گزارش طراحی و ارزیابی Multimodal RAG System

تمرین سوم درس پردازش زبان های طبیعی

علی شریفی 403206532

رئوفه رضایی 403207012

نسرین صادقلو 402208954

یاسان حسن زاد 403206995

بهار 1403

فهرست مطالب

مقدمه

با افزایش داده‌های چندرسانه‌ای در حوزه‌هایی مانند فرهنگ، تاریخ و آموزش، نیاز به سیستم‌هایی که بتوانند متن و تصویر را به‌صورت یکپارچه پردازش کنند اهمیت زیادی یافته است. مدل‌های چندوجهی با ترکیب داده‌های متنی و تصویری بازنمایی‌های دقیق‌تری ارائه می‌دهند و عملکرد بهتری در جستجو و پاسخ‌گویی دارند.

یکی از رویکردهای کلیدی در این زمینه، بازیابی تقویت‌شده با تولید (RAG) است که در آن مدل بازیاب به یک مدل مولد متصل می‌شود تا پاسخ‌ها علاوه بر مدل زبانی، بر داده‌های بازیابی‌شده نیز تکیه داشته باشند. در این پروژه پس از گردآوری و پیش‌پردازش داده‌ها، یک سیستم چندوجهی مبتنی بر RAG پیاده‌سازی شده و در نهایت مجموعه‌ای از پرسش‌های ارزیابی روی این سیستم و همچنین روی دو مدل چندوجهی پایه اجرا و مقایسه می‌گردد.

جمع آوری داده ها

فرآیند جمع‌آوری داده‌ها در این پروژه با هدف ایجاد یک مجموعه‌ی چندوجهی برای حوزه‌ی منابع طبیعی و سایت‌های گردشگری انجام گرفته است. داده‌ها به دو بخش اصلی تقسیم شده‌اند: داده‌های آموزش و داده‌های ارزیابی.

داده آموزش

برای آموزش مدل، مجموعه‌ای شامل ۱۱۰۰ نمونه داده گردآوری و سازمان‌دهی شده است. این داده‌ها از منابع عمومی و معتبر از جمله Wikipedia و کتاب‌های جغرافیای استانی استخراج شده‌اند. در فرآیند انتخاب، تمرکز اصلی بر انسجام معنایی میان متن و تصویر بوده است تا مجموعه بتواند مبنای مناسبی برای آموزش یک مدل چندوجهی دقیق باشد. داده‌ها از استان‌های اصفهان، فارس، بوشهر، چهارمحال و بختیاری، هرمزگان و کهگیلویه و بویراحمد گردآوری شده‌اند.

هر نمونه داده در قالب ساختار JSON ذخیره گردیده و به‌صورت زیر تعریف می‌شود:

```
{
   "id":1 ,
   "context": "",
   "image": ""
}
```

این داده‌ها در Huggingface تحت عنوان alisharifi/tourist-attractions-text-image قرارداده‌شده‌اند. به تصاویر نیز از طریق این لینک می‌توانید دسترسی داشته باشید.

داده ارزیابی

به‌منظور سنجش عملکرد مدل، مجموعه‌ای از سؤالات چندگزینه‌ای طراحی و سازمان‌دهی شده است. این مجموعه شامل دو نوع سؤال است:

1. سوالات متنی به تعداد ۵۰
2. سوالات متن و تصویر به تعداد ۳۰

تمام سؤالات طراحی‌شده در دو دسته‌ی موضوعی اصلی قرار دارند تا دقت مدل در هر حوزه به‌طور جداگانه ارزیابی شود:

1. جاذبه های طبیعی
2. جاذبه های ساخت انسان

آموزش مدل

پیش‌پردازش داده‌ها

متون فارسی موجود در دیتاست با استفاده از کتابخانه‌ی Hazm نرمال شدند. این مرحله شامل یکپارچه‌سازی کاراکترها، حذف فاصله‌های اضافی و اصلاح علائم نگارشی بود.

تنظیمات آموزش

پارامترهای کلیدی آموزش به شرح زیر تعیین شدند:

- تعداد داده بازیابی شده از دیتابیس: ۳ عدد به صورت پیش‌فرض
- استفاده از میانگین‌گیری (Fusion) بین embedding متن و تصویر
- نرمال‌سازی بردارهای embedding برای افزایش دقت جستجو

مدل و فرایند آموزش

مدل انتخاب‌شده برای استخراج بازنمایی‌ها، M-CLIP بود. برای داده‌های متنی از نسخه‌ی چندزبانه‌ی CLIPM-CLIP/XLM-Roberta-Large-Vit-B-32 استفاده شد و برای داده‌های تصویری مدل CLIP ViT-B/32 به کار رفت. در هر مرحله، متن و تصویر به‌طور جداگانه به مدل داده شدند و embedding آن‌ها استخراج گردید. سپس، با روش Fusion (میانگین‌گیری و نرمال‌سازی)، بازنمایی چندوجهی واحد ساخته شد.

ما در نهایت موفق به fine-tune کردن مدل نشدیم، چرا که زیرساخت موردنیاز برای این کار فراهم نبود. حتی با استفاده از روش‌هایی مانند LoRA نیز fine-tune کردن این مدل‌های بزرگ میسر نبود.

ساخت ایندکس برای بازیابی

پس از استخراج embeddingها، تمامی بردارها در یک FAISS Index ذخیره شدند. این ساختار جستجوی سریع و بهینه در داده‌های چندوجهی را ممکن می‌سازد. علاوه بر این، یک فایل نگاشت ایجاد شد تا هر بردار به داده‌ی اصلی (متن و تصویر) مرتبط شود.

