مقدمه

در این تمرین، هدف بررسی توانایی مدل‌های زبانی بزرگ (LLMs) در تولید و تحلیل داده‌های تخصصی پزشکی است .این فرآیند شامل سه گام اصلی است:

۱- تولید داده مصنوعی واقع‌گرایانه

۲- فاین‌تیون مدل LLAMA با تکنیک LORA

۳- ارزیابی عملکرد مدل قبل و بعد از آموزش مجدد.

گام ۱ :تولید داده مصنوعی

در مرحله اول با استفاده از داده‌های معتبر پزشکی، ۱۰۰ جفت علائم و تشخیص تولید شد. داده‌ها به زبان فارسی و در قالب JSON ذخیره شدند. GPT-4O و پرامپت زیر برای تولید این داده‌ها استفاده گردید:

پرامپت تولید داده:

من یک فایل PDF پزشکی دارم که شامل اطلاعات کامل در مورد علائم، بیماری‌ها، و توضیحاتی در مورد شرایط مختلف پزشکی است.

لطفاً با استفاده از محتوای این کتاب، ۱۰ جفت نمونه‌ی زیر تولید کن (به زبان فارسی):

- یک :"symptom" که شامل توضیحی کوتاه و بالینی درباره علائم یک بیمار باشد (مثل سن، جنس، وضعیت جسمی، شکایت اصلی و نشانه‌ها)
- یک :"diagnosis" که تشخیص قطعی یا محتمل همان وضعیت باشد

قالب خروجی باید به صورت JSON باشد، مانند نمونه زیر:

```
{
"symptom": " مرد ۵۰ ساله با تاری دید ناگهانی، بدون درد در یک چشم، بدون سابقه بیماری قلبی،
".مراجعه کرده است
"diagnosis": "انسداد شریان مرکزی شبکیه"
},
```

لطفاً فقط از اطلاعات واقعی موجود در محتوای فایل PDF استفاده کن و اگر لازم است، علائم را به‌صورت ترکیبی اما واقعی تنظیم کن

نمونه‌ای از داده تولیدشده:

```
{
    "symptom": "مرد ۵۰ ساله با بی‌حسی، ضعف یک طرف بدن و اختلال گفتار ناگهانی مراجعه کرده است.",
    "diagnosis": "سکته مغزی ایسکمیک"
}
```

گام ۲ :فاین‌تیون مدل با Lora

مدل پایه unsloh/Llama-۳.۲-۱B انتخاب شد و با استفاده از تکنیک Lora آموزش مجدد داده شد .در این روش، تنها برخی از وزن‌های مدل تغییر داده شده و هزینه محاسباتی کاهش می‌یابد .

تکنیک Lora به‌جای به‌روزرسانی تمام وزن‌های مدل، فقط وزن‌های لایه‌های خاصی را تغییر می‌دهد. این روش باعث می‌شود نیاز به حافظه و توان محاسباتی بسیار کمتر شده و سرعت آموزش افزایش یابد، در حالی‌که کیفیت مدل حفظ می‌شود.

پیکربندی آموزش مدل:

-مدل پایه unsloh/Llama-۳.۲-۱B

-تکنیک Lora با هدف کاهش مصرف منابع

-لایه‌های مورد تنظیم q_proj و v_proj

-پارامترهای Lora

- r = ۸
- Lora alpha= ۱۶
- Lora dropout = ۰.۰۵

-تنظیمات آموزش

- تعداد epoch: ۲
- batch size مؤثر: ۱۶ (با استفاده از gradient accumulation)
- توکن طول: ۳۸۴

- یادگیری نرخ: ۴e-۵

برای آماده‌سازی داده‌ها، هر نمونه به قالب زیر تبدیل شد:

علائم بیمار: مردی با سردرد، تهوع و تاری دید مراجعه کرده است .
تشخیص: تومور مغزی

سپس با استفاده از توکنایزر مدل، این متن‌ها به ورودی عددی تبدیل شده و فرآیند آموزش با استفاده از کتابخانه transformers و peft اجرا شد. در پایان، مدل fine-tuned ذخیره گردید.

گام ۳ :ارزیابی مدل

برای ارزیابی عملکرد مدل، ۱۰۰ سؤال چهارگزینه‌ای پزشکی طراحی شد. داده‌ها به زبان فارسی و در قالب JSON ذخیره شدند. GPT-4O و پرامپت زیر برای تولید این داده‌ها استفاده گردید:

پرامپت تولید داده:
من یک فایل PDF پزشکی دارم که شامل اطلاعات کامل در مورد علائم، بیماری‌ها، و توضیحاتی در مورد شرایط مختلف پزشکی است.

