Vol. 5, No. 17, Jan-Mar 2026, (57-83)
Journal of New Achievements in Electrical, Computer and Technology (JNAECT)
مجله دستاوردهای نوین در برق،کامپیوتر و فناوری
دوره ۵، شماره ۱۷، زمستان ۱۴۰۴(۸۳-۵۷)
journal homepage: https://jnaect.ir

مروری بر روش‌های خودکار صحت‌سنجی مراجع مقالات

رضا صیدمرادی[۱]، مهدی نقوی[۲]

[۱] دانشجوی کارشناسی ارشد دانشگاه جامع امام حسین(ع)، [۲] استادیار دانشگاه جامع امام حسین(ع)

reza.seidmoradi@ihu.ac.ir

mnaghavai@ihu.ac.ir

چکیده

صحت‌سنجی ادعاهای علمی یکی از چالش‌های کلیدی در پردازش زبان طبیعی و هوش مصنوعی است که به دلیل حجم بالای اطلاعات و پیچیدگی روابط میان داده‌ها، نیازمند روش‌های دقیق و توضیح‌پذیر می‌باشد. پژوهش‌های اخیر از رویکردهای متنوعی شامل روش‌های با داده محدود، بازیابی شواهد و ارتقای روند، مبتنی بر جستجوی خارجی، استدلال ساخت‌یافته، خود ا ستدلالی و خودپالایی، یادگیری مت ضاد، مدل‌های ا ستنتاج زبان طبیعی و روش‌های تو ضیح‌پذیر ا ستفاده کرده‌اند. همچنین، تولید و استفاده از مجموعه داده‌های تخصصی و دوزبانه نقش مهمی در ارتقاء عملکرد این سیستم‌ها داشته است. در این مقاله، با دسته‌بندی و تحلیل روش‌ها و مجموعه داده‌های موجود، نقاط قوت و محدودیت‌های هر رویکرد بررسی شده و زمینه‌های قابل بهبود برای پژوهش‌های آتی مشخص گردیده است.

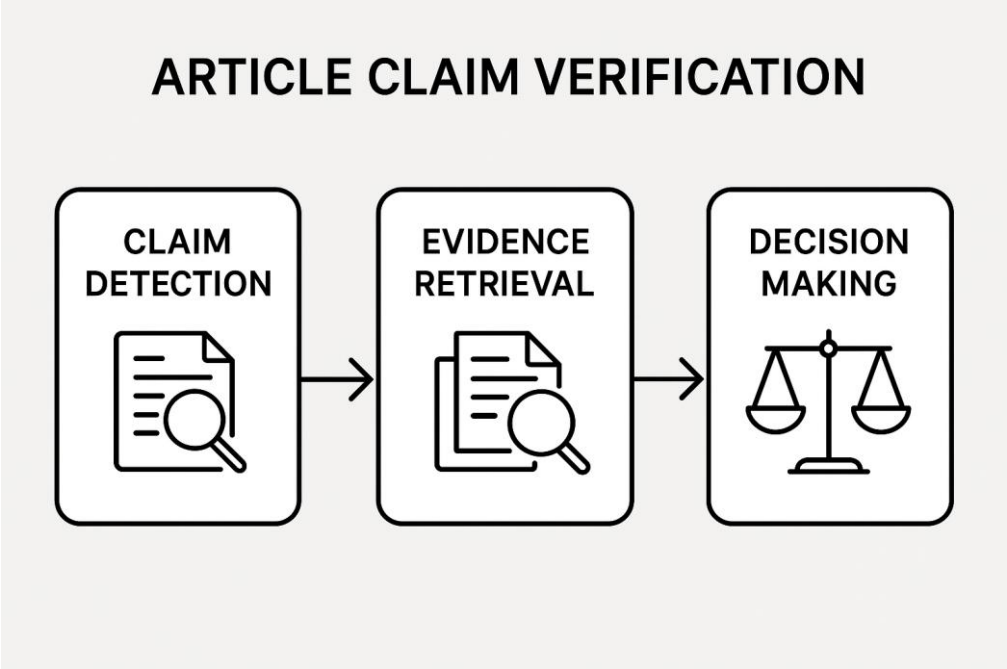
واژه‌های کلیدی: صحت‌سنجی علمی، راستی‌آزمایی خودکار، بازیابی شواهد، مدل‌های زبانی بزرگ، یادگیری متضاد، استدلال ساختاریافته، تولید تقویت‌شده با بازیابی، استنتاج زبان طبیعی

۱. مقدمه

با توسعه‌ی چشم‌گیر حجم مقالات علمی در دهه‌های اخیر، ارزیابی دقیق ادعاهای علمی به ویژه در زمینه‌های حساس مانند پزشکی و داروسازی تبدیل به فرآیندی دشوار و زمان‌بر شده است. مطالعات نشان می‌دهند که حتی برخی ادعاهای پرارجاع در حوزه‌های بالینی ممکن است در مراحل بعدی رد شوند. از سوی دیگر، نرخ بالای خطاهای استنادی که گاه تا ۲۵٪ گزارش شده است[۱][۲]، نشان‌دهنده‌ی نیاز جدی به ابزارهای دقیق و خودکار برای ارزیابی ادعاهاست.

در پاسخ به این چالش، ابزارهای خودکار برای صحت‌سنجی ادعاها شکل گرفته‌اند. رویکرد این ابزارها که در شکل ۱ نشان داده شده است، به صورت کلی به سه مرحله تقسیم می‌شود: تشخیص ادعا، بازیابی شواهد و تصمیم‌گیری در مورد نتیجه. یکی از

نمونه‌های برجسته در فضای پزشکی، سیستم CliVER است که با استفاده از تکنیک‌های تولید تقویت شده با بازیابی و چارچوب PICO، قادر به بازیابی اسناد بالینی مرتبط، انتخاب جملات مستدل و طبقه‌بندی ادعاها به حالت‌های «اثبات‌شده»، «ردشده» یا «نامشخص» است. این سیستم در مجموعه داده‌ی CoVERt به امتیاز F1 برابر با ۰,۹۲ دست یافت[۳].



شکل ۱- مراحل کلی فرایند صحت‌سنجی مراجع مقالات

افزون بر این، مطالعات تطبیقی و تحلیل‌های مروری، چارچوب‌های جدید بر پایه‌ی مدل‌های زبان بزرگ را تحلیل می‌کنند. این منابع ساختار کلی روند کار صحت‌سنجی شامل بازیابی، استدلال، و تنظیم دقیق را معرفی کرده و به لزوم دستیابی به سیستم‌هایی با تفسیرپذیری بهتر اشاره دارند[۴].

رویکردهای دیگری مانند FactDetect با تمرکز بر استخراج «حقایق کوتاه» از شواهد، زمینه‌ای برای بهبود عملکرد و توضیح‌پذیری فراهم کرده‌اند. با ادغام این حقایق کوتاه و استدلال چندمرحله‌ای AugFactDetect، عملکرد zero-shot مدل‌های زبانی بزرگ نیز به‌طور معنی‌داری ارتقا یافته است[۵]. در مواجهه با ادعاهای علمی پیچیده، رویکردهایی نظیر PrunE با بهره‌گیری از گراف‌های شواهد و تکنیک‌های هرس مسیر تلاش کرده‌اند تا ارتباط‌های کلیدی حفظ شوند و از ورود اطلاعات غیرمرتبط جلوگیری گردد[۶].

در نهایت، در حوزه زیست‌پزشکی سامانه‌ی SciClaims یک ساختار انتها به انتها مبتنی بر مدل زبانی بزرگ ارائه می‌دهد که استخراج ادعا، بازیابی شواهد و ارزیابی آن‌ها را بدون نیاز به تنظیم دقیق انجام می‌دهد و در مقایسه با مدل‌های پیشین نتایج قابل‌توجهی ارائه می‌کند.

۲. مجموعه‌داده‌ها

تولید و انتشار مجموعه‌داده‌های استاندارد، نقش کلیدی در آموزش و ارزیابی مدل‌های صحت‌سنجی ادعا دارد. این مجموعه‌داده‌ها با تأکید بر کیفیت شواهد، امکان مقایسه دیگر روش‌ها و توسعه مدل‌های دقیق‌تر را فراهم می‌کنند.

تورن و همکاران[۷] مجموعه‌داده‌ی FEVER را معرفی کردند که شامل ادعاهای استخراج‌شده از ویکی‌پدیا همراه با شواهد برچسب‌خورده است. این مجموعه، یکی از منابع استاندارد در پژوهش‌های صحت‌سنجی به‌شمار می‌رود و امکان آموزش و ارزیابی مدل‌ها را در تشخیص تأیید، رد یا نامطمئن بودن ادعا فراهم می‌کند.

