DOR: 20.1001.1.2538161.1404.10.2.3.1

DOI: 10.22034/csj.2025.520613.1091

تابستان ۱۴۰۴، دوره ۱۰، شماره ۲، پیاپی ۳۷

دریافت مقاله: ۱۴۰۴/۰۲/۱۲

پذیرش مقاله: ۱۴۰۴/۰۳/۱۳

نوع مقاله: پژوهشی

شناسایی فاکتورهای سبک زندگی در متون زیستی پزشکی با استفاده از مدل‌های زبانی بزرگ

اسماعیل نورانی*

استادیار دانشکده فناوری اطلاعات و مهندسی کامپیوتر دانشگاه شهید مدنی آذربایجان، تبریز، ایران

پست الکترونیکی: ac.nourani@azaruniv.ac.ir

چکیده

شناسایی موجودیت‌های نام‌گذاری شده (NER) در متون علمی، یکی از چالش‌های کلیدی در پردازش زبان طبیعی است. روش‌های سنتی NER، مانند مدل‌های مبتنی بر یادگیری ماشین نظارت شده، به حجم بالایی از داده‌های برچسب‌گذاری شده نیاز دارند که تهیه آن‌ها بسیار زمان‌بر و پرهزینه است. در مقابل، مدل‌های زبانی بزرگ (LLMs)، با کمترین نیاز به داده‌های برچسب‌گذاری شده، فرصتی برای کاربردهای پر هزینه قبلی فراهم آورده‌اند. شناسایی عوامل سبک زندگی در میلیون‌ها مقاله علمی گذشته، یکی از این کاربردهای مغفول مانده است که می‌تواند موجب سازماندهی دانش موجود در زمینه ارتباط بین بیماری‌ها و سبک زندگی شده و زمینه را برای استفاده از چنین پایگاه دانشی در سیاست‌های سلامت عمومی و یا حتی درمان‌های پزشکی فراهم آورد. در این مقاله برای نخستین بار یک راهکار سه مرحله‌ای مبتنی بر LLM به نام LSF-NER توسعه داده شده است که فاکتورهای سبک زندگی را از متون زیستی پزشکی استخراج می‌کند. ارزیابی نتایج آزمایش‌ها نشان می‌دهد که مدل آموزش دیده در این تحقیق، علی‌رغم استفاده از منابع محاسباتی کمتر، عملکردی قابل مقایسه و امیدوارکننده نسبت به مدل‌هایی مانند GPT-4o از خود نشان داده است و نوید امکان‌پذیر بودن ادامه تحقیقات بدون نیاز به زیرساخت‌های محاسباتی پر هزینه را می‌دهد.

کلمات کلیدی: متن کاوی، شناسایی فاکتورهای سبک زندگی، مدل‌های زبانی بزرگ، NER، یادگیری ماشین

۱- مقدمه

شناسایی موجودیت‌های نام‌گذاری شده (NER)[1] یکی از زیرشاخه‌های اساسی در پردازش زبان طبیعی است که شامل شناسایی و دسته‌بندی موجودیت‌ها در متن می‌شود. رویکردهای سنتی NER، مانند روش‌های مبتنی بر قواعد با چالش‌هایی در تعمیم‌پذیری با حوزه‌ها و زبان‌های مختلف روبه‌رو بوده‌اند. هرچند رویکردهای یادگیری عمیق با توانایی درک وابستگی‌های متنی، عملکرد بهتری نسبت به روش‌های سنتی یادگیری ماشین ارائه می‌دهند، اما نیاز به داده‌های برچسب‌گذاری‌شده گسترده دارند [۱-۳]. هر چند تلاش‌هایی صورت گرفته تا از طریق انتقال دانش بین کاربردهای مختلف این محدودیت برطرف شده و نیاز به داده برچسب دار کمتر شود [٤، ٥] ولی توفیق قابل‌توجهی حاصل نشده است.

1-Named Entity Recognition (NER)

* نویسندهٔ مسئول

ظهـــور مدل‌های زبانی بزرگ[۲]، ماننـــد GPT-3، افق‌های جدیدی برای بهبود قابلیت‌های NER باز کرده است[٦]. این مدل‌ها، با پیش‌آموزش گســـترده، پتانسیل قابل توجهی در بهبود عملکرد NER نشـــان داده‌اند و در برخی موارد حتی موفق به دســـتیابی به بیش از ۱۰۰ درصد افزایش در دقت نسبت به روش‌های سنتی شده‌اند[۷]. البته این زمینه، خالی از چالـــش نبوده و نکات مهمی را در این خصوص باید در نظـــر گرفت. مدل‌های زبانی بـــزرگ عمدتاً برای تولید متن طراحی شده‌اند، که با ماهیت برچسب‌گذاری توالی در NER متفاوت است و این تفاوت اساسی می‌تواند منجر به عملکرد غیر بهینه شـــود [۸]. همچنین LLMها منابع محاســـباتی و ذخیره‌ســـازی قابل‌توجهی نیاز دارند، کـــه می‌تواند مانعی برای پیاده‌ســـازی آنهـــا، به‌ویژه در محیط‌هـــای با منابع محدود، باشـــد. انـــدازه بزرگ این مدل‌هـــا همچنین فرآیند آموزش و یا حتی تطبیق آنهـــا برای کاربردهای خاص را پیچیده می‌کند [۸].

چالـــش دیگر این که توســـعه NER بـــرای کاربردهای خاص به دلیل داده‌های محدود و نیاز به تطبیق LLM با آن کاربرد خاص ممکن است دچار مشکل شود. این چالش در سناریوهای بین رشته‌ای که داده‌های برچسب‌گذاری‌شده کمیاب اســـت و نیاز به تطبیق و تنظیـــم دقیق[۳] LLM وجود دارد، تشـــدید می‌شـــود. در نهایت چالش ذاتی LLMها که می‌تواند جواب نادرست ولی با اعتماد به نفس بالا تولید کند نیز وجود دارد که در کاربرد جاری می‌تواند با روش‌هایی مانند راستی‌آزمایی تا حدود زیادی برطرف شود.

بـــه طورکلی مدل‌هـــای زبانی بـــزرگ دو تکنیک اصلی برای بهبود شناســـایی موجودیت‌های نام‌گذاری‌شده ارائه می‌دهنـــد: یادگیری در زمینه[۴] و تنظیـــم دقیق یادگیری در زمینه امکان گرفتن جواب سریع را بدون تغییرات قابل توجه در مدل فراهـــم می‌کند، در حالی که تنظیم دقیق بهبودهای چشمگیری، به‌ویژه در کاربردهای خاص، ارائه می‌دهد. به عنوان مثال، در حوزه زیستی پزشکی، مدل‌های LLM برای تبدیل NER از یک وظیفه برچسب‌گذاری توالی به یک وظیفه تولید متن تطبیق داده شـــده‌اند و موفق به دســـتیابی دقت بالاتری در مقایســـه با قابلیت‌های یادگیری چندنمونه‌ای[۵] مدل‌هایی مانند GPT-4 شده‌اند [۹].

البته باید در نظر گرفت که چنین بهبودی به قیمت تلاش برای آموزش مدل‌های زبانی بزرگ میسر شده که جدا از نیاز به تهیه داده برچســـب‌دار برای آموزش مدل، نیاز به زیرساخت ســـخت‌افزاری گزافی خواهد داشت. لذا در این مقاله روشی ارایه شـــده که با حداقل سخت‌افزار و با اتکا بر دستور متنی[٦] (فرایند ایجاد و پالایش دستورالعمل‌هایی که به مدل‌های هوش مصنوعی داده می‌شود تا دقیق‌ترین پاسخ‌ها به دســـت آید.)، یادگیری در زمینه و نهایتا تنظیم دقیق LLM، اقدام به شناســـایی موجودیت‌های متن می‌کند. به طور خاص در این تحقیق برای شناسایی موجودیت‌هایی در متون زیســـتی پزشـــکی اقدام شده اســـت که به‌عنوان فاکتورهایی موثر در سبک زندگی مطرح می‌شوند.

ســـبک زندگی نقش مهمی در سلامـــت ایفا می‌کند و بر سلامت جســـمی و روانی تأثیرگذار اســـت. شناسایی این عوامل در مقالات زیستی پزشکی برای ساختاردهی دانش موجود و هموار کردن راه برای کشـــفیات جدید ضروری اســـت. ســـبک زندگی ســـالم، خطر ابتلا به دیابـــت نوع ۲، بیماری‌های قلبی- عروقی و مرگ‌ومیر کلی را به‌طور قابل توجهی کاهش می‌دهد [۱۰، ۱۱]. از طرفی انجام فعالیت‌های بدنی، مشارکت اجتماعی و رژیم غذایی سالم با سلامت مغز مرتبط است [۱۲]. عوامل سبک زندگی تأثیرات قابل‌توجهی حتی بر سلامت روان دارند، به‌عنوان مثال خواب نامناسب به‌عنوان یک عامل خطر شناخته شده برای اختلالات روانی محسوب می‌شود [۱۳].

با توجه به هرم جمعیتی کشـــور و پیش‌بینی وضعیت آینده، طی چند دهۀ آتی ترکیب جمعیتی به ســـمت افزایش میانگین سنی و پدیده پیری جمعیت حرکت خواهد کرد. لذا عادات ســـالم ســـبک زندگی در میانسالی، مانند حفظ وزن

2-Large Language Models - LLMs
3-Fine Tuning
4-In-context learning
5-Few-shot Learning
6-Prompt Engineering

مناسب، عدم سیگار کشیدن و فعالیت بدنی منظم، به پیری سالم کمک می‌کنند [۱۴] و وضعیت سیستم درمانی کشور را از بحران پیش رو رهایی می‌بخشد.

