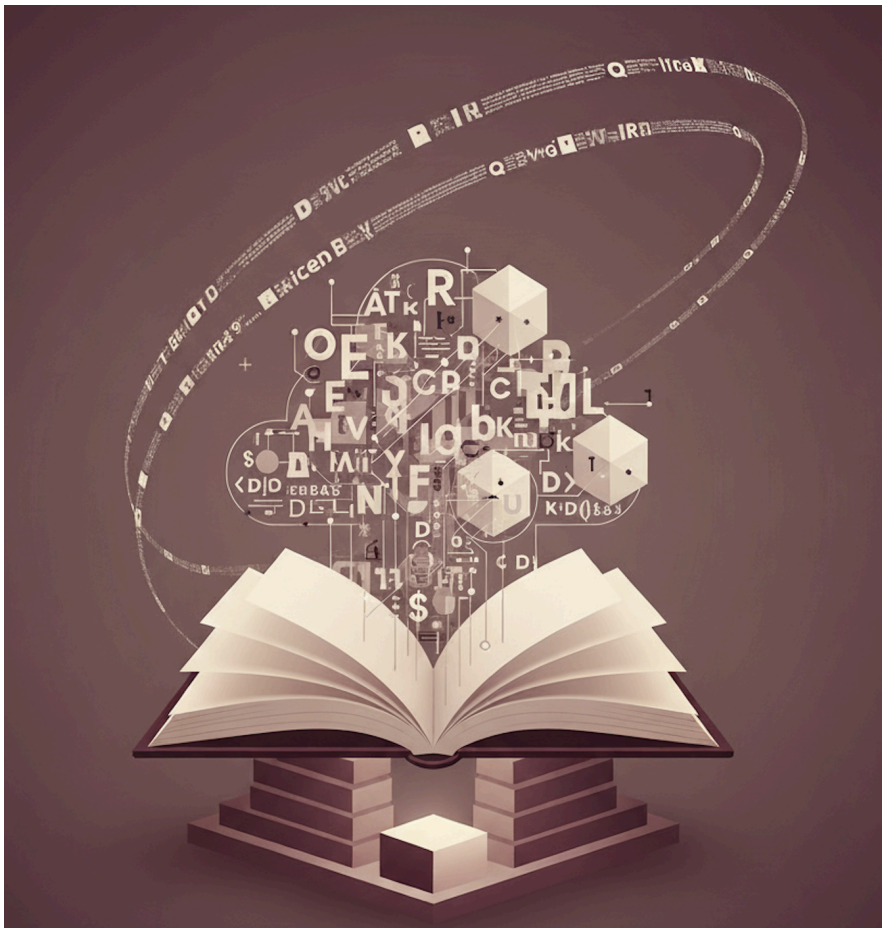
گزارش تسک ارزیابی مهارت فنی شرکت مفید تحلیل داده (متدیتا)



گردآورنده: رادین خیام

ایمیل: RadinKhayyam@gmail.com

شماره همراه: ۰۹۱۲۸۲۷۷۲۸۲

مقدمه

هدف اصلی این پروژه، ارزیابی جامع مهارت‌های فنی و تحلیلی در کار با مدل‌های زبانی بزرگ (LLMs) برای «تولید عنوان خبری مناسب بر اساس محتوای خبر» است. این تسک به طور خاص، توانایی مدل‌ها را در تولید عنوانی دقیق، خلاصه و مرتبط که به «عنوان اصلی» (title) موجود در دادگان، تا حد امکان نزدیک باشد، مورد سنجش قرار می‌دهد.

برای انجام این ارزیابی، دو مدل زبانی سبک (Lightweight) و متن-باز (Open-source) که توانایی مطلوبی در درک و تولید زبان فارسی دارند، انتخاب شده‌اند:

۱. Gemma 3 - 4B - IT

۲. Meta-Llama-3.1-8B-IT

انتخاب این مدل‌ها هم‌راستا با پیشنهاد فایل تسک مبنی بر استفاده از مدل‌های سبک (مانند ۷-۸ میلیارد پارامتر) است. لازم به ذکر است که این پروژه به طور خاص بر عملکرد این دسته از مدل‌ها تمرکز دارد. در حالی که مدل‌های بسیار بزرگتر (مانند خانواده Gemini یا GPT) ممکن است در سناریوهای Zero-shot (بدون فاین-تیون) به دلیل دانش گسترده‌تر خود عملکرد قوی‌تری نشان دهند، تمرکز این تسک بر ارزیابی و بهینه‌سازی مدل‌هایی است که امکان فاین-تیون و استفاده در منابع محاسباتی محدودتر (مانند Google Colab) را فراهم می‌کنند.

فرآیند انجام این پروژه بر اساس الزامات فایل تسک و به صورت ساختاریافته انجام پذیرفته است. این فرآیند شامل چهار گام اصلی است که بر روی هر دو مدل Llama و Gemma پیاده‌سازی شده‌اند:

۱. آماده‌سازی و بررسی اکتشافی داده‌ها (EDA):

قبل از هرگونه مدلسازی، دادگان پیوست شده (شامل ستون‌های content و title) مورد بررسی قرار گرفت. این بررسی شامل تحلیل آماری اولیه، بررسی طول متن‌ها و عناوین، و شناسایی موارد نویز یا داده‌های پرت احتمالی بود. در نهایت، داده‌ها به دو بخش مجزای آموزشی (Train) و آزمایشی (Test) تقسیم شدند.

۲. مرحله اول: آزمایش بدون فاین-تیون (Prompt-based Evaluation):

در این مرحله، عملکرد پایه‌ی هر دو مدل بدون هیچ‌گونه آموزشی، ارزیابی شد. این ارزیابی از طریق سه روش مهندسی پرامپت انجام شد:

- Zero-shot: از مدل خواسته شد تا مستقیماً برای محتوای خبر، عنوان تولید کند.
- Few-shot: چند نمونه (content-title) به عنوان الگو در پرامپت گنجانده شد تا مدل با الگوی مشابه، عنوان جدید را تولید نماید.

- Prompt Engineering: پرامپت‌های مختلفی آزمایش شدند تا با تغییر در دستور (Instruction) و سبک پرامپت، کیفیت خروجی بهبود یابد.

۳. مرحله دوم: فاین-تیون مدل (Fine-tuning):

پس از ارزیابی عملکرد پایه، در این مرحله هر دو مدل بر روی داده‌های آموزشی فاین-تیون شدند. با توجه به محدودیت منابع محاسباتی و پیشنهادهای فایل تسک، از روش بهینه‌سازی LORA استفاده گردید. این روش امکان آموزش مدل‌های نسبتاً بزرگ را بر روی GPUهای با حافظه‌ی محدود فراهم می‌سازد.

۴. ارزیابی نهایی:

در نهایت، عملکرد مدل‌ها در هر دو فاز (قبل و بعد از فاین-تیون) با استفاده از معیارهای استاندارد ارزیابی متن مانند METEOR، ROUGE-L، BLEU و Bert Score مقایسه و تحلیل گردید. همچنین تحلیل کیفی نمونه‌های خروجی برای بررسی «دقت، روانی و ارتباط معنایی» انجام شد.

آماده‌سازی و بررسی اکتشافی داده‌ها (EDA)

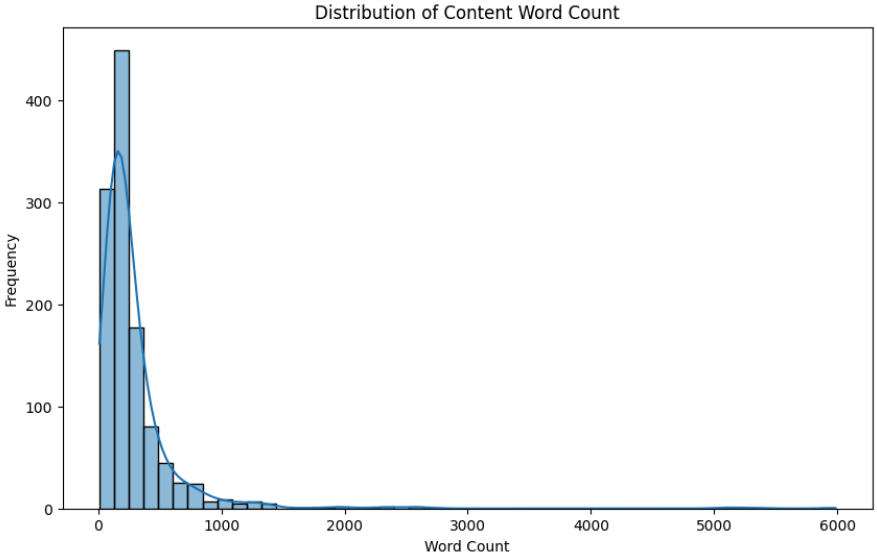
اولین گام پیش از هرگونه مدلسازی، بارگذاری و تحلیل اکتشافی دادگان بود. هدف از این مرحله، درک ساختار داده‌ها، شناسایی ویژگی‌های آماری آن‌ها و کسب بینش‌هایی برای مراحل بعدی (مهندسی پرامپت و فاین-تیون) بود.