پیاده‌سازی بازیابی چندوجهی

برای تولید پاسخ نهایی مبتنی بر داده‌های بازیابی‌شده، از مدل زبانی بزرگ چندوجهی LLaVA-v1.6-Mistral-7B استفاده گردید. این مدل توانایی پردازش هم‌زمان متن و تصویر را داشته و در این پروژه برای پاسخ‌گویی به پرسش‌های کاربر به کار رفت.

Pipeline نهایی به این صورت طراحی شد:

- در حالت کوئری صرفاً متنی، embedding متن محاسبه شده و در FAISS جستجو می‌شود.
- در حالت کوئری متنی–تصویری، embedding متن و تصویر استخراج و با روش Fusion ترکیب می‌گردد.
- نزدیک‌ترین نمونه‌ها از ایندکس FAISS بازیابی شده و همراه با پرسش کاربر به مدل زبانی ارسال می‌شوند.
- مدل LVM پاسخی جامع و یکپارچه همراه با توضیحات تکمیلی درباره‌ی نتایج بازیابی‌شده تولید می‌کند.

ارزیابی

در این بخش، عملکرد سه مدل مختلف در سناریوهای گوناگون ارزیابی گردید. مجموعه‌ی آزمون شامل موارد زیر است:

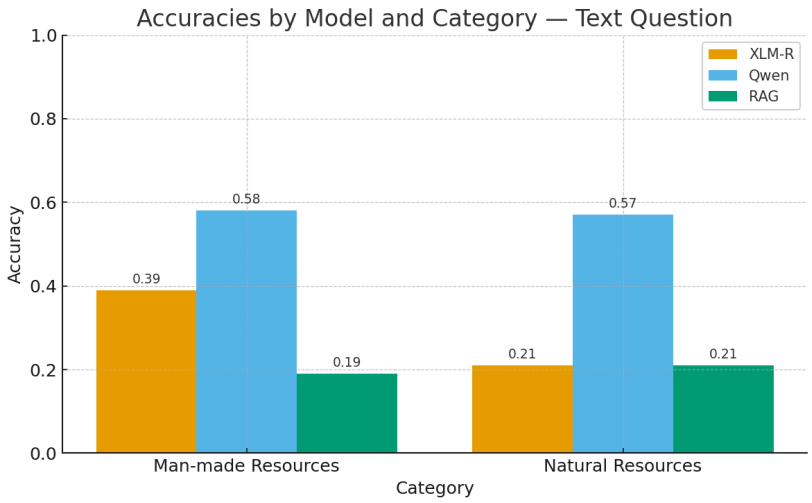
- ۵۰ سؤال چندگزینه‌ای متنی به زبان فارسی
- ۳۰ سؤال چندگزینه‌ای چندوجهی (متنی–تصویری)
- و یک سناریوی جایگزین که در آن به‌جای تصویر، یک راهنمای متنی ارائه شد.

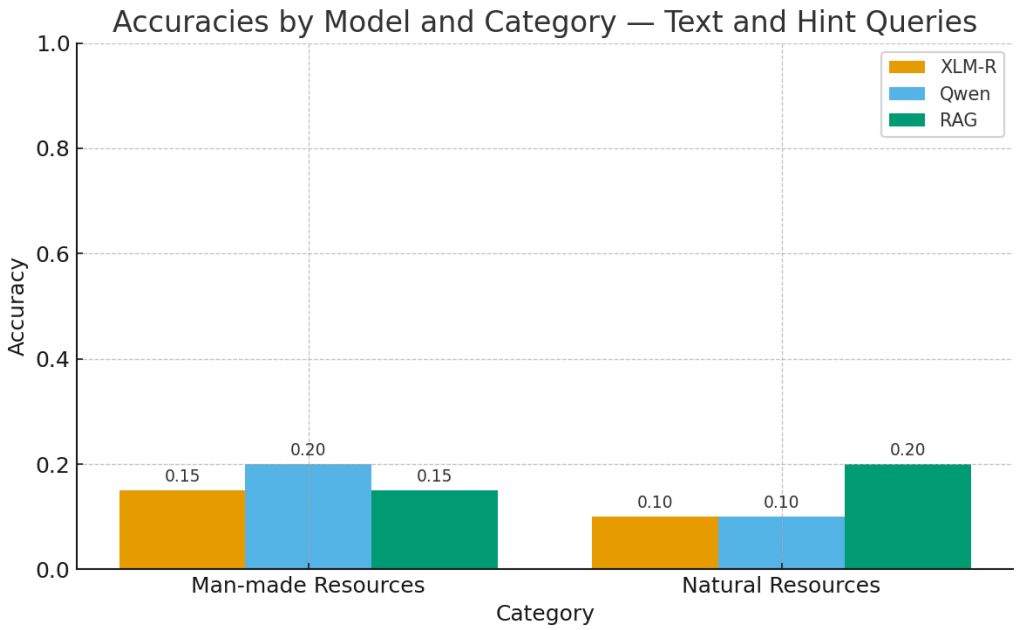
به‌منظور تحلیل دقیق‌تر، سؤالات در دو حوزه‌ی موضوعی جاذبه‌های طبیعی و جاذبه‌های ساخت بشر دسته‌بندی شدند تا امکان گزارش دقت حوزه‌ای فراهم گردد.
مدل‌های مورد ارزیابی عبارت بودند از:

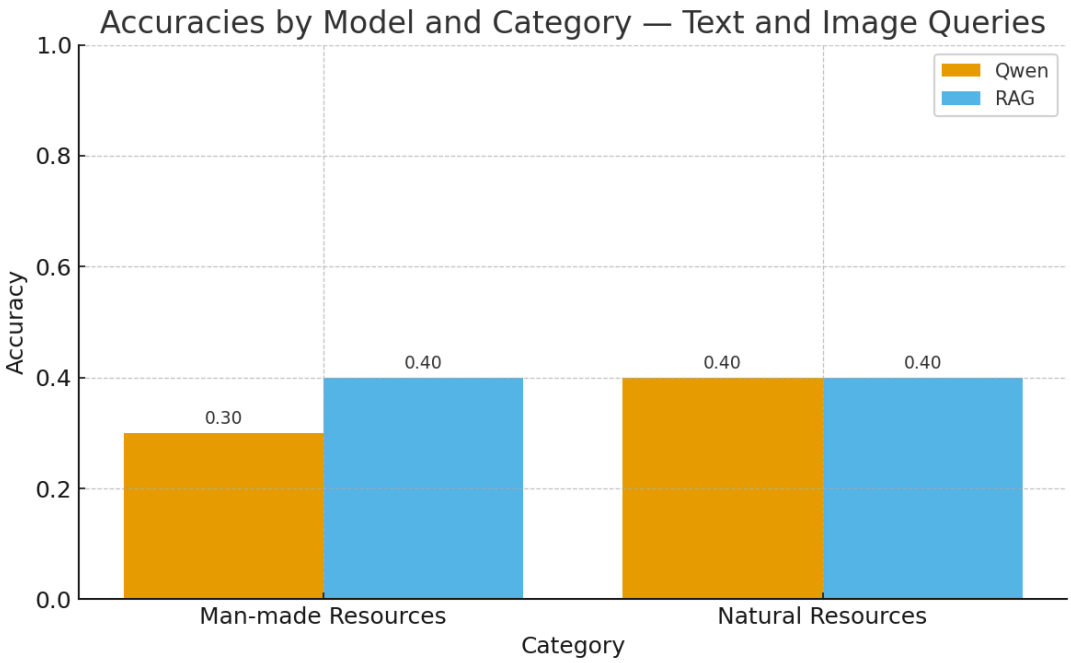
- XLM-R
- Qwen
- RAG

نتایج ارزیابی حوزه‌ای

میزان دقت مدل‌ها به تفکیک حوزه‌ی موضوعی در سه سناریوی «پرسش متنی»، «پرسش متنی همراه با راهنما» و «پرسش متنی–تصویری» به‌صورت نمودارهای زیر ارائه شده است:



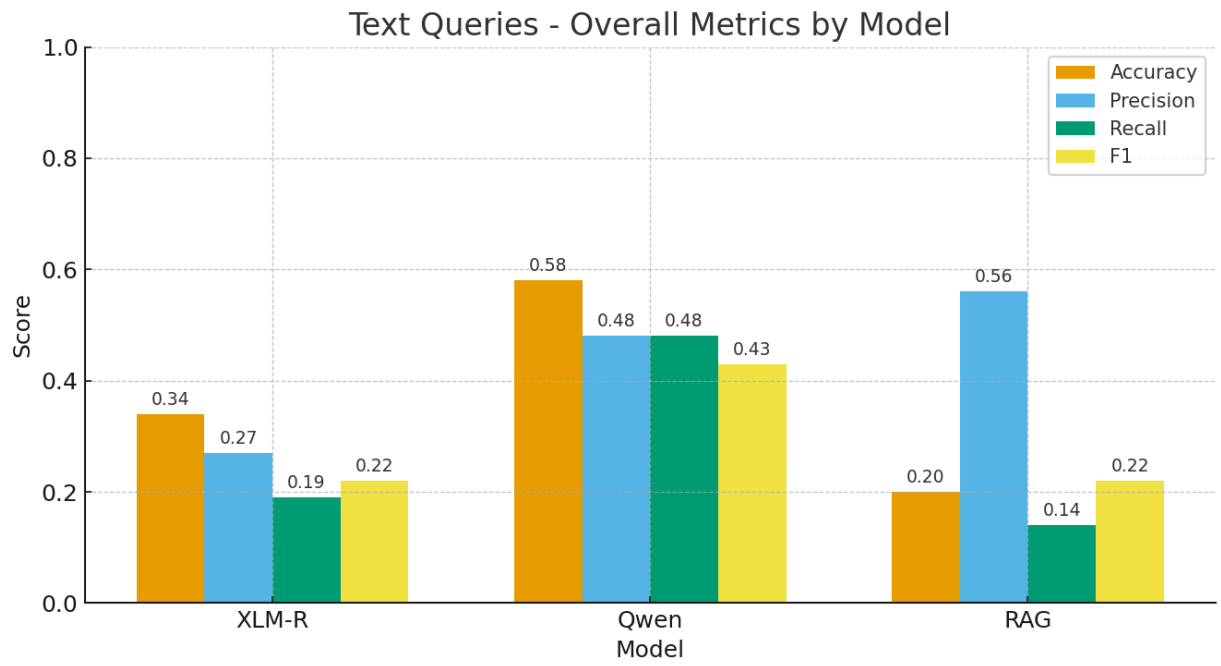


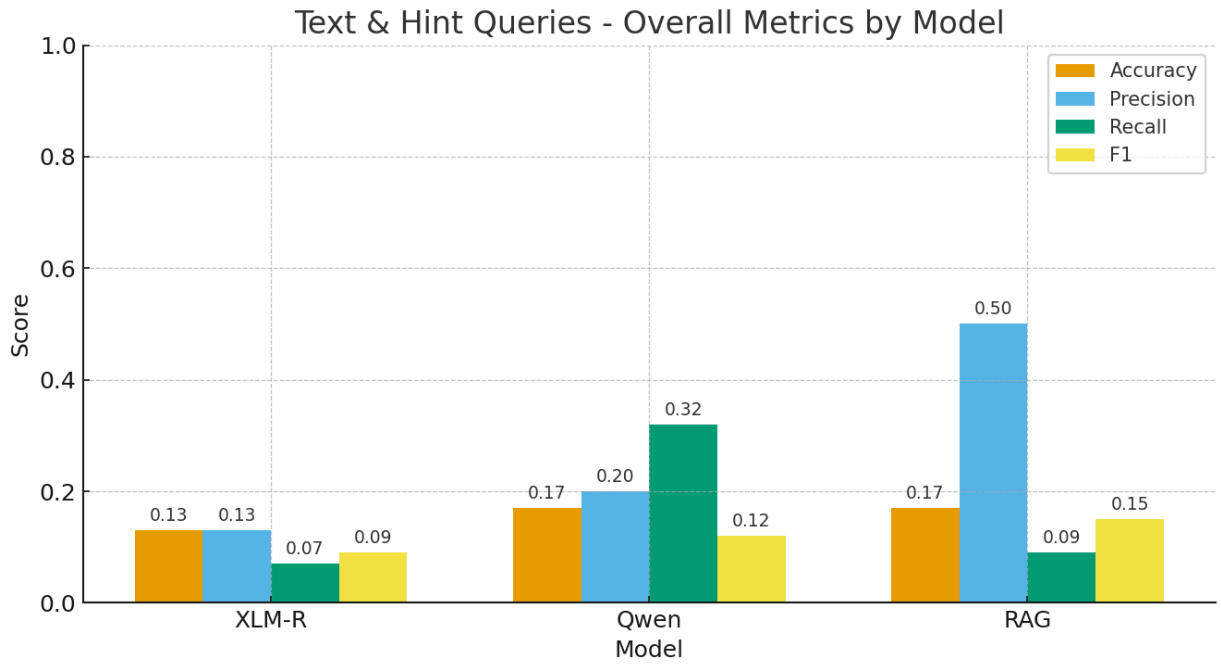


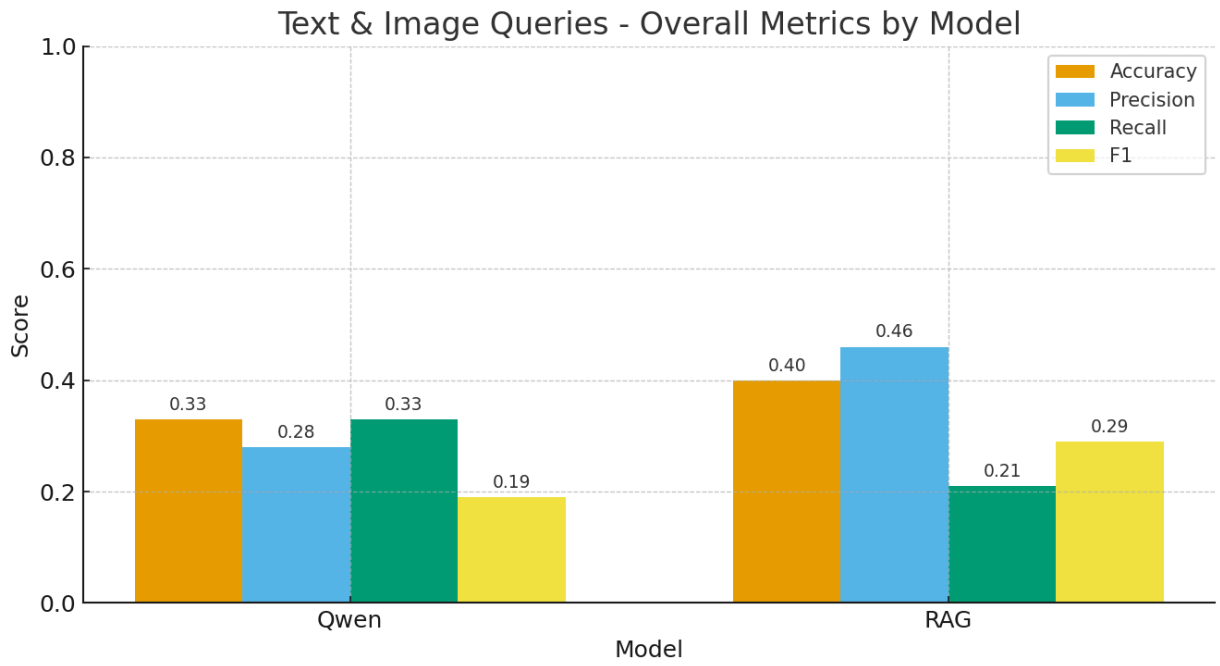
در پرسش‌های متنی، Qwen بهترین عملکرد را داشت، در حالی که XLM-R دقت متوسطی نشان داد و RAG برخلاف انتظار ضعیف‌تر بود. در سناریوی همراه با راهنما، دقت همه‌ی مدل‌ها کاهش یافت و این نوع سرنخ کمکی به بهبود پاسخ‌گویی نکرد. در پرسش‌های متنی–تصویری، RAG برترین مدل بود و توانست از داده‌های چندوجهی بهتر استفاده کند. به‌طور کلی، Qwen در متون و RAG در داده‌های چندوجهی برتری دارند و XLM-R در همه‌ی سناریوها ضعیف‌تر عمل کرده است. دقت مدل‌ها در هر دو حوزه و در هر سه سناریو تفاوت چشمگیری نداشت و روند عملکرد آن‌ها تقریباً مشابه بود.