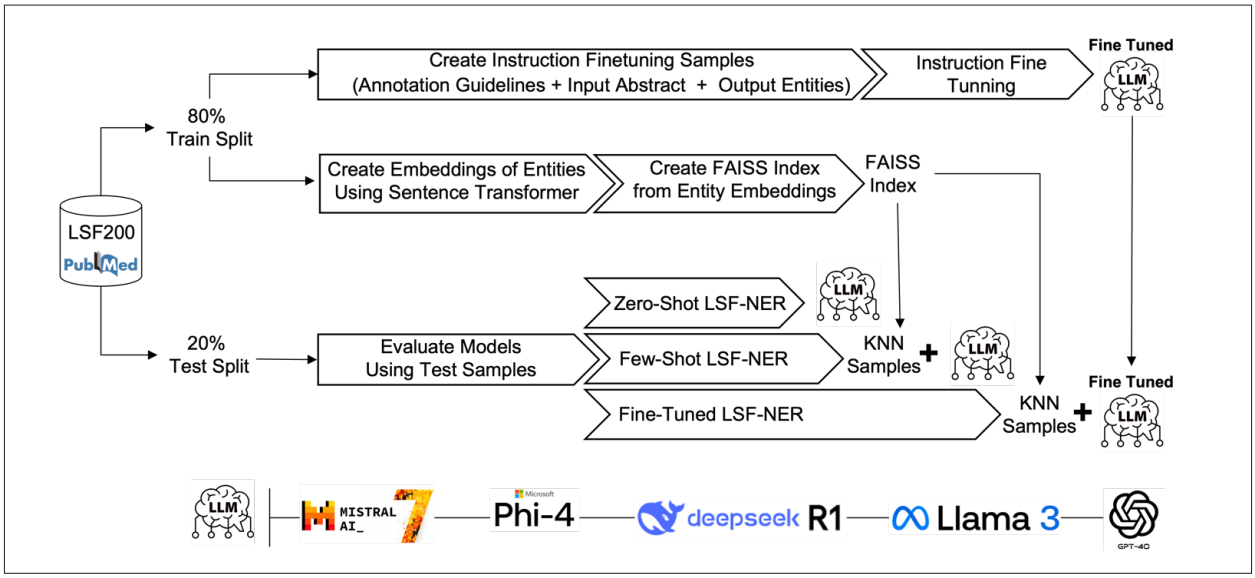
مواردی که ذکر شد، تنها چند مثال محدود از اهمیت سبک زندگی و تاثیر آن بر بیماری‌ها بود و ناگفته پیداست که این تاثیرات در مقالات علمی پزشکی سالیان اخیر به وفور مورد بررسی قرار گرفته است. در ادامه به اهمیت پرداختن به مطالعات قبلی و نظم‌دهی به دانش موجود می‌پردازیم. شناسایی و طبقه‌بندی سیستماتیک عوامل سبک زندگی در میلیون‌ها مقاله زیستی پزشکی به سازماندهی دانش موجود کمک می‌کند و آن را برای تحقیقات بیشتر و کاربرد در سیاست‌های سلامت عمومی قابل دسترس می‌سازد. همچنین شناسایی عوامل سبک زندگی در مقالات علمی گذشته می‌تواند شکاف‌های موجود در تحقیقات کنونی را برجسته کرده و آغازگر مطالعات و نوآوری‌های جدید در درمان‌های پزشکی از طریق درک تاثیر متقابل پیچیده بین عوامل سبک زندگی و بیماری‌ها باشد]۱۵[. در نهایت شناسایی عوامل سبک زندگی در مقالات پزشکی سبب پشتیبانی از توسعه دستورالعمل‌ها و درمان‌های مبتنی بر شواهد شده و به ارائه‌دهندگان خدمات سلامت کمک می‌کند تا اصلاحات سبک زندگی را برای بهبود نتایج درمانی توصیه کنند[۱۶].

نکته حایز اهمیت و عجیب آن است که بر خلاف دلایل ذکر شده و اهمیت غیرقابل انکار سبک زندگی، همچنان فاکتورهای سبک زندگی و نحوه تاثیر آنها بر بیماری‌ها از مقالات علمی پزشکی در قالب یک پایگاه داده و یا پایگاه دانش منسجم استخراج نشده است [۱۷]. تحقیق جاری در جهت توسعه چنین پایگاه دانشی بوده و قدم اول را بر می‌دارد که همان شناسایی فاکتورهای سبک زندگی در مقالات زیستی پزشکی می‌باشد.

در استخراج فاکتورهای سبک زندگی، چالش‌های متعددی وجود دارد. تنوع و پیچیدگی این فاکتورها، که به‌طور سیستماتیک در متون علمی پزشکی توصیف نمی‌شوند، شناسایی آن‌ها را برای سیستم‌های NER دشوار می‌کند. به نحوی که بسیاری از این اصطلاحات می‌توانند در متون مختلف، معانی متفاوتی داشته باشند. لذا جهت حل مشکل می‌توان به راه‌کارهایی که از فهم متن و موضوع بهره می‌برند و در کاربردهای دیگر موفق ظاهر شده‌اند استفاده کرد. به‌عنوان مثال تشخیص چندین مفهوم زیستی مبتنی بر دستور متنی[۷] با بهره‌گیری از اطلاعات زمینه‌ای، در مدل BioPRO منجر به ارایه یک مدل NER موفق شده است [۱۸]. فقدان داده‌های برچسب‌گذاری‌شده نیز مزید علت بوده تا در زمینه کشف فاکتورهای زیستی از متون پزشکی شاهد پیشرفت محسوسی در سالیان اخیر نباشیم که با ارایه اولین مجموعه داده به نام LSF200 در یکی از تحقیقات اخیر [۱۹] تا حدودی مرتفع شده است و زمینه را برای تحقیقات بعدی فراهم کرده است.

تحقیق جاری از این نظر حایز اهمیت است که با بهره‌گیری از تحولات اخیر در حوزه پردازش متن به سراغ یکی از زمینه‌های مغفول در تحقیقات پزشکی رفته که در واقع هموار کردن مسیر برای قدم‌های بعدی در زمینه ایجاد پایگاه‌های دانش ساخت‌یافته در خصوص ارتباط سبک زندگی با بیماری‌هاست. در این تحقیق بدون نیاز به داده‌های برچسب‌دار در مقیاس بزرگ که به دلیل نیاز به اختصاص زمان زیاد و تخصص انسانی همواره به عنوان چالش مطرح بوده و تنها با مجموعه داده بسیار محدود LSF200، مدل زبانی بزرگی آماده استخراج موجودیت‌های مرتبط به سبک زندگی شده است. لذا نوآوری‌های اصلی تحقیق جاری عبارت است از تنظیم دقیق[۸] مدل‌های زبانی بزرگ با استفاده از نمونه‌های بسیار محدود، استفاده از روش K نزدیکترین همسایه[۹] برای پیدا کردن شبیه‌ترین نمونه‌ها جهت استفاده در یادگیری چندنمونه‌ای و در نهایت به کارگیری این روش‌ها برای اولین بار برای استخراج طیف گسترده‌ای از جوانب مختلف سبک زندگی از متون علمی پزشکی می‌باشد.

7-Prompt
8-Fine Tuning
9-K-Nearest Neighbor (KNN)



شکل ۱. کلیات راهکار پیشنهادی (LSF-NER) برای استخراج فاکتورهای سبک زندگی از متون علمی

۲- روش پیشنهادی

در این مطالعه، برای شناسایی فاکتورهای سبک زندگی در متون علمی پزشکی، یک راهکار نوین سه‌نسخه‌ای به نام LSF-NER پیشنهاد شده است که مبتنی بر مدل‌های زبانی بزرگ بوده و اولین روشی است که برای استخراج فاکتورهای سبک زندگی از متون مختلف از LLM استفاده می‌کند. این رویکرد شامل نسخه‌های یادگیری بدون نمونه[10]، یادگیری با چند نمونه، و تنظیم دقیق مدل است. در شکل ۱ کلیات روش پیشنهادی و تنوع LLMهای مورد استفاده نمایش داده شده است و در ادامه هر یک از اجزای LSF-NER به تفصیل توضیح داده خواهدشد.

