

 <p>دانشگاه علم و صنعت ایران</p> <p>دانشکده مهندسی کامپیوتر</p> <p>مدرس: دکتر مرضیه داوود آبادی</p>		مبانی پردازش زبان طبیعی	به نام او
نمره	<p><b>بازنویسی پرس و جو</b></p> <p><b>استفاده از LLM های فارسی برای Query Paraphrasing</b></p>		
---	<p>در این تمرین، شما باید از چندین مدل زبان بزرگ (LLM) برای حل مسئله Query Paraphrasing یا تشخیص جمله‌های مترادف استفاده کنید. دیتاست Query Paraphrasing شامل زوج‌های سؤال است که باید تعیین کنید آیا این دو سؤال مترادف یکدیگر هستند یا خیر. هدف این تمرین، آشنایی با مدل‌های زبان بزرگ و ارزیابی عملکرد آنها در تشخیص جمله‌های مترادف است. شما با استفاده از سه مدل مختلف و در سه سناریوی مختلف (Zero-shot، Five-shot و One-shot) عملکرد این مدل‌ها را مقایسه خواهید کرد. این تمرین به شما کمک می‌کند تا با Prompt Engineering و In Context Learning آشنا شوید. در این تمرین نیاز به آموزش مدل نیست و فقط از تعامل با مدل‌ها تست‌های مرتبط را انجام دهید.</p> <p><b>1 اهمیت تشخیص جمله‌های مترادف</b></p> <p>تشخیص جمله‌های مترادف یکی از مسائل مهم در پردازش زبان طبیعی است. این مسئله در کاربردهای مختلفی مانند جستجوی اطلاعات، پاسخ به سوالات، و سیستم‌های توصیه‌گر اهمیت دارد. به عنوان مثال، در یک سیستم جستجوی اطلاعات، تشخیص اینکه دو سؤال مختلف به یک موضوع اشاره دارند، می‌تواند به بهبود نتایج جستجو کمک کند.</p> <p><b>مدل‌های مورد استفاده</b></p> <p>شما باید از مدل‌های زیر استفاده کنید:</p> <ul style="list-style-type: none"> <li>meta-llama/Meta-Llama-3-8B-Instruct</li> <li>ViraIntelligentDataMining/PersianLLaMA-13B-Instruct</li> <li>universitytehran/PersianMind-v1.0</li> </ul> <p>بچه‌ها ورودی هر سه تا مدل باید اینجوری باشه؟ f"{question1}&lt;sep&gt;{question2}" بعد خروجی هر سه تا هم 0 و 1 هست؟ این لینک رو ببین <a href="https://huggingface.co/docs/transformers/main/en/chat_templating#what-are-generation-prompts">https://huggingface.co/docs/transformers/main/en/chat_templating#what-are-generation-prompts</a></p>		

استفاده کنید <https://huggingface.co/microsoft/Phi-3-mini-4k-instruct> برای تمرین شماره ۷، به جای مدل دوم می‌تونید از مدل همچنین برای لود مدل لاما می‌تونید از لینک زیر استفاده کنید:

<https://medium.com/@rohanvermaAI/llama-3-what-we-know-and-how-to-use-it-in-free-collab-24ec5d6058ff>

اگه هم درخواستتون مورد تایید قرار نگیرد از مدل زیر استفاده کنید:

<https://huggingface.co/NousResearch/Meta-Llama-3-8B-Instruct>

## سناریوهای آزمایش

برای هر مدل، شما باید سه سناریوی زیر را آزمایش کنید:

- **Zero-shot Learning**: در این سناریو، مدل بدون هیچ داده آموزشی، باید تشخیص دهد که آیا دو سؤال مترادف هستند یا خیر. این سناریو نشان می‌دهد که مدل تا چه حد می‌تواند بدون آموزش قبلی، وظیفه را انجام دهد.
- **One-shot Learning**: در این سناریو، یک مثال آموزشی به مدل داده می‌شود و سپس مدل باید تشخیص دهد که آیا دو سؤال مترادف هستند یا خیر. این سناریو نشان می‌دهد که مدل تا چه حد می‌تواند از یک مثال آموزشی یاد بگیرد.
- **Five-shot Learning**: در این سناریو، پنج مثال آموزشی به مدل داده می‌شود و سپس مدل باید تشخیص دهد که آیا دو سؤال مترادف هستند یا خیر. این سناریو نشان می‌دهد که مدل تا چه حد می‌تواند از چند مثال آموزشی یاد بگیرد.

## دیتاست

دیتاست Query Paraphrasing شامل زوج‌های سؤال به همراه برچسب (0 یا 1) است که نشان می‌دهد آیا دو سؤال مترادف هستند یا خیر.

<https://github.com/persiannlp/parsinlu/tree/master/data/qqp>

## نکات مهم

- سعی کنید بهترین Prompt را برای هر مدل پیدا کنید.
- با استفاده از مثال‌های آموزشی، مدل‌ها را بهبود دهید و عملکرد آنها را ارزیابی کنید.
- روند کار، پارامترهای استفاده شده برای تعامل با مدل و نتایج به دست آمده را به دقت تحلیل کنید و تفاوت عملکرد مدل‌ها را در سناریوهای مختلف بررسی و گزارش کنید.

موفق باشید.