نتایج کلی

علاوه بر دقت حوزه‌ای، شاخص‌های کلی شامل Recall، Precision، Accuracy و F1-score برای هر مدل در هر سناریو محاسبه و در نمودارهای زیر ارائه شده است:



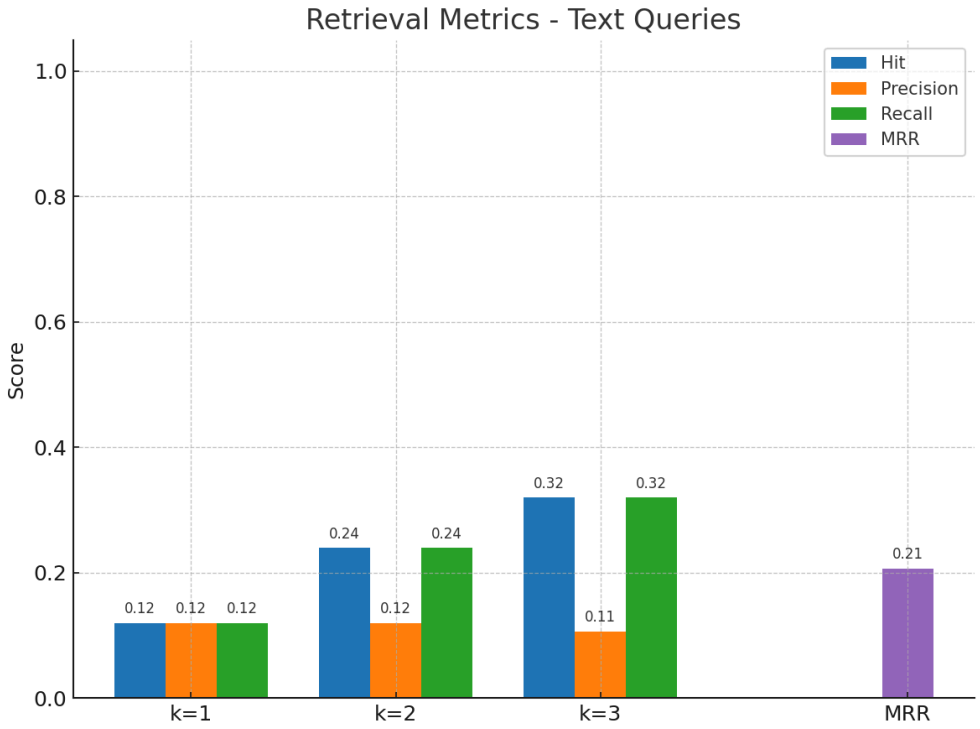


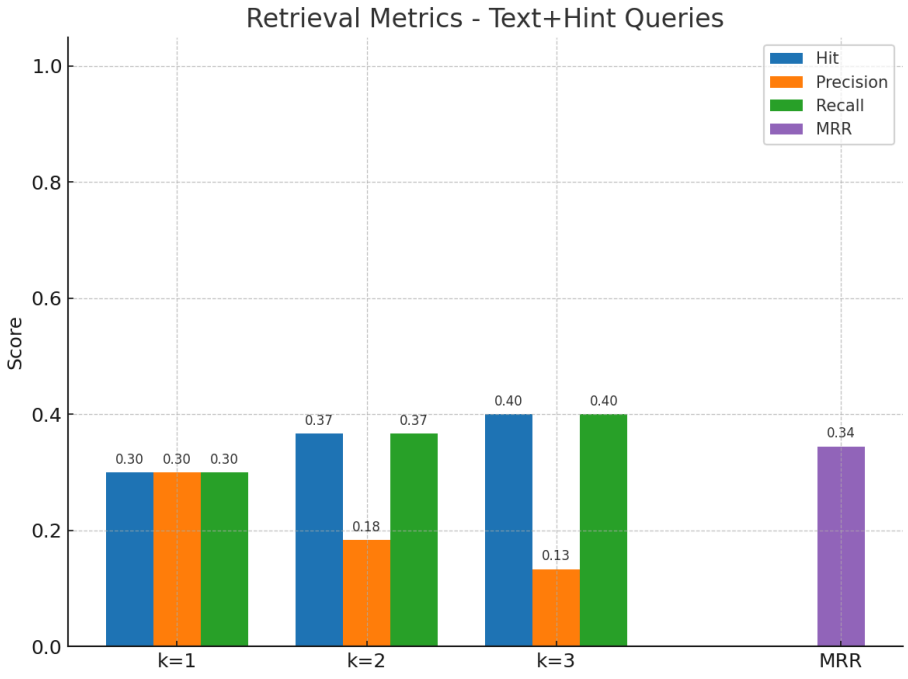


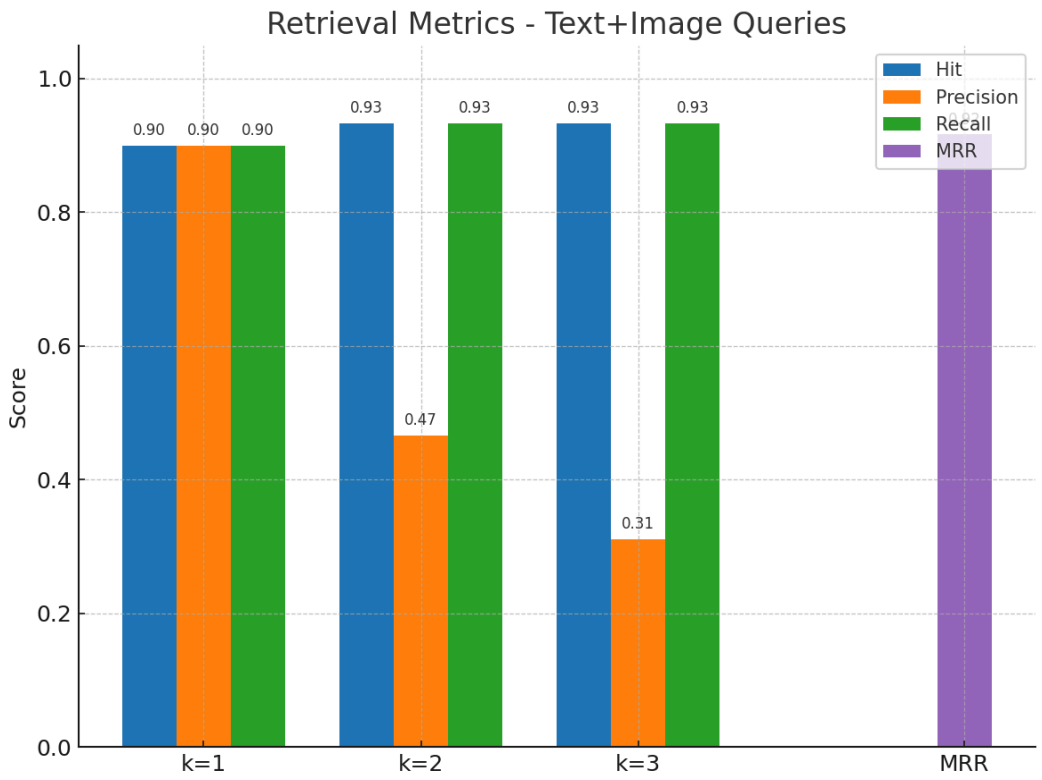
نتایج کلی نشان داد که در پرسش‌های متنی، Qwen بهترین عملکرد را با تعادل مناسب میان Accuracy، Precision و Recall به دست آورد، در حالی که XLM-R عملکرد متوسطی داشت و RAG به دلیل Recall پایین ضعیف‌تر عمل کرد. در سناریوی همراه با راهنما، دقت همه‌ی مدل‌ها افت کرد و این نوع سرنخ کارآمد نبود. در پرسش‌های متنی–تصویری، RAG توانست با Accuracy و Precision بالاتر، برتری خود را در استفاده از داده‌های چندوجهی نشان دهد، در حالی که Qwen عملکرد ضعیف‌تری داشت. به‌طور کلی، Qwen در متون و RAG در داده‌های چندوجهی بهترین نتایج را ارائه دادند و XLM-R در همه‌ی سناریوها ضعیف‌تر بود.

