دانشگاه صنعتی شریف
دانشکده مهندسی کامپیوتر

پایان‌نامه کارشناسی
مهندسی کامپیوتر

طراحی عامل هوش‌مصنوعی متخصص برای استخراج اطلاعات طرح‌های تأمین مالی جمعی از محتوای HTML

نگارش

زهرا آذر

استاد راهنما

دکتر محمدامین فضلی

شهریور ۱۴۰۴

سپاس

از استاد بزرگوارم که با کمک‌ها و راهنمایی‌های بی‌دریغشان، مرا در به سرانجام رساندن این پایان‌نامه یاری داده‌اند، تشکر و قدردانی می‌کنم. همچنین از همکاران عزیزی که با راهنمایی‌های خود در بهبود نگارش این نوشتار سهیم بوده‌اند، صمیمانه سپاسگزارم.

چکیده

هدف اصلی این پژوهش، طراحی، ساخت و ارزیابی یک **عامل متخصص** برای استخراج اطلاعات ساختارمند با دقت بالا از صفحات وب تأمین مالی جمعی است. برای دستیابی به این هدف، یک فرایند پیشرفته شامل پاک‌سازی هوشمند محتوای HTML، استخراج چندمرحله‌ای و ادغام هوشمند نتایج پیاده‌سازی شد. به منظور سنجش کارایی این عامل، عملکرد آن با دو راهبرد ساده‌تر (یک عامل پایه و یک عامل تابع‌محور) که به عنوان معیار پایه عمل می‌کنند، مقایسه گردید. نتایج تجربی بر اساس داوری انسانی نشان داد که عامل متخصص با دستیابی به دقت **۹۰/۹%**، برتری چشمگیری نسبت به عامل‌های پایه دارد و کارایی راهبرد پیشرفته را اثبات می‌کند. به عنوان یک هدف ثانویه، قابلیت اطمینان الگوواره «مدل زبانی بزرگ به عنوان داور» نیز ارزیابی شد. نتایج نشان داد که داور خودکار با توافق بالا با داوری انسانی (امتیاز F1 بالای ۹۰٪)، ابزاری معتبر برای ارزیابی‌های مقیاس‌پذیر در آینده است.

کلیدواژه‌ها: استخراج اطلاعات از وب، عامل هوش مصنوعی، معماری عامل‌محور، پاک‌سازی HTML، ادغام هوشمند، مدل زبانی بزرگ به عنوان داور، تأمین مالی جمعی.

فهرست مطالب

فهرست جداول

فصل ۱

مقدمه

این پژوهش به طراحی، پیاده‌سازی و ارزیابی یک سامانه هوشمند عامل‌محور برای استخراج خودکار و ساختارمند اطلاعات از صفحات وب «تأمین مالی جمعی» می‌پردازد. با توجه به رشد روزافزون سکوهای تأمین مالی جمعی و نیاز مبرم به تحلیل داده‌های مرتبط با طرح‌های سرمایه‌گذاری، فرآیند دستی استخراج اطلاعات به دلیل زمان‌بر بودن، هزینه بالا و مستعد خطا بودن، دیگر پاسخگوی نیازهای موجود نیست. این گزارش یک راهکار مبتنی بر مدل‌های زبانی بزرگ را معرفی می‌کند که با بهره‌گیری از معماری عامل‌محور، رویکردهای متفاوتی برای استخراج اطلاعات را پیاده‌سازی و مقایسه می‌نماید.

۱-۱ تعریف مسئله

۱-۱-۱ بیان مسئله در سطح کلان

استخراج اطلاعات ساخت‌یافته از صفحات وب غیرساخت‌یافته (HTML) یکی از مسائل کلاسیک و در عین حال چالش‌برانگیز است. تعدد قالب‌ها، وجود نویزهای فراوان (اسکریپت‌ها، استایل‌ها، تبلیغات و ...)، ناهمگنی نشانه‌گذاری‌ها و تفاوت زبان و نگارش محتوا باعث می‌شود تبدیل HTML به داده‌های تمیز و ساختارمند نیازمند رویکردهای هوشمندانه باشد.

۱-۱-۲ بیان مسئله در دامنه خاص

در این پروژه دامنه هدف، سکوهای ایرانی طرح‌های «تأمین مالی جمعی» است. هدف، استخراج خودکار اطلاعات کلیدی هر طرح شامل نام طرح، شرکت متقاضی، سود مورد انتظار، مدت، نوع تضمین، وضعیت و مبلغ سرمایه‌گذاری از محتوای HTML خام صفحه هر طرح است؛ به‌نحوی که خروجی نهایی ساخت‌یافته و قابل تحلیل باشد.

۱-۲ اهمیت موضوع

- **کارایی و مقیاس‌پذیری**: جایگزینی فرایند دستی پرهزینه با پردازش خودکار، امکان پوشش سکوهای متعدد و طرح‌های بسیار را در زمان کمتری فراهم می‌کند.
- **پایایی و یکنواختی**: تبدیل داده‌های ناهمگن به ساختار واحد، تحلیل‌پذیری و مقایسه‌پذیری را افزایش می‌دهد.
- **اهمیت صحت اطلاعات برای تحلیل و انتشار**: دقت استخراج برای تحلیل‌های آتی، گزارش‌دهی عمومی، تصمیم‌سازی سرمایه‌گذاران و حتی انطباق‌های نظارتی حیاتی است؛ خطا در داده‌های پایه می‌تواند به برداشت‌ها و تصمیم‌های اشتباه منجر شود.
- **پایش پیوسته و به موقع اطلاعات**: در پایش دستی اطلاعات، از زمان انتشار اطلاعات تا دسترسی سامانه به اطلاعات مورد نیاز تاخیر زمانی زیادی وجود دارد. در دامنه‌ای مانند دامنه‌ی سرمایه‌گذاری و تامین مالی جمعی، دسترسی سریع و به موقع به طرح‌هایی که تازه منتشره شده‌اند اهمیت پیدا می‌کند. استخراج خودکار اطلاعات می‌تواند زمان دسترسی به اطلاعات را به شدت کاهش دهد.

۱-۳ چالش‌ها

- ناهمگونی شدید قالب و نشانه‌گذاری صفحات وب میان سکوهای مختلف
- زبان فارسی و تغییرات نگارشی/زبانی در متن‌ها
- نویزهای ساختاری (script، style، ads) و محتوایی در HTML
- نیاز به توازن بین دقت استخراج، زمان پردازش و هزینه توکنی

۱-۴ اهداف پژوهش

هدف اصلی این پژوهش، طراحی و ساخت یک **عامل متخصص** با دقت بالا برای استخراج اطلاعات ساختارمند از صفحات وب تأمین مالی جمعی بود. این هدف از طریق ترکیب سه روش کلیدی دنبال شد: (۱) پیش‌پردازش و پاک‌سازی هوشمند محتوای HTML برای کاهش نویز و هزینه، (۲) استخراج اطلاعات در چندین دور با پرامپت‌های متنوع برای افزایش پوشش و دقت، و (۳) پس‌پردازش و ادغام هوشمند نتایج برای دستیابی به یک خروجی نهایی بهینه.

