سیستم Topic Restriction و Moderation مبتنی بر Embeddingبا معماری Dual-Centroid

1.چارچوب مسئله و منطق طراحی

در ابتدای پروژه، هیچ مدل زبانی بزرگ (LLM) یا معماری عامل (Agent) مشخصی برای اعمال گاردریل تعریف نشده بود. با توجه به تأکید سرپرست پروژه بر موضوعاتی مانند تعدیل محتوا(Moderation) ، محدودسازی موضوعی (Topic Restriction)و زیرساخت‌های ایمنی، پروژه به‌گونه‌ای تعریف شد که به جای وابستگی به یک مدل یا کاربرد خاص، به‌عنوان یک مؤلفه عمومی و قابل‌استفاده مجدد در سیستم‌های مختلف هوش مصنوعی عمل کند.

در تعریف کلی، محتوای ناایمن (unsafe) می‌تواند شامل دو سطح متفاوت باشد: نخست، محتوای ذاتاً نامناسب مانند خشونت یا محتوای جنسی که معمولاً با مدل‌های عمومی و از پیش‌تعریف‌شده قابل شناسایی هستند. برای این دسته، مدل‌هایی مانند Llama Guard به دلیل اجرای محلی و تأخیر پایین، گزینه‌های مناسبی محسوب می‌شوند. در سطح دیگر، محتوایی قرار می‌گیرد که ذاتاً نامناسب نیست، اما با سیاست‌های مشخص سیستم یا سازمان در تضاد است. تمرکز این پروژه بر این نوع از محتوا بوده است، به‌ویژه در حوزه‌های سیاسی و مذهبی که نیازمند تعریف مرزهای تصمیم‌گیری سفارشی و حساس به context هستند.

در مرحله طراحی، چندین راهکار موجود برای تعدیل محتوا مورد بررسی قرار گرفتند. سرویس‌های تجاری تعدیل محتوا مانندOpenAI Moderation ، Azure Content Safetyو Google Cloud Content Moderation به‌عنوان گزینه‌های اصلی ارزیابی شدند. اگرچه این سرویس‌ها قابلیت‌های قوی برای سناریوهای عمومی ارائه می‌کنند، اما وابستگی به سرویس‌های خارجی، هزینه عملیاتی و افزایش تأخیر پاسخ‌گویی از محدودیت‌های اصلی آن‌ها محسوب می‌شود. علاوه بر این، در سناریوهای چندمرحله‌ای، مجموع latency ناشی از چند فراخوانی می‌تواند از محدودیت زمانی طراحی‌شده برای سیستم (کمتر از یک ثانیه) فراتر رود. همچنین میزان کنترل‌پذیری این سرویس‌ها برای سیاست‌های خاص، به‌ویژه در حوزه محتوای فارسی سیاسی و مذهبی، محدود ارزیابی شد.

مدل‌های محلی مانند Llama Guard نیز بررسی شدند. این مدل‌ها عملکرد مناسبی در تشخیص دسته‌های عمومی محتوای ناایمن، از جمله خشونت و محتوای جنسی، دارند. با این حال، آن‌ها عمدتاً بر اساس مجموعه‌ای از سیاست‌های از پیش تعریف‌شده طراحی شده‌اند و انعطاف‌پذیری محدودی برای پیاده‌سازی سیاست‌های سفارشی و محدودسازی موضوعی ارائه می‌دهند.

چارچوب NVIDIA NeMo Guardrails نیز در ادامه مورد بررسی قرار گرفت. اگرچه NeMo Guardrails ابزار مناسبی برای تعریف قوانین مکالمه و کنترل رفتار سیستم‌های مبتنی بر LLM فراهم می‌کند، اما تمرکز اصلی آن بر ارکستراسیون و مدیریت جریان تعامل در سیستم‌های عامل‌محور است، نه تشخیص مستقل و کم‌هزینه محتوای ناایمن در سطح متن. از آنجا که در این پروژه مدل یا عامل مشخصی به‌عنوان هسته سیستم وجود نداشت و هدف

اصلی، طراحی یک مؤلفه سبک برای Topic Restriction و Content Moderation با تأخیر پایین بود، استفاده از NeMo Guardrailsپیچیدگی اضافی نسبت به نیاز مسئله ایجاد می‌کرد. همچنین اعمال سیاست‌های سفارشی همچنان به یک ماژول طبقه‌بندی یا تشخیص در لایه زیرین وابسته است.