بررسی کیفیت بازیابی

به‌منظور ارزیابی کیفیت بازیابی مستقل از عملکرد مدل مولد، چهار معیار Hit@k، Precision@k، Recall@k و MRR محاسبه شدند. نمودارهای زیر مقادیر به‌دست‌آمده برای سناریو های مختلف را نشان می‌دهند:







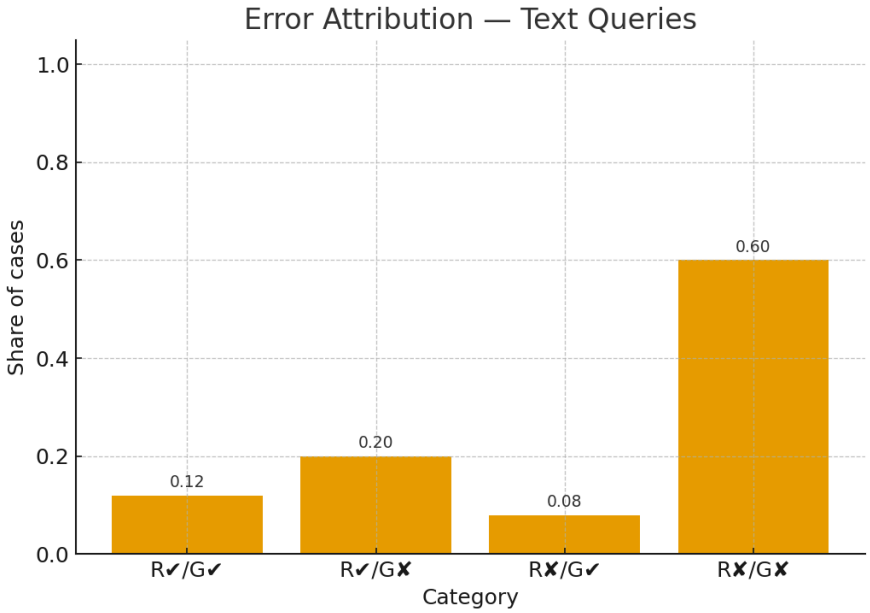
نتایج نشان داد که در سناریوی متنی، بازیاب عملکرد ضعیفی داشت و تنها در حدود یک‌سوم پرسش‌ها پاسخ درست را در سه نتیجه‌ی اول یافت.. در سناریوی متنی همراه با راهنما، Recall کمی بهبود یافت (0.40) اما Precision تغییر محسوسی نداشت. در مقابل، در سناریوی متنی–تصویری بازیاب بهترین عملکرد را ارائه داد؛ Recall@3 به حدود 0.93 و Precision@3 به 0.31 رسید که نشان‌دهنده‌ی توان بالاتر آن در بهره‌گیری از داده‌های چندوجهی است.