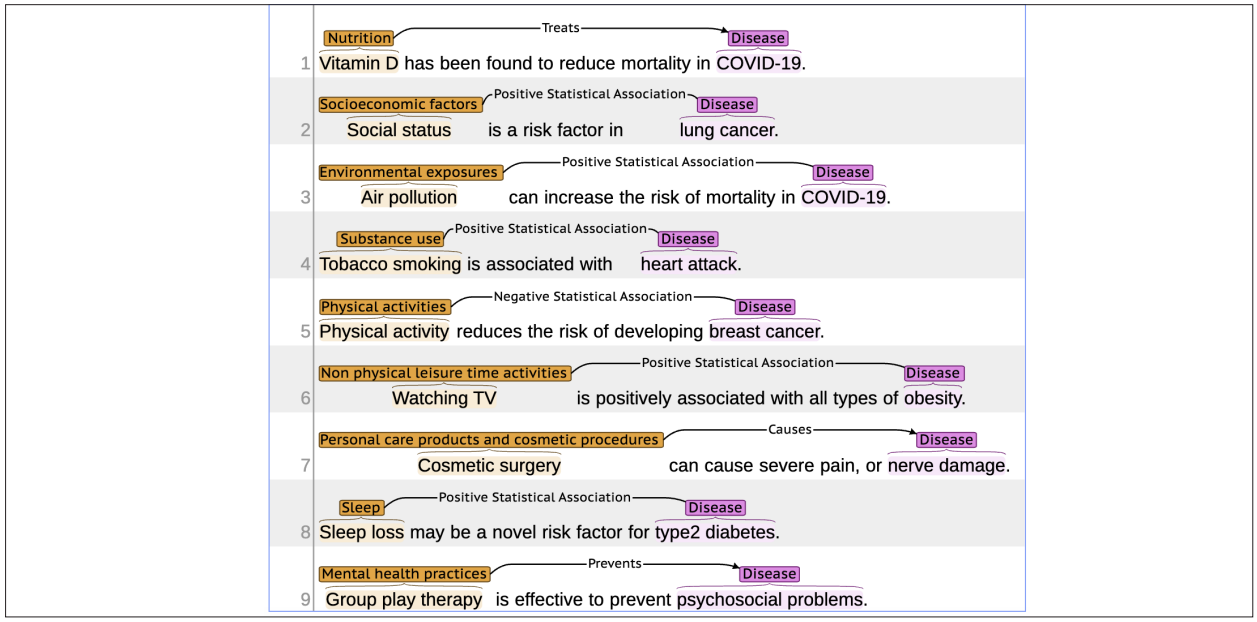
هر چند که مزیت اصلی استفاده از LLMها مرتفع کردن نیاز به داده‌های برچسب‌دار در مقیاس بزرگ می‌باشد ولی همچنان نیاز به مجموعه داده‌های کوچکی از مثال‌های برچسب‌دار جهت استفاده در دستور متنی به‌عنوان مثال برای یادگیری با چند نمونه، ارزیابی عملکرد LLM و در نهایت آموزش LLM در قالب تنظیم دقیق مبتنی بر دستور[11] وجود دارد. در این مطالعه از مجموعه‌داده‌های ایجاد شده در مطالعه قبلی [۱۹] به نام LSF200 استفاده شده است که مشخصات آن درجدول ۱ قابل مشاهده است.

جدول ۱. مشخصات مجموعه‌داده‌های مورد استفاده در این مطالعه (LSF200) جهت طراحی مثال، ارزیابی و آموزش مدل

دسته‌بندی	درصد	تعداد موجودیت غیرتکراری	تعداد موجودیت کل
تغذیه	٪ ۲۷/۳۶	۲۷۰	۴۴۵
عوامل اقتصادی- اجتماعی	٪ ۱۷/۶۳	۱۷۴	۲۷۵
مواجه‌های محیطی	٪۱۳/۳۷	۱۳۲	۱۹۹
مصرف مواد	٪ ۷/۶۰	۷۵	۱۸۷
فعالیت‌های بدنی	٪ ۸/۹۲	۸۸	۲۲۲
مراقبت‌های شخصی و زیبایی	٪ ۳/۵۵	۳۵	۵۶
فعالیت‌های فراغتی غیر فیزیکی	٪ ۲/۷۴	۲۷	۶۱
خواب	٪ ۵/۷۸	۵۷	۱۱۰
تمرینات سلامت روان	٪ ۲/۹۴	۲۹	۷۴
موجودیت‌های متفرقه نامرتبط	٪ ۱۰/۱۳	۱۰۰	۲۴۷
مجموع	٪ ۱۰۰	۹۸۷	۱۸۷۶

مجموعه‌داده‌های LSF200 در برگیرنده چکیده ۲۰۰ مقاله از پایگاه داده PubMed به همراه موجودیت‌های برچسب‌گذاری شده است. این موجودیت‌ها شامل ۹ دسته از فاکتورهای مختلف سبک زندگی می‌باشد که رویهم رفته ۹۸۷ فاکتور منحصر به فرد و تاثیرگذار سبک زندگی را در بر می‌گیرد. متون برچسب‌گذاری شده در این

10-Zero-Shot Learning
11-Instruction Fine-Tuning



شکل ۲. نحوه برچسب‌گذاری فاکتورهای سبک زندگی و ارتباطشان با بیماری‌ها در محیط BRAT

مجموعه‌داده‌ها به دو بخش آموزش و آزمایش به ترتیب ۸۰ درصد و ۲۰ درصد تقسیم شده و در این مقاله مورد استفاده قرار گرفته است. به نحوی که نمونه‌های موجود در بخش آموزش برای دو هدف طراحی مثال برای دستور متنی در یادگیری چندنمونه‌ای و همچنین جهت تنظیم دقیق LLM مبتنی بر دستور جهت انطباق با کاربرد NER استفاده خواهد شد. از نمونه‌های موجود در بخش آزمایش، جهت ارزیابی دقت پاسخ‌های LLM که همان فاکتورهای سبک زندگی کشف شده از متن است استفاده شده است. جهت برچسب‌گذاری مجموعه‌داده‌های فوق از ابزار BRAT [۲۰] استفاده شده است و نمونه‌ای از محیط این ابزار برای برچسب‌گذاری دسته‌های مختلف فاکتورهای سبک زندگی و چگونگی ارتباط بین آنها و بیماری‌های مختلف را در شکل ۲ مشاهده می‌کنید.

۲-۱- یادگیری بدون نمونه

در مرحله ابتدایی، از قابلیت‌های LLM برای انجام برچسب‌گذاری بدون ارائه هیچ نمونه‌ای از داده‌های برچسب‌گذاری شده استفاده می‌شود. در این حالت، مدل تنها بر اساس دستورالعمل‌های برچسب‌گذاری[۱۲] که برای آموزش برچسب‌زن‌های انسانی نیز کاربرد دارند، عمل می‌کند. بدین معنی که حتی انسان نیز برای آن که متنی را دریافت و به صورت صریح بداند که هدف اصلی، برچسب‌زنی و استخراج چه عباراتی است باید قوانین و شرایط خاص را بداند و به صورت هماهنگ اقدام به برچسب‌زنی کند. این مطالعه [۱۹] برای اولین بار چنین قوانینی را به صورت جامع برای استخراج ۹ جنبه مختلف سبک زندگی آماده کرده[۱۳] که در تحقیق جاری از آن استفاده شده و دستور متنی مرحله اول تولید شده است. در این مطالعه، برای استخراج موجودیت‌های مرتبط با سبک زندگی از متون علمی، یک سامانه شناسایی موجودیت‌های نامدار (NER) مبتنی بر LLM ارایه شده است و بدون شک نخستین قدم در استفاده از LLMها طراحی یک دستور متنی است که بتواند انتظاراتمان را بدون ابهام و دقیق مشخص کند. در ادامه به بخش‌های مختلف دستور متنی طراحی شده در شکل ۳ می‌پردازیم. این راهنما مشخص می‌کند که منظور ما از سبک زندگی چیست و چه جنبه‌هایی از سبک زندگی مدنظر این مطالعه خواهد بود و همچنین قالب خروجی مورد انتظار چیست.

12-Annotation Guidelines

13-https://esmaeilnourani.github.io/lifestylefactors-annotation-docs/entities

جنبه‌های مختلف سبک زندگی مدنظر در این مطالعه: عوامل سبک زندگی شامل جنبه‌های مختلفی هستند که می‌توانند بر سلامت و رفاه تأثیرگذار باشند. این دسته‌ها عبارتند از:

۱- **تغذیه**: رژیم‌های غذایی، گروه‌های غذایی، روش‌های آماده‌سازی غذا، مکمل‌های غذایی و غیره

۲- **عوامل اجتماعی-اقتصادی**: تأثیرات اجتماعی و اقتصادی بر سلامت مانند درآمد، تحصیلات و شغل و غیره

۳- **مواجهه‌های محیطی**: تماس با آلودگی‌های هوا یا آب، خطرات محیط کار و غیره

۴- **مصرف مواد**: مصرف دخانیات، الکل، داروهای غیرمجاز

۵- **فعالیت‌های بدنی**: ورزش‌ها و فعالیت‌های فیزیکی

۶- **مراقبت‌های شخصی و زیبایی**: بهداشت فردی، جراحی‌های زیبایی، مراقبت‌های شخصی

۷- **فعالیت‌های فراغتی غیر فیزیکی**: فعالیت‌های اوقات فراغت که نیاز به تحرک فیزیکی ندارند مانند تماشای تلویزیون یا بازی‌های ویدئویی.

۸- **خواب**: الگوهای خواب، کیفیت و عادات خواب

۹- **تمرین‌های سلامت روان**: فعالیت‌های مرتبط با سلامت عاطفی و روانی

ملاحظات: عبارات کلی که به‌طور مستقیم با سلامت در ارتباط نیستند یا به یکی از دسته‌های فوق تعلق ندارند، حذف می‌شوند.

قالب خروجی: برای نمایش موجودیت‌های استخراج شده از قالب BRAT استفاده می‌شود. یعنی به ازای هر متن چکیده مقاله ورودی، هر موجودیت استخراج شده از متن شامل موارد زیر است:

برچسب: که می‌تواند در برگیرنده یکی از گروه‌های ۹ گانه سبک زندگی باشد

موقعیت شروع: محل شروع موجودیت در متن

موقعیت پایان: محل پایان موجودیت در متن

متن: عبارت استخراج شده

انطباق با متن: موجودیت‌ها باید دارای زمینه‌ای مرتبط با سلامت باشند و از استخراج عبارات مبهم یا کلی که فاقد این ارتباط هستند، اجتناب شود. همچنین موجودیت‌ها باید به‌صورت مختصر و دقیق انتخاب شوند تا مفهوم اصلی را به‌درستی منعکس کنند.