برای ارزیابی میزان اثربخشی عامل متخصص، عملکرد آن با دو راهبرد ساده‌تر (عامل پایه و عامل تابع‌محور) که به عنوان خط پایه عمل می‌کردند، مقایسه شد. این تحلیل مقایسه‌ای، موازنه بین دقت، سرعت و پیچیدگی پیاده‌سازی را به وضوح نشان می‌دهد. به عنوان یک هدف پژوهشی ثانویه، کارایی الگوواره «مدل زبانی بزرگ به عنوان داور» نیز برای اعتبارسنجی خودکار نتایج مورد بررسی قرار گرفت تا پتانسیل آن برای تسریع چرخه‌های ارزیابی در تحقیقات آینده سنجیده شود.

۱-۵ ساختار پایان‌نامه

ساختار این پایان‌نامه به شرح زیر است: فصل دوم به مفاهیم اولیه و تعاریف کلیدی مانند مدل‌های زبانی بزرگ، معماری عامل‌محور و روش‌های ارزیابی می‌پردازد. فصل سوم، با مرور کارهای پیشین در حوزه استخراج اطلاعات از وب، جایگاه این پژوهش را مشخص می‌کند. در فصل چهارم، راه‌حل پیشنهادی شامل معماری فنی، پیاده‌سازی سه عامل استخراج و روش‌شناسی ارزیابی آن‌ها به تفصیل تشریح می‌شود. فصل پنجم نتایج تجربی را تحلیل کرده و عملکرد عامل‌ها و داور خودکار و انسانی را مقایسه می‌کند. در نهایت، فصل ششم به نتیجه‌گیری، جمع‌بندی دستاوردها، بیان محدودیت‌ها و ارائه پیشنهادهایی برای کارهای آینده اختصاص دارد.

فصل ۲

مفاهیم اولیه و تعاریف

در این فصل، مفاهیم بنیادی و چارچوب‌های نظری که اساس معماری و روش‌شناسی این پژوهش را تشکیل می‌دهند، معرفی می‌شوند. تمرکز بر تعریف مفاهیمی است که برای درک عمیق راهکار ارائه‌شده ضروری هستند، پیش از آنکه جزئیات پیاده‌سازی در فصل‌های بعدی مورد بحث قرار گیرند.

۲-۱ طرح تأمین مالی جمعی

تأمین مالی جمعی روشی نوین برای جذب سرمایه است که در آن، کارآفرینان و صاحبان ایده، طرح‌های خود را از طریق سکوهای آنلاین به تعداد زیادی از سرمایه‌گذاران بالقوه عرضه می‌کنند. هر فرد می‌تواند با سرمایه‌گذاری مبالغ کوچک در این طرح‌ها مشارکت کرده و در ازای آن، متناسب با مدل طرح، سود، سهام یا پاداش دریافت کند [۱]. این سازوکار، دسترسی به منابع مالی را برای کسب‌وکارهای نوپا تسهیل کرده و علاوه بر تأمین سرمایه، به عنوان ابزاری برای اعتبارسنجی بازار و ایجاد جامعه اولیه حامیان عمل می‌کند. استخراج دقیق اطلاعات از صفحات این طرح‌ها برای تحلیل بازار و تصمیم‌گیری سرمایه‌گذاران از اهمیت بالایی برخوردار است.

۲-۲ استخراج اطلاعات ساختارمند

استخراج اطلاعات شاخه‌ای از پردازش زبان طبیعی است که هدف آن استخراج خودکار اطلاعات ساختارمند از متون غیرساختارمند است. در چارچوب این پروژه، "اطلاعات ساختارمند" به داده‌هایی اطلاق می‌شود که در یک ساختار از پیش‌تعریف‌شده (مانند یک JSON با کلیدهای مشخص) قرار می‌گیرند

و "متون غیرساختارمند" همان محتوای خام صفحات HTML است. هدف نهایی، تبدیل یک سند HTML پیچیده و پُرنویز به یک رکورد داده تمیز و قابل استفاده در پایگاه داده است.

۲-۳ مدل‌های زبانی بزرگ (LLMs)

مدل‌های زبانی بزرگ، سیستم‌های هوش مصنوعی مبتنی بر شبکه‌های عصبی عمیق هستند که بر روی حجم عظیمی از داده‌های متنی آموزش دیده‌اند. این مدل‌ها که معماری بسیاری از آن‌ها بر پایه ترنسفورمر است [۲]، با روش‌هایی مانند پیش‌آموزشی عمیق [۳، ۴] توانایی‌های قابل توجهی کسب کرده‌اند. ویژگی کلیدی آن‌ها، توانایی درک، تولید، خلاصه‌سازی و استدلال بر روی زبان طبیعی است.

۲-۳-۱ مدل Gemini برای استخراج اطلاعات

خانواده مدل‌های Gemini که توسط گوگل توسعه داده شده [۵]، به طور خاص برای کاربردهای چندوجهی و درک متون طولانی طراحی شده است. مدل‌های این خانواده، به ویژه گونه‌های سبک‌تر مانند Gemini Flash، توازن بسیار خوبی بین هزینه، سرعت و دقت برقرار می‌کنند. قابلیت کلیدی آن‌ها برای این پروژه، پشتیبانی قوی از **فراخوانی تابع** است که به مدل اجازه می‌دهد خروجی خود را مستقیماً در یک ساختار از پیش‌تعریف‌شده و معتبر تولید کند. این ویژگی، Gemini را به ابزاری ایده‌آل برای وظیفه اصلی این پژوهش، یعنی استخراج اطلاعات ساختارمند، تبدیل می‌کند.

۲-۳-۲ مدل Gemma برای ارزیابی

خانواده مدل‌های Gemma نیز از مدل‌های متن‌باز و قدرتمند گوگل هستند [۶]. در این پروژه از مدل Gemma-3-27b-it، که یک نسخه آموزش‌دیده برای پیروی از دستورالعمل است، برای وظیفه **داوری خودکار** استفاده شده است. انتخاب یک مدل نسبتاً سبک‌تر برای این وظیفه عمدی بوده است؛ چرا که وظایف اعتبارسنجی و قضاوت دودویی (صحیح/غلط) معمولاً به اندازه استخراج اولیه نیازمند توان محاسباتی بالا نیستند و استفاده از مدل‌های بهینه‌تر می‌تواند هزینه ارزیابی‌های مقیاس‌بزرگ را به شدت کاهش دهد.

۲-۴ فراخوانی تابع

فراخوانی تابع یک قابلیت پیشرفته در مدل‌های زبانی بزرگ است که به آن‌ها اجازه می‌دهد خروجی خود را به جای متن آزاد، در قالب یک فراخوانی تابع ساختارمند تولید کنند [۷]. در این روش، یک ساختار تابع (شامل نام، توضیحات و پارامترها با نوع مشخص) به مدل ارائه می‌شود. مدل پس از پردازش ورودی، مؤلفه‌های این تابع را با اطلاعات استخراج‌شده پر کرده و یک شیء ساختارمند بازمی‌گرداند. این مکانیزم، پایایی و قابلیت اطمینان خروجی‌های ساختارمند را به شدت افزایش می‌دهد و به کاهش خطای توهم در مدل کمک می‌کند [۸].

