به نام خدا

گزارش پروژه داور هوش مصنوعی

توسط: آرمان خلیلی

مسئول کارفرما: خانم معصومه زارع شرکت: پیام پرداز

تاریخ شروع: 1404/04/7 تاریخ پایان: 1404/04/21

ایمیل: armankhalilieng@gmail.com شماره تماس:+989113687998

فهرست

مقدمه

پیشرفت سریع مدل‌های زبانی بزرگ(LLMs) ، چالش مهمی را به وجود آورده است: چگونه می‌توان کیفیت خروجی‌های این مدل‌ها را به شیوه‌ای کارآمد و پایدار ارزیابی کرد؟ ارزیابی سنتی مبتنی بر انسان، اگرچه استاندارد طلایی محسوب می‌شود، اما فرآیندی کند، پرهزینه و مستعد قضاوت‌های ذهنی است. این موضوع به یک گلوگاه در چرخه توسعه مدل تبدیل شده و مانع از تکرار و بهبود سریع می‌شود. پروژه **»داور هوش مصنوعی (AI Adjudicator)** «برای مقابله با این چالش طراحی شده است تا یک سیستم خودکار برای پیش‌بینی قابل‌اعتماد ترجیحات انسانی در انتخاب بین دو پاسخ رقیبِ تولیدشده توسط هوش مصنوعی ایجاد کند.

هدف اصلی این پروژه، توسعه یک مدل پیش‌بینی‌کننده پیشرفته است که نقش یک «داور» را ایفا می‌کند. این مدل با تحلیل یک پرسش (prompt) و دو پاسخ کاندید(response_a و response_b)، برنده را(A ، Bیا مساوی) با دقت بالا مشخص می‌نماید. برای دستیابی به این هدف، ما یک سیستم چندمرحله‌ای مهندسی کرده‌ایم که داده‌های خام را به یک مدل قدرتمند و غنی از ویژگی‌ها برای ارزیابی تبدیل می‌کند.

این مستند به تشریح سه فاز کلیدی پروژه می‌پردازد:

۱ .**فاز اول: بنیان‌گذاری و پاک‌سازی داده :(Data Foundation & Governance)** در این فاز، ما با پیاده‌سازی یک Pipeline دقیق برای پاک‌سازی و اعتبارسنجی داده، یک (Single Source of Truth) ایجاد می‌کنیم. این فرآیند، داده‌های خام را از چندین مرحله فیلتر—شامل استانداردسازی، اعتبارسنجی محتوا و حذف موارد تکراری از نظر معنایی—عبور می‌دهد تا یک مجموعه داده «طلایی» با خلوص بالا تولید کند که برای آموزش مدل، کاملاً قابل‌اعتماد است.

۲ .**فاز دوم: دسته‌بندی پرامپت ها با روش تقطیر دانش معلم-شاگرد (Teacher-Student Distillation):**با درک این موضوع که همه پرسش‌ها یکسان نیستند، ابتدا یک طبقه‌بند سریع و دقیق برای پرسش‌ها توسعه دادیم. این امر از طریق متدولوژی «معلم-شاگرد» محقق شد؛ در این روش، یک مدل زبان بزرگ و قدرتمند اما کند (Zephyr-7B) به یک مدل کوچک‌تر و سریع‌تر مبتنی بر (RoBERTa)

آموزش می‌دهد تا پرسش‌ها را با دقت بالا به ده دسته مجزا (مانند *اطلاعات عمومی*، *برنامه‌نویسی* یا *محتوای خلاقانه*) طبقه‌بندی کند.

۳ .**فاز سوم: مدل داور - مهندسی ویژگی و مدل‌سازی ترکیبی (Ensemble Modeling):** این فاز، قلب سیستم است. ما مجموعه‌ای جامع شامل بیش از ۲۰ ویژگی برای هر پاسخ مهندسی کرده‌ایم که سیگنال‌های مختلفی از پیچیدگی زبانی و ارتباط معنایی گرفته تا ساختار کد را در بر می‌گیرد. این ویژگی‌ها، نیروی محرکه یک مدل ترکیبی (Ensemble) از مدل‌های LightGBMهستند که شامل یک «مدل جهانی (Universal Model) «برای کاربردهای عمومی و یک «متخصص کد (Code Expert) « برای پرسش‌های برنامه‌نویسی است. این مدل‌ها با هم ترکیب شده تا یک قضاوت نهایی و دقیق ارائه دهند.

این پروژه که بر پایه مجموعه‌ای از فناوری‌های قدرتمند مانند Hugging Face Transformers، PyTorch، SentenceTransformersو LightGBMساخته شده، ابزاری توانمند برای شتاب‌دهی به فرآیند توسعهLLM ها فراهم می‌کند. این مستند به عنوان یک راهنمای جامع برای معماری این سیستم، از مرحله دریافت داده اولیه تا ارزیابی عملکرد تفکیک‌شده مدل نهایی، عمل خواهد کرد.

فاز اول: بنیان‌گذاری و حاکمیت داده(Data Foundation & Governance)

پایه‌ی هر سیستم یادگیری ماشین قابل‌اعتماد، داده‌های باکیفیت و قابل‌اطمینان است. هدف فاز اول، ساخت یک خط لوله (pipeline) خودکار و تکرارپذیر برای تبدیل داده‌های خام و پر از نویز به یک مجموعه داده «طلایی (Golden Dataset) «پاک‌سازی‌شده و نسخه‌بندی‌شده بود. این مجموعه داده طلایی به عنوان «منبع حقیقت واحد (Single Source of Truth) «برای تمام مراحل بعدی مهندسی ویژگی و آموزش مدل عمل کرده و از بروز خطا جلوگیری می‌کند.