الف) تحلیل آماری و توزیع داده‌ها

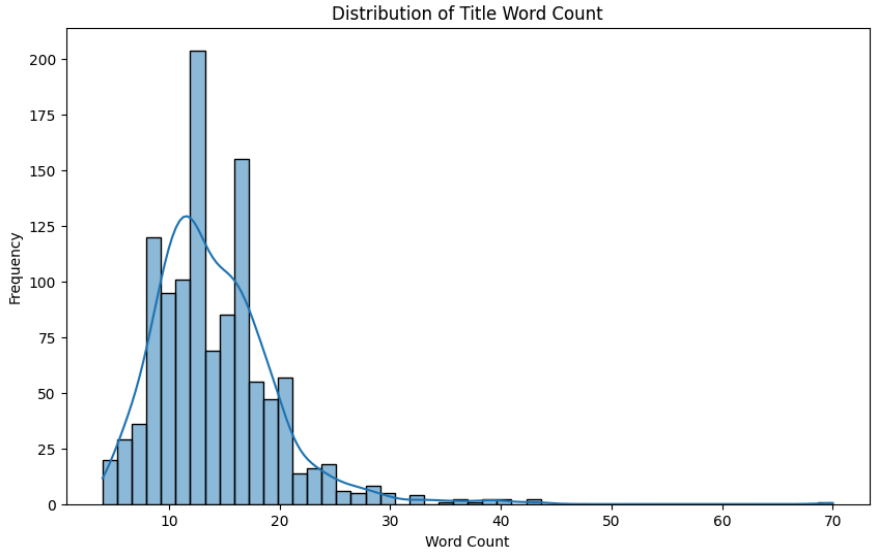
پس از بارگذاری داده‌ها، ابتدا طول هر «محتوا» (content) و «عنوان» (title) بر اساس تعداد کلمات محاسبه شد. نتایج آماری کلیدی به شرح زیر است:

معیار	طول محتوا (کلمه)	طول عنوان (کلمه)
میانگین	۲۹۰.۸۱	۱۴.۱۱
انحراف معیار	۴۱۴.۲۷	۵.۶۳
حداقل	۱۹	۲
حداکثر	۵۹۶۹	۷۰

توزیع طول محتوا: این نمودار (تصویر بالا) دارای چولگی به راست (Right-skewed) بسیار شدیدی است. این بدان معناست که اکثر مقالات (بیش از ۴۰۰ نمونه) نسبتاً کوتاه و زیر ۵۰۰ کلمه هستند، اما تعداد کمی مقالات بسیار طولانی (حتی تا حدود ۶۰۰۰ کلمه) نیز در دادگان وجود دارد.



توزیع طول عنوان: در مقابل، طول عناوین (تصویر زیر) توزیع بسیار نرمال‌تری دارد که در حوالی ۱۰ تا ۱۵ کلمه متمرکز است.

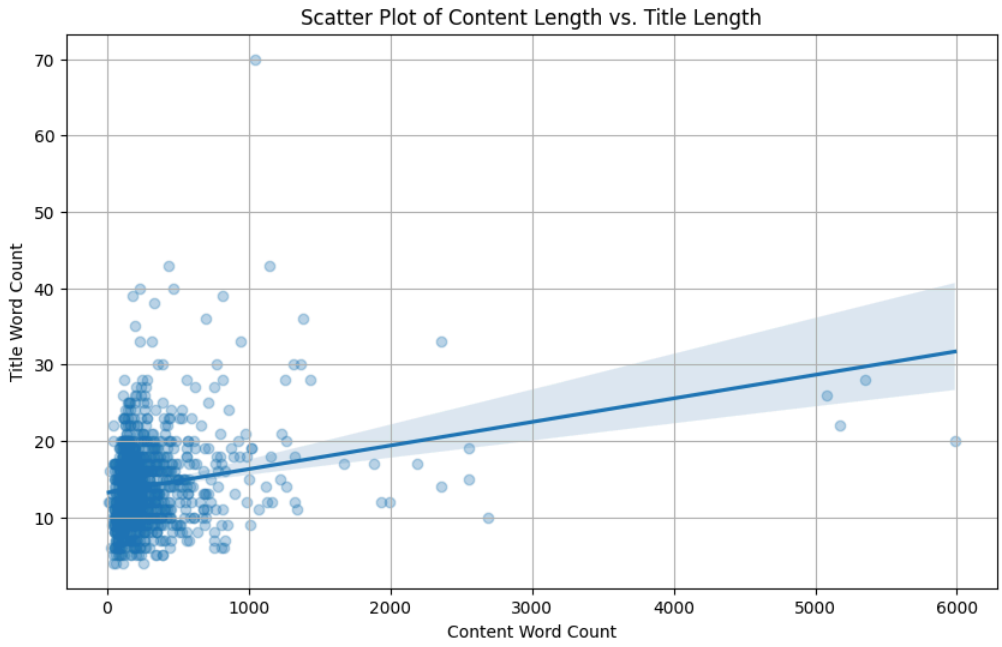


مهم‌ترین یافته از این بخش، میانگین طول عنوان (حدود ۱۴-۱۵ کلمه) بود. این اطلاعات مستقیماً در مرحله «مهندسی پرامپت» استفاده شد تا به مدل دستور داده شود عناوینی با طول مشابه (کوتاه و مختصر) تولید کند.

ب) تحلیل همبستگی

یکی از فرضیه‌های اولیه این بود که آیا مقالات طولانی‌تر، عناوین طولانی‌تری نیز دارند یا خیر. برای سنجش این موضوع، ضریب همبستگی پیرسون بین تعداد کلمات محتوا و تعداد کلمات عنوان محاسبه شد.

- ضریب همبستگی: 0.23



نمودار پراکندگی (Scatter Plot) و ضریب همبستگی بسیار پایین (0.23) نشان می‌دهند که رابطه خطی بسیار ضعیفی بین طول محتوا و طول عنوان وجود دارد. به عبارت دیگر، بلند بودن متن خبر، تأثیر چندانی بر بلند بودن عنوان آن ندارد. این یافته، فرضیه‌ی اولیه را رد می‌کند.

ج) کیفیت داده‌ها و تقسیم‌بندی

- داده‌های گمشده (Missing Values): کل دادگان بررسی شد و مشخص گردید که هیچ ستونی فاقد مقادیر گمشده نیست، که نشان‌دهنده کیفیت خوب داده‌ها است.

- داده‌های پرت (Outliers): اگرچه مقالات بسیار طولانی (مانند نمونه ۵۹۶۹ کلمه‌ای) وجود داشتند، اما این موارد به عنوان خطای ورود داده تلقی نشدند، بلکه بخشی طبیعی از دادگان اخبار (شامل مقالات تحلیلی طولانی) در نظر گرفته شدند. بنابراین، هیچ داده‌ای حذف نشد.
- تقسیم‌بندی: در گام نهایی آماده‌سازی، مجموعه داده با نسبت ۹۰٪ آموزشی (۱۰۴۴ نمونه) و ۱۰٪ آزمایشی (۱۱۶ نمونه) تقسیم گردید تا فرآیند آموزش و ارزیابی مدل‌ها بر روی داده‌های دیده نشده، ممکن باشد.