۲-۵ معماری عامل‌محور

یک سیستم عامل‌محور، سیستمی است که از مجموعه‌ای از عامل‌های مستقل و هوشمند تشکیل شده است [۹، ۱۰]. هر عامل قادر است محیط خود را درک کرده و برای رسیدن به اهداف مشخصی، به صورت مستقل عمل کند. در این پژوهش، از این معماری برای پیاده‌سازی و مقایسه راهبردهای مختلف استخراج استفاده شده است. هر "عامل" یک رویکرد خاص برای استخراج اطلاعات را کپسوله می‌کند. این طراحی امکان تحلیل مستقل هر رویکرد را فراهم می‌سازد:

- **عامل پایه**: نماینده رویکرد سریع و ساده مبتنی بر پرامپت.
- **عامل تابع‌محور**: نماینده رویکردی که بر پایایی و یکنواختی ساختاری خروجی تمرکز دارد و از قابلیت فراخوانی تابع مدل زبانی استفاده می‌کند.
- **عامل متخصص**: نماینده رویکردی پیچیده‌تر که با پیش‌پردازش و پس‌پردازش، به دنبال حداکثرسازی دقت است.

مقایسه این سه عامل، امکان درک عمیق موازنه بین دقت، سرعت و پیچیدگی پیاده‌سازی را فراهم می‌آورد.

۲-۶ مدل زبانی بزرگ به عنوان داور

ارزیابی عملکرد سیستم‌های استخراج اطلاعات به طور سنتی به داوری انسانی وابسته است که فرآیندی کند، پرهزینه و مقیاس‌ناپذیر است. الگوواره «**مدل زبانی بزرگ به عنوان داور**» یک راهکار نوین برای این

چالش است [۱۱]. در این رویکرد، از یک مدل زبانی توانمند (در این پروژه Gemma) به عنوان یک داور خودکار برای ارزیابی خروجی‌های تولیدشده توسط عامل‌های دیگر استفاده می‌شود. این داور با دریافت متن منبع، خروجی استخراج‌شده و یک معیار ارزیابی دقیق، قضاوت می‌کند که آیا خروجی صحیح است یا خیر. مزایای اصلی این روش شامل مقیاس‌پذیری بالا، سرعت و یکنواختی در ارزیابی است که آن را به ابزاری کارآمد برای پایش مداوم و آزمایش‌های گسترده تبدیل می‌کند.

۷-۲ ماتریس اغتشاش و معیارهای ارزیابی

ماتریس اغتشاش ابزاری استاندارد برای ارزیابی عملکرد سیستم‌های طبقه‌بندی است [۱۲]. در یک مسئله طبقه‌بندی دودویی (مانند تشخیص "صحیح" یا "نادرست" بودن یک فیلد استخراج‌شده)، این ماتریس چهار مقدار کلیدی را نمایش می‌دهد:

- **مثبت واقعی**: تعداد مواردی که به درستی "صحیح" پیش‌بینی شده‌اند.
- **مثبت کاذب**: تعداد مواردی که به اشتباه "صحیح" پیش‌بینی شده‌اند (در حالی که نادرست بوده‌اند).
- **منفی واقعی**: تعداد مواردی که به درستی "نادرست" پیش‌بینی شده‌اند.
- **منفی کاذب**: تعداد مواردی که به اشتباه "نادرست" پیش‌بینی شده‌اند (در حالی که صحیح بوده‌اند).

بر اساس این مقادیر، معیارهای عملکردی مهمی مانند دقت (Accuracy)، صحت (Precision)، بازیابی (Recall) و معیار F1 محاسبه می‌شوند که تصویر جامعی از عملکرد سیستم ارزیابی ارائه می‌دهند.

فصل ۳

مرور کارهای پیشین

استخراج ساختارمند اطلاعات از محتوای وب یک مسئله سنتی و پرچالش در علوم کامپیوتر است که با ظهور مدل‌های زبانی بزرگ، وارد مرحله جدیدی شده است. این فصل به مرور روندهای اصلی پژوهشی مرتبط با این پروژه می‌پردازد، از رویکردهای سنتی مبتنی بر الگو و یادگیری ماشین گرفته تا الگووارهای نوین مبتنی بر مدل زبانی بزرگ، فراخوانی تابع و ارزیابی خودکار.

۳-۱ رویکردهای سنتی در استخراج اطلاعات از وب

پیش از ظهور مدل‌های زبانی بزرگ، استخراج اطلاعات از وب عمدتاً بر سه رویکرد استوار بود. یکی از جامع‌ترین بررسی‌ها در این زمینه توسط فرارا و همکاران انجام شده است [۱۳]. این رویکردها شامل سیستم‌های مبتنی بر قاعده و الگو بودند که از عبارات منظم یا مسیرهای XPath برای شناسایی داده‌ها استفاده می‌کردند. این روش‌ها، که اغلب در قالب «پوشش‌دهنده»ها پیاده‌سازی می‌شدند [۱۴]، برای وب‌سایت‌هایی با ساختار ثابت دقیق و سریع بودند، اما در برابر تغییرات ساختاری وب‌سایت شکننده عمل می‌کردند. رویکرد دیگر، استفاده از یادگیری ماشین نظارت‌شده بود که در آن مدل‌هایی مانند میدان‌های تصادفی شرطی [۱۵] برای یادگیری الگوهای استخراج از روی داده‌های برچسب‌خورده به کار گرفته می‌شدند. چالش اصلی در اینجا، نیاز به حجم بالای داده‌های آموزشی باکیفیت بود. برای کاهش این هزینه، روش‌های یادگیری با نظارت ضعیف یا دور نیز توسعه یافتند که از منابع جانبی برای تولید خودکار داده‌های آموزشی پُرنویز استفاده می‌کردند [۱۶].

۳-۲ تحول با مدل‌های زبانی بزرگ

ظهور مدل‌های زبانی بزرگ، همانطور که در پیمایش جامع ژائو و همکاران تشریح شده است [۱۷]، الگوواره استخراج اطلاعات را متحول کرد. این مدل‌ها با توانایی درک عمیق زبان طبیعی و پیروی از دستورالعمل‌های پیچیده، امکان استخراج اطلاعات را بدون نیاز به آموزش خاص دامنه فراهم کردند. این توانایی‌ها اغلب از طریق یادگیری درون‌ـمتنی بروز می‌یابند [۱۸]. یکی از کلیدی‌ترین قابلیت‌های این مدل‌ها، توانایی تولید خروجی‌های ساختارمند مانند JSON است. این قابلیت، نیاز به طراحی پوشش‌دهنده‌های پیچیده یا برچسب‌گذاری داده‌های انبوه را تا حد زیادی برطرف می‌کند.