الهندی و همکاران[۸] با توسعه‌ی مجموعه‌داده LIAR-PLUS، دلایل و توجیه‌های انسانی ارائه‌شده در مقالات حقیقت‌سنجی را نیز به داده‌های موجود افزودند و دو نوع طبقه‌بندی دودسته‌ای درست/نادرست و شش‌دسته‌ای از کاملاً درست تا کاملاً مخالف ایجاد کردند.

آگنشتاین و همکاران[۹] بزرگ‌ترین مجموعه‌داده‌ی حقیقت‌سنجی به نام MultiFC را از ۲۶ وب‌سایت معتبر گردآوری کردند که شامل ادعاها، منابع مرتبط و فراداده غنی با برچسب‌های روزنامه‌نگاران خبره است. این مجموعه‌داده علاوه بر فراهم‌کردن یک آزمون واقعی و چندرشته‌ای، امکان پژوهش روی ترکیب مدارک و مدل‌سازی فراداده را ایجاد می‌کند.

اوستروسکی و همکاران[۱۰] با معرفی PolitiHop نخستین گام در صحت‌سنجی چندمرحله‌ای ادعاهای سیاسی واقعی برداشتند. این مجموعه‌داده شامل ۵۰۰ ادعای سیاسی همراه با جملات شواهد چندمرحله‌ای است که بیش از ۷۵٪ آن‌ها به بیش از یک جمله برای ارزیابی نیاز دارند.

وادن و همکاران[۱۱] برای نخستین بار وظیفه‌ی صحت‌سنجی ادعاهای علمی را مطرح کردند و مجموعه‌داده‌ی SciFact را با حدود ۱۴۰۰ ادعای کارشناسی همراه با چکیده‌های برچسب‌دار و استدلال ارائه دادند که به‌عنوان یک بستر چالش‌برانگیز و راهبردی برای پژوهش‌های آینده در صحت‌سنجی علمی معرفی شد.

ساته و همکاران[۱۲] برای ارتقای کیفیت سنجش حقایق، مجموعه‌داده‌ی WikiFactCheck English را معرفی کردند که شامل بیش از ۱۲۴ هزار سه‌تایی ادعا-متن-حقیقت و بیش از ۳۴هزار ادعای واقعی ردشده است. این مجموعه‌داده، بزرگ‌ترین منبع واقعی برای بررسی حقیقت بوده و شرایطی نزدیک‌تر به سناریوهای عملی نسبت به داده‌های ساختگی مانند FEVER یا SNLIفراهم می‌کند.

چاکرابارتی و همکاران[۱۳] مجموعه‌داده‌ای شامل ۴۰۸۶ ادعای واقعی درباره همه‌گیری COVID-19 ایجاد کرده‌اند که هر ادعا همراه با شواهد مرتبط و نسخه‌های مخالف[1]خودکار تولیدشده از همان شواهد است. این مجموعه برخلاف داده‌های پیشین به‌طور کامل به روش خودکار گردآوری شده و علاوه بر ادعاهای علمی، شامل ادعاهای عمومی و ساده‌شده نیز هست. ویژگی مذکور، این مجموعه را به منبعی ارزشمند برای بررسی و مقابله با اطلاعات نادرست عمومی در زمینه COVID-19 تبدیل می‌کند.

کارامولگکو[۱۴] دو مجموعه‌داده جدید شامل ادعاها و شواهد تأییدکننده برای آموزش مدل‌های تفکیک و صحت‌سنجی مبتنی بر BERT ارائه کرد.

نورگارد و درچینسکی[۱۵] مجموعه‌داده DanFEVER را به‌عنوان نخستین منبع غیرانگلیسی در حوزه صحت‌سنجی ادعا معرفی کردند. این مجموعه‌داده به زبان دانمارکی و با الگوی FEVER طراحی شده و شامل ادعاها، شواهد انسانی‌برچسب‌خورده از ویکی‌پدیا و منابع محلی، و برچسب‌هایی مانند تأییدشده، ردشده یا اطلاعات ناکافی است.

ژیانگ و همکاران[۱۶] مجموعه‌داده HoVer را معرفی کردند که برای ارزیابی توانایی مدل‌ها در استخراج شواهد از چندین مقاله و صحت‌سنجی ادعاها طراحی شده است. مجموعه‌داده شامل ادعاهای چندبخشی با وابستگی‌های طولانی است که استدلال

[1] counter-claim

چندمرحله‌ای را پیچیده می‌کند و مدل‌های آزموده در HoVer باید شواهد را از حداکثر چهار سند مختلف استخراج کنند. HoVer با بیش از ۲۶هزار ادعا، جامعه پژوهش را به توسعه سامانه‌های صحت‌سنجی چندمرحله‌ای واقعی ترغیب می‌کند.

ساروتی و همکاران[۱۷] مجموعه‌داده HEALTHVER را معرفی کردند که برای ارزیابی صحت ادعاهای حوزه سلامت طراحی شده است. این مجموعه‌داده شامل ۱۴۳۳۰ جفت ادعا-استناد است که ادعاهای واقعی بازنویسی‌شده از پرسش‌های-COVID 19به همراه شواهد استخراج‌شده از مقالات علمی را شامل می‌شود و به‌صورت دستی با برچسب‌های تأیید، رد یا خنثی برچسب‌گذاری شده است. فرایند ساخت داده شامل جمع‌آوری ادعاها، بازیابی و رتبه‌بندی مقالات با مدل مبتنی بر T5، و سپس برچسب‌گذاری دستی روابط میان ادعا و شواهد بوده است.

خان و همکاران[۱۸] مجموعه‌داده‌ای نوین شامل ۳۳۷۲۱ ادعا به همراه مقالات بازبینی سازمان‌های صحت‌سنجی و مقالات پایه مرتبط ارائه کردند. این مجموعه‌داده بازتابی واقعی‌تر از فرآیند صحت‌سنجی فراهم می‌کند، زیرا همانند کار کارشناسان، ابتدا به بررسی ادعا و مقالات بازبینی پرداخته و سپس از منابع پایه برای استدلال استفاده می‌شود. در حالی که بسیاری از مجموعه‌داده‌های پیشین صرفاً شامل ادعا و شواهد خلاصه بودند.

ولادیکا و همکاران[۱۹] مجموعه‌داده دوزبانه HealthFC را برای صحت‌سنجی ادعاهای حوزه سلامت معرفی کردند. این مجموعه شامل ۷۵۰ ادعا به زبان‌های انگلیسی و آلمانی است که توسط متخصصان پزشکی برچسب‌گذاری شده و با شواهد معتبر بالینی تأیید شده‌اند و دارای استدلال‌های استخراج‌شده، توضیحات مختصر و فراداده غنی است.

تن و همکاران[۲۰] روشی برای تولید خودکار داده‌های صحت‌سنجی علمی ارائه کردند که پرسش‌های چهارگزینه‌ای را به ادعاهای خبری تبدیل می‌کند. با این روش، دو مجموعه‌داده‌ی بزرگ در حوزه‌ی پزشکی به نام MedFact با ۱۵۰هزار نمونه و علوم عمومی به نام GsciFact با ۳۲هزار نمونه ساخته شد.

کائو و همکاران[۲۱] مجموعه‌داده SciNews را شامل ۲۴۰۰ مقاله علمی معرفی کردند تا توانایی مدل‌ها در شناسایی اطلاعات نادرست در اخبار علمی ارزیابی شود. معماری‌ها و راهبردهای مختلف مهندسی دستور برای استفاده از مدل‌های زبانی بزرگ در تشخیص تحریف یافته‌های علمی در رسانه‌ها برای این مجموعه‌داده بررسی شد.

دیگلمن و همکاران[۲۲] مجموعه‌داده‌ای هدفمند برای صحت‌سنجی ادعاهای مرتبط با تغییرات اقلیمی معرفی کردند که شامل ۱۵۳۵ ادعای واقعی است. برای هر ادعا، پنج جمله شاهد از ویکی‌پدیا استخراج و توسط کارشناسان به‌صورت اثبات‌کننده، ردکننده یا اطلاعات ناکافی برچسب‌گذاری شد و در مجموع ۷۶۷۵ جفت ادعا-شاهد ایجاد شد.