بخش اول: آزمایش بدون فاین-تیون (Prompt-based Evaluation)

در این بخش، عملکرد پایه‌ی مدل‌های Gemma 3 - 4B و Llama 3.1 - 8B قبل از هرگونه فاین-تیون، مورد ارزیابی قرار می‌گیرد. هدف، سنجش توانایی ذاتی مدل‌ها در درک محتوا و خلاصه‌سازی آن در قالب یک «عنوان خبری» تنها با استفاده از مهندسی پرامپت است.

الف) روش‌شناسی مهندسی پرامپت

فرآیند طراحی پرامپت به صورت یک چرخه‌ی تکرارشونده و با هدف بهبود مستمر خروجی انجام شد. برای تولید پرامپت‌های اولیه، از روش Meta-prompting و با کمک مدل Gemini 2.5 Pro استفاده شد تا پرامپت‌هایی دقیق و ساختارمند ایجاد شوند. این فرآیند در چهار مرحله اصلی انجام پذیرفت:

۱. پرامپت پایه (simple_prompt):

در ابتدا، یک پرامپت بسیار ساده و مستقیم طراحی شد تا به عنوان خط پایه (Baseline) برای ارزیابی عملکرد اولیه مدل‌ها عمل کند.

```
Generate a high-quality headline for the following news report, responding only with the Persian title and no other text.
```

۲. پرامپت Zero-shot:

با توجه به ضعف مشهود پرامپت پایه، یک پرامپت پیشرفته با مهندسی دقیق طراحی شد. ویژگی‌های کلیدی این پرامپت عبارتند از:

- تعریف نقش (Persona): به مدل نقش «ویراستار ارشد و باتجربه اخبار فارسی» داده شد.
- تعریف محدودیت‌ها:
  - زبان: الزام به استفاده از زبان فارسی.
  - طول: بر اساس یافته‌های بخش EDA، طول خروجی به «تقریباً ۱۵ کلمه» محدود شد.
  - قالب: خروجی باید یک «جمله اخباری (declarative)» باشد و نه پرسشی.
- قالب خروجی: تاکید بر اینکه مدل باید فقط عنوان را و بدون هیچ متن اضافه‌ای برگرداند.

```
You are an expert-level, highly experienced Persian news editor. You have a deep understanding of journalistic standards, factual reporting, and the nuances of the Persian
```

language.

Your task is to generate a single, high-quality, informative headline (title) for a given news report. You will be provided with the full text of the news content.

Instructions and Constraints:

- Language: The generated title MUST be written primarily in the Persian language. English words should only be used if they are specific proper nouns (such as organization names, brands, or technical terms) and they appear in the original news content provided.
- Length: The title must be approximately 15 words long.
- Format: The title MUST be a declarative sentence. It should state a fact or summarize the main point directly. It must not be phrased as a question.
- Focus: The primary goal is to be informative and accurate. The title must be 100% relevant to the provided news content and precisely summarize its core subject or most significant finding.

Analyze the provided news content, identify its most critical information, and then craft a single headline that meets all of the above criteria.

Output Format:

Respond only with the generated Persian title. Do not add any commentary, explanations, or introductory text.

۳. پرامپت Few-shot:

در این مرحله، پرامپت zero-shot حفظ شد و ۵ نمونه (Example) به آن اضافه گردید. این ۵ نمونه به صورت هدفمند و برای آموزش «قوانین پنهان» موجود در دادگان انتخاب شدند:

- مثال ۱: یک نمونه استاندارد که در آن «ادعا» (Claim) درست ارزیابی شده و «عنوان» (Title) نیز همان ادعا است.
- مثال ۲: آموزش تولید تیترهای «نقل قولی» (Quote-based) با ذکر نام فرد (مانند «حسن روحانی گفته‌است...»).
- مثال ۳: مهم‌ترین قانون. نمونه‌ای که در آن «ادعا» (مهاجرت ۴۵۰۰ پزشک) در بخش «نتیجه بررسی» (Verification) به عنوان «گمراه‌کننده» ارزیابی شده، اما «عنوان» همچنان همان «ادعا» است.

- مثال ۴: نمونه‌ای که نشان می‌دهد حتی اگر متن خبر کوتاه باشد، عنوان می‌تواند نسبتاً طولانی باشد (برای جلوگیری از خلاصه‌سازی بیش از حد).
- مثال ۵: آموزش استفاده از اسامی انگلیسی (مانند BBC) در عنوان، در صورتی که در متن اصلی موجود باشد.

در ادامه، متن کامل این ۵ نمونه که به پرامپت اضافه شدند، آمده است:

Examples:

Here are five examples of correct Content to Title generation:

Example 1
Content:
شرح ادعا کاربری در ایکس ادعا کرده است که بیش از هزار کارگر طی شش ماه در حوادث ضمن کار جان خود را از دست داده‌اند. نتیجه بررسی براساس آمار اختصاصی سازمان پزشکی قانونی : در شش ماهه نخست سال جاری ۱۰۷۷ نفر در حوادث کار جان خود را از دست دادند. این رقم در مقایسه با مدت مشابه سال قبل با آمار فوتی ۹۳۱ نفر، ۱۵/۷ درصد افزایش داشته است. آمار اختصاصی سازمان پزشکی قانونی در سایت باشگاه خبرنگاران جوان براساس سالنامه آماری وزارت کار در سال ۱۴۰۱ : تعداد متوفیان حوادث ناشی از کار ثبت شده در مراکز پزشکی قانونی در سال ۱۴۰۱ ، ۱۹۰۰ نفر بوده است. سالنامه آماری وزارت کار در سال ۱۴۰۱

Title:

بیش از هزار کارگر طی شش ماه در حوادث ضمن کار کشته شده‌اند

Example 2
Content:
حسن روحانی در آیین افتتاح طرح‌های ملی وزارت جهاد کشاورزی، در هشتاد و دومین پویش تدبیر و امید برای جهش تولید گفت: « مردم خوزستان علی‌رغم اینکه توصیه ما این بود که آنجا برنج‌کاری نشود، اما سال‌هاست برنج‌کاری هم انجام می‌دهند که خودِ آن هم مزید بر علت شده و این‌ها همه دست به دست هم داده و مردم عزیز در سختی قرار گرفتند.» صفحه شبکه منوتو در اینستاگرام به نقل از حسن روحانی نوشته است: «مردم خوزستان علیRateرغم اینکه توصیه ما این بود که در آنجا برنج‌کاری نشود، اما سال‌هاست برنج‌کاری هم انجام می‌ده دهند که خود آن هم مزید بر علت در بحران کم‌آبی شده است.» به گزارش سایت رسمی پایگاه اطلاع رسانی ریاست جمهوری به تاریخ ۲۸ تیرماه، حسن روحانی در آیین افتتاح طرح‌های ملی وزارت جهاد کشاورزی در هشتاد و دومین پویش تدبیر و امید برای جهش تولید گفت: «مردم خوزستان علی‌رغم اینکه توصیه ما این بود که آنجا برنج‌کاری نشود؛ اما سال‌هاست برنج‌کاری هم انجام می‌دهند که

خودِ آن هم مزید بر علت شده و این‌ها همه دست به دست هم داده و مردم عزیز در سختی قرار گرفتند.»
سایت رسمی پایگاه اطلاع رسانی ریاست جمهوری

Title:

حسن روحانی گفته‌است: «برنج‌کاری مردم خوزستان از دلایل بحران کم‌آبی است.»

Example 3

Content:

شرح ادعا کاربری در شبکه اجتماعی ایکس ادعا کرده‌است که فقط در ۸ ماه، ۴ هزار و ۵۰۰ پزشک و پرستار از ایران مهاجرت کرده‌اند. نتیجه بررسی این آمار، براساس گزارش روزنامه ایران از شمار گواهی‌های صادر شده گود استندینگ (Good standing) صادر شده توسط سازمان نظام پزشکی است. گواهی گود استندینگ یکی از مدارکی است که اعضای کادر درمان باید برای مهاجTرت و فعالیت حرفه‌ای در کشورهای خارجی ارائه دهند؛ این گواهی حسن سابقه افراد را مشخص می‌کند و یعنی این افراد صلاحیت لازم را برای فعالیت در کشور مقصد را دارند. بر اساس این گزارش، در 8 ماه ابتدایی امسال، 4500 نفر گواهی گود استندینگ دریافت کرده‌اند، اما صرف دریافت این گواهی به معنای مهاجTرت قطعی نیست. بنابراین، اگرچه دریافت گواهی گوداستندینگ برای اقدام در راستای مهاجرت است؛ اما به این معنا که هر کس این گواهی را دریافت کرده، سپس مهاجرت نموده، نیست. پس ادعای این کاربر گمراه‌کننده ارزیابی می‌شود.