۳-۲-۱ فراخوانی تابع

فراخوانی تابع، که توسط شرکت OpenAI به عنوان یک قابلیت کلیدی معرفی شد [۱۹]، یک گام تکاملی مهم در استخراج ساختارمند با مدل‌های زبانی بزرگ است. در این رویکرد، به جای توصیف ساختار خروجی در پرامپت متنی، یک «ابزار» یا «تابع» با پارامترهای مشخص و نوع‌بندی‌شده به مدل معرفی می‌شود. مدل پس از تحلیل متن ورودی، این تابع را با مقادیر استخراج‌شده فراخوانی می‌کند. این مکانیزم پایایی خروجی را افزایش داده، احتمال توهم مدل را کاهش می‌دهد و اعتبارسنجی خودکار را تسهیل می‌کند. در این پروژه، عامل‌های تابع‌محور و متخصص از این قابلیت برای تضمین کیفیت و یکنواختی خروجی بهره می‌برند.

۳-۳ ارزیابی خودکار با مدل زبانی بزرگ به عنوان داور

ارزیابی کیفیت سیستم‌های استخراج اطلاعات به طور سنتی نیازمند قضاوت انسانی است که فرآیندی کند و پرهزینه است. الگوواره «مدل زبانی بزرگ به عنوان داور» یک راهکار نوین برای این چالش است، که چانگ و همکاران در پیمایش خود به تفصیل به آن پرداخته‌اند [۲۰]. در این رویکرد، یک مدل زبانی بزرگ و توانمند وظیفه مقایسه خروجی استخراج‌شده با متن منبع را بر عهده می‌گیرد و بر اساس یک معیار از پیش‌تعریف‌شده، درستی یا نادرستی آن را قضاوت می‌کند. با این حال، باید توجه داشت که این داور نیز ممکن است خطا کند. بنابراین، کالیبره کردن آن با استفاده از یک مجموعه داده طلایی که توسط انسان ارزیابی شده و محاسبه معیارهایی مانند ماتریس اغتشاش برای سنجش توافق بین دو داور، امری ضروری است.

۳-۴ جمع‌بندی

این پروژه در تقاطع چندین حوزه پژوهشی قرار دارد. با بهره‌گیری از توانایی‌های نوین مدل‌های زبانی بزرگ در درک و ساختاردهی اطلاعات و ترکیب آن با تکنیک‌های مهندسی مانند پاک‌سازی داده و ارزیابی خودکار، یک راهکار جامع و عملی برای چالش استخراج اطلاعات از وب ارائه می‌دهد. معماری سه‌عاملی این پروژه نیز امکان بررسی موازنه بین سادگی، پایایی و دقت را فراهم می‌آورد و یک چارچوب آزمایشی غنی برای تحلیل عملکرد رویکردهای مختلف فراهم می‌کند.

فصل ۴

بیان راه‌حل پیشنهادی

در این فصل، جزئیات فنی پیاده‌سازی و معماری سامانه استخراج اطلاعات تشریح می‌شود. سامانه بر یک معماری عامل‌محور استوار است که توسط یک چارچوب ارزیابی جامع برای سنجش عملکرد پشتیبانی می‌شود.

۴-۱ معماری فنی و جزئیات پیاده‌سازی

سامانه از دو بخش اصلی تشکیل شده است: هسته استخراج که وظیفه پردازش HTML و استخراج داده را بر عهده دارد و چارچوب ارزیابی که برای سنجش و مقایسه نتایج به کار می‌رود.

۴-۱-۱ هسته استخراج: عامل‌های هوشمند

سه عامل با راهبردهای متفاوت برای استخراج اطلاعات پیاده‌سازی شده‌اند تا امکان تحلیل موازنه‌های مختلف بین سرعت، دقت و پیچیدگی فراهم شود. الگوهای کامل پرامپت استفاده‌شده برای هر عامل در پیوست ب آمده است.

- **عامل پایه**: این عامل به عنوان خط مبنا عمل می‌کند و از یک پرامپت مستقیم (پیوست ب) برای استخراج اطلاعات در قالب JSON از HTML خام بهره می‌برد. این روش فاقد مرحله پیش‌پردازش یا فراخوانی تابع است و سادگی را در اولویت قرار می‌دهد.
- **عامل تابع‌محور**: این عامل با بهره‌گیری از قابلیت فراخوانی تابع، ساختار داده مورد نظر را به عنوان

یک تابع به مدل معرفی می‌کند تا خروجی ساختارمند و یکنواخت تضمین شود. پرامپت این عامل (پیوست ب) بسیار مختصر است و تنها از مدل می‌خواهد که از تابع معرفی‌شده استفاده کند.

- **عامل متخصص**: این عامل از یک فرایند چندمرحله‌ای برای دستیابی به حداکثر دقت استفاده می‌کند که در ادامه به تفصیل تشریح می‌شود.

۴-۱-۲ جزئیات عملکرد عامل متخصص

عامل متخصص برای دستیابی به بالاترین دقت، فرآیندی سه‌مرحله‌ای را طی می‌کند:

مرحله ۱: پاک‌سازی هوشمند HTML اولین گام، آماده‌سازی محتوای HTML برای پردازش بهینه توسط مدل زبانی است. این مرحله با هدف کاهش نویز و حجم داده ورودی، بدون از دست دادن اطلاعات کلیدی، انجام می‌شود. فرآیند پاک‌سازی شامل حذف کامل تگ‌های <script> و <style>، حذف کامنت‌های HTML، حذف ویژگی‌های غیرضروری مانند class و id که صرفاً برای استایل‌دهی به کار می‌روند و در نهایت، نرمال‌سازی فضاهای خالی برای کاهش تعداد توکن‌های غیرضروری است. این کار نه تنها هزینه فراخوانی مدل را کاهش می‌دهد، بلکه با ارائه یک ورودی تمیزتر، به مدل کمک می‌کند تا بر روی محتوای معنایی تمرکز کند.

مرحله ۲: استخراج چندمرحله‌ای با پرامپت‌های متنوع به جای اتکا به یک بار استخراج، عامل متخصص سه دور استخراج مستقل را با استفاده از سه پرامپت متفاوت (که در پیوست ب آمده‌اند) اجرا می‌کند. هر پرامپت با هدف خاصی طراحی شده است تا مدل را به تمرکز بر جنبه‌های متفاوتی از وظیفه وادارد و از این طریق، احتمال خطا را کاهش داده و جامعیت نتیجه را افزایش دهد. این راهبرد تضمین می‌کند که اگر اطلاعاتی در یک دور به دلیل پیچیدگی خاصی از دید مدل پنهان بماند، در دورهای بعدی با نگاهی متفاوت، شانس استخراج آن افزایش یابد.

- **پرامپت دور اول (استخراج جامع)**: هدف این پرامپت، انجام یک برداشت اولیه و گسترده از اطلاعات است. دستورالعمل‌های آن عمومی هستند و مدل را تشویق می‌کنند تا تمام اطلاعات قابل استخراج را بدون سخت‌گیری بیش از حد پیدا کند. این مرحله مانند یک تور بزرگ عمل می‌کند که ممکن است جزئیاتی را از قلم بیندازد، اما تصویر کلی را به دست می‌آورد و تضمین می‌کند که مقادیر اولیه برای اکثر فیلدها شناسایی شوند.