۲.۱ .قیف پالایش داده‌ها(The Data Filtration Funnel)

برای دستیابی به این هدف، یک قیف پالایش داده چندمرحله‌ای طراحی شد. هر مرحله مانند یک دروازه کیفیت عمل کرده و داده‌ها را به صورت سیستماتیک پاک‌سازی و فیلتر می‌کند. این فرآیند با بیش از ۵۷,۰۰۰ رکورد خام آغاز شده و با دقت آن‌ها را پالایش می‌کند تا با حذف ورودی‌های کم‌کیفیت و زائد، به یک مجموعه داده نهایی با خلوص بالا دست یابد.

۲.۲ .مراحل خط لوله

مرحله ۱: دریافت و استانداردسازی(Ingestion and Standardization)

خط لوله با خواندن فایل خام Dataset.csvآغاز می‌شود. مراحل اولیه بر پاک‌سازی ساختاری و نرمال‌سازی متمرکز است:

- **تجزیه JSON:** ستون‌های prompt، response_aو response_bکه به صورت رشته‌های JSON ذخیره شده بودند، تجزیه (parse) شده تا محتوای متنی خالص استخراج شود.
- **یکپارچه‌سازی برنده** :ستون‌های برنده که به صورت وان-هات (one-hot) بودند (winner_model_a, winner_model_b, winner_tie)، برای شفافیت بیشتر در یک ستون دسته‌بندی‌شده به نام winnerادغام می‌شوند.
- **حذف موارد تکراری اولیه** :ردیف‌هایی با idتکراری حذف می‌شوند.
- **حذف مقادیر پوچ** :هر ردیفی که در ستون‌های حیاتی دارای داده گمشده باشد، حذف می‌گردد.

این مرحله اولیه تضمین می‌کند که داده‌ها از یک ساختار یکپارچه و قابل‌پیش‌بینی برای پردازش‌های بعدی برخوردار باشند.

مرحله ۲: پاک‌سازی و نرمال‌سازی متن(Text Cleaning and Normalization)

پس از استانداردسازی ساختار داده، این مرحله بر پاک‌سازی محتوای متنی تمرکز دارد. مجموعه‌ای از توابع نرمال‌سازی بر روی هر فیلد متنی اعمال می‌شود:

- **نرمال‌سازی یونیکد** :کتابخانه ftfyبرای رفع مشکلات احتمالی کدبندی متن (mojibake) استفاده می‌شود.
- **حذف تگ‌های HTML**: کتابخانه BeautifulSoupبرای حذف هرگونه تگ HTML باقی‌مانده از متن، که در داده‌های جمع‌آوری‌شده از وب رایج است، به کار می‌رود.
- **حذف URL و فاصله‌های اضافی** :از عبارات باقاعده (Regular Expressions) برای حذف آدرس‌های وب و تبدیل تمام فاصله‌های خالی به یک فاصله واحد استفاده می‌شود تا نویز برای مدل‌های NLP کاهش یابد.

مرحله ۳: فیلتر آماری و اکتشافی(Statistical & Heuristic Filtering)

این مرحله داده‌های پرت آماری را که به احتمال زیاد نمونه‌های کم‌کیفیت هستند، حذف می‌کند. طول متن در ستون‌های prompt، response_aو response_bتحلیل شده و هر ردیفی که طول آن خارج از محدوده **صدک اول و نود و نهم** باشد، حذف می‌شود. این کار به طور مؤثری پاسخ‌های بسیار کوتاه)مانند "ok" یا "نمی‌دانم ("یا پاسخ‌های بسیار طولانی و پر از نویز را که می‌توانند بر آموزش مدل تأثیر منفی بگذارند، حذف می‌کند.

مرحله ۴: اعتبارسنجی زبان و محتوا(Language and Content Validation)

این یک دروازه کیفیت حیاتی است که با استفاده از تحلیل‌های زبانی، اطمینان حاصل می‌کند که هر فیلد متنی حاوی محتوای معنادار است. هر نمونه متنی باید تمام بررسی‌های زیر را با موفقیت پشت سر بگذارد (که با استفاده از langdetectو NLTKانجام می‌شود:(

1. **تشخیص زبان** :متن باید به عنوان زبان انگلیسی (en) شناسایی شود.

2. **نسبت علائم نگارشی** :متن نباید عمدتاً از علائم نگارشی تشکیل شده باشد.
3. **نسبت ایست‌واژه‌ها** :(Stopwords) متن نباید تقریباً به طور کامل از ایست‌واژه‌های رایج تشکیل شده باشد.
4. **ساختار گرامری** :متن باید حداقل شامل یک اسم و یک فعل باشد که از طریق تگ‌گذاری نوع کلمات (POS tagging) شناسایی می‌شود. این روش ساده اما مؤثر، جملات بی‌معنی یا ناقص را فیلتر می‌کند.

مرحله ۵: حذف موارد تکراری از نظر معنایی(Semantic Deduplication)

حذف موارد تکراری ساده کافی نیست، زیرا پرسش‌هایی که بازنویسی شده یا کمی تغییر کرده‌اند را نادیده می‌گیرد و این امر می‌تواند منجر به نشت داده (Data Leakage) بین مجموعه داده‌های آموزش و آزمون شود. این مرحله نهایی فیلترینگ با روش زیر این مشکل را برطرف می‌کند:

1. تولید بردار بازنمایی (embedding) برای تمام پرسش‌ها با استفاده از مدل all-MiniLM-L-6-v2 از کتابخانه .SentenceTransformers
2. استفاده از تابع util.paraphrase_miningبرای یافتن کارآمد جفت پرسش‌هایی که شباهت معنایی بالایی دارند (شباهت کسینوسی ≥ 0.95).
3. حذف یکی از پرسش‌ها از هر جفت تقریبا تکراری که شناسایی شده است.