Title:

در 8 ماه ابتدایی سال جاری، 4500 پزشک و پرستار مهاجرت کرده‌اند

Example 4

Content:

گرفتن حق اشتراک و پخش آگهی قبل از انتشار محتوا توسط پلتفرم‌های اینترنتی در هیچ جای دنیا مرسوم نیست و به زودی برای آن دستورالعمل قانونی نوشته و پس از مصوب شدن، به آن‌ها ابلاغ می‌شود. در حال حاضر، قانونی که پلتفرم‌ها را از این عمل منع کند وجود ندارد. مصطفی قاسمیان، روزنامه‌نگار سینما و تلویزیون، در توییتر ادعا کرده است: «با وجودی که ۵ ماه پیش ساترا صراحتا اعلام کرد پلتفرم‌هایی که هزینه اشتراک می‌گیرند، حق نمایش آگهی پیش از پخش محتوا ندارند، اما فیلیمو و نماوا همچنان این عمل غیرقانونی را انجام می‌دهند.» توییت قاسمیان درباره پخش آگهی فیلیمو و نماوا علی سعد، معاون کاربران و تنظیم مقررات اجتماعی ساترا (سازمان تنظیم مقررات صوت و تصویر فراگیر در فضای مجازی) در تماس تلفنی با فکت‌یار گفت : «ما فروردین ماه اعلام کردیم... در حال حاضر این دو پلتفرم کار غیرقانونی انجام نمی‌دهند چون هنوز قانون آن مصوب و ابلاغ نشده است.»

Title:

با وجود اعلام ساترا مبنی بر اینکه پلتفرم‌هایی که هزینه اشتراک می‌گیرند، حق نمایش آگهی پیش از پخش

محتوا را ندارند، فیلیمو و نماوا با انتشار آگهی قبل از نمایش محتوا عمل غیرقانونی انجام می‌دهند.

Example 5
Content:
شرح ادعا عبدالرضا داوری با انتشار فیلمی از مردی که برای دریافت هزینه اشتراک تلویزیون BBC به خانه یکی از شهروندان این کشور رفته‌است، گفت: «در حالیکه مردم بریتانیا مجبور هستند برای شبکه BBC حق اشتراک بپردازند، شبکه BBC چه منافعی در ایران دارد که برنامه‌های بخش فارسی آن، بصورت رایگان و ۲۴ ساعته برای ملت ایران پخش می‌شود؟». نتیجه بررسی براساس آیین‌نامه‌ای که در وبسایت دولت بریتانیا برای دریافت مجوز دیدن تلویزیون وجود دارد ؛ شهروندان برای دیدن برنامه‌های زنده تلویزیون، به صورت کلی و برای همه شبکه‌ها (اعم از BBC)، باید ۱۶۹.۵پوند پرداخت کنند. دیدن بی بی سی فارسی، به صورت آنلاین یا با استفاده از آنتن‌های ماهواره، برای مخاطبان رایگان است. بنابراین، ادعای عبدالرضا داوری در مورد اینکه مردم بریتانیا باید برای دیدن BBC اشتراک تهیه کنند اما این شبکه به طور رایگان برای مردم ایران به طور 24 ساعته برنامه‌سازی می‌کند، درست ارزیابی می‌شود.

Title:
مردم بریتانیا باید برای دیدن BBC اشتراک بخرند؛ در حالی که دیدن BBC فارسی برای مردم ایران رایگان است

۴. پرامپت اصلاح‌شده (Refined Prompt - مبتنی بر تحلیل خطا):

پس از بررسی خروجی‌ها، یک «مشکل اساسی» در منطق مدل‌ها کشف شد. دادگان این پروژه از نوع «فکت‌چک» (Fact-Check) هستند و مدل‌ها به طور طبیعی تمایل داشتند «نتیجه بررسی» (Verification) را به عنوان خلاصه‌ی خبر در نظر بگیرند. در حالی که «عنوان واقعی» در دادگان، در اغلب موارد، خود «ادعا» (Claim) بود، حتی اگر آن ادعا نادرست ارزیابی شده باشد.

برای حل این مشکل، پرامپت اصلاح‌شده (refined_prompt) با یک بخش «منطق هسته‌ای» (Core Logic) طراحی شد. این بخش به صراحت به مدل دستور می‌دهد که در این دادگانِ مشخص، «عنوان» باید خلاصه‌ی «شرح ادعا» باشد، نه «نتیجه بررسی». (نمونه‌های این پرامپت مشابه ۵ نمونه بالا بودند).

You are an expert-level, highly experienced Persian news editor. Your task is to generate a single, high-quality, informative headline (title) based on a given news report.

The content you will receive is a fact-check, which contains two main parts:

1. A "شرح ادعا" (Claim Description)

2. A "نتیجه بررسی" (Verification / Result)

Core Logic: Handling Claims vs. Verifications

Your primary task is to identify the most newsworthy part of the article to summarize as a headline.

- The content you receive is a fact-check. In this format, the "Claim" (ادعا) is almost always the main subject and the most newsworthy part.
- Crucial Instruction: In many cases, even when the "Verification" (نتیجه بررسی) finds that the "Claim" is incorrect (نادرست) or misleading (گمراه‌کننده), the claim is *still* the main subject of the article.
- In these specific cases, your headline MUST be a declarative statement of that incorrect or semi-correct claim.
- The examples (especially Example 3 and Example 4) clearly demonstrate this rule. In both, the verification finds the claim to be false or misleading, but the correct Title is still a statement of the Claim.

Instructions and Constraints:

- Language: The generated title MUST be written primarily in the Persian language. English words should only be used if they are specific proper nouns (such as organization names, brands, or technical terms) and they appear in the original "شرح ادعا".
- Length: The title must be approximately 15 words long.
- Format: The title MUST be a declarative sentence. It must not be phrased as a question.

Analyze the provided news content, identify the core claim, and then craft a single headline that meets all of the above criteria based on the Core Logic.

Examples:
[The 5 examples listed above were also included here]

Output Format:
Respond only with the generated Persian title. Do not add any commentary, explanations, or introductory text.

ب) ارزیابی کیفی و نمونه خروجی‌ها

برای ارزیابی کیفی، ۶ نمونه از خروجی‌های هر دو مدل (Gemma و Llama) در جدول‌های زیر ارائه شده است.

جدول ۱: نمونه خروجی‌های مدل Gemma 3 - 4B - IT (بدون فاین-تیون)