زو و همکاران[۲۳] مسئله‌ی اعتبارسنجی ادعاهای سلامت در اخبار چینی را با طراحی یک مجموعه‌داده‌ی دوزبانه شامل ۸۶۴۷ ادعا و بیش از ۷۶۳ هزار مقاله پزشکی بررسی کرده‌اند. این مجموعه‌داده دارای یک نسخه‌ی طلایی شامل ۱۴۳۹ ادعا با استناد مستقیم و یک نسخه‌ی نقره‌ای گسترش‌یافته با خوشه‌بندی اخبار، بازنویسی ادعاها با ChatGPT Paraphraser و مقالات مشابه PubMedاست.

۳. روش‌های خودکار صحت‌سنجی منابع

روش‌های خودکار صحت‌سنجی منابع[۲]به مجموعه تکنیک‌ها و الگوریتم‌هایی گفته می‌شود که هدفشان بررسی اعتبار یک ادعا یا محتوا بر اساس منابع موجود است. این فرآیند معمولاً با استخراج ادعا[۳]آغاز می‌شود که در آن جملات یا گزاره‌های قابل بررسی از متن شناسایی و به شکل ساختاریافته آماده می‌شوند. پس از آن، مرحله بازیابی شواهد مرتبط[۴]دنبال می‌شود که سیستم سعی می‌کند مستندات یا جملات پشتیبان مرتبط با ادعا را از منابع معتبر مانند مقالات علمی، ویکی‌پدیا یا پایگاه‌های داده تخصصی پیدا کند.

در نهایت مرحله تحلیل و تصمیم‌گیری در مورد ادعا[۵]می‌تواند به روش‌های مختلف که در ادامه مرور می‌شود انجام شود.

این چارچوب‌ها به پژوهشگران امکان می‌دهند تا صحت ادعاها را به‌صورت خودکار و مقیاس‌پذیر ارزیابی کنند و نقش مهمی در کاهش خطاهای انسانی و افزایش اعتماد به تحلیل داده‌های متنی علمی ایفا کنند.

۱،۳. تحلیل ادعا

بخشی از پژوهش‌ها به تحلیل ادعاها و چالش‌های مرتبط با صحت‌سنجی آن‌ها پرداخته‌اند. این پژوهش‌ها معمولا ویژگی‌های ادعا، روابط موجودیت‌ها و معیارهای قابل‌اعتبار بودن را بررسی می‌کنند و دیدگاه‌های مفهومی برای توسعه سامانه‌های صحت‌سنجی ارائه می‌دهند.

اشتمباخ و همکاران[۲۴] نشان دادند که انتخاب پایگاه دانش نقش تعیین‌کننده‌ای در صحت‌سنجی ادعا دارد. سیستم‌ها مشروط به دسترسی به پایگاه دانش مناسب می‌توانند بدون تغییر مدل به حوزه جدید منتقل شوند. «بهترین» پایگاه دانش وجود ندارد و عملکرد به هم‌پوشانی دامنه‌ای بستگی دارد. ترکیب چند پایگاه دانش معمولاً کمکی نمی‌کند و حتی مضر است. همچنین، نمره اعتماد مدل[۶]می‌تواند به‌عنوان شاخصی برای پیش‌بینی کارایی پایگاه دانش در غیاب برچسب‌های دستی به‌کار رود.

پن و همکاران[۲۵] در مروری جامع، نشان دادند که مدل‌های صحت‌سنجی ادعا در شرایط zero-shot و few-shot برای حوزه‌های فاقد داده‌های برچسب‌خورده به‌ویژه در مواجهه با ادعاهای خاص یا شواهد متفاوت عملکرد ضعیفی دارند. آن‌ها عوامل مؤثر بر این ضعف مانند اندازه داده، طول شواهد و نوع ادعا را شناسایی کرده و پیشنهاد کردند که پیش‌آموزش در حوزه هدف و تولید خودکار داده‌ها می‌تواند تعمیم‌پذیری را بهبود بخشد.

وورل و همکاران[۲۶] نشان دادند که در صحت‌سنجی ادعاهای پزشکی، نوع موجودیت‌ها و روابط میان آن‌ها نقشی اساسی دارد. ادعاهایی با موجودیت‌های مبهم یا روابط پیچیده، چالش‌برانگیزتر بوده و بررسی آن‌ها برای مدل‌های خودکار و حتی ارزیابان انسانی دشوارتر است.

ژانگ و آیوز[۲۷] با تعریف گراف منشأ ادعا به بررسی منبع و سیر تحول ادعاها پرداختند. این ساختار برای مدل‌سازی مسیر پیدایش و تغییر ادعا طراحی شده و عمدتاً به‌صورت یک وظیفه‌ی استخراج اطلاعات با کمک مدل استلزام متنی پیاده‌سازی شده است.

---

۲ Automatic Fact Verification
۳ Claim Extraction
۴ Evidence Retrieval
۵ Claim Verification
۶ Confidence Score

۲،۳. روش‌های مبتنی بر داده محدود

این روش‌ها، با غلبه بر چالش داده‌ی محدود سعی داشته‌اند ادعاها را اعتبارسنجی کنند و اساس این روش‌های صحت‌سنجی، مهندسی ویژگی، Few-shot و مدل‌های سازگار با مجموعه‌داده‌های کم‌نمونه و کوچک هستند.

سی و همکاران[۲۸] مدلی دو مرحله‌ای ارائه کردند که در مرحله نخست، با استفاده از نسخه‌ای اصلاح‌شده از TF-IDF آموزش‌دیده بر داده‌های خبرگزاری‌های معتبر، ادعاهای قابل بررسی در مناظرات ریاست‌جمهوری آمریکا شناسایی شد و در مرحله دوم، با بهره‌گیری از ویژگی‌هایی مانند برچسب‌های دستوری و شباهت کسینوسی، صحت ادعاها ارزیابی گردید. استفاده از خوشه‌بندی بدون‌نظارت برای تولید داده‌های جدید نقطه قوت مهم این روش محسوب می‌شود.

یانگ و همکاران[۲۹] روشی برای صحت‌سنجی ادعاها ارائه کردند که از داده‌های مثبت و بدون برچسب[۷]استفاده می‌کند. در این روش، ادعاهای تأییدشده توسط منابع معتبر به‌عنوان داده مثبت در نظر گرفته می‌شوند و سایر داده‌ها بدون برچسب باقی می‌مانند. سپس با بهره‌گیری از شبکه‌های مولد متخاصم[۸] داده‌های مصنوعی تولید و برچسب‌گذاری می‌شوند تا فضای آموزش غنی‌تری فراهم شود.

هاتوا و همکاران[۳۰] نیز با الهام‌گیری از همین تکنیک، چارچوبی سه‌گانه شامل سه جفت مولد-متمایزکننده ارائه کردند که به ترتیب داده‌های مثبت، داده‌های منفی و ترکیب نهایی برچسب‌گذاری را تولید و ارزیابی می‌کنند.

رایت و همکاران[۳۱] روشی چندمرحله‌ای برای خودکارسازی تولید ادعاهای علمی معتبر ارائه کردند. این رویکرد با ساده‌سازی جملات علمی پیچیده به ادعاهای پایه و قابل استناد، چالش‌هایی مانند کمبود داده و دشواری زبان علمی را برطرف می‌کند. آن‌ها همچنین روشی به نام KBIN برای ایجاد ادعاهای منفی معرفی کردند که انسجام و اعتبار بالاتری دارد.

ژنگ و گائو[۳۲] روش ProToCo را معرفی کردند که با تأکید بر سازگاری پاسخ‌ها در نسخه‌های مختلف ادعا، دقت مدل‌های زبانی بزرگ در صحت‌سنجی ادعا را با حداقل داده برچسب‌خورده افزایش می‌دهد. این رویکرد با ایجاد تغییرات معنایی ساده در ادعاها و اعمال محدودیت سازگاری، عملکرد مدل‌ها را در شرایط few-shot و zero-shot بهبود می‌بخشد.

چادلاپاکا و همکاران[۳۳] چارچوبی برای اعتبارسنجی ادعاها در متون داستانی معرفی کرده‌اند؛ جایی که حقیقت عینی وجود ندارد و داده‌های آموزشی بسیار محدود است. آن‌ها با طراحی داده‌های مصنوعی و به‌کارگیری ترکیب مدل‌های زبانی و شبکه‌های گرافی، روشی انعطاف‌پذیر ارائه کردند که در برخی موارد حتی از نویسندگان انسانی نیز عملکرد بهتری نشان داده است.