این کار تضمین می‌کند که مجموعه داده نهایی از مسائل معنایی متمایز تشکیل شده و به یک ارزیابی مدل معتبرتر منجر می‌شود.

۲.۳ .اعتبارسنجی نهایی و خروجی نسخه‌بندی‌شده

پس از عبور از کل قیف پالایش، DataFrameنهایی با استفاده از کتابخانه **pandera**در برابر یک اسکیمای سخت‌گیرانه اعتبارسنجی می‌شود. این مرحله یک «قرارداد داده (Data Contract) «را اعمال می‌کند و تضمین می‌نماید که هر ستون دارای نوع داده صحیح بوده و از محدودیت‌های مشخص‌شده (مانند یکتا بودن idیا تعلق winnerبه یکی از سه مقدار مجاز) پیروی می‌کند.

پس از اعتبارسنجی موفق، مجموعه داده پاک‌سازی‌شده در مسیر high_purity_golden_datasets/به عنوان یک فایل Parquet با برچسب زمانی در نام آن (مانند high_purity_golden_data_v_2025-07-05T14-27-23.parquet) ذخیره می‌شود. این استراتژی نسخه‌بندی، یک تاریخچه غیرقابل‌تغییر و قابل‌ردیابی از داده‌های آموزشی فراهم می‌کند که یکی از الزامات کلیدی حاکمیت داده است.

۲.۴ .گزارش خلاصه پالایش

خط لوله با تولید گزارشی که تأثیر هر مرحله فیلترینگ را کمی‌سازی می‌کند، به پایان می‌رسد. این گزارش یک نمای کلی و شفاف از فرآیند پاک‌سازی داده و نرخ نگه‌داشت کلی داده‌ها ارائه می‌دهد.

مرحله پالایش	تعداد رکوردهای باقی‌مانده	درصد حذف‌شده از مرحله قبل
۰ .رکوردهای خام اولیه	۵۷,۴۷۷	۰.۰۰٪
۱ .پس از دریافت و استانداردسازی	۵۷,۴۳۹	۰.۰۷٪
۲ .پس از نرمال‌سازی متن	۵۷,۴۳۹	۰.۰۰٪
۳ .پس از فیلتر آماری	۵۴,۱۹۱	۵.۶۵٪
۴ .پس از اعتبارسنجی محتوا	۴۵,۶۶۸	۱۵.۷۳٪
۵ .پس از حذف تکرارهای معنایی (نهایی)	**۳۹,۹۶۶**	**۱۲.۴۹٪**

مجموعه داده «طلایی» نهایی **۶۹.۵۳٪** از رکوردهای اصلی را حفظ کرده است که نشان‌دهنده زیرمجموعه باکیفیتی است که برای ساخت مدل «داور هوش مصنوعی» استفاده می‌شود.

فاز دوم: طبقه‌بندی پرسش با روش تقطیر دانش معلم-شاگرد

یک تصمیم معماری کلیدی برای پروژه «داور هوش مصنوعی»، فراتر رفتن از یک مدل یکپارچه (monolithic)و حرکت به سمت رویکرد **«ترکیبی از متخصصان (Mixture of Experts - MoE)»** است. این طراحی به مدل‌های «متخصص» اجازه می‌دهد تا حوزه‌های خاص (مانند کدنویسی یا محتوای خلاقانه) را با دقت بالاتری پردازش کنند. یک پیش‌نیاز حیاتی برای این معماری، وجود یک سیستم سریع و قابل‌اعتماد برای طبقه‌بندی پرسش‌های ورودی و هدایت آن‌ها به متخصص مربوطه است. فاز دوم به طور کامل بر ساخت این جزء کلیدی متمرکز است.

برای جلوگیری از فرآیند پرهزینه و زمان‌بر برچسب‌گذاری دستی هزاران پرسش، ما یک استراتژی پیشرفته **تقطیر دانش معلم-شاگرد (Teacher-Student Distillation)** را پیاده‌سازی کردیم.

۳.۱ .پارادایم معلم-شاگرد

این پارادایم از یک مدل بزرگ و قدرتمند «معلم» برای تولید برچسب‌های باکیفیت استفاده می‌کند و سپس از این برچسب‌ها برای آموزش یک مدل «شاگرد» کوچک‌تر و سریع‌تر بهره می‌برد که برای استفاده در محیط عملیاتی مناسب است.

- **مدل معلم**HuggingFaceH4/zephyr-7b-beta :

یک مدل قدرتمند ۷ میلیارد پارامتری به دلیل توانایی بالا در استدلال و پیروی از دستورالعمل‌ها به عنوان معلم انتخاب شد. برای اجرای این مدل در محدودیت‌های یک GPU ابری استاندارد، از روش کوانتیزه‌سازی ۴ بیتی با کتابخانه bitsandbytesاستفاده شد که به طور قابل‌توجهی حافظه مورد نیاز را بدون افت شدید عملکرد کاهش داد.

- **مدل شاگرد**roberta-base :

برای مدل شاگرد، roberta-baseانتخاب شد که یک مدل ترنسفورمر بسیار کوچک‌تر اما بسیار کارآمد است. اندازه جمع‌وجور آن، سرعت استنتاج (inference) بسیار بالایی را تضمین می‌کند و آن را برای وظیفه طبقه‌بندی آنی که باید قبل از منطق اصلی داوری انجام شود، ایده‌آل می‌سازد.