- **پرامپت دور دوم (تأکید بر دقت و کامل بودن)**: این پرامپت مدل را به بازبینی دقیق‌تر و وسواس‌گونه‌تر وامی‌دارد. با استفاده از کلماتی مانند «با دقت» و «کامل»، به مدل گفته می‌شود که به جزئیات توجه کرده و از استخراج متن‌های خلاصه‌شده یا ناقص خودداری کند. این مرحله برای فیلدهایی که نیازمند متن کامل هستند (مانند بخش تضمین‌ها) یا مقادیری که ممکن است در نگاه اول ناقص دیده شوند، حیاتی است.
- **پرامپت دور سوم (چک‌لیست کیفیت)**: این پرامپت به عنوان یک مرحله بازبینی نهایی عمل می‌کند. با ارائه یک چک‌لیست صریح، مدل موظف می‌شود خروجی خود را بر اساس معیارهای کیفی مشخصی مانند پر بودن فیلدهای اجباری، کامل بودن متن‌ها و صحت واحد اعداد، اعتبارسنجی کند. این پرامپت، احتمال بروز خطاهای رایج مانند مقادیر خالی یا جایگزین را به حداقل می‌رساند و به نوعی مدل را به کنترل کیفیت خروجی خود وامی‌دارد.

مرحله ۳: ادغام هوشمند نتایج پس از اتمام سه دور استخراج، نتایج به دست آمده باید در یک خروجی واحد و بهینه تجمیع شوند. فرآیند ادغام هوشمند به این صورت عمل می‌کند که ابتدا استخراجی که بالاترین امتیاز اطمینان اولیه را دارد به عنوان نتیجه پایه در نظر گرفته می‌شود. سپس، این نتیجه پایه با نتایج دو دور دیگر مقایسه می‌شود. برای هر فیلد، اگر مقدار آن در نتیجه پایه خالی یا دارای نشانه‌های خطا (مانند null یا not found) باشد، با مقدار معتبر از دورهای دیگر جایگزین می‌شود. همچنین، اگر یک دور دیگر متنی طولانی‌تر و کامل‌تر برای یک فیلد متنی پیدا کرده باشد، آن مقدار جایگزین نسخه کوتاه‌تر می‌شود. این فرآیند تضمین می‌کند که خروجی نهایی، کامل‌ترین و صحیح‌ترین اطلاعات ممکن از مجموع سه دور استخراج را در خود جای داده است.

۴-۱-۳ روش‌شناسی ارزیابی

داده‌های مورد استفاده در ارزیابی از سکوهای فعال تأمین مالی جمعی در ایران گردآوری شده‌اند. هر نمونه داده، صفحه وب توضیحات یک طرح واقعی در یکی از این سکوها است. عامل‌های طراحی‌شده، محتوای HTML خام این صفحات را به عنوان ورودی دریافت می‌کنند و تلاش می‌کنند تا اطلاعات کلیدی طرح را در یک ساختار خروجی یکسان استخراج کنند. این ساختار شامل اطلاعاتی نظیر نام طرح، نام متقاضی، سود، مدت و وضعیت طرح است. جزئیات کامل ساختار خروجی مورد انتظار در پیوست آ آمده است.

برای سنجش دقیق عملکرد عامل‌های استخراج، از یک روش‌شناسی ارزیابی دومرحله‌ای استفاده شده است که شامل ارزیابی انسانی به عنوان معیار اصلی و ارزیابی خودکار به عنوان یک پژوهش روش‌شناختی ثانویه است.

ارزیابی انسانی: حقیقت زمینه‌ای مبنای اصلی برای سنجش دقت عامل‌ها، ارزیابی دقیق توسط انسان است. برای این منظور، یک مجموعه داده تهیه شد. در این فرآیند، خروجی هر سه عامل برای تمام پروژه‌های مجموعه آزمایشی به صورت دستی بررسی و صحت هر فیلد استخراج‌شده تأیید یا رد گردید. این مجموعه داده برچسب‌خورده، به عنوان حقیقت زمینه‌ای عمل می‌کند و تمامی معیارهای عملکردی نهایی، از جمله امتیاز F1، بر اساس مقایسه با آن محاسبه شده‌اند. این روش، با وجود زمان‌بر بودن، بالاترین سطح اطمینان را برای ارزیابی عملکرد واقعی عامل‌ها فراهم می‌کند.

ارزیابی خودکار: سنجش کارایی الگوواره داور خودکار به عنوان یک هدف پژوهشی ثانویه، این تحقیق به بررسی کارایی و قابلیت اطمینان الگوواره «مدل زبانی بزرگ به عنوان داور» پرداخت. در این راستا، یک داور خودکار با استفاده از مدل Gemma-3-27b-it پیاده‌سازی شد. وظیفه این داور، مقایسه خودکار خروجی عامل‌ها با پاسخ‌های صحیح از پیش تعیین‌شده و قضاوت در مورد صحت آن‌ها بود. هدف از این بخش، ارزیابی این موضوع بود که آیا می‌توان از یک داور خودکار به عنوان جایگزینی مقیاس‌پذیر برای داوری انسانی در پژوهش‌های آینده استفاده کرد یا خیر. نتایج عملکرد این داور، که در فصل نتایج ارائه شده است، از طریق مقایسه قضاوت‌های آن با حقیقت زمینه‌ای (یعنی داوری انسانی) به دست آمده است تا میزان توافق و قابلیت اطمینان آن به صورت کمی سنجیده شود.

فصل ۵

نتایج تجربی

در این فصل، نتایج به‌دست‌آمده از اجرای سه عامل بر روی مجموعه داده ارزیابی و تحلیل می‌شود. ابتدا عملکرد عامل‌ها در استخراج اطلاعات مقایسه شده و سپس، به عنوان یک تحلیل روش‌شناختی، عملکرد داور خودکار در مقایسه با داوری انسانی سنجیده می‌شود.

۵-۱ تحلیل کارایی پیش‌پردازش: کاهش توکن

یکی از فرضیه‌های اصلی در طراحی عامل متخصص، تأثیر مثبت پاک‌سازی HTML بر کاهش هزینه و افزایش تمرکز مدل بود. جدول ۵-۱ تأثیر این گام را بر تعداد توکن‌های ورودی نشان می‌دهد.

جدول ۵-۱: مقایسه حجم توکن ورودی قبل و بعد از پاک‌سازی HTML

شاخص توکن	مقدار
مجموع توکن خام (ورودی عامل پایه)	۲,۹۲۱,۷۶۲
مجموع توکن پاک‌شده (ورودی عامل متخصص)	۷۶۵,۰۰۶
کاهش کل توکن	**۲,۱۵۶,۷۵۶ (۷۳٫۸۲%)**
میانگین توکن به‌ازای هر پروژه (خام)	۵۶,۱۸۸
میانگین توکن به‌ازای هر پروژه (پاک‌شده)	۱۴,۷۱۲
میانگین کاهش توکن ورودی به ازای هر پروژه	۴۱,۴۷۶

تحلیل نتایج نشان می‌دهد که گام پیش‌پردازش به طور میانگین حجم ورودی را **بیش از ۷۳ %** کاهش داده است. این کاهش چشمگیر نه تنها هزینه فراخوانی مدل‌های زبانی را به شدت کاهش می‌دهد، بلکه با حذف

اطلاعات نامرتبط، به مدل کمک می‌کند تا بر روی محتوای اصلی تمرکز کرده و دقت استخراج را بهبود بخشد، که این موضوع در بخش بعدی مشهود است.