ژنگ[۳۴] سه رویکرد نوآورانه معرفی کرده است: SEED[۹]با بهره‌گیری از تفاوت‌های معنایی جفت ادعا–شواهد در شرایط کم‌نمونه، MAPLE[۱۰]با تحلیل تکاملی زبان و یک مدل seq2seq کوچک، و Active PETs که با یادگیری فعال با استفاده از مدل‌های مختلف PET[۱۱]داده‌های بدون‌برچسب را بهینه برچسب‌گذاری می‌کند.

لی و همکاران[۳۵]، به‌جای استفاده از مجموعه‌ای بزرگ از شواهد، تلاش می‌کنند گروه‌های حداقلی شواهد یا MEG[۱۲]ها شناسایی شوند که به‌طور کامل از ادعا پشتیبانی کنند. این گروه‌ها باید سه ویژگی اصلی داشته باشند: کافی بودن، غیرتکراری بودن و حداقلی بودن. به عبارت دیگر، هر گروه باید به‌طور کامل از ادعا پشتیبانی کند، هیچ‌یک از شواهد درون گروه نباید

---

۷ Positive Unlabeled Learning
۸ Generative Adversarial Network
۹ Semantic Embedding Element-wise Difference
۱۰ Micro Analysis of Pairwise Language Evolution
۱۱ Pattern Exploiting Training
۱۲ Minimal Evidence Groups

تکراری باشند، و تعداد شواهد درون گروه باید به حداقل برسد. برای شناسایی این گروه‌ها، پژوهشگران از مدل زبانی PaLM 2L استفاده کرده‌اند که تعیین می‌کنند آیا یک گروه شواهد از ادعا پشتیبانی می‌کند یا خیر.

۳،۳. روش‌های مبتنی بر بازیابی شواهد و ارتقای روند

مطالعات اخیر نشان می‌دهند که دقت نهایی صحت‌سنجی به‌طور مستقیم با کیفیت شواهد بازیابی‌شده و طراحی کارآمد روند کار مرتبط است.

با توجه به این موضوع، اشتمباخ و نیومن[۳۶] برای چالش FEVER 2.0 رویکردی دو مرحله‌ای ارائه دادند که در آن شواهد نه‌تنها براساس ادعا بلکه با جستجوی جانبی از دل شواهد قبلی نیز گسترش می‌یابد.

جینگ ما و همکاران[۳۷]، یک رویکرد انتها به انتها مبتنی بر شبکه‌ی سلسله‌مراتبی با مکانیزم توجه[۱۳]ارائه داده‌اند که با تمرکز مرحله‌ای بر شواهد مرتبط، صحت‌سنجی ادعاها را پیش‌بینی می‌کند.

جی ژو و همکاران[۳۸] چارچوب GEAR مبتنی بر شبکه گرافی[۱۴]را ارائه کرده‌اند که با تجمیع و استدلال روی مدارک بازیابی‌شده، اطلاعات میان مدارک را منتقل و ادعاها را به‌صورت صحیح، متناقض یا نیازمند اطلاعات بیشتر ارزیابی می‌کند.

لئو و همکاران[۳۹] نیز به روشی مبتنی بر شبکه توجه گرافی با ساختار کرنل‌محور تمرکز کرده‌اند که با وزن‌دهی دقیق‌تر به ادعا و شواهد، امکان استنتاج پیچیده و صحت‌سنجی مؤثرتر ادعاها را فراهم می‌کند.

سلیمانی و همکاران[۴۰] در چارچوب چالش FEVER رویکردی سه‌مرحله‌ای معرفی کردند که در آن از دو مدل BERT تنظیم‌دقیق‌شده استفاده شد: یکی برای بازیابی جملات استنادی و دیگری برای طبقه‌بندی ادعاها به سه حالت تأیید، رد یا ناکافی. درنهایت با بررسی روش‌های آموزش زیان نقطه‌ای و جفتی و نیز نقش کاوش منفی سخت[۱۵]نشان دادند که سیستم پیشنهادی، عملکرد برجسته‌ای دارد.

پن و همکاران[۴۱]، رویکردی مؤثر به نام QACG[۱۶]برای صحت‌سنجی بدون نیاز به داده‌های آموزشی ارائه کردند. در این روش، با استفاده از مدل‌های زبانی، از متون، پرسش-پاسخ تولید و سپس به جملات خبری تبدیل می‌شوند. بر این اساس، سه نوع ادعا شامل تأییدشده، ردشده و اطلاعات ناکافی به‌طور خودکار ساخته می‌شوند و برای آموزش مدل به‌کار می‌روند.

ژانگ و همکاران[۴۲]، به بررسی عمیق سه زیرفرایند کلیدی شامل بازیابی خلاصه[۱۷]انتخاب جمله‌ی استدلال[۱۸]و پیش‌بینی موضع[۱۹]در قالب یک چارچوب واحد و جامع پرداختند. این چارچوب با استفاده از یک مدل زبانی بزرگ هر سه مرحله را همزمان آموزش می‌دهد و با افزودن قاعده‌مندسازی[۲۰]بین نمرات توجه، بازیابی خلاصه و انتخاب استدلال، اطلاعات بین مراحل به‌طور مؤثر به اشتراک گذاشته می‌شود.

---

۱۳ Hierarchical Attention Network
۱۴ Graph Attention Network
۱۵ Hard negative mining
۱۶ Question Answering for Claim Generation
۱۷ Abstract retrieval
۱۸ Rationale selection
۱۹ stance prediction
۲۰ Regularization

ساندریال و همکاران[۴۳] رویکردی دو مرحله‌ای برای صحت‌سنجی ادعاهای مرتبط با همه‌گیری کووید-۱۹ پیشنهاد کردند. در این روش، ابتدا با ترکیب BM25 و مدل‌های مبدل، اسناد مرتبط بازیابی و رتبه‌بندی می‌شوند و سپس با بهره‌گیری از معماری استلزام متنی صحت ادعا تعیین می‌شود. استفاده از این چارچوب ترکیبی همراه با مدل زبانی موجب بهبود چشمگیر عملکرد نسبت به سیستم‌های پایه شد.

آتاناسووا و همکاران[۴۴] برای نخستین‌بار بر ضرورت توانایی سیستم‌های صحت‌سنجی در تشخیص «عدم کفایت شواهد» تأکید کردند و با استفاده از روش حذف بخش‌هایی از شواهد و آزمایش روی مدل‌های مبدل مختلف نشان دادند که نبود برخی اجزای کلیدی متن موجب خطای سیستم می‌شود. با معرفی مجموعه‌داده تشخیصی SufficientFacts1 که توسط کارشناسان برای ارزیابی کفایت شواهد برچسب‌گذاری شده بود، نتایج حاکی از ضعف مدل‌ها در شناسایی عدم کفایت شواهد، به‌ویژه در حذف قیدها[۲۱]بود. در نهایت، با ترکیب روش حذف متن با آموزش سه‌گانه[۲۲]یک راهبرد افزایشی مبتنی بر یادگیری تقابلی[۲۳]برای بهبود این توانایی ارائه شد.

پانکوفسکی و همکاران[۴۵] رویکردی دو مرحله‌ای برای شناسایی و صحت‌سنجی ادعاهای مشکوک درباره کووید-۱۹ ارائه کردند. ابتدا جملات مشکوک در متون با استفاده از مدل‌های زبانی پیش‌آموزش‌دیده شناسایی می‌شوند و سپس در گام دوم، صحت آن‌ها با استفاده از سرویس Google Fact Check Tools بررسی می‌شود.

مونجیوی و گنگمی[۴۶] رویکردی برای بهبود بازیابی شواهد ارائه کردند که بر پایه ساخت یک گراف از کل مجموعه اسناد استوار است. در این گراف، قطعات متن به‌عنوان گره و بر اساس اشاره‌های مشترک[۲۴]به هم متصل می‌شوند و جست‌وجوی شواهد در زیرگراف‌های مرتبط انجام می‌گیرد.

وادن و همکاران[۴۷] روش MultiVerS را معرفی کردند که با رویکرد یادگیری چندوظیفه‌ای، هم برچسب صحت‌سنجی ادعا و هم استدلال‌های مرتبط را پیش‌بینی می‌کند. این مدل با استفاده از Longformer امکان رمزگذاری همزمان ادعا و متن کامل سند را فراهم می‌سازد و با بهره‌گیری از نظارت ضعیف، توانایی تعمیم در شرایط کم‌داده را نشان می‌دهد.

