دانشگاه صنعتی امیرکبیر
(پلی‌تکنیک تهران)
دانشکدهٔ مهندسی کامپیوتر

پایان‌نامه کارشناسی

مهندسی کامپیوتر

توسعهٔ سامانه‌ای برای تحلیل احساسات ضمنی با استفاده از زنجیرهٔ تفکر چندمرحله‌ای

نگارش

فاطمه درج

استاد راهنما

دکتر مصطفی حقیر چهرقانی

تیر ۱۴۰۵

بسم الله الرحمن الرحیم

صفحه فرم ارزیابی و تصویب پایان نامه- فرم تأیید اعضاء کمیته دفاع

در این صفحه فرم دفاع یا تایید و تصویب پایان نامه موسوم به فرم کمیته دفاع- موجود در پرونده آموزشی- را قرار دهید.

نکات مهم:

- نگارش پایان نامه/رساله باید به زبان فارسی و بر اساس آخرین نسخه دستورالعمل و راهنمای تدوین پایان نامه های دانشگاه صنعتی امیرکبیر باشد.(دستورالعمل و راهنمای حاضر)
- رنگ جلد پایان نامه/رساله چاپی کارشناسی، کارشناسی ارشد و دکترا باید به ترتیب مشکی، طوسی و سفید رنگ باشد.
- چاپ و صحافی پایان نامه/رساله بصورت پشت و رو(دورو) بلامانع است و انجام آن توصیه می شود.

دانشگاه صنعتی امیرکبیر
(پلی‌تکنیک تهران)

به نام خدا

تعهدنامه اصالت اثر

تاریخ: تیر ۱۴۰۵

اینجانب **فاطمه درج** متعهد می‌شوم که مطالب مندرج در این پایان‌نامه حاصل کار پژوهشی اینجانب تحت نظارت و راهنمایی اساتید دانشگاه صنعتی امیرکبیر بوده و به دستاوردهای دیگران که در این پژوهش از آنها استفاده شده است مطابق مقررات و روال متعارف ارجاع و در فهرست منابع و مآخذ ذکر گردیده است. این پایان‌نامه قبلاً برای احراز هیچ مدرک هم‌سطح یا بالاتر ارائه نگردیده است.

در صورت اثبات تخلف در هر زمان، مدرک تحصیلی صادر شده توسط دانشگاه از درجه اعتبار ساقط بوده و دانشگاه حق پیگیری قانونی خواهد داشت.

کلیه نتایج و حقوق حاصل از این پایان‌نامه متعلق به دانشگاه صنعتی امیرکبیر می‌باشد. هرگونه استفاده از نتایج علمی و عملی، واگذاری اطلاعات به دیگران یا چاپ و تکثیر، نسخه‌برداری، ترجمه و اقتباس از این پایان نامه بدون موافقت کتبی دانشگاه صنعتی امیرکبیر ممنوع است. نقل مطالب با ذکر مآخذ بلامانع است.

فاطمه درج

امضا

تقدیم به پدر و مادرم،

که با مهر، شکیبایی و حمایت همیشگی‌شان،

مسیر آموختن را برایم هموار کردند.

سپاس‌گزاری

از استاد راهنمای ارجمندم، جناب آقای دکتر مصطفی حقیر چهرقانی، به‌سبب راهنمایی‌های علمی، دقت نظر و حمایت ایشان در مسیر انجام این پژوهش صمیمانه سپاس‌گزاری می‌کنم. همچنین از خانواده‌ام که در دوران تحصیل با شکیبایی و پشتیبانی همیشگی همراه من بودند، قدردانی می‌کنم.

فاطمه درج

تیر ۱۴۰۵

فهرست مطالب

عنوان | صفحه

۱ مقدمه ۱

۱-۱ بیان مسئله و اهمیت پژوهش ۲

۱-۲ اهداف پژوهش ۳

۱-۳ پرسش‌های پژوهش ۳

۱-۴ نوآوری‌ها و دستاوردها ۴

۱-۵ روش کلی پژوهش ۴

۱-۶ ساختار پایان‌نامه ۵

۲ تعاریف اولیه ۶

۲-۱ تحلیل احساسات، جنبه و قطبیت ۷

۲-۲ تحلیل احساسات ضمنی و چالش‌های آن ۷

۲-۳ مدل‌های زبانی و مهندسی پرامپت ۸

۲-۴ استدلال چندمرحله‌ای، بازبینی و خودسازگاری ۸

۲-۵ کنترل‌گر و انتخاب چندمنبعی ۹

۲-۶ معیارهای ارزیابی ۹

۳ ادبیات تحقیق ۱۱

۳-۱ تحلیل احساسات ضمنی و مجموعه‌داده‌ها ۱۲

۳-۲ استدلال و بازبینی در روش‌های مولد ۱۲

۳-۳ روش‌های THOR، SAoT و خودسازگاری ۱۳

۳-۴ انتخاب منبع و شکاف تحقیقاتی ۱۳

۴ کار پیشنهادی ۱۵

۴-۱ صورت‌بندی مسئله و نمای کلی سامانه ۱۶

۴-۲ تولید منابع پیش‌بینی ۱۶

۴-۳ بازبینی آگاه از نوع خطا و کنترل‌گر میانی ۱۸

۴-۴ انتخاب منبع محافظت‌شده و تنظیم‌شده با دادهٔ آموزش ۱۹

۴-۵ انتخاب‌گرهای جایگزین ۲۰
۴-۶ پیاده‌سازی و جمع‌بندی ۲۰

۵ نتایج آزمایش‌ها ۲۲
۵-۱ مقدمه ۲۳
۵-۲ تنظیمات آزمایش ۲۳
۵-۲-۱ مجموعه‌داده و تقسیم ارزیابی ۲۳
۵-۲-۲ مدل‌ها و زیرساخت اجرا ۲۴
۵-۲-۳ پروتکل ارزیابی و جلوگیری از نشت داده ۲۴
۵-۳ روش‌های مورد مقایسه و معیارهای ارزیابی ۲۵
۵-۳-۱ روش‌های مورد مقایسه ۲۵
۵-۳-۲ معیارهای ارزیابی ۲۶
۵-۴ نتایج روش‌های اصلی روی آزمون کامل ۲۶
۵-۴-۱ مقایسهٔ کلی روش‌ها ۲۶
۵-۴-۲ تحلیل ماتریس‌های درهم‌ریختگی ۲۸
۵-۴-۳ تحلیل عملکرد بر اساس دامنه ۲۹
۵-۵ تحلیل انتخاب‌گر و مؤلفه‌های سامانه ۳۰
۵-۵-۱ توزیع منابع و سود یا زیان نسبت به مسیر مستقیم ۳۰
۵-۵-۲ مقایسهٔ تصمیم میانی، انتخاب‌گرهای فراسطح و کران اوراکل ۳۱
۵-۵-۳ اعتبارسنجی زنجیرهٔ نهایی ۳۱
۵-۶ تحلیل کیفی و خطا ۳۲
۵-۶-۱ نمونه‌های اصلاح‌شده ۳۲
۵-۶-۲ نمونه‌های تضعیف‌شده ۳۲
۵-۶-۳ الگوهای خطای باقی‌مانده ۳۳
۵-۷ مقایسهٔ تکمیلی با Gemini 2.5 Flash ۳۴
۵-۷-۱ ساخت زیرمجموعهٔ مشترک ۳۴
۵-۷-۲ مقایسهٔ سنجه‌های دو مدل ۳۴
۵-۷-۳ ماتریس‌های درهم‌ریختگی Gemini ۳۵

۵-۸ ملاحظات اجرایی و محدودیت‌ها ۳۶
۵-۹ جمع‌بندی ۳۷

۶ نتیجه‌گیری و کارهای آتی ۳۸
۶-۱ جمع‌بندی پژوهش ۳۹
۶-۲ پاسخ به پرسش‌های پژوهش ۴۰
۶-۳ دستاوردهای اصلی ۴۱
۶-۳-۱ دستاوردهای علمی ۴۱
۶-۳-۲ دستاوردهای فنی و مهندسی ۴۲
۶-۴ محدودیت‌های پژوهش ۴۲
۶-۵ کارهای آتی ۴۳
۶-۵-۱ توسعه و بهبود انتخاب‌گر منبع ۴۳
۶-۵-۲ بهبود بازبینی و استدلال هدف‌محور ۴۳
۶-۵-۳ تحلیل خطاهای مشترک و گسترش منابع ۴۳
۶-۵-۴ گسترش ارزیابی به داده‌ها و مدل‌های دیگر ۴۴
۶-۵-۵ کاهش هزینهٔ اجرا ۴۴
۶-۵-۶ ارزیابی انسانی و ریزتنظیم ۴۴
۶-۶ نتیجه‌گیری نهایی ۴۴

منابع و مراجع ۴۶

پیوست: جزئیات بازتولید آزمایش‌ها ۴۸

واژه‌نامهٔ فارسی به انگلیسی ۵۱

واژه‌نامهٔ انگلیسی به فارسی ۵۲

فهرست اشکال

شکل صفحه

۴-۱ معماری سامانهٔ پیشنهادی؛ کنترل‌گر یک تصمیم میانی و انتخاب‌گر منبع تولیدکنندهٔ پاسخ نهایی است. ۱۷

۵-۱ مقایسهٔ دقت کلی و امتیاز اف‌یک ماکرو هشت روش اصلی روی ۴۴۲ نمونهٔ آزمون. . . ۲۸

۵-۲ ماتریس‌های درهم‌ریختگی پیش‌بینی مستقیم و سامانهٔ نهایی Qwen3 8B روی آزمون کامل. ۲۹

۵-۳ مقایسهٔ امتیاز اف‌یک ماکرو Qwen3 8B و Gemini 2.5 Flash روی زیرمجموعهٔ مشترک ۹۰ نمونه‌ای آزمون. ۳۵

۵-۴ ماتریس‌های درهم‌ریختگی پیش‌بینی مستقیم و سیاست انتخاب تنظیم‌شده برای Gemini 2.5 Flash روی آزمون مشترک. ۳۵

فهرست جداول

جدول صفحه

۳-۱ مقایسهٔ پژوهش‌های محوری مرتبط با تحلیل احساسات ضمنی. ۱۴

۴-۱ رده‌های بازبینی تشخیصی و نقش تصمیمی آن‌ها. ۱۹

۵-۱ توزیع نمونه‌های ضمنی در دو دامنه و دو بخش داده. ۲۴

۵-۲ نتایج روش‌های اصلی، از خط پایهٔ کلاسیک تا سامانهٔ نهایی، روی مجموعهٔ کامل آزمون. ۲۶

۵-۳ امتیاز اف‌یک کلاس‌ها برای پیش‌بینی مستقیم و سامانهٔ نهایی Qwen3 8B. ۲۹

۵-۴ عملکرد پیش‌بینی مستقیم و سامانهٔ نهایی به تفکیک دامنه در بخش آزمون. ۲۹

۵-۵ توزیع منبع منتخب در ۴۴۲ نمونهٔ آزمون. ۳۰

۵-۶ مقایسهٔ سیاست‌های انتخاب منبع روی ۴۴۲ نمونهٔ آزمون Qwen3 8B. ۳۱

فهرست نمادها

نماد	مفهوم
s	جملهٔ ورودی
t	هدف مشخص‌شده در جمله
a	جنبهٔ مرتبط با هدف
o	نظر ضمنی استنتاج‌شده
y	برچسب قطبیت طلایی
$\hat{y}$	برچسب قطبیت پیش‌بینی‌شده
P	صحت پیش‌بینی
R	بازخوانی
F_1	امتیاز اف‌یک
π	سیاست انتخاب منبع
$\mathcal{D}_{\text{train}}$	مجموعهٔ آموزش
$\mathcal{D}_{\text{test}}$	مجموعهٔ آزمون
$\mathcal{S}$	مجموعهٔ منابع پیش‌بینی نامزد

فصل اول

مقدمه

۱-۱ بیان مسئله و اهمیت پژوهش

تحلیل احساسات[۱] نگرش موجود در متن را معمولاً در سه کلاس مثبت، منفی و خنثی تعیین می‌کند. در احساس صریح، واژه‌ای مانند «عالی» جهت ارزیابی را آشکار می‌سازد؛ اما در احساس ضمنی[۲] قطبیت از پیامد، مقایسه یا انتظار استنباط می‌شود. برای نمونه، جملهٔ «باتری فقط دو ساعت دوام آورد» بدون صفت منفی، نارضایتی از باتری را منتقل می‌کند. تشخیص چنین نگرشی به بافت، دانش عمومی و رعایت دامنهٔ هدف نیاز دارد و با تطبیق سادهٔ واژگان حل نمی‌شود [۱، ۲].

مدل‌های زبانی بزرگ می‌توانند پیش‌بینی مستقیم یا یک زنجیرهٔ استدلال تولید کنند. روش استدلال سه‌مرحله‌ای THOR[۳] مسئله را به استخراج جنبه، استنتاج نظر ضمنی و تعیین قطبیت تقسیم می‌کند [۱]. بااین‌حال، طولانی‌ترشدن پاسخ تضمین‌کنندهٔ صحت نیست: خطای یک مرحلهٔ میانی می‌تواند به برچسب نهایی منتقل شود و یک توضیح روان نیز ممکن است بر فرضی نادرست تکیه کند. بنابراین مسئلهٔ این پژوهش فقط تولید استدلال نیست، بلکه مدیریت اختلاف میان پیش‌بینی مستقیم، استدلالی و تشخیصی است.

سامانهٔ پیشنهادی چند منبع پیش‌بینی را تولید و سپس با سیاستی تنظیم‌شده فقط بر پایهٔ دادهٔ آموزش، منبع نهایی را انتخاب می‌کند. ارزیابی اصلی بر نسخهٔ ضمنی و برچسب‌خوردهٔ SCAPT[۴] از دادهٔ SemEval 2014 انجام می‌شود. پس از پاک‌سازی، داده شامل ۲۱۸۸ نمونه در دو دامنهٔ لپ‌تاپ و رستوران است: ۱۷۴۶ نمونهٔ آموزش و ۴۴۲ نمونهٔ آزمون. تقسیم رسمی در تمام مراحل ثابت می‌ماند و هیچ برچسب طلایی آزمون در تنظیم انتخاب‌گر استفاده نمی‌شود.

ارزش کاربردی پژوهش در این است که دیدگاه‌های ضمنی بخش مهمی از تجربهٔ واقعی کاربران را حمل می‌کنند، اما با روش‌های متکی به واژگان آشکار به‌آسانی از دست می‌روند. از نظر روش‌شناختی نیز تفکیک تولید منابع، بازبینی و انتخاب نهایی امکان می‌دهد سهم هر مؤلفه و منشأ هر پاسخ ردیابی شود.

---

[۱] Sentiment Analysis

[۲] Implicit Sentiment Analysis (ISA)

[۳] Three-hop Reasoning؛ Reasoning Implicit Sentiment with Chain-of-Thought Prompting مقالهٔ در معرفی‌شده

[۴] Supervised Contrastive Pre-Training (SCAPT)

۱-۲ اهداف پژوهش

هدف اصلی، طراحی و ارزیابی یک سامانهٔ چندمنبعی برای تحلیل احساسات ضمنی است که بدون ریزتنظیم وزن مدل، پیش‌بینی مستقیم، استدلال چندمرحله‌ای، بازبینی ساختاریافته و انتخاب منبع را ترکیب کند. اهداف فرعی عبارت‌اند از:

- ساخت دادهٔ سه‌کلاسهٔ ضمنی با حفظ جدایی آموزش و آزمون و ایجاد خط پایهٔ کلاسیک و پیش‌بینی مستقیم؛
- پیاده‌سازی مسیر THOR با خودسازگاری سه‌اجرایی و بازبینی آگاه از نوع خطا؛
- تفکیک تصمیم میانی کنترل‌گر از منبع واقعی پاسخ نهایی و یادگیری سیاست محافظت‌شدهٔ انتخاب از دادهٔ آموزش؛
- ارزیابی کمی و کیفی مؤلفه‌ها و مقایسهٔ تکمیلی Qwen3 8B و Gemini 2.5 Flash روی زیرمجموعهٔ مشترک.