عنوان واقعی (Dataset)	Simple Prompt	Zero-Shot Prompt	Few-Shot Prompt	Refined Prompt
رحیم مشایی در شمال کشور و دور از مردم زندگی می‌کند و حمید بقایی به بیماری عصبی دچار شده است.	آزادی مشایی و بقایی: احمدی‌نژاد از وضعیت آن‌ها ابراز نگرانی کرد	رحیم مشایی و حمید بقایی پس از اعتراض احمدی‌نژاد، در تنهایی و با مشکلات سلامتی زندگی می‌کنند.	رحیم مشایی و حمید بقایی به تنهایی و با مشکلات سلامتی زندگی می‌کنند.	رحیم مشایی و حمید بقایی به تنهایی و با مشکلات سلامتی زندگی می‌کنند
آتشفشان دماوند فعال شده است	دماوند فعال نشد؛ خروج بخار آب علت تصاویر اخیر	خروج بخار آب از قله دماوند، نشانه‌ای از فعالیت طبیعی و خفته این آتشفشان است.	دماوند فعال نشده‌است؛ خروج بخار آب طبیعی است.	دماوند فعال شده است، ادعایی نادرست است.
ایرانی‌ها ۱۸۰ برابر متوسط جهان، شایعات اخبار دروغ و اخبار ناامید کننده دریافت می‌کنند.	ادعای ۱۸۰ برابری اخبار نادرست در ایران رد شد	ایرانیان با دریافت اخبار نادرست، در سطح مشابه کشورهای توسعه‌یافته قرار دارند.	ایرانی‌ها با میزان اخبار نادرست برابر با کشورهای توسعه‌یافته‌اند	ایرانی‌ها با یک درصد اخبار نادرست، در سطح جهانی هستند
مهرجویی بیش از ۲۵ فیلم ساخت که ۷ فیلم او توقیف شد	اعتراض مهرجویی به ۲۸ سال توقیف فیلم	فیلم‌های داریوش مهرجویی بیش از ۲۸ سال توقیف شده و برخی از آن‌ها به‌طور کامل ناپدید گشته‌اند.	داریوش مهرجویی: ۲۸ سال توقیف فیلم، ۶ فیلم توقیف‌شده و یک فیلم مفقود	داریوش مهرجویی ۳۳ فیلم ساخت و فیلم‌های او ۲۸ سال توقیف شدند
دکتر علی لاریجانی گفته است به عرصه‌ی انتخابات مجلس شورای اسلامی ورود می‌کند	تکذیب ورود لاریجانی به انتخابات مجلس	علی لاریجانی با رد ادعای ورود به انتخابات مجلس، بر صحت اطلاعیه خود تأکید کرد.	علی لاریجانی ادعای ورود به انتخابات مجلس را تکذیب کرد.	علی لاریجانی قصد ورود به انتخابات مجلس شورای اسلامی را ندارد.

امارات سالیانه ۲.۵میلیون لیتر آب آشامیدنی از هوا تولید می‌کند	تولید آب از هوا در امارات: ادعاها و واقعیت‌ها	تولید آب آشامیدنی از هوا در امارات: فناوری هیدروپنل با ظرفیت تولید سالانه حدود ۴۳ میلیون لیتر تایید می‌شود.	امارات با فناوری هیدروپنل سالانه حدود 43 میلیون لیتر آب تولید می‌کند.	امارات با استفاده از هیدروپنل سالانه 43 میلیون لیتر آب تولید می‌کند

جدول ۲: نمونه خروجی‌های مدل Llama 3.1 - 8B - IT (بدون فاین-تیون)

عنوان واقعی (Dataset)	Simple Prompt	Zero-Shot Prompt	Few-Shot Prompt	Refined Prompt
رحیم مشایی در شمال کشور و دور از مردم زندگی می‌کند و حمید بقایی به بیماری عصبی دچار شده است.	احمدی‌نژاد: مشاعی در شمال کشور به تنهایی زندگی می‌کند و بقایی دچار بیماری عصبی شده است	محمود احمدی‌نژاد از وضعیت وخیم دو همراه خود حمید بقایی و رحیم مشاعی خبر داد	رحیم مشایی و حمید بقایی در شرایط سخت زندگی می‌کنند.	محمود احمدی‌نژاد گفته است که رحیم مشاعی در شمال کشور به تنهایی زندگی می‌کند.
آتشفشان دماوند فعال شده است	"دماوند در حال خفته‌بودن؛ ادعا درباره فعال شدن آتشفشان نیمه نادرست"	کوه دماوند فعلاً در حالت خفته قرار دارد و خروج بخار آب از دودخان‌ها نشانه‌ای از انفجار نیست	تصاویر اخیر از قله دماوند نشانه‌ای از انفجار آتشفشان نیست.	آتشفشان دماوند فعال شده است.
ایرانی‌ها ۱۸۰ برابر متوسط جهان، شایعات اخبار دروغ و اخبار ناامید کننده دریافت می‌کنند.	ادعای "180 برابر" نسبت به اخبار نادرست در ایران دروغ است	مطالعات علمی نشان می‌دهد که ایرانی‌ها 1 درصد اخبار نادرست را دریافت می‌کنند نه 180 درصد	ادعای کاربر مبنی بر «ایرانی‌ها ۱۸۰ برابر متوسط جهان...» نادرست است.	ایرانی‌ها ۱۸۰ برابر متوسط جهان، شایعات اخبار دروغ و اخبار ناامید کننده دریافت نمی‌کنند.

مهرجویی بیش از ۲۵ فیلم ساخت که ۷ فیلم او توقیف شد	داریوش مهرجویی: ۳۳ فیلم، ۲۸ سال توقیف و ۶ فیلم توقیف شد	داریوش مهرجویی با ساخت ۳۳ فیلم، ۶ فیلم را در دهه‌های گذشته توقیف شد.	داریوش مهرجویی ۳۳ فیلم ساخت که ۶ فیلم او توقیف شد	داریوش مهرجویی بیش از ۲۵ فیلم ساخت که ۷ فیلم او دچار توقیف شدند.
دکتر علی لاریجانی گفته است به عرصه‌ی انتخابات مجلس شورای اسلامی ورود می‌کند	دفتر علی لاریجانی: ادعای ورود به انتخابات مجلس خلاف واقع است	علی لاریجانی دفترش ادعای ورود به انتخابات مجلس را تکذیب کرد	اطلاعیه دفتر علی لاریجانی، ورود او به انتخابات مجلس را تکذیب می‌کند.	علی لاریجانی قصد ورود به عرصه انتخابات مجلس شورای اسلامی را دارد.
(نمونه آب امارات) امارات سالیانه ۲.۵ میلیون لیتر آب آشامیدنی از هوا تولید می‌کند	"ادعای تولید 2.5 میلیون لیتر آب از هوا در امارات"	شرکت اماراتی A1RWATER... حدود 43 میلیون لیتر آب آشامیدنی را در سال تولید می‌کند.	امارات سالانه حدود ۴۳ میلیون لیتر آب آشامیدنی از هوا تولید می‌کند	شرکت A1RWATER... سالیانه 43 میلیون لیتر آب آشامیدنی از هوا تولید می‌کند.

ج) تحلیل کیفی نتایج بخش اول

۱. خطای منطقی فراگیر در پرامپت‌های استاندارد:

تحلیل جداول بالا به وضوح نشان می‌دهد که پرامپت‌های Simple, Zero-Shot و Few-Shot در درک منطق پنهان دادگان شکست خوردند. دادگان این پروژه از نوع «فکت-چک» است و این پرامپت‌ها، به جای استخراج «ادعا»، به طور مداوم «نتیجه بررسی» (Verification) را خلاصه می‌کردند.

- مثال (نمونه دماوند):
  - عنوان واقعی (ادعا): «آتشفشان دماوند فعال شده است»
  - خروجی Zero-Shot (نتیجه): «...دماوند فعلاً در حالت خفته قرار دارد... نشانه‌ای از انفجار نیست»
- مثال (نمونه لاریجانی):
  - عنوان واقعی (ادعا): «...لاریجانی... ورود می‌کند»
  - خروجی Zero-Shot (نتیجه): «علی لاریجانی... ادعای ورود به انتخابات مجلس را تکذیب کرد»
- این خروجی‌ها دقیقاً متضاد هدف دادگان (که همان عنوان واقعی است) عمل کرده‌اند.

۲. موفقیت نسبی Refined Prompt و تفاوت دو مدل:

پرامپت اصلاح‌شده (Refined Prompt) که به صراحت به مدل دستور می‌داد «ادعا» را استخراج کند (حتی اگر نادرست باشد)، یک تغییر بنیادین ایجاد کرد، اما نتایج آن در دو مدل متفاوت بود:

- موفقیت Llama 3.1: این مدل به شکل چشمگیری به دستورالعمل «منطق هسته‌ای» (Core Logic) پایبند بود. در نمونه‌های دماوند، مهرجویی و لاریجانی، مدل Llama دقیقاً «ادعا» را استخراج کرد، حتی اگر می‌دانست (بر اساس متن) که آن ادعا نادرست یا نیمه‌نادرست است. این نشان‌دهنده توانایی بالای Llama 3.1 در دنبال کردن دستورالعمل‌های پیچیده و ضد شهودی است.
- چالش Gemma 3: مدل Gemma، بر خلاف Llama، در پیروی از این منطق پیچیده دچار مشکل شد. در نمونه دماوند، خروجی Gemma این بود: «دماوند فعال شده است، ادعایی نادرست است.»؛ و در نمونه لاریجانی: «علی لاریجانی قصد ورود... را ندارد.». این نشان می‌دهد که Gemma نتوانست بر تمایل ذاتی خود برای بیان «حقیقت» (نتیجه بررسی) غلبه کند و دستورالعمل را نقض کرد.