هدف این فاز، «تقطیر» دانش طبقه‌بندی از مدل بزرگ Zephyr به مدل چابک RoBERTa است.

۳.۲ .فرآیند برچسب‌گذاری: وظیفه معلم

وظیفه مدل معلم، طبقه‌بندی پرسش‌های استخراج‌شده از مجموعه داده «طلایی» (که در فاز اول ایجاد شد) به یکی از ده دسته از پیش تعریف‌شده بود. برای اطمینان از کیفیت بالا و خروجی ساختاریافته، یک پرامپت مبتنی بر **زنجیره-تفکر (Chain-of-Thought - CoT)** مهندسی شد. این پرامپت نه تنها درخواست طبقه‌بندی را مطرح می‌کرد، بلکه مدل را ملزم می‌نمود تا استدلال (Reasoning)و یک امتیاز اطمینان (confidence)را در یک قالب JSON کاملاً مشخص ارائه دهد.

نمونه‌ای از ساختار پرامپت:CoT

```
: You are an expert text classification assistant. First, provide a brief
`Reasoning`. Then, on a new line, provide ONLY a valid JSON object.

Categories:
- Code & Programming: ...
- Creative Content: ...
...

Output Format:
Reasoning: [Your one-sentence reasoning for the classification.]
{"prompt_type": "The Best-Fit Category", "confidence": 0.0-1.0}

User: [The prompt to be classified]
```

این رویکرد ساختاریافته به ما امکان داد تا خروجی مدل را به صورت برنامه‌ریزی‌شده تجزیه کرده و فقط برچسب‌های با اطمینان بالا (اطمینان ≥ 0.80) را فیلتر کنیم.

به دلیل محدودیت‌های سخت‌افزاری در محیط اجرا، فرآیند برچسب‌گذاری پس از تولید **۳۰۰ نمونه با اطمینان بالا** به صورت دستی متوقف شد. این مجموعه داده کوچک‌تر و گزینش‌شده سپس برای آموزش مدل شاگرد مورد استفاده قرار گرفت. توزیع کلاس‌ها در این مجموعه نامتوازن بود که نتیجه مستقیم توزیع طبیعی داده‌ها در منبع اصلی است:

- **پرسش‌های فنی** :۱۰۴ نمونه
- **اطلاعات واقعی** :۸۰ نمونه

- **تبدیل داده/محتوا** :۳۷ نمونه
- **محتوای خلاقانه** :۲۴ نمونه
- **مشاوره و راهنمایی شخصی** :۱۶ نمونه
- **کدنویسی و برنامه‌نویسی** :۱۵ نمونه
- **معماهای ریاضی و منطقی** :۱۱ نمونه
- **نقش‌آفرینی و شخصیت‌پردازی** :۶ نمونه
- **تجاری و حرفه‌ای** :۵ نمونه
- **تعامل عمومی** :۲ نمونه

۳.۳ .فرآیند تقطیر: آموزش شاگرد

با در اختیار داشتن داده‌های برچسب‌گذاری‌شده توسط معلم، به سراغ آموزش مدل شاگرد رفتیم.

۱ .**تقسیم داده‌ها** :مجموعه داده ۳۰۰ نمونه‌ای به دو بخش آموزش (۲۴۰ نمونه) و آزمون (۶۰ نمونه) تقسیم شد. نکته حیاتی این است که این تقسیم به صورت **طبقه‌بندی‌شده (stratified)** بر اساس prompt_typeانجام شد تا اطمینان حاصل شود که عدم توازن کلاس‌ها به نسبت در هر دو مجموعه حفظ می‌شود و ارزیابی واقعی‌تری صورت می‌گیرد.
۲ .**تنظیم دقیق (Fine-Tuning)**: مدل roberta-baseبه مدت ۳ دوره (epoch) با استفاده از کتابخانه Trainerدر Hugging Face آموزش داده شد. از هایپرپارامترهای استاندارد و آموزش با دقت ترکیبی (fp16)برای افزایش سرعت استفاده گردید.

۳.۴ .نتایج و خروجی‌های نهایی

پس از آموزش، مدل شاگرد بر روی مجموعه آزمون ارزیابی شد. نتایج، چالش قابل‌توجه آموزش بر روی یک مجموعه داده کوچک و نامتوازن را به خوبی نشان می‌دهد:

متریک	مقدار
زیان ارزیابی(Evaluation Loss)	1.727
دقت(Accuracy)	0.3667
امتیاز F1 وزن‌دار	0.2151

تفسیر نتایج :دقت و امتیاز F1 پایین مدل، نتیجه‌ای قابل‌پیش‌بینی با توجه به داده‌های آموزشی محدود است. مدل الگوهایی را آموخته، اما به طور قابل‌توجهی کم‌آموزش‌دیده (under-trained) است و در کلاس‌های کم‌تعداد با چالش مواجه است. با این حال، این فاز با موفقیت امکان‌پذیری خط لوله تقطیر دانش معلم-شاگرد را اثبات می‌کند. انتظار می‌رود با تخصیص منابع محاسباتی بیشتر برای تکمیل فرآیند برچسب‌گذاری، عملکرد مدل شاگرد به طور چشمگیری افزایش یابد.

خروجی‌های اصلی این فاز عبارتند از:

۱ .**مجموعه داده برچسب‌گذاری‌شده** :فایل roberta_teacher_labeled.csvکه شامل ۳۰۰ نمونه استفاده‌شده برای آموزش است.

۲ .**مدل طبقه‌بند** :مدل roberta-10-class-classifierکه به صورت یک آرتیفکت ذخیره شده و برای استفاده در فاز سوم آماده است.