شکل ۲ دستور متنی اصلی استفاده شده در مرحله اول یعنی یادگیری بدون نمونه که به عنوان ورودی LLM جهت شناسایی دقیق فاکتورهای سبک زندگی طراحی کرده ایم را نشان می‌دهد.

```
"""
You are an expert Named Entity Recognition (NER) system.
Your task is to accept Text as input and extract LIFESTYLE entities.
Below are Guidelines to help you what kinds of named entities to extract for LIFESTYLE label.
LIFESTYLE Entity Guidelines:

1. Valid Categories:
Lifestyle factors encompass various aspects that can impact health and well-being by causing, preventing, treating, increasing the risk, decreasing the risk, or having any association with diseases. These factors include:

Nutrition
Socioeconomic Factors
Environmental Exposures
Substance Use
Physical Activities
Non-physical Leisure Activities
Personal Care & Cosmetic Procedures
Sleep
Mental Health Practices

2. Exclusions:
Exclude general terms and phrases unless explicitly tied to health outcomes or clearly belong to one of the above mentioned valid categories or is the name of the category itslef

3. Output Format:
Extract entities in BRAT format with proper offsets. Each entity should include:
```

Label: LIFESTYLE

Start offset: Position in the text where the entity begins

End offset: Position where the entity ends

Text: The extracted entity itself

4. Context Matching:

Extract entities with a clear health-related context.

Do not Extract vague or general word or phrases unless they are health related in the given context.

5. Entities must be concise and capture only key phrases

Query: Using the above guidelines, identify and extract lifestyle factor entities (label: LIFESTYLE) from the following Abstract. Ensure strict compliance with the guidelines, including entity definitions, exclusions, and formatting. Return the extracted entities in the BRAT format ordered as they occur in the Abstract and ensure start and end offset of the entities are correct . \n"

"""

شکل ۳. دستور متنی مرحله اول که در برگیرنده قوانین برچسب گذاری برای ۹ جنبه مختلف سبک زندگی و قالب خروجی مورد انتظار است

یادگیری بدون نمونه می‌تواند نقاط قوت و ضعف مدل در درک مفاهیم پیچیده و شناسایی موجودیت‌های سبک زندگی را مشخص کند. این روش به‌ویژه در سناریوهایی که داده‌های برچسب‌خورده کمیاب هستند، مفید واقع می‌شود. لذا راهکار LSF-NER که در این مقاله ارایه می‌شود در سه نسخه می‌باشد که نسخه اول، بدون استفاده از مثال و فقط بر اساس دستورالعمل و راهنمای برچسب‌گذاری توصیف شده در مهندسی دقیق انتظار دارد که LLM بتواند فاکتورهای سبک زندگی مورد انتظار را منطبق بر قالب توصیف شده در دستور متنی کشف کند. شکل ۳ نمونه‌ای از موجودیت‌های کشف شده از متن چکیده یک مقاله را در قالب BRAT نمایش می‌دهد که قالب مورد انتظار ما از LLM می‌باشد.

T1	Nutrition	0	9	Vitamin D
T2	Disease	48	56	COVID-19
T3	Socioeconomic_factors	58	71	Social status
T4	Disease	92	103	lung cancer
T5	Environmental_exposures	105	118	Air pollution
T6	Disease	157	165	COVID-19
T7	Substance_use	167	182	Tobacco smoking
T8	Disease	202	214	heart attack
T9	Physical_activities	216	233	Physical activity
T10	Disease	265	278	breast cancer
T11	Non_physical_leisure_time_activities	280	291	Watching TV
T12	Disease	335	342	obesity
T13	Personal_care_products_and_cosmetic_procedures	344	360	Cosmetic surgery
T14	Disease	387	399	nerve damage
T15	Sleep	401	411	Sleep loss
T16	Disease	443	457	type2 diabetes
T17	Mental_health_practices	459	477	Group play therapy

شکل ۴. نمونه‌ای از موجودیت‌های کشف شده از متن چکیده یک مقاله در قالب BRAT

همان‌طور که در شکل ٤ نمایش داده شده، هر موجودیت کشف شده در یک سطر قرار داد و در ستون اول، یک شناسه مثلا T۱ و در ستون دوم نوع موجودیت کشف شده مانند Nutrition، در ستون‌های سوم و چهارم موقعیت شروع و پایان موجودیت کشف شده در داخل متن ذکر شده و در نهایت ستون آخر عبارت متنی که کشف شده را نمایش می‌دهد. به عنوان مثال در سطر اول Vitamin D به عنوان یک مکمل غذایی در دسته Nutrition کشف شده و در متن در موقعیت ۰ تا ۹ قرار دارد. البته باید توجه داشت که موجودیت‌هایی که از نوع Disease می‌باشد صرفا جهت توضیح اضافه شده و هدف LSF-NER کشف بیماری‌ها نیست.

۲-۲- یادگیری با چند نمونه

تجربه نشان داده که اگر در دستور متنی چند نمونه از ورودی و خروجی مورد انتظار قرار داده شود می‌توان نتایج با دقت بهتری کسب کرد. لذا نسخه دوم[۱٤] ارایه شده در این مقاله با استفاده از مثال‌هایی به بهبود دستور متنی می‌پردازد و باید توجه داشته باشیم که دستور متنی یک فرایند بهبود تکراری است، بدین معنی که ابتدا از یک دستور متنی اولیه شروع کرده و با ایجاد تغییراتی با استفاده از تکنیک‌های شناخته شده‌ای به بهبود دقت خروجی کمک می‌کنیم. نمونه‌های اضافه شده به

14-Few-Shot LSF-NER

مدل کمک می‌کننـد تا انتظارات ما از قالب خروجی و انواع موجودیت‌های موردنظر برای استخراج را بهتر درک کند. این راهکار در کاربردهای متعددی همچون پزشـکی نتایج امیدوارکننده‌ای از خود نشـان داده است [۲۱, ۲۲]. انتخاب نمونه‌ها به‌گونه‌ای انجام می‌شود که تنوع فاکتورهای سبک زنـدگی مانند تغذیه، عوامل محیطی، و وضعیت اجتماعی-اقتصادی را پوشـش دهد. اما رسیدن به این هدف آسـان نیسـت، چرا که سـبک زندگی مدنظر در ایـن مطالعه در برگیرنـده طیف وسیعی از ۹ جنبه مختلف بوده که هر یک از آنها نیز زیر شـاخه‌های متعددی دارند. لذا نشان دادن مثال‌هایی که ارتباط مسـتقیمی با متن مورد سوال نداشته باشـند ممکن اسـت برای LLM گمراه‌کننده باشد. در این مطالعه روش جدیدی ارایه شـده کـه از مجموعه‌داده‌های محدودی به نام LSF200 که جهت طراحی مثال‌ها در اختیار داریم شبیه‌ترین مثال‌ها به متن مورد سوال فعلی را انتخاب کرده و در دستور متنی استفاده کنیم.

۲-۲-۱- بهبود دسـتور متنی به کمـک روش k نزدیکترین همسایه

یادگیری با چند نمونه به مدل امکان می‌دهد که الگوهای پیچیده‌تر را شناسایی کرده و دقت شناسایی موجودیت‌ها را افزایـش دهد. این روش به‌ویژه در حوزه‌هایی که دارای تنوع بالای داده هستند، کارایی بالایی دارد. مهمترین چالش برای تنوع بالا زمانی اسـت که بخواهیم از مثال‌های ثابتی در مهندسی دقیق استفاده کنیم بدون توجه به متن ورودی. برای متناسـب کـردن مثال‌هـای مورد اسـتفاده با متن ورودی در این مطالعه از روش KNN استفاده می‌کنیم تا K شبیه‌ترین مثال به متن ورودی در مجموعه‌داده‌های مثال‌ها را پیدا کرده و در دستور متنی استفاده کنیم.

ناگفتـه پیداسـت کـه لازمـه ایـن روش داشـتن مجموعه‌داده‌هـای از مثال‌هـای از پیش برچسـب‌گذاری شـده اسـت. در این تحقیق از مجموعه‌داده‌های LSF200 اسـتفاده شده اسـت، به نحوی که از ۸۰ درصد متون این مجموعه‌داده‌ها برای تولید مثال و افزودن آن به دسـتور متـنی و از ۲۰ درصد مابقی برای ارزیابی مدل اسـتفاده می‌کنیم. بایسـتی توجه شود که در این مجموعه‌داده‌ها به ازای هر چکیده مقاله (۲۰۰ چکیده) دو فایل ارایه شده است، فایل اول حاوی متن خام چکیـده بوده و فایل دوم حاوی موجودیت‌های استخراج شـده (فاکتورهای سبک زندگی) از متن اسـت که به قالب BRAT ذخیره شـده. فرآیند تولید دسـتور متنی و یافتن مناسـب‌ترین مثال‌ها با تکنیک‌های جستجوی مشـابهت برداری در ادامه شرح داده می‌شود. این روش شامل مراحل زیر است:

۲-۲-۱-۱- تبدیل موجودیت‌های برچسـب‌گذاری شـده در مجموعه‌داده‌ها به بردارهای عددی

ابتدا فایل‌های حاوی برچسـب‌های موجودیت‌ها (شکل ٤) کـه به ازای هـر متن چکیـده در قالب فایـل مجزایی در مجموعه‌داده‌هـای LSF200 قـرار دارنـد و حـاوی برچسـب‌گذاری‌های انسـانی بـه قالب BRAT می‌باشـد اسـتفاده شده و موجودیت‌های کشـف شده از هر چکیده مقاله که در سـتون آخر هر فایل قرار گرفته‌اند اسـتخراج می‌شـوند. سـپس برای تبدیل ایـن موجودیت‌های متنی بـه بردارهای عددی[15]، از مدل مبـدّل جملات[16] با معماری all-MiniLM-L6-v2 اسـتفاده می‌شود. این مدل قادر است مفاهیم معنایی جملات و عبارات را به بردارهای عددی در فضای برداری نگاشـت کند. هدف این مرحله این است که بتوانیم برای هر فایل چکیده برداری داشـته باشیم که در برگیرنده موجودیت‌های کشف شـده از آن چکیده باشد. برای این نگاشـت از روش‌های متعددی می‌توان اسـتفاد کرد. مزیت اسـتفاده از مبدّل جملات آن است که بر خلاف مدل‌هـایی ماننـد Word2Vec کـه تمرکز آن‌هـا بر روی کلمات منفرد اسـت، مدل‌های مبدّل جملات قادرند مفاهیم پیچیده‌تـر مانند عبارات چندکلمه‌ای را به شـکل دقیق‌تری بـه بردارهای معنایی تبدیل کنند. این ویژگی برای کشـف موجودیت‌های مرتبط با فاکتورهای سبک زندگی که اغلب به صورت عبارات چندکلمه‌ای هستند، بسیار حیاتی است.

15-Embedding
16-Sentence Transformer

در خصوص انتخاب مدل all-MiniLM-L6-v2 نیز در این تحقیق بر اساس ارزیابی کارایی آن انجام شده است. این مدل توازن خوبی از نظر سرعت پردازش و دقت عملکرد داشته و متناسب با سایز مجموعه‌داده‌ها و سرعت دلخواه، می‌توان از سایر مدل‌های مبدّل جملات نیز بسته به نیاز پروژه از لیست بهترین مدل‌های موجود انتخاب دیگری داشت[۱۷].

۲-۲-۱-۱- ساخت شاخص

بردارهای به‌دست آمده از مرحله قبل به ازای هر فایل چکیده مقاله به کمک کتابخانه FAISS[۲۳]، که توسط فیسبوک توسعه داده شده و برای جستجوی سریع و کارآمد در داده‌های برداری طراحی شده، به یک شاخص برداری تبدیل می‌شوند. در حقیقت FAISS یکی از ساده‌ترین و کارآمدترین انواع جستجوی شباهت را ارایه می‌دهد که بر اساس محاسبه فاصله اقلیدسی[۱۸] بین بردارها عمل می‌کند. این روش فاصله بین دو بردار را به صورت زیر محاسبه می‌کند:

$$L = k\_argmin_{i=۰:l} ||x - y_i||_۲ \qquad (۱)$$

که در آن همه بردارها n بعدی هستند وx بردار حاصل از نگاشت متن مورد سوال فعلی و y بردار حاصل از نگاشت متون موجود در مجموعه‌داده‌ها است که می‌خواهیم k شبیه‌ترین از آنها به بردار مورد آزمایش یعنی x یافته شود. این شاخص تمامی بردارها را در حافظه نگه می‌دارد و به همین دلیل سرعت بسیار بالایی در جستجوی مشابهت دارد، اما مصرف حافظه آن نسبتاً زیاد است. این ویژگی‌ها باعث می‌شوند که این روش برای مجموعه‌های داده‌ای کوچک تا متوسط انتخاب مناسبی باشد.

۲-۲-۱-۳- استخراج اولیه موجودیت‌های متون جدید

برای متن مورد جستجو که قاعدتا یکی از ۲۰ درصد مقالات موجود در بخش آزمایش مجموعه‌داده‌ها می‌باشد، ابتدا یک LLM بدون هیچ مثالی در دستور متنی مورد استفاده قرار می‌گیرد. این مدل به‌صورت مستقیم موجودیت‌های اولیه متن را کشف می‌کند. موجودیت‌های کشف شده در این مرحله قرار نیست به‌عنوان خروجی نهایی استفاده شوند و هدف آن است که تخمینی از موضوع متن مورد جستجو به دست آوریم و بر اساس آن شبیه‌ترین متون را از شاخص ایجاد شده استخراج کنیم.

۲-۲-۱-۴- جستجوی مشابه‌ترین متون در شاخص

موجودیت‌های کشف شده از متن مورد جستجو در مرحله قبل نیز به بردارهای عددی تبدیل می‌شوند و با شاخص FAISS مقایسه می‌شوند. این مقایسه به منظور یافتن K متن که شبیه‌ترین موجودیت‌ها را به موجودیت‌های اولیه استخراج شده از متن مورد سوال دارد می‌باشد. توجه داشته باشید که در همه این مراحل می‌توانستیم به جای آن که موجودیت‌های متون را به بردار نگاشت کنیم، خود متن خام را به بردار معادل نگاشت کنیم و هنگام جستجو نیز بردار حاصل از نگاشت متن مورد جستجو را مدنظر قرار دهیم و نیازی به استفاده از یک LLM برای استخراج اولیه موجودیت‌ها نباشد. مشکل این روش آن است که متن کامل چکیده یک مقاله حاوی اطلاعات و عبارات متعددی است که ممکن است مستقیما با موضوع موجودیت‌های مدنظر هماهنگ نباشد و اگر معیار جستجو و پیدا کردن متون مشابه بر اساس آنها تنظیم شود برای LLM گمراه‌کننده باشد.

۲-۲-۱-۵- ایجاد دستور متنی مبتنی بر چند نمونه

در این حالت، نمونه‌های مشابه یافته‌شده به‌عنوان داده‌های آموزشی محدود در یک دستور متنی جدید قرار داده می‌شوند. دستور متنی جدید در حقیقت متشکل از دستور متنی موجود در شکل ۳ به همراه k مثال یافته شده در این مرحله می‌باشد. نحوه اضافه کردن این مثال‌ها به دستور متنی نیز بدین شکل است که ابتدا متن خام هر مثال از فایل txt موجود در مجموعه‌داده‌ها قرار داده می‌شود و در ادامه موجودیت‌های برچسب‌گذاری شده انسانی موجود

17-https://huggingface.co/spaces/mteb/leaderboard
18-L2 Distance

در فایل متناظر ann آن چکیده اضافه می‌شود و این عمل برای هر یکی از k مثال تکرار می‌شود و در نهایت بعد از اتمام مثال‌ها، متن خام مورد سوال به دستور متنی اضافه می‌شود و موجودیت‌های متن مورد سوال قرار می‌گیرد انتظار بر آن است که LLM بتواند به کمک مثال‌های مشابه ارایه شده در دستور متنی فرآیند کشف موجودیت‌های متن جدید را با دقت بالاتری انجام دهد. استفاده از تکنیک یادگیری با چند نمونه موجب می‌شود مدل درک بهتری از موضوع متن مورد جستجو پیدا کند و همچنین فرمت خروجی مورد انتظار را بتواند از مثال‌های فراهم شده در دستور متنی استخراج کند.

۲-۳- آموزش مدل با استفاده از QLoRA [19]

در مرحله نهایی، به منظور تطبیق دقیق‌تر LLM با کاربرد خاص موردنظر، فرآیند تنظیم دقیق وزن‌های LLM انجام می‌شود. برای این منظور از روش تنظیم دقیق مبتنی بر دستورالعمل[20] استفاده می‌شود[۲۴].

در این مرحله، مدل با نمونه‌های متنوعی از متون علمی پزشکی آموزش می‌بیند تا توانایی خود را در شناسایی دقیق موجودیت‌های مرتبط با فاکتورهای سبک زندگی بهبود دهد. این فرآیند نیازمند منابع محاسباتی قابل توجهی است، اما می‌تواند دقت مدل را در کاربردهای خاص به‌طور چشمگیری افزایش دهد. این فرآیند شامل آماده‌سازی داده‌ها و انجام فرآیند تنظیم دقیق با استفاده از تکنیک QLoRA است که در ادامه توضیح داده می‌شود.

۲-۳-۱- آماده‌سازی مجموعه‌داده‌ها

آماده‌سازی مجموعه‌داده‌ها برای فرآیند تنظیم دقیق به صورت زوج‌های دستور[21] انجام می‌شود. این فرآیند با استفاده از داده‌های واقعی برچسب‌گذاری شده در مجموعه‌داده‌های LSF200 انجام می‌شود که متن چکیده به‌صورت ورودی و موجودیت‌های استخراج شده به عنوان خروجی‌های مورد انتظار آماده شده‌اند. هر نمونه شامل سه نقش است:

سیستم: در این نقش، دستورالعمل‌ها یا راهنماهای برچسب‌گذاری و مهندسی دقیق‌های از پیش آماده شده ارائه می‌شوند. محتوای ارایه شده در این بخش، نقش تعیین‌کننده در هدایت مدل به سمت درک صحیح وظیفه دارد.

کاربر: محتوای این بخش شامل متن خام (چکیده مقالات) است که به‌عنوان ورودی به مدل داده می‌شود و هدف اصلی استخراج فاکتورهای سبک زندگی از این متن است.

دستیار: در این نقش، پاسخ مدل شامل موجودیت‌های استخراج‌شده از متن ورودی است که در قالب BRAT ارائه می‌شود. این پاسخ‌ها به مدل کمک می‌کنند تا الگوهای دقیق استخراج اطلاعات را یاد بگیرد.