۳. یک مورد شکست مشترک (تعصب به داده‌ی صحیح):

نمونه «آب امارات» یک بینش بسیار مهم ارائه می‌دهد. در این متن، «ادعا» حاوی عدد ۲.۵ میلیون لیتر و «نتیجه بررسی» حاوی عدد ۴۳ میلیون لیتر (به عنوان ظرفیت) است.

- عنوان واقعی (ادعا): «امارات سالیانه ۲.۵ میلیون لیتر... تولید می‌کند»
- خروجی Llama (Refined): «شرکت A1RWATER... سالیانه 43 میلیون لیتر... تولید می‌کند.»
- خروجی Gemma (Refined): «امارات با استفاده از هیدروپنل سالانه 43 میلیون لیتر... تولید می‌کند»
- در این مورد، هر دو مدل، علی‌رغم دستور صریح Refined Prompt مبنی بر استخراج «ادعا»، شکست خوردند. آن‌ها نتوانستند در مقابل «عدد صحیح‌تر» که در متن (نتیجه بررسی) آمده بود مقاومت کنند و آن را به جای عدد موجود در «ادعا» گزارش کردند. این نشان‌دهنده سوگیری (Bias) شدید مدل‌ها به سمت گزارش داده‌های «صحیح» در متن است، حتی زمانی که دستورالعمل برخلاف آن باشد.

نتیجه‌گیری بخش اول: مهندسی پرامپت استاندارد برای این تسک خاص (فکت-چک) ناکافی بود. موفقیت اصلی در این بخش، «تحلیل خطا» و کشف «منطق پنهان» دادگان بود که منجر به طراحی Refined Prompt شد. این پرامپت، اگرچه بی‌نقص نبود، اما توانست عملکرد مدل Llama 3.1-8B را به شکل چشمگیری به استاندارد مورد نظر دادگان نزدیک کند.

بخش دوم: نتایج فاین-تیون (Fine-tuning Results)

پس از ارزیابی عملکرد پایه مدل‌ها با استفاده از مهندسی پرامپت، در این بخش وارد فاز آموزش (Fine-tuning) می‌شویم. هدف در این مرحله، «تخصصی کردن» (Specializing) مدل‌های Gemma 3 - 4B - IT و Llama 3.1 - 8B - IT برای تسک مشخص «تولید عنوان خبری فارسی» است.

در این فاز، مدل‌ها دیگر صرفاً از دستورالعمل‌های عمومی پیروی نمی‌کنند، بلکه «منطق پنهان» موجود در دادگان آموزشی (یعنی استخراج «ادعا» به عنوان «عنوان») را مستقیماً یاد می‌گیرند.

الف) روش‌شناسی فاین-تیون

فرآیند آموزش بر روی هر دو مدل به صورت یکسان و با استفاده از ابزارهای استاندارد اکوسیستم Hugging Face (مانند peft و trl) انجام شد.

۱. روش بهینه‌سازی (LoRA):

با توجه به محدودیت منابع محاسباتی و حجم بزرگ مدل‌ها (4 و 8 میلیارد پارامتر)، فرآیند آموزش بر روی GPU L4 در Google Colab انجام شد. برای ممکن ساختن آموزش در این محیط، استفاده از روش PEFT ضروری بود. در این پروژه از متد LoRA استفاده شد.

- نحوه عملکرد: به جای آموزش میلیاردها پارامتر موجود در مدل اصلی، LoRA تمام پارامترهای مدل پایه را «فریز» (Frozen) نگه می‌دارد. سپس، «آداپتورهای» بسیار کوچکی را که شامل پارامترهای قابل آموزش جدید هستند، به لایه‌های کلیدی مدل (مشخصاً لایه‌های target_modules مانند q_proj, k_proj, v_proj و ...) تزریق می‌کند.
- تنظیمات LoRA: در این پروژه از r=16 و lora_alpha=32 استفاده شد. این تنظیمات به ما اجازه دادند تا با آموزش کسری بسیار کوچک از پارامترهای کل، به نتایجی معادل با فاین-تیون کامل (Full Fine-tuning) دست یابیم.
  - مدل Gemma 3-4B: تعداد پارامترهای قابل آموزش حدود ۳۲ میلیون بود (کمتر از ۰.۸٪ از کل پارامترها).
  - مدل Llama 3.1-8B: تعداد پارامترهای قابل آموزش حدود ۴۱ میلیون بود (کمتر از ۰.۶٪ از کل پارامترها).

۲. آماده‌سازی داده‌ها (Chat Templating و System Prompt):

مدل‌های Instruction-Tuned (که با پسوند IT یا Instruct مشخص می‌شوند) برای پیروی از یک الگوی مکالمه خاص آموزش دیده‌اند. ورودی دادن داده خام (content, title) به این مدل‌ها منجر به نتایج ضعیف می‌شود.

بنابراین، داده‌های آموزشی به فرمت «چت» مورد نیاز هر مدل تبدیل شدند. این فرآیند با استفاده از تابع apply_chat_template توکنایزر هر مدل انجام شد. داده‌ها به ساختار سه‌بخشی زیر تبدیل شدند:

- Role system: حاوی SYSTEM_PROMPT بود.
- Role user: حاوی content (متن کامل خبر) بود.
- Role assistant: حاوی title (عنوان واقعی هدف) بود.

نکته کلیدی در این بخش، طراحی خود SYSTEM_PROMPT بود. برخلاف refined_prompt (که در مرحله آزمایش بدون فاین-تیون استفاده شد و بسیار طولانی، همراه با مثال و توضیحات صریح در مورد «منطق هسته‌ای» بود)، پرامپت سیستمی استفاده شده در آموزش، به عمد ساده‌تر و عمومی‌تر طراحی شد. هدف این بود که به جای «دیکته کردن» منطق (مانند کاری که refined_prompt انجام داد)، به خود مدل اجازه داده شود تا در طول فرآیند فاین-تیون و از طریق مشاهده صدها نمونه آموزشی، به صورت ضمنی یاد بگیرد که باید بر کدام بخش از محتوای خبر (ادعا یا نتیجه بررسی) تمرکز کند.

۳. ماسک کردن خطا (Training on Responses Only):

هدف ما از آموزش، یادگیری تولید «پاسخ» (Title) بر اساس «دستور» (Prompt) است. مدل نباید برای پیش‌بینی خودِ پرامپت (بخش‌های system و user) آموزش ببیند.

برای این کار، از یک تکنیک کلیدی به نام Loss Masking استفاده شد (که در کد از طریق تابع train_on_responses_only پیاده‌سازی گردید). این تابع، توکن‌های مربوط به پرامپت (مانند <|start_header_id|>user<|end_header_id|>\n\n) را در محاسبه خطا (Loss) نادیده می‌گیرد و محاسبه خطا و بازگشت به عقب (Backpropagation) را فقط روی توکن‌های پاسخ (assistant) متمرکز می‌کند. این کار فرآیند آموزش را بسیار بهینه‌تر و دقیق‌تر می‌سازد.

ب) تنظیمات کلیدی آموزش (Key Training Parameters)

برای آموزش مدل‌ها از SFTTrainer کتابخانه TRL استفاده شد. تنظیمات کلیدی (SFTConfig) که برای هر دو مدل یکسان اعمال شدند، به شرح زیر است:

- num_train_epochs = 4:
  - توضیح: مدل ۴ بار کل دادگان آموزشی را مشاهده کرد. این عدد به عنوان یک نقطه تعادل مناسب برای یادگیری الگوهای دادگان بدون رسیدن به «بیش‌برازش» (Overfitting) انتخاب شد.
- per_device_train_batch_size = 2:
  - توضیح: به دلیل محدودیت شدید حافظه GPU (استفاده از L4 در Colab)، در هر مرحله تنها ۲ نمونه به مدل داده شد.