فاز سوم: مدل داور - مهندسی ویژگی و مدل‌سازی ترکیبی

این فاز، قلب پروژه «داور هوش مصنوعی» را تشکیل می‌دهد؛ جایی که تمام کارهای زیربنایی از فازهای قبل، به آموزش و ارزیابی مدل پیش‌بینی‌کننده نهایی ختم می‌شود. هدف اصلی، ساخت مدلی است که بتواند با دقت بالا و بر اساس مجموعه‌ای غنی از ویژگی‌های مهندسی‌شده، برنده را بین دو پاسخ تولیدشده توسط هوش مصنوعی پیش‌بینی کند. این فاز در دو پیکربندی اجرا شد:

- **خط لوله کوتاه (USE_FULL_PIPELINE = False):** یک اجرای سریع و متمرکز بر توسعه که از مجموعه داده کوچک و از پیش برچسب‌گذاری‌شده (۳۰۰ نمونه) فاز دوم استفاده می‌کند.
- **خط لوله کامل (USE_FULL_PIPELINE = True):** شبیه‌سازی خط لوله نهایی، با استفاده از زیرمجموعه بزرگ‌تری از داده‌های طلایی (۵۰۰ نمونه) و طبقه‌بندی آنی پرسش‌ها.

۴.۱ .مهندسی ویژگی پیشرفته

قدرت پیش‌بینی مدل داور به مجموعه‌ای جامع و معنادار از ویژگی‌ها وابسته است. ما یک استخراج‌کننده ویژگی توسعه دادیم که معیارهای متنوعی را برای هر پاسخ محاسبه می‌کند. سپس از این معیارها برای محاسبه تفاوت یا «دلتا (Delta) «بین دو پاسخ استفاده می‌شود. این رویکرد **ویژگی‌های دلتا(Δ-feature)** ، محور اصلی مدل‌سازی است، زیرا مستقیماً مسئله را به شکل یک مقایسه درمی‌آورد.

ویژگی‌ها در چند دسته طبقه‌بندی می‌شوند:

- **ویژگی‌های عمومی و خوانایی (Universal & Readability):** آمارهای متنی پایه که نمایی کلی از پاسخ ارائه می‌دهند.
    - *ویژگی‌ها:* تعداد کلمات، تعداد جملات، سطح خوانایی Flesch-Kincaid.
- **ویژگی‌های معنایی و ارتباط (Semantic & Relevance):** ویژگی‌های پیشرفته NLP که رابطه معنایی بین پرسش و پاسخ را می‌سنجند.

- *مدل‌های مورد استفاده:* SBERT و BGE-largeبرای شباهت کسینوسی، Cross-Encoderبرای امتیازدهی ارتباط، و BERTScoreبرای امتیاز F1 مبتنی بر شباهت توکن‌ها.

- **تحلیل تخصصی کد (Code-Specific Analysis):** برای ساخت یک متخصص واقعی، ساختار کدهای موجود در پاسخ‌ها را تحلیل می‌کنیم. به جای وابستگی به ابزارهای پیچیده مانند tree-sitter، از ترکیبی عمل‌گرایانه از ماژول داخلی پایتون ast(درخت نحو انتزاعی) و کتابخانه radonاستفاده کردیم.
  - *ویژگی‌ها:* پیچیدگی سایکلوماتیک، تعدادimport ها، تعداد توابع و یک پرچم باینری برای وجود کد.

- **تحلیل استدلال و موجودیت‌ها (Argumentative & Entity Analysis):** ویژگی‌هایی برای درک ساختار استدلال و هم‌پوشانی مفاهیم کلیدی.
  - *ویژگی‌ها:* تعداد نشانگرهای گفتمانی) مانند ("however", "therefore" و شباهت جاکارد موجودیت‌های نام‌گذاری‌شده (استخراج‌شده توسط spaCy)بین دو پاسخ.

۴.۲ .معماری ترکیبی از متخصصان(Ensemble of Experts)

برای مدیریت ماهیت متنوع پرسش‌ها، ما یک معماری **ترکیبی از متخصصان** با استفاده از LightGBM پیاده‌سازی کردیم LightGBM .یک چارچوب گرادیان بوستینگ سریع و کارآمد است که برای داده‌های جدولی بسیار مناسب است.

۱ .**مدل جهانی (Universal Model):** این مدل بر روی کل مجموعه داده آموزشی با استفاده از ویژگی‌های عمومی (تمام ویژگی‌ها به جز موارد تخصصی کد) آموزش داده می‌شود. این مدل به عنوان یک پایه قدرتمند عمل کرده و تمام پیش‌بینی‌های غیرمرتبط با کد را انجام می‌دهد.
۲ .**متخصص «کدنویسی و برنامه‌نویسی»:** این مدل تخصصی *فقط* بر روی پرسش‌هایی که به عنوان Code & Programmingطبقه‌بندی شده‌اند، آموزش می‌بیند. این مدل از مجموعه کامل ویژگی‌ها، از جمله ویژگی‌های غنی تحلیل کد، استفاده می‌کند که به آن امکان می‌دهد درک عمیق‌تری از ویژگی‌های

یک پاسخ کد خوب پیدا کند.

۳ .**منطق پیش‌بینی** :هنگام ارزیابی یک نمونه جدید، سیستم ابتدا دسته آن را بررسی می‌کند.

اگر دسته Code & Programmingباشد و مدل متخصص کد با موفقیت آموزش دیده باشد، پیش‌بینی نهایی یک میانگین وزنی از احتمالات خروجی مدل جهانی و مدل متخصص خواهد بود (با وزن بالاتر ۰.۷ برای متخصص).