لذا به ازای هر جفت فایل متن خام و فایل حاوی موجودیت‌های استخراج شده، فقط یک رکورد در فایل JSON با این سه نقش ایجاد می‌شود و این فایل به تعداد نمونه‌های موجود در بخش آموزش مجموعه‌داده‌های LSF200 رکورد خواهد داشت. و در نهایت فایل JSON آماده شده همان مجموعه‌داده‌ها حاوی زوج‌های دستور خواهد بود که در آموزش مدل استفاده می‌شود.

نکته حایز اهمیت این است که برخلاف تصور، تعداد نمونه‌های لازم برای آموزش LLM می‌تواند تعدادی در حد چند صد یا چند هزار باشد و نیاز به داشتن چندین میلیون نمونه آموزشی نیست. علت این است که قرار نیست تمام وزن‌های موجود در LLM به‌روزرسانی و تنظیم شود و تنها ماتریس کوچکی از وزن‌ها به LLM اضافه خواهد شد تا برای وظیفه و کاربرد خاص جاری تطبیق داده شود.

۲-۳-۲- فرآیند تنظیم دقیق با استفاده از QLoRA

روش QLoRA [۲۵]، با ترکیب رقمی‌سازی ٤ بیتی و LoRA[22] [۲٦] امکان تنظیم دقیق مدل‌های بزرگ را با مصرف کمتر منابع محاسباتی فراهم می‌کند. این روش به مدل اجازه می‌دهد تا با حفظ عملکرد بالا، مصرف حافظه

19-Quantized Low-Rank Adaptation
20-Instruction fine-tuning
21-Instruction pairs
22-Low-Rank Adaptation

و نیازهای محاسباتی را به طور قابل‌توجهی کاهش دهد. روش LoRA به جای به‌روزرسانی مستقیم تمام پارامترهای مدل، تغییرات را در یک فضای کم‌رتبه[۲۳] اعمال می‌کند. ایده اصلی این روش بر این فرض استوار است که تغییرات مورد نیاز برای تطبیق مدل به یک وظیفه جدید می‌توانند در یک زیرفضای کم‌رتبه نمایش داده شوند فرمول کلی LoRA به شکل زیر است.

$$W = W_{\cdot} + \Delta W \qquad (۲)$$

که در آن W ماتریس نهایی وزن‌های مدل بعد از آموزش است، $W_{\cdot}$ ماتریس وزن‌های اولیه مدل است، و ΔW تنظیمات لازم برای وزن‌هاست که بعد از انجام آموزش لازم است اعمال شود و ابعاد آن برابر با ابعاد ماتریس وزن‌های اولیه یعنی $W_{\cdot}$ است. نکته اصلی LoRa آن است که می‌توان ماتریس تنظیمات، یعنی ΔW را از طریق دو ماتریس با ابعاد بسیار کوچک‌تر بازسازی کرد که هزینه محاسباتی بسیار پایین‌تری خواهد داشت.

لذا می‌توان فرمول فوق را به شکل زیر بازنویسی کرد

$$W = W_{\cdot} + BA \qquad (۳)$$

که در آن A و B ماتریس‌های کم‌رتبه هستند، به نحوی که ابعاد ماتریس A عبارت است از $r \times d$ و ابعاد ماتریس B عبارت است از $d \times r$ و d ابعاد ماتریس اولیه وزن‌هاست. ضرب این دو ماتریس به مدل اجازه می‌دهد تا تغییرات لازم را در ابعاد کوچک‌تر اعمال کند، که باعث کاهش تعداد پارامترهای قابل آموزش می‌شود. به طور معمول r بسیار کوچک‌تر از d است، بنابراین تعداد پارامترهای جدید به طور قابل‌توجهی کمتر از تعداد پارامترهای ماتریس اصلی است.

روش QLoRA علاوه بر استفاده از ایده LoRA، از رقمی‌سازی ٤ بیتی برای کاهش مصرف حافظه استفاده می‌کند. در این روش، وزن‌های اصلی مدل به صورت فشرده شده ذخیره می‌شوند و تنها لایه‌های کم‌رتبه[۲٤] به‌روزرسانی می‌شوند. فرمول QLoRA به شکل زیر است:

$$W^* = Q(W) + BA \qquad (٤)$$

که در آن $Q(W)$ نسخه رقمی شده وزن‌های مدل اصلی است و BA ماتریس‌های تغییرات کم‌رتبه که مدل را برای وظیفه جدید تطبیق می‌دهد. در رقمی‌سازی، دقت نمایش وزن‌های مدل به ٤ بیت کاهش یافته، که باعث کاهش قابل توجه مصرف حافظه می‌شود. در پیاده‌سازی انجام گرفته در این تحقیق مدل Llama ۳ که ۸ میلیارد پارامتر دارد انتخاب شده و به روش QLoRA و با انتخاب $r = ۱۶$ به شکل ناباورانه‌ای تنها بر روی کولب[۲۵] با کارت گرافیکی T4 و نسخه رایگان که ۱۵ گیگابایت وی‌رم دارد اقدام به آموزش مدل کرده‌ایم. برای درک میزان کاهش تعداد پارامتر مورد نیاز، کافی است در نظر بگیریم که در ابعاد مدل اصلی $d = ۴۰۹۶$ می‌باشد و با انتخاب $r = ۱۶$، تعداد پارامترهای جدید برای دو ماتریس A و B تنها معادل $۲ \times ۱۶ \times ۴۰۹۶ = ۱۳۱/۰۷۲$ خواهد بود، در حالی که به‌روزرسانی مستقیم مدل اصلی نیازمند ۸ میلیارد پارامتر می‌بود. در حالی که تعداد پارامترهای LoRA به طور قابل‌توجهی کمتر است، QLoRA با اعمال رقمی‌سازی بر روی وزن‌های اصلی مدل نیز به کاهش مصرف حافظه کمک می‌کند. اگرچه رقمی‌سازی مستقیماً تعداد پارامترها را کاهش نمی‌دهد، اما با کاهش تعداد بیت‌های موردنیاز برای نمایش هر پارامتر، مصرف حافظه را به طور چشمگیری کاهش می‌دهد. این ترکیب از رقمی‌سازی و تطبیق کم‌رتبه[۲٦] موجب می‌شود که QLoRA بتواند مدل‌های بزرگ را به طور کارآمد و با منابع محدودتر آموزش دهد.

۳- ارزیابی عملکرد مدل

برای ارزیابی عملکرد LSF-NER و توانایی آن در استخراج فاکتورهای سبک زندگی از مجموعه آزمون مجموعه‌داده‌های LSF200 استفاده شده است و هر سه نسخه LSF-NER مورد ارزیابی قرار گرفته است. بدین‌منظور از چندین LLM بهره برده‌ایم که هم شامل LLMهای متن‌باز همچون Llama می‌باشد و هم استفاده از LLMهای تجاری همچون GPT4 از طریق API. مشخصات این مدل‌ها در

23-Low-rank
24-LoRA adapters
25-Colab
26-Low-Rank Adaptation

جدول ۲. جزییات LLMهای استفاده شده در این تحقیق

نام مدل	ارایه کننده	تعداد پارامتر	روش کوانتیزیشن	لینک دسترسی
Mistral	Mistral AI	7 billion	4-bit	https://huggingface.co/TheBloke/Mistral-7B-Instruct-v0.2-GGUF
Phi-4	Microsoft	14 billion	4-bit	https://huggingface.co/microsoft/phi-4
DeepSeek R1	DeepSeek	14 billion	4-bit	https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Qwen-14B
Llama 3	Meta	8 billion	4-bit	https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct
Llama 3 (Fine Tuned)	Meta	8 billion	4-bit	https://huggingface.co/Esmaeil/llama-3-8b-Instruct-bnb-4bit-FineTuned
GPT4o	OpenAI	N/A	N/A	N/A

جدول ۲ ارایه شده است. مدل GPT-4o در این جدول به دلیل عدم دسترسی عمومی و نبود اطلاعات دقیق در مورد تعداد پارامتر و رقمی‌سازی آن، بدون مقدار مشخصی برای این ویژگی‌ها درج شده است.

همان‌طور که در جدول ۲ ذکر شده است، در این تحقیق از پنج LLM مختلف استفاده شده است، البته LLM ششم نیز از طریق آموزش Llama ۳ تولید شده و در مجموع شش مدل مختلف ارایه شده است. همه مدل‌های متن‌باز با رقمی‌سازی ٤ بیتی ارائه شده‌اند. این روش رقمی‌سازی باعث کاهش حجم مدل‌ها می‌شود و امکان بارگذاری و اجرای آنها روی سخت‌افزارهای محدودتر، از جمله کامپیوترهای شخصی را فراهم می‌کند. بیشتر این مدل‌ها در نسخه‌های اصلی خود شامل ۸ تا ١٤ میلیارد پارامتر هستند، اما نسخه‌های چکیده[۲۷] شده (مانند DeepSeek R1) نیز در این لیست وجود دارند. این کاهش اندازه، بهینه‌سازی پردازش و کاهش نیازهای سخت‌افزاری را ممکن می‌سازد. در این تحقیق، آزمایش‌ها روی کامپیوتر Mac M1 با ۳۲ گیگابایت رم آزمایش شده‌اند، این سیستم با بهره‌گیری از معماری حافظه یکپارچه، امکان دسترسی به حدود ۲۰ گیگابایت وی رم مجازی را فراهم می‌کند که برای اجرای این مدل‌های رقمی شده کافی است. البته لازم به ذکر است که برای آموزش مدل Llama 3 از Colab با کارت گرافیکی T4 و نسخه رایگان که ۱۵ گیگابایت وی‌رم دارد استفاده شده است.

