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

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 همیشگی شان،

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 دج

تیر ۱۴۰۵

چکیده

تشخیص احساس در متن‌هایی که فاقد واژه‌های ارزیابانهٔ آشکار هستند، به درک بافت، دانش عمومی و استنتاج رابطهٔ میان رویداد توصیف‌شده و هدف نیاز دارد. این پایان‌نامه سامانه‌ای چندمنبعی و مبتنی بر پرامپت برای تحلیل احساسات ضمنی ارائه می‌کند. سامانه پیش‌بینی مستقیم مدل زبانی را با استدلال چندمرحله‌ای الهام‌گرفته از روش THOR، بازبینی ساختاریافتهٔ خطا و سیاست انتخاب منبع ترکیب می‌کند. سیاست انتخاب نهایی با استفاده از دادهٔ آموزش تنظیم می‌شود. این سیاست بدون دسترسی به برچسب‌های طلایی آزمون، یکی از منابع مستقیم، استدلالی یا تشخیصی را برمی‌گزیند. ارزیابی اصلی روی داده‌های SCAPT مبتنی بر SemEval 2014 و در دو دامنهٔ لپ‌تاپ و رستوران انجام شده است. مجموعهٔ اصلی شامل ۲۱۸۸ نمونه است که ۱۷۴۶ نمونه در بخش آموزش و ۴۴۲ نمونه در بخش آزمون قرار دارند. اجرای اصلی با Qwen3 8B و از طریق Ollama انجام شده است. دقت سامانهٔ نهایی در بخش آزمون از ۰٫۶۷۸۷۳۳ به ۰٫۷۲۳۹۸۲ و امتیاز اف‌یک ماکرو از ۰٫۶۷۴۰۷۵ به ۰٫۷۱۹۱۱۹ رسیده است. افزون بر این، مقایسه‌ای تکمیلی با Gemini 2.5 Flash روی زیرمجموعهٔ متوازن ۹۰ نمونه‌ای انجام شده است. برای مدل، داده و تنظیم آزمایشی بررسی‌شده در این پایان‌نامه، نتایج نشان می‌دهند که ترکیب کنترل‌شدهٔ منابع پیش‌بینی از اتکای ثابت به یک مسیر مناسب‌تر است.

واژه‌های کلیدی:

تحلیل احساسات ضمنی، مدل‌های زبانی بزرگ، زنجیرهٔ تفکر، استدلال چندمرحله‌ای، انتخاب منبع

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 ۱۰

۳ ادبیات تحقیق ۱۱

۳-۱ تحلیل احساسات ضمنی و مجموعه‌داده‌ها ۱۲

۳-۲ استدلال و بازبینی در روش‌های مولد ۱۲

۳-۳ روش‌های THOR، SAoT و خودسازگاری ۱۳

۳-۴ انتخاب منبع و شکاف تحقیقاتی ۱۳

۴ کار پیشنهادی ۱۵

۴-۱ صورت‌بندی مسئله و نمای کلی سامانه ۱۶

۴-۲ تولید منابع پیش‌بینی ۱۷

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

۴-۴ انتخاب منبع محافظت‌شده و تنظیم‌شده با دادهٔ آموزش ۱۸
۴-۵ انتخاب‌گرهای جایگزین ۲۰
۴-۶ ساختار پیاده‌سازی سامانه ۲۰
۴-۷ پیاده‌سازی و جمع‌بندی ۲۱

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

۵-۷-۱ ساخت زیرمجموعهٔ مشترک ۳۳
۵-۷-۲ مقایسهٔ سنجه‌های دو مدل ۳۴
۵-۷-۳ ماتریس‌های درهم‌ریختگی Gemini ۳۵
۵-۸ ملاحظات اجرایی و محدودیت‌ها ۳۵
۵-۹ جمع‌بندی ۳۶

۶ نتیجه‌گیری و کارهای آتی ۳۷
۶-۱ جمع‌بندی پژوهش ۳۸
۶-۲ پاسخ به پرسش‌های پژوهش ۳۸
۶-۳ دستاوردهای اصلی ۳۹
۶-۳-۱ دستاوردهای علمی ۳۹
۶-۳-۲ دستاوردهای فنی و مهندسی ۳۹
۶-۴ محدودیت‌های پژوهش ۳۹
۶-۵ کارهای آتی ۴۰
۶-۵-۱ توسعه و بهبود انتخاب‌گر منبع ۴۰
۶-۵-۲ بهبود بازبینی و استدلال هدف‌محور ۴۰
۶-۵-۳ تحلیل خطاهای مشترک و گسترش منابع ۴۰
۶-۵-۴ گسترش ارزیابی به داده‌ها و مدل‌های دیگر ۴۰
۶-۵-۵ کاهش هزینهٔ اجرا ۴۰
۶-۵-۶ ارزیابی انسانی و ریزتنظیم ۴۱
۶-۶ نتیجه‌گیری نهایی ۴۱

منابع و مراجع ۴۲

پیوست: راهنمای بازتولید آزمایش‌ها ۴۴

واژه‌نامهٔ فارسی به انگلیسی ۴۶

واژه‌نامهٔ انگلیسی به فارسی ۴۷

فهرست اشکال

شکل صفحه

۴-۱ معماری سامانهٔ پیشنهادی؛ کنترل‌گر یک تصمیم میانی و انتخاب‌گر منبع تولیدکنندهٔ پاسخ نهایی است. ۱۶

۵-۱ مقایسهٔ دقت کلی و امتیاز اف‌یک ماکرو هشت روش اصلی روی ۴۴۲ نمونهٔ آزمون. . . ۲۷

۵-۲ ماتریس‌های درهم‌ریختگی پیش‌بینی مستقیم و سامانهٔ نهایی Qwen3 8B روی آزمون کامل. ۲۸

۵-۳ مقایسهٔ امتیاز اف‌یک ماکرو Qwen3 8B و Gemini 2.5 Flash روی زیرمجموعهٔ مشترک ۹۰ نمونه‌ای آزمون. ۳۴

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

فهرست جداول

جدول صفحه

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

۴-۱ رده‌های بازبینی تشخیصی و نقش تصمیمی آن‌ها. ۱۸

۴-۲ لایه‌های پیاده‌سازی و جریان ورودی/خروجی آن‌ها. ۲۰

۵-۱ توزیع نمونه‌های ضمنی در دو دامنه و دو بخش داده. ۲۳

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

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

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

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

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

۱ مدل و تنظیمات اجرای خروجی‌های ذخیره‌شده. ۴۴

۲ حداقل مراحل بازتولید گزارش. ۴۵

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
$F_۱$	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
[۳] Three-hop Reasoning
[۴] Supervised Contrastive Pre-Training

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

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

مثبتی وجود ندارد.

