| --- | --- | --- |
| کاربرد و اهمیت مساله: یکی از مشکلات مهم مدل‌های زبانی مخصوصا مدل‌های زبانی بزرگ تولید متون با کیفیت از نظر دستور زبانی و به ظاهر درست می‌باشد در حالی که این متون شامل مفاهیم و گزاره‌های اشتباه هستند. در این پروژه تلاش می‌شود این موارد در پرسش و پاسخ به زبان فارسی تشخیص و برطرف شوند. | | |
| دادگان مورد استفاده | <https://github.com/sajjjadayobi/PersianQA> | |
| ورودی و خروجی‌های مساله | متن کوئری و پاسخ تولید شده توسط LLM | مشخص کردن Factual بودن یا نبودن پاسخ به همراه دلیل یا span قسمت ناصحیح |
| داده ارزیابی | تقسیم دادگان مورد استفاده به قسمت‌های Train/Test/Validation یا استفاده از Cross-Validation. | |
| معیار ارزیابی | selfCheckGPT score, Precision, Recall, Rough, BLEU | |
| مدل پیشنهادی | بخش تشخیص: استفاده از selfCheckGPT  <https://arxiv.org/pdf/2303.08896>  بخش تصحیح:  استفاده از قابلیت خود مدل زبانی و تولید افزوده بازیابی(RAG) و تنظیم دقیق مدل (Finetune) برای تشخیص و برطرف توهمات در مدل زبانی در زبان فارسی | |
| مدل‌های پایه برای مقایسه | T5, GPT3, Dorna Persian LLM, PersianMind, Bert-Multilingual | |
| مساله تولید متن | بخش تشخیص:  مرحله ۱: تولید پاسخ برای پرسش مطرح شده از متن  مرحله ۲: پاسخ تولید شده و پاسخ واقعی به selfCheckGPT داده می شود تا میزان ناسازگاری مشخص شود.  بخش تصحیح:  استفاده از RAG‌ به منظور تولید پاسخ صحیح | |
| استفاده از مدل‌های زبانی بزرگ | تولید خروجی متن | |
| دموی پیشنهادی | اجرای اولیه بخش تشخیص:  در زیر نمونه‌ای از مجموعه داده پرسش و پاسخ PersianQA آورده شده است که به ازای هر متن تعدادی سوال و پاسخ مطرح شده است. از مدل Bert-Multilingual به عنوان مدل پایه برای سنجش میزان توهم استفاده شده است. برای برخی سوالات در متن متناظرش پاسخی وجود ندارد اما LLM پاسخ‌هایی تولید کرده است . در ستون scoresent-score-nli امتیاز selfCheckGPT گزارش شده است. امتیاز selfCheckGPT هر چقدر بالاتر باشد نشان دهنده ناسازگار بودن جواب مدل و جواب واقعی پرسش است.      <https://cdn-uploads.huggingface.co/production/uploads/62f4ac43567dbf9a39f75474/rKUcS3txndMbUHLIYgTG0.gif> | |
| اکانت گیت‌هاب نماینده | https://github.com/elahe-badali | |
| اکانت هاگینگ‌فیس نماینده | https://huggingface.co/elahebadali | |