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 مقدمه ۱۲
۳-۲ تحول رویکردهای تحلیل احساسات ۱۲
۳-۲-۱ روش‌های مبتنی بر واژگان و یادگیری ماشین ۱۲
۳-۲-۲ روش‌های مبتنی بر شبکه‌های عصبی و مدل‌های زبانی ۱۳
۳-۳ پژوهش‌های تحلیل احساسات مبتنی بر جنبه و ضمنی ۱۴
۳-۳-۱ مجموعه‌داده‌ها و چارچوب‌های ارزیابی ۱۴
۳-۳-۲ روش‌های پیش از رویکردهای مولد ۱۴
۳-۳-۳ چالش‌های مشترک گزارش‌شده ۱۵
۳-۴ مهندسی پرامپت ۱۶

۳-۵ زنجیرهٔ تفکر در پژوهش‌های پردازش زبان طبیعی ۱۶
۳-۶ روش THOR برای تحلیل احساسات ضمنی ۱۷
۳-۶-۱ تجزیهٔ استدلال به جنبه، نظر و قطبیت ۱۷
۳-۶-۲ خودسازگاری و بازبینی در THOR ۱۸
۳-۶-۳ نقاط قوت و محدودیت‌های THOR ۱۸
۳-۷ بازبینی و خوداصلاحی در تحلیل احساسات ضمنی ۱۹
۳-۸ خودسازگاری در استدلال چندمرحله‌ای ۱۹
۳-۹ کنترل‌گرها و انتخاب منبع پیش‌بینی ۲۰
۳-۱۰ مقایسهٔ پژوهش‌های مرتبط ۲۱
۳-۱۱ شکاف تحقیقاتی و جایگاه پژوهش حاضر ۲۱
۳-۱۲ جمع‌بندی ۲۲

۴ کار پیشنهادی ۲۴
۴-۱ مقدمه ۲۵
۴-۲ صورت‌بندی مسئله ۲۵
۴-۲-۱ ورودی، خروجی و نمادگذاری ۲۵
۴-۲-۲ اصول طراحی ۲۶
۴-۳ نمای کلی سامانهٔ پیشنهادی ۲۷
۴-۴ آماده‌سازی داده ۲۸
۴-۴-۱ منبع داده و واحد تحلیل ۲۸
۴-۴-۲ استخراج و پاک‌سازی نمونه‌ها ۲۸
۴-۵ خط پایهٔ کلاسیک ۲۹
۴-۵-۱ بازنمایی TF-IDF و رگرسیون لجستیک ۲۹
۴-۶ تولید منابع اولیهٔ پیش‌بینی ۳۰
۴-۶-۱ پیش‌بینی مستقیم ۳۰
۴-۶-۲ استدلال چندمرحله‌ای مبتنی بر THOR ۳۰
۴-۶-۳ خودسازگاری سه‌اجرایی ۳۲
۴-۷ بازبینی آگاه از نوع خطا ۳۲

۴-۷-۱ شرط فعال‌سازی بازبینی ۳۲
۴-۷-۲ رده‌بندی خطاها ۳۳
۴-۷-۳ قالب خروجی و ترمیم پاسخ ۳۳
۴-۸ کنترل‌گر تصمیم‌گیری میانی ۳۴
۴-۹ انتخاب منبع محافظت‌شده و تنظیم‌شده با دادهٔ آموزش ۳۵
۴-۹-۱ تعریف پروفایل انتخاب ۳۵
۴-۹-۲ یادگیری سیاست از بخش آموزش ۳۵
۴-۹-۳ تنظیم آستانه‌های محافظ با اعتبارسنجی داخلی ۳۶
۴-۹-۴ تولید پیش‌بینی نهایی ۳۶
۴-۱۰ انتخاب‌گرهای یادگرفتنی جایگزین ۳۷
۴-۱۰-۱ انتخاب‌گر فراسطح مبتنی بر درخت ۳۷
۴-۱۰-۲ رتبه‌بند لجستیک منبع ۳۷
۴-۱۱ جزئیات پیاده‌سازی و بازتولیدپذیری ۳۸
۴-۱۱-۱ معماری نرم‌افزار ۳۸
۴-۱۱-۲ مدل‌ها و رابط اجرا ۳۸
۴-۱۱-۳ ذخیره، ادامهٔ اجرا و اعتبارسنجی ۳۹
۴-۱۲ جمع‌بندی ۳۹

۵ نتایج آزمایش‌ها ۴۱
۵-۱ مقدمه ۴۲
۵-۲ تنظیمات آزمایش ۴۲
۵-۲-۱ مجموعه‌داده و تقسیم ارزیابی ۴۲
۵-۲-۲ مدل‌ها و زیرساخت اجرا ۴۳
۵-۲-۳ پروتکل ارزیابی و جلوگیری از نشت داده ۴۳
۵-۳ روش‌های مورد مقایسه و معیارهای ارزیابی ۴۴
۵-۳-۱ روش‌های مورد مقایسه ۴۴
۵-۳-۲ معیارهای ارزیابی ۴۵
۵-۴ نتایج روش‌های اصلی روی آزمون کامل ۴۵

۵-۴-۱ مقایسهٔ کلی روش‌ها ۴۵
۵-۴-۲ تحلیل ماتریس‌های درهم‌ریختگی ۴۷
۵-۴-۳ تحلیل عملکرد بر اساس دامنه ۴۸
۵-۵ تحلیل انتخاب‌گر و مؤلفه‌های سامانه ۴۹
۵-۵-۱ توزیع منابع و سود یا زیان نسبت به مسیر مستقیم ۴۹
۵-۵-۲ مقایسهٔ تصمیم میانی، انتخاب‌گرهای فراسطح و کران اوراکل ۵۰
۵-۵-۳ اعتبارسنجی زنجیرهٔ نهایی ۵۰
۵-۶ تحلیل کیفی و خطا ۵۱
۵-۶-۱ نمونه‌های اصلاح‌شده ۵۱
۵-۶-۲ نمونه‌های تضعیف‌شده ۵۱
۵-۶-۳ الگوهای خطای باقی‌مانده ۵۲
۵-۷ مقایسهٔ تکمیلی با Gemini 2.5 Flash ۵۳
۵-۷-۱ ساخت زیرمجموعهٔ مشترک ۵۳
۵-۷-۲ مقایسهٔ سنجه‌های دو مدل ۵۳
۵-۷-۳ ماتریس‌های درهم‌ریختگی Gemini ۵۴
۵-۸ ملاحظات اجرایی و محدودیت‌ها ۵۵
۵-۹ جمع‌بندی ۵۶

۶ نتیجه‌گیری و کارهای آتی ۵۷
۶-۱ جمع‌بندی پژوهش ۵۸
۶-۲ پاسخ به پرسش‌های پژوهش ۵۹
۶-۳ دستاوردهای اصلی ۶۰
۶-۳-۱ دستاوردهای علمی ۶۰
۶-۳-۲ دستاوردهای فنی و مهندسی ۶۱
۶-۴ محدودیت‌های پژوهش ۶۱
۶-۵ کارهای آتی ۶۲
۶-۵-۱ توسعه و بهبود انتخاب‌گر منبع ۶۲
۶-۵-۲ بهبود بازبینی و استدلال هدف‌محور ۶۲

۶-۵-۳ تحلیل خطاهای مشترک و گسترش منابع ۶۲
۶-۵-۴ گسترش ارزیابی به داده‌ها و مدل‌های دیگر ۶۳
۶-۵-۵ کاهش هزینهٔ اجرا ۶۳
۶-۵-۶ ارزیابی انسانی و ریزتنظیم ۶۳
۶-۶ نتیجه‌گیری نهایی ۶۳

منابع و مراجع ۶۵

پیوست: جزئیات بازتولید آزمایش‌ها ۶۷

واژه‌نامهٔ فارسی به انگلیسی ۷۰

واژه‌نامهٔ انگلیسی به فارسی ۷۱

فهرست اشکال

شکل صفحه

۴-۱ معماری کلی سامانهٔ پیشنهادی. ۲۷

۵-۱ مقایسهٔ دقت کلی و امتیاز اف‌یک ماکرو هشت روش اصلی روی ۴۴۲ نمونهٔ آزمون. . . ۴۷

۵-۲ ماتریس‌های درهم‌ریختگی پیش‌بینی مستقیم و سامانهٔ نهایی Qwen3 8B روی آزمون کامل. ۴۸

۵-۳ مقایسهٔ امتیاز اف‌یک ماکرو Qwen3 8B و Gemini 2.5 Flash روی زیرمجموعهٔ مشترک ۹۰ نمونه‌ای آزمون. ۵۴

۵-۴ ماتریس‌های درهم‌ریختگی پیش‌بینی مستقیم و سیاست انتخاب تنظیم‌شده برای Gemini 2.5 Flash روی آزمون مشترک. ۵۴

فهرست جداول

جدول صفحه

۳-۱ مقایسهٔ پژوهش‌های محوری مرتبط با تحلیل احساسات ضمنی. ۲۱

۴-۱ انواع خروجی بازبینی تشخیصی و نقش آن‌ها در تصمیم‌گیری. ۳۳

۴-۲ مؤلفه‌های اصلی پیاده‌سازی سامانهٔ پیشنهادی. ۳۸

۵-۱ توزیع نمونه‌های ضمنی در دو دامنه و دو بخش داده. ۴۳

۵-۲ نتایج روش‌های اصلی، از خط پایهٔ کلاسیک تا سامانهٔ نهایی، روی مجموعهٔ کامل آزمون. ۴۵

۵-۳ امتیاز اف‌یک کلاس‌ها برای پیش‌بینی مستقیم و سامانهٔ نهایی Qwen3 8B. ۴۸

۵-۴ عملکرد پیش‌بینی مستقیم و سامانهٔ نهایی به تفکیک دامنه در بخش آزمون. ۴۸

۵-۵ توزیع منبع منتخب در ۴۴۲ نمونهٔ آزمون. ۴۹

۵-۶ مقایسهٔ سیاست‌های انتخاب منبع روی ۴۴۲ نمونهٔ آزمون Qwen3 8B. ۵۰

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

۳-۱ مقدمه

فصل دوم مفاهیم و معیارهای لازم برای تحلیل احساسات ضمنی را تعریف کرد. در این فصل، مسیر تحول پژوهش‌ها و نسبت آن‌ها با مسئلهٔ پایان‌نامه بررسی می‌شود. هدف، تکرار تعریف‌ها نیست؛ بلکه باید روشن شود هر جریان پژوهشی کدام بخش مسئله را حل کرده و چه محدودیتی باقی گذاشته است.

مرور از روش‌های عمومی تحلیل احساسات آغاز می‌شود. سپس پژوهش‌های مبتنی بر جنبه و مجموعه‌داده‌های احساس ضمنی بررسی می‌شوند.

در ادامه، مهندسی پرامپت، زنجیرهٔ تفکر، خودسازگاری، روش THOR و چارچوب تحلیل احساسات مبتنی بر تفکر SAoT[۱] تحلیل می‌شوند. پایان فصل نیز به مقایسهٔ روش‌ها و استخراج شکاف پژوهشی اختصاص دارد.

مقایسهٔ عددی نتایج مقاله‌ها باید با احتیاط انجام شود. پژوهش‌ها از مدل‌های پایه، نسخهٔ داده، شیوهٔ نمونه‌گیری و تنظیمات متفاوت استفاده کرده‌اند. بنابراین، اعداد گزارش‌شده در یک مقاله لزوماً با اعداد مقالهٔ دیگر قابل مقایسهٔ مستقیم نیستند. تمرکز این فصل بر ایدهٔ روش، نوع اطلاعات مصرف‌شده و سازوکار تولید تصمیم است.

۳-۲ تحول رویکردهای تحلیل احساسات

۳-۲-۱ روش‌های مبتنی بر واژگان و یادگیری ماشین

پژوهش‌های نخستین تحلیل احساسات بیشتر بر واژه‌های دارای بار عاطفی تکیه داشتند. در این روش‌ها، متن با استفاده از فرهنگ واژگان، بسامد واژه‌ها یا ویژگی‌های سطحی نمایش داده می‌شد. سپس یک قاعده یا طبقه‌بند آماری قطبیت را تعیین می‌کرد [۶، ۷].

این رویکردها ساده، سریع و تا حدی تفسیرپذیر بودند. بااین‌حال، معنای واژه به بافت و هدف وابسته است. منفی‌سازها، تضاد، کنایه و چندمعنایی نیز رابطهٔ ثابت میان واژه و قطبیت را مختل می‌کنند. این مشکل در تحلیل مبتنی بر جنبه جدی‌تر است، زیرا یک جمله ممکن است دربارهٔ دو هدف، دو نگرش متفاوت بیان کند.