27-Distill

نوآوری این تحقیق ارایه راهکاری به نام LSF-NER برای استخراج فاکتورهای سبک زندگی از متون علمی است. این راهکار در سه نسخه متفاوت ارایه شده است که شامل یادگیری بدون نمونه، یادگیری با چند نمونه، و آموزش مدل (تنظیم دقیق) برای کاربرد NER می‌باشد. جدول ۳ دقت LSF-NER را در هر سه نسخه ارایه شده به کمک مدل‌های مختلف نمایش می‌دهد.

در این مقاله، برای ارزیابی عملکرد مدل، از سه معیار استاندارد Recall، Precision و F1-score استفاده شده است. این معیارها بر اساس مقایسه نتایج پیش‌بینی شده مدل با موجودیت‌های موجود در دیتاست LSF200 به نحو زیر محاسبه شده است:

(۵)

$$Precesion = \frac{TP}{TP+FP},\ Recall = \frac{TP}{TP+FN},\ F1 = ۲ * \frac{Precesion * Recall}{Precision + recall}$$

در این معیارها:

TP (True Positive) موجودیتی است که هم در مجموعه‌داده‌های LSF200 وجود دارد و هم به درستی توسط مدل شناسایی شده است.

FP (False Positive) موجودیتی است که توسط مدل شناسایی شده اما در مجموعه‌داده‌های مرجع وجود ندارد

FN (False Negative) موجودیتی است که در مجموعه‌داده‌های مرجع وجود دارد اما توسط مدل شناسایی نشده است.

جدول ۳. ارزیابی دقت LSF-NER با استفاده از LLMهای مختلف و سه رویکرد ارایه شده

نام مدل	F1		Recall		Precision	
	Three-Shot	Zero-Shot	Three-Shot	Zero-Shot	Three-Shot	Zero-Shot
Mistral	۰/۴۳۷	۰/۴۷۹	۰/۶۵۶	۰/۶۰۱	۰/۳۲۸	۰/۳۹۸
Phi-4	۰/۵۴۰	۰/۴۲۶	۰/۷۰۶	۰/۴۲۹	۰/۴۳۷	۰/۴۲۲
Llama-3	۰/۵۴۸	۰/۴۴۶	۰/۵۴۶	۰/۴۱۷	۰/۵۴۹	۰/۴۷۹
Llama-3 (Fine-Tuned)	۰/۶۳۴	۰/۵۷۰	۰/۶۸۱	۰/۶۲۶	۰/۵۹۴	۰/۵۲۳
DeepSeek-R1	۰/۶۱۳	۰/۵۳۳	۰/۶۰۷	۰/۴۱۷	۰/۶۱۹	۰/۷۳۹
GPT-4o	۰/۶۸۱	۰/۵۸۰	۰/۶۹۹	۰/۵۴۶	۰/۶۶۳	۰/۶۱۸

مهمترین نکته‌ای که ارزش توجه دارد، مقیاس مدل‌های متن باز در مقایسه با GPT-4o می‌باشد. تعداد پارامترهای این مدل‌های متن‌باز قابل مقایسه با GPT-4o نیست. هر چند که جزییات رسمی از سوی شرکت OpenAI ارایه نشده ولی تخمین‌های غیررسمی حاکی از آن است که تعداد این پارامترها به بیش از ۲۰۰ میلیارد می‌رسد. لذا عملکرد قابل قبول مدل‌های متن‌باز در مقایسه با GPT-4o به شدت ارزشمند است، هر چند که بهترین دقت کسب شده از آن GPT-4o با F1= 0.681 می‌باشد.

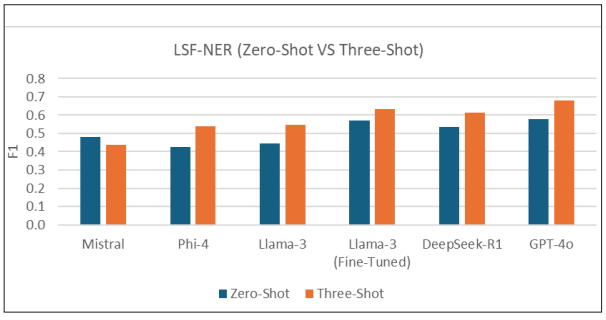
نوآوری مهم این تحقیق، آموزش مدل Llama 3 در قالب تنظیم دقیق مبتنی بر دستور است و این مدل آموزش دیده همان‌طور که لینک دسترسی به آن در جدول ۲ قرار داده شده قابل بارگیری است. عملکرد این مدل آموزش‌دیده برتر از دیگر مدل‌های متن باز و بسیار نزدیک به عملکرد GPT-4o می‌باشد. هر چند شاید این طور به نظر آید که این مدل برای هدف خاص فعلی روی مجموعه‌داده‌های موجود برچسب دار آموزش دیده و مقایسه آن به GPT-4o شاید عادلانه به نظر نرسد ولی باید توجه کرد که در بسیاری از کاربردهای امروزی به راحتی می‌توان مجموعه‌داده‌های برچسب دار محدودی را آماده کرد و از LLMهای متن باز در دسترس که حتی قابلیت اجرا بر روی ماشین شخصی را دارند بهره برد و با آموزش از طریق همان مجموعه‌داده‌های محدود، دقت مدل را به میزان قابل توجه و قابل مقایسه با مدل‌های تجاری رساند (۰/۶۳۴ در مقایسه به ۰/۶۸۱) بدون آن که مجبور به استفاده از مدل‌های تجاری و پرداخت هزینه‌های گزاف شد. بهبود مدل آموزش دیده Llama در مقایسه با مدل آموزش ندیده، بسیار قابل توجه است. به نحوی که معیار F1 از ۰/۵۴۸ به ۰/۶۳۴ رسیده است که معادل بیش از ۱۵ درصد بهبود دقت است.

نکته دیگری که در عملکرد DeepSeek-R1 قابل توجه است، دستیابی به بهترین دقّت است، بدین معنی که موجودیت‌هایی که توسط این مدل کشف شده با دقت حدود ۷۴ درصد صحیح است. جواب‌هایی که با این مدل تولید می‌شود، همواره با بخش استدلال شروع می‌شود که در داخل برچسب <think/> قرار می‌گیرد. بررسی محتویات داخل این برچسب به شدت مفید بود و نشان می‌دهد که چگونه مدل به جواب نهایی رسیده است و می‌توان با تحلیل این استدلال در یک فرایند بهبود تکراری اقدام به اصلاح دستور متنی کرد و نتایج بهتری کسب کرد. داشتن چنین توضیحات دقیقی از نحوه تصمیم‌گیری مدل برای مثال‌های مختلف مزیت بسیار بزرگی به نام توضیح‌پذیری[۲۸] را ارایه می‌دهد که مورد نیاز بسیاری از کاربردهای امروزی است. در اکثر موارد مدل‌های یادگیری ماشین همچون جعبه سیاهی هستند که نمی‌توان دلیل و توضیحی برای جواب‌های ارایه شده توسط آنها پیدا کرد و ارایه مراحل استدلال انجام گرفته توسط مدل، می‌تواند پیشرفت چشمگیری در این زمینه باشد.

نوآوری دیگر تحقیق جاری، ارایه راهکاری برای یافتن نزدیک‌ترین مثال‌ها به مثال مورد سوال و اضافه کردن

28-Explainability

آنها به مهندسی دقیق است تا از این طریق راهنمایی بهتری به LLM ارایه شـود، به‌خصوص آن که فاکتورهای سبک زندگی بسیار متنوع هسـتند و نشـان دادن متن مشابه به متن جاری در دسـتور متنی می‌تواند بسیار مفید بود و به کشف موجودیت‌های جدید کمک کند. تعداد مثال‌های اضافه شده به مدل دلخواه هست ولی تعداد مثال‌ها معمولا زیر ۵ اسـت، چرا که اگر قرار است مدل با دیدن مثال‌هایی بهبود یابد، نهایتا ۵ مثال کافی اسـت و در این تحقیق بر حسـب تجربه ۳ نزدیک‌ترین همسـایه بر اساس روش پیشنهادی از مجموعه‌داده‌های آموزش کشـف شـده و به دسـتور متنی اضافه شده است. شکل ۵ میزان بهبود دقت مدل‌های مختلف به بهره‌گیری از ۳ مثال اضافه شده به دستور متنی را نمایش می‌دهد.



شکل ۵. مقایسه دقت LSF-NER با استفاده از مدل‌های مختلف و دو رویکرد دستور متنی بدون نمونه و سه نمونه‌ای

همان‌طور که در شـکل ۵ مشـاهده می‌شود به غیر از Mistral عملکر تمام مدل‌ها در نسـخه Three-Shot بهبود داشـته و حتی در مورد Mistral نیز که F1 بهبودی تجربه نکرده، طبق جدول ۳ میزان Recall بهبود ۵ درصدی داشته و ایـن برای همه مدل‌هـا صدق می‌کند کـه بهبود Recall قابل‌توجهی همان‌طور که انتظـار می‌رود تجربه کرده‌اند. به عبارت دیگر ارایه مثال‌های متنوع و البته شـبیه به مثال جاری، کمک زیادی به افزایش تنوع موجودیت‌های کشـف شـده کرده و البته این ممکن اسـت با ریسک کاهش دقت همراه شـده و موجودیت‌های کشف شده صحیح نباشد[۲۹]. به همین منظور است که در این تحقیق تعداد مثال‌هایی که در دستور متنی اضافه شده تنها به سه شبیه‌ترین همسایه محدود شده و از ارایه مثال‌های نامرتبط که می‌تواند برای مدل گمراه کننده باشد، پرهیز شده است.