۱-۳ پرسش‌های پژوهش

۱. پیش‌بینی مستقیم و استدلال چندمرحله‌ای در تحلیل احساسات ضمنی چه تفاوتی دارند و آیا استدلال به‌تنهایی همواره موجب بهبود می‌شود؟

۲. آیا بازبینی ساختاریافتهٔ نوع خطا و کنترل‌گر صریح، اختلاف منابع را به تصمیمی قابل‌کنترل‌تر تبدیل می‌کنند؟

۳. آیا انتخاب منبع بر اساس الگوهای دادهٔ آموزش در آزمون از اتکای ثابت به یک منبع بهتر عمل می‌کند؟

۴. کدام الگوهای خطا، از جمله ازدست‌رفتن احساس، تفسیر بیش‌ازحد و جابه‌جایی دامنهٔ هدف، نقش اصلی را در شکست سامانه دارند؟

۵. انتخاب مدل و پشتانهٔ اجرا چه ارتباطی با عملکرد مسیر مستقیم، استدلال و انتخاب منبع دارد؟

۱-۴ نوآوری‌ها و دستاوردها

نوآوری پژوهش ارائهٔ الگوریتم بنیادی تازه برای مدل‌های زبانی نیست، بلکه طراحی و ارزیابی ترکیبی قابل‌ردیابی برای مدیریت منابع مکمل است. دستاوردهای اصلی شامل تبدیل بازبینی آزاد به تشخیص ساختاریافتهٔ خطا، جداسازی کنترل‌گر از انتخاب نهایی، و طراحی سیاست محافظت‌شده‌ای است که بر پایهٔ برچسب‌های مستقیم و THOR، نوع خطا، اطمینان تشخیصی و دامنه عمل می‌کند. گروه‌های کم‌پشتیبانی، کم‌حاشیه یا دیده‌نشده به پیش‌بینی مستقیم بازمی‌گردند.

در آزمون ۴۴۲ نمونه‌ای، سامانهٔ نهایی دقت را از ۰/۶۷۸۷۳۳ به ۰/۷۲۳۹۸۲ و اف‌یک ماکرو را از ۰/۶۷۴۰۷۵ به ۰/۷۱۹۱۱۹ رساند؛ ۲۶ خطای مسیر مستقیم را اصلاح کرد و شش پاسخ درست آن را تضعیف کرد. این نتیجه برتری مستقل THOR را نشان نمی‌دهد، زیرا مسیر استدلالی به‌تنهایی ضعیف‌تر از پیش‌بینی مستقیم بود؛ دستاورد محوری، انتخاب کنترل‌شدهٔ منبع در تنظیم آزمایشی حاضر است.

۱-۵ روش کلی پژوهش

پس از پاک‌سازی داده، خط پایهٔ TF-IDF و رگرسیون لجستیک و سپس منابع زبانی تولید می‌شوند. منبع مستقیم فقط جمله و هدف را دریافت می‌کند. مسیر THOR جنبه و نظر ضمنی را استخراج و استدلال قطبیت را به برچسب تبدیل می‌کند؛ سه اجرای مستقل آن با رأی اکثریت تجمیع می‌شوند. هنگام اختلاف مستقیم و THOR، بازبین نوع خطا، برچسب پیشنهادی و اطمینان را گزارش می‌کند و کنترل‌گر یک تصمیم میانی می‌سازد.

در مرحلهٔ نهایی، نمونه‌های آموزش بر اساس پنج سیگنال عملکردی گروه‌بندی و منبع مناسب هر گروه تعیین می‌شود. واگذاری به منبع غیرمستقیم تنها با پشتیبانی و حاشیهٔ کافی و آستانه‌های انتخاب‌شده در اعتبارسنجی داخلی انجام می‌شود. فصل چهارم فرمول‌ها و قواعد دقیق این فرایند را ارائه می‌کند. ارزیابی با دقت و اف‌یک ماکرو، تحلیل کلاس و دامنه، مطالعهٔ انتخاب‌گر و نمونه‌های کیفی انجام می‌شود. نتیجهٔ اصلی متعلق به Qwen3 8B روی آزمون کامل است؛ مقایسهٔ Gemini 2.5 Flash فقط شاهدی تکمیلی روی ۹۰ نمونهٔ آزمون مشترک است و اثر علّی ظرفیت مدل را نشان نمی‌دهد.

۱-۶ ساختار پایان‌نامه

فصل دوم مفاهیم لازم را تعریف می‌کند؛ فصل سوم ادبیات و شکاف پژوهشی را می‌کاود؛ فصل چهارم روش پیشنهادی را شرح می‌دهد؛ فصل پنجم نتایج و خطاها را تحلیل می‌کند؛ و فصل ششم به پرسش‌ها پاسخ می‌دهد و محدودیت‌ها و کارهای آتی را جمع‌بندی می‌کند.

فصل دوم

تعاریف اولیه

این فصل مفاهیمی را تعریف می‌کند که مستقیماً در ادبیات، روش و ارزیابی استفاده می‌شوند.

۲-۱ تحلیل احساسات، جنبه و قطبیت

تحلیل احساسات نگرش متن نسبت به یک موجودیت را استخراج می‌کند. این پژوهش مسئله را در سطح هدف یا جنبه بررسی می‌کند؛ بنابراین ورودی شامل جمله s و هدف t است و خروجی یکی از برچسب‌های $\mathcal{Y} = \{$مثبت، منفی، خنثی$\}$ خواهد بود:

$$f(s,t) \longrightarrow y, \qquad y \in \mathcal{Y}. \tag{۲-۱}$$

هدف[۱] موجودیتی است که قطبیت نسبت به آن سنجیده می‌شود و جنبه[۲] ویژگی یا بُعد مرتبط با آن است. نظر[۳] محتوایی است که ارزیابی را منتقل می‌کند. رابطهٔ این عناصر را می‌توان با (t, a, o, y) نمایش داد؛ در احساس ضمنی، جنبه یا نظر ممکن است در متن آشکار نباشد. تحلیل احساسات مبتنی بر جنبه[۴] به همین دلیل از برچسب‌گذاری کل سند دقیق‌تر است و در وظیفهٔ SemEval 2014 نیز با هدف‌های مشخص ارزیابی شده است [۳].

قطبیت مثبت مطلوبیت، قطبیت منفی نامطلوبیت و قطبیت خنثی نبود شواهد کافی برای ارزیابی دو جهت دیگر را نشان می‌دهد. خنثی با ابهام یکسان نیست: خنثی یک کلاس معنایی است، اما ابهام میزان قطعیت تفسیر را توصیف می‌کند.

۲-۲ تحلیل احساسات ضمنی و چالش‌های آن

در احساس صریح، عبارت ارزیابانه مسیر مستقیمی به قطبیت می‌سازد. در تحلیل احساسات ضمنی، نگرش از رویداد، پیامد یا انتظار استنباط می‌شود. برای مثال، در جملهٔ «پشتیبانی با نخستین تماس مشکل را حل کرد»، رویداد حل مشکل نسبت به هدف «پشتیبانی» نگرش مثبت می‌سازد، هرچند صفت مثبتی وجود ندارد.

[۱] Target
[۲] Aspect
[۳] Opinion
[۴] Aspect-Based Sentiment Analysis (ABSA)

این مسئله چهار چالش اصلی دارد: بازیابی شواهد غیرواژگانی، استفاده از دانش عمومی و وابسته به دامنه، رعایت محدودهٔ هدف در جمله‌های چندموجودیتی، و تفکیک نگرش ضمنی از توصیف واقعاً خنثی. استنتاج بیش‌ازحد یک گزارهٔ بی‌طرف را احساسی می‌کند و استنتاج ناکافی نگرش واقعی را از دست می‌دهد. همچنین یک توضیح زبانی منسجم لزوماً نشانهٔ استنتاج درست نیست؛ ازاین‌رو برچسب و مسیر استدلال باید جداگانه بررسی شوند [۱، ۲].

۲-۳ مدل‌های زبانی و مهندسی پرامپت

مدل زبانی بزرگ[۵] با دریافت دستور متنی می‌تواند برچسب یا توضیح تولید کند. پرامپت[۶] وظیفه، ورودی، مقادیر مجاز و شکل پاسخ را مشخص می‌کند و بخشی از روش آزمایشی است. در پرامپت صفرنمونه[۷] نمونهٔ حل‌شده‌ای در ورودی ارائه نمی‌شود و وزن‌های مدل نیز، برخلاف ریزتنظیم، تغییر نمی‌کنند.

خروجی آزاد برای توضیح مناسب است، اما پردازش خودکار آن دشوار است. خروجی ساختاریافته فیلدهایی مانند برچسب، نوع خطا و اطمینان را از پیش تعیین می‌کند. اعتبار نحویِ قالب با اعتبار معناییِ مقادیر متفاوت است؛ بنابراین □□□□□□ باید فقط برچسب معتبر و یکتا را بپذیرد و پاسخ مبهم را ناشناخته در نظر بگیرد.

۲-۴ استدلال چندمرحله‌ای، بازبینی و خودسازگاری

زنجیرهٔ تفکر[۸] مسئله را به تصمیم‌های میانی تقسیم می‌کند [۴]. برای احساس ضمنی، این مراحل می‌توانند استخراج جنبه، استنتاج نظر و تعیین قطبیت باشند. THOR همین تجزیهٔ سه‌مرحله‌ای را تخصصی می‌کند [۱]. مزیت آن ایجاد ردپای قابل‌بررسی است، اما خطای نخستین مرحله می‌تواند به بقیهٔ زنجیره منتقل شود و ردپای تولیدشده نیز گزارش قطعی سازوکار درونی مدل نیست.

بازبینی، متن، هدف و پاسخ اولیه را دوباره بررسی می‌کند. تشخیص خطا با اصلاح آن یکسان نیست: بازبین ممکن است ناسازگاری را درست تشخیص دهد، ولی جایگزین نادرستی پیشنهاد کند. همچنین اطمینان کلامی مدل، بدون آزمون کالیبراسیون، احتمال تجربی صحت محسوب نمی‌شود. بنابراین

[۵] Large Language Model (LLM)
[۶] Prompt
[۷] Zero-shot Prompting
[۸] Chain-of-Thought (CoT)

بازبینی در این پژوهش یک منبع اطلاعاتی است، نه پاسخ قطعی.

خودسازگاری[۹] یک مسیر غیرقطعی را K بار اجرا و برچسب‌ها را با رأی اکثریت تجمیع می‌کند [۵]:

$$\hat{y}_{\text{vote}} = \arg\max_{c \in \mathcal{Y}} \sum_{i=1}^{K} \mathbb{I}\big(y^{(i)} = c\big) \,. \tag{۲-۲}$$

توافق بیشتر اجراها صحت را تضمین نمی‌کند، زیرا مسیرها ممکن است خطای نظام‌مند مشترکی داشته باشند. سیاست رفع تساوی و انتخاب ردپای نماینده باید پیش از ارزیابی ثابت شود.

۲-۵ کنترل‌گر و انتخاب چندمنبعی

سامانهٔ چندمنبعی برای یک ورودی چند برچسب، مانند پیش‌بینی مستقیم، استدلالی و تشخیصی، تولید می‌کند. کنترل‌گر[۱۰] توافق، نوع خطا و اطمینان را به یک تصمیم میانی تبدیل می‌کند؛ انتخاب‌گر منبع[۱۱] مشخص می‌کند کدام منبع برچسب نهایی را تأمین کند. اگر $s_j(s,t)$ خروجی منبع j و $\mathbf{z}$ سیگنال‌های تصمیم باشد، آنگاه:

$$j^* = \pi(\mathbf{z}), \qquad \hat{y}_{\text{final}} = s_{j^*}(s,t). \tag{۳-۲}$$

این تفکیک منشأ پاسخ را روشن می‌کند و مانع یکسان‌گرفتن پیشنهاد بازبین با خروجی نهایی می‌شود. سیاست داده‌محور باید فقط با آموزش یا اعتبارسنجی ساخته شود؛ استفاده از برچسب طلایی آزمون نشت داده ایجاد می‌کند و برآورد عملکرد را خوش‌بینانه می‌سازد.

۲-۶ معیارهای ارزیابی

برای کلاس c، مقادیر TP_c، FP_c و FN_c به‌ترتیب پاسخ درست در کلاس، انتساب نادرست به کلاس و ازدست‌دادن نمونهٔ واقعی کلاس‌اند. ماتریس درهم‌ریختگی توزیع این جابه‌جایی‌ها را نشان می‌دهد. دقت

[۹] Self-Consistency

[۱۰] Controller

[۱۱] Source Selector

کلی[۱۲] نسبت همهٔ پیش‌بینی‌های درست است:

$$\text{Accuracy} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}(\hat{y}_i = y_i)\,. \tag{۴-۲}$$

صحت پیش‌بینی[۱۳]، بازخوانی[۱۴] و اف‌یک[۱۵] هر کلاس به صورت زیر محاسبه می‌شوند:

$$P_c = \frac{TP_c}{TP_c + FP_c}, \tag{۵-۲}$$

$$R_c = \frac{TP_c}{TP_c + FN_c}, \tag{۶-۲}$$

$$F_{1,c} = \frac{2P_cR_c}{P_c + R_c}. \tag{۷-۲}$$

اف‌یک ماکرو میانگین اف‌یک کلاس‌ها است:

$$\text{Macro-}F_1 = \frac{1}{C}\sum_{c=1}^{C}F_{1,c}. \tag{۸-۲}$$

در اف‌یک ماکرو هر کلاس مستقل از فراوانی خود وزن برابر دارد؛ بنابراین گزارش هم‌زمان آن با دقت کلی، ضعف روی کلاس کم‌تعداد را آشکارتر می‌کند. این سنجه‌ها صحت برچسب را می‌سنجند، نه درستی استدلال یا انتخاب منبع؛ فصل پنجم به همین دلیل تحلیل خطا و منشأ پاسخ را نیز گزارش می‌کند.

[۱۲] Accuracy
[۱۳] Precision
[۱۴] Recall
[۱۵] F1-score

فصل سوم

ادبیات تحقیق

این فصل پژوهش‌هایی را مرور می‌کند که مستقیماً به احساس ضمنی، استدلال مولد، بازبینی و انتخاب منبع مرتبط‌اند. تعریف‌های پایه در فصل دوم آمده‌اند و در اینجا تمرکز بر تحول راه‌حل‌ها و شکاف پژوهشی است.

۳-۱ تحلیل احساسات ضمنی و مجموعه‌داده‌ها

روش‌های نخستین تحلیل احساسات از فرهنگ واژگان و ویژگی‌های سطحی و سپس طبقه‌بندهای آماری استفاده می‌کردند [۶، ۷]. شبکه‌های عصبی و مدل‌های ازپیش‌آموزش‌دیده بازنمایی بافتی قوی‌تری ساختند و روش‌های گرافی نیز رابطهٔ نحوی هدف و عبارت نظر را مدل کردند [۸]. بااین‌حال، شواهد احساس ضمنی گاه اصلاً در واژه یا رابطهٔ نحوی آشکار حضور ندارند و به دانش عمومی نیاز دارند.

دادهٔ SemEval 2014 یکی از مبناهای تحلیل احساسات هدف‌محور در دامنه‌های لپ‌تاپ و رستوران است [۳]. لی و همکاران نمونه‌های صریح و ضمنی آن را تفکیک و روش پیش‌آموزش تقابلی نظارت‌شده SCAPT[۱] را پیشنهاد کردند [۹]. این روش بازنمایی را به تمایز قطبیت‌های ضمنی حساس می‌کند، اما به آموزش ویژه نیاز دارد و ردپای زبانی استنتاج را نشان نمی‌دهد. پژوهش‌های علّی نیز برای کاهش میان‌برهای آماری پیشنهاد شده‌اند، ولی همچنان بر اصلاح یک مدل واحد تمرکز دارند [۱۰].

چالش مشترک این رویکردها تمایز میان شواهد کافی و بیش‌تفسیر است. یک مدل ممکن است پیامد واقعی را از دست بدهد یا برعکس، گزارهٔ خنثی را احساسی تلقی کند. جابه‌جایی نگرش میان چند هدف نیز حتی با بازنمایی قوی باقی می‌ماند؛ مسئله‌ای که در تحلیل خطای پژوهش حاضر بررسی می‌شود.

۳-۲ استدلال و بازبینی در روش‌های مولد

مهندسی پرامپت نشان می‌دهد تعریف وظیفه، ترتیب اطلاعات و قالب پاسخ بر رفتار مدل اثر می‌گذارند [۱۱]. زنجیرهٔ تفکر با تولید مراحل میانی پیش از پاسخ نهایی، حل مسائل چندمرحله‌ای را آشکارتر کرد [۴]. در تحلیل احساسات ضمنی، این ایده اجازه می‌دهد مدل پیش از قطبیت، هدف، جنبه، رویداد و پیامد را بررسی کند.