یادگیری ماشین نظارت‌شده بخشی از وابستگی به قواعد دستی را کاهش داد. ویژگی‌هایی مانند چندواژه، نقش دستوری و فاصلهٔ واژه تا هدف به طبقه‌بند داده می‌شدند. بااین‌حال، کیفیت مدل همچنان

[۱] Sentiment Analysis of Thinking (SAoT)

به مهندسی ویژگی و شباهت میان داده‌های آموزش و آزمون وابسته بود. چنین بازنمایی‌هایی معمولاً رابطهٔ دوربرد میان هدف و قرینهٔ احساسی را ناقص ثبت می‌کردند.

این محدودیت در احساس ضمنی بنیادی‌تر است. ممکن است هیچ واژهٔ مثبت یا منفی آشکاری در جمله وجود نداشته باشد. در این حالت، افزایش اندازهٔ فرهنگ واژگان به تنهایی مسئله را حل نمی‌کند. مدل باید پیامد رویداد، انتظار متعارف یا رابطهٔ آن با هدف را نیز استنتاج کند.

۳-۲-۲ روش‌های مبتنی بر شبکه‌های عصبی و مدل‌های زبانی

شبکه‌های عصبی امکان یادگیری بازنمایی متن را بدون تعریف دستی همهٔ ویژگی‌ها فراهم کردند. مدل‌های بازگشتی، پیچشی و سازوکار توجه برای تمرکز بر بخش‌های مرتبط با هدف به کار رفتند. در تحلیل مبتنی بر جنبه، نمایش هدف معمولاً همراه با بازنمایی جمله ساخته می‌شد تا قطبیت نسبت به همان هدف تعیین شود.

مدل‌های گرافی گام دیگری در این مسیر بودند. برای نمونه، شبکهٔ توجه گراف رابطه‌ای[۲] از رابطهٔ نحوی میان واژه‌ها برای پیوند هدف و نشانه‌های نظر استفاده می‌کند [۸]. این روش‌ها وابستگی‌های ساختاری را بهتر از ویژگی‌های سطحی نمایش می‌دهند. بااین‌حال، نمودار نحوی لزوماً دانشی را که در متن بیان نشده است فراهم نمی‌کند.

مدل‌های زبانی ازپیش‌آموزش‌دیده، دانش زبانی گسترده‌تری را وارد مسئله کردند. سازگارسازی این مدل‌ها برای داده‌های برچسب‌خورده، عملکرد تحلیل مبتنی بر جنبه را بهبود داد. پژوهش‌های بعدی نیز اهداف پیش‌آموزش ویژه‌ای برای حساس‌کردن بازنمایی به هدف و قطبیت طراحی کردند [۹].

مدل‌های زبانی بزرگ، افزون بر بازنمایی، قابلیت تولید پاسخ و توضیح را در اختیار می‌گذارند. مسئله دیگر فقط نگاشت یک بردار به برچسب نیست. پژوهشگر می‌تواند صورت مسئله، فضای برچسب و مراحل تحلیل را با زبان طبیعی بیان کند. این تحول، مسیر استفاده از پرامپت و استدلال زنجیره‌ای را در تحلیل احساسات گشود.

[۲] Relational Graph Attention Network

۳-۳ پژوهش‌های تحلیل احساسات مبتنی بر جنبه و ضمنی

۳-۳-۱ مجموعه‌داده‌ها و چارچوب‌های ارزیابی

رقابت SemEval 2014 یکی از مبناهای مهم تحلیل احساسات مبتنی بر جنبه است [۳]. این رقابت داده‌هایی از دامنه‌های لپ‌تاپ و رستوران ارائه کرد. وظایف آن شامل شناسایی جنبه و تعیین قطبیت نسبت به هدف بود. ساختار هدف‌محور این داده‌ها امکان بررسی جمله‌هایی را فراهم کرد که چند موضوع یا چند نگرش دارند.

بااین‌حال، داده‌های اصلی صرفاً برای مطالعهٔ احساس ضمنی طراحی نشده بودند. لی و همکاران نمونه‌های صریح و ضمنی را از یکدیگر تفکیک کردند. آن‌ها همچنین پیش‌آموزش تقابلی نظارت‌شده[۳] را برای یادگیری احساس ضمنی در تحلیل مبتنی بر جنبه پیشنهاد کردند [۹]. روش پیشنهادی آن‌ها با نام SCAPT شناخته می‌شود. نسخهٔ تفکیک‌شدهٔ داده نیز در پژوهش‌های بعدی بازاستفاده شده است.

اهمیت این تفکیک در نوع شواهد موجود است. نمونه‌های صریح معمولاً نشانه‌ای نزدیک و آشکار برای نظر دارند. در مقابل، نمونهٔ ضمنی مدل را وادار می‌کند از توصیف رویداد یا ویژگی به نگرش برسد. به همین دلیل، عملکرد مناسب روی کل دادهٔ مبتنی بر جنبه الزاماً به معنای توانایی مناسب روی زیرمجموعهٔ ضمنی نیست.

دو دامنهٔ لپ‌تاپ و رستوران نیز الگوهای استنتاجی متفاوتی دارند. در رستوران، سرعت خدمت، زمان انتظار و کیفیت غذا با انتظارات روزمره پیوند دارند. در لپ‌تاپ، عمر باتری، وزن، قیمت و سازگاری نرم‌افزاری اهمیت بیشتری دارند. یک روش موفق باید هم ساختار مشترک مسئله و هم تفاوت دانش دامنه‌ای را در نظر بگیرد.

۳-۳-۲ روش‌های پیش از رویکردهای مولد

روش SCAPT با ساخت جفت‌های تقابلی، مدل را به تمایز نمونه‌های دارای قطبیت متفاوت هدایت می‌کند [۹]. این جهت‌گیری از پیش‌آموزش عمومی هدفمندتر است. مدل می‌آموزد بازنمایی‌هایی بسازد که برای تشخیص احساس نسبت به جنبه مفید باشند. مزیت این روش آن است که دانش مسئله در پارامترها و فضای بازنمایی تثبیت می‌شود.

در کنار آن، روش‌های مبتنی بر ساختار نحوی تلاش کردند مسیر میان هدف و عبارت نظر را مدل

[۳] Supervised Contrastive Pre-training

کنند [۸]. مداخلهٔ علّی نیز برای کاهش اتکای مدل به هم‌بستگی‌های سطحی پیشنهاد شد [۱۰]. این پژوهش نشان می‌دهد بخشی از خطاهای احساس ضمنی از میان‌برهای آماری ناشی می‌شوند. مدل ممکن است به یک واژه یا الگوی پرتکرار تکیه کند، بدون آنکه رابطهٔ واقعی شواهد و قطبیت را دریابد.

این خانواده از روش‌ها معمولاً به دادهٔ برچسب‌خورده و آموزش ویژه نیاز دارد. خروجی آن‌ها نیز اغلب یک برچسب نهایی است. در نتیجه، تشخیص اینکه مدل کدام جنبه یا پیامد ضمنی را مبنای تصمیم قرار داده دشوار می‌شود. همچنین، انتقال به دامنه‌ای تازه می‌تواند به سازگارسازی مجدد نیاز داشته باشد.

ورود مدل‌های مولد، پرسش پژوهش را تغییر داد. اکنون می‌توان از مدل خواست رابطهٔ پنهان میان هدف، ویژگی ضمنی و نگرش را به زبان طبیعی آشکار کند. این قابلیت جای روش‌های بازنمایی‌محور را به طور کامل نمی‌گیرد، اما ابزار تازه‌ای برای صورت‌بندی و بررسی فرایند استنتاج فراهم می‌کند.

۳-۳-۳ چالش‌های مشترک گزارش‌شده

نخستین چالش، نبود قرینهٔ احساسی صریح است. مدل باید بفهمد یک رویداد برای هدف مطلوب، نامطلوب یا بی‌اثر است. این داوری به دانش عمومی و انتظارهای دامنه‌ای وابسته است. یک رویداد واحد نیز ممکن است برای دو هدف پیامد متفاوتی داشته باشد.

چالش دوم، نسبت‌دادن احساس به هدف درست است. جمله ممکن است چند موجودیت یا چند ویژگی داشته باشد. اگر مدل شواهد مربوط به یک هدف را به هدف دیگر منتقل کند، قطبیت نهایی نادرست می‌شود. روش‌های نحوی این مسئله را کاهش می‌دهند، اما استنتاج ضمنی همیشه با رابطهٔ نحوی مستقیم هم‌راستا نیست.

چالش سوم، تمایز خنثی از احساس ضعیف و ضمنی است. نبود واژهٔ ارزیابانه می‌تواند مدل را به پیش‌بینی خنثی سوق دهد. از سوی دیگر، جست‌وجوی افراطی برای معنای پنهان ممکن است جملهٔ واقعاً خنثی را احساسی جلوه دهد. بنابراین، روش باید هم از حذف شواهد ضمنی و هم از تفسیر بیش‌ازحد جلوگیری کند.

چالش چهارم، نبود تضمین برای وفاداری توضیح است. یک مدل مولد می‌تواند پس از انتخاب برچسب، توضیحی ظاهراً قانع‌کننده بسازد. پس وجود متن استدلال به تنهایی اثبات نمی‌کند که تصمیم از همان مسیر حاصل شده است. ارزیابی برچسب، سازگاری مراحل و منشأ تصمیم باید از یکدیگر تفکیک شوند.

۳-۴ مهندسی پرامپت

مرور روش‌های پرامپت نشان می‌دهد که صورت‌بندی وظیفه، نمونه‌های راهنما و قالب پاسخ بر رفتار مدل اثر می‌گذارند [۱۱]. در تحلیل احساسات، یک پرامپت حداقلی فقط جمله، هدف و برچسب‌های مجاز را مشخص می‌کند. پرامپت غنی‌تر می‌تواند تعریف قطبیت، قید هدف‌محوری یا مراحل تحلیل را نیز اضافه کند.

پرامپت مستقیم یک خط پایهٔ ضروری است. این روش نشان می‌دهد مدل بدون الزام به تولید توضیح تا چه اندازه مسئله را حل می‌کند. اگر روش استدلالی با خط پایهٔ مستقیم مقایسه نشود، نمی‌توان اثر واقعی مراحل افزوده را سنجید. همچنین، هر افزایش عملکرد را نمی‌توان صرفاً به طولانی‌ترشدن پرامپت نسبت داد.

قالب خروجی نیز اهمیت دارد. پاسخ آزاد ممکن است حاوی چند برچسب، توضیح مبهم یا عبارتی خارج از فضای برچسب باشد. قالب ساختاریافته، استخراج خودکار و تشخیص پاسخ نامعتبر را آسان‌تر می‌کند. بااین‌حال، سخت‌گیری قالب نباید به اندازه‌ای باشد که ظرفیت تحلیل مدل را محدود کند.

پرامپت صفرنمونه هزینه آماده‌سازی مثال را کاهش می‌دهد، اما به تفسیر مدل از دستور وابسته است. پرامپت چندنمونه می‌تواند مرز میان برچسب‌ها را روشن کند. بااین‌حال، گزینش مثال‌ها ممکن است سوگیری ایجاد کند. در هر دو حالت، متن دستور، ترتیب اطلاعات و نحوهٔ نمایش هدف باید ثابت و مستند باشند.

در احساس ضمنی، پرامپت خوب باید مدل را به شواهد مرتبط با هدف هدایت کند. دستور کلی «احساس را تشخیص بده» ممکن است نگرش کلی جمله را جایگزین نگرش هدف‌محور کند. ازاین‌رو، پژوهش‌های جدیدتر وظیفه را به پرسش‌های کوچک‌تر تقسیم کرده‌اند.

۳-۵ زنجیرهٔ تفکر در پژوهش‌های پردازش زبان طبیعی

وی و همکاران نشان دادند که افزودن نمونه‌های دارای مراحل میانی می‌تواند توانایی استدلال مدل‌های زبانی بزرگ را آشکار کند [۴]. ایدهٔ اصلی آن‌ها تولید چند گام زبان طبیعی پیش از پاسخ نهایی بود. این روش ابتدا در مسئله‌های حسابی، نمادین و استدلال عقل سلیم بررسی شد.

اهمیت این پژوهش برای تحلیل احساسات ضمنی در شباهت ساختاری وظایف است. در هر دو حالت، پاسخ نهایی از مشاهدهٔ یک نشانهٔ صریح به دست نمی‌آید. مدل باید اطلاعات متن را با دانش

زمینه‌ای ترکیب و رابطهٔ چندمرحله‌ای را دنبال کند. بااین‌حال، مراحل مناسب برای هر وظیفه باید بر اساس ساختار همان مسئله طراحی شوند.