ژانگ و همکاران[۴۸] رویکردی برای بهبود بازیابی اولیه شواهد پیشنهاد کردند که در این روش، دانش یک مدل دقیق به یک مدل سبک منتقل می‌شود تا در مرحله‌ی بازیابی سریع، ضمن حفظ کارایی، دقت بیشتری به دست آید.

جعفری و آلن[۵] رویکردی ارائه کرده‌اند که با تمرکز بر شناسایی اجزای واقعی در متن و سپس مقایسه‌ی آن‌ها با شواهد معتبر، فرایند صحت‌سنجی ادعا را بهینه می‌سازد. این روش با ترکیب استخراج دقیق عبارات کلیدی و ارزیابی بر اساس مقایسه با شواهد معتبر، دقت و پایداری بیشتری در برابر اطلاعات غیرمرتبط فراهم می‌کند.

کورینا وهمکاران[۴۹] نیز روش QECV[۲۵] معرفی کرده‌اند که با ترکیب سه تکنیک متنوع تک‌مرحله‌ای، چندمرحله‌ای و سبک کارشناس صحت‌سنجی[۲۶]در تولید پرسش برای بازیابی شواهد و بهره‌گیری از مدل‌های زبانی بزرگ در ارزیابی، فرایند صحت‌سنجی ادعاها را در محیط‌های واقعی بهبود می‌بخشد.

---

[۲۱] Adverbial modifier
[۲۲] tri-training
[۲۳] Contrastive self-learning
[۲۴] co-mention
[۲۵] Question Enrichment Claim Verification
[۲۶] Fact-checker Style

چن و همکاران[۵۰] یک گردش کار پنج مرحله‌ای را با استفاده از شواهد موجود قبل از انتشار ارائه کردند. این روش شامل تفکیک ادعا به زیرسؤالات، بازیابی اسناد، استخراج بخش‌های مرتبط، خلاصه‌سازی متمرکز و قضاوت نهایی است.

در نهایت، ژائو و همکاران[۵۱] با معرفی چارچوب PACAR[۲۷]سعی در برنامه‌ریزی انتخاب شواهد بر اساس نوع و ترتیب و سپس استدلال عملیاتی سفارشی با استفاده از مدل GPT-3.5-turbo، دقت و توضیح‌پذیری صحت‌سنجی ادعاها را نسبت به روش‌های متکی صرف بر مدل‌های زبانی افزایش دهند.

۴،۳. روش‌های مبتنی بر جستجوی خارجی (RAG)[۲۸]

در این روش، مدل‌های زبانی از بازیابی شواهد مرتبط برای تقویت تصمیم‌گیری خود بهره می‌برند و پاسخ‌ها را با اطلاعات واقعی مقایسه می‌کنند. این رویکرد توانسته دقت صحت‌سنجی را در سناریوهای پیچیده به طور قابل توجهی افزایش دهد.

پرادیپ و همکاران[۵۲]، سامانه‌ی VERT5ERINI را ارائه کرده‌اند؛ یک چارچوب سه‌مرحله‌ای مبتنی بر T5 که شامل بازیابی چکیده‌ها، انتخاب جملات شواهد و پیش‌بینی برچسب نهایی است.

شلیختکرول و همکاران[۵۳] برای نخستین بار مسئله صحت‌سنجی ادعاها را در محیط دامنه‌باز مبتنی بر جداول بررسی کردند و رویکردی را معرفی کردند که مرکب از رتبه‌بندی مجدد و صحت‌سنجی است: ابتدا جدول‌های نامناسب با یک جستجوی ساده حذف شده و سپس از طریق مدل مبتنی بر Roberta تنظیم دقیق‌شده، ادعا مقابل مجموعه‌ای از جدول‌های مرتبط بررسی می‌شود.

هاگستروم و همکاران[۵۴] با هدف ارزیابی نحوه استفاده مدل‌های زبانی از زمینه در سناریوهای واقعی، مجموعه‌داده‌ای به نام DRUID[۲۹]معرفی کرده‌اند که شامل پرسش‌های واقعی و زمینه‌های بازیابی‌شده است و به صورت دستی برای تشخیص موضع[۳۰] برچسب‌گذاری شده‌اند. پژوهش نشان داد که مجموعه‌داده‌های مصنوعی مانند CounterFact و ConflictQA اغلب بازنمایی غیرواقعی از زمینه‌ها دارند و نتایج گمراه‌کننده ایجاد می‌کنند. این پژوهش همچنین با معرفی معیار پیشنهادی ACU[۳۱]اهمیت استفاده از داده‌های واقعی برای بهبود ارزیابی سیستم‌های مبتنی بر RAG را برجسته ساخت.

عمر و همکاران[۵۵] رویکردی مبتنی بر تولید تقویت‌شده با بازیابی (RAG) پیشنهاد کرده‌اند که در آن شواهد مرتبط از پایگاه‌های داده معتبر بازیابی و به مدل زبانی منتقل می‌شوند تا با ترکیب این اطلاعات، تصمیم نهایی درباره صحت ادعا ارائه شود. لئو و همکاران[۳] همین روش را با استفاده از مدل زبانی SciBERT ارائه داده‌اند که صحت‌سنجی ادعاها را همراه با توضیحات مستند امکان‌پذیر می‌سازد.

---

[۲۷] Planning and Customized Action Reasoning
[۲۸] Retrieval-Augmented Generation
[۲۹] Dataset of Retrieved Unreliable, Insufficient, and Difficult-to-understand contexts
[۳۰] Stance detection
[۳۱] Augmented Context Utilization

۵،۳. روش‌های مبتنی بر یادگیری متضاد[۳۲]

روش‌های یادگیری متضاد با آموزش مدل برای تشخیص شباهت‌ها و تفاوت‌ها بین ادعا و شواهد، به بهبود انتخاب شواهد مرتبط و دقت صحت‌سنجی کمک می‌کنند.

سریرام و همکاران[۵۶] روشی بر همین اساس ارائه کرده‌اند که نمونه‌های مثبت (اسناد مرتبط) را به هم نزدیک و نمونه‌های منفی (اسناد نامرتبط) را از هم دور می‌کند.

۶،۳. روش‌های مبتنی بر خود استدلالی[۳۳] و خودپالایی[۳۴]

در این فرایند، مدل‌های زبانی ابتدا یک پاسخ یا تحلیل اولیه ارائه می‌کنند و سپس با مرور و بازاندیشی روی همان خروجی، آن را بازبینی و بهبود می‌دهند. این روش باعث افزایش دقت و انسجام استدلال در صحت‌سنجی ادعاها می‌شود، چرا که مدل می‌تواند خطاهای خود را شناسایی و اصلاح کند.

یانگ و همکاران[۵۷] روشی مبتنی بر استدلال داخلی مدل با نام خود استدلالی سازگار با برچسب[۳۵] معرفی کرده‌اند که علاوه بر صحت‌سنجی ادعاها، توضیحات مرتبط و قابل تفسیر برای تصمیم مدل تولید می‌کند.

کائو و ین[۵۸] در چارچوب MAGIC[۳۶] تولید چندین استدلال مرتبط برای هر ادعا و مکانیزم خودپالایی، کیفیت خروجی‌ها را افزایش می‌دهند و در مقایسه با مدل‌های پایه دقت و قابلیت تعمیم بالاتری ارائه می‌کنند.

شیا و همکاران[۵۹] رویکردی سه‌مرحله‌ای مبتنی بر خود استدلالی برای بهبود عملکرد مدل‌های زبان بزرگ معرفی کردند. این روش شامل ارزیابی ارتباط، انتخاب شواهد مرتبط و تحلیل مسیر استدلالی است.

خو و همکاران[۶۰] نیز در رویکرد HaHa[۳۷] از فرایند راستی‌آزمایی انسانی الهام گرفته‌اند که ادعاهای پیچیده را به زیرادعاهای ساده تقسیم کرده و با تکمیل اطلاعات و اعتبارسنجی مرحله‌ای، صحت ادعاها را تعیین می‌کند.

۷،۳. روش‌های استدلال ساختاریافته[۳۸]

در این رویکرد، مدل‌های زبانی فرآیند صحت‌سنجی را به مراحل گام‌به‌گام تقسیم می‌کنند و هر مرحله به‌طور منطقی بر مرحله قبل استوار است. این ساختاردهی باعث افزایش شفافیت و قابل‌تبیین بودن نتایج می‌شود و اجازه می‌دهد تا مسیر استدلال مدل به‌صورت واضح دنبال و ارزیابی شود.