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 تجزیه‌گر 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

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

اطمینان کلامی مدل، بدون آزمون کالیبراسیون، احتمال تجربی صحت محسوب نمی‌شود. بنابراین 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.

[۹] Self-Consistency

[۱۰] Controller

[۱۱] Source Selector

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 کلی[۱۲] نسبت همهٔ پیش‌بینی‌های درست است:

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

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

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

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

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

اف‌یک ماکرو میانگین اف‌یک کلاس‌ها است:

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

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

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

فصل سوم

ادبیات تحقیق

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

۳-۱ تحلیل احساسات ضمنی و مجموعه‌داده‌ها

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

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

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

۳-۲ استدلال و بازبینی در روش‌های مولد

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

[۱] Supervised Contrastive Pre-Training

در عین حال، استدلال زبانی محدودیت‌هایی دارد. یک مرحلهٔ آغازین نادرست می‌تواند بقیهٔ زنجیره را منحرف کند و انسجام متن تضمین‌کنندهٔ صحت نیست. بازبینی و خوداصلاحی برای کشف ناسازگاری پاسخ اولیه توسعه یافته‌اند، اما بازبین نیز ممکن است پاسخ درست را بی‌دلیل تغییر دهد. بنابراین پذیرش اصلاح باید از تولید نقد جدا باشد؛ تمایزی که مبنای کنترل‌گر و انتخاب‌گر این پایان‌نامه است.

۳-۳ روش‌های THOR، SAoT و خودسازگاری

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

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

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

۳-۴ انتخاب منبع و شکاف تحقیقاتی

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

جدول حرکت ادبیات را از یادگیری بازنمایی به آشکارسازی و بازبینی استدلال نشان می‌دهد. این مؤلفه‌ها مکمل‌اند، اما هیچ‌کدام به‌تنهایی همهٔ نیازها را پوشش نمی‌دهد: بازنمایی قوی توضیح ندارد،

[۲] Sentiment Analysis of Thinking

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

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

استدلال ممکن است منحرف شود، رأی اکثریت خطای مشترک را تثبیت می‌کند و بازبینی می‌تواند بیش‌اصلاح کند.

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

فصل چهارم

کار پیشنهادی

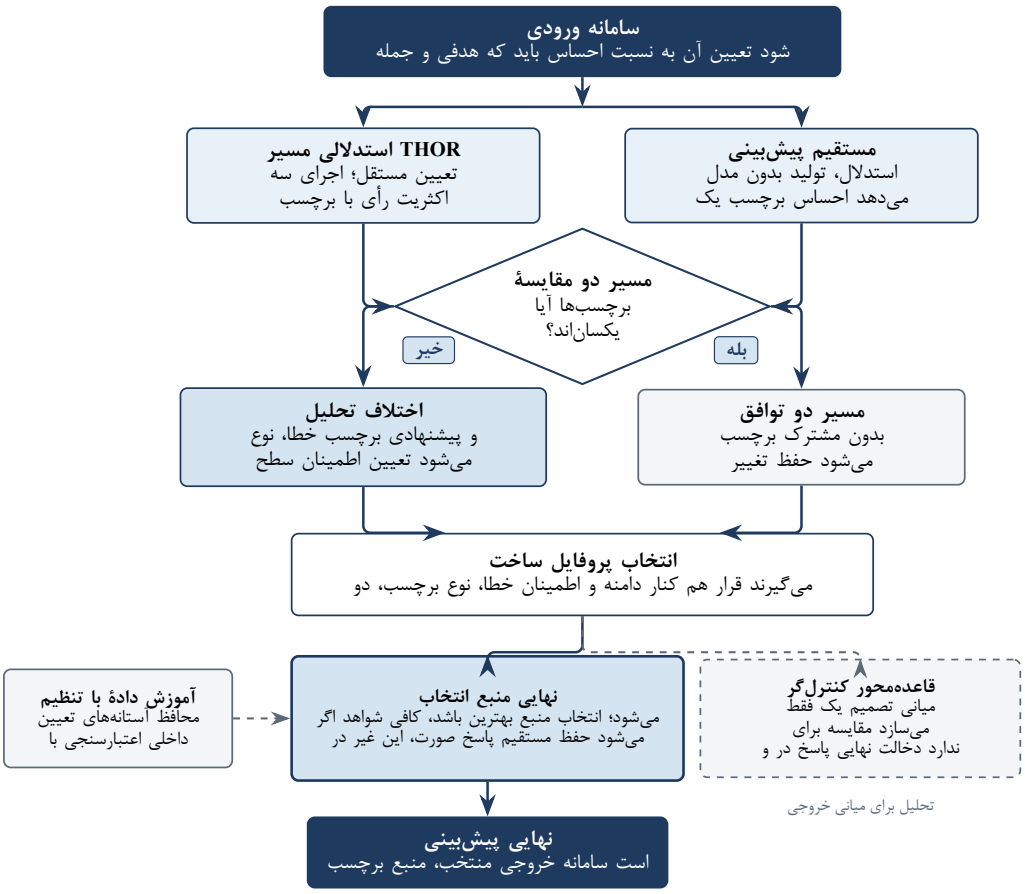
این فصل سامانهٔ چندمنبعی پژوهش را صورت‌بندی می‌کند. اجزای اصلی عبارت‌اند از تولید پیش‌بینی مستقیم و استدلالی، بازبینی اختلاف، کنترل میانی و سیاست مستقل انتخاب پاسخ نهایی.