زنجیرهٔ تفکر را نباید با صحت تضمین‌شده یکی دانست. مراحل تولیدی می‌توانند حاوی فرض نادرست، جابه‌جایی هدف یا تناقض باشند. طولانی‌ترشدن استدلال نیز الزاماً کیفیت را افزایش نمی‌دهد. هر مرحلهٔ اضافی یک فرصت تازه برای اصلاح، اما هم‌زمان یک نقطهٔ تازه برای خطا ایجاد می‌کند.

در ادبیات تحلیل احساسات ضمنی، این ایده از یک دستور عمومی به تجزیهٔ وظیفه تبدیل شد. به جای درخواست مبهم برای «فکرکردن مرحله‌به‌مرحله»، پژوهش‌ها پرسش‌هایی دربارهٔ جنبه، نظر و رابطهٔ آن‌ها با قطبیت مطرح کردند. روش THOR نمونهٔ شاخص این صورت‌بندی ساختاریافته است.

۳-۶ روش THOR برای تحلیل احساسات ضمنی

فی و همکاران روش THOR را برای استدلال دربارهٔ احساس ضمنی ارائه کردند [۱]. این روش مسئله را به سه پرسش وابسته تقسیم می‌کند. هر پاسخ، زمینهٔ لازم برای مرحلهٔ بعد را فراهم می‌سازد. هدف آن است که مدل پیش از انتخاب قطبیت، اجزای پنهان رابطهٔ احساسی را آشکار کند.

۳-۶-۱ تجزیهٔ استدلال به جنبه، نظر و قطبیت

مرحلهٔ نخست، جنبه یا ویژگی ضمنی مرتبط با هدف را استنتاج می‌کند. ممکن است جمله نام هدف را داشته باشد، اما ویژگی ارزیابی‌شده را آشکار بیان نکند. تشخیص این ویژگی فضای تفسیر را محدود می‌کند. برای مثال، توصیف مدت کارکرد دستگاه می‌تواند به جنبه‌ای مانند باتری مربوط باشد.

مرحلهٔ دوم، نظر ضمنی نسبت به جنبه را استخراج می‌کند. این نظر لزوماً یک صفت موجود در جمله نیست. مدل باید پیامد ویژگی یا رویداد را با انتظارات رایج مقایسه کند. پاسخ مرحلهٔ دوم به متن، هدف و جنبه‌ای که در مرحلهٔ نخست ساخته شده وابسته است.

مرحلهٔ سوم، قطبیت را از رابطهٔ میان هدف، جنبه و نظر تعیین می‌کند. این ترتیب، جهش مستقیم از جمله به برچسب را به چند تصمیم قابل مشاهده تبدیل می‌کند. در نتیجه، پژوهشگر می‌تواند محل احتمالی انحراف را دقیق‌تر بررسی کند.

وابستگی مراحل هم مزیت و هم محدودیت است. اگر جنبه درست تشخیص داده شود، مرحلهٔ نظر راهنمای متمرکزتری دریافت می‌کند. اما خطای مرحلهٔ نخست می‌تواند به مراحل بعد منتقل شود. بنابراین، صحت هر گام و سازگاری میان گام‌ها اهمیت دارد.

۳-۶-۲ خودسازگاری و بازبینی در THOR

روش THOR برای کاهش ناپایداری تولید از نمونه‌گیری چندباره و رأی‌گیری استفاده می‌کند. ایدهٔ خودسازگاری آن است که مسیرهای استدلال متنوع تولید شوند و پاسخ پرتکرار انتخاب شود [۵]. این سازوکار وابستگی نتیجه به یک اجرای تصادفی را کاهش می‌دهد.

در صورت وجود دادهٔ نظارت‌شده، مقالهٔ THOR سازوکاری برای بازبینی استدلال نیز مطرح می‌کند. نمونهٔ خطادار می‌تواند همراه با مسیر اصلاح‌شده به مدل ارائه شود. این بخش نشان می‌دهد که تولید استدلال و اصلاح استدلال دو مسئلهٔ جدا هستند. بااین‌حال، دسترسی به نمونهٔ اصلاح‌شده یا نظارت مناسب همیشه ممکن نیست.

خودسازگاری نیز تضمین‌کنندهٔ پاسخ درست نیست. اگر یک سوگیری در بیشتر مسیرها تکرار شود، رأی اکثریت همان خطا را تثبیت می‌کند. افزون بر این، چند اجرای یک روش ممکن است تنوع معنایی کافی نداشته باشند. بنابراین، توافق باید به عنوان نشانهٔ ثبات تفسیر شود، نه اثبات صحت.

۳-۶-۳ نقاط قوت و محدودیت‌های THOR

نقطهٔ قوت اصلی THOR، انطباق مراحل استدلال با ساختار تحلیل احساسات مبتنی بر جنبه است. روش به طور مشخص می‌پرسد چه چیزی درباره هدف ارزیابی شده و نظر پنهان چیست. این طراحی از یک دستور عمومی دقیق‌تر است و ردپایی برای تحلیل خطا فراهم می‌کند.

نقطهٔ قوت دوم، امکان تفکیک خطاها است. خطا می‌تواند از تشخیص جنبه، استنتاج نظر یا نگاشت نظر به قطبیت ناشی شود. این تفکیک برای طراحی بازبین و تحلیل کیفی مفید است. همچنین، پاسخ استدلالی می‌تواند در کنار پیش‌بینی مستقیم به عنوان منبعی مکمل نگهداری شود.

محدودیت نخست، انتشار خطا در زنجیره است. هر مرحلهٔ بعدی به پاسخ‌های پیشین وابسته است. محدودیت دوم، هزینهٔ بیشتر اجرای چندمرحله‌ای و چندباره است. محدودیت سوم، امکان تولید توضیحی روان اما ناسازگار با شواهد جمله است.

افزون بر این، مقالهٔ اصلی عمدتاً کیفیت یک مسیر استدلالی را ارزیابی می‌کند. اختلاف میان پاسخ مستقیم و پاسخ استدلالی خود به مسئله‌ای مستقل تبدیل می‌شود. اگر یکی از دو منبع در گروهی از نمونه‌ها بهتر باشد، انتخاب ثابت یک منبع می‌تواند بخشی از اطلاعات مفید را از بین ببرد.

۳-۷ بازبینی و خوداصلاحی در تحلیل احساسات ضمنی

بازبینی از مدل می‌خواهد پاسخ اولیه را دوباره بررسی کند. این کار می‌تواند با پرسشی کلی یا مجموعه‌ای از معیارهای مشخص انجام شود. بازبینی کلی انعطاف‌پذیر است، اما خروجی آن ممکن است مبهم باشد. بازبینی ساختاریافته، نوع ناسازگاری و پیشنهاد اصلاح را جدا می‌کند.

دوان و وانگ روش SAoT را برای تحلیل احساسات ضمنی پیشنهاد کردند [۲]. این روش ابتدا جنبه‌ها و نظرهای ضمنی را تحلیل می‌کند. سپس مدل تحلیل خود را بازنگری می‌کند و در پایان قطبیت را نتیجه می‌گیرد. در این طراحی، بازاندیشی بخشی صریح از زنجیرهٔ تصمیم است.

تفاوت مهم SAoT و THOR در جایگاه بازبینی است. THOR بر تجزیهٔ گام‌به‌گام و خودسازگاری تأکید دارد. SAoT یک مرحلهٔ بازاندیشی را پیش از نتیجه‌گیری نهایی وارد می‌کند. هر دو روش می‌کوشند استنتاج پنهان را آشکار کنند، اما کنترل خطا را به شکل متفاوتی صورت‌بندی می‌کنند.

بازبینی می‌تواند خطای اولیه را اصلاح کند، ولی ممکن است پاسخ درست را نیز تغییر دهد. مدل بازبین همان محدودیت‌های دانش و سوگیری مدل تولیدکننده را دارد. اگر دستور بازبینی صرفاً مدل را به تغییر پاسخ ترغیب کند، اصلاح ظاهری جای بررسی شواهد را می‌گیرد. بنابراین، فعال‌سازی بازبینی و پذیرش پیشنهاد آن باید از هم جدا باشند.

پژوهش‌های یادشده نشان می‌دهند که «دوباره فکرکردن» به تنهایی معیار تصمیم نیست. خروجی بازبینی باید قابل استخراج و قابل مقایسه با منابع دیگر باشد. همچنین، لازم است موارد توافق، اختلاف و نبود شواهد کافی از یکدیگر تفکیک شوند.

۳-۸ خودسازگاری در استدلال چندمرحله‌ای

وانگ و همکاران خودسازگاری را برای بهبود استدلال زنجیره‌ای معرفی کردند [۵]. به جای انتخاب نخستین مسیر، چند مسیر با نمونه‌گیری تصادفی تولید می‌شوند. سپس پاسخ نهایی با تجمیع پاسخ‌ها تعیین می‌شود. این روش از تنوع مسیرهای ممکن برای کاهش خطای یک اجرای منفرد استفاده می‌کند.

خودسازگاری یک روش استدلال مستقل نیست. کیفیت آن به پرامپت، مدل پایه و تنوع نمونه‌گیری وابسته است. اگر همهٔ اجراها از یک برداشت نادرست آغاز شوند، رأی‌گیری کمکی نمی‌کند. اگر پاسخ‌ها بسیار پراکنده باشند، رأی اکثریت نیز ممکن است شکننده باشد.

در تحلیل احساسات ضمنی، دو سطح تجمیع قابل تصور است. در سطح نخست، برچسب نهایی هر

اجرای کامل رأی می‌دهد. در سطح دوم، پاسخ هر مرحلهٔ استدلال جداگانه تجمیع می‌شود. سطح دوم می‌تواند خطای میانی را زودتر مهار کند، اما به نرمال‌سازی پاسخ‌های زبانی نیاز دارد.

رأی‌گیری همچنین اطلاعاتی فراتر از برچسب فراهم می‌کند. نسبت آرا می‌تواند نشانهٔ ثبات باشد و مسیرهای مخالف می‌توانند برای تحلیل اختلاف نگهداری شوند. بااین‌حال، این نشانه نباید بدون کالیبراسیون به احتمال صحت تعبیر شود.

۳-۹ کنترل‌گرها و انتخاب منبع پیش‌بینی

بیشتر روش‌های مرورشده یک مسیر اصلی را برای تولید پاسخ نهایی تعیین می‌کنند. روش بازنمایی‌محور از طبقه‌بند، روش مستقیم از پاسخ کوتاه و روش زنجیره‌ای از نتیجهٔ استدلال استفاده می‌کند. با وجود این، هیچ مسیر واحدی در همهٔ نمونه‌ها برتری تضمین‌شده ندارد.

پیش‌بینی مستقیم معمولاً کوتاه‌تر و کم‌هزینه‌تر است. این مسیر گاهی از دانش درونی مدل به شکل مؤثری استفاده می‌کند، بدون آنکه در مراحل میانی منحرف شود. مسیر استدلالی در نمونه‌هایی مفید است که رابطهٔ رویداد و قطبیت به توضیح چندمرحله‌ای نیاز دارد. بازبینی نیز ممکن است ناسازگاری مشخصی را آشکار کند.

این تفاوت، مسئلهٔ انتخاب منبع را مطرح می‌کند. کنترل‌گر باید شواهدی مانند توافق، اعتبار قالب و نشانه‌های خطا را بررسی کند. سپس یک سیاست انتخاب تعیین می‌کند کدام منبع پاسخ نهایی را تأمین کند. این نگاه با انتخاب همیشگی پاسخ طولانی‌تر یا آخرین پاسخ متفاوت است.

ادبیات THOR و SAoT اهمیت مراحل میانی و بازبینی را نشان می‌دهد [۱، ۲]. بااین‌حال، این پژوهش‌ها انتخاب داده‌محور میان پیش‌بینی مستقیم، استدلالی و تشخیصی را به عنوان مسئلهٔ محوری صورت‌بندی نمی‌کنند. ازاین‌رو، مدیریت اختلاف منابع همچنان جای بررسی دارد.

سیاست انتخاب باید بدون استفاده از برچسب طلایی آزمون ساخته شود. در غیر این صورت، انتخاب بهترین منبع پس از مشاهدهٔ پاسخ درست به برآوردی غیرواقعی منجر می‌شود. همچنین، رفتار سیاست در وضعیت‌های دیده‌نشده باید روشن باشد. جزئیات سیاست این پایان‌نامه در فصل چهارم ارائه می‌شود.

۳-۱۰ مقایسهٔ پژوهش‌های مرتبط

جدول ۳-۱ پژوهش‌های محوری را از نظر هدف، سازوکار و محدودیت مرتبط مقایسه می‌کند. منظور از محدودیت، نفی دستاورد پژوهش نیست. این ستون نشان می‌دهد برای ساخت یک سامانهٔ چندمنبعی چه مسئله‌ای همچنان باقی می‌ماند.