در عین حال، استدلال زبانی محدودیت‌هایی دارد. یک مرحلهٔ آغازین نادرست می‌تواند بقیهٔ زنجیره

[۱] Supervised Contrastive Pre-Training

را منحرف کند و انسجام متن تضمین‌کنندهٔ صحت نیست. بازبینی و خوداصلاحی برای کشف ناسازگاری پاسخ اولیه توسعه یافته‌اند، اما بازبین نیز ممکن است پاسخ درست را بی‌دلیل تغییر دهد. بنابراین پذیرش اصلاح باید از تولید نقد جدا باشد؛ تمایزی که مبنای کنترل‌گر و انتخاب‌گر این پایان‌نامه است.

۳-۳ روش‌های THOR، SAoT و خودسازگاری

فی و همکاران THOR را برای تحلیل احساس ضمنی ارائه کردند [۱]. روش ابتدا جنبهٔ مرتبط را شناسایی، سپس نظر ضمنی را استنتاج و در پایان قطبیت را تعیین می‌کند. این تجزیه رابطهٔ پنهان رویداد و نگرش را قابل مشاهده می‌سازد، ولی اگر جنبه یا نظر اشتباه باشد، خطا به برچسب منتقل می‌شود.

THOR برای کاهش نوسان تولید از خودسازگاری استفاده می‌کند. در خودسازگاری چند مسیر نمونه‌گیری و پاسخ پرتکرار انتخاب می‌شود [۵]. این سازوکار وابستگی به یک اجرای تصادفی را کاهش می‌دهد، اما خطای مشترک مسیرها را حذف نمی‌کند. در پروژهٔ حاضر سه اجرای کامل مسیر استدلال رأی می‌دهند؛ این پیکربندی با ایدهٔ پژوهش اصلی همسو است، نه ادعای بازتولید دقیق تمام تنظیمات رسمی آن.

دوان و وانگ روش SAoT[۲] را پیشنهاد کردند [۲]. این روش پس از تحلیل جنبه و نظر، یک مرحلهٔ بازاندیشی را پیش از قطبیت وارد می‌کند. تفاوت اصلی آن با THOR جایگاه صریح بازبینی است. هر دو روش مراحل پنهان تصمیم را آشکار می‌کنند، اما هیچ‌یک انتخاب داده‌محور میان پاسخ مستقیم، استدلالی و تشخیصی را مسئلهٔ اصلی خود قرار نمی‌دهد.

۳-۴ انتخاب منبع و شکاف تحقیقاتی

جدول ۳-۱ پژوهش‌های محوری را از نظر سازوکار و محدودیت مرتبط با یک سامانهٔ چندمنبعی مقایسه می‌کند.

جدول حرکت ادبیات را از یادگیری بازنمایی به آشکارسازی و بازبینی استدلال نشان می‌دهد. این مؤلفه‌ها مکمل‌اند، اما هیچ‌کدام به‌تنهایی همهٔ نیازها را پوشش نمی‌دهد: بازنمایی قوی توضیح ندارد، استدلال ممکن است منحرف شود، رأی اکثریت خطای مشترک را تثبیت می‌کند و بازبینی می‌تواند

[۲] Sentiment Analysis of Thinking (SAoT)

جدول ۳-۱: مقایسهٔ پژوهش‌های محوری مرتبط با تحلیل احساسات ضمنی.

پژوهش	سازوکار	محدودیت مرتبط
SCAPT [۹]	پیش‌آموزش تقابلی برای احساس ضمنی	نیازمند آموزش ویژه و فاقد ردپای زبانی
مداخلهٔ علّی [۱۰]	کاهش اتکا به هم‌بستگی سطحی	تمرکز بر یک مدل، نه انتخاب منابع
زنجیرهٔ تفکر [۴]	تولید مراحل زبانی پیش از پاسخ	عمومی و مستعد انتشار خطا
خودسازگاری [۵]	رأی‌گیری چند مسیر استدلال	امکان تثبیت خطای مشترک
THOR [۱]	استنتاج جنبه، نظر و قطبیت	نبود انتخاب تطبیقی میان منابع
SAoT [۲]	تحلیل، بازاندیشی و قطبیت	خطر تغییر پاسخ درست و نبود قاعدهٔ پذیرش مستقل

بیش‌اصلاح کند.

شکاف تحقیقاتی، مدیریت اختلاف این منابع است. پژوهش حاضر به‌جای فرض برتری ثابت یک مسیر، ابتدا پاسخ مستقیم، استدلالی و تشخیصی را جدا تولید می‌کند؛ سپس با سیگنال‌هایی که فقط از دادهٔ آموزش آموخته شده‌اند، منبع نهایی را انتخاب می‌کند. شرط‌های پشتیبانی و حاشیه نیز انتخاب‌های کم‌شاهد را به مسیر مستقیم بازمی‌گردانند. فصل چهارم این سیاست محافظت‌شده و تمایز میان کنترل میانی و انتخاب نهایی را صورت‌بندی می‌کند.

فصل چهارم

کار پیشنهادی

این فصل سامانهٔ چندمنبعی پژوهش را صورت‌بندی می‌کند. اجزای اصلی عبارت‌اند از تولید پیش‌بینی مستقیم و استدلالی، بازبینی اختلاف، کنترل میانی و سیاست مستقل انتخاب پاسخ نهایی.

۴-۱ صورت‌بندی مسئله و نمای کلی سامانه

برای هر نمونه، جمله s_i، هدف t_i و در زمان آموزش برچسب y_i در دسترس است، به‌طوری‌که

$$\mathcal{Y} = \{\square\square\square\square\square\square\square\square\square, \square\square\square\square\square\square\square\square\square, \square\square\square\square\square\square\square\}.$$

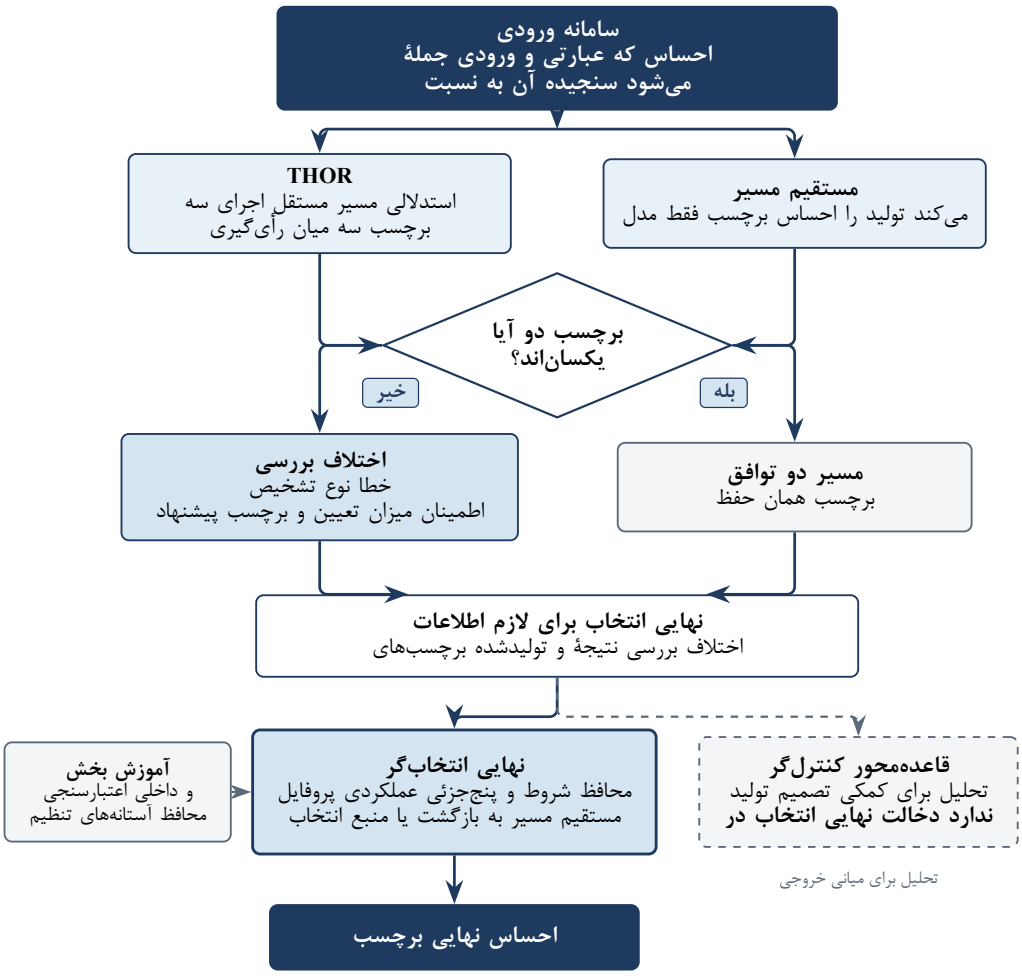
هدف، ساخت $\hat{y}_i$ بدون استفاده از برچسب نمونه‌های آزمون است. شکل ۴-۱ جریان اطلاعات را نشان می‌دهد.

داده از نسخهٔ برچسب‌خوردهٔ SCAPT مبتنی بر SemEval 2014 می‌آید [۳، ۹]. پس از حذف قطبیت‌های نامعتبر و گزینش نمونه‌های ضمنی، ۱۷۴۶ سطر آموزش و ۴۴۲ سطر آزمون باقی می‌ماند. نبود مقدار مفقود در فیلدهای ضروری پیش از فیلتر بررسی شده است. ستون ID به‌تنهایی یکتا نیست؛ تطبیق پایدار با کلید مرکب [id, source_sentence_id, sentence, target, from, to, polarity] انجام می‌شود.

پیش از روش‌های زبانی، خط پایهٔ کلاسیک از TF-IDF واژه‌ای و نویسه‌ای و رگرسیون لجستیک استفاده می‌کند. مقدار C فقط با اعتبارسنجی پنج‌لایه روی آموزش انتخاب و مدل منتخب روی کل آموزش بازبرازش می‌شود. این خط پایه نشان می‌دهد اطلاعات واژگانی بدون استنتاج مولد تا چه حد مسئله را حل می‌کنند.

۴-۲ تولید منابع پیش‌بینی

منبع مستقیم جمله و هدف را با دمای صفر به مدل می‌دهد و برچسب $\hat{y}_i^{(D)}$ را تولید می‌کند. منبع استدلالی از منطق THOR الهام می‌گیرد و مراحل زیر را اجرا می‌کند:



شکل ۴-۱: معماری سامانهٔ پیشنهادی؛ کنترل‌گر یک تصمیم میانی و انتخاب‌گر منبع تولیدکنندهٔ پاسخ نهایی است.

$$a_i = f_\theta(P_A(s_i, t_i))\,, \tag{۱-۴}$$

$$o_i = f_\theta(P_O(s_i, t_i, a_i))\,, \tag{۲-۴}$$

$$r_i = f_\theta(P_R(s_i, t_i, o_i))\,, \tag{۳-۴}$$

$$\hat{y}_i^{(T)} = f_\theta(P_Y(r_i))\,. \tag{۴-۴}$$

ورودی صریح پرامپت قطبیت، (s, t, o) است؛ جنبه می‌تواند غیرمستقیم از طریق استخراج نظر اثر بگذارد. استخراج جنبه، نظر و استدلال با دمای 0.7 انجام می‌شود، اما تبدیل استدلال به برچسب برای کاهش تغییرپذیری دمای صفر دارد. □□□□□□ فقط زمانی خروجی را می‌پذیرد که دقیقاً یک برچسب معتبر با مرز کلمه یافت شود؛ در غیر این صورت unknown ثبت می‌شود.

برای خودسازگاری سه‌اجرایی SC3[۱]، کل مسیر استدلال سه بار اجرا و برچسب دارای بیشترین رأی انتخاب می‌شود:

$$\hat{y}_i^{(T)} = \arg\max_{c \in \mathcal{Y}} \sum_{k=۱}^{۳} \mathbb{I}\Big(\hat{y}_{i,k}^{(T)} = c\Big)\,. \tag{۵-۴}$$

خروجی نامعتبر کنار گذاشته می‌شود؛ در تساوی، برچسبی که زودتر در ترتیب اجراهای مساوی ظاهر شده انتخاب و نخستین ردپای همسو با رأی به‌عنوان نماینده ذخیره می‌شود.

۴-۳ بازبینی آگاه از نوع خطا و کنترل‌گر میانی

بازبینی فقط هنگام اختلاف برچسب مستقیم و THOR فعال می‌شود. بازبین جمله، هدف، دو برچسب و ردپای استدلالی را دریافت و نوع خطا، برچسب پیشنهادی و اطمینان low/medium/high را در قالب ساختاریافته تولید می‌کند. رده‌ها در جدول ۴-۱ آمده‌اند.

کنترل‌گر قاعده‌محور در توافق، همان برچسب مشترک را نگه می‌دارد. در اختلاف، اصلاح فقط برای نوع خطای قابل اصلاح، برچسب تشخیصی معتبر و اطمینان دست‌کم متوسط پذیرفته می‌شود؛ در غیر

[۱] Self-Consistency with three reasoning runs

جدول ۴-۱: رده‌های بازبینی تشخیصی و نقش تصمیمی آن‌ها.

شناسه	معنا	نقش
missed_implicit_negative	ازدست‌رفتن نگرش منفی ضمنی	قابل اصلاح
missed_implicit_positive	ازدست‌رفتن نگرش مثبت ضمنی	قابل اصلاح
neutral_overinterpretation	نسبت‌دادن قطبیت به گزارهٔ خنثی	قابل اصلاح
target_scope_shift	انتقال نگرش از هدفی دیگر	قابل اصلاح
aspect_opinion_mismatch	ناسازگاری جنبه و نظر	قابل اصلاح
reasoning_label_inconsistency	ناسازگاری استدلال و برچسب	قابل اصلاح
no_error	تشخیص‌ندادن خطای قابل اصلاح	غیر اصلاحی
insufficient_evidence	نبود شاهد کافی برای تغییر	غیر اصلاحی

این صورت مسیر مستقیم مبنای بازگشت است. اگر مستقیم نامعتبر باشد، THOR و سپس تشخیصی بررسی می‌شوند. دلیل ماشینی هر تصمیم ذخیره می‌شود. این خروجی صرفاً نقطهٔ مقایسه است و مستقیماً پاسخ نهایی سامانه محسوب نمی‌شود.

۴-۴ انتخاب منبع محافظت‌شده و تنظیم‌شده با دادهٔ آموزش

نمونه‌ها با پنج سیگنال در پروفایل عملکردی زیر گروه‌بندی می‌شوند:

$$p_i = \left(\hat{y}_i^{(D)}, \hat{y}_i^{(T)}, e_i, c_i, d_i\right), \tag{۶-۴}$$

که e_i نوع خطا، c_i اطمینان و d_i دامنه است. برای منابع $\mathcal{S} = \{D, T, E\}$، بهترین منبع هر پروفایل فقط با شمار پاسخ‌های درست آموزش تعیین می‌شود:

$$\pi(p) = \arg\max_{s \in \mathcal{S}} \sum_{i \in \mathcal{D}_{\text{train}}} \mathbb{I}(p_i = p)\mathbb{I}(\hat{y}_i^{(s)} = y_i). \tag{۷-۴}$$

در تساوی، ترتیب مستقیم، THOR و تشخیصی اعمال می‌شود. برای جلوگیری از تصمیم کم‌شاهد، واگذاری به منبع غیرمستقیم چهار شرط دارد: حداقل پشتیبانی پروفایل، حاشیهٔ برتری نسبت به مستقیم، حاشیه نسبت به منبع دوم و حداقل بهبود نسبی. آستانه‌ها با ده تقسیم اعتبارسنجی داخلی آموزش و با حفظ گروه‌های دامنه و قطبیت انتخاب می‌شوند؛ معیار نخست اف‌یک ماکرو و سپس دقت است. پیکربندی نهایی حداقل پشتیبانی ۱۰، حاشیهٔ دو پاسخ درست نسبت به مستقیم، حاشیهٔ صفر نسبت به

منبع دوم و بهبود نسبی 0.05 دارد. پروفایل دیده‌نشده یا نامطمئن به مستقیم بازمی‌گردد.