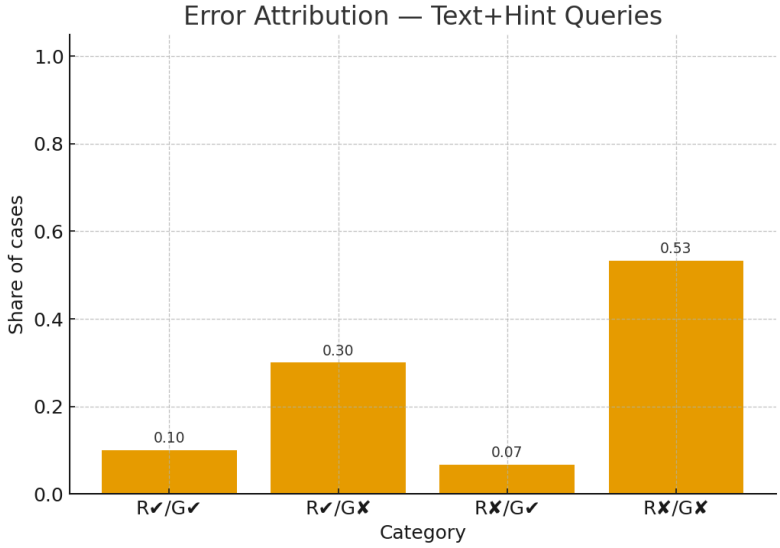
ردیابی خطا

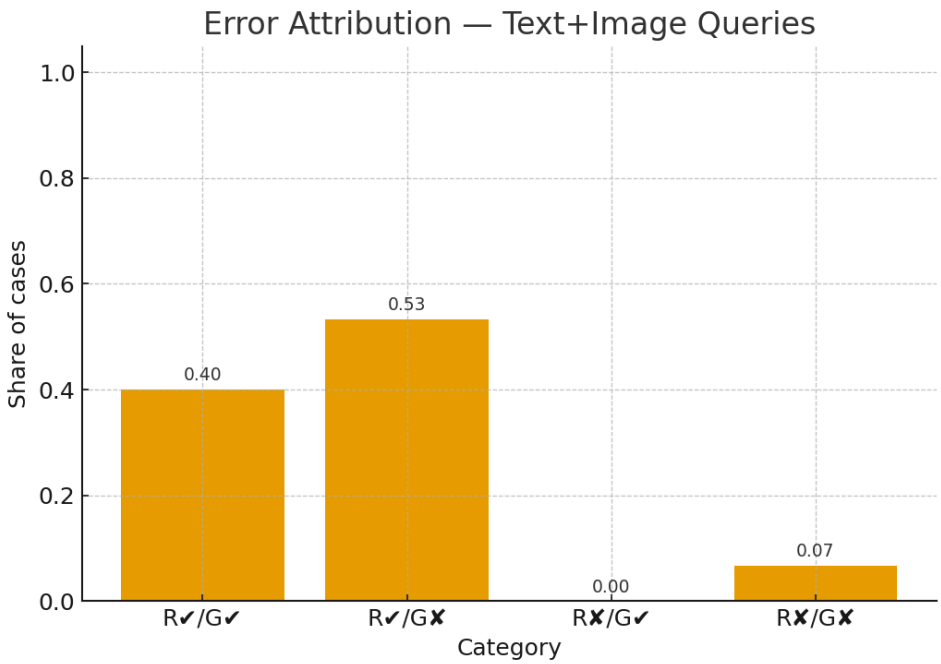
برای شناسایی منبع خطا در سامانه، خروجی‌ها به چهار حالت تقسیم شدند:

- بازیابی صحیح و تولید صحیح
- بازیابی صحیح اما تولید غلط
- بازیابی غلط اما تولید صحیح
- بازیابی غلط و تولید غلط

نمودارهای زیر سهم هر دسته را در سه سناریوی پرسش متنی، متنی–راهنما و متنی–تصویری نشان می‌دهند.







به طور کلی، نتایج نشان داد که ضعف اصلی سامانه در سناریوهای متنی و متنی–راهنما بیشتر به کیفیت بازیابی برمی‌گردد، در حالی که در سناریوی متنی–تصویری عملکرد به شکل قابل توجهی بهبود یافته است. با این حال، همچنان بخشی از خطاها ناشی از مولد است که حتی در حضور شواهد درست نیز پاسخ نادرست تولید می‌کند. بنابراین، برای ارتقای سیستم لازم است هم بهبود ماژول بازیاب و هم تقویت بخش مولد مدنظر قرار گیرد.