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

پژوهش	ایده و سازوکار	محدودیت مرتبط
لی و همکاران؛ SCAPT [۹]	پیش‌آموزش تقابلی نظارت‌شده برای یادگیری بازنمایی حساس به احساس ضمنی	نیازمند آموزش ویژه و فاقد ردپای زبانی برای استنتاج
وانگ و همکاران [۱۰]	مداخلهٔ علّی برای کاهش اتکا به هم‌بستگی‌های سطحی	تمرکز بر بازنمایی یک مدل، نه انتخاب میان منابع
وی و همکاران؛ زنجیرهٔ تفکر [۴]	تولید مراحل زبانی میانی پیش از پاسخ نهایی	مراحل عمومی و غیرتخصصی برای احساس ضمنی
وانگ و همکاران؛ خودسازگاری [۵]	نمونه‌گیری چند مسیر و انتخاب پاسخ پرتکرار	امکان تثبیت خطای مشترک با رأی اکثریت
فی و همکاران؛ THOR [۱]	استنتاج مرحله‌ای جنبه، نظر و قطبیت همراه با خودسازگاری	انتشار خطا میان مراحل و نبود انتخاب تطبیقی منبع
دوان و وانگ؛ SAoT [۲]	تحلیل جنبه و نظر، بازاندیشی و سپس استنتاج قطبیت	امکان تغییر پاسخ درست و نبود سیاست مستقل برای پذیرش اصلاح

روند کلی جدول از یادگیری بازنمایی به آشکارسازی استدلال حرکت می‌کند. SCAPT و روش علّی بر بهبود مدل آموزش‌دیده تمرکز دارند. زنجیرهٔ تفکر و THOR ساخت تصمیم را به زبان طبیعی قابل مشاهده می‌کنند. خودسازگاری و SAoT نیز به ترتیب ثبات و بازبینی را وارد فرایند می‌کنند.

بااین‌حال، هیچ‌یک از این مؤلفه‌ها به تنهایی همهٔ نیازها را پوشش نمی‌دهد. بازنمایی قوی ممکن است توضیحی نداشته باشد. توضیح چندمرحله‌ای ممکن است منحرف شود. رأی اکثریت ممکن است خطای مشترک را تقویت کند و بازبینی ممکن است پاسخ درست را تغییر دهد. این ناهمگونی، انگیزهٔ ترکیب کنترل‌شدهٔ منابع را فراهم می‌کند.

۳-۱۱ شکاف تحقیقاتی و جایگاه پژوهش حاضر

مرور ادبیات چهار شکاف مرتبط را نشان می‌دهد. نخست، بسیاری از روش‌ها یک مسیر را پاسخ نهایی و قطعی فرض می‌کنند. درحالی‌که پیش‌بینی مستقیم و استدلالی می‌توانند در نمونه‌های متفاوت برتری داشته باشند. اختلاف آن‌ها صرفاً مزاحمت نیست و می‌تواند اطلاعات تشخیصی فراهم کند.

دوم، بازبینی اغلب به شکل یک مرحلهٔ تولیدی دیگر تعریف می‌شود. تولید پیشنهاد اصلاح با پذیرش آن یکسان نیست. سامانه باید تشخیص دهد بازبینی چه نوع خطایی را گزارش کرده و تا چه اندازه قابل اتکاست. بدون این جداسازی، خوداصلاحی می‌تواند به خودتخریبی تبدیل شود.

سوم، خودسازگاری عمدتاً مسیرهای یک روش را تجمیع می‌کند. این سازوکار برای کاهش نوسان مفید است، اما تفاوت میان منابع ناهمگون را مدل نمی‌کند. پاسخ مستقیم، استدلال زنجیره‌ای و بازبینی تشخیصی ماهیت یکسانی ندارند. رأی برابر میان آن‌ها نیز لزوماً بهترین سیاست نیست.

چهارم، انتخاب پس از مشاهدهٔ برچسب آزمون فاقد اعتبار تجربی است. لازم است سیاست انتخاب فقط با داده‌های مجاز ساخته شود و برای وضعیت ناشناخته رفتار مشخصی داشته باشد. همچنین، منشأ هر پاسخ نهایی باید قابل ردیابی بماند تا سهم واقعی هر منبع ارزیابی شود.

پژوهش حاضر در همین نقطه قرار می‌گیرد. هدف، ابداع یک مدل زبانی یا روش بنیادی تازه برای زنجیرهٔ تفکر نیست. این پایان‌نامه پیش‌بینی مستقیم، استدلال چندمرحله‌ای و بازبینی تشخیصی را منابع مکمل در نظر می‌گیرد. سپس مسئلهٔ اصلی را به انتخاب کنترل‌شدهٔ منبع نهایی تبدیل می‌کند.

این جایگاه با THOR و SAoT پیوند دارد، اما با آن‌ها یکسان نیست. از THOR ساختار استدلال جنبه، نظر و قطبیت الهام گرفته می‌شود. ایدهٔ بازبینی نیز برای تشخیص ناسازگاری به کار می‌رود. تفاوت اصلی در جداسازی تصمیم میانی از منبع نهایی و ارزیابی سیاست انتخاب است.

جزئیات این سیاست در این فصل ارائه نمی‌شود. فصل چهارم معماری، مسیرهای پیش‌بینی، بازبین و سازوکار انتخاب را به طور دقیق شرح می‌دهد. فصل پنجم نیز نشان می‌دهد هر مؤلفه و هر منبع در داده‌های آزمایشی چه رفتاری داشته است.

۳-۱۲ جمع‌بندی

این فصل سیر پژوهش را از ویژگی‌های سطحی به بازنمایی‌های آموخته‌شده و سپس استدلال مولد مرور کرد. روش‌های واژه‌محور و آماری برای نشانه‌های آشکار مناسب‌تر بودند. مدل‌های عصبی، گرافی و پیش‌آموزش تقابلی رابطهٔ هدف و متن را بهتر مدل کردند، اما همچنان به آموزش ویژه وابسته بودند.

زنجیرهٔ تفکر امکان آشکارسازی مراحل استنتاج را فراهم کرد. THOR این ایده را با تجزیهٔ جنبه، نظر و قطبیت برای احساس ضمنی تخصصی کرد. خودسازگاری نوسان اجرا را کاهش داد و SAoT بازاندیشی را وارد زنجیره کرد. در عین حال، انتشار خطا، توضیح غیرهمسو و اصلاح نادرست همچنان باقی ماندند.

نتیجهٔ اصلی مرور آن است که هیچ منبعی همواره برتر نیست. مسئلهٔ پژوهش حاضر فقط تولید یک

استدلال دیگر نیست؛ بلکه باید اختلاف منابع را مدیریت و منشأ پاسخ نهایی را روشن کند. فصل بعد بر پایهٔ این شکاف، کار پیشنهادی و اجزای اجرایی سامانه را تشریح می‌کند.

فصل چهارم

کار پیشنهادی

۴-۱ مقدمه

مرور فصل پیش نشان داد که استدلال چندمرحله‌ای می‌تواند شواهد پنهان احساس را آشکار کند، اما همیشه از پیش‌بینی مستقیم دقیق‌تر نیست. مسیر استدلال ممکن است یک گزارهٔ خنثی را بیش از اندازه تفسیر کند یا خطای یک مرحلهٔ میانی را به برچسب نهایی انتقال دهد. از سوی دیگر، پیش‌بینی مستقیم نیز در نمونه‌هایی که به دانش عمومی و استنتاج پیامد نیاز دارند، بخشی از اطلاعات لازم را بازیابی نمی‌کند. بنابراین، تکیهٔ ثابت بر یکی از این دو مسیر برای تمام نمونه‌ها مناسب نیست.

کار پیشنهادی این پایان‌نامه، مسئله را به‌صورت تولید چند منبع پیش‌بینی و انتخاب کنترل‌شدهٔ منبع نهایی صورت‌بندی می‌کند. سه منبع اصلی عبارت‌اند از پیش‌بینی مستقیم، پیش‌بینی استدلالی مبتنی بر THOR و برچسب پیشنهادیِ بازبینی تشخیصی. ابتدا دو مسیر مستقیم و استدلالی به‌طور مستقل اجرا می‌شوند. در صورت اختلاف، یک بازبین ساختاریافته نوع خطا، برچسب پیشنهادی و سطح اطمینان را تولید می‌کند. سپس یک کنترل‌گر قاعده‌محور تصمیمی میانی می‌سازد و در مرحلهٔ نهایی، انتخاب‌گری محافظت‌شده که فقط با بخش آموزش تنظیم شده است، یکی از سه منبع را برمی‌گزیند.

این فصل نحوهٔ پیاده‌سازی این ایده را شرح می‌دهد. ابتدا مسئله و نمادگذاری آن مشخص می‌شود. سپس معماری کلی، آماده‌سازی داده، تولید منابع پیش‌بینی، بازبینی خطا و دو سطح تصمیم‌گیری توضیح داده می‌شوند. در پایان نیز جزئیات نرم‌افزاری و سازوکارهای بازتولیدپذیری ارائه می‌شوند. نتایج کمی و مقایسهٔ روش‌ها به فصل پنجم واگذار شده‌اند.

۴-۲ صورت‌بندی مسئله

۴-۲-۱ ورودی، خروجی و نمادگذاری

هر نمونهٔ ورودی با زوج (s_i, t_i) نمایش داده می‌شود که در آن، s_i جملهٔ ورودی و t_i هدف مشخص‌شده در جمله است. دامنه با d_i و بخش داده با z_i نشان داده می‌شود. دامنه یکی از دو مقدار لپ‌تاپ یا رستوران و بخش داده یکی از دو مقدار آموزش یا آزمون است. برچسب طلایی نمونهٔ i نیز با y_i نمایش داده می‌شود.

فضای برچسب‌ها شامل سه قطبیت مثبت، منفی و خنثی است:

$$\mathcal{Y} = \{\text{positive}, \text{negative}, \text{neutral}\}. \tag{۱-۴}$$

هدف سامانه، تولید برچسب نهایی $\hat{y}_i^* \in \mathcal{Y}$ برای هر زوج جمله و هدف است. محدودکردن خروجی به این فضای سه‌عضوی، ارزیابی یکسان همهٔ مسیرها را ممکن می‌کند. اگر پاسخ خام مدل با هیچ‌یک از برچسب‌های مجاز تطبیق نداشته باشد، در مرحلهٔ پردازش با مقدار داخلی unknown علامت‌گذاری می‌شود. این مقدار یک کلاس مسئله نیست و فقط برای آشکارکردن خروجی نامعتبر به کار می‌رود.

تمرکز این پایان‌نامه بر احساس هدف‌محور ضمنی است. ازاین‌رو، سامانه نباید صرفاً قطبیت کلی جمله را تعیین کند. شواهد باید نسبت به هدف t_i تفسیر شوند و نبود واژهٔ احساسی آشکار نیز نباید به‌طور خودکار به برچسب خنثی بینجامد. برای نمونه، توصیف انتظار طولانی ممکن است بدون صفت احساسی، نگرشی منفی نسبت به خدمت مورد نظر منتقل کند.