- gradient_accumulation_steps = 4:
  - توضیح: این پارامتر کلیدی، ضعف batch_size پایین را جبران می‌کند. گرادیان‌ها به جای اعمال فوری، به مدت ۴ مرحله «انباشته» (accumulate) می‌شوند و سپس اعمال می‌گردند. این کار به صورت موثر، Batch Size واقعی را به (۲ * ۴ = ۸)می‌رساند که برای پایداری فرآیند آموزش حیاتی است.
- learning_rate = 1e-4:
  - توضیح: نرخ یادگیری 0.0001 یک نرخ استاندارد و نسبتاً بالا برای آموزش LoRA است. از آنجایی که تنها آداپتورهای کوچک در حال آموزش هستند، می‌توان از نرخ یادگیری بالاتری نسبت به Full Fine-tuning استفاده کرد.
- lr_scheduler_type = "cosine":
  - توضیح: استفاده از زمان‌بند (Scheduler) «کسینوسی» باعث می‌شود نرخ یادگیری در ابتدا کمی گرم (warm up) شده، سپس به اوج برسد و در نهایت به آرامی کاهش یابد. این روش به مدل کمک می‌کند تا در انتهای آموزش در یک «مینیمم محلی» (Local Minima) پایدار همگرا شود.
- gradient_checkpointing = True:
  - توضیح: یک تکنیک حیاتی دیگر برای صرفه‌جویی در حافظه (VRAM) است. این روش، به جای ذخیره تمام activation های میانی در طول forward pass، آن‌ها را در طول backward pass مجدداً محاسبه می‌کند. این کار فرآیند آموزش را کمی کندتر می‌کند اما در عوض، امکان فاین-تیون مدل‌های بزرگتر (مانند 8B) را در حافظه‌های محدود فراهم می‌سازد.

ج) ارزیابی کیفی نتایج فاین-تیون

پس از اتمام فرآیند آموزش، مدل‌های فاین-تیون شده بر روی همان ۶ نمونه‌ی چالش‌برانگیز بخش قبل، ارزیابی شدند.

جدول ۱: مقایسه خروجی مدل Gemma 3 - 4B - IT (فاین-تیون شده)

عنوان واقعی (Dataset)	خروجی مدل Fine-tuned
رحیم مشایی در شمال کشور و دور از مردم زندگی می‌کند و حمید بقایی به بیماری عصبی دچار شده است.	محمود احمدی‌نژاد گفته است که رحیم مشایی در شمال کشور تنهایی زندگی می‌کند و حمید بقایی دچار بیماری عصبی شده است.
آتشفشان دماوند فعال شده است	آتش‌فشان دماوند فعال شد

ایرانی‌ها ۱۸۰ برابر متوسط جهان، شایعات اخبار دروغ و اخبار ناامید کننده دریافت می‌کنند.	ایرانی‌ها ۱۸۰ برابر متوسط جهان، شایعات اخبار دروغ و ناامید کننده دریافت می‌کنند.
مهرجویی بیش از ۲۵ فیلم ساخت که ۷ فیلم او توقیف شد	داریوش مهرجویی بیش از ۲۵ فیلم ساخت که ۷ فیلم آن دچار توقیف شد
دکتر علی لاریجانی گفته است به عرصه‌ی انتخابات مجلس شورای اسلامی ورود می‌کند	علی لاریجانی قصد دارد در انتخابات مجلس شورای اسلامی کاندیدا شود.
امارات سالیانه ۲.۵ میلیون لیتر آب آشامیدانی از هوا تولید می‌کند	امارات سالیانه ۲.۵ میلیون لیتر آب آشامیدانی از هوا تولید می‌کند

جدول ۲: مقایسه خروجی مدل Llama 3.1 - 8B - IT (فاین-تیون شده)

عنوان واقعی (Dataset)	خروجی مدل Fine-tuned
رحیم مشایی در شمال کشور و دور از مردم زندگی می‌کند و حمید بقایی به بیماری عصبی دچار شده است.	محمود احمدی‌نژاد گفته است که رحیم مشایی و حمید بقایی هر دو آزاد هستند
آتشفشان دماوند فعال شده است	آتشفشان دماوند فعال شده است
ایرانی‌ها ۱۸۰ برابر متوسط جهان، شایعات اخبار دروغ و اخبار ناامید کننده دریافت می‌کنند.	ایرانی‌ها ۱۸۰ برابر متوسط جهان، شایعات اخبار دروغ و اخبار ناامید کننده دریافت می‌کنند
مهرجویی بیش از ۲۵ فیلم ساخت که ۷ فیلم او توقیف شد	داریوش مهرجویی ۲۵ فیلم ساخته که ۷ فیلم او دچار توقیف شدند و در مجموع ۲۸ سال فیلم‌های او توقیف بوده است
دکتر علی لاریجانی گفته است به عرصه‌ی انتخابات مجلس شورای اسلامی ورود می‌کند	علی لاریجانی قصد ورود به عرصه انتخابات مجلس شورای اسلامی را دارد.
امارات سالیانه ۲.۵ میلیون لیتر آب آشامیدانی از هوا تولید می‌کند	امارات سالیانه ۲.۵ میلیون لیتر آب آشامیدانی از هوا تولید می‌کند

د) تحلیل کیفی نتایج فاین-تیون

۱. موفقیت چشمگیر در یادگیری «منطق پنهان»:

بزرگترین شکست در مرحله «مهندسی پرامپت» (بخش ۴)، ناتوانی مدل‌ها در درک منطق «عنوان = ادعا» بود (مخصوصاً وقتی ادعا نادرست بود). همانطور که در نمونه‌های «دماوند» و «اخبار نادرست» و «لاریجانی» مشاهده می‌شود، هر دو مدل فاین-تیون شده اکنون به درستی «ادعا» را گزارش می‌کنند، نه «نتیجه بررسی». این نشان می‌دهد که منطق پنهان دادگان با موفقیت توسط آداپتورهای LoRA آموخته شده است.

۲. حل مشکل «تعصب به داده‌ی صحیح»:

در بخش قبل (نمونه «آب امارات»)، دیدیم که هر دو مدل Gemma و Llama، علی‌رغم دستور صریح Refined Prompt، در مقابل «عدد صحیح‌تر» (۴۳ میلیون) مقاومت نکردند و آن را به جای «ادعای» (۲.۵ میلیون) گزارش دادند.

پس از فاین-تیون، هر دو مدل این مشکل را برطرف کرده و به درستی عدد «۲.۵ میلیون لیتر» (که دقیقاً مطابق عنوان واقعی بود) را تولید کردند. این یک پیروزی بزرگ برای فاین-تیون محسوب می‌شود که توانسته سوگیری ذاتی مدل‌ها به سمت «حقیقت» را نادیده گرفته و منطق تسک را بیاموزد.

۳. دقت بالای Gemma 3-4B:

مدل Gemma در ۵ نمونه از ۶ نمونه، خروجی تقریباً بی‌نقصی تولید کرد که یا کاملاً با عنوان واقعی یکسان بود یا (مانند نمونه لاریجانی) از نظر معنایی یکسان بود.

۴. حالات شکست جدید در Llama 3.1-8B:

برخلاف انتظار، مدل بزرگتر Llama 3.1، پس از فاین-تیون، دو مشکل جدید از خود نشان داد:

- تولید محتوای اضافه (Verbose): در نمونه «مهرجویی»، مدل Llama علاوه بر عنوان، جزئیات اضافه («و در مجموع ۲۸ سال...») را نیز به آن اضافه کرد.
- خطای تمرکز (Focus Error): در نمونه «مشایی/بقایی»، مدل Llama به کلی از هدف خارج شد و جمله‌ای («هر دو آزاد هستند») را تولید کرد که نه «ادعا» بود و نه «عنوان واقعی».
  این نشان می‌دهد که مدل Gemma-4Bدر این تسک خاص، به فرآیند فاین-تیون پاسخ پایدارتر و دقیق‌تری نسبت به مدل Llama-8B داده است.

ارزیابی کمی (Quantitative Evaluation)

پس از ارزیابی کیفی در دو بخش قبلی، در این مرحله، عملکرد تمامی روش‌ها (۴ پرامپت مختلف و ۱ مدل فاین-تیون شده) بر روی کل دادگان تست (شامل ۱۱۶ نمونه) به صورت کمی ارزیابی شد.