۵-۲ مقایسه عملکرد عامل‌های استخراج

برای ارزیابی و مقایسه، هر سه عامل بر روی مجموعه داده‌ای یکسان شامل ۵۲ نمونه طرح از سکوهای مختلف تأمین مالی جمعی اجرا شدند. جدول ۵-۲ عملکرد سه عامل را بر اساس معیارهای کلیدی دقت (بر اساس داوری انسانی) و میانگین زمان پردازش به ازای هر پروژه مقایسه می‌کند.

جدول ۵-۲: مقایسه عملکرد سه عامل استخراج اطلاعات

عامل	دقت خروجی (درصد)	میانگین زمان پردازش (ثانیه)
عامل پایه	۷۸/۰	۲/۹۳
عامل تابع‌محور	۷۶/۹	۲/۰۶
عامل متخصص	۹۰/۹	۷/۵۹

از نتایج فوق می‌توان چند نکته کلیدی را استنتاج کرد:

- **برتری عامل متخصص**: عامل متخصص با اختلاف قابل توجهی (بیش از ۱۲ درصد نسبت به عامل پایه و ۱۴ درصد نسبت به عامل تابع‌محور) دقیق‌ترین نتایج را تولید کرده است. این برتری مستقیماً به معماری چندمرحله‌ای آن، به ویژه گام پاک‌سازی HTML و ادغام هوشمند نتایج، نسبت داده می‌شود.

- **موازنه سرعت و دقت**: عامل تابع‌محور و عامل پایه دقت تقریباً یکسانی داشتند، اما عامل تابع‌محور به طور قابل توجهی سریع‌تر بود. این ویژگی، عامل تابع‌محور را به گزینه‌ای مناسب برای کاربردهایی تبدیل می‌کند که در آن‌ها سرعت پاسخ‌دهی بر دقت مطلق اولویت دارد.

- **قدرت پیروی از دستورالعمل در مدل**: دقت بالای عامل پایه (٪۷۸)، با وجود عدم استفاده از فراخوانی تابع، نشان‌دهنده قدرت بالای مدل Gemini در پیروی از دستورالعمل‌ها و تولید خروجی با ساختار صحیح است. این نتیجه نشان می‌دهد که حتی بدون سازوکارهای سخت‌گیرانه فراخوانی تابع، مدل قادر است در اکثر موارد، خروجی JSON معتبر و مطابق با الگوی درخواستی تولید کند.

- **هزینه دقت**: افزایش چشمگیر دقت در عامل متخصص با هزینه زمانی بالاتری همراه است. این زمان اضافی صرف مراحل پیش‌پردازش، چندین دور فراخوانی مدل و پس‌پردازش (ادغام) می‌شود.

۵-۳ ارزیابی عملکرد داور خودکار

به عنوان هدف پژوهشی ثانویه، عملکرد داور خودکار در مقایسه با داوری انسانی سنجیده شد تا قابلیت اطمینان آن برای پژوهش‌های آتی مشخص شود. جدول ۵-۳ ابتدا امتیازهای دقت اختصاص‌داده‌شده به هر عامل توسط هر دو داور را نمایش می‌دهد تا میزان هم‌خوانی کلی آن‌ها مشخص شود.

جدول ۵-۳: مقایسه امتیاز دقت هر عامل از دیدگاه داور انسانی و داور خودکار

عامل	امتیاز داور انسانی (درصد)	امتیاز داور خودکار (درصد)
عامل پایه	۷۸/۰	۶۵/۷
عامل تابع‌محور	۷۶/۹	۶۹/۰
عامل متخصص	۹۰/۹	۸۲/۷

این مقایسه نشان می‌دهد که اگرچه داور خودکار در تشخیص عامل برتر (عامل متخصص) با داور انسانی هم‌نظر است، اما به طور کلی تمایل دارد امتیازهای پایین‌تری را ثبت کند.

برای تحلیل دقیق‌تر عملکرد داور خودکار، جدول ۵-۴ ماتریس اغتشاش و معیارهای کلیدی حاصل از مقایسه قضاوت آن با داور انسانی را نشان می‌دهد.

جدول ۵-۴: ماتریس اغتشاش و معیارهای ارزیابی داور خودکار

معیار	مقدار
دقت (Accuracy)	۰/۸۵۴
دقت (Precision)	۰/۹۶۵
بازخوانی (Recall)	۰/۸۵۳
معیار F1	۰/۹۰۵
مثبت‌های درست (TP)	۷۶۳
منفی‌های درست (TN)	۱۶۹
مثبت‌های کاذب (FP)	۲۸
منفی‌های کاذب (FN)	۱۳۲

نتایج ماتریس اغتشاش نشان می‌دهد که داور خودکار با امتیاز F1 بالای ۹۰٪، توافق بالایی با داوری انسانی دارد. این سطح از قابلیت اطمینان، استفاده از آن را به عنوان یک ابزار کارآمد برای ارزیابی سریع و در مقیاس بزرگ در پژوهش‌های آتی توجیه می‌کند.

فصل ۶

نتیجه‌گیری و کارهای آینده

این پژوهش یک معماری عامل‌محور جامع برای استخراج اطلاعات ساختارمند از صفحات وب تأمین مالی جمعی ارائه کرد. تمرکز اصلی بر طراحی و مقایسه سه راهبرد متفاوت استخراج اطلاعات بود که در قالب عامل‌های مستقل پیاده‌سازی شدند. نتایج به وضوح نشان داد که راهبردهای پیشرفته‌تر، مانند پیش‌پردازش هوشمند محتوا و ادغام نتایج چندمرحله‌ای که در عامل متخصص به کار گرفته شد، تأثیر چشمگیری بر افزایش دقت دارند، در حالی که روش‌های ساده‌تر مزایای خود را در سرعت و سهولت پیاده‌سازی حفظ می‌کنند.