۴-۲-۲ اصول طراحی

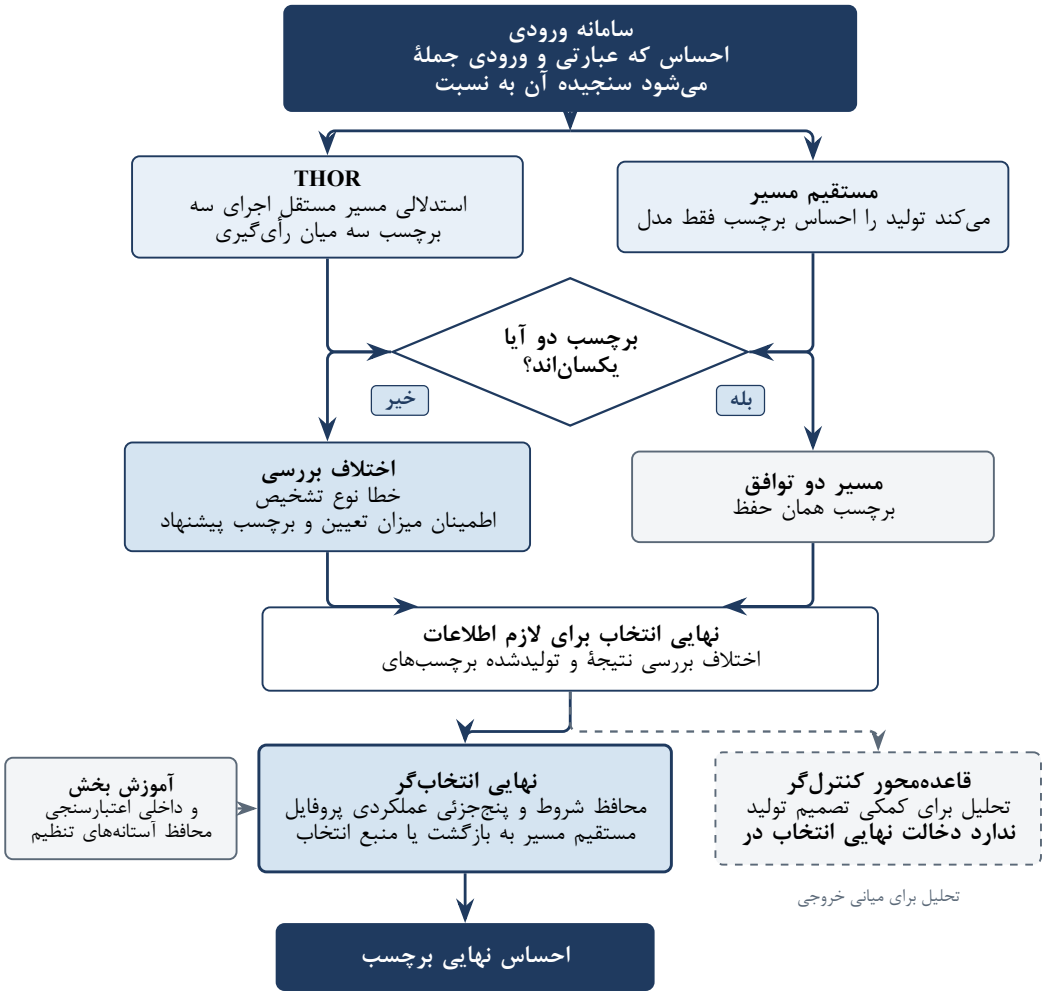
معماری پیشنهادی بر چهار اصل استوار است. اصل نخست، حفظ یک خط پایهٔ ساده و قابل تفسیر است. پیش‌بینی مستقیم مشخص می‌کند مدل بدون زنجیرهٔ میانی چه برچسبی تولید می‌کند. این مسیر هم مبنای مقایسه است و هم در موقعیت‌های نامطمئن، گزینهٔ بازگشت محافظه‌کارانه را فراهم می‌کند.

اصل دوم، جداسازی استدلال از تصمیم نهایی است. مسیر THOR شواهد را در چند گام آشکار می‌کند، اما خروجی آن به‌طور خودکار پاسخ نهایی سامانه محسوب نمی‌شود. اصل سوم، فعال‌سازی بازبینی بر پایهٔ اختلاف است. هنگامی که پیش‌بینی مستقیم و استدلالی یکسان‌اند، تولید نقد تازه معمولاً اطلاعات ضروری ایجاد نمی‌کند. بازبینی برای نمونه‌هایی اجرا می‌شود که دو مسیر دربارهٔ آن‌ها اختلاف دارند.

اصل چهارم، جلوگیری از استفاده از برچسب طلایی آزمون در انتخاب منبع است. سیاست انتخاب فقط با نمونه‌های بخش آموزش و برچسب‌های همان بخش ساخته می‌شود. پس از تثبیت سیاست، همان قاعده بدون تغییر بر بخش آزمون اعمال می‌شود. این جداسازی مانع نشت داده و برآورد خوش‌بینانهٔ عملکرد می‌شود.

۴-۳ نمای کلی سامانهٔ پیشنهادی

شکل ۴-۱ جریان اطلاعات را در سامانه نشان می‌دهد. ورودی مشترک به دو شاخهٔ مستقیم و استدلالی فرستاده می‌شود. پس از مقایسهٔ دو برچسب، در حالت توافق برچسب مشترک حفظ می‌شود و در حالت اختلاف، بازبینی آگاه از نوع خطا فعال می‌گردد. سیگنال‌های حاصل در اختیار کنترل‌گر میانی و انتخاب‌گر نهایی قرار می‌گیرند. دادهٔ آموزش نیز از مسیری جداگانه برای تنظیم پروفایل‌های عملکردی و آستانه‌های محافظ وارد انتخاب‌گر می‌شود.



شکل ۴-۱: معماری کلی سامانهٔ پیشنهادی.

دو سطح تصمیم‌گیری شکل، نقش یکسانی ندارند. کنترل‌گر قاعده‌محور سازگاری خروجی‌ها و قابل اعتماد بودن اصلاح را بررسی می‌کند و یک تصمیم میانی قابل ردیابی می‌سازد. انتخاب‌گر نهایی، خروجی کنترل‌گر را به‌عنوان منبع چهارم در نظر نمی‌گیرد؛ بلکه با استفاده از الگوی آموخته‌شده در بخش آموزش،

مستقیماً یکی از منابع مستقیم، THOR یا تشخیصی را انتخاب می‌کند. این تفکیک مانع آن می‌شود که یک تصمیم میانی با برچسب نهایی سامانه اشتباه گرفته شود.

۴-۴ آماده‌سازی داده

۴-۴-۱ منبع داده و واحد تحلیل

داده‌های مورد استفاده از نسخهٔ برچسب‌خورده‌ای می‌آیند که در پژوهش SCAPT از داده‌های SemEval 2014 ساخته شده است [۳، ۹]. این داده دو دامنهٔ لپ‌تاپ و رستوران را پوشش می‌دهد. واحد تحلیل، یک هدف در یک جمله است؛ بنابراین، اگر جمله چند هدف داشته باشد، هر هدف می‌تواند یک نمونهٔ مستقل ایجاد کند. این تعریف با ماهیت هدف‌محور مسئله سازگار است.

برای هر نمونه، شناسه، دامنه، بخش داده، متن جمله، عبارت هدف، جایگاه هدف، قطبیت و نشانگر صریح یا ضمنی استخراج می‌شوند. در صورت وجود، اطلاعات واژهٔ نظر نیز نگه‌داری می‌شود. این فیلدها امکان هم‌ترازی خروجی مسیرهای مختلف و تحلیل خطا در مراحل بعد را فراهم می‌کنند.

۴-۴-۲ استخراج و پاک‌سازی نمونه‌ها

فایل‌های مبدأ در قالب XML خوانده می‌شوند. هر مدخل هدف به یک سطر جدولی تبدیل می‌شود و فاصلهٔ ابتدا و انتهای هدف برای کنترل هم‌ترازی حفظ می‌شود. پیش از آزمایش، نبود مقدار مفقود در ستون‌های ضروری شامل شناسه، شناسه جملهٔ مبدأ، دامنه، بخش داده، متن جمله، هدف، بازهٔ هدف، قطبیت و نشانگر ضمنی‌بودن بررسی شد و در نسخهٔ مورداستفاده هیچ ردیف ناقصی مشاهده نشد. سپس فقط قطبیت‌های مثبت، منفی و خنثی و نمونه‌های دارای نشانگر معتبر صریح یا ضمنی نگه داشته شدند.

پس از تبدیل، نمونه‌هایی انتخاب می‌شوند که در برچسب‌گذاری SCAPT ضمنی شناخته شده‌اند. این فیلتر برای هر دو دامنه و هر دو بخش داده یکسان اعمال می‌شود. تقسیم اصلی آموزش و آزمون نیز بدون جابه‌جایی حفظ می‌شود. در نتیجهٔ این فرایند، مجموعهٔ اصلی شامل ۲۱۸۸ نمونه است که ۱۷۴۶ نمونه در بخش آموزش و ۴۴۲ نمونه در بخش آزمون قرار دارند.

پاک‌سازی داده فقط به آماده‌سازی متن محدود نیست. ستون ID به‌تنهایی شناسه‌ای یکتا برای هر ردیف نیست. برای اتصال پایدار خروجی‌ها از کلید مرکب شامل id، source_sentence_id، متن جمله، هدف، ابتدا و انتهای بازهٔ هدف و قطبیت استفاده می‌شود. یکتایی این کلید مرکب، شمار سطرها، بخش

داده و برچسب طلایی پیش از ادغام فایل‌های پیش‌بینی کنترل می‌شوند. این کنترل از جابه‌جایی خاموش سطرها میان مسیرهای مستقیم، استدلالی و تشخیصی جلوگیری می‌کند و بدون تغییر یا بازشماره‌گذاری شناسه‌های اصلی انجام می‌شود.

۴-۵ خط پایهٔ کلاسیک

۴-۵-۱ بازنمایی TF-IDF و رگرسیون لجستیک

در نخستین گام آزمایش‌های این پژوهش، پیش از به‌کارگیری مدل‌های زبانی، یک خط پایهٔ کلاسیک و کم‌هزینه ساخته شد تا سطح مرجعی برای دشواری مسئله فراهم شود. این خط پایه جزئی از معماری سامانهٔ پیشنهادی نیست و فقط یک روش مقایسه‌ای مستقل است. برای هر نمونه، عبارت هدف و متن جمله پس از یکسان‌سازی فاصله‌ها با جداکنندهٔ [SEP] به‌صورت target [SEP] sentence به هم متصل شدند.

از متن حاصل، دو دستهٔ ویژگی استخراج شد: ویژگی‌های واژه‌ای TF-IDF شامل تک‌واژه‌ها و دوواژه‌ها، و ویژگی‌های نویسه‌ای درون‌واژه‌ای شامل دنباله‌های سه تا پنج نویسه‌ای. الحاق این دو بازنمایی به رگرسیون لجستیک با وزن‌دهی متوازن کلاس‌ها، سقف ۲۰۰۰ تکرار و بذر ثابت ۴۲ داده شد. به‌دلیل نامتوازن‌بودن توزیع سه کلاس، انتخاب مدل بر اساس اف‌یک ماکرو انجام شد تا هر کلاس سهم یکسانی در معیار انتخاب داشته باشد.

ضریب منظم‌سازی از میان مقادیر $C \in \{0.1, 1, 10\}$ با اعتبارسنجی متقاطع طبقه‌بندی‌شدهٔ پنج‌لایه روی ۱۷۴۶ نمونهٔ آموزش انتخاب شد. در هر لایه، واژگان و وزن‌های TF-IDF فقط روی چهار لایهٔ آموزشی برازش شدند. معیار اصلی، میانگین اف‌یک ماکروی اعتبارسنجی بود؛ در صورت تساوی، دقت بیشتر و سپس مقدار کوچک‌تر C اولویت داشت. این فرایند مقدار $C = 1$ را انتخاب کرد. سپس خط لولهٔ تازه‌ای با همین مقدار روی کل بخش آموزش برازش و فقط یک‌بار روی ۴۴۲ نمونهٔ آزمون ارزیابی شد. بنابراین، هیچ متن یا برچسبی از آزمون در ساخت ویژگی یا انتخاب ابرپارامتر دخالت نداشت.

۴-۶ تولید منابع اولیهٔ پیش‌بینی

۴-۶-۱ پیش‌بینی مستقیم

پیش‌بینی مستقیم ساده‌ترین منبع سامانه است. پرامپت این مسیر، جمله، هدف و سه برچسب مجاز را به مدل ارائه می‌کند. از مدل خواسته می‌شود فقط یک برچسب تولید کند و هیچ توضیح اضافی ننویسد. این اجرا صفرنمونه است و وزن‌های مدل در آن تغییر نمی‌کنند.

اگر P_D پرامپت مستقیم، f_θ مدل زبانی و $\mathcal{N}$ تابع نرمال‌سازی خروجی باشد، برچسب مستقیم به شکل زیر تعریف می‌شود:

$$\hat{y}_i^{(D)} = \mathcal{N}(f_\theta(P_D(s_i, t_i)))\,. \tag{۴-۲}$$

در اجرای اصلی، دمای تولید این مسیر صفر و سقف پاسخ کوتاه در نظر گرفته شده است. پارسر پاسخ را به حروف کوچک تبدیل و فاصله‌های اضافی را یکسان می‌کند؛ سپس سه برچسب معتبر را با رعایت مرز واژه جست‌وجو می‌کند. خروجی فقط هنگامی پذیرفته می‌شود که دقیقاً یک برچسب یکتای مثبت، منفی یا خنثی در پاسخ وجود داشته باشد. اگر هیچ برچسبی یافت نشود یا پاسخ بیش از یک برچسب متفاوت داشته باشد، مقدار داخلی unknown ثبت می‌شود تا خروجی غایب یا مبهم با یک قطبیت واقعی اشتباه نشود.

این مسیر دو کارکرد دارد. نخست، خط پایهٔ حداقلی برای سنجش ارزش مراحل پیچیده‌تر است. دوم، در کنترل‌گر و انتخاب‌گر نقش یک منبع پایدار را دارد. کوتاهی پرامپت و تولید قطعی باعث می‌شود تفاوت آن با مسیر استدلالی به شکل روشنی قابل بررسی باشد.