با توجه به این محدودیت‌ها، یک معماری مبتنی بر Embedding برای سیستم تعدیل محتوا انتخاب شد. این رویکرد امکان استقرار محلی، تأخیر پایین، کنترل کامل بر مرزهای تصمیم‌گیری، و پشتیبانی از زبان‌های فارسی و انگلیسی را بدون نیاز به Fine-Tuning گسترده فراهم می‌کند.

2. معماری سیستم

این سیستم یک معماری سبک مبتنی بر embedding-space geometry برای طبقه‌بندی متون به دو کلاس safe و unsafeاست. تصمیم‌گیری در این چارچوب بدون نیاز به fine-tuning و صرفاً بر اساس شباهت برداری به نمایش‌های پروتوتایپی کلاس‌ها انجام می‌شود.

a. ساخت دیتاست ها

در این پروژه، داده‌ها از چند منبع گردآوری و به سه مجموعه اصلی تقسیم شدند. ابتدا یک مجموعه داده safe شامل نمونه‌های عمومی مانند متونfactual ، پرسش‌های عمومی و محتوای غیرمخاطره‌آمیز مورد استفاده قرار گرفت. در کنار آن، مجموعه‌ای مبتنی بر policy برای نمونه‌های unsafe شامل محتوای نقض‌کننده سیاست‌ها مانند محتوای خشونت‌آمیز، حساس سیاسی یا غیرمجاز ساخته شد.

برای بهبود دقت مرز تصمیم، مجموعه سومی تحت عنوان hard_mixed شامل نمونه‌هایی نزدیک به مرز سیاست و دارای ابهام معنایی ایجاد شد. در نهایت، این مجموعه با داده‌های safe ترکیب شد تا توزیع کلاس safe به جای یک ناحیه فشرده، به یک ناحیه گسترده و مرزی‌پوش تبدیل شود و از حساسیت بیش از حد مدل به کلیدواژه‌های خاص جلوگیری گردد.

b. ساختcentroid ها و نمایش هندسی

هر متن ابتدا توسط یک encoder به فضای embedding نگاشت می‌شود. سپس برای هر کلاس به جای استفاده از یک centroidواحد، چندین centroid برای حفظ تنوع درون‌کلاسی ساخته می‌شود. اینcentroid ها از طریق تقسیم داده به چند بخش (chunking) و محاسبه میانگین بردارها در هر بخش تولید شده و در ادامه نرمال‌سازی می‌شوند تا تنها جهت بردار در فضای embedding حفظ شود.

این طراحی موجب می‌شود هر کلاس به جای یک نقطه منفرد، به مجموعه‌ای از نماینده‌های محلی در فضای embeddingتبدیل شود که نواحی مختلف معنایی کلاس را پوشش می‌دهند.

c. تجمیع شباهت و pipeline تصمیم‌گیری

در مرحله inference ، ورودی به embedding تبدیل شده و شباهت آن با تمام centroid های هر کلاس بر اساس cosine similarity محاسبه می‌شود. به جای استفاده از بیشینه ساده (max pooling) ، از تابع Log-Sum-Exp (LSE) برای تجمیع شباهت‌ها استفاده شده است.

این انتخاب باعث می‌شود تصمیم نهایی صرفاً بر اساس نزدیک‌ترین centroid نباشد، بلکه چند centroid نزدیک به‌صورت هم‌زمان در تصمیم‌گیری اثرگذار باشند، در حالی که centroid های دورتر وزن کمتری دارند. در نتیجه، رفتار مدل نسبت به نویز و نقاط مرزی پایدارتر می‌شود.

در نهایت، دو مقدار شباهت برای safe و unsafe محاسبه می‌شود و اختلاف آن‌ها به‌صورت یک مقدار margin تعریف می‌شود:

- $margin > 0$ نشان‌دهنده گرایش به unsafe
- $margin \leq 0$ نشان‌دهنده گرایش به safe

در انتها، با اعمال یک threshold روی این مقدار، خروجی نهایی سیستم به صورت binary classification تعیین می‌شود.

3. ارزیابی و یافته‌های آزمایشی

a. ارزیابی اولیه (Sanity Check)

در مرحله ابتدایی، سیستم بر روی تقسیم داخلی داده‌های Safe و Unsafe مورد ارزیابی قرار گرفت تا صحت عملکرد پایه‌ای مدل و پایداری مرز تصمیم بررسی شود. این مرحله صرفاً برای اطمینان از رفتار صحیح scoring function و صحت جداسازی اولیه بین دو کلاس انجام شد. معیار تصمیم در این مرحله، اختلاف شباهت بین کلاس‌ها (margin) و threshold روی آن بوده است.

b. ارزیابی دنیای واقعی (Out-of-Distribution Evaluation)