برای سایر دسته‌ها، پیش‌بینی مدل جهانی به طور مستقیم استفاده می‌شود.

یک عنصر حیاتی در فرآیند آموزش، استفاده از GroupKFoldبر اساس ستون promptاست. این کار تضمین می‌کند که تمام داده‌های مربوط به یک پرسش واحد، یا در مجموعه آموزش یا در مجموعه آزمون قرار می‌گیرند. این یک محافظ مهم در برابر **نشت داده** است و ارزیابی صادقانه‌تری از توانایی تعمیم مدل فراهم می‌کند.

۴.۳ .ارزیابی مدل و تحلیل نتایج

مدل در هر دو پیکربندی "کوتاه" و "کامل" آموزش و ارزیابی شد.

الف) نتایج خط لوله کوتاه(USE_FULL_PIPELINE = False)

- **داده‌ها** :۳۰۰ نمونه برچسب‌دار، تقسیم‌شده به ۲۴۰ نمونه آموزش و ۶۰ نمونه آزمون.
- **دقت کلی** :٪۱۰۰.۰۰

گزارش طبقه‌بندی:

	precision	recall	f1-score	support
B Wins	1.00	1.00	1.00	16
Tie	1.00	1.00	1.00	21
A Wins	1.00	1.00	1.00	23
accuracy			1.00	60

عملکرد تفکیک‌شده بر اساس دسته:

category	Accuracy	Test_Sample_Count
Business & Professional	1.0	1
Code & Programming	1.0	4
Creative Content	1.0	4
Data/Content Transformation	1.0	7
Factual Information	1.0	17
Math & Logic Puzzles	1.0	3
Personal Advice & Guidance	1.0	2
Technical Inquiry	1.0	22

ب) نتایج خط لوله کامل(USE_FULL_PIPELINE = True)

- **داده‌ها** : ۵۰۰ نمونه طبقه‌بندی‌شده، تقسیم‌شده به ۴۰۰ نمونه آموزش و ۱۰۰ نمونه آزمون.
- دقت کلی :٪۱۰۰.۰۰

گزارش طبقه‌بندی:

	precision	recall	f1-score	support
B Wins	1.00	1.00	1.00	36
Tie	1.00	1.00	1.00	32
A Wins	1.00	1.00	1.00	32

accuracy	1.00	100

عملکرد تفکیک‌شده بر اساس دسته:

Category	Accuracy	Test_Sample_Count
Factual Information	1.0	7
Technical Inquiry	1.0	93

تحلیل دقت ٪۱۰۰: شناسایی یک راه‌حل ساده‌انگارانه

اگرچه دقت ٪۱۰۰ در نگاه اول یک موفقیت برجسته به نظر می‌رسد، اما میتواند نشان دهنده یک مشکل باشد، که اغلب به مشکلات اساسی در داده‌ها اشاره دارد. بررسی دقیق‌تر تأیید می‌کند که در اینجا نیز همین‌طور است.

امتیازات کامل، نشان‌دهنده آموزش و آزمون بر روی **زیرمجموعه‌های داده‌ای کوچک، همگن و احتمالاً غیرنماینده** هستند. مدل LightGBM، به دلیل کارایی بالا، الگوهای ساده و قطعی ("میان‌برهایی") را که در این مجموعه داده‌های کوچک وجود دارد، کشف کرده است. برای مثال، ممکن است یک ویژگی مانند delta_word_count برای رسیدن به دقت ٪۱۰۰ در این مجموعه آزمون خاص کافی باشد.

این پدیده نوعی **بیش برازش (Overfitting)** بر روی ویژگی‌های تصادفی یک مجموعه داده محدود است. اگرچه استفاده از GroupKFold از نشت مستقیم پرسش جلوگیری می‌کند، اما نمی‌تواند مشکل بنیادین یک مسئله ساده‌انگارانه ناشی از خود داده‌ها را حل کند. مدل، معنای پیچیده یک پاسخ «خوب» را یاد نگرفته، بلکه صرفاً ساده‌ترین مسیر را در داده‌های موجود پیدا کرده است.

نتیجه‌گیری :معماری خط لوله و منطق مهندسی ویژگی‌ها صحیح است. با این حال، دقت ٪۱۰۰ گزارش‌شده یک نتیجه مصنوعی ناشی از داده‌های محدود استفاده‌شده در این اجراها است. برای به دست آوردن یک معیار واقعی از عملکرد مدل، باید آن را بر روی مجموعه داده کامل و متنوع تولید شده در فاز اول آموزش داد و ارزیابی کرد .

خلاصه نتایج و ارزیابی

این پروژه با یک رویکرد متوالی و چندمرحله‌ای ساختار یافته است که در آن، خروجی هر فاز به عنوان ورودی حیاتی برای فاز بعدی عمل می‌کند. بنابراین، ارزیابی سیستم نهایی باید به عنوان یک ارزیابی جامع از تمام اجزای آن، از کیفیت داده گرفته تا دقت پیش‌بینی نهایی، در نظر گرفته شود.

- **فاز اول: بنیان‌گذاری داده** :خط لوله پالایش داده بسیار مؤثر عمل کرد و یک مجموعه داده خام با **۵۷,۴۷۷** رکورد را به یک مجموعه داده «طلایی» با خلوص بالا و **۳۹,۹۶۶** رکورد تبدیل نمود. این نرخ نگه‌داشت **۶۹.۵۳٪**، نشان‌دهنده حذف موفقیت‌آمیز موارد تکراری، متون غیرانگلیسی و ورودی‌های ناقص گرامری است. فایل Parquet نسخه‌بندی‌شده حاصل، یک پایه پایدار و قابل‌اعتماد برای تمام وظایف بعدی فراهم می‌کند.