پس از تثبیت آستانه‌ها، سیاست از کل آموزش آموخته و برای تولید پیش‌بینی نهایی بدون تغییر روی آزمون اعمال می‌شود:

$$\hat{y}_i^* = \hat{y}_i^{(\pi_g(p_i))}. \tag{۸-۴}$$

نام منبع نیز کنار برچسب ذخیره می‌شود. برچسب طلایی آزمون نه در ساخت پروفایل، نه در تنظیم آستانه و نه در انتخاب هر نمونه نقش دارد؛ فقط پس از تثبیت سیاست برای محاسبهٔ سنجه‌ها خوانده می‌شود.

۴-۵ انتخاب‌گرهای جایگزین

دو انتخاب‌گر یادگرفتنی نیز به‌عنوان مقایسه بررسی شدند. فرادسته‌بند درختی برچسب منبع را از خروجی منابع، توافق‌ها، نوع خطا، اطمینان، دامنه و ویژگی‌های رأی می‌آموزد. هدف آموزشی نخستین منبع درست در ترتیب ثابت است؛ بنابراین نمونه‌های چندصحیح به منبع زودتر متمایل می‌شوند. مدل و ابرپارامتر فقط با تقسیم داخلی آموزش انتخاب و سپس روی کل آموزش بازبرازش می‌شوند.

رتبه‌بند لجستیک هر زوج نمونه□منبع را یک ردیف می‌داند و احتمال درست‌بودن منبع را برآورد می‌کند؛ منبع با بیشترین امتیاز انتخاب می‌شود. اگر مدل برازش‌پذیر نباشد یا امتیاز معتبر تولید نکند، بازگشت مستقیم اجرا می‌شود. این روش‌ها از برچسب آزمون استفاده نمی‌کنند، اما همان‌طور که فصل پنجم نشان می‌دهد، یادگرفتنی‌بودن به‌تنهایی تضمین‌کنندهٔ عبور از خط پایهٔ مستقیم نیست.

۴-۶ پیاده‌سازی و جمع‌بندی

هر مرحله خروجی □□□ شامل کلید نمونه، پاسخ خام، برچسب نرمال‌شده و سیگنال‌های لازم را جدا ذخیره می‌کند. اجراهای طولانی قابل ادامه‌اند و پیش از ارزیابی، شمار ردیف‌ها، کلید مرکب، دامنه، تقسیم و برچسب‌ها میان فایل‌ها تطبیق داده می‌شوند. فهرست فایل‌ها و فرمان‌های بازتولید در پیوست آمده است.

در این معماری، استدلال منبع مکمل می‌سازد، بازبین اختلاف را ساختارمند می‌کند و انتخاب‌گر مستقل، پاسخ نهایی را با محافظ‌های آموخته‌شده از آموزش تعیین می‌کند. فصل پنجم عملکرد این اجزا و الگوهای موفقیت و شکست آن‌ها را ارزیابی می‌کند.

فصل پنجم

نتایج آزمایش‌ها

۵-۱ مقدمه

فصل چهارم سامانهٔ پیشنهادی را به‌صورت تولید چند منبع پیش‌بینی و انتخاب کنترل‌شدهٔ منبع نهایی صورت‌بندی کرد. در آن معماری، پیش‌بینی مستقیم، استدلال چندمرحله‌ای THOR و بازبینی تشخیصی نقش‌های متفاوتی دارند و خروجی کنترل‌گر قاعده‌محور نیز تصمیمی میانی است، نه پاسخ نهایی سامانه. پرسش اصلی این فصل آن است که این جداسازی در عمل چه اثری بر تشخیص احساسات ضمنی دارد و آیا انتخاب منبع بر پایهٔ بخش آموزش می‌تواند از اتکای ثابت به یک مسیر بهتر عمل کند.

برای پاسخ به این پرسش، ارزیابی در دو سطح انجام می‌شود. در سطح نخست، آزمایش‌ها با یک خط پایهٔ کلاسیک آغاز می‌شوند و سپس پیکربندی‌های مبتنی بر Qwen3 8B روی مجموعهٔ کامل آزمون مقایسه می‌شوند. این سطح برای سنجش فاصله با یک مرجع کم‌هزینه، مقایسهٔ روش‌ها، تحلیل کلاس‌ها، بررسی رفتار انتخاب‌گر و مطالعهٔ خطا به کار می‌رود. سطح دوم، مقایسه‌ای تکمیلی میان Qwen3 8B و Gemini 2.5 Flash روی زیرمجموعه‌ای مشترک و متوازن است. از آنجا که این زیرمجموعه کوچک‌تر از آزمون اصلی است، نتایج آن فقط برای مقایسهٔ دو پشتانهٔ مدل در شرایط یکسان تفسیر می‌شوند و جایگزین نتیجهٔ اصلی نیستند.

ادامهٔ فصل ابتدا تنظیمات آزمایش و روش‌های مورد مقایسه را مشخص می‌کند. سپس نتایج کمی Qwen3 8B، ماتریس‌های درهم‌ریختگی، عملکرد دامنه‌ای و رفتار انتخاب‌گر تحلیل می‌شوند. پس از بررسی نمونه‌های کیفی و خطاهای باقی‌مانده، مقایسهٔ تکمیلی با Gemini 2.5 Flash ارائه می‌شود. در پایان نیز محدودیت‌های ارزیابی و یافته‌های اصلی فصل جمع‌بندی می‌شوند.

۵-۲ تنظیمات آزمایش

۵-۲-۱ مجموعه‌داده و تقسیم ارزیابی

همان‌گونه که در فصل چهارم توضیح داده شد، داده‌های اصلی از نمونهٔ برچسب‌خورده و ضمنی SCAPT مبتنی بر SemEval 2014 به دست آمده‌اند. هر سطر شامل یک جمله و یک هدف مشخص است و برچسب آن یکی از سه قطبیت مثبت، منفی یا خنثی است. جدول ۵-۱ توزیع نمونه‌ها را در دو دامنهٔ لپ‌تاپ و رستوران نشان می‌دهد. تقسیم اصلی آموزش و آزمون بدون جابه‌جایی حفظ شده است.

در بخش آزمون، ۸۲ نمونه منفی، ۲۴۸ نمونه خنثی و ۱۱۲ نمونه مثبت وجود دارد. بنابراین، کلاس خنثی بیش از نیمی از آزمون را تشکیل می‌دهد. این توزیع سبب می‌شود که دقت کلی به‌تنهایی برای

جدول ۵-۱: توزیع نمونه‌های ضمنی در دو دامنه و دو بخش داده.

دامنه	آموزش	آزمون	مجموع
لپ‌تاپ	۷۱۶	۱۷۵	۸۹۱
رستوران	۱۰۳۰	۲۶۷	۱۲۹۷
مجموع	۱۷۴۶	۴۴۲	۲۱۸۸

تفسیر عملکرد کافی نباشد و رفتار هر سه کلاس نیز جداگانه بررسی شود.

۵-۲-۲ مدل‌ها و زیرساخت اجرا

اجرای اصلی با Qwen3 8B و از طریق Ollama انجام شده است. پیش‌بینی مستقیم با تولید قطعی و پاسخ کوتاه اجرا می‌شود. مسیر سازگارشدهٔ THOR کل زنجیرهٔ استدلال را سه بار اجرا می‌کند و برچسب نهایی این مسیر از رأی اکثریت به دست می‌آید. دمای مراحل استدلال ۰/۷ و دمای تبدیل استدلال به برچسب صفر است. بازبینی تشخیصی فقط در صورت اختلاف پیش‌بینی مستقیم و استدلالی فعال می‌شود.

مقایسهٔ تکمیلی از Gemini 2.5 Flash استفاده می‌کند. برای جلوگیری از مخلوط‌شدن تفاوت مدل با تفاوت داده، هر دو مدل روی زیرمجموعهٔ مشترک یکسان اجرا شده‌اند. وزن هیچ‌یک از مدل‌ها تغییر نکرده و همهٔ روش‌ها مبتنی بر پرامپت هستند. پاسخ خام، برچسب نرمال‌شده، اجزای استدلال، اطلاعات تشخیصی و منبع منتخب در فایل‌های جداگانه ذخیره شده‌اند.

۵-۲-۳ پروتکل ارزیابی و جلوگیری از نشت داده

انتخاب‌گر نهایی فقط از برچسب‌های طلایی بخش آموزش برای سنجش اعتبار منابع در پروفایل‌های مشاهده‌شده استفاده می‌کند. پس از ساخت سیاست، همان نگاشت بدون تغییر روی بخش آزمون اعمال می‌شود. برچسب طلایی آزمون فقط پس از تولید پیش‌بینی نهایی و برای محاسبهٔ معیارها خوانده می‌شود. انتخاب‌گرهای تنظیم‌شده با اعتبارسنجی نیز تقسیم‌های داخلی خود را فقط از بخش آموزش می‌سازند.

برای خط پایهٔ کلاسیک نیز همین اصل جداسازی رعایت شده است. انتخاب مقدار C و برازش بازنمایی‌های TF-IDF فقط در اعتبارسنجی پنج‌لایهٔ بخش آموزش انجام شد. پس از انتخاب $C = 1$، مدل روی کل آموزش بازبرازش شد و بخش آزمون فقط برای گزارش نهایی معیارها به کار رفت.

پیش از ارزیابی، هم‌ترازی سطرها میان خروجی‌های مستقیم، استدلالی، تشخیصی و نهایی کنترل

شده است. هر فایل ۲۱۸۸ سطر دارد، کلید مرکب نمونه در هر فایل یکتا است و تمام ۲۱۸۸ برچسب نهایی در فضای سه‌کلاسه معتبرند. ستون ID به‌تنهایی یکتا نیست و در این کنترل همراه شناسه جملهٔ مبدأ، متن، هدف، بازهٔ هدف و قطبیت به کار می‌رود. این کنترل مانع آن می‌شود که اختلاف ترتیب سطرها یا یک خروجی نامعتبر به‌صورت تغییر عملکرد تفسیر شود.

۵-۳ روش‌های مورد مقایسه و معیارهای ارزیابی

۵-۳-۱ روش‌های مورد مقایسه

برای نشان‌دادن مسیر پژوهش از یک مرجع کلاسیک تا سامانهٔ نهایی، هشت پیکربندی در آزمایش اصلی مقایسه می‌شوند:

۱. **خط پایهٔ کلاسیک:** پیش از آزمایش مدل‌های زبانی، ویژگی‌های واژه‌ای و نویسه‌ای TF-IDF به رگرسیون لجستیک متوازن داده می‌شوند و ضریب منظم‌سازی فقط با بخش آموزش انتخاب می‌شود.

۲. **پیش‌بینی مستقیم Qwen3 8B:** مدل فقط یکی از سه برچسب مجاز را بدون تولید زنجیرهٔ استدلال برمی‌گرداند.

۳. **THOR ساده‌شده:** مسیر چندمرحله‌ای نخستین پیاده‌سازی است که یک زنجیرهٔ استدلال برای هر نمونه تولید می‌کند.

۴. **THOR SC3 سازگارشدهٔ سه‌اجرایی:** سازگارسازی ویژهٔ این پژوهش است که کل مسیر را سه بار اجرا و برچسب‌های نهایی را با رأی اکثریت تجمیع می‌کند.

۵. **بازبینی ساده:** خروجی مسیر استدلالی بدون رده‌بندی ساختاریافتهٔ نوع خطا بازبینی می‌شود.

۶. **کنترل‌گر استاندارد:** بازبینی آگاه از نوع خطا[۱] روی THOR ساده‌شده اعمال و تصمیم میانی قاعده‌محور تولید می‌شود. اختصار ETC در شکل‌ها به همین کنترل‌گر آگاه از نوع خطا اشاره دارد.

۷. **کنترل‌گر روی THOR SC3:** همان قواعد با خروجی استدلالی سه‌اجرایی به کار می‌روند.

[۱] Error-Type-Aware Reflection; ETC

۸. **سامانهٔ نهایی انتخاب منبع:** سیاست محافظت‌شده و تنظیم‌شده با اعتبارسنجی‌های داخلی آموزش، یکی از منابع مستقیم، استدلالی یا تشخیصی را برای هر پروفایل انتخاب می‌کند و در وضعیت‌های کم‌پشتیبانی یا کم‌حاشیه به مسیر مستقیم بازمی‌گردد.

دو پیکربندی ششم و هفتم خروجی کنترل‌گر میانی را ارزیابی می‌کنند. بنابراین، نباید آن‌ها را با پیکربندی هشتم که پیش‌بینی نهایی سامانه است یکسان دانست. این تفکیک نشان می‌دهد آیا اجرای قواعد ثابت به‌تنهایی کافی است یا انتخاب منبع متناسب با وضعیت نمونه ارزش افزوده ایجاد می‌کند.

۵-۳-۲ معیارهای ارزیابی

تعریف دقت کلی، صحت پیش‌بینی، بازخوانی، امتیاز اف‌یک هر کلاس و امتیاز اف‌یک ماکرو در روابط ۲-۴ تا ۲-۸ فصل دوم ارائه شد. در این فصل، دقت کلی برای سنجش نسبت کل پاسخ‌های درست و امتیاز اف‌یک ماکرو برای وزن‌دادن یکسان به سه کلاس گزارش می‌شود. نزدیک‌بودن تغییر این دو معیار نشان می‌دهد که بهبود فقط حاصل غلبهٔ کلاس خنثی نیست؛ بااین‌حال، برای روشن‌شدن محل تغییرها، ماتریس درهم‌ریختگی و اف‌یک هر کلاس نیز بررسی می‌شوند.

۵-۴ نتایج روش‌های اصلی روی آزمون کامل

۵-۴-۱ مقایسهٔ کلی روش‌ها

جدول ۵-۲ نتایج هشت روش را روی ۴۴۲ نمونهٔ آزمون گزارش می‌کند. ترتیب روش‌ها مسیر آزمایش را از خط پایهٔ کلاسیک، پیش‌بینی مستقیم مدل زبانی، روش‌های استدلالی و بازبینی تا انتخاب منبع نهایی دنبال می‌کند. نسخهٔ THOR سه‌اجرایی نیز بلافاصله پس از نسخهٔ ساده‌شده قرار گرفته است.

جدول ۵-۲: نتایج روش‌های اصلی، از خط پایهٔ کلاسیک تا سامانهٔ نهایی، روی مجموعهٔ کامل آزمون.

روش	دقت کلی	اف‌یک ماکرو
خط پایهٔ کلاسیک TF-IDF + Logistic Regression	۰/۵۴۷۵۱۱	۰/۵۱۴۰۸۲
پیش‌بینی مستقیم Qwen3 8B	۰/۶۷۸۷۳۳	۰/۶۷۴۰۷۵
THOR ساده‌شده	۰/۵۹۹۵۴۸	۰/۵۷۸۴۳۷
THOR SC3 سازگارشدهٔ سه‌اجرایی	۰/۵۹۰۴۹۸	۰/۶۰۰۶۹۰
بازبینی ساده	۰/۶۱۵۳۸۵	۰/۶۰۶۷۳۲
کنترل‌گر استاندارد	۰/۶۶۰۶۳۳	۰/۶۵۹۴۳۰
کنترل‌گر روی THOR SC3	۰/۶۶۰۶۳۳	۰/۶۶۲۱۰۸
سامانهٔ نهایی انتخاب منبع	۰/۷۲۳۹۸۲	۰/۷۱۹۱۱۹

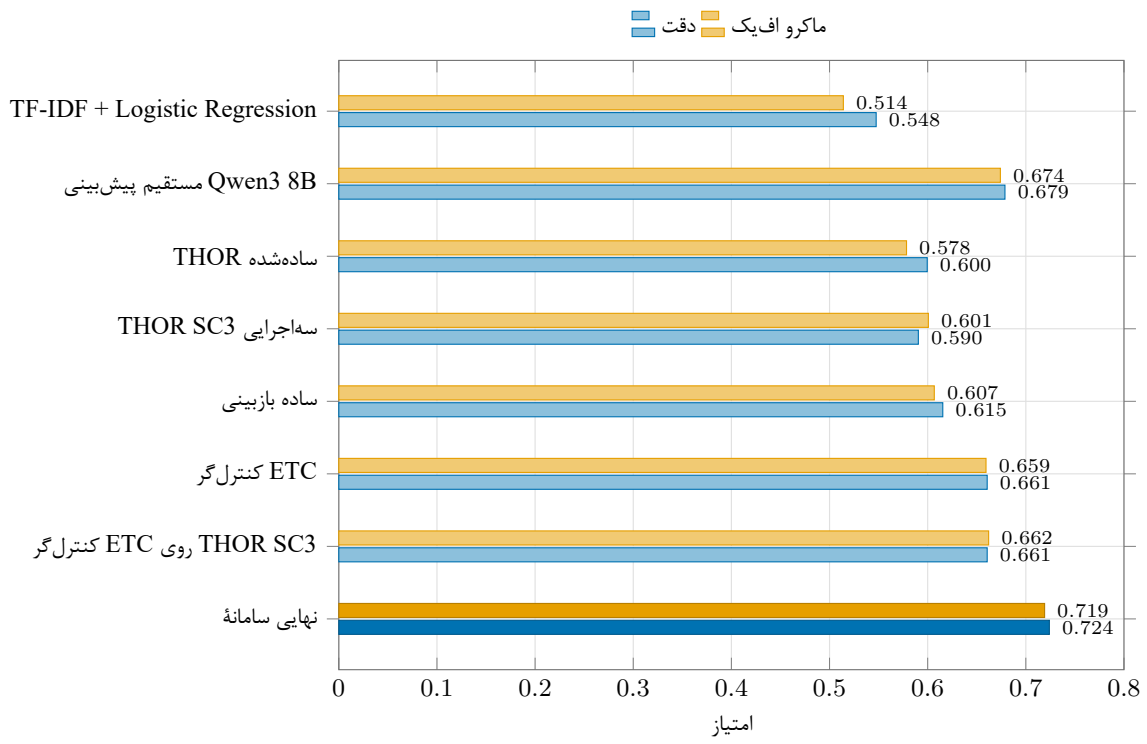
خط پایهٔ کلاسیک در نخستین آزمایش به دقت ۰/۵۴۷۵۱۱ و اف‌یک ماکرو ۰/۵۱۴۰۸۲ رسید. اف‌یک کلاس منفی در این روش ۰/۳۷۴۳۳۲ بود و نشان داد که تشخیص احساس ضمنی منفی دشوارترین بخش این بازنمایی کلاسیک است. پیش‌بینی مستقیم Qwen3 8B دقت را حدود ۱۳/۱۲ واحد درصد و اف‌یک ماکرو را حدود ۱۶/۰۰ واحد درصد نسبت به این مرجع افزایش داد. این مقایسه نشان می‌دهد که حتی پیش از افزودن مراحل استدلال و انتخاب منبع، مدل زبانی اطلاعات معنایی بیشتری از بازنمایی واژگانی و نویسه‌ای استخراج می‌کند.