با استفاده از اطلاعات دقیق و معتبر پزشکی، لطفاً ۱۰ سؤال چهارگزینه‌ای در حوزه علائم و بیماری‌ها تولید کن. هر سؤال باید دارای ساختار زیر باشد:

* "question": یک پرسش واضح و دقیق درباره‌ی علائم، تشخیص بیماری، عوارض بیماری یا تفکیک بیماری‌ها باشد.

* "options": چهار گزینه که فقط یکی از آن‌ها درست است.

* "answer": فقط شماره پاسخ صحیح از بین چهار گزینه، بدون تکرار سؤال.

لطفاً فقط اطلاعات واقعی و علمی را استفاده کن و هیچ‌گونه سؤال ساختگی یا نادرست نساز.

نمونه‌ای از داده تولیدشده:

```
{
    "question": "دلیل شایع تنگی نفس در بیماران با نارسایی قلبی چیست؟",
    "options": [
      "کمبود اکسیژن محیطی",
      "تجمع مایع در ریه‌ها",
      "کمبود قند خون",
      "تب و لرز"
    ],
    "answer": 2
  }
```

سپس با هر دو مدل (قبل و بعد از fine-tune) سوالات بررسی شدند. پاسخ صحیح از مدل دریافت و با کلید پاسخ‌ها مقایسه گردید .

پرامپت ارزیابی:

در ادامه یک سؤال چهارگزینه‌ای پزشکی آمده است .فقط یکی از گزینه‌ها صحیح است .لطفاً فقط شماره گزینه صحیح را بنویس.

سؤال :کدام‌یک از علائم زیر مربوط به کم‌خونی فقر آهن است؟

۱.تب شدید

۲.خستگی و رنگ‌پریدگی

۳.درد مفاصل

۴.تعریق شبانه

پاسخ:

نمونه خروجی مدل:


```
{
  "question": "کدام‌یک از علائم زیر مربوط به کم‌خونی فقر آهن است؟",
  "options": ["شدید تب", "پریدگی‌رنگ و خستگی", "مفاصل درد", "شبانه تعریق"],
  "answer": ۲,
  "predicted": "۲",
  "is_correct": true
}
```

نتایج ارزیابی

نتایج ارزیابی مدل در دو حالت به شرح زیر است:

مدل	دقت
LLAMA BASE	34.07%
Fine-Tuned LLAMA	34.07%

نتیجه‌گیری

دقت مدل پایه و مدل آموزش‌دیده یکسان باقی ماند. این موضوع می‌تواند دلایل مختلفی داشته باشد. در ادامه به مهم‌ترین دلایل و راه‌حل‌های پیشنهادی پرداخته شده است:

۱) داده‌های آموزشی ناکارآمد:

اگر داده‌های موجود در فایل ورودی خیلی ساده، تکراری یا بی‌ربط به سؤالات ارزیابی باشند، ممکن است مدل از قبل پاسخ‌ها را بداند و آموزش جدید تأثیر نگذارد.

۲) مدت آموزش ناکافی:

تنظیم ۲ ایپوک ممکن است برای یادگیری مناسب کافی نباشد.

۳) نشت اطلاعات بین داده آموزشی و ارزیابی:

اگر ساختار سؤالات ارزیابی بسیار شبیه به داده آموزشی باشد، ممکن است مدل حتی بدون آموزش جدید بتواند پاسخ دهد. در این حالت آموزش عملاً چیزی به مدل اضافه نمی‌کند.

۴) سؤالات ارزیابی ساده یا پرتکرار:

اگر سؤالات طراحی‌شده دارای جواب‌های واضح یا پرتکرار باشند، مدل پایه نیز به راحتی می‌تواند آن‌ها را پاسخ دهد و تفاوت قابل‌ملاحظه‌ای با مدل آموزش‌دیده مشاهده نمی‌شود.

پیشنهادهای بهبود

۱. تعداد epoch ها را افزایش دهیم.

۲. لاگ‌های آموزش را بررسی کنیم که آیا loss کاهش یافته یا ثابت مانده است.
۳. سؤالات ارزیابی جدید طراحی کنیم که از داده‌های آموزشی مستقل باشند.
۴. نمونه‌های دشوارتر را برای ارزیابی مدل استفاده کنیم.

به دلیل محدودیت زمانی و نیاز به صرف زمان قابل‌توجه برای آموزش‌های مجدد، تنظیم دقیق‌تر پارامترها و استفاده از داده‌های بیشتر، اجرای پیشنهادهای بهبود و آزمایش‌های تکمیلی امکان‌پذیر نبود.