۴-۱ صورت‌بندی مسئله و نمای کلی سامانه

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

$$\mathcal{Y} = \{\text{positive}, \text{negative}, \text{neutral}\}.$$

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



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

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

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

۴-۲ تولید منابع پیش‌بینی

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

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

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

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

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

ورودی صریح پرامپت قطبیت، (s, t, o) است؛ جنبه می‌تواند غیرمستقیم از طریق استخراج نظر اثر بگذارد. استخراج جنبه، نظر و استدلال با دمای ۰٫۷ انجام می‌شود، اما تبدیل استدلال به برچسب برای کاهش تغییرپذیری دمای صفر دارد. تجزیه‌گر فقط زمانی خروجی را می‌پذیرد که دقیقاً یک برچسب معتبر با مرز کلمه یافت شود؛ در غیر این صورت unknown ثبت می‌شود.

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

[۱] Self-Consistency with three reasoning runs

انتخاب می‌شود:

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

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

۴-۳ بازبینی آگاه از نوع خطا و کنترل‌گر میانی

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

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

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

کنترل‌گر قاعده‌محور در توافق، همان برچسب مشترک را نگه می‌دارد. در اختلاف، اصلاح فقط برای نوع خطای قابل اصلاح، برچسب تشخیصی معتبر و اطمینان دست‌کم متوسط پذیرفته می‌شود؛ در غیر این صورت مسیر مستقیم مبنای بازگشت است. اگر مستقیم نامعتبر باشد، THOR و سپس تشخیصی بررسی می‌شوند. دلیل ماشینی هر تصمیم ذخیره می‌شود. این خروجی صرفاً نقطهٔ مقایسه است و مستقیماً پاسخ نهایی سامانه محسوب نمی‌شود.

۴-۴ انتخاب منبع محافظت‌شده و تنظیم‌شده با دادهٔ آموزش

نمونه‌ها با پنج سیگنال در پروفایل عملکردی زیر گروه‌بندی می‌شوند:

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

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

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

در تساوی، ترتیب مستقیم، THOR و تشخیصی اعمال می‌شود. برای جلوگیری از تصمیم کم‌شاهد، واگذاری به منبع غیرمستقیم چهار شرط دارد: حداقل پشتیبانی پروفایل، حاشیهٔ برتری نسبت به مستقیم، حاشیه نسبت به منبع دوم و حداقل بهبود نسبی. آستانه‌ها با ده تقسیم اعتبارسنجی داخلی آموزش و با حفظ گروه‌های دامنه و قطبیت انتخاب می‌شوند؛ معیار نخست اف‌یک ماکرو و سپس دقت است. پیکربندی نهایی حداقل پشتیبانی ۱۰، حاشیهٔ دو پاسخ درست نسبت به مستقیم، حاشیهٔ صفر نسبت به منبع دوم و بهبود نسبی ۰٫۰۵ دارد. پروفایل دیده‌نشده یا نامطمئن به مستقیم بازمی‌گردد.

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

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

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

۴-۵ انتخاب‌گرهای جایگزین

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

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

۴-۶ ساختار پیاده‌سازی سامانه

هستهٔ قابل‌استفادهٔ مجدد سامانه در پوشهٔ src به‌صورت کلاس و تابع پیاده‌سازی شده است؛ در مقابل، انتخاب‌گرها و اجراهای دسته‌ای عمدتاً تابع‌ها و اسکریپت‌های پوشهٔ experiments هستند، نه کلاس‌های مستقل. جدول ۴-۲ مرز این لایه‌ها و قرارداد داده‌ای اصلی آن‌ها را خلاصه می‌کند.

جدول ۴-۲: لایه‌های پیاده‌سازی و جریان ورودی/خروجی آن‌ها.

لایه	کلاس یا فایل	ورودی/خروجی	مسئولیت
داده و پیکربندی	`data_loader.py` و `experiment_config.py`	XML/متغیرهای محیطی → CSV پاک و مسیرها و شناسه‌های اجرا	استخراج سطرهای SCAPT، اعتبارسنجی فیلدهای ضروری، پالایش برچسب‌ها و یکسان‌سازی تنظیمات مسیر و مدل
فراخوان مدل	PromptRunner، OllamaPromptRunner، HFSeq2SeqPromptRunner و OpenAICompatiblePromptRunner	پرامپت و پارامترهای تولید → متن پاسخ	انتخاب پس‌زمینهٔ Ollama، مدل متنی HF یا سازگار با OpenAI و تحویل رابط یکنواخت run به لایه‌های بالاتر
استدلال و بازبینی	THORPipeline، SimpleReflectionPipeline، ErrorTypeReflectionPipeline و `controller.py`	جمله، هدف و برچسب‌ها/ردپاها → پیش‌بینی، تشخیص و تصمیم کنترل‌گر	تولید ردپای جنبه، نظر و قطبیت؛ بازبینی ساده یا تشخیصی؛ و اعمال سیاست بازگشت در تصمیم میانی
خط پایهٔ کلاسیک	`run_tfidf_logreg_baseline.py`	آموزش و آزمون CSV → پیش‌بینی، نتایج اعتبارسنجی و سنجه‌ها	ساخت ویژگی‌های واژه‌ای و نویسه‌ای TF-IDF، آموزش رگرسیون لجستیک متوازن، انتخاب C فقط با اعتبارسنجی پنج‌لایه روی آموزش و بازبرازش مدل منتخب بر کل آموزش
انتخاب منبع	`apply_etc_policy.py`، `run_meta_selector.py` و `run_logistic_source_ranker.py`	خروجی‌های مستقیم، THOR و تشخیصی → منبع و برچسب منتخب	اعمال سیاست محافظت‌شده، مقایسهٔ فرادسته‌بند و رتبه‌بند لجستیک؛ این مؤلفه‌ها اسکریپت‌ها و تابع‌های آزمایشی‌اند
ارزیابی، تجمیع و شکل‌ها	`evaluator.py`، `final_results.py`، `run_final_pipeline.py` و `generate_thesis_result_figures.py`	CSVهای پیش‌بینی → دقت، Macro-F1، جدول سنجه، گزارش اعتبارسنجی و شکل‌ها	`evaluator.py` سنجه‌های مشترک را محاسبه می‌کند؛ `final_results.py` هم‌ترازی و تجمیع را کنترل می‌کند و اسکریپت‌ها گزارش و شکل می‌سازند