۴-۶-۲ استدلال چندمرحله‌ای مبتنی بر THOR

منبع دوم از ساختار THOR الهام گرفته است [۱]. در این مسیر، مدل پیش از برچسب نهایی سه مسئلهٔ وابسته را حل می‌کند: تعیین جنبهٔ مرتبط، استنتاج نظر و توضیح جهت قطبیت. هر مرحله، جمله، هدف و خروجی لازم از مراحل پیشین را دریافت می‌کند. به این ترتیب، برچسب فقط نتیجهٔ یک دستور کلی نیست و مسیر رسیدن به آن نیز ثبت می‌شود.

در گام جنبه، مدل باید مشخص کند کدام ویژگی یا بُعد از هدف در جمله مطرح شده است. پاسخ

این مرحله کوتاه نگه داشته می‌شود تا از تولید توضیح‌های پراکنده جلوگیری شود. اگر جنبهٔ مشخصی قابل استخراج نباشد، مقدار عمومی ثبت می‌شود. خروجی این مرحله با a_i نشان داده می‌شود:

$$a_i = f_\theta(P_A(s_i, t_i)) \,. \tag{۴-۳}$$

در گام نظر، مدل رابطهٔ رویداد توصیف‌شده با هدف و جنبه را تفسیر می‌کند. پرامپت از دانش عمومی استفاده می‌کند، اما هم‌زمان مدل را از تفسیر افراطی جمله‌های صرفاً خبری، فهرست‌گونه یا نامرتبط بازمی‌دارد. اگر شاهد کافی برای نگرش وجود نداشته باشد، باید نبود شاهد احساسی را بیان کند. این قید برای کاهش تبدیل خودکار هر رویداد به احساس طراحی شده است:

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

در گام قطبیت، مدل بررسی می‌کند که نظر استنتاج‌شده چه پیامدی برای هدف دارد و چرا این پیامد مثبت، منفی یا خنثی است. این پاسخ یک توضیح زبانی است و هنوز برچسب نهایی محسوب نمی‌شود:

$$r_i = f_\theta(P_R(s_i, t_i, o_i)) \,. \tag{۴-۵}$$

در پرامپت این گام، جنبه به‌صورت ورودی صریح تکرار نمی‌شود؛ اثر آن به‌طور غیرمستقیم از طریق نظر o_i که در گام پیشین با توجه به جنبه استخراج شده است، وارد تصمیم می‌شود. در پایان، یک پرامپت جداگانه توضیح قطبیت را به برچسبی از فضای $\mathcal{Y}$ تبدیل می‌کند. جداسازی توضیح و برچسب، احتمال مخلوط‌شدن متن استدلال با خروجی قابل ارزیابی را کاهش می‌دهد:

$$\hat{y}_i^{(T)} = \mathcal{N}(f_\theta(P_L(s_i, t_i, r_i))) \,. \tag{۴-۶}$$

۴-۶-۳ خودسازگاری سه‌اجرایی

تولید زنجیرهٔ استدلال با نمونه‌گیری غیرقطعی ممکن است میان اجراها تغییر کند. برای کاهش وابستگی نتیجه به یک مسیر منفرد، کل زنجیرهٔ THOR در یک سازگارسازی ویژهٔ این پژوهش سه بار اجرا می‌شود و رأی اکثریت برچسب‌های نهایی، خروجی خودسازگاری سه‌اجرایی SC3[۱] را می‌سازد. مراحل استخراج جنبه، نظر و تولید استدلال قطبیت با دمای ۰/۷ اجرا می‌شوند، اما پرامپت جداگانهٔ تبدیل استدلال به برچسب نهایی برای کاهش تغییرپذیری دمای صفر دارد. این طراحی از ایدهٔ خودسازگاری الهام می‌گیرد [۵] و بازتولید دقیق سازوکار مرحله‌به‌مرحلهٔ پژوهش اصلی THOR نیست. هر اجرا جنبه، نظر، استدلال قطبیت و برچسب مستقل خود را تولید می‌کند.

اگر برچسب اجرای kام با $\hat{y}_{i,k}^{(T)}$ نمایش داده شود، برچسب تجمیع‌شده از بیشترین تعداد رأی معتبر به دست می‌آید:

$$\hat{y}_i^{(T)} = \arg\max_{c \in \mathcal{Y}} \sum_{k=۱}^{۳} \mathbb{I}\left(\hat{y}_{i,k}^{(T)} = c\right). \tag{۴-۷}$$

پیش از رأی‌گیری، خروجی‌های نامعتبر کنار گذاشته می‌شوند. اگر هیچ اجرای معتبری وجود نداشته باشد، خروجی تجمیع‌شده unknown است. در حالت تساوی، برچسبی انتخاب می‌شود که زودتر در ترتیب اجراهای مساوی ظاهر شده باشد. مسیر نمایندهٔ ذخیره‌شده نیز نخستین اجرای دارای برچسب منتخب است. ثبت هر سه پاسخ خام و شمار رأی‌ها، امکان بازبینی رفتار این سازوکار را حفظ می‌کند.

۴-۷ بازبینی آگاه از نوع خطا

۴-۷-۱ شرط فعال‌سازی بازبینی

بازبینی زمانی فعال می‌شود که برچسب مستقیم و برچسب THOR با یکدیگر متفاوت باشند. در حالت برابری، بازبینی اجرا نمی‌شود و برای نمونه وضعیت بدون خطا ثبت می‌گردد. اگر برچسب مشترک معتبر باشد، کنترل‌گر همان برچسب را حفظ می‌کند. این شرط، هزینه و خطر بیش‌اصلاحی را محدود می‌کند؛ زیرا بازبین فقط در نقطه‌ای وارد می‌شود که شواهدی از ناسازگاری تصمیم وجود دارد.

[۱] Self-Consistency with three reasoning runs; SC3

بازبین جمله، هدف، پیش‌بینی مستقیم، مسیر استدلال THOR و برچسب آن را دریافت می‌کند. وظیفهٔ آن تولید یک توضیح آزاد نیست. خروجی باید سه جزء ساختاریافته داشته باشد: نوع خطا، برچسب پیشنهادی و سطح اطمینان. این تفکیک باعث می‌شود نقد مدل به سیگنال‌های قابل استفاده در قواعد تصمیم تبدیل شود.

۴-۷-۲ رده‌بندی خطاها

رده‌های خطا بر اساس الگوهای محتمل در تحلیل احساس ضمنی طراحی شده‌اند. جدول ۴-۱ این رده‌ها و اثر تصمیمی آن‌ها را نشان می‌دهد. شش رده قابل اصلاح‌اند و دو رده به‌تنهایی مجوز تغییر پاسخ را نمی‌دهند.

جدول ۴-۱: انواع خروجی بازبینی تشخیصی و نقش آن‌ها در تصمیم‌گیری.

شناسه	توصیف	نقش
missed_implicit_negative	نادیده‌گرفتن پیامد یا نگرش منفیِ بیان‌نشده به‌صورت صریح	قابل اصلاح
missed_implicit_positive	نادیده‌گرفتن پیامد یا نگرش مثبتِ بیان‌نشده به‌صورت صریح	قابل اصلاح
neutral_overinterpretation	نسبت‌دادن قطبیت به گزارهٔ خنثی یا فاقد شاهد کافی	قابل اصلاح
target_scope_shift	انتقال نگرش از هدف واقعی جمله به موجودیت یا هدفی دیگر	قابل اصلاح
aspect_opinion_mismatch	ناسازگاری میان جنبهٔ انتخاب‌شده و نظر استنتاج‌شده	قابل اصلاح
reasoning_label_ inconsistency	ناسازگاری برچسب نهایی با استدلال قطبیت تولیدشده	قابل اصلاح
no_error	تشخیص‌ندادن خطای قابل اتکا در مسیر بررسی‌شده	غیر اصلاحی
insufficient_evidence	ناکافی‌بودن شواهد برای داوری مطمئن میان منابع	غیر اصلاحی

رده‌های «از دست‌رفتن احساس مثبت یا منفی» به ناتوانی در بازیابی نگرش ضمنی مربوط‌اند. «بیش‌تفسیر خنثی» حالت معکوس را پوشش می‌دهد؛ یعنی مدل از یک واقعیت، نگرشی تولید کرده است که شواهد کافی ندارد. «تغییر دامنهٔ هدف» و «ناهمخوانی جنبه و نظر» خطاهای هدف‌محوری را مشخص می‌کنند. «ناسازگاری استدلال و برچسب» نیز حالتی است که متن استدلال یک جهت و برچسب نهایی جهت دیگری را نشان می‌دهد.

۴-۷-۳ قالب خروجی و ترمیم پاسخ

خروجی بازبین در قالب JSON درخواست می‌شود. فیلد error_type یکی از شناسه‌های جدول، فیلد label یکی از سه قطبیت و فیلد confidence یکی از سطوح low، medium یا high است. این محدودیت،

تبدیل پاسخ زبانی به داده‌ای قابل پردازش را ساده می‌کند.

پارسر ابتدا قالب کامل JSON را بررسی می‌کند و سپس الگوهای متنی نزدیک به قالب را نیز می‌پذیرد. نوع خطای ناشناخته به «شاهد ناکافی» و سطح اطمینان نامعتبر به سطح پایین نگاشت می‌شود. اگر برچسب پیشنهادی نامعتبر باشد، یک پرامپت ترمیمی فقط برای بازتولید برچسب اجرا می‌شود. اگر تلاش ترمیم نیز برچسب معتبری نسازد، مقدار unknown باقی می‌ماند و اصلاح در کنترل‌گر پذیرفته نمی‌شود.

۴-۸ کنترل‌گر تصمیم‌گیری میانی

کنترل‌گر قاعده‌محور از پیش‌بینی مستقیم، پیش‌بینی THOR و خروجی بازبین استفاده می‌کند. هدف آن ساخت تصمیمی ردیابی‌پذیر است که دلیل هر انتخاب را نیز ثبت کند. این مؤلفه به داده‌های طلایی آزمون دسترسی ندارد و قواعد آن برای تمام نمونه‌ها یکسان‌اند.

نخست، اگر دو پیش‌بینی معتبر مستقیم و استدلالی برابر باشند، همان برچسب مشترک نگه داشته می‌شود. دوم، اگر نوع خطا در مجموعهٔ خطاهای قابل اصلاح باشد، برچسب تشخیصی معتبر باشد و اطمینان دست‌کم متوسط گزارش شود، کنترل‌گر برچسب پیشنهادی را می‌پذیرد. صرف تولید یک نقد برای تغییر پاسخ کافی نیست؛ نوع خطا و اطمینان نیز باید شرط پذیرش را برآورده کنند.

در سایر حالت‌ها، قاعدهٔ بازگشت محافظه‌کارانه اجرا می‌شود. پیکربندی اصلی در اختلاف‌ها به پیش‌بینی مستقیم بازمی‌گردد و گزینهٔ no_error را مجوز ترجیح THOR نمی‌داند. اگر برچسب مستقیم نامعتبر باشد، کنترل‌گر به‌ترتیب از برچسب معتبر THOR و سپس برچسب تشخیصی استفاده می‌کند. اگر هیچ منبع معتبری وجود نداشته باشد، خروجی داخلی unknown ثبت می‌شود.

برای هر تصمیم، یک دلیل ماشینی مانند توافق منابع، پذیرش نوع خطای مشخص یا بازگشت به منبع مستقیم ذخیره می‌شود. این فیلد امکان تحلیل فراوانی قواعد و بررسی نمونه‌های تغییرکرده را فراهم می‌کند. بااین‌حال، خروجی کنترل‌گر پیش‌بینی نهایی سامانه نیست. نقش آن جمع‌بندی قاعده‌محور سیگنال‌ها و فراهم‌کردن یک نقطهٔ مقایسه برای ارزیابی فصل پنجم است.

۴-۹ انتخاب منبع محافظت‌شده و تنظیم‌شده با دادهٔ آموزش

۴-۹-۱ تعریف پروفایل انتخاب

هیچ‌یک از منابع برای تمام نمونه‌ها برتر نیست. برای تبدیل این مشاهده به یک سیاست عملی، نمونه‌ها بر اساس پنج سیگنال در پروفایل‌های عملکردی گسسته گروه‌بندی می‌شوند. پروفایل نمونهٔ i به شکل زیر است:

$$p_i = \left(\hat{y}_i^{(D)}, \hat{y}_i^{(T)}, e_i, c_i, d_i\right), \tag{۴-۸}$$

که در آن، e_i نوع خطای تشخیصی، c_i سطح اطمینان بازبین و d_i دامنه است. کنار هم قرارگرفتن برچسب‌های مستقیم و THOR نوع توافق یا اختلاف دو منبع را آشکار می‌کند؛ نوع خطا و اطمینان ماهیت اختلاف را توصیف می‌کنند؛ و دامنه تفاوت الگوهای لپ‌تاپ و رستوران را در نظر می‌گیرد.