سامانهٔ نهایی نسبت به خط پایهٔ کلاسیک حدود ۱۷/۶۵ واحد درصد در دقت و ۲۰/۵۰ واحد درصد در اف‌یک ماکرو برتری دارد. بااین‌حال، این فاصله نباید فقط به معماری استدلالی پیشنهادی نسبت داده شود؛ زیرا خط پایهٔ کلاسیک و روش‌های مبتنی بر مدل زبانی از خانواده‌ها و ظرفیت‌های متفاوتی استفاده می‌کنند. نقش این آزمایش فراهم‌کردن یک مرجع کلاسیک بازتولیدپذیر است و این روش جایگزین مقایسه با رمزگذارهای ازپیش‌آموزش‌دیده و تنظیم‌دقیق‌شده نیست.

نتایج نشان می‌دهند که افزودن استدلال چندمرحله‌ای به‌تنهایی بهبود ایجاد نکرده است. دقت هر دو نسخهٔ THOR از پیش‌بینی مستقیم کمتر است. خودسازگاری سه‌اجرایی نیز اگرچه اف‌یک ماکرو را نسبت به THOR ساده‌شده از ۰/۵۷۸۴۳۷ به ۰/۶۰۰۶۹۰ رسانده، دقت کلی را اندکی کاهش داده است. این تفاوت نشان می‌دهد که رأی‌گیری، توزیع خطا میان کلاس‌ها را تغییر داده است، اما ضعف مسیر استدلالی را به‌طور کامل برطرف نمی‌کند.

بازبینی ساده نسبت به هر دو مسیر استدلالی بهتر عمل کرده، ولی همچنان از خط پایهٔ مستقیم پایین‌تر است. کنترل‌گرهای قاعده‌محور بخش مهمی از این فاصله را جبران می‌کنند، اما هیچ‌یک از خط پایهٔ مستقیم عبور نمی‌کنند. در مقابل، سامانهٔ نهایی محافظت‌شده به دقت ۰/۷۲۳۹۸۲ و اف‌یک ماکرو ۰/۷۱۹۱۱۹ رسیده است. نسبت به پیش‌بینی مستقیم، افزایش دقت حدود ۴/۵۲ واحد درصد و افزایش اف‌یک ماکرو حدود ۴/۵۰ واحد درصد است.

شکل ۵-۱ همین مقایسهٔ هشت‌روش را از خط پایهٔ کلاسیک تا سامانهٔ نهایی نمایش می‌دهد. بااین‌حال، نزدیک‌بودن دقت و اف‌یک ماکرو به‌تنهایی برای داوری دربارهٔ همهٔ کلاس‌ها کافی نیست؛ جدول ۵-۳ نشان می‌دهد که امتیاز اف‌یک هر سه کلاس منفی، خنثی و مثبت نسبت به پیش‌بینی مستقیم افزایش یافته است.



شکل ۵-۱: مقایسهٔ دقت کلی و امتیاز اف‌یک ماکرو هشت روش اصلی روی ۴۴۲ نمونهٔ آزمون.

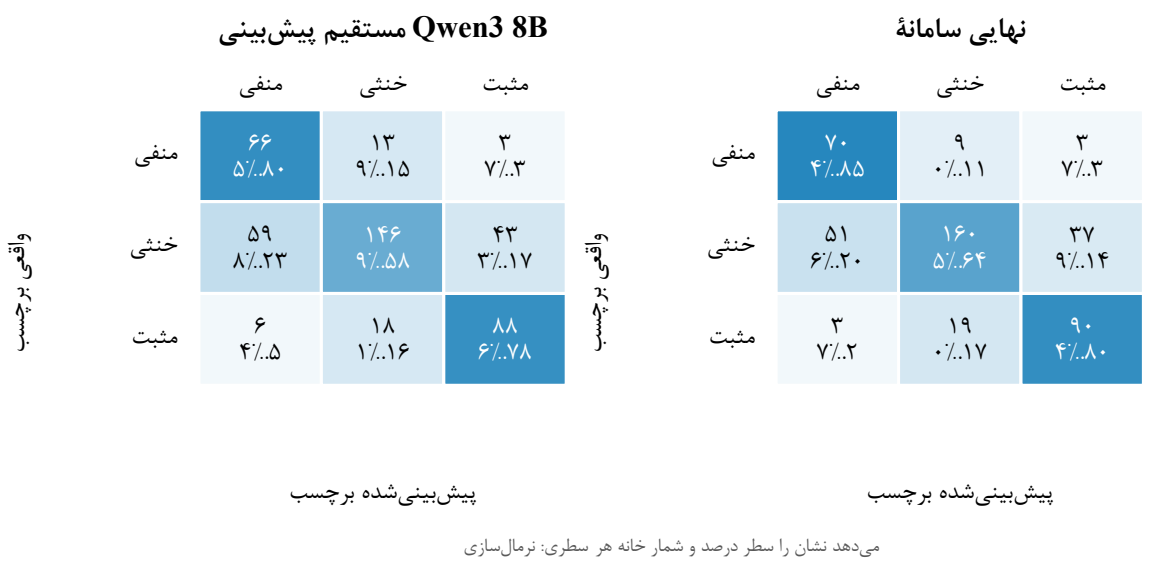
۵-۴-۲ تحلیل ماتریس‌های درهم‌ریختگی

برای مشخص‌شدن منشأ بهبود، شکل ۵-۲ ماتریس‌های درهم‌ریختگی پیش‌بینی مستقیم و سامانهٔ نهایی را مقایسه می‌کند. سطرها برچسب طلایی و ستون‌ها برچسب پیش‌بینی‌شده‌اند و درصدهای هر خانه نسبت به مجموع همان سطر محاسبه شده‌اند.

در کلاس منفی، شمار پاسخ‌های درست از ۶۶ به ۷۰ رسیده و بازخوانی از ۸۰/۵ درصد به ۸۵/۴ درصد افزایش یافته است. هم‌زمان، خطای منفی به خنثی از ۱۳ به ۹ نمونه کاهش یافته است. در کلاس خنثی، شمار پاسخ‌های درست از ۱۴۶ به ۱۶۰ رسیده و بازخوانی از ۵۸/۹ درصد به ۶۴/۵ درصد رسیده است. کاهش خنثی‌های اشتباه پیش‌بینی‌شده به‌عنوان منفی، از ۵۹ به ۵۱ نمونه، مهم‌ترین تغییر این کلاس است.

در کلاس مثبت نیز شمار پاسخ‌های درست از ۸۸ به ۹۰ افزایش یافته است. خطای مثبت به منفی از ۶ به ۳ نمونه کاهش یافته، هرچند خطای مثبت به خنثی از ۱۸ به ۱۹ رسیده است. بنابراین، بهبود نهایی حاصل حذف یکنواخت همهٔ خطاها نیست؛ انتخاب‌گر بخشی از خطاها را اصلاح کرده و هم‌زمان تعداد محدودی خطای تازه ایجاد کرده است.

جدول ۵-۳ نشان می‌دهد که اف‌یک هر سه کلاس نسبت به پیش‌بینی مستقیم افزایش یافته است.



می‌دهد نشان را سطر درصد و شمار خانه هر سطری: نرمال‌سازی

شکل ۵-۲: ماتریس‌های درهم‌ریختگی پیش‌بینی مستقیم و سامانهٔ نهایی Qwen3 8B روی آزمون کامل.

بیشترین افزایش مربوط به کلاس منفی است و کلاس خنثی نیز با وجود فراوانی بیشتر، بهبود قابل توجهی دارد.

جدول ۵-۳: امتیاز اف‌یک کلاس‌ها برای پیش‌بینی مستقیم و سامانهٔ نهایی Qwen3 8B.

کلاس	پیش‌بینی مستقیم	سامانهٔ نهایی
منفی	۰/۶۱۹۷۱۸	۰/۶۷۹۶۱۲
خنثی	۰/۶۸۷۰۵۹	۰/۷۳۳۹۴۵
مثبت	۰/۷۱۵۴۴۷	۰/۷۴۳۸۰۲

۵-۴-۳ تحلیل عملکرد بر اساس دامنه

جدول ۵-۴ نتایج دو دامنه را جدا می‌کند. سامانهٔ نهایی در هر دو دامنه از پیش‌بینی مستقیم بهتر است؛ بنابراین، افزایش سنجهٔ کل فقط از یک دامنه ناشی نشده است.

جدول ۵-۴: عملکرد پیش‌بینی مستقیم و سامانهٔ نهایی به تفکیک دامنه در بخش آزمون.

دامنه	تعداد	مستقیم		نهایی	
		دقت	اف‌یک ماکرو	دقت	اف‌یک ماکرو
لپ‌تاپ	۱۷۵	۰/۶۸۵۷۱۴	۰/۶۷۳۶۴۷	۰/۷۲۰۰۰۰	۰/۷۰۳۵۹۶
رستوران	۲۶۷	۰/۶۷۴۱۵۷	۰/۶۷۱۶۵۰	۰/۷۲۶۵۹۲	۰/۷۲۵۸۰۶

در دامنهٔ لپ‌تاپ، افزایش اف‌یک ماکرو حدود ۳/۰ واحد درصد است. این افزایش در دامنهٔ رستوران به حدود ۵/۴۲ واحد درصد می‌رسد. متن‌های رستوران در بسیاری از موارد پیامدهایی مانند انتظار،

قیمت، کیفیت خدمت و رفتار کارکنان را توصیف می‌کنند؛ ازاین‌رو، مسیر استدلالی در بخشی از این نمونه‌ها اطلاعات مکمل بیشتری نسبت به پیش‌بینی مستقیم فراهم کرده است. بااین‌حال، این مشاهده یک رابطهٔ علّی را اثبات نمی‌کند و فقط با الگوی خطاهای همین آزمون سازگار است.

۵-۵ تحلیل انتخاب‌گر و مؤلفه‌های سامانه

۵-۵-۱ توزیع منابع و سود یا زیان نسبت به مسیر مستقیم

انتخاب منبع به معنای تغییر قطعی وضعیت صحت نیست؛ ممکن است دو منبع برچسب‌های متفاوت اما هر دو نادرست تولید کرده باشند. جدول ۵-۵ نشان می‌دهد که سیاست محافظت‌شده در ۴۰۹ نمونهٔ آزمون منبع مستقیم و فقط در ۳۳ نمونه منبع THOR را انتخاب کرده است. برچسب تشخیصی در هیچ نمونهٔ آزمون منبع نهایی نبوده است. بنابراین، بهبود آزمون از واگذاری محدود به پروفایل‌هایی حاصل شده که پشتیبانی و حاشیهٔ آموزشی کافی داشته‌اند، نه از اتکای گسترده به مسیر استدلالی یا برچسب بازبین.

جدول ۵-۵: توزیع منبع منتخب در ۴۴۲ نمونهٔ آزمون.

منبع منتخب	تعداد نمونه
پیش‌بینی مستقیم	۴۰۹
THOR	۳۳
برچسب تشخیصی	صفر

مقایسهٔ مستقیم برچسب نهایی با خط پایه نشان می‌دهد که دو روش در ۴۱۰ نمونه نتیجهٔ صحت یکسان داشته‌اند: هر دو در ۲۹۴ نمونه درست و در ۱۱۶ نمونه نادرست بوده‌اند. سامانهٔ نهایی ۲۶ خطای پیش‌بینی مستقیم را اصلاح کرده و در ۶ نمونه، پاسخ درست مستقیم را به پاسخ نادرست تبدیل کرده است. حاصل خالص این جابه‌جایی، ۲۰ پاسخ درست بیشتر و افزایش شمار پاسخ‌های صحیح از ۳۰۰ به ۳۲۰ است.

هر ۳۳ نمونهٔ واگذارشده به THOR برچسبی متفاوت از مسیر مستقیم داشته‌اند و در ۳۲ نمونه وضعیت صحت نیز تغییر کرده است؛ یک نمونه میان دو پاسخ نادرست جابه‌جا شده است. این تمرکز نشان می‌دهد شروط محافظ بیشتر انتخاب‌های بی‌اثر نسخهٔ بدون محافظ را حذف کرده‌اند، درحالی‌که همان ۲۶ اصلاح و ۶ تضعیف اصلی حفظ شده‌اند.

۵-۵-۲ مقایسهٔ تصمیم میانی، انتخاب‌گرهای فراسطح و کران اوراکل

جدول ۵-۶ سامانهٔ نهایی را با چند نقطهٔ مقایسه نشان می‌دهد. کنترل‌گر قاعده‌محور همان خروجی میانی فصل چهارم است. انتخاب‌گر فراسطح با یک درخت تصمیم تنظیم‌شده روی اعتبارسنجی و رتبه‌بند لجستیک نیز تلاش می‌کنند منبع مناسب را با ویژگی‌های بیشتری پیش‌بینی کنند. اوراکل از برچسب طلایی برای انتخاب بهترین منبع هر نمونه استفاده می‌کند و فقط کران بالای تحلیلی است؛ این روش در کاربرد واقعی قابل استفاده نیست.

جدول ۵-۶: مقایسهٔ سیاست‌های انتخاب منبع روی ۴۴۲ نمونهٔ آزمون Qwen3 8B.

روش	دقت کلی	اف‌یک ماکرو
پیش‌بینی مستقیم	۰/۶۷۸۷۳۳	۰/۶۷۴۰۷۵
کنترل‌گر قاعده‌محور	۰/۶۶۰۶۳۳	۰/۶۶۲۱۰۸
رتبه‌بند لجستیک منبع	۰/۷۱۲۶۷۰	۰/۶۹۷۹۱۳
انتخاب‌گر فراسطح مبتنی بر درخت	۰/۷۱۹۴۵۷	۰/۷۱۳۴۵۲
انتخاب پروفایلی بدون محافظ (حذف مؤلفه)	۰/۷۲۳۹۸۲	۰/۷۱۹۲۰۴
سامانهٔ نهایی محافظت‌شده	۰/۷۲۳۹۸۲	۰/۷۱۹۱۱۹
اوراکل انتخاب منبع	۰/۷۸۵۰۶۸	۰/۷۷۹۵۱۱

کنترل‌گر قاعده‌محور از خط پایهٔ مستقیم پایین‌تر است؛ پس قواعد ثابت پذیرش اصلاح برای ساخت پاسخ نهایی کافی نیستند. انتخاب‌گرهای یادگرفتنی از خط پایه عبور کرده‌اند. نسخهٔ پروفایلی بدون محافظ اف‌یک ماکرویی فقط ۰/۰۰۰۰۸۵ بیشتر از روش اصلی دارد، اما ۱۴۲ نمونه را به THOR واگذار می‌کند؛ در مقابل، آستانه‌های روش محافظت‌شده فقط با اعتبارسنجی آموزش انتخاب شده‌اند و شمار واگذاری را به ۳۳ کاهش می‌دهند. بنابراین، روش محافظت‌شده با عملکرد عملاً برابر و تصمیم‌های محافظه‌کارانه‌تر به‌عنوان سامانهٔ اصلی در نظر گرفته می‌شود و نسخهٔ بدون محافظ نقش مطالعهٔ حذف مؤلفه را دارد. فاصلهٔ اف‌یک ماکرو سامانهٔ نهایی با اوراکل حدود ۶/۰۴ واحد درصد است.