مسیر اجرا با بارگذاری داده و ساخت اجرای مستقیم آغاز می‌شود؛ سپس THORPipeline ردپای استدلالی و پیش‌بینی استدلالی را می‌سازد و در صورت اختلاف، ErrorTypeReflectionPipeline خروجی تشخیصی تولید می‌کند. `controller.py` تنها تصمیم میانی را می‌سازد و `apply_etc_policy.py` یا انتخاب‌گرهای آزمایشی، منبع نهایی را تعیین می‌کنند. در پایان، `run_final_pipeline.py` خروجی‌های مرحله‌ای را به سنجه‌ها و گزارش یکپارچه تبدیل می‌کند و `generate_thesis_result_figures.py` داده‌های گزارش‌شده را به شکل‌های پایان‌نامه تبدیل می‌کند.

کنترل‌های پیاده‌سازی ابتدا ستون‌های ضروری و نبود مقدار مفقود را بررسی می‌کنند. تجزیهٔ برچسب فقط یک برچسب معتبر را می‌پذیرد و خروجی مبهم را unknown ثبت می‌کند؛ تجزیهٔ تشخیصی نیز در صورت برچسب نامعتبر یک نوبت ترمیم را می‌آزماید. کنترل‌گر در نبود شاهد کافی به مسیر مستقیم بازمی‌گردد. خروجی مرحله‌های مولد به‌صورت CSV شامل پاسخ خام، برچسب نرمال‌شده و سیگنال‌های لازم ذخیره می‌شود؛ خط پایهٔ TF-IDF به‌جای پاسخ زبانی، متن مدل، پیش‌بینی و نتایج اعتبارسنجی را ذخیره می‌کند. تطبیق فایل‌ها با کلید مرکب، یکتایی کلید و شمار سطرها کنترل می‌شود و اجراهای مستقیم و استدلالیِ طولانی از خروجی ذخیره‌شده ادامه‌پذیرند.

۷-۴ پیاده‌سازی و جمع‌بندی

هر مرحله متناسب با نوع اجرا خروجی CSV خود را جدا ذخیره می‌کند: مرحله‌های مولد پاسخ خام، برچسب نرمال‌شده و سیگنال‌های لازم را ثبت می‌کنند و خط پایهٔ کلاسیک پیش‌بینی‌ها و نتایج اعتبارسنجی را. اجراهای طولانی قابل ادامه‌اند و پیش از ارزیابی، شمار ردیف‌ها، کلید مرکب، دامنه، تقسیم و برچسب‌ها میان فایل‌ها تطبیق داده می‌شوند. فهرست فایل‌ها و فرمان‌های بازتولید در پیوست آمده است.

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

فصل پنجم

نتایج آزمایش‌ها

۵-۱ مقدمه

این فصل اثر جداسازی تولید منابع، کنترل میانی و انتخاب پاسخ نهایی را می‌سنجد. ارزیابی اصلی، خط پایهٔ کلاسیک و پیکربندی‌های Qwen3 8B را روی ۴۴۲ نمونهٔ آزمون مقایسه و رفتار کلاسی، دامنه‌ای و انتخاب‌گر را تحلیل می‌کند. مقایسهٔ تکمیلی Qwen3 8B و Gemini 2.5 Flash فقط روی زیرمجموعهٔ مشترک ۹۰ نمونه‌ای انجام می‌شود و جایگزین نتیجهٔ اصلی نیست.

۵-۲ تنظیمات آزمایش

۵-۲-۱ مجموعه‌داده و تقسیم ارزیابی

داده‌های اصلی از نمونهٔ ضمنی SCAPT مبتنی بر SemEval 2014 به دست آمده‌اند. جدول ۵-۱ توزیع سه‌کلاسه را در دو دامنه نشان می‌دهد؛ تقسیم اصلی آموزش و آزمون بدون جابه‌جایی حفظ شده است.

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

آزمون شامل ۸۲ نمونهٔ منفی، ۲۴۸ خنثی و ۱۱۲ مثبت است؛ ازاین‌رو، افزون بر دقت، اف‌یک ماکرو و رفتار هر کلاس بررسی می‌شود.

۵-۲-۲ مدل‌ها و زیرساخت اجرا

اجرای اصلی با Qwen3 8B و Ollama انجام شد. پیش‌بینی مستقیم قطعی است؛ THOR SC3 زنجیره را سه بار با دمای استدلال ۰٫۷ اجرا و با رأی اکثریت تجمیع می‌کند، درحالی‌که تبدیل استدلال به برچسب دمای صفر دارد. بازبینی فقط هنگام اختلاف فعال می‌شود. مقایسهٔ تکمیلی نیز بدون تغییر وزن مدل‌ها، Gemini 2.5 Flash و Qwen3 8B را روی شناسه‌های یکسان می‌سنجد.

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، شمار و هم‌ترازی ۲۱۸۸ سطر، یکتایی کلید مرکب معرفی‌شده در فصل چهارم و اعتبار سه‌کلاس کنترل شد تا اختلاف ترتیب یا خروجی نامعتبر به‌صورت تغییر عملکرد تفسیر نشود.

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 با اختصار ETC[1] روی THOR ساده‌شده اعمال و تصمیم میانی قاعده‌محور تولید می‌شود؛ ETC در شکل‌ها نیز به همین کنترل‌گر اشاره دارد.

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

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.

[1] Error-Type-Aware Reflection

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.