تحلیل نتایج

در این بخش، به تحلیل نتایج به‌دست‌آمده پرداخته می‌شود تا تفاوت عملکرد مدل‌ها در سناریوهای مختلف، کیفیت بازیابی، نقش promptها و چالش‌های زبانی و چندرسانه‌ای به‌طور دقیق‌تر بررسی گردد.

- آیا خروجی ها در حالت تصویری، متنی یا ترکیبی تفاوت معناداری دارند؟ بله، خروجی‌ها در سه سناریو تفاوت معناداری داشتند. در حالت متنی، دقت پایین‌تر بود و مدل‌ها بیشتر دچار خطا در بازیابی یا تولید شدند. در حالت متنی–راهنما، عملکرد حتی افت بیشتری داشت و نشان داد که سرنخ‌های متنی برای مدل‌ها گیج‌کننده بوده‌اند. در مقابل، در حالت متنی–تصویری، نتایج به‌طور قابل توجهی بهتر بود و مدل RAG توانست با استفاده مؤثر از تصاویر، پاسخ‌های دقیق‌تری ارائه دهد.

- آیا بازیابی موفق و مرتبط انجام شده؟ بازیابی در همه‌ی سناریوها به یک اندازه موفق نبود. در حالت متنی، میزان موفقیت پایین بود و بسیاری از پرسش‌ها اسناد مرتبطی در نتایج اولیه نداشتند. در حالت متنی–راهنما نیز وضعیت مشابه بود و خطاهای بازیابی سهم بالایی داشتند. در مقابل، در حالت متنی–تصویری، بازیاب عملکرد بسیار بهتری داشت و در اغلب موارد توانست داده‌های مرتبط و صحیح را در صدر نتایج قرار دهد.

- نقش prompt ها یا تنظیمات مدل چه بوده است؟ Promptها و تنظیمات مدل نقش مهمی در کیفیت خروجی داشتند. در سناریوی متنی، استفاده از دستورالعمل‌های ساده و مستقیم باعث شد مدل عملکرد بهتری نشان دهد. اما در سناریوی راهنما، به دلیل ابهام و پیچیدگی در promptها، مدل‌ها نتوانستند ارتباط درستی میان داده‌های بازیابی‌شده و پرسش برقرار کنند. در حالت متنی–تصویری، طراحی prompt چندوجهی به مدل کمک کرد تا از ترکیب متن و تصویر بهره‌ی مؤثرتری ببرد.

- آیا خروجی تولیدشده مبتنی بر اسناد بازیابی شده بوده یا صرفاً حدس مدل بوده است؟ در بسیاری از موارد، خروجی تولیدشده مستقیماً بر اساس اسناد بازیابی‌شده نبود و مدل صرفاً بر پایه‌ی دانش عمومی یا حدس خود پاسخ داده است. این موضوع به‌ویژه در سناریوهای متنی و متنی–راهنما مشاهده شد، جایی که بازیابی ضعیف یا نامرتبط بود. در مقابل، در حالت متنی–تصویری، هم‌پوشانی بیشتری میان محتوای بازیابی‌شده و پاسخ نهایی دیده شد که نشان می‌دهد مدل در این سناریو وابستگی بیشتری به شواهد داشت.

- چه چالش هایی در زمینه زبان، ساختار سوال یا محتوای چندرسانه‌ای مشاهده شده است؟ چند چالش اساسی مشاهده شد. در بخش زبان فارسی، وجود خطاهای نگارشی و تنوع در سبک نوشتار باعث کاهش کیفیت بازنمایی‌ها شد. در ساختار سؤال‌ها، طولانی بودن یا ترکیب چند مفهوم در یک پرسش موجب سردرگمی مدل گردید. در محتوای چندرسانه‌ای نیز هم‌ترازی دقیق بین متن و تصویر اهمیت بالایی داشت و هرجا این هم‌ترازی ضعیف بود، خروجی نادرست تولید شد.

نتیجه گیری

در این تمرین عملکرد مدل‌ها در سه سناریوی متنی، متنی–راهنما و متنی–تصویری مورد ارزیابی قرار گرفت. نتایج نشان داد که در حالت متنی، مدل Qwen بهترین عملکرد را داشت اما همچنان با خطاهای قابل توجه در بازیابی و تولید مواجه بود. افزودن راهنمای متنی نه‌تنها کمکی نکرد، بلکه باعث افت دقت شد و بیانگر ضعف مدل‌ها در تفسیر سرنخ‌های متنی است. در مقابل، سناریوی متنی–تصویری بیشترین بهبود را به همراه داشت و مدل RAG توانست با استفاده مؤثر از داده‌های تصویری، پاسخ‌های دقیق‌تری ارائه کند.

بررسی کیفیت بازیابی نیز نشان داد که خطاها عمدتاً ناشی از ضعف در انتخاب اسناد مرتبط در سناریوهای متنی بوده و در بسیاری از موارد حتی با وجود بازیابی صحیح، مدل مولد از شواهد به درستی بهره‌برداری نکرده است. همچنین تحلیل خطا نشان داد که طراحی prompt و هم‌ترازی دقیق داده‌های چندوجهی نقش مهمی در موفقیت سامانه دارد.

به طور کلی، نتایج بیانگر آن است که استفاده از داده‌های چندرسانه‌ای می‌تواند کارایی سیستم‌های چندوجهی را به طور چشمگیری افزایش دهد، اما برای رسیدن به عملکرد پایدار لازم است بهبودهایی هم در بخش بازیاب و هم در بخش مولد صورت گیرد.