۵-۵-۳ اعتبارسنجی زنجیرهٔ نهایی

اعتبارسنجی فقط به شمار سطرها محدود نشده است. کلیدهای نمونه در چهار خروجی مستقیم، THOR، کنترل‌گر و انتخاب‌گر بررسی شده‌اند. هیچ کلید تکراری وجود ندارد، ۲۱۸۸ سطر مستقیم و کنترل‌گر با یکدیگر هم‌ترازند، ۲۱۸۸ سطر THOR با کنترل‌گر تطبیق دارند و تمام سطرهای کنترل‌گر و انتخاب‌گر نیز هم‌ترازند. ستون‌های تشخیصی در فایل نهایی حاضرند و هیچ برچسبی خارج از مثبت، منفی و خنثی مشاهده نشده است. در نتیجه، سنجه‌های گزارش‌شده بر یک زنجیرهٔ کامل و قابل ردیابی محاسبه شده‌اند.

۵-۶ تحلیل کیفی و خطا

۵-۶-۱ نمونه‌های اصلاح‌شده

در بخش آزمون، ۲۶ نمونه وجود دارد که پیش‌بینی مستقیم در آن‌ها نادرست و سامانهٔ نهایی درست است. نمونهٔ زیر به دامنهٔ رستوران و هدف price مربوط است:

For the price you pay for the food here, you'd expect it to be at least on par with other Japanese restaurants.

برچسب طلایی این نمونه منفی است. مسیر مستقیم آن را خنثی پیش‌بینی کرده، اما THOR رابطهٔ میان قیمت پرداخت‌شده و برآورده‌نشدن انتظار را منفی استنتاج کرده است. انتخاب‌گر منبع THOR را برگزیده و خطای «از دست‌رفتن احساس منفی ضمنی» اصلاح شده است. این نمونه با تعریف فصل دوم سازگار است: واژهٔ احساسی صریح برای هدف وجود ندارد و قطبیت از مقایسهٔ هزینه و انتظار به دست می‌آید.

نمونهٔ اصلاح‌شدهٔ دیگر به هدف OS X Mountain Lion مربوط است:

Got this Mac Mini with OS X Mountain Lion for my wife.

پیش‌بینی مستقیم برچسب مثبت تولید کرده است، در حالی که جمله فقط مالکیت و دریافت دستگاه را بیان می‌کند و نگرش مشخصی نسبت به سیستم‌عامل ندارد. مسیر استدلالی نبود شاهد احساسی را تشخیص داده و برچسب خنثی تولید کرده است. انتخاب همین منبع از بیش‌تفسیر یک گزارهٔ خبری جلوگیری کرده است.

۵-۶-۲ نمونه‌های تضعیف‌شده

در مقابل، شش نمونه وجود دارد که پیش‌بینی مستقیم درست بوده و سامانهٔ نهایی آن‌ها را نادرست کرده است. نمونهٔ زیر برای هدف food برچسب منفی دارد:

I have never in my life sent back food before, but I simply had to, and the waiter argued with me over this.

مسیر مستقیم منفی را به‌درستی تشخیص داده است، اما مسیر THOR عبارت بازگرداندن غذا را به‌عنوان شاهد ناکافی تفسیر کرده و خنثی داده است. انتخاب‌گر نیز به دلیل شباهت پروفایل این نمونه با مواردی که THOR در آموزش بهتر بوده، منبع استدلالی را انتخاب کرده است. در اینجا خطای مرحلهٔ استدلال به انتخاب نهایی منتقل شده است.

نمونهٔ بعدی اهمیت محدودهٔ هدف را نشان می‌دهد:

> Unfortunately, we chose this spot for lunch as we had done a lot of walking and ended up at the South St Seaport.

هدف نمونه lunch و برچسب طلایی آن خنثی است. مسیر مستقیم خنثی را حفظ کرده، اما مسیر استدلالی نشانهٔ منفی ابتدای جمله و وضعیت کلی انتخاب مکان را به هدف ناهار منتقل کرده است. این خطا نمونه‌ای از جابه‌جایی دامنهٔ هدف است: وجود احساس منفی در جمله برای منفی‌بودن هر هدف کافی نیست.

۵-۶-۳ الگوهای خطای باقی‌مانده

نمونه‌های کیفی سه الگوی اصلی را آشکار می‌کنند. الگوی نخست، از دست‌رفتن پیامد ضمنی است؛ یعنی مدل رابطهٔ میان انتظار، رویداد و ارزیابی را بازیابی نمی‌کند. الگوی دوم، بیش‌تفسیر گزارهٔ خنثی است؛ در این حالت مدل از رویدادی که فقط اطلاعات توصیفی دارد، قطبیت می‌سازد. الگوی سوم، انتقال نگرش میان هدف‌های مختلف یک جمله است. مسیر چندمرحله‌ای می‌تواند الگوی نخست را اصلاح کند، اما اگر جنبه یا هدف را نادرست تثبیت کند، دو الگوی بعدی را تشدید می‌کند.

وجود ۱۱۶ نمونه‌ای که هم پیش‌بینی مستقیم و هم سامانهٔ نهایی در آن‌ها نادرست‌اند نشان می‌دهد که مسئله فقط انتخاب میان منابع فعلی نیست. بخشی از خطاها میان منابع مشترک‌اند و برای اصلاح آن‌ها به استدلال هدف‌محور دقیق‌تر، دانش زمینه‌ای مناسب‌تر یا منابع پیش‌بینی متنوع‌تر نیاز است. در عین حال، فاصلهٔ اوراکل با سامانهٔ نهایی نشان می‌دهد که بهبود خود انتخاب‌گر نیز هنوز ظرفیت قابل توجهی دارد.

۵-۷ مقایسهٔ تکمیلی با Gemini 2.5 Flash

۵-۷-۱ ساخت زیرمجموعهٔ مشترک

برای مقایسهٔ دو مدل، زیرمجموعه‌ای متوازن با ۲۴۰ نمونه ساخته شده است. این زیرمجموعه ۱۵۰ نمونهٔ آموزش و ۹۰ نمونهٔ آزمون دارد و بخش آزمون شامل ۳۰ نمونه از هر کلاس است. تمام مسیرهای مستقیم، THOR سه‌اجرایی و انتخاب منبع برای هر دو مدل روی همین شناسه‌ها اجرا شده‌اند. بنابراین، اختلاف نتایج این بخش از تفاوت جمعیت آزمون ناشی نمی‌شود.

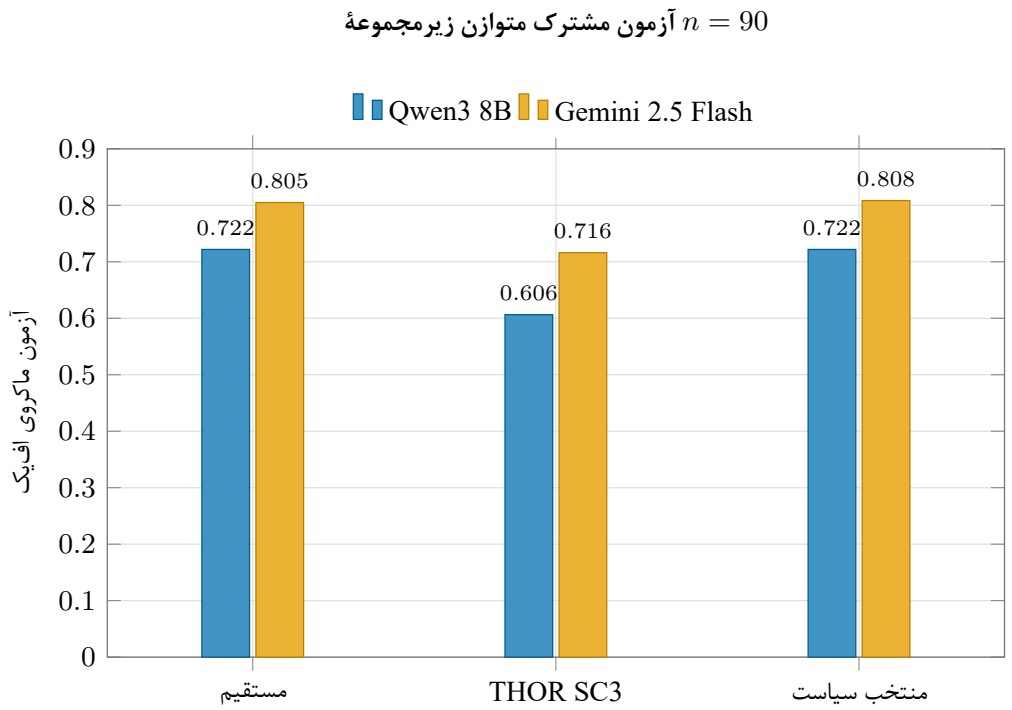
بااین‌حال، اندازهٔ ۹۰ نمونه برای نتیجه‌گیری کلی دربارهٔ برتری یک مدل کافی نیست. هدف این آزمایش بررسی رفتار دو پشتانه در یک چارچوب و مجموعهٔ یکسان است. نتیجهٔ اصلی پایان‌نامه همچنان همان ارزیابی Qwen3 8B روی ۴۴۲ نمونهٔ آزمون کامل است.

۵-۷-۲ مقایسهٔ سنجه‌های دو مدل

شکل ۵-۳ امتیاز اف‌یک ماکرو دو مدل را در سه پیکربندی مشترک نشان می‌دهد. پیکربندی انتخاب منبع این بخش با تقسیم‌های اعتبارسنجی ساخته‌شده از بخش آموزش تنظیم شده است و هیچ برچسب طلایی آزمون در انتخاب سیاست دخالت ندارد.

در مسیر مستقیم، اف‌یک ماکرو Qwen3 8B برابر ۰/۷۲۱۹۰۳ و مقدار Gemini 2.5 Flash برابر ۰/۸۰۴۸۸۶ است. در مسیر THOR SC3، مقادیر به‌ترتیب ۰/۶۰۶۴۵۰ و ۰/۷۱۶۱۰۹ هستند. بنابراین، مسیر استدلالی برای هر دو مدل از پیش‌بینی مستقیم ضعیف‌تر است، هرچند افت مشاهده‌شده در پشتانهٔ Gemini کمتر است.

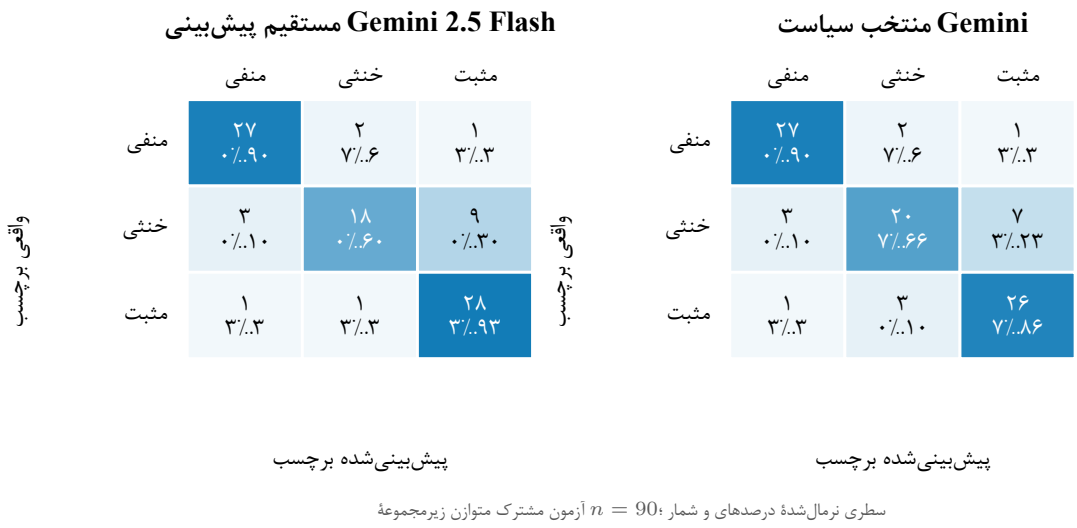
پس از تنظیم سیاست انتخاب، اف‌یک ماکرو Qwen3 8B برابر ۰/۷۲۱۹۰۳ باقی مانده و برای Gemini 2.5 Flash به ۰/۸۰۸۳۴۰ رسیده است. در این زیرمجموعه، انتخاب منبع برای Qwen نسبت به مسیر مستقیم سود خالصی ایجاد نکرده و بهبود Gemini نیز محدود است. ازاین‌رو، نتیجه شکل بیش از آنکه اثر علّی ظرفیت مدل را نشان دهد، بیانگر وابستگی سود مراحل بعدی به انتخاب مدل، پشتانهٔ اجرا و الگوی خطای مشاهده‌شده در همین تنظیم آزمایشی است.



شکل ۵-۳: مقایسهٔ امتیاز اف‌یک ماکرو Qwen3 8B و Gemini 2.5 Flash روی زیرمجموعهٔ مشترک ۹۰ نمونه‌ای آزمون.

۵-۷-۳ ماتریس‌های درهم‌ریختگی Gemini

شکل ۵-۴ تغییر رفتار کلاسی Gemini را پس از انتخاب منبع نشان می‌دهد. دقت کلی هر دو پیکربندی برابر ۰/۸۱۱۱۱۱ است و هر کدام ۷۳ پاسخ درست دارند، اما توزیع خطاها یکسان نیست.



شکل ۵-۴: ماتریس‌های درهم‌ریختگی پیش‌بینی مستقیم و سیاست انتخاب تنظیم‌شده برای Gemini 2.5 Flash روی آزمون مشترک.

در کلاس منفی، هر دو روش ۲۷ نمونهٔ درست دارند. در کلاس خنثی، شمار پاسخ‌های درست از ۱۸ به ۲۰ افزایش یافته و خطای خنثی به مثبت از ۹ به ۷ کاهش یافته است. در مقابل، پاسخ‌های درست کلاس مثبت از ۲۸ به ۲۶ کاهش یافته و خطای مثبت به خنثی از یک به سه رسیده است. این جابه‌جایی توضیح می‌دهد که چرا دقت ثابت مانده، ولی اف‌یک ماکرو اندکی تغییر کرده است. انتخاب منبع تعادل میان کلاس‌های خنثی و مثبت را عوض کرده، نه اینکه تعداد کل پاسخ‌های درست را افزایش دهد.

۵-۸ ملاحظات اجرایی و محدودیت‌ها

مسیر مستقیم برای هر نمونه یک پاسخ کوتاه تولید می‌کند و کم‌هزینه‌ترین پیکربندی است. مسیر THOR SC3 کل زنجیرهٔ استدلال را سه بار اجرا می‌کند و بازبینی تشخیصی نیز برای موارد اختلاف فراخوانی اضافی دارد. بنابراین، افزایش عملکرد سامانهٔ نهایی با هزینهٔ محاسباتی بیشتر همراه است. از آنجا که زمان اجرا در محیط‌های مختلف به سخت‌افزار، بار سرویس و پشتانهٔ مدل وابسته است و آزمایش زمان‌سنجی کنترل‌شده‌ای انجام نشده، در این فصل ادعای عددی دربارهٔ سرعت یا هزینه ارائه نمی‌شود.