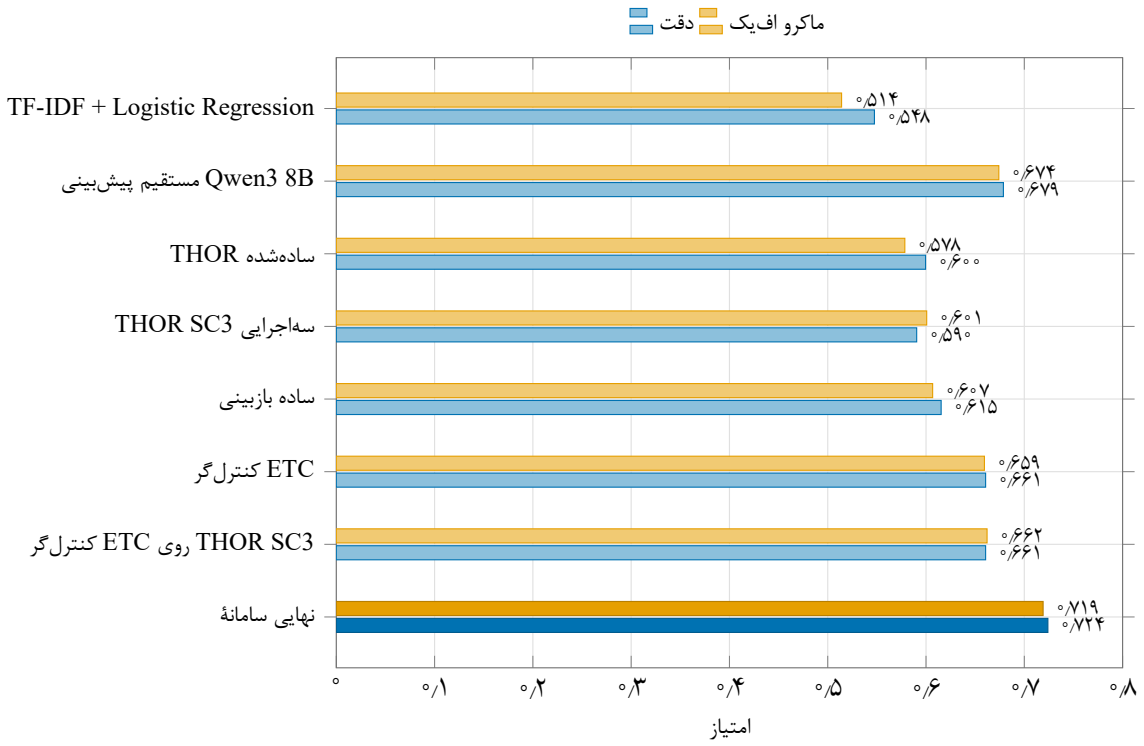
سامانهٔ نهایی نسبت به خط پایهٔ کلاسیک حدود ۱۷٫۶۵ واحد درصد در دقت و ۲۰٫۵۰ واحد درصد در اف‌یک ماکرو برتری دارد. بااین‌حال، این فاصله نباید فقط به معماری استدلالی پیشنهادی نسبت داده شود؛ زیرا خط پایهٔ کلاسیک و روش‌های مبتنی بر مدل زبانی از خانواده‌ها و ظرفیت‌های متفاوتی استفاده می‌کنند. نقش این آزمایش 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.

نهایی سامانهٔ

واقعی برچسب \ پیش‌بینی‌شده برچسب	منفی	خنثی	مثبت
منفی	۷۰ ۸۵/۴%	۹ ۱۱/۰%	۳ ۳/۷%
خنثی	۵۱ ۲۰/۶%	۱۶۰ ۶۴/۵%	۳۷ ۱۴/۹%
مثبت	۳ ۲/۷%	۱۹ ۱۷/۰%	۹۰ ۸۰/۴%

Qwen3 8B مستقیم پیش‌بینی

واقعی برچسب \ پیش‌بینی‌شده برچسب	منفی	خنثی	مثبت
منفی	۶۶ ۸۰/۵%	۱۳ ۱۵/۹%	۳ ۳/۷%
خنثی	۵۹ ۲۳/۸%	۱۴۶ ۵۸/۹%	۴۳ ۱۷/۳%
مثبت	۶ ۵/۴%	۱۸ ۱۶/۱%	۸۸ ۷۸/۶%

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.

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. 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.

رتبه‌بند لجستیک منبع. هر زوج نمونه و منبع یک ردیف آموزشی است؛ مدل احتمال درست‌بودن هر منبع را برآورد می‌کند و منبع با بیشترین امتیاز انتخاب می‌شود.

انتخاب‌گر فراسطح مبتنی بر درخت. این مدل با ویژگی‌های خروجی منابع، توافق، نوع خطا، اطمینان، دامنه و رأی‌ها مستقیماً نام منبع مناسب را پیش‌بینی می‌کند؛ انتخاب مدل فقط با تقسیم داخلی آموزش انجام می‌شود.

اوراکل انتخاب منبع. اوراکل با دیدن برچسب طلایی برای هر نمونه یک منبع درست را برمی‌گزیند؛ بنابراین فقط کران بالای تحلیلی است و هنگام استنتاج قابل استفاده نیست.

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

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

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

خروجی‌های مستقیم، THOR، کنترل‌گر و انتخاب‌گر هیچ کلید تکراری یا برچسب خارج از سه کلاس ندارند و همهٔ ۲۱۸۸ سطر آن‌ها با کلید مرکب هم‌ترازند. ستون‌های تشخیصی نیز در فایل نهایی حاضرند؛ بنابراین، سنجه‌ها بر زنجیره‌ای 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