- **فاز دوم: طبقه‌بندی پرسش** :خط لوله تقطیر دانش معلم-شاگرد با موفقیت پیاده‌سازی شد. با این حال، به دلیل محدودیت‌های محاسباتی، مدل» LLM معلم» تنها برای تولید **۳۰۰** برچسب با اطمینان بالا استفاده شد، در حالی که برای یک آموزش قوی به هزاران نمونه نیاز بود. در نتیجه مستقیم این محدودیت، عملکرد مدل «شاگرد RoBERTa «بر روی مجموعه آزمون، متوسط بود و به **امتیاز F1 وزن‌دار ۰.۲۱۵** دست یافت. اگرچه دقت پایین است، اما این فاز با موفقیت کارایی جریان کاری تقطیر دانش را تأیید کرد و انتظار می‌رود با تکمیل فرآیند برچسب‌گذاری، عملکرد طبقه‌بند به شدت بهبود یابد.

- **فاز سوم: مدل ترکیبی داور** :مدل نهایی داور تحت دو سناریو آزمایش شد که هر دو تحت تأثیر مجموعه داده‌های کوچک موجود بودند:
  - **خط لوله کوتاه (۳۰۰ نمونه)** :به **دقت ۱۰۰٪** بر روی مجموعه آزمون ۶۰ نمونه‌ای خود دست یافت.
  - **خط لوله کامل (۵۰۰ نمونه)** :نیز به **دقت ۱۰۰٪** بر روی مجموعه آزمون ۱۰۰ نمونه‌ای خود رسید.

همانطور که در بخش قبل توضیح داده شد، این امتیازات کامل نشان‌دهنده یک مدل بی‌نقص نیستند، بلکه حاکی از **بیش برازش (overfitting) بر روی یک مجموعه داده کوچک و غیرنماینده** هستند. مدل LightGBMتوانسته است میان‌برهای ساده و قطعی موجود در داده‌ها را پیدا کند که یک پدیده رایج هنگام آموزش یک مدل قدرتمند بر روی یک فضای مسئله محدود و همگن است. نکته کلیدی، خود امتیاز دقت نیست، بلکه نمایش موفقیت‌آمیز مهندسی ویژگی پیشرفته و معماری ترکیبی از متخصصان است. خط لوله به درستی یک مدل جهانی و یک «متخصص کد» را آموزش داد و با هم ترکیب کرد و مفهوم معماری را اثبات نمود.

راه‌حل ایده‌آل برای جلوگیری از بیش برازش (overfitting) مشاهده‌شده، آموزش مدل‌ها بر روی یک مجموعه داده بسیار بزرگ‌تر و متنوع‌تر است. با این حال، این امر به دلیل محدودیت‌های ذاتی پلتفرم Google Colabامکان‌پذیر نبود. استفاده طولانی‌مدت از منابع پردازشی قدرتمند مانند T4 GPU به طور مکرر منجر به قطعی کرنل (kernel) و توقف اجرا می‌شد و مانع از تکمیل فرآیند آموزش بر روی کل مجموعه داده می‌گردید. از سوی دیگر، اجرای این وظایف سنگین مهندسی ویژگی و آموزش بر روی CPU، به شدت زمان‌بر و غیرعملی بود.

در تلاش برای کاهش این محدودیت‌ها در فاز دوم، زمان قابل‌توجهی صرف آزمایش API مدل‌های LLM خارجی (شامل OpenAIاز GPT-4o-mini ، TogetherAIزDeepSeek-V3 ، Llama3از Groq و Gemini از Google) برای وظیفه برچسب‌گذاری داده‌ها شد. متأسفانه این تلاش‌ها نیز موفقیت‌آمیز نبودند، زیرا محدودیت‌های استفاده در سطح رایگان هر یک از این سرویس‌ها به سرعت به پایان می‌رسید و از تولید یک مجموعه داده برچسب‌گذاری‌شده به اندازه کافی بزرگ جلوگیری می‌کرد. بنابراین، مدل‌های فاز سوم به ناچار بر روی زیرمجموعه‌های محدودی از داده‌ها آموزش دیدند و نتایج باید به عنوان یک **اثبات مفهوم (Proof of Concept)**برای معماری سیستم تفسیر شوند، نه یک معیار قطعی از عملکرد بالقوه آن.

نتیجه‌گیری کلی از نتایج :پروژه با موفقیت اجزای معماری کلیدی سیستم «داور هوش مصنوعی» را ساخت و اعتبارسنجی کرد. بنیان داده قوی است، مهندسی ویژگی جامع است و چارچوب مدل‌سازی ترکیبی در جای خود قرار دارد. گلوگاه فعلی عملکرد، به وضوح اندازه و تنوع داده‌های آموزشی مورد استفاده در مراحل نهایی مدل‌سازی شناسایی شده است که نتیجه مستقیم محدودیت منابع در اجرای فازهای ۲ و ۳ می‌باشد.

نتیجه‌گیری و بهبودهای آینده

پروژه «داور هوش مصنوعی» با موفقیت به هدف اصلی خود یعنی طراحی و پیاده‌سازی یک سیستم کامل برای ارزیابی خودکار کیفیت پاسخ‌های رقیب LLM دست یافته است. در طول سه فاز متمایز، ما یک خط لوله داده قدرتمند، یک مجموعه ویژگی پیچیده و یک معماری مدل‌سازی ترکیبی از متخصصان قابل‌توسعه ایجاد کردیم. این پروژه، امکان‌پذیری استفاده از یک مدل سبک مبتنی بر ویژگی مانند LightGBMرا برای بازتولید مؤثر قضاوت انسانی به اثبات رساند.