برای بررسی تعمیم‌پذیری سیستم، از دیتاست Reddit Politics به‌عنوان داده خارج از توزیع (OOD) استفاده شد. این مجموعه شامل متن‌های کوتاه، مبهم و دارای بار سیاسی است و به دلیل وجود noise ، sarcasm و ambiguity برای ارزیابی robustness مناسب است. همچنین نمونه‌های ترکیبی شامل محتوای factual همراه با موضوعات سیاسی برای بررسی رفتار مرز تصمیم در شرایط نزدیک به boundary مورد استفاده قرار گرفتند. نتایج نشان داد که اگرچه عملکرد نسبت به داده داخلی کاهش می‌یابد، اما رفتار مدل پایدار باقی می‌ماند و جداسازی نسبی بین کلاس‌ها حفظ می‌شود

c. نتایج عملکرد

در مجموعه داده داخلی، سیستم عملکرد زیر را نشان داد:

Precision: 0.97

Recall: 0.90

F1 Score: 0.94

در ارزیابی(OOD (Reddit Politics :

Accuracy: 0.83

AUC: 0.74

همچنین تحلیل امتیازها نشان داد:

میانگین risk در کلاس0.48 :safe

میانگین risk در کلاس0.66 :unsafe

این نتایج نشان‌دهنده separation قابل قبول بین دو کلاس هستند، با وجود overlap در نواحی مرزی.

d. کارایی سیستم(Latency)

سیستم از نظر کارایی دارای تأخیر پردازشی پایین و مناسب برای کاربرد real-time است. میانگین latency در سطح APIبین 75 تا 82 میلی‌ثانیه برای هر درخواست گزارش شده است که شامل مراحلembedding generation ، محاسبه similarityو تصمیم‌گیری نهایی می‌شود

.

4. مسیرهای توسعه آینده

در ادامه مسیر توسعه، چند جهت مشخص برای بهبود عملکرد سیستم قابل تعریف است. اولین محور، بهبود کیفیت جداسازی در فضای embedding است که می‌تواند از طریق lightweight fine-tuning یا استفاده از روش‌های contrastive learningبر روی داده‌های مرزی انجام شود. هدف از این بهبود، کاهش overlap میان کلاس‌های safe و unsafeو افزایش تفکیک‌پذیری در نواحی مرزی تصمیم است.

در سطح تجمیع(aggregation) ، استفاده از روش‌های پیشرفته‌تر مانند attention pooling یا learned weighting به جای LSE می‌تواند باعث بهبود تفکیک میانcentroid های informative و noisy شود و در نتیجه پایداری تصمیم‌گیری در نواحی ambiguous را افزایش دهد.

از منظر معماری، اضافه کردن یک لایه بازبینی دومرحله‌ای (two-stage classification) یا ترکیب سیستم با یک مدل sequence-levelمانند LLM-based guard می‌تواند در کاهش false positive در نمونه‌های مبهم مؤثر باشد. همچنین انجام calibration دقیق‌تر روی threshold می‌تواند کنترل بهتری بر trade-off میان precision و recall فراهم کند، به‌ویژه در شرایط خارج از توزیع.(OOD)

از منظر روش‌های مدل‌سازی، رویکردهای سبک مبتنی بر few-shot learning مانند SetFit نیز می‌توانند در این چارچوب مورد بررسی قرار گیرند، خصوصاً در شرایطی که داده‌های برچسب‌خورده محدود بوده و نیاز به راهکارهای کم‌هزینه و قابل استقرار وجود دارد.

در نهایت، این پروژه تحت محدودیت‌های زمانی و محاسباتی طراحی و پیاده‌سازی شده و از روش‌هایی مانند training سنگین، fine-tuning گسترده و مدل‌های بزرگ چند میلیارد پارامتری صرف‌نظر شده است. همچنین برخی رویکردهای جایگزین که در مراحل اولیه طراحی بررسی شدند، از جمله supervised transformer classifiers، full contrastive fine-tuning و cross-encoderهای سنگین، به دلیل محدودیت latency و منابع محاسباتی کنار گذاشته شدند، هرچند همچنان می‌توانند در نسخه‌های آینده از نظر دقت عملکردی مورد بررسی قرار گیرند.

6. جمع‌بندی ارزیابی

در مجموع، رویکرد مبتنی بر embedding و dual-centroid بدون نیاز به fine-tuning قادر است عملکرد پایدار و قابل قبولی در سناریوهای داخلی و خارج از توزیع ارائه دهد. محدودیت اصلی مشاهده‌شده مربوط به کاهش دقت در داده‌های بسیار ambiguous و خارج از توزیع است که ناشی از overlap طبیعی در فضای embedding است.