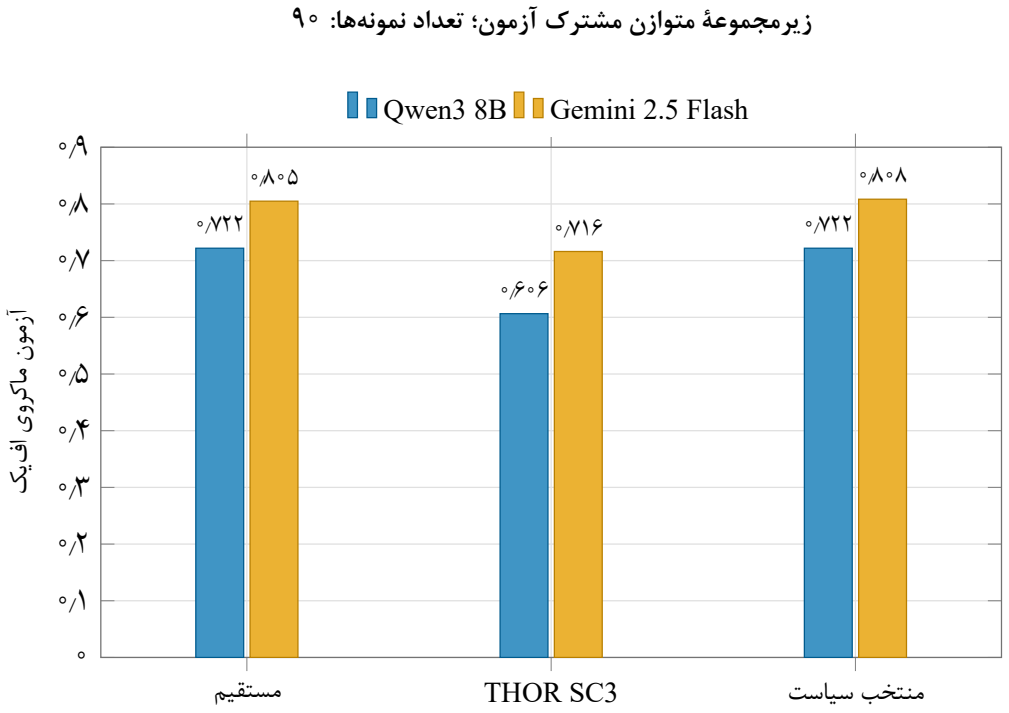
۵-۷-۱ ساخت زیرمجموعهٔ مشترک

برای مقایسهٔ دو مدل، زیرمجموعه‌ای متوازن با ۲۴۰ نمونه ساخته شده است. این زیرمجموعه ۱۵۰ نمونهٔ آموزش و ۹۰ نمونهٔ آزمون دارد. ۱۵۰ نمونهٔ آموزش فقط برای تنظیم سیاست انتخاب به کار می‌روند و تمام سنجه‌ها، شکل‌ها و ادعاهای مقایسه‌ای این زیربخش فقط بر ۹۰ نمونهٔ آزمون مشترک (۳۰ نمونه از هر کلاس) گزارش می‌شوند. 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.



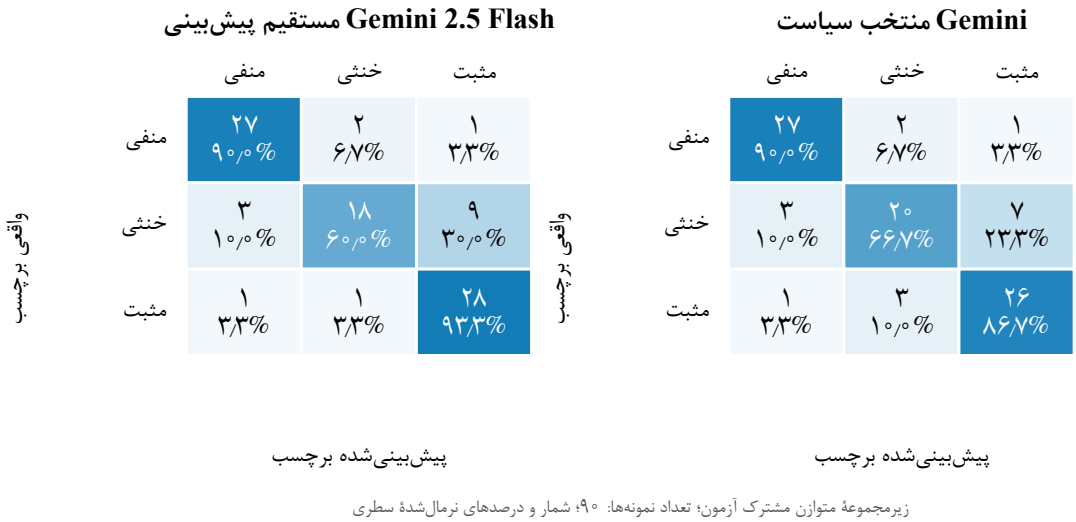
شکل ۵-۳: مقایسهٔ امتیاز اف‌یک ماکرو Qwen3 8B و Gemini 2.5 Flash روی زیرمجموعهٔ مشترک ۹۰ نمونه‌ای آزمون.

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.

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.



زیرمجموعهٔ متوازن مشترک آزمون؛ تعداد نمونه‌ها: ۹۰؛ شمار و درصدهای نرمال‌شدهٔ سطری

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

استدلال و کنترل‌گر ثابت به‌تنهایی از پیش‌بینی مستقیم عبور نکردند، اما انتخاب محافظت‌شدهٔ منبع، دقت را از ۰٫۶۷۸۷۳۳ به ۰٫۷۲۳۹۸۲ و اف‌یک ماکرو را از ۰٫۶۷۴۰۷۵ به ۰٫۷۱۹۱۱۹ رساند: ۲۶ اصلاح در برابر شش تضعیف. وجود ۱۱۶ خطای مشترک و فاصلهٔ حدود شش واحد درصدی با اوراکل، ظرفیت بهبود منابع و انتخاب‌گر را نشان می‌دهد. مقایسهٔ محدود Gemini نیز وابستگی سود مراحل بعدی به مدل و جمعیت ارزیابی را برجسته کرد.

فصل ششم

نتیجه‌گیری و کارهای آتی

۶-۱ جمع‌بندی پژوهش

این پژوهش سامانه‌ای چندمنبعی برای تحلیل احساسات ضمنی هدف‌محور طراحی و ارزیابی کرد. پیش‌بینی مستقیم در کنار مسیر استدلالی الهام‌گرفته از THOR قرار گرفت؛ بازبین هنگام اختلاف، نوع خطا و اطمینان را ساختارمند کرد و کنترل‌گر تصمیمی میانی ساخت. سپس سیاستی محافظت‌شده، که فقط با اعتبارسنجی داخلی آموزش تنظیم شده بود، منبع پاسخ نهایی را انتخاب کرد. این تفکیک مانع یکسان‌انگاری ردپای استدلال، تصمیم کنترل‌گر و پیش‌بینی نهایی می‌شود.