نوآوری‌های کلیدی این پروژه عبارتند از:

- **قیف پالایش داده چندمرحله‌ای** :ایجاد یک مجموعه داده «طلایی» قابل‌اعتماد با حذف سیستماتیک نویز و موارد تکراری معنایی.
- **تقطیر دانش معلم-شاگرد به صورت عملی** :ایجاد یک جریان کاری برای انتقال دانش از یک LLMبزرگ به یک طبقه‌بند سریع و آماده برای محیط عملیاتی.
- **مجموعه قابل‌توسعه از متخصصان** :ساخت یک چارچوب که یک مدل عمومی را با متخصصان حوزه‌های خاص ترکیب می‌کند، همراه با یک اثبات مفهوم موفق برای پرسش‌های «کدنویسی و برنامه‌نویسی.«
- **اعتبارسنجی دقیق** :پیاده‌سازی GroupKFoldبرای جلوگیری از نشت داده و اطمینان از اینکه ارزیابی مدل، قوی و قابل‌اعتماد است.

اگرچه نتایج اولیه بر روی زیرمجموعه‌های کوچک داده به طور مصنوعی بالا بود، اما معماری زیربنایی، مستحکم و آماده برای مقیاس‌پذیری است. مسیر پیش رو روشن و همسو با نقشه راه اولیه پروژه است و بر مقیاس‌دهی داده‌ها، عملیاتی‌سازی سیستم و ایجاد یک چرخه بهبود مستمر تمرکز دارد.

نقشه راه و بهبودهای آینده:

مراحل زیر، مسیر استراتژیک برای تکامل از نمونه اولیه فعلی به یک سیستم کاملاً عملیاتی و در سطح سازمانی را مشخص می‌کند.

۱ .**مقیاس‌دهی داده و آموزش مجدد (اولویت فوری):**

تکمیل برچسب‌گذاری داده‌ها :تخصیص منابع GPU لازم برای اجرای کامل خط لوله برچسب‌گذاری فاز ۲ بر روی کل مجموعه داده طلایی ۳۹,۹۶۶ رکوردی. این مهم‌ترین گام بعدی برای آموزش یک مدل واقعاً قابل‌تعمیم است.

آموزش مجدد مدل داور :با استفاده از مجموعه داده کاملاً برچسب‌گذاری‌شده، مدل‌های جهانی و متخصص از فاز ۳ را مجدداً آموزش داده تا یک معیار عملکرد واقعی به دست آید.

۲ .**بهبودهای مدل (گسترش فاز ۳):**

گسترش مدل ترکیبی :توسعه مدل‌های متخصص بیشتر برای سایر دسته‌های کلیدی شناسایی‌شده در فاز ۲، مانند «محتوای خلاقانه»، «اطلاعات واقعی» و «پرسش‌های فنی.«

کاوش در معماری‌های پیشرفته :حرکت از LightGBMبه سمت مدل‌های مبتنی بر ترنسفورمر که با یک تابع زیان رتبه‌بندی زوجی (pairwise ranking loss) آموزش داده می‌شوند و ممکن است بتوانند تفاوت‌های معنایی عمیق‌تری را درک کنند.

۳ .**عملیاتی‌سازی و استقرار (فاز ۴ پیش رو):**

کانتینرسازی مدل :بسته‌بندی کل خط لوله مدل (طبقه‌بند و مدل ترکیبی داور) در یک کانتینر داکر با استفاده از یک چارچوب ارائه خدمات مانند BentoMLیا .FastAPI

استقرار به عنوان میکروسرویس :استقرار کانتینر بر روی یک زیرساخت مقیاس‌پذیر) مانند Kubernetesیا (AWS SageMaker Endpoints برای ارائه یک API نسخه-بندی‌شده برای استنتاج آنی.

آزمون بار :(Load Testing) آزمودن API تحت فشار برای اطمینان از برآورده کردن الزامات تأخیر (latency)و توان عملیاتی (throughput) برای موارد استفاده عملیاتی مانند.RLHF

۴ .**عملیات زنده و بهبود مستمر (فاز ۵ پیش رو):**

پیاده‌سازی نظارت :(Monitoring) راه‌اندازی داشبوردها) مثلاً با Grafana و (Evidently AI برای نظارت بر معیارهای عملیاتی و به طور حیاتی، نظارت بر انحراف داده و مفهوم.(data and concept drift)

ایجاد یک سیستم انسان-در-حلقه :(HITL) ایجاد فرآیندی برای ارسال پیش‌بینی‌های با اطمینان پایین

به ارزیابان انسانی. این بازخورد برای اندازه‌گیری دقت واقعی در محیط زنده و ایجاد داده‌های باکیفیت برای آموزش مجدد مستمر استفاده خواهد شد و تضمین می‌کند که مدل در طول زمان سازگار باقی می‌ماند.

چارچوب آزمون A/B: پیاده‌سازی یک چارچوب قهرمان/رقیب (champion/challenger) برای آزمایش و استقرار ایمن نسخه‌های جدید مدل داور بدون ایجاد اختلال در سرویس.

با دنبال کردن این نقشه راه، «داور هوش مصنوعی» می‌تواند از یک اثبات مفهوم موفق به یک ابزار عملیاتی اصلی تبدیل شود که توسعه LLM را شتاب بخشیده و هزینه‌ها و زمان مرتبط با ارزیابی دستی را به میزان قابل‌توجهی کاهش می‌دهد.