۴-۹-۲ یادگیری سیاست از بخش آموزش

مجموعهٔ منابع نامزد با $\mathcal{S} = \{D, T, E\}$ نمایش داده می‌شود. این نمادها به‌ترتیب نمایندهٔ پیش‌بینی مستقیم، پیش‌بینی THOR و برچسب تشخیصی‌اند. برای هر پروفایل مشاهده‌شده در بخش آموزش، تعداد پاسخ‌های درست هر منبع محاسبه می‌شود. منبع دارای بیشترین شمار پاسخ درست، منبع منتخب آن پروفایل است:

$$\pi(p) = \arg\max_{s \in \mathcal{S}} \sum_{i \in \mathcal{D}_{\text{train}}} \mathbb{I}(p_i = p)\, \mathbb{I}\left(\hat{y}_i^{(s)} = y_i\right). \tag{۴-۹}$$

در این محاسبه، برچسب طلایی فقط برای سطرهای آموزش خوانده می‌شود. در صورت برابری امتیاز منابع، ترتیب ثابت مستقیم، THOR و تشخیصی اعمال می‌شود؛ بنابراین، منبع مستقیم نخستین اولویت تساوی است. بااین‌حال، انتخاب صرفاً بر پایهٔ بیشترین شمار پاسخ درست برای پروفایل‌های کم‌نمونه ناپایدار است. ازاین‌رو، سیاست اصلی پیش از واگذاری نمونه به منبعی غیر از مستقیم، چهار شرط محافظ را بررسی می‌کند.

۴-۹-۳ تنظیم آستانه‌های محافظ با اعتبارسنجی داخلی

شرط نخست حداقل پشتیبانی پروفایل، یعنی شمار نمونه‌های آموزشی همان پروفایل، است. شرط دوم حداقل حاشیهٔ برتری منبع منتخب نسبت به منبع مستقیم، شرط سوم حاشیهٔ آن نسبت به دومین منبع و شرط چهارم حداقل بهبود نسبی نسبت به منبع مستقیم است. اگر هر شرط برقرار نباشد یا پروفایل در آموزش دیده نشده باشد، سامانه به پیش‌بینی مستقیم بازمی‌گردد.

مقادیر این آستانه‌ها از روی آزمون انتخاب نمی‌شوند. بخش آموزش با حفظ گروه‌های دامنه و قطبیت، برای بذرهای صفر تا ۹ به دو زیرمجموعهٔ ساخت سیاست و اعتبارسنجی داخلی تقسیم می‌شود. برای هر پیکربندی، اف‌یک ماکرو و سپس دقت اعتبارسنجی محاسبه می‌شوند و میانگین ده تقسیم مبنای انتخاب است. پیکربندی منتخب از پروفایل پنج‌جزئی بالا و آستانه‌های حداقل پشتیبانی ۱۰، حداقل حاشیهٔ دو پاسخ درست نسبت به منبع مستقیم، حاشیهٔ صفر نسبت به منبع دوم و حداقل بهبود نسبی 0.05 استفاده می‌کند. پس از تثبیت این مقادیر، سیاست از کل بخش آموزش آموخته و بدون تغییر روی آزمون اعمال می‌شود.

۴-۹-۴ تولید پیش‌بینی نهایی

پس از تعیین سیاست محافظت‌شده π_g، برچسب نهایی هر نمونه از منبع منتخب پروفایل آن خوانده می‌شود:

$$\hat{y}_i^* = \hat{y}_i^{(\pi_g(p_i))}. \qquad (۱۰-۴)$$

علاوه بر برچسب نهایی، نام منبع منتخب نیز ذخیره می‌شود. این جداسازی برای تحلیل رفتار انتخاب‌گر ضروری است؛ زیرا می‌توان مشخص کرد هر پروفایل به کدام منبع واگذار شده و انتخاب منبع نسبت به پیش‌بینی مستقیم چه سود یا زیانی داشته است. خروجی کنترل‌گر در این فرمول وارد نمی‌شود و فقط به‌عنوان تصمیم میانی جداگانه باقی می‌ماند.

مزیت این سیاست نسبت به یک قاعدهٔ ثابت آن است که اعتبار نسبی منابع را در وضعیت‌های تکرارشونده از دادهٔ آموزش برآورد می‌کند و با شروط محافظ از تصمیم‌های مبتنی بر شواهد اندک می‌پرهیزد. برچسب طلایی آزمون نه در ساخت پروفایل، نه در انتخاب آستانه‌ها و نه در انتخاب منبع هر

نمونه نقش ندارد و فقط پس از تثبیت پیکربندی برای محاسبهٔ سنجه‌های نهایی خوانده می‌شود. مقایسهٔ کمی سیاست محافظت‌شده و نسخهٔ بدون محافظ در فصل پنجم ارائه خواهد شد.

۴-۱۰ انتخاب‌گرهای یادگرفتنی جایگزین

۴-۱۰-۱ انتخاب‌گر فراسطح مبتنی بر درخت

برای بررسی اینکه آیا یک مدل یادگرفتنی می‌تواند روابط پیچیده‌تری را بیاموزد، برای هر نمونهٔ آموزش یک برچسب منبع ساخته می‌شود. منابع به ترتیب مستقیم، THOR و تشخیصی بررسی می‌شوند و نخستین منبعی که پاسخ درست دارد، هدف آموزشی است؛ اگر هیچ منبعی درست نباشد، منبع مستقیم به‌عنوان بازگشت ثبت می‌شود. این ترتیب در نمونه‌هایی که چند منبع هم‌زمان درست‌اند، هدف را به سمت منبع زودتر متمایل می‌کند و یکی از محدودیت‌های این صورت‌بندی است.

ویژگی‌ها شامل برچسب‌های منابع مستقیم، THOR، تشخیصی و کنترل‌گر، نوع خطا، اطمینان تشخیصی، دامنه، فعال‌شدن بازبینی، توافق زوجی منابع و ویژگی‌های رأی‌گیری SC3 مانند بیشترین رأی، حاشیهٔ رأی، سهم رأی برتر و شمار برچسب‌های یکتا هستند. بخش آموزش با بذر ۴۲ و به‌صورت طبقه‌بندی‌شده به ۷۵ درصد برازش و ۲۵ درصد اعتبارسنجی داخلی تقسیم می‌شود. نامزدها شامل رگرسیون لجستیک متوازن، درخت‌های تصمیم با عمق‌های سه و پنج و حداقل اندازهٔ برگ ۱۰ یا ۲۰، و گرادیان‌بوستینگ با ۱۰۰ درخت عمق دو و نرخ یادگیری 0.05 هستند. مدل با بیشترین اف‌یک ماکروی اعتبارسنجی انتخاب و سپس روی کل بخش آموزش بازبرازش می‌شود؛ هیچ برچسب آزمونی در این فرایند به کار نمی‌رود.

۴-۱۰-۲ رتبه‌بند لجستیک منبع

در صورت‌بندی دوم، هر نمونهٔ آموزش به سه ردیف نامزد، یکی برای هر منبع، گسترش می‌یابد؛ بنابراین، ۱۷۴۶ نمونهٔ آموزش ۵۲۳۸ ردیف نامزد می‌سازند. ویژگی‌های مشترک نمونه همراه با نام منبع، برچسب پیشنهادی آن و توافق آن با سایر خروجی‌ها به رگرسیون لجستیک داده می‌شوند. هدف دودویی مشخص می‌کند آیا برچسب همان منبع با پاسخ طلایی برابر است. مدل با C=1، وزن‌دهی متوازن کلاس‌ها، سقف ۳۰۰۰ تکرار و بذر ۴۲ برازش می‌شود و در زمان پیش‌بینی، منبع دارای بیشترین امتیاز انتخاب می‌گردد. اگر امتیاز قابل استفاده‌ای وجود نداشته باشد، مسیر مستقیم گزینهٔ بازگشت است.

۴-۱۱ جزئیات پیاده‌سازی و بازتولیدپذیری

۴-۱۱-۱ معماری نرم‌افزار

پیاده‌سازی به مؤلفه‌های مستقل تقسیم شده است تا هر خروجی پیش از ورود به مرحلهٔ بعد قابل بررسی باشد. جدول ۴-۲ نقش فایل‌های اصلی را خلاصه می‌کند. این جداسازی امکان اجرای مجدد یک مرحله را بدون بازتولید تمام پاسخ‌های پرهزینه فراهم می‌سازد.

جدول ۴-۲: مؤلفه‌های اصلی پیاده‌سازی سامانهٔ پیشنهادی.

فایل یا مؤلفه	وظیفه
src/data_loader.py	خواندن داده، استخراج هدف‌ها و ساخت جدول یکپارچه
experiments/ run_tfidf_logreg_ baseline.py	ساخت خط پایهٔ کلاسیک، تنظیم C فقط روی آموزش و ذخیرهٔ پیش‌بینی‌های آزمون
src/prompt_runner.py	ارسال پرامپت، دریافت پاسخ و یکسان‌سازی رابط مدل‌ها
src/thor_pipeline.py	اجرای مراحل جنبه، نظر، قطبیت و تولید برچسب THOR
src/reflection_pipeline.py	تولید و پردازش بازبینی ساختاریافتهٔ نوع خطا
src/controller.py	اعمال قواعد کنترل‌گر و ثبت دلیل تصمیم میانی
experiments/apply_etc_policy.py	یادگیری پروفایل‌ها از بخش آموزش و انتخاب منبع نهایی
src/evaluator.py src/final_results.py	اعتبارسنجی برچسب‌ها و محاسبهٔ خروجی‌های ارزیابی

پرامپت‌ها در فایل‌های متنی مستقل نگه‌داری می‌شوند و کد اجرا فقط مقادیر جمله، هدف و خروجی مراحل پیشین را در آن‌ها جای‌گذاری می‌کند. این روش، نسخهٔ دستورها را از منطق برنامه جدا می‌سازد. پارامترهایی مانند دما، سقف توکن، شمار اجراها، مسیر خروجی و حالت ادامهٔ اجرا نیز به‌صورت تنظیمات ثبت می‌شوند.

۴-۱۱-۲ مدل‌ها و رابط اجرا

اجرای اصلی با مدل Qwen3 8B و از طریق Ollama انجام شده است. مسیر تکمیلی نیز همان مراحل سامانه را با Gemini 2.5 Flash از طریق رابط سازگار با OpenAI اجرا می‌کند. یک لایهٔ مشترک، ساخت پیام، فراخوانی مدل و استخراج متن پاسخ را برای این دو پشتانه یکسان می‌کند. در هیچ‌یک از این اجراها وزن مدل تغییر داده نمی‌شود و تفاوت رفتار از پرامپت، نمونه‌گیری و سیاست انتخاب ناشی است.

تفاوت تنظیمات مسیر مستقیم و استدلالی آگاهانه است. پیش‌بینی مستقیم با تولید قطعی و پاسخ کوتاه اجرا می‌شود تا یک خط پایهٔ پایدار بسازد. مسیر خودسازگاری از نمونه‌گیری با دمای غیرصفر

استفاده می‌کند تا مسیرهای متفاوتی ایجاد شوند. تمام تنظیمات همراه خروجی ذخیره می‌شوند تا مقایسه به پیکربندی مشخصی ارجاع داشته باشد.

۴-۱۱-۳ ذخیره، ادامهٔ اجرا و اعتبارسنجی

خروجی هر مرحله در فایل CSV جداگانه ذخیره می‌شود. پاسخ خام مدل، برچسب نرمال‌شده، اجزای استدلال، شمار رأی‌ها، نوع خطا، اطمینان، تصمیم کنترل‌گر و منبع منتخب در ستون‌های مستقل قرار می‌گیرند. نگه‌داری پاسخ خام مانع از آن می‌شود که خطای پارس یا نرمال‌سازی بدون امکان بازبینی باقی بماند.

اجراهای طولانی از سرگیری می‌شوند. پیش از پردازش هر سطر، وجود خروجی معتبر آن مرحله بررسی می‌شود و فقط نمونه‌های ناقص دوباره اجرا می‌شوند. ذخیرهٔ دوره‌ای نیز اثر قطع ارتباط یا توقف فرایند را محدود می‌کند. این سازوکار به‌ویژه برای سه اجرای THOR و بازبینی اختلاف‌ها اهمیت دارد.

پیش از تولید نتایج نهایی، هم‌ترازی سطرها میان فایل‌ها اعتبارسنجی می‌شود. شمار کل نمونه‌ها، شناسه‌ها، جمله، هدف، دامنه، بخش داده و برچسب طلایی باید یکسان باشند. سطر تکراری، جابه‌جایی کلید یا برچسب نهایی خارج از فضای مجاز به‌عنوان خطا گزارش می‌شود. پس از عبور از این کنترل‌ها، فایل یکپارچهٔ نهایی مبنای محاسبهٔ معیارها و ترسیم نمودارهای فصل پنجم قرار می‌گیرد.