۶-۱ جمع‌بندی دستاوردهای اصلی

- **دستیابی به دقت بالا از طریق عامل متخصص**: دستاورد اصلی این پژوهش، طراحی و پیاده‌سازی موفق عامل متخصص است که با دستیابی به دقت **۹۰/۹%**، کارایی خود را در استخراج اطلاعات پیچیده به اثبات رساند. این نتیجه، ارزش ترکیب راهبردهای پیشرفته مانند پاک‌سازی هوشمند HTML، استخراج چندمرحله‌ای و ادغام هوشمند نتایج را تأیید می‌کند.
- **تأثیر حیاتی پیش‌پردازش در بهینه‌سازی**: نشان داده شد که مرحله پاک‌سازی HTML در عامل متخصص، با کاهش بیش از **۷۳%** در تعداد توکن‌های ورودی، نه تنها هزینه محاسباتی را به شدت کاهش می‌دهد، بلکه با حذف نویز، به عنوان یک عامل کلیدی در افزایش دقت نهایی نیز عمل می‌کند.
- **ارائه یک چارچوب تحلیلی برای مقایسه راهبردها**: این پژوهش با مقایسه عامل متخصص با دو

عامل پایه، یک چارچوب کامل برای تحلیل موازنه بین دقت، سرعت و هزینه توکنی ارائه می‌دهد. این چارچوب به تصمیم‌گیری آگاهانه برای انتخاب راهبرد مناسب بر اساس نیازهای خاص هر کاربرد کمک می‌کند.

- **اعتبارسنجی روش‌شناسی ارزیابی مقیاس‌پذیر**: به عنوان یک دستاورد روش‌شناختی، این پژوهش نشان داد که می‌توان از یک داور خودکار مبتنی بر مدل زبانی بزرگ برای ارزیابی کارآمد و در مقیاس بزرگ استفاده کرد. داور خودکار با دستیابی به امتیاز F1 بالای ۹۰٪ در مقایسه با داوری انسانی، قابلیت اطمینان خود را به عنوان ابزاری برای تسریع چرخه‌های آزمایش و ارزیابی در پژوهش‌های آتی به اثبات رساند.

۶-۲ محدودیت‌های پژوهش

علی‌رغم نتایج مثبت، این پژوهش با محدودیت‌هایی نیز همراه بود که مسیر را برای تحقیقات آتی هموار می‌کند:

- **تعمیم‌پذیری به دامنه‌های دیگر**: اگرچه معماری ارائه‌شده انعطاف‌پذیر است، اما عملکرد آن به طور خاص بر روی دامنه تأمین مالی جمعی با زبان فارسی سنجیده شده است. ارزیابی عملکرد آن بر روی دامنه‌های دیگر با ساختارها و زبان‌های متفاوت نیازمند تحقیقات بیشتر است.
- **وابستگی به طراحی پرامپت**: کیفیت نتایج در عامل متخصص تا حد زیادی به کیفیت طراحی پرامپت‌ها وابسته است. مهندسی پرامپت یک فرآیند تکراری است و ممکن است پرامپت‌های بهینه‌تری نیز وجود داشته باشند.

۶-۳ پیشنهادها برای کارهای آینده

بر اساس یافته‌ها و محدودیت‌های این پژوهش، مسیرهای متعددی برای تحقیقات آینده قابل تصور است:

- **توسعه عامل‌های ترکیبی و تطبیق‌پذیر**: طراحی عاملی که بتواند به صورت خودکار و بر اساس پیچیدگی صفحه HTML، بین راهبردهای مختلف (مثلاً سوئیچ از حالت پایه به متخصص) انتخاب کند، می‌تواند موازنه بهینه‌ای بین سرعت و دقت ایجاد کند.

- **بهبود الگوریتم ادغام هوشمند:** می‌توان الگوریتم ادغام در عامل متخصص را با در نظر گرفتن معیارهای پیچیده‌تر، مانند تحلیل معنایی مقادیر استخراج‌شده یا استفاده از مدل‌های زبانی برای قضاوت بین گزینه‌های مختلف، بهبود بخشید.

- **تحلیل عمیق‌تر خطاهای داور خودکار:** بررسی مواردی که داور خودکار با داور انسانی اختلاف نظر دارد (به ویژه موارد مثبت و منفی کاذب) می‌تواند به شناسایی نقاط ضعف مدل داور و بهبود دستورالعمل‌های ارزیابی آن منجر شود.

- **استفاده از مدل‌های چندوجهی:** بررسی استفاده از مدل‌هایی که علاوه بر متن، قادر به درک ساختار بصری صفحه وب نیز هستند، می‌تواند به استخراج اطلاعاتی که صرفاً از طریق تحلیل متنی قابل دستیابی نیستند، کمک کند.

در نهایت، این پژوهش نشان داد که با ترکیب هوشمندانه تکنیک‌های پردازش زبان طبیعی، معماری نرم‌افزار و روش‌شناسی ارزیابی دقیق، می‌توان به راه‌حل‌های کارآمد و قابل اعتمادی برای چالش‌های پیچیده استخراج اطلاعات از وب دست یافت.

Bibliography

[1] E. Mollick. The dynamics of crowdfunding: An exploratory study. *Journal of business venturing*, 29(1):1–16, 2014.

[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[4] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. In *Advances in neural information processing systems*, volume 33, pages 1877–1901, 2020.

[5] Google. Introducing gemini: our largest and most capable ai model. https://blog.google/technology/ai/google-gemini-ai/, December 2023.

[6] Google. Gemma: Open models based on gemini research and technology. https://blog.google/technology/developers/gemma-open-models/, February 2024.

[7] G. Team et al. Gemini: A family of highly capable multimodal models. Technical report, Google, 2023.

[8] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. Bang, A. Madotto, and P. Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023.

[9] M. Wooldridge and N. R. Jennings. Intelligent agents: Theory and practice. *The knowledge engineering review*, 10(2):115–152, 1995.

[10] S. J. Russell and P. Norvig. *Artificial intelligence: a modern approach*. Prentice Hall, 2010.

[11] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Brooks, E. Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*, 2023.

[12] D. M. Powers. Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. *Journal of Machine Learning Technologies*, 2(1):37–63, 2011.

[13] E. Ferrara, P. De Meo, G. Fiumara, and R. Baumgartner. Web data extraction, applications and techniques: A survey. *Knowledge-Based Systems*, 70:301–323, 2014.

[14] N. Kushmerick. Wrapper induction for information extraction. In *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence*, pages 729–744. Morgan Kaufmann Publishers Inc., 1997.

[15] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the eighteenth international conference on machine learning, ICML '01*, pages 282–289, 2001.

[16] M. Mintz, S. Bills, R. Snow, and D. Jurafsky. Distant supervision for relation extraction without labeled data. In *Proceedings of the joint conference of the 47th annual meeting of the ACL and the 4th international joint conference on natural language processing of the AFNLP*, pages 1003–1011, 2009.

[17] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023.

[18] Q. Dong, L. Li, D. Dai, C. Zheng, Z. Wu, B. Chang, X. Sun, J. Li, and Z. Sui. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*, 2023.

[19] OpenAI. Function calling and other api updates. https://openai.com/blog/function-calling-and-other-api-updates, June 2023.

[20] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Zhu, H. Chen, and X. Xie. A survey on evaluation of large language models. *arXiv preprint arXiv:2307.03109*, 2023.