در آزمون کامل SCAPT با Qwen3 8B، استدلال چندمرحله‌ای و کنترل‌گر ثابت به‌تنهایی از پیش‌بینی مستقیم بهتر نبودند؛ بااین‌حال، خطاهای منابع کاملاً هم‌پوشان نبود. سیاست انتخاب توانست از این مکمل‌بودن استفاده کند. بنابراین، یافتهٔ محوری برتری یک پرامپت خاص نیست، بلکه ضرورت جداسازی تولید پاسخ‌های مکمل از انتخاب منبع قابل اعتماد است.

۶-۲ پاسخ به پرسش‌های پژوهش

۱. **تفاوت پیش‌بینی مستقیم و استدلال چندمرحله‌ای.** مسیر مستقیم کوتاه‌تر، کم‌هزینه‌تر و در تنظیم اصلی دقیق‌تر بود. THOR ردپایی قابل بررسی ساخت، اما خطای جنبه یا نظر را نیز به برچسب منتقل کرد؛ پس افزودن مراحل استدلال به‌خودی‌خود بهبود را تضمین نمی‌کند.

۲. **نقش بازبینی ساختاریافته و کنترل‌گر.** بازبین اختلاف را به نوع خطا، پیشنهاد اصلاح و اطمینان تبدیل کرد و این اطلاعات برای پروفایل‌بندی مفید بود. بااین‌حال، کنترل‌گر از خط پایهٔ مستقیم پایین‌تر ماند و برچسب تشخیصی در آزمون منبع نهایی نشد؛ در نتیجه، برتری مستقل آن‌ها اثبات نشده است.

۳. **اثر انتخاب منبع تنظیم‌شده با آموزش.** دقت از ۰٫۶۷۸۷۳۳ به ۰٫۷۲۳۹۸۲ و اف‌یک ماکرو از ۰٫۶۷۴۰۷۵ به ۰٫۷۱۹۱۱۹ رسید. سیاست ۲۶ خطا را اصلاح و شش پاسخ درست را تضعیف کرد؛ یعنی ۲۰ پاسخ درست بیشتر، بدون استفاده از برچسب آزمون در انتخاب.

۴. **الگوهای اصلی خطا.** سه الگوی غالب عبارت بودند از ازدست‌رفتن پیامد ضمنی، بیش‌تفسیر گزارهٔ خنثی و انتقال نگرش به هدف نادرست. ۱۱۶ خطای مشترک مستقیم و نهایی نشان دادند که انتخاب بهتر، بدون تنوع یا کیفیت بیشتر منابع، همهٔ خطاها را رفع نمی‌کند.

۵. **اثر انتخاب مدل و پشتانهٔ اجرا.** در آزمون مشترک ۹۰ نمونه‌ای، Gemini از اجرای متناظر Qwen امتیاز بالاتری داشت، ولی THOR برای هر دو مدل از مسیر مستقیم ضعیف‌تر بود و سود انتخاب منبع محدود ماند. این مشاهده به مدل، زیرساخت و جمعیت حاضر مقید است و اثر علّی ظرفیت مدل یا رتبه‌بندی عمومی دو مدل را اثبات نمی‌کند.

۶-۳ دستاوردهای اصلی

۶-۳-۱ دستاوردهای علمی

نخست، استدلال چندمرحله‌ای به‌عنوان منبع مکمل و نه پاسخ قطعی صورت‌بندی شد. دوم، بازبینی، کنترل میانی و انتخاب نهایی از یکدیگر جدا شدند. سوم، مکمل‌بودن خطاهای منابع با سیاستی آموخته‌شده و محافظت‌شده، بدون نشت اطلاعات آزمون، به بهبود عملی تبدیل شد؛ هرچند فاصلهٔ باقی‌مانده با اوراکل نشان می‌دهد انتخاب منبع هنوز مسئله‌ای باز است.

۶-۳-۲ دستاوردهای فنی و مهندسی

خط لولهٔ انتهابه‌انتها تقسیم رسمی داده را حفظ و خروجی خام و نرمال‌شدهٔ مراحل مستقیم، استدلالی، تشخیصی، کنترل‌گر و نهایی را جدا ذخیره می‌کند. کنترل خودکار شمار، یکتایی کلید مرکب، هم‌ترازی سطرها و اعتبار برچسب‌ها، ردیابی هر تصمیم و تکرار ارزیابی را در محدودهٔ تنظیمات گزارش‌شده ممکن می‌سازد.

۶-۴ محدودیت‌های پژوهش

دادهٔ اصلی انگلیسی و محدود به دامنه‌های لپ‌تاپ و رستوران است؛ مقایسهٔ Gemini 2.5 Flash نیز فقط ۹۰ نمونهٔ آزمون را پوشش می‌دهد. مسیر استدلالی غیرقطعی و پرهزینه‌تر است، اما اجرای مستقل کامل، ارزیابی انسانی وفاداری ردپا و زمان‌سنجی کنترل‌شده انجام نشده‌اند. همچنین، برچسب بازبین در آزمون منبع نهایی نبود و گروه‌های کم‌نمونه با بازگشت محافظت شده‌اند، نه اینکه مسئلهٔ کمبود شاهد حذف شده باشد. اوراکل فقط کران تحلیلی است و مسیر سازگارشدهٔ THOR نیز بازتولید کامل تنظیم رسمی آن پژوهش نیست؛ ازاین‌رو، نتیجه به چارچوب پرامپت‌محور، مدل‌ها و داده‌های بررسی‌شده محدود است.

۶-۵ کارهای آتی

۶-۵-۱ توسعه و بهبود انتخاب‌گر منبع

پایداری پروفایل‌ها و آستانه‌ها باید با تقسیم‌ها و بازنمونه‌گیری بیشتر سنجیده شود. هموارسازی گروه‌های کم‌نمونه و فرادسته‌بندهای محافظ می‌توانند علاوه بر میانگین عملکرد، شمار پاسخ‌های تضعیف‌شده را نیز معیار انتخاب قرار دهند.