نخستین محدودیت تجربی، تمرکز مجموعهٔ اصلی بر داده‌های انگلیسی SCAPT و دو دامنهٔ لپ‌تاپ و رستوران است. تعمیم نتیجه به زبان‌ها، دامنه‌ها یا شیوهٔ برچسب‌گذاری دیگر به ارزیابی مستقل نیاز دارد. دوم، مقایسهٔ Gemini فقط روی ۹۰ نمونهٔ آزمون مشترک انجام شده است و نباید با نتیجهٔ کامل Qwen در یک جمعیت واحد تلقی شود. سوم، مسیر استدلالی غیرقطعی است و خودسازگاری فقط بخشی از تغییرپذیری پاسخ را مهار می‌کند.

محدودیت دیگر به منبع تشخیصی مربوط است. اگرچه بازبین نوع خطا و اطمینان را برای ساخت پروفایل فراهم می‌کند، برچسب تشخیصی در هیچ نمونهٔ آزمون منبع نهایی نبوده است. پس شواهد این آزمایش از نقش اطلاعات تشخیصی در انتخاب میان مستقیم و THOR پشتیبانی می‌کنند، اما برتری مستقل برچسب تشخیصی را نشان نمی‌دهند. همچنین، فاصلهٔ باقی‌مانده با اوراکل نباید به‌عنوان عملکرد دست‌یافتنی قطعی تفسیر شود؛ اوراکل از اطلاعاتی استفاده می‌کند که هنگام پیش‌بینی در دسترس نیست.

۵-۹ جمع‌بندی

این فصل نشان داد که استدلال چندمرحله‌ای به‌تنهایی از پیش‌بینی مستقیم دقیق‌تر نیست. هر دو نسخهٔ THOR و نیز بازبینی ساده روی آزمون کامل Qwen3 8B پایین‌تر از خط پایهٔ مستقیم قرار گرفتند. کنترل‌گر قاعده‌محور بخشی از خطاها را مهار کرد، اما خروجی آن نیز از خط پایه عبور نکرد. این نتایج ضرورت جداسازی تولید استدلال از انتخاب پاسخ نهایی را تأیید می‌کنند.

سامانهٔ نهایی محافظت‌شدهٔ انتخاب منبع، دقت آزمون را از ۰/۶۷۸۷۳۳ به ۰/۷۲۳۹۸۲ و اف‌یک ماکرو را از ۰/۶۷۴۰۷۵ به ۰/۷۱۹۱۱۹ رساند. بهبود در هر دو دامنه و هر سه کلاس مشاهده شد. انتخاب‌گر ۲۶ خطای مستقیم را اصلاح و شش پاسخ درست را تضعیف کرد و در نهایت ۲۰ پاسخ درست بیشتر ساخت. بااین‌حال، ۱۱۶ نمونهٔ مشترک نادرست و فاصلهٔ حدود شش واحد درصدی اف‌یک ماکرو با اوراکل نشان می‌دهند که هم منابع پیش‌بینی و هم سازوکار انتخاب هنوز قابل بهبودند.

مقایسهٔ تکمیلی نشان داد که Gemini 2.5 Flash روی زیرمجموعهٔ مشترک از Qwen3 8B سنجه‌های بالاتری دارد، اما انتخاب منبع در این جمعیت کوچک فقط بهبود محدودی برای Gemini ایجاد کرده و برای Qwen تغییری در اف‌یک ماکرو نداده است. بنابراین، اثر مراحل استدلال و انتخاب به الگوی خطای مدل پایه و جمعیت ارزیابی وابسته است. فصل ششم بر پایهٔ این یافته‌ها نتیجه‌گیری کلی پژوهش، محدودیت‌های اصلی و مسیرهای کار آتی را بیان می‌کند.

فصل ششم

نتیجه‌گیری و کارهای آتی

۶-۱ جمع‌بندی پژوهش

تحلیل احساسات ضمنی با تحلیل متن‌هایی سروکار دارد که نگرش نویسنده را بدون اتکای مستقیم به واژه‌های احساسی آشکار منتقل می‌کنند. در چنین متن‌هایی، قطبیت ممکن است از پیامد یک رویداد، برآورده‌شدن یا نشدن انتظار، مقایسه، توصیه یا رابطهٔ میان هدف و بافت استنباط شود. دشواری مسئله زمانی بیشتر می‌شود که یک جمله چند موجودیت داشته باشد یا گزارشی خنثی در کنار نشانه‌ای احساسی قرار گیرد. ازاین‌رو، سامانهٔ تحلیل باید هم احساس بیان‌نشده را بازیابی کند و هم از نسبت‌دادن نگرش به هدفی نادرست یا تفسیر افراطی یک واقعیت خنثی بپرهیزد.

هدف این پژوهش طراحی و ارزیابی سامانه‌ای چندمنبعی برای تحلیل احساسات ضمنی هدف‌محور بود. سامانهٔ پیشنهادی یک پیش‌بینی مستقیم را در کنار مسیر استدلال چندمرحله‌ای الهام‌گرفته از THOR قرار داد. هنگام اختلاف این دو مسیر، بازبینی ساختاریافته نوع خطا، برچسب پیشنهادی و سطح اطمینان را تولید کرد و کنترل‌گر قاعده‌محور تصمیمی میانی ساخت. در مرحلهٔ نهایی، سیاستی محافظت‌شده که آستانه‌های آن فقط با اعتبارسنجی‌های داخلی بخش آموزش تنظیم شده بود، از میان منابع مستقیم، استدلالی و تشخیصی منبع مناسب را انتخاب کرد. این تفکیک اهمیت دارد؛ زیرا خروجی کنترل‌گر پاسخ نهایی سامانه نیست و انتخاب نهایی نیز به برچسب‌های طلایی آزمون دسترسی ندارد.

ارزیابی اصلی با Qwen3 8B و از طریق Ollama روی آزمون کامل SCAPT انجام شد. نتایج نشان دادند که افزودن مراحل استدلالی به‌تنهایی دقت را تضمین نمی‌کند. هر دو پیکربندی استدلالی بررسی‌شده از پیش‌بینی مستقیم ضعیف‌تر بودند و کنترل‌گر قاعده‌محور نیز به‌تنهایی از خط پایه عبور نکرد. بااین‌حال، منابع مختلف دقیقاً در نمونه‌های یکسان خطا نمی‌کردند. سیاست انتخاب منبع توانست از این مکمل‌بودن استفاده کند و عملکرد نهایی را در هر دو دامنه و هر سه کلاس بهبود دهد.

بنابراین، یافتهٔ محوری پژوهش برتری مستقل یک زنجیرهٔ استدلال یا یک پرامپت خاص نیست. نتیجهٔ اصلی این است که در تحلیل احساسات ضمنی، تولید استدلال و انتخاب پاسخ نهایی باید دو مسئلهٔ جدا در نظر گرفته شوند. استدلال چندمرحله‌ای می‌تواند شواهد و پاسخ‌های مکمل فراهم کند، اما استفاده از آن باید تحت کنترل سازوکاری باشد که از دادهٔ آموزش می‌آموزد هر منبع در چه شرایطی قابل اعتمادتر است.

۶-۲ پاسخ به پرسش‌های پژوهش

بر پایهٔ نتایج فصل پنجم، پاسخ پرسش‌های مطرح‌شده در فصل نخست به شرح زیر است:

۱. **تفاوت پیش‌بینی مستقیم و استدلال چندمرحله‌ای.** پیش‌بینی مستقیم مسیر کوتاه‌تر و کم‌هزینه‌تری فراهم کرد و در ارزیابی اصلی از هر دو نسخهٔ استدلالی دقیق‌تر بود. در مقابل، مسیر THOR تصمیم را به تشخیص جنبه، استنتاج نظر و تعیین قطبیت تقسیم کرد و ردپایی قابل بررسی ساخت. این ردپا برای تحلیل و تولید منبع مکمل مفید بود، اما افزایش تعداد مراحل باعث انتقال خطای جنبه یا نظر به برچسب نهایی نیز شد. پس پاسخ بخش دوم پرسش منفی است: استدلال چندمرحله‌ای به‌تنهایی همواره موجب بهبود نمی‌شود و در تنظیم آزمایشی این پژوهش از پیش‌بینی مستقیم ضعیف‌تر بود.

۲. **نقش بازبینی ساختاریافته و کنترل‌گر.** بازبینی ساختاریافته اختلاف میان دو مسیر را به اطلاعاتی قابل استفاده شامل نوع خطا، پیشنهاد اصلاح و اطمینان تبدیل کرد. این اطلاعات امکان داد تصمیم‌ها بر اساس رده‌های مشخصی مانند احساس ازدست‌رفته، بیش‌تفسیر خنثی و جابه‌جایی دامنهٔ هدف بررسی شوند. بااین‌حال، خروجی کنترل‌گر قاعده‌محور هنوز از خط پایهٔ مستقیم پایین‌تر بود و برچسب تشخیصی نیز در هیچ نمونهٔ آزمون به‌عنوان منبع نهایی انتخاب نشد. بنابراین، بازبینی و کنترل‌گر اختلاف را ردیابی‌پذیرتر و قابل‌کنترل‌تر کردند، اما شواهد موجود برتری مستقل پاسخ تشخیصی یا قواعد ثابت کنترل‌گر را نشان نمی‌دهند. فایدهٔ عملی اطلاعات تشخیصی در این پژوهش، کمک به ساخت پروفایل انتخاب میان منابع مستقیم و استدلالی بود.

۳. **اثر انتخاب منبع تنظیم‌شده با دادهٔ آموزش.** پاسخ این پرسش در مجموعهٔ اصلی مثبت است. سامانهٔ نهایی بدون استفاده از برچسب طلایی آزمون، دقت را از ۰/۶۷۸۷۳۳ به ۰/۷۲۳۹۸۲ و امتیاز اف‌یک ماکرو را از ۰/۶۷۴۰۷۵ به ۰/۷۱۹۱۱۹ رساند. این سیاست ۲۶ خطای پیش‌بینی مستقیم را اصلاح کرد و در مقابل، ۶ پاسخ درست مستقیم را تضعیف کرد. حاصل خالص، ۲۰ پاسخ درست بیشتر بود. در نتیجه، انتخاب وابسته به الگوهای آموخته‌شده از بخش آموزش از اتکای ثابت به مسیر مستقیم یا استدلالی مناسب‌تر بود. این نتیجه به سیاست، داده و مدل بررسی‌شده مقید است و به معنای برتری هر انتخاب‌گر در هر مجموعهٔ دیگری نیست.

۴. **الگوهای اصلی خطا.** تحلیل کیفی سه الگوی برجسته را نشان داد. در حالت نخست، مدل پیامد یا انتظار ضمنی را بازیابی نمی‌کند و احساس مثبت یا منفی را از دست می‌دهد. در حالت دوم،

یک گزارهٔ خبری یا خنثی بیش‌ازحد تفسیر می‌شود و قطبیتی بدون شاهد کافی شکل می‌گیرد. در حالت سوم، نگرش مربوط به یک موجودیت یا وضعیت کلی به هدف تعیین‌شده منتقل می‌شود. مسیر استدلالی می‌تواند الگوی نخست را اصلاح کند، اما در صورت تثبیت نادرست هدف یا جنبه، دو الگوی دیگر را تشدید می‌کند. همچنین، پیش‌بینی مستقیم و سامانهٔ نهایی در ۱۱۶ نمونهٔ آزمون هر دو نادرست بودند؛ ازاین‌رو، بخشی از خطاها با انتخاب بهتر میان منابع موجود رفع نمی‌شود.

۵. **اثر انتخاب مدل و پشتانهٔ اجرا.** این مقایسه روی ۹۰ نمونهٔ مشترک آزمون انجام شد. در آن، Gemini در هر دو مسیر مستقیم و استدلالی امتیاز بالاتری از اجرای متناظر Qwen داشت. بااین‌حال، مسیر THOR برای هر دو مدل از پیش‌بینی مستقیم همان مدل ضعیف‌تر بود. انتخاب منبع برای Gemini بهبودی محدود ایجاد کرد و برای Qwen در این زیرمجموعه تغییری در اف‌یک ماکرو نداد. چون خانواده مدل، زیرساخت اجرا و سایر مشخصات نیز هم‌زمان متفاوت‌اند، این مشاهده فقط رابطهٔ انتخاب مدل و الگوی خطای آن با میزان سود مراحل بعدی را در تنظیم حاضر نشان می‌دهد و اثر علّی ظرفیت مدل را اثبات نمی‌کند. با توجه به اندازهٔ محدود آزمون مشترک، این نتیجه شاهدی تکمیلی است و نباید به‌عنوان رتبه‌بندی عمومی دو مدل یا تضمین رفتار آن‌ها در داده‌های دیگر تفسیر شود.

۶-۳ دستاوردهای اصلی

۶-۳-۱ دستاوردهای علمی

نخستین دستاورد علمی پژوهش، صورت‌بندی استدلال چندمرحله‌ای به‌عنوان منبعی مکمل به جای پاسخ قطعی سامانه است. کیفیت ظاهری ردپای استدلال صحت برچسب را تضمین نمی‌کند؛ ازاین‌رو، تولید تفسیر و انتخاب پاسخ دو وظیفهٔ جدا هستند. دستاورد دوم، تفکیک بازبینی، کنترل میانی و انتخاب نهایی است. بازبین خطای احتمالی را ساختارمند می‌کند، کنترل‌گر تصمیمی قابل ردیابی می‌سازد و انتخاب‌گر با تکیه بر الگوهای آموزش منبع پاسخ را تعیین می‌کند.

دستاورد دیگر، ارزیابی انتخاب منبع بدون نشت اطلاعات آزمون است. پروفایل‌ها و آستانه‌های محافظ فقط با بخش آموزش ساخته و تنظیم و سپس پس از تثبیت روی آزمون اعمال شدند. بهبود حاصل نشان داد که می‌توان مکمل‌بودن خطاهای منابع را بدون دسترسی به پاسخ آزمون به سود عملی تبدیل کرد.

در عین حال، فاصلهٔ سامانهٔ نهایی با اوراکل نشان می‌دهد که تشخیص منبع مناسب هنوز مسئله‌ای باز است.

۶-۳-۲ دستاوردهای فنی و مهندسی

از نظر فنی، پروژه خط لوله‌ای انتهابه‌انتها از خواندن داده تا گزارش نهایی فراهم کرد. تقسیم رسمی آموزش و آزمون حفظ شد و خروجی‌های مستقیم، استدلالی، تشخیصی، کنترل‌گر و نهایی جداگانه ذخیره شدند. پیش از ارزیابی نیز شمار و یکتایی ردیف‌ها، هم‌ترازی شناسه، جمله، هدف و دامنه و اعتبار برچسب‌ها به‌صورت خودکار کنترل شد.

اجرای اصلی با Qwen3 8B و Ollama در یک چارچوب واحد انجام شد. برای مقایسهٔ تکمیلی نیز خروجی دو مدل روی شناسه‌های مشترک سنجیده شد تا تفاوت جمعیت آزمون وارد مقایسه نشود. این اجزا امکان ردیابی خروجی هر مرحله و تکرار ارزیابی در محدودهٔ تنظیمات گزارش‌شده را فراهم می‌کنند.

۶-۴ محدودیت‌های پژوهش

مجموعهٔ اصلی شامل متن‌های انگلیسی و فقط دو دامنهٔ لپ‌تاپ و رستوران است؛ بنابراین، تعمیم نتیجه به زبان فارسی، شبکه‌های اجتماعی یا حوزه‌های تخصصی به آزمایش مستقل نیاز دارد. نتیجهٔ اصلی نیز به اجرای کامل Qwen3 8B تعلق دارد. مقایسهٔ Gemini 2.5 Flash فقط روی ۹۰ نمونهٔ آزمون مشترک انجام شد و شاهدی تکمیلی است.

مسیر استدلال چندمرحله‌ای غیرقطعی است. خودسازگاری بخشی از نوسان برچسب را مهار می‌کند، اما خطاهای مشترک اجراها را از بین نمی‌برد. ردپای استدلال نیز ارزیابی انسانی مستقلی از نظر صحت و وفاداری نداشته است. از سوی دیگر، مراحل چندگانه فراخوانی و متن بیشتری تولید می‌کنند و چون زمان‌سنجی کنترل‌شده انجام نشده، نسبت دقیق بهبود به زمان یا هزینه مشخص نیست.