---

[۳۲] Contrastive learning
[۳۳] Self-Reasoning
[۳۴] Self-Refinement
[۳۵] Label-adaptive Self-Rationalization
[۳۶] Multi-Argument Generation with Self-Refinement
[۳۷] Human fAct-cHecking imitation based Fact Verification
[۳۸] Structured Reasoning

ژائو و همکاران[۶۱] چارچوبی مبتنی بر معماری مبدل ارائه کردند که با افزودن سازوکار گام توجه اضافی[۳۹]توانایی مدل زبانی برای استدلال چندشاهدی[۴۰]را به طور مؤثری تقویت می‌کند. این روش با بهره‌گیری از نوعی توجه اضافی امکان پیمایش میان بخش‌های مرتبط متن را فراهم می‌آورد و در نتیجه استنتاج پیچیده‌تر و فراسندیافته‌تری را ممکن می‌سازد.

ژنگ و همکاران[۶۲] مدل QMUL-SDS را برای بازشناسی ادعاهای علمی ارائه کردند که بر پایه طبقه‌بندی دودویی گام‌به‌گام طراحی شده است. این مدل در سه مرحله شامل بازیابی چکیده، انتخاب جملات استنادی مرتبط و پیش‌بینی برچسب ادعا بر اساس ساختار تصمیم دو مرحله‌ای شرط کافی بودن میزان اطلاعات شواهد عمل می‌کند.

پن و همکاران[۶۳] رویکرد ProgramFC[۴۱]معرفی کردند که با تجزیه ادعاهای پیچیده و چندمرحله‌ای به گام‌های کوچک، صحت‌سنجی را شفاف و قابل توضیح می‌سازد. در این روش، مدل‌های زبانی بزرگ مانند Codex یا GPT-3.5 برای هر ادعا یک برنامه استنتاجی شامل زیرفرایندهایی مانند پرسش، تأیید جزء ادعا یا تصمیم‌گیری منطقی تولید می‌کنند و سپس هر زیرفرایند به یک تابع اختصاصی واگذار و اجرا می‌شود.

گانگ و همکاران[۶۴] چارچوبی به نام STRIVE ارائه می‌دهند که با تکیه بر سه مؤلفه‌ی اصلی بازیابی شواهد، استدلال چندمرحله‌ای ساختاریافته، و خودبهبودی به اعتبارسنجی ادعاها می‌پردازد. این روش با ترکیب تحلیل مرحله‌ای و بازخورد داخلی، توانسته در ادعاهای پیچیده و داده‌های چندمنبعی دقت بیشتری نسبت به روش‌های متداول ارائه دهد.

روش Bidev[۴۲]توسط لئو و همکاران[۶۵] نیز بر اعتبارسنجی ادعاهای پیچیده تمرکز دارد و با تجزیه آن‌ها به اجزای ساده‌تر و تحلیل هم‌زمان و دوطرفه‌ی ادعا و شواهد مستقل، دقت و اعتبار صحت‌سنجی را افزایش می‌دهد.

۸،۳. روش‌های مبتنی بر استنتاج زبان طبیعی[۴۳]

در این روش، مدل‌های استنتاج زبان طبیعی روی داده‌های مرتبط با صحت‌سنجی ادعاها آموزش دیده‌اند تا بتوانند رابطه منطقی بین یک ادعا و شواهد آن را تشخیص دهند. این کار باعث می‌شود مدل‌ها به‌صورت خودکار تشخیص دهند که شواهد ارائه‌شده، ادعا را تأیید، رد یا نامطمئن می‌کند و دقت نهایی افزایش یابد.

برای مثال، کاسپردیچ و همکاران[۶۶] از مدل‌های استنتاج زبان طبیعی تنظیم دقیق‌شده برای اعتبارسنجی ادعاهای علمی استفاده کرده‌اند.

۹،۳. روش‌های قابل توضیح

روش‌های قابل توضیح تلاش می‌کنند نه‌تنها صحت ادعاها را بسنجند، بلکه دلایل و ارتباط بین ادعا و شواهد را برای کاربران روشن کنند تا اعتماد و شفافیت در فرایند صحت‌سنجی افزایش یابد.

---

[۳۹] Extra Hop Attention
[۴۰] Multi-evidence reasoning
[۴۱] Program-Guided Fact-Checking
[۴۲] Bilateral Defusing Verification for Complex Claim Fact-Checking
[۴۳] Natural Language Inference

اشتمباخ و اش[۶۷] به‌جای اتکای صرف بر شواهد بازیابی‌شده، با بهره‌گیری از یادگیری کم‌نمونه و بدون نیاز به بازآموزی مدل، توانایی مدل زبانی بزرگ GPT-3 را در تولید توضیحات خودکار مختصر و آموزنده برای ارزیابی صحت ادعاها بررسی کردند.

وو و همکاران[۶۸]، مدل EHIAN[۴۴] را برای صحت‌سنجی قابل‌توضیح معرفی کردند که هم بر استخراج قطعه‌های کلیدی شواهد و هم بر اعتبار منبع تأکید دارد. این مدل با استفاده از دو لایه توجه سلسله‌مراتبی، هم ارتباط عمیق ادعا و مقالات مرتبط را تقویت می‌کند و هم با درنظرگرفتن ویژگی‌های منبع، شواهد معتبرتر را شناسایی کرده و اثر محتوای افراطی را کاهش می‌دهد.

گوراپو و همکاران[۶۹] چارچوب عصبی ExClaim را معرفی کرده‌اند که علاوه بر تعیین درستی یا نادرستی ادعا، دلایل تصمیم را به‌صورت متنی و قابل‌فهم ارائه می‌دهد. این رویکرد در راستای هوش مصنوعی توضیح‌پذیر (Explainable AI (XAI طراحی شده و هدف آن، فراتر از دقت صرف، ایجاد اعتماد انسانی از طریق توضیحات معتبر و قابل‌تأیید است.

لیانگ[۷۰] و ولادیکا و هاکایووا[۷۱] با تمرکز بر اعتبارسنجی ادعاهای پزشکی، رویکردهایی گام‌به‌گام برای اعتبارسنجی ادعاهای پزشکی ارائه کرده‌اند که علاوه بر تعیین صحت ادعا، دلایل تصمیم‌گیری مدل را نیز برای کاربر قابل فهم می‌سازند.

تن و همکاران[۷۲] نیز در این زمینه روش خود را با ایده تحلیل و محاسبه همبستگی بین ادعا و شواهد ارائه داده‌اند.

۴. مقایسه

با توجه به اهمیت صحت‌سنجی ادعاهای علمی و پیچیدگی چالش‌های موجود در این حوزه، پژوهش‌های اخیر تلاش کرده‌اند تا روش‌های مختلفی را برای بهبود دقت، توضیح‌پذیری و کارایی ارائه دهند. در ادامه، در جدول ۱ به بررسی و مقایسه این رویکردها پرداخته می‌شود تا نقاط قوت، محدودیت‌ها و زمینه‌های قابل بهبود هر روش مشخص گردد و دید جامعی از وضعیت فعلی پژوهش‌ها ارائه شود.

جدول ۱- مقایسه روش‌های صحت‌سنجی مراجع مقالات

پژوهش	سال انتشار	مجموعه‌داده	مدل / ابزار	دقت / خطا	مزایا	معایب
سی و همکاران[۲۸]	۲۰۲۰	Fact Checking Master	GBM	۹۸,۶	• دقت بسیار بالا • مجموعه‌داده واقعی	• عملکرد ضعیف برای دسته‌بندی «نیمه درست»
یانگ و همکاران[۲۹]	۲۰۲۰	FEVER	GAN	-	استفاده از یادگیری بر اساس داده‌های مثبت	امکان بیش‌برازش یا تولید مصنوعی بیش از حد داده‌ها

[۴۴] Evidence-Aware Hierarchical Interactive Attention Networks

هاتوا و همکاران[۳۰]	۲۰۲۱	FEVER	GAN	۵۱	بهبود قابل توجه در امتیاز F1 نسبت به BERT	مدل محدود به دو کلاس تأییدشده و ردشده است
رایت و همکاران[۳۱]	۲۰۲۲	CLAIMGEN-BART	BART	-	• نخستین مطالعه در زمینه تولید ادعاهای علمی • معرفی روش KBIN برای تولید مؤثر نمونه‌های منفی	محدود به حوزه‌ی زیست‌پزشکی
ژنگ و همکاران[۳۲]	۲۰۲۳	• SciFACT • FEVER • VitaminC	T-Few	• ۳۴,۲ • ۷۷,۴ • ۴۳,۹	• بهبود قطعیت داستانی مدل با ایجاد قیدهای سازگاری[۴۵] • اثبات برتری قوی نسبت به مدل‌های few-shot و zero-shot استاندارد	وابستگی به تولید متغیرهای ادعا
چادالاپاکا و همکاران[۳۳]	۲۰۲۳	مجموعه‌داده مصنوعی	• C-BERT • U-BERT	• ۰,۱۹۵ • ۰,۰۰۸۷۱	• اولین مطالعه ساختارمند در صحت‌سنجی ادعاهای داستانی • ترکیب معماری‌های قوی BERT، GATv2 و DNN	نیاز به منابع مصرفی زیاد