۶-۵-۲ بهبود بازبینی و استدلال هدف‌محور

برجسته‌کردن بازهٔ هدف، سنجش ارتباط شاهد با هدف و آزمون سازگاری جنبه، نظر و برچسب می‌تواند جابه‌جایی دامنهٔ هدف را کاهش دهد. اطمینان بازبین نیز باید با داده‌ای مستقل کالیبره و با کاهش خطاهای هدف‌محور ارزیابی شود.

۶-۵-۳ تحلیل خطاهای مشترک و گسترش منابع

تحلیل ۱۱۶ خطای مشترک باید سهم دانش عمومی، ابهام برچسب و چندهدف‌بودن را روشن کند. منابعی با استنتاج یا دانش متفاوت را می‌توان افزود، مشروط به اینکه مطالعهٔ حذف مؤلفه، اثر تنوع واقعی را از افزایش صرف فراخوانی جدا کند.

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

ارزیابی روی زبان فارسی، دامنه‌ها و مجموعه‌داده‌های دیگر و چند آزمون مستقل برای مدل‌ها، پایداری نتیجه را روشن می‌کند. افزون بر سنجهٔ نهایی، تغییر الگوی خطا و نیاز به تنظیم دوبارهٔ انتخاب‌گر باید گزارش شود.

۶-۵-۵ کاهش هزینهٔ اجرا

می‌توان مراحل پرهزینه را فقط هنگام عدم‌اطمینان یا ابهام هدف فعال و صرفه‌جویی در فراخوانی را همراه افت یا بهبود عملکرد گزارش کرد. زمان، توکن، حافظه و هزینهٔ سرویس نیز باید در شرایط کنترل‌شده اندازه‌گیری شوند.

۶-۵-۶ ارزیابی انسانی و ریزتنظیم

ردپاها باید از نظر صحت شاهد، ارتباط با هدف، سازگاری قطبیت و وفاداری ارزیابی انسانی شوند. با دادهٔ بزرگ‌تر، ریزتنظیم پارامترکم نیز قابل بررسی است، اما باید تحت تقسیم یکسان با خط پایهٔ پرامپت‌محور مقایسه شود تا خطای برچسب‌های خودتولیدشده پنهان نماند.

۶-۶ نتیجه‌گیری نهایی

افزودن مراحل بیشتر به‌تنهایی تحلیل احساسات ضمنی را بهبود نداد؛ ارزش استدلال زمانی ظاهر شد که منبعی مکمل بود و انتخاب پاسخ از تولید آن جدا شد. چارچوب پیشنهادی اختلاف منابع را ساختارمند و پاسخ نهایی را با سیاستی آموخته‌شده از آموزش تعیین می‌کند. بهبود مشاهده‌شده به تنظیم حاضر مقید است، اما توجه هم‌زمان به شاهد، محدودهٔ هدف و اعتمادپذیری نسبی منابع مسیر روشنی برای ادامهٔ پژوهش فراهم می‌سازد.

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: راهنمای بازتولید آزمایش‌ها

این پیوست مسیر کوتاه بازسازی اعداد، جدول‌ها و شکل‌های گزارش‌شده را مشخص می‌کند. خروجی‌های مورد ارزیابی در مخزن ذخیره شده‌اند؛ پس بازسازی گزارش از اجرای دوبارهٔ استنتاج جدا است.

داده و کلید نمونه

دادهٔ ضمنی پاک‌سازی‌شده شامل ۲۱۸۸ سطر است: ۱۷۴۶ سطر آموزش و ۴۴۲ سطر آزمون. هم‌ترازی فایل‌ها با کلید مرکب زیر کنترل می‌شود:

```
[id, source_sentence_id, sentence, target, from, to, polarity]
```

در مقایسهٔ Gemini، ۱۵۰ سطر آموزش فقط برای تنظیم به کار می‌رود و نتایج گزارش‌شده فقط بر ۹۰ نمونه آزمون مشترک است.

مدل‌ها و تنظیمات تولید

جدول ۱: مدل و تنظیمات اجرای خروجی‌های ذخیره‌شده.

مؤلفه	تنظیم
مدل اصلی	qwen3:8b با Ollama
پیش‌بینی مستقیم	دمای صفر
مسیر استدلالی	دمای ۰/۷؛ برچسب نهایی با دمای صفر
خودسازگاری SC3	سه نمونه‌گیری درون هر پیش‌بینی و رأی اکثریت
مقایسهٔ تکمیلی	Gemini 2.5 Flash روی ۹۰ نمونه آزمون مشترک

جدول ۲: حداقل مراحل بازتولید گزارش.

مرحله	اسکریپت	خروجی	نیاز به اجرای مدل
اعتبارسنجی و جدول	experiments/run_final_pipeline.py	results/final_results_table.csv	خیر
شکل‌های پایان‌نامه	experiments/generate_thesis_result_figures.py	قالب__تمپلیت__پایان_نامه_امیرکبیر_thesis_template_of_Amirkabir/Images/Chapter5	خیر
آزمون سند	tests/test_thesis_finalization.py	گزارش unittest	خیر
ساخت پایان‌نامه	scripts/build_thesis.ps1	قالب__تمپلیت__پایان_نامه_امیرکبیر_thesis_template_of_Amirkabir/AUTthesis.pdf	خیر

مراحل بازتولید

فایل‌های اصلی

- data/processed/semeval14_scapt_isa_only_clean.csv
- data/processed/gemini_model_comparison_subset_train150_test90.csv
- results/direct_isa_predictions.csv
- results/thor_originalish_sc3_isa_predictions.csv
- results/etc_thor_originalish_sc3_guarded_tuned_selected_isa_predictions.csv

فرمان‌های حداقلی

فرمان‌ها از ریشهٔ مخزن در 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 -B -m unittest tests.test_thesis_finalization -v
powershell.exe -NoProfile -ExecutionPolicy Bypass `
  -File .\scripts\build_thesis.ps1
```

بازسازی جدول‌ها و شکل‌ها مدل زبانی را دوباره اجرا نمی‌کند و به اجرای LLM نیاز ندارد؛ اما تولید دوبارهٔ خروجی خام به پشتانهٔ مدل و اجرای اسکریپت‌های استنتاج وابسته است.

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 90 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