واژه‌نامه

الف

الگوواره ..... Paradigm

معماری عامل‌محور ... Agent-based Architecture

استخراج اطلاعات ..... Information Extraction

ادغام هوشمند ..... Intelligent Merging

صحت ..... Accuracy

امتیاز اف‌ـ۱ ..... F1-Score

پ

مهندسی پرامپت ..... Prompt Engineering

پاک‌سازی اچ‌تی‌ام‌ال ..... HTML Cleaning

دقت ..... Precision

ت

عامل‌های ترکیبی و تطبیق‌پذیر ... Hybrid/Adaptive Agents

توکن‌سازی ..... Tokenization

مثبت درست ..... True Positive (TP)

منفی درست ..... True Negative (TN)

تامین مالی جمعی ..... Crowdfunding

ج

فراخوانی تابع ..... Function Calling

خ

مثبت کاذب ..... False Positive (FP)

منفی کاذب ..... False Negative (FN)

ح

حقیقت زمینه‌ای ..... Ground Truth

د

مجموعه داده استاندارد ..... Benchmark Dataset

ر

بازخوانی ..... Recall

ط

تأمین مالی جمعی ..... Crowdfunding

ع

عامل متخصص ..... Expert Agent

عامل تابع‌محور ..... Function-based Agent

عامل پایه ..... Basic Agent

م

مدل زبانی بزرگ . Large Language Model (LLM)
مدل زبانی بزرگ به عنوان داور ... LLM-as-a-Judge
مدل‌های چندوجهی ..... Multimodal Models

ماتریس اغتشاش ..... Confusion Matrix
مؤلفه‌ی تابع ..... Function Parameter

پیوست آ

قالب پیکربندی استخراج

در این پیوست، ساختار فایل پیکربندی که برای تعریف فیلدهای هدف جهت استخراج اطلاعات به کار می‌رود، تشریح شده است.

این محتوا به عامل‌ها کمک می‌کند تا ساختار دقیق خروجی را درک کرده و اعتبارسنجی‌های لازم را انجام دهند. نمونه‌ای از ساختار این فایل در ادامه آمده است:

```
{
  "function_name": "extract_single_project",
  "object_name": "project",
  "fields": {
    "name": {"type": "string", "required": true},
    "company": {"type": "string", "required": true},
    "profit": {"type": "number", "required": true},
    "guarantee": {"type": "string", "required": true},
    "investment_amount": {"type": "string"},
    "duration": {"type": "string"},
    "status": {"type": "string"}
  }
}
```

در ادامه، توضیح هر یک از این فیلدهای اطلاعاتی آمده است:

- name: عنوان رسمی طرح تأمین مالی.
- company: نام شرکت، نهاد یا شخصی که متقاضی جذب سرمایه است.
- profit: نرخ سود پیش‌بینی‌شده برای سرمایه‌گذاران که معمولاً به صورت درصد بیان می‌شود.
- guarantee: نوع و جزئیات تضمین بازپرداخت اصل و سود سرمایه (مانند ضمانت‌نامه بانکی، بیمه و ...).
- investment amount: مبلغ کل سرمایه مورد نیاز طرح یا حداقل مبلغ قابل سرمایه‌گذاری.
- duration: مدت زمان اجرای طرح یا دوره بازپرداخت سرمایه.
- status: وضعیت فعلی طرح (مانند در حال جذب سرمایه، تکمیل‌شده، فعال).

پیوست ب

الگوهای کامل پرامپت

در این پیوست، الگوهای کامل پرامپت استفاده‌شده برای هر یک از سه عامل استخراج اطلاعات به منظور شفاف‌سازی و بازتولیدپذیری پژوهش ارائه شده است.

ب‌ـ۱ پرامپت عامل پایه

Basic Agent Prompt

```
Extract the {object_description} from the following HTML content.

TARGET FIELDS TO EXTRACT:
- name: The title or name of the project (required)
- company: The company associated with the project (required)
- profit: A number between 0 and 100 representing the profit
 percentage (required)
- guarantee: Description of guarantees offered (required)
- investment_amount: Minimum or total investment amount required
- duration: Project duration or investment period
- status: Current status of the project (active, completed, pending,
 etc.)

Please provide the extracted data in this JSON format:
{{
    "{object_name}": {{
        "field_name": "extracted_value_or_null"
    }}
}}
- Return only the JSON response, no additional text

HTML CONTENT:
{html_content}
```

ب‌ـ۲ پرامپت عامل تابع‌محور

Function-oriented Agent Prompt

```
Use the function `{function_name}` to return the {object_description}
from the following HTML. Only use the function.

{html_content}
```

ب‌ـ۳ پرامپت‌های عامل متخصص

ب‌ـ۳-۱ پرامپت دور اول: استخراج جامع

Expert Agent: Round 1 (Comprehensive Extraction)

```
Extract the {object_description} from the HTML using the
`{function_name}` function.

INSTRUCTIONS:
- Extract ALL available information for each field
- Use exact text from the HTML when possible
- If a field is not found, set it to null
- Pay special attention to numerical values (preserve formatting and
 units)
- Look for complete text in guarantee/description fields

HTML:
{html_content}
```

ب_۳-۲ پرامپت دور دوم: دقت و کامل بودن

Expert Agent: Round 2 (Accuracy and Completeness)

```
Carefully extract the {object_description} using
`{function_name}`. Focus on COMPLETENESS and ACCURACY.


EXTRACTION STRATEGY:
- Scan the ENTIRE HTML content systematically
- Look for both visible text and data attributes
- For text fields, capture the FULL text, not abbreviated versions
- For numerical fields, preserve original formatting and units
- Double-check each field before finalizing


HTML:
{html_content}
```

ب_۳-۳ پرامپت دور سوم: چک‌لیست کیفیت

Expert Agent: Round 3 (Quality Checklist)

```
Perform a THOROUGH extraction of {object_description} using
 `{function_name}`.

QUALITY CHECKLIST:
  All required fields are extracted
  Text fields contain complete, untruncated information
  Numbers include proper units and formatting
  No placeholder or generic values
  Information matches what's actually in the HTML

Be extremely careful and methodical. Extract exactly what you see in
 the HTML.

HTML:
{html_content}
```

Abstract

This research pursues two primary objectives. The first is to design and evaluate an agent-based system for the **structured information extraction** of crowdfunding projects from HTML content. To this end, three distinct strategies are implemented and compared as independent agents (Basic, Function, and Expert) based on human-annotated ground truth. Experimental results demonstrate the significant superiority of the Expert agent, which achieves **90.9%** accuracy by leveraging HTML cleaning and intelligent merging. The second objective is to assess the efficacy of the "LLM-as-a-Judge" paradigm. For this, an automated judge based on the lightweight Gemma model was implemented and its performance was compared against human evaluation. The results show that the automated judge is highly reliable for future research, achieving an F1-score above 90%. In summary, this study provides a comprehensive analysis of trade-offs in extraction strategies while also validating the use of an automated judge for such validation tasks.

Keywords: Information Extraction, HTML, Agent-based Architecture, LLM-as-a-Judge, Model Evaluation, Crowdfunding.

Sharif University of Technology

Department of Computer Engineering

B.Sc. Thesis

Designing an Expert AI Agent for Extracting Crowdfunding Project Information from HTML

By:

Zahra Azar

Supervisor:

Dr. MohammadAmin Fazli

September 2025