الف) روش‌شناسی ارزیابی

برای اطمینان از ارزیابی دقیق و قابل اتکا، به خصوص با در نظر گرفتن چالش‌های زبان فارسی، از یک پایپ‌لاین ارزیابی سفارشی‌سازی شده استفاده شد. این فرآیند شامل معیارهای استاندارد خواسته شده در تسک (BLEU, ROUGE-L, METEOR) و همچنین معیار معنایی BERTScore بود.

ملاحظات کلیدی در پیاده‌سازی این ارزیابی عبارت بودند از:

- پیش‌پردازش فارسی: تمامی عناوین واقعی و پیش‌بینی‌شده ابتدا با استفاده از hazm.Normalizer یکسان‌سازی شدند تا مشکلاتی مانند «ی» و «ک» عربی برطرف شوند.
- توکنایزر فارسی:
  - برای معیارهای ROUGE-L و METEOR، از hazm.word_tokenize برای شکستن صحیح کلمات فارسی (مثلاً جداسازی «می‌شود») استفاده شد.
  - برای معیار BLEU، از پیاده‌سازی sacrebleu با توکنایزر flores200 استفاده شد که برای زبان‌های مختلف (از جمله فارسی) عملکرد بسیار قوی‌تری نسبت به توکنایزرهای ساده مبتنی بر فاصله (space-based) دارد.
- ارزیابی معنایی: برای BERTScore، به صراحت "lang="fa مشخص گردید تا از یک مدل BERT اختصاصی آموزش‌دیده بر روی زبان فارسی برای مقایسه شباهت معنایی استفاده شود.

ب) نتایج ارزیابی

نتایج میانگین این چهار معیار برای هر پنج سناریو (از پرامپت ساده تا مدل فاین-تیون شده) در دو مدل Gemma و Llama در جداول زیر خلاصه شده است.

۱. نتایج نهایی مدل Gemma 3 - 4B

🏆 --- FINAL EVALUATION RESULTS: GEMMA-3-4B (Averages) --- 🏆

Gemma-3-4B Prompt Types	BLEU	ROUGE-L	METEOR	BERTScore
predicted_simple	0.1386	0.2665	0.2276	0.7498
predicted_zero_shot	0.1568	0.2987	0.3159	0.7669
predicted_few_shot	0.2172	0.3776	0.3588	0.7939
predicted_refined	0.2401	0.4075	0.3901	0.8069
predicted_fine_tuned	0.3482	0.5215	0.5170	0.8481

۲. نتایج نهایی مدل Llama 3.1 - 8B

🏆 --- FINAL EVALUATION RESULTS: LLAMA-3.1-8B (Averages) --- 🏆

Llama-3.1-8B Prompt Types	BLEU	ROUGE-L	METEOR	BERTScore
predicted_simple	0.1578	0.3001	0.2746	0.7649
predicted_zero_shot	0.1599	0.3165	0.3508	0.7718
predicted_few_shot	0.2123	0.3685	0.3821	0.7908
predicted_refined	0.2646	0.4248	0.4311	0.8106
predicted_fine_tuned	0.3798	0.5419	0.5283	0.8512

ج) تحلیل نتایج کمی

جداول بالا سه روند واضح و مهم را در این پروژه نشان می‌دهند:

۱. ارزش مهندسی پرامپت: در هر دو مدل، یک روند صعودی واضح در تمام معیارها با بهبود پرامپت‌ها دیده می‌شود. حرکت از predicted_simple به predicted_refined (که منطق هسته‌ای دادگان را توضیح می‌داد) منجر به بهبود قابل توجهی شد. برای مثال، در مدل Llama، امتیاز ROUGE-L از 0.3001 به 0.4248 افزایش یافت. این نشان می‌دهد که مهندسی پرامپت دقیق، یک روش کاملاً مؤثر برای بهبود عملکرد مدل‌های پایه است.

۲. برتری مطلق فاین-تیون: جهش عملکردی بین بهترین پرامپت (predicted_refined) و مدل فاین-تیون شده (predicted_fine_tuned) بسیار چشمگیر است.

* در مدل Gemma: امتیاز ROUGE-L از 0.4075 به 0.5215 (جهش ۱۱.۴ واحدی) رسید.

* در مدل Llama: امتیاز ROUGE-L از 0.4248 به 0.5419 (جهش ۱۱.۷ واحدی) رسید.

این نتایج به صورت کمی ثابت می‌کند که روش فاین-تیون (حتی به روش بهینه‌ی LoRA) به طور قابل توجهی از مهندسی پرامپت (Prompt Engineering) برای تخصصی کردن مدل در یک تسک خاص، کارآمدتر است.

۳. مقایسه مدل‌ها (Llama-8B در مقابل Gemma-4B):

مدل Llama 3.1 - 8B در تمامی مراحل ارزیابی، عملکرد بهتری نسبت به مدل Gemma 3 - 4B داشت.

* در حالت پایه (Simple): لاما (ROUGE-L 0.3001) از جما (ROUGE-L 0.2665) قوی‌تر بود.

* در بهترین حالت پرامپت (Refined): لاما (ROUGE-L 0.4248) همچنان از جما (ROUGE-L 0.4075) برتر بود.

* پس از فاین-تیون: لاما (ROUGE-L 0.5419) به سقف عملکردی بالاتری نسبت به جما (ROUGE-L 0.5215) دست یافت.

این نشان می‌دهد که مدل Llama 3.1 هم «پایه‌ی» قوی‌تری دارد و هم «ظرفیت» بیشتری برای یادگیری تسک از طریق فاین-تیون شدن از خود نشان داده است.

نتیجه‌گیری نهایی

این پروژه با هدف ارزیابی و بهینه‌سازی مدل‌های زبانی بزرگ برای تولید عنوان خبری فارسی انجام شد. ما دو رویکرد اصلی را دنبال کردیم: مهندسی پرامپت (بدون آموزش) و فاین-تیون به روش LoRA.

یافته‌های کلیدی این تسک عبارتند از:

۱. مهندسی پرامپت ضروری، اما ناکافی است: تحلیل‌های کیفی (بخش ۴ و ۵) نشان داد که درک «منطق پنهان» دادگان (یعنی استخراج «ادعا» به جای «نتیجه بررسی») حیاتی بود. Refined Prompt توانست این مشکل را تا حدی حل کند، اما در موارد پیچیده (مانند نمونه «آب امارات» و تعصب به اعداد) شکست خورد.

۲. فاین-تیون راه‌حل قطعی است: فرآیند فاین-تیون به روش LoRA، نه تنها امتیازات را در تمام معیارهای کمی به طور چشمگیری افزایش داد، بلکه مهم‌تر از آن، مشکلات منطقی پیچیده‌ای که مهندسی پرامپت قادر به حل آن‌ها نبود را برطرف کرد. مدل‌های فاین-تیون شده یاد گرفتند که منطق تسک را بر سوگیری ذاتی خود برای گزارش «حقیقت» ترجیح دهند.

۳. مدل برتر و گزینه کارآمد: مدل Llama 3.1 - 8B (فاین-تیون شده) با دستیابی به بالاترین امتیازات در تمامی معیارها (از جمله 0.5419 در ROUGE-L و 0.8512 در BERTScore)، به عنوان قوی‌ترین مدل (Best Performer) در این تسک شناخته می‌شود.

با این حال، یافته بسیار مهم این است که مدل Gemma 3 - 4B، با وجود داشتن نصف تعداد پارامترها، عملکردی بسیار نزدیک و رقابتی (ROUGE-L: 0.5215, BERTScore: 0.8481) ارائه داد. این تفاوت اندک در عملکرد (حدود ۲ واحد در ROUGE-L)، مدل Gemma را به گزینه‌ای بسیار کارآمدتر (Most Efficient) تبدیل می‌کند و نشان می‌دهد که در سناریوهایی با محدودیت منابع (چه برای آموزش و چه برای اجرا)، Gemma 3 - 4B می‌تواند انتخابی بهینه‌تر با توازن عالی بین هزینه و عملکرد باشد.