۴-۱۲ جمع‌بندی

در این فصل، سامانهٔ پیشنهادی برای تحلیل احساسات ضمنی تشریح شد. سامانه ابتدا دو دیدگاه متفاوت تولید می‌کند: پیش‌بینی مستقیم و استدلال چندمرحله‌ای THOR با خودسازگاری سه‌اجرایی. اختلاف این دو مسیر، بازبینی ساختاریافته‌ای را فعال می‌کند که نوع خطا، برچسب پیشنهادی و سطح اطمینان را گزارش می‌دهد. کنترل‌گر قاعده‌محور از این سیگنال‌ها یک تصمیم میانی قابل ردیابی می‌سازد.

پیش‌بینی نهایی با سیاستی جداگانه انتخاب می‌شود. این سیاست، نمونه‌های آموزش را بر اساس برچسب‌های مستقیم و THOR، نوع خطا، اطمینان و دامنه گروه‌بندی می‌کند و برای هر پروفایل، مناسب‌ترین منبع را از میان مستقیم، THOR و تشخیصی می‌سنجد. واگذاری به منبع غیرمستقیم فقط با پشتیبانی و حاشیهٔ کافی انجام می‌شود و در سایر حالت‌ها قاعدهٔ بازگشت مستقیم فعال است. برچسب طلایی آزمون در ساخت پروفایل، تنظیم آستانه یا انتخاب منبع به کار نمی‌رود.

تفکیک مسیرهای تولید، بازبینی، کنترل میانی و انتخاب نهایی، امکان بررسی سهم هر مؤلفه را فراهم

می‌کند. فصل پنجم این مؤلفه‌ها را روی داده‌های آزمون مقایسه می‌کند، ماتریس‌های درهم‌ریختگی را تحلیل می‌کند و رفتار انتخاب‌گر را در دو اجرای Qwen3 8B و Gemini 2.5 Flash بررسی خواهد کرد.

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

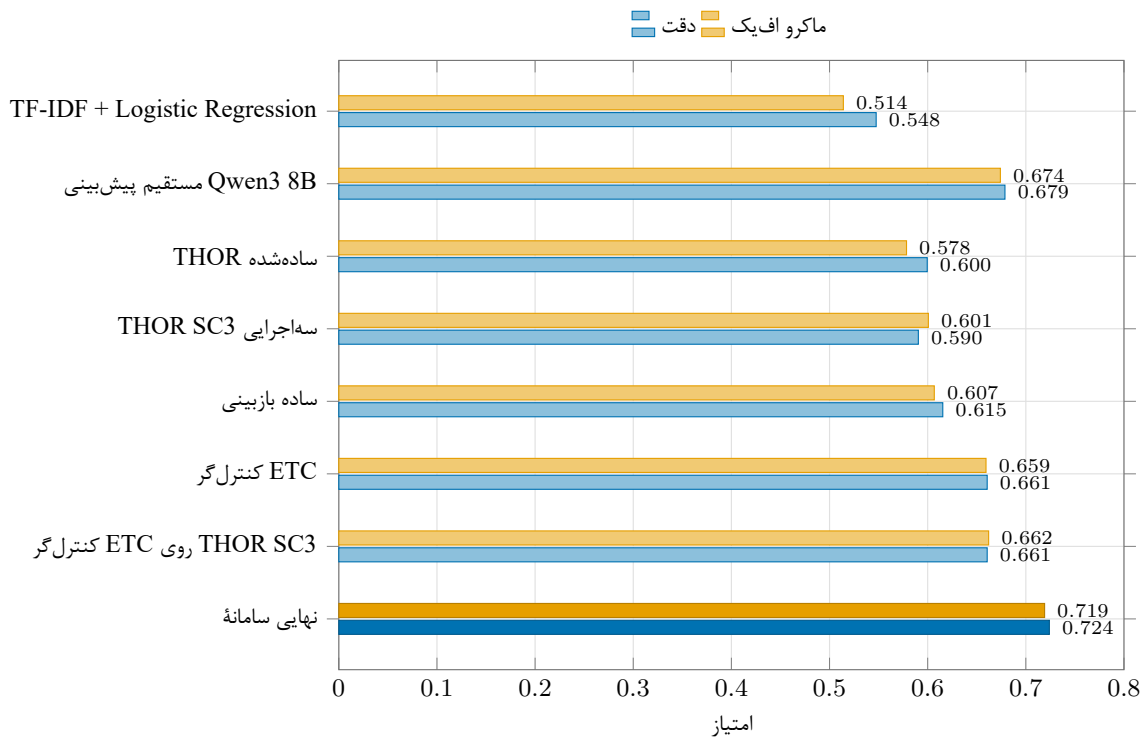
خط پایهٔ کلاسیک در نخستین آزمایش به دقت ۰/۵۴۷۵۱۱ و اف‌یک ماکرو ۰/۵۱۴۰۸۲ رسید. اف‌یک کلاس منفی در این روش ۰/۳۷۴۳۳۲ بود و نشان داد که تشخیص احساس ضمنی منفی دشوارترین بخش این بازنمایی کلاسیک است. پیش‌بینی مستقیم Qwen3 8B دقت را حدود ۱۳/۱۲ واحد درصد و اف‌یک ماکرو را حدود ۱۶/۰۰ واحد درصد نسبت به این مرجع 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.

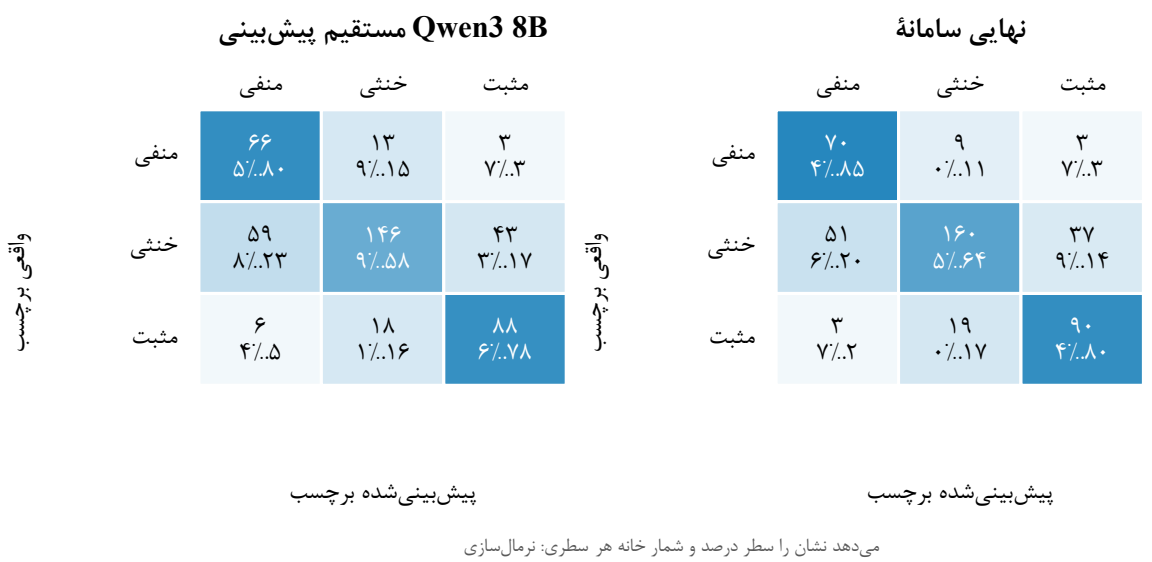
۵-۴-۲ تحلیل ماتریس‌های درهم‌ریختگی

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.

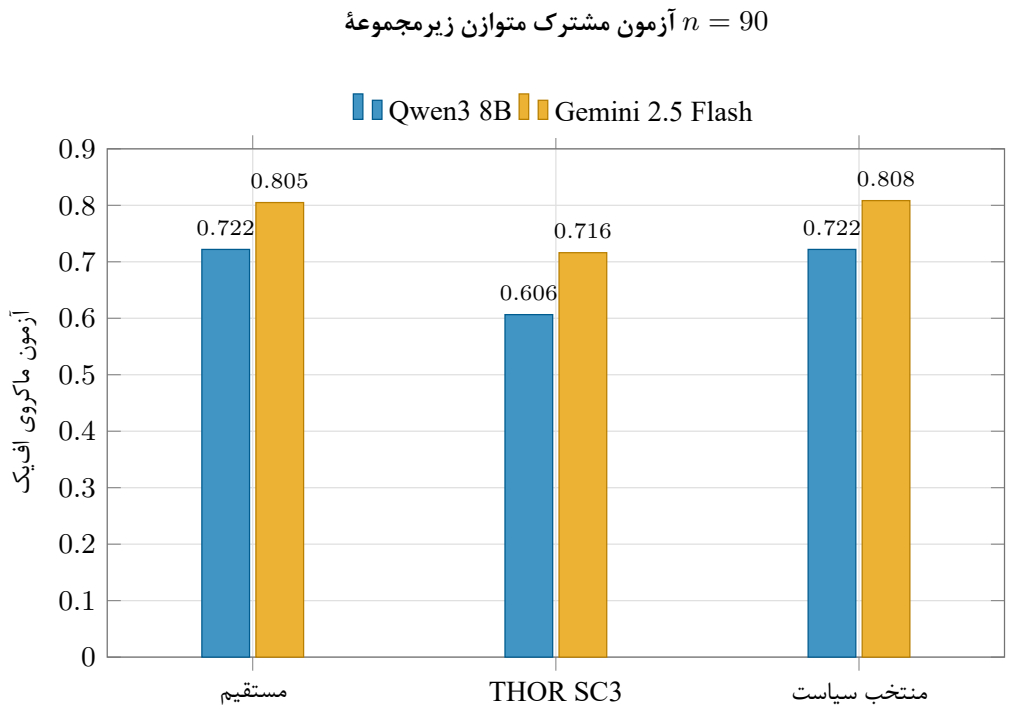
بااین‌حال، اندازهٔ ۹۰ نمونه برای نتیجه‌گیری کلی دربارهٔ برتری یک مدل کافی نیست. هدف این آزمایش بررسی رفتار دو پشتانه در یک چارچوب و مجموعهٔ یکسان است. نتیجهٔ اصلی پایان‌نامه همچنان همان ارزیابی Qwen3 8B روی ۴۴۲ نمونهٔ آزمون کامل است.

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.

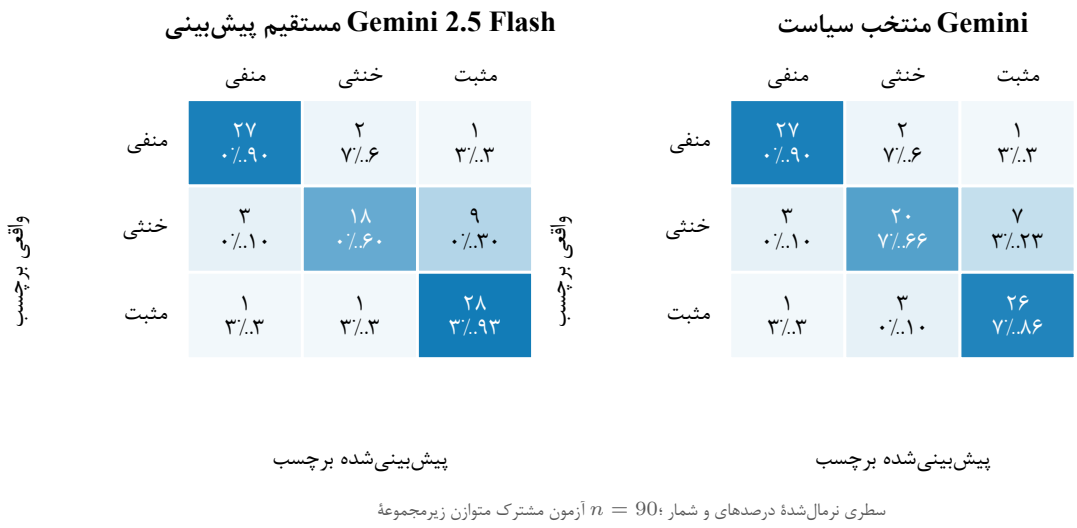
پس از تنظیم سیاست انتخاب، اف‌یک ماکرو Qwen3 8B برابر ۰/۷۲۱۹۰۳ باقی مانده و برای Gemini 2.5 Flash به ۰/۸۰۸۳۴۰ رسیده است. در این زیرمجموعه، انتخاب منبع برای Qwen نسبت به مسیر مستقیم سود خالصی ایجاد نکرده و بهبود Gemini نیز محدود است. ازاین‌رو، نتیجه شکل بیش از آنکه اثر علّی ظرفیت مدل را نشان دهد، بیانگر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