29-False Positive

نتیجه‌گیری

امروزه اسـتفاده از مدل‌های زبانی بزرگ همه زوایای یادگیـری ماشین را در بر گرفته و بـر خلاف این تصور که این فناوری به زیرسـاخت سخت افزاری بسیار گزافی نیاز داشـته و برای کاربردهای تحقیقاتی دور از دسترس است، در این تحقیق نشان دادیم که با سخت‌افزار محدودی می‌تـوان از عمده مزایای این فنـاوری بهره برد. نوآوری این پژوهـش در ارایه اولین مدلی اسـت کـه به صورت خاص برای اسـتخراج فاکتورهای سـبک زندگی از متون علمی طراحی شده اسـت. رویکرد پیشنهادی ما با ترکیب سه مرحله‌ای یادگیری بدون نمونه، یادگیری با چند نمونه، و نهایتـا آموزش مـدل، امکان شناسـایی دقیق و کارآمد فاکتورهای سـبک زندگی در متون علمی پزشکی را فراهم می‌کند و مدل آموزش داده شده در این تحقیق در دسترس محققین قرار گرفته است. این رویکرد نه تنها از قابلیت‌های مدل‌های زبانی بزرگ متن‌باز بهره می‌برد، بلکه چالش‌های مربوط به کمبـود داده‌های برچسـب‌خورده را نیز تا حد زیادی برطرف می‌کند. قدم بعدی، اسـتخراج ارتباطات بین بیماری‌ها و فاکتورهای سـبک زندگی بـوده و در نهایت می‌توان تمـامی ارتباطات موجود در ده‌هـا میلیون مقاله پزشکی را استخراج و تبدیل به پایگاه دانش ساخت‌یافته‌ای کرد که زمینه را برای یافته‌های نوین پزشکی فراهم می‌کند

مراجع

1. Cho, H., & Lee, H. 2019. “Biomedical named entity recognition using deep neural networks with contextual information,” BMC Bioinformatics, vol. 20, no. 1, p. 735, Dec. doi: 10.1186/s12859-019-3321-4.
2. Liu, X. 2019. “Deep Recurrent Neural Network for Protein Function Prediction from Sequence,” arXiv:1701.08318, Jan. 2017, Accessed: Dec. 03, [Online]. Available: http://arxiv.org/abs/1701. 08318
3. Song, B., Li, F., Liu, Y. & Zeng, X. 2021. “Deep learning methods for biomedical named entity recognition: a

survey and qualitative comparison," Briefings in Bioinformatics, vol. 22, no. 6, p. bbab282, Nov. doi: 10.1093/bib/bbab282.

4. Jia, C., Liang, X., & Zhang, Y. 2019. "Cross-Domain NER using Cross-Domain Language Modeling," in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy: Association for Computational Linguistics, pp. 2464–2474. doi: 10.18653/v1/P19-1236.
5. Chen, X. et al., 2023. "One Model for All Domains: Collaborative Domain-Prefix Tuning for Cross-Domain NER", arXiv. doi: 10.48550/ARXIV.2301.10410.
6. Wang, S. et al., 2023. "GPT-NER: Named Entity Recognition via Large Language Models," arXiv. doi: 10.48550/ARXIV.2304.10428.
7. Luo, Z., Yang, Y., Qi, R., Fang, Z., Guo, Y. & Wang, Y. 2023. "Incorporating Large Language Models into Named Entity Recognition: Opportunities and Challenges," in 2023 4th International Conference on Computer, Big Data and Artificial Intelligence (ICCBD+AI), Guiyang, China: IEEE, Dec. pp. 429–433. doi: 10.1109/ICCBD-AI62252.2023.00079.
8. Zhang, Z., Zhao, Y., Gao, H. & Hu, M. 2024. "LinkNER: Linking Local Named Entity Recognition Models to Large Language Models using Uncertainty," in Proceedings of the ACM Web Conference 2024, Singapore Singapore: ACM, May 2024, pp. 4047–4058. doi: 10.1145/3589334.3645414.
9. Keloth, V. K.et al., 2024. "Advancing entity recognition in biomedicine via instruction tuning of large language models," Bioinformatics, vol. 40, no. 4, p. btae163, Mar. 2024, doi: 10.1093/bioinformatics/ btae163.
10. Zhang, Y. et al., 2020."Combined lifestyle factors and risk of incident type 2 diabetes and prognosis among individuals with type 2 diabetes: a systematic review and meta-analysis of prospective cohort studies," Diabetologia, vol. 63, no. 1, pp. 21–33, Jan. 2020, doi: 10.1007/s00125-019-04985-9.
11. Zhang, Y.-B.et al., 2021. "Combined lifestyle factors, all-cause mortality and cardiovascular disease: a systematic review and meta-analysis of prospective cohort studies," J Epidemiol Community Health, vol. 75, no. 1, pp. 92–99, Jan. 2021, doi: 10.1136/jech-2020-214050.
12. Randolph, J. J.et al., 2024. "Integrating Lifestyle Factor Science into Neuropsychological Practice: A National Academy of Neuropsychology Education Paper," Archives of Clinical Neuropsychology, vol. 39, no. 2, pp. 121–139, Feb. 2024, doi: 10.1093/arclin/acad078.
13. Firth, J. et al., 2020."A meta-review of 'lifestyle psychiatry': the role of exercise, smoking, diet and sleep in the prevention and treatment of mental disorders," World Psychiatry, vol. 19, no. 3, pp. 360–380, Oct. 2020, doi: 10.1002/wps.20773.
14. Atallah N. et al., 2018. "How Healthy Lifestyle Factors at Midlife Relate to Healthy Aging," Nutrients, vol. 10, no. 7, p. 854, Jun. 2018, doi: 10.3390/nu10070854.
15. Jirout et al., J. 2019. "How Lifestyle Factors Affect Cognitive and Executive Function and the Ability to Learn in Children," Nutrients, vol. 11, no. 8, p. 1953, Aug. 2019, doi: 10.3390/nu11081953.
16. Martins, L. C. G., Lopes, M. V. D. O., Diniz, C. M., & Guedes, N. G. 2021. "The factors related to a sedentary lifestyle: A meta-analysis review," Journal of Advanced Nursing, vol. 77, no. 3, pp. 1188–1205, Mar. 2021, doi: 10.1111/jan.14669.
17. Nourani, E. et al., 2025. "LSD600: the first corpus of biomedical abstracts annotated with lifestyle–disease relations," Database, vol. 2025, p. baae129, Jan. 2025, doi: 10.1093/database/baae129.
18. Zhu, T., Qin, Y., Feng, M., Chen, Q., Hu, B. & Xiang, Y. 2024. "BioPRO: Context-Infused Prompt Learning for Biomedical Entity Linking," IEEE/ACM Trans. Audio Speech Lang. Process., vol. 32, pp. 374–385, 2024, doi: 10.1109/TASLP.2023.3331149.
19. Nourani, E. et al., 2024. "Lifestyle factors in the biomedical literature: an ontology and comprehensive resources for named entity recognition," Bioinformatics, vol. 40, no. 11, p. btae613, Nov. 2024, doi: 10.1093/bioinformatics/btae613.
20. Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., & Tsujii, J. 2012."brat: a Web-based Tool for NLP-Assisted Text Annotation," in Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics, F. Segond, Ed., Avignon, France: Association for Computational Linguistics, Apr. 2012, pp. 102–107. Accessed: Oct. 22, 2024. [Online]. Available: https://aclanthology.org/E12-2021
21. Hu, Y. et al., 2024. "Improving large language models for clinical named entity recognition via prompt engineering," Journal of the American Medical Informatics Association, vol. 31, no. 9, pp. 1812–1820, Sep. 2024, doi: 10.1093/jamia/ocad259.
22. Zeghidi, H. & Moncla, L. 2024. "Evaluating Named Entity Recognition Using Few-Shot Prompting with Large Language Models," Sep. 04, 2024, arXiv: arXiv:2408.15796. doi: 10.48550/arXiv.2408.15796.
23. Johnson, J., Douze, M. & Jégou, H. 2021. "Billion-Scale Similarity Search with GPUs," IEEE Transactions on Big Data, vol. 7, no. 3, pp. 535–547, Jul. 2021, doi: 10.1109/TBDATA.2019.2921572.
24. Ouyang, L.et al., 2022. "Training language models to follow instructions with human feedback," Mar. 04, 2022, arXiv: arXiv:2203.02155. doi: 10.48550/arXiv.2203.02155.
25. Dettmers, T., Pagnoni, A., Holtzman, A. & Zettlemoyer, L. 2023. "QLORA: efficient finetuning of quantized LLMs," in Proceedings of the 37th International Conference on Neural Information Processing Systems, in NIPS '23. Red Hook, NY, USA: Curran Associates Inc., Dec. 2023, pp. 10088–10115.
26. Hu, E. J.et al., 2021. "LoRA: Low-Rank Adaptation of Large Language Models," Oct. 16, 2021, arXiv: arXiv:2106.09685. doi: 10.48550/arXiv.2106.09685.