[۴۵] Consistency Constraints

ژنگ و همکاران[۳۴]	۲۰۲۴	• SciFACT • FEVER • Climate FEVER	T5-small	• ۶۴,۵۹ • ۳۶,۲۲ • ۴۴,۱۲	نیاز پایین به منابع محاسباتی نسبت به مدل‌های زبانی بزرگ	-
لی و همکاران[۳۵]	۲۰۲۴	• WiCE • SciFact	PaLM-2L	• ۶۴ • ۵۹,۱	تکرارپذیری و بهبود قابل توجه عملکرد نسبت به دستور مستقیم مدل زبانی	پیچیدگی محاسباتی
اشتمباخ و همکاران[۳۶]	۲۰۱۹	FEVER	BERT	۶۸,۴۶	بهره‌گیری از BERTتنظیم دقیق شده	پیچیدگی محاسباتی
ما و همکاران[۳۷]	۲۰۱۹	• Snopes • PolitiFact	HAN	• ۷۵,۹ • ۳۴	بازنمایی سطح جمله با توجه سلسله‌مراتبی و استنتاجی	پیچیدگی محاسباتی
ژو و همکاران[۳۸]	۲۰۱۹	FEVER	BERT	۶۷,۱	استدلال بین شواهد از طریق گراف	پیچیدگی محاسباتی
لئو و همکاران[۳۹]	۲۰۲۰	FEVER	GAN	۷۰,۳۸	تمرکز بهتر روی شواهد مرتبط با استفاده از کرنل‌های یال و گره	پیچیدگی محاسباتی
سلیمانی و همکاران[۴۰]	۲۰۲۰	FEVER	BERT	۶۹,۷	نرخ بازخوانی ۸۷,۱٪	پیچیدگی محاسباتی
پن و همکاران[۴۱]	۲۰۲۱	FEVER	RoBERTa	۷۷	کاهش نیاز به داده‌های برچسب‌خورده‌ی دستی	وابستگی به کیفیت مولد سوال
ژنگ و همکاران[۴۲]	۲۰۲۱	SciFact	• RoBERTa • BioBERT	• ۵۷.۵ • ۶۳,۲۷	اشتراک اطلاعات بین وظایف مختلف با قاعده‌مندسازی	پیچیدگی پیاده‌سازی
سندریال و همکاران[۴۳]	۲۰۲۲	CORD-19	BART	۴۵,۷۳	• بهبود قابل توجه بازیابی با روش ترکیبی	پیچیدگی و نیاز به منابع بالا برای اجرا

					• ساختار دو مرحله‌ای همراه با API قابل‌استفاده	
آتاناسوا و همکاران[۴۴]	۲۰۲۲	• FEVER • HoVer • VitaminC	• BERT • RoBERTa • ALBERT	• ۸۰,۷۵ • ۸۸,۶۹ • ۸۳,۳۸	بهره‌گیری از تشخیص نقص شواهد برای تقویت مدل‌های صحت‌سنجی	• پیچیدگی پیاده‌سازی • محدودیت تعمیم‌پذیری
پانکوسکی و همکاران[۴۵]	۲۰۲۲	• CORD-19 • FakeCovid • CoAID	• BERT • DistilBERT • SciBERT • RoBERTa	• ۹۸,۰۵ • ۹۷,۸۱ • ۹۷,۵۵ • ۹۷,۵	• دقت بسیار بالا در شناسایی جملات مشکوک • فرایند خودکار و استفاده از منابع بومی	اعتبارسنجی وابسته به نتایج API خارجی
مونجیووی و همکاران[۴۶]	۲۰۲۲	FEVER	GAN	۷۰,۲	توانایی استخراج شواهد پراکنده از چند سند با رویکرد گرافی	پیچیدگی محاسباتی
وادن و همکاران[۴۷]	۲۰۲۲	• HealthVer • COVIDFact • SCIFACT	RoBERTa-large	• ۷۸,۴ • ۷۷,۶ • ۷۰,۹	• ادغام اطلاعات کامل سند برای تصمیم‌گیری دقیق • قابلیت سازگاری و آموزش با داده‌های کم	پیچیدگی محاسباتی
ژنگ و همکاران[۴۸]	۲۰۲۳	SciFact	• BioBERT • RoBERTA-Large • RERRFACT	• ۷۴,۰۹ • ۷۶,۵۸ • ۸۷,۱۴	استفاده از تکنیک تقطیر دانش برای اجرای سریع بازیابی اولیه	پیچیدگی پیاده‌سازی
کورینا و همکاران[۴۹]	۲۰۲۴	AVeriTeC	• T5 • Mistral • Mixtral • GPT3.5 • GPT4	۴۲	تقویت کیفیت شواهد با سؤال‌های غنی‌شده	پیچیدگی در تولید سؤال

چن و همکاران[۵۰]	۲۰۲۴	ClaimDecomp	GPT-3	۶۰	تولید خلاصه‌های تولیدی وفادار به متن اصلی	امتیاز پایین در بازیابی شواهد
ژائو و همکاران[۵۱]	۲۰۲۴	• FEVEROUS • SciFact	gpt-3.5-turbo	• ۹۴,۴۳ • ۷۵,۰۶	استدلال مرحله‌ای و برنامه‌ریزی پویا	پیچیدگی محاسباتی
پردایپ و همکاران[۵۲]	۲۰۲۱	SciFact	T5	۷۶,۱۴	• یکپارچگی در سه مرحله • تعمیم‌پذیری بالا	پیچیدگی محاسباتی
شلیختکرول و همکاران[۵۳]	۲۰۲۱	TabFact	RoBERTa	۷۷,۶	اولین مدل حوزه باز	پیچیدگی اجرا
هاگستروم و همکاران[۵۴]	۲۰۲۴	DRUID	• Llama 3.1-8B • Pythia-6.9B	۷۱	ارائه مجموعه داده‌ی واقعی برای تحلیل رفتار RAG	محدودیت در دامنه ویژگی‌ها
عمر[۵۵]	۲۰۲۴	AVeriTeC	BERT	۲۷	تولید سؤال دقیق	هزینه محاسباتی
سریرام و همکاران[۵۶]	۲۰۲۵	• FEVER • ClaimDecomp • HotpotQA	GPT-4	• ۵۷ • ۳۲ • ۳۲	انتقال‌پذیری عملکرد به مجموعه‌های دیگر	محدودیت در تمرکز فقط بر مرحله‌ی بازیابی
یانگ و همکاران[۵۷]	۲۰۲۵	• PubHealth • AVeriTeC	• T5 • GPT-4-turbo • GPT-3.5-turbo • Llama-3-8B	• ۶۵,۶ • ۶۵,۷	کارایی few-shot برابر با مدل‌های تنظیم دقیق شده	پیچیدگی اجرایی بیشتر روش دو مرحله‌ای نسبت به تنظیم دقیق مستقیم
کائو و همکاران[۵۸]	۲۰۲۵	XClaimCheck	XLM-RoBERTa-Base	۲۶,۲۳	افزایش تطبیق استدلال با شواهد با سازوکار خودپالایی	عملکرد مدل Seq2Seq-based نسبتاً ضعیف
شیا و همکاران[۵۹]	۲۰۲۵	FEVER	• LLAMA2 • GPT-4	۸۳,۹	افزایش قابل توجه در قابلیت اطمینان و ردیابی پاسخ‌ها	خطر فزاینده در مراحل استدلال طولانی
خو و همکاران[۶۰]	۲۰۲۴	HOVER	GPT-3.5-turbo	۶۶,۵۷	عملکرد بهتر روی ادعاهای پیچیده	وابستگی به شواهد بیرونی
ژائو و همکاران[۶۱]	۲۰۲۵	HotpotQA	BERT	۶۶,۳	طراحی ساده و یکپارچه	وابستگی به تعداد گام‌ها و تنظیم مراحل