اطلاعات تشخیصی در ساخت پروفایل‌ها نقش داشتند، اما برچسب پیشنهادی بازبین در هیچ نمونهٔ آزمون منبع نهایی نبود؛ پس کارایی مستقل آن اثبات نشده است. برخی پروفایل‌ها کم‌نمونه‌اند، ولی سیاست اصلی برای پشتیبانی کمتر از ۱۰ یا حاشیهٔ ناکافی به مسیر مستقیم بازمی‌گردد؛ بنابراین، خطر تصمیم‌های مبتنی بر گروه‌های کوچک محدود شده، نه اینکه کم‌نمونه‌بودن آن گروه‌ها حذف شده باشد. فاصلهٔ سامانهٔ نهایی با اوراکل فقط کرانی تحلیلی است و نباید عملکرد تضمین‌شدهٔ انتخاب‌گر آینده تلقی شود.

در نهایت، مسیر THOR برای مدل، داده و زیرساخت پروژه سازگار شده و بازتولید کامل تنظیم رسمی پژوهش اصلی نیست. وزن مدل‌ها ثابت مانده و اثر ریزتنظیم نیز بررسی نشده است. یافته‌های این پایان‌نامه در نتیجه به چارچوب پرامپت‌محور و انتخاب منبع محدود می‌شوند.

۶-۵ کارهای آتی

۶-۵-۱ توسعه و بهبود انتخاب‌گر منبع

می‌توان شمار تقسیم‌های اعتبارسنجی را افزایش داد و پایداری انتخاب پروفایل‌ها و آستانه‌های محافظ را با بازنمونه‌گیری سنجید. هموارسازی الگوهای کم‌نمونه و ادغام پروفایل‌های مشابه نیز قابل بررسی‌اند. فرادسته‌بندهای کنترل‌شده می‌توانند با ویژگی‌های تنظیم‌شده از بخش آموزش و قید پرهیز از اصلاح کم‌اطمینان توسعه یابند. ارزیابی آن‌ها باید علاوه بر میانگین عملکرد، شمار پاسخ‌های درستی را که تضعیف می‌شوند گزارش کند و جدایی آموزش و آزمون را حفظ کند.

۶-۵-۲ بهبود بازبینی و استدلال هدف‌محور

برای کاهش جابه‌جایی دامنهٔ هدف، می‌توان بازهٔ هدف و شواهد مرتبط را در تمام مراحل برجسته کرد و پیش از تعیین قطبیت، ارتباط شاهد با هدف را سنجید. آزمون سازگاری میان جنبه، نظر و برچسب نیز می‌تواند استدلال‌های ناهم‌جهت را آشکار کند. بازبین آینده بهتر است شواهد موافق و مخالف را تفکیک و نبود شاهد کافی را معتبر تلقی کند. اطمینان تشخیصی نیز باید با دادهٔ اعتبارسنجی کالیبره و اثر تغییرها با کاهش خطاهای هدف‌محور سنجیده شود.

۶-۵-۳ تحلیل خطاهای مشترک و گسترش منابع

تحلیل دقیق‌تر ۱۱۶ نمونهٔ مشترک نادرست می‌تواند سهم دانش عمومی، ابهام برچسب، ساختار نحوی و چندهدف‌بودن جمله را روشن کند. برای این موارد باید منابعی با شیوهٔ استنتاج یا دانش متفاوت، مانند پرامپت هدف‌محور مستقل یا بازیابی دانش زمینه‌ای، بررسی شوند. افزودن هر منبع باید با مطالعهٔ حذف مؤلفه همراه باشد تا اثر تنوع واقعی از افزایش صرف تعداد فراخوانی‌ها جدا شود.

۶-۵-۴ گسترش ارزیابی به داده‌ها و مدل‌های دیگر

ارزیابی روی مجموعه‌داده‌ها، دامنه‌ها و زبان‌های دیگر، به‌ویژه متن فارسی، می‌تواند پایداری سیاست انتخاب را روشن کند. مقایسهٔ مدل‌ها نیز باید روی چند زیرمجموعهٔ مستقل یا آزمون کامل انجام شود. در این بررسی، علاوه بر امتیاز نهایی، تغییر الگوهای خطا، مکمل‌بودن منابع و نیاز انتخاب‌گر به تنظیم دوباره با دادهٔ آموزش اهمیت دارد.

۶-۵-۵ کاهش هزینهٔ اجرا

مراحل پرهزینه می‌توانند فقط هنگام نامطمئن‌بودن پیش‌بینی مستقیم یا وجود نشانهٔ ابهام هدف فعال شوند. چنین دروازه‌ای باید کاهش فراخوانی را همراه با تغییر عملکرد گزارش کند. اندازه‌گیری کنترل‌شدهٔ زمان، شمار توکن، حافظه و هزینهٔ سرویس نیز امکان مقایسهٔ پیکربندی‌ها برای کاربردهای برخط، دسته‌ای و کم‌منبع را فراهم می‌کند.

۶-۵-۶ ارزیابی انسانی و ریزتنظیم

ردپای استدلال باید از نظر صحت شواهد، ارتباط با هدف، سازگاری با قطبیت و وفاداری ارزیابی انسانی شود. در صورت فراهم‌شدن دادهٔ بزرگ‌تر و برچسب‌های معتبر برای جنبه، نظر، شواهد و نوع خطا، ریزتنظیم پارامترکم نیز قابل بررسی است. این مسیر باید با خط پایهٔ پرامپت‌محور و تحت تقسیم یکسان مقایسه شود؛ زیرا آموزش با برچسب‌های خودتولیدشده ممکن است خطای مدل مولد را بازتولید کند.

۶-۶ نتیجه‌گیری نهایی

این پایان‌نامه نشان داد که افزودن مراحل بیشتر به یک پرامپت، به‌تنهایی تحلیل احساسات ضمنی را بهبود نمی‌دهد. پیش‌بینی مستقیم مرجعی قوی بود و استدلال چندمرحله‌ای، با وجود ایجاد ردپای قابل تحلیل، خطاهای تازه‌ای نیز وارد کرد. ارزش آن زمانی آشکار شد که به‌عنوان منبعی مکمل به کار رفت و انتخاب پاسخ از تولید استدلال جدا شد.

دستاورد اصلی پژوهش چارچوبی برای مدیریت اختلاف منابع است: خطای احتمالی ساختارمند می‌شود و پاسخ نهایی با سیاستی جداگانه و آموخته‌شده از آموزش انتخاب می‌گردد. بهبود مشاهده‌شده به داده و تنظیم حاضر مقید است، اما نشان می‌دهد توجه هم‌زمان به شواهد، محدودهٔ هدف و اعتمادپذیری

نسبی منابع مسیر مناسبی برای ادامهٔ پژوهش است.

منابع و مراجع

[1] Fei, Hao, Li, Bobo, Liu, Qian, Bing, Lidong, Li, Fei, and Chua, Tat-Seng. Reasoning implicit sentiment with chain-of-thought prompting. in Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pp. 1171–1183. Association for Computational Linguistics, 2023.

[2] Duan, Zhihua and Wang, Jialin. Implicit sentiment analysis based on chain-of-thought prompting. arXiv preprint arXiv:2408.12157, 2024.

[3] Pontiki, Maria, Galanis, Dimitris, Pavlopoulos, John, Papageorgiou, Harris, Androutsopoulos, Ion, and Manandhar, Suresh. Semeval-2014 task 4: Aspect based sentiment analysis. in Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pp. 27–35. Association for Computational Linguistics, 2014.

[4] Wei, Jason, Wang, Xuezhi, Schuurmans, Dale, Bosma, Maarten, Ichter, Brian, Xia, Fei, Chi, Ed H., Le, Quoc V., and Zhou, Denny. Chain-of-thought prompting elicits reasoning in large language models. in Advances in Neural Information Processing Systems, vol. 35, pp. 24824–24837, 2022.

[5] Wang, Xuezhi, Wei, Jason, Schuurmans, Dale, Le, Quoc V., Chi, Ed H., Narang, Sharan, Chowdhery, Aakanksha, and Zhou, Denny. Self-consistency improves chain of thought reasoning in language models. in International Conference on Learning Representations, 2023.

[6] Pang, Bo and Lee, Lillian. Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval, 2(1–2):1–135, 2008.

[7] Wankhade, Mayur, Rao, Annavarapu Chandra Sekhara, and Kulkarni, Chaitali. A survey on sentiment analysis methods, applications, and challenges. Artificial Intelligence Review, 55(7):5731–5780, 2022.

[8] Wang, Kai, Shen, Weizhou, Yang, Yunyi, Quan, Xiaojun, and Wang, Rui. Relational graph attention network for aspect-based sentiment analysis. in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 3229–3238. Association for Computational Linguistics, 2020.

[9] Li, Zhengyan, Zou, Yicheng, Zhang, Chong, Zhang, Qi, and Wei, Zhongyu. Learning implicit sentiment in aspect-based sentiment analysis with supervised contrastive pre-training. in Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp. 246–256. Association for Computational Linguistics, 2021.

[10] Wang, Siyin, Zhou, Jie, Sun, Changzhi, Ye, Junjie, Gui, Tao, Zhang, Qi, and Huang, Xuanjing. Causal intervention improves implicit sentiment analysis. in Proceedings of the 29th International Conference on Computational Linguistics, pp. 6966–6977. International Committee on Computational Linguistics, 2022.

[11] Liu, Pengfei, Yuan, Weizhe, Fu, Jinlan, Jiang, Zhengbao, Hayashi, Hiroaki, and Neubig, Graham. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. ACM Computing Surveys, 55(9), 2023.

پیوست: جزئیات بازتولید آزمایش‌ها

این پیوست اطلاعات لازم برای بازسازی ارزیابی‌های گزارش‌شده را خلاصه می‌کند. خروجی‌های خام و میانی هر مرحله به‌صورت فایل‌های جداگانه ذخیره شده‌اند؛ بنابراین، بازتولید جدول‌ها و شکل‌های نهایی به اجرای دوبارهٔ مدل زبانی نیاز ندارد.

داده و تقسیم‌بندی

دادهٔ اصلی از نمونهٔ برچسب‌خورده و ضمنی SCAPT مبتنی بر SemEval 2014 به دست آمده است. پس از پاک‌سازی، ۲۱۸۸ نمونه شامل ۱۷۴۶ نمونهٔ آموزش و ۴۴۲ نمونهٔ آزمون باقی مانده است. تقسیم رسمی آموزش و آزمون در تمام مراحل حفظ می‌شود. برای تطبیق پایدار ردیف‌ها، کلید مرکب شامل id، شناسه جملهٔ منبع، متن جمله، هدف، بازهٔ هدف و قطبیت به کار می‌رود.

مدل‌ها و تنظیمات تولید

اجرای اصلی از مدل qwen3:8b از طریق Ollama استفاده می‌کند. پیش‌بینی مستقیم با دمای صفر تولید می‌شود. در مسیر THOR، مراحل استخراج جنبه، استخراج نظر ضمنی و تولید استدلال قطبیت با دمای ۰/۷ اجرا می‌شوند؛ مرحلهٔ تبدیل استدلال به برچسب نهایی دمای صفر دارد. برای خودسازگاری سه‌اجرایی، سه مسیر استدلالی برای هر نمونه تولید و برچسب‌ها با رأی اکثریت تجمیع می‌شوند. مقایسهٔ تکمیلی Gemini 2.5 Flash فقط روی زیرمجموعهٔ متوازن ۲۴۰ نمونه‌ای انجام شده است.

انتخاب منبع

سیاست اصلی، انتخاب‌گر پروفایلی محافظت‌شده و تنظیم‌شده با اعتبارسنجی داخلی بخش آموزش است. ویژگی‌های پروفایل شامل پیش‌بینی مستقیم، پیش‌بینی THOR، نوع خطای تشخیصی، سطح اطمینان و دامنه است. واگذاری به منبع غیرمستقیم فقط در صورت پشتیبانی و حاشیهٔ کافی انجام می‌شود؛ در غیر این صورت، پیش‌بینی مستقیم قاعدهٔ بازگشت است. هیچ برچسب طلایی آزمون در ساخت پروفایل‌ها، تنظیم آستانه‌ها یا انتخاب منبع نهایی استفاده نمی‌شود.

فایل‌های اصلی خروجی

- `results/direct_isa_predictions.csv`
- `results/thor_originalish_sc3_isa_predictions.csv`
- `results/final_results_table.csv`
- `results/tfidf_logreg_predictions.csv`

وجود واژهٔ originalish در نام داخلی بعضی فایل‌های خروجی صرفاً برای سازگاری با اجراهای ذخیره‌شده است و نام علمی روش محسوب نمی‌شود.

فرمان‌های بازتولید گزارش

فرمان‌های زیر در ریشهٔ مخزن و در محیط ویندوز اجرا می‌شوند:

```
.\.venv\Scripts\python.exe -B `
  experiments\run_final_pipeline.py
.\.venv\Scripts\python.exe -B `
  experiments\generate_thesis_result_figures.py
.\.venv\Scripts\python.exe -m unittest discover `
  -s tests -p "test_*.py" -v
powershell.exe -NoProfile -ExecutionPolicy Bypass `
```

```
-File .\scripts\build_thesis.ps1
```

فرمان run_final_pipeline.py پیش‌بینی‌های ذخیره‌شده را اعتبارسنجی و جدول نهایی را بازسازی می‌کند و مدل زبانی را دوباره اجرا نمی‌کند. فرمان generate_thesis_result_figures.py نیز شکل‌ها را از همان خروجی‌های ذخیره‌شده می‌سازد. اجرای دوبارهٔ استنتاج مستلزم راه‌اندازی پشتانهٔ مدل و اجرای جداگانهٔ اسکریپت‌های مراحل مستقیم، استدلالی و تشخیصی است و برای بازسازی اعداد این پایان‌نامه ضروری نیست.

واژه‌نامهٔ فارسی به انگلیسی

ارزیابی . . . . . Evaluation

استدلال چندمرحله‌ای . . . . . Multi-Stage Reasoning

انتخاب منبع . . . . . Source Selection

بازبینی تشخیصی . . . . . Diagnostic Review

بازخوانی . . . . . Recall

پرامپت . . . . . Prompt

پیش‌بینی مستقیم . . . . . Direct Prediction

تحلیل احساسات ضمنی Implicit Sentiment Analysis

تحلیل احساسات مبتنی بر جنبه . . . Aspect-Based Sentiment Analysis

خودسازگاری . . . . . Self-Consistency

دقت . . . . . Accuracy

زنجیرهٔ تفکر . . . . . Chain of Thought

صحت پیش‌بینی . . . . . Precision

مدل زبانی بزرگ . . Large Language Model

ماتریس درهم‌ریختگی . . Confusion Matrix

اف‌یک ماکرو Macro-F1

واژه‌نامهٔ انگلیسی به فارسی

مدل زبانی بزرگ . . Large Language Model

اف‌یک ماکرو Macro-F1

استدلال چندمرحله‌ای . . . . . Multi-Stage Reasoning

صحت پیش‌بینی . . . . . Precision

پرامپت . . . . . Prompt

بازخوانی . . . . . Recall

خودسازگاری . . . . . Self-Consistency

انتخاب منبع . . . . . Source Selection

دقت . . . . . Accuracy

تحلیل احساسات مبتنی بر جنبه . . . Aspect-Based Sentiment Analysis

زنجیرهٔ تفکر . . . . . Chain of Thought

ماتریس درهم‌ریختگی . . Confusion Matrix

بازبینی تشخیصی . . . . . Diagnostic Review

پیش‌بینی مستقیم . . . . . Direct Prediction

ارزیابی . . . . . Evaluation

تحلیل احساسات ضمنی Implicit Sentiment Analysis

Abstract

Detecting sentiment in texts without explicit evaluative words requires contextual understanding, commonsense knowledge, and inference about the relationship between a described event and its target. This thesis presents a prompt-based, multi-source system for implicit sentiment analysis. The system combines a language model's direct prediction with multi-stage reasoning inspired by THOR, structured error diagnosis, and a source-selection policy. The final selection policy is tuned using the training data and, without access to gold test labels, selects one of the direct, reasoning, or diagnostic sources. The main evaluation uses the SCAPT-labelled data derived from SemEval 2014 in the laptop and restaurant domains. The main dataset contains 2,188 examples, with 1,746 training examples and 442 test examples. The primary experiments use Qwen3 8B through Ollama. On the test set, the final system increases accuracy from 0.678733 to 0.723982 and Macro-F1 from 0.674075 to 0.719119. A supplementary comparison with Gemini 2.5 Flash is also conducted on a balanced subset of 240 examples. For the model, data, and experimental setting examined in this thesis, the results indicate that controlled source selection is more effective than relying on one fixed prediction path. **Key Words:**

Implicit Sentiment Analysis, Large Language Models, Chain-of-Thought, Multi-Stage Reasoning, Source Selection