ژنگ و همکاران[۶۲]	۲۰۲۱	SciFact	BioBERT-large	۵۵,۳۵	بهبود چشمگیر در دقت و F1 نسبت به مدل پایه	قدرت تشخیص پایین داده‌های با اطلاعات ناکافی
پن و همکاران[۶۳]	۲۰۲۳	• FEVEROUS • HoVer	• T5 • RoBERTa-large	• ۵۹,۶۶ • ۵۴,۲۷	توانایی استدلال چندمرحله‌ای با ترکیب شواهد متنی و ساختاریافته	پیچیدگی محاسباتی
گانگ و همکاران[۶۴]	۲۰۲۵	HOVER	Llama-3-8B-Instruct	۷۰,۵۵	شروع استدلال با داده‌های برچسب‌خورده کم	عملکرد پایین در روش‌های ساده بدون ساختار
لیو و همکاران[۶۵]	۲۰۲۵	• HOVER • Feverous	gpt-3.5-turbo	• ۷۰,۶۳ • ۹۲,۳۹	بهبود چشمگیر عملکرد در مواجهه با ادعاهای پیچیده و چندمرحله‌ای	وابستگی به عملکرد و توانایی مدل زبانی در نقش‌های مختلف
کاسپردیک و همکاران[۶۶]	۲۰۲۴	• SciFact • HealthVer	• RoBERTa-large • DeBERTa-large	۸۸	تنظیم دقیق شده برای تمایز ادعاهای حامی، مخالف یا فاقد اطلاعات	نیاز به منابع مصرفی زیاد
اشتمباخ و همکاران[۶۷]	۲۰۲۵	FEVER	• BERT • RoBERTa	۶۵	شفافیت و قابلیت تفسیر بالا	پیچیدگی محاسباتی
وو و همکاران[۶۸]	۲۰۲۱	• PolitiFact • Snopes	HAN	• ۳۶,۲ • ۷۹,۳	قابلیت توضیح‌پذیری از طریق سازوکار توجه	پیچیدگی معماری بیشتر نسبت به مدل‌های ساده‌تر
گوراپو و همکاران[۶۹]	۲۰۲۵	FEVER	BART	۷۰	قابلیت توضیح‌پذیری	دقت عددی پایین‌تر از مدل‌های پیشرفته‌تر
لیانگ و همکاران[۷۰]	۲۰۲۵	• SciFact • NLI4CT	GPT4o-mini	• ۹۴ • ۸۶	شفافیت تصمیم‌گیری مدل	وابستگی به کیفیت بازیابی مطالعات توسط کاربر انسانی
ولادیکا و همکاران[۷۱]	۲۰۲۵	• HealthFC • CoVERT • SciFact	• GPT 4o-mini • LLaMa 3.1 • Mixtral 8x7B • DeBERTa	• ۸۱,۷ • ۷۸,۵ • ۸۱,۲	ساختار شفاف و قابل ردیابی	وابستگی به کیفیت پرسش‌ها و شواهد تولیدشده

تن و همکاران[۷۲]	۲۰۲۵	• FEVER • CorFEVER	• Llama-2 • GPT-4	۹۳,۵۷	تفسیرپذیری بالا	وابستگی به توانایی مدل زبانی

۵. نتیجه‌گیری

با توجه به گسترش روزافزون تولید دانش و اطلاعات علمی، صحت‌سنجی ادعاها به یکی از ضروری‌ترین نیازهای پژوهش علمی تبدیل شده است. مرور انجام‌شده در این مقاله نشان می‌دهد که مسیر تکامل سامانه‌های صحت‌سنجی از مدل‌های سنتی مبتنی بر قوانین و ویژگی‌ها به‌سمت مدل‌های زبانی بزرگ و چارچوب‌های استدلالی خودکار حرکت کرده است. در این میان، تلفیق سه مؤلفه‌ی کلیدی یعنی بازیابی مؤثر شواهد، استدلال چندمرحله‌ای، و توضیح‌پذیری خروجی نقش تعیین‌کننده‌ای در عملکرد سامانه‌ها دارد.

نتایج بررسی‌ها نشان می‌دهد که هرچند مدل‌های زبانی بزرگ توانسته‌اند بهبود چشمگیری در دقت و تفسیرپذیری ایجاد کنند، اما چالش‌هایی همچون کمبود داده‌های تخصصی و دوزبانه، سوگیری در مجموعه‌داده‌ها، پیچیدگی محاسباتی، و دشواری در تشخیص شواهد ناکافی همچنان پابرجاست. از سوی دیگر، روش‌های نوین مبتنی بر خوداستدلالی، یادگیری متضاد و جستجوی تقویت‌شده با بازیابی نشان داده‌اند که با ترکیب تحلیل مرحله‌ای و بازخورد خودکار، می‌توان به تعادل میان دقت، کارایی و شفافیت دست یافت.

در مجموع، مسیر آینده‌ی پژوهش در صحت‌سنجی خودکار ادعاهای علمی به‌سوی سیستم‌های ترکیبی، چندزبانه و توضیح‌پذیر پیش می‌رود که بتوانند با بهره‌گیری از مدل‌های زبانی بزرگ و دانش تخصصی حوزه‌های مختلف، ارزیابی دقیق‌تر، مستدل‌تر و قابل اعتمادتری ارائه دهند. ایجاد مجموعه‌داده‌های جامع، ارزیابی میان‌رشته‌ای و طراحی چارچوب‌های اخلاقی برای کاهش خطاهای استنتاجی نیز از الزامات کلیدی برای پیشرفت در این حوزه به شمار می‌آیند.

۶. منابع

1. S. Mogull., “Accuracy of cited ‘facts’ in medical research articles: A review of study methodology and recalculation of quotation error rate,” *journals.plos.orgSA MogullPLoS One, 2017•journals.plos.org*, vol. 12, no. 9, Sep. 2017, doi: 10.1371/JOURNAL.PONE.0184727.

2. C. Baethge, H. Jergas, “Systematic review and meta-analysis of quotation inaccuracy in medicine,” *SpringerC Baethge, H JergasResearch Integrity and Peer Review, 2025•Springer*, vol. 10, no. 1, Jul. 2025, doi: 10.1186/S41073-025-00173-Z.

3. H. Liu *et al.*, “Retrieval augmented scientific claim verification,” *academic.oup.comH Liu, A Soroush, JG Nestor, E Park, B Idnay, Y Fang, J Pan, S Liao, M Bernard, Y PengJAMIA open, 2024•academic.oup.com*, 2024, doi: 10.1093/jamiaopen/ooae021.

4. A. Dmonte, R. Oruche, M. Zampieri, P. Calyam, and I. Augenstein, “Claim verification in the age of large language models: A survey,” *arxiv.orgA Dmonte, R Oruche, M Zampieri, P Calyam, I AugensteinarXiv preprint arXiv:2408.14317, 2024•arxiv.org*, Accessed: Feb. 25, 2025. [Online]. Available: https://arxiv.org/abs/2408.14317

5. N. Jafari and J. Allan, “Robust Claim Verification Through Fact Detection,” Jul. 2024, Accessed: Feb. 26, 2025. [Online]. Available: http://arxiv.org/abs/2407.18367

6. L. Fang, D. Fu, and V. T.-T. A. A. for S. Hypothesis, “Automatic Scientific Claims Verification with Pruned Evidence Graph,” *openreview.netL Fang, D Fu, VI TorvikTowards Agentic AI for Science: Hypothesis Generation, Comprehension …•openreview.net*, Accessed: Aug. 27, 2025. [Online]. Available: https://openreview.net/forum?id=cYAFwjY2bY

7. J. Thorne, A. Vlachos, C. Christodoulopoulos, and A. Mittal, “FEVER: A large-scale dataset for fact extraction and verification,” *NAACL HLT 2018 - 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, vol. 1, pp. 809–819, 2018, doi: 10.18653/v1/n18-1074.

8. T. Alhindi, S. Petridis, … S. M. the first workshop on fact, and undefined 2018, “Where is your evidence: Improving fact-checking by justification modeling,” *aclanthology.orgT Alhindi, S Petridis, S MuresanProceedings of the first workshop on fact extraction and verification, 2018•aclanthology.org*, pp. 85–90, 2018, Accessed: Mar. 31, 2025. [Online]. Available: https://aclanthology.org/W18-5513/

9. I. Augenstein *et al.*, “MultIFC: A real-world multi-domain dataset for evidence-based fact checking of claims,” *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, pp. 4685–4697, 2019, doi: 10.18653/v1/d19-1475.

10. W. Ostrowski *et al*, “Multi-hop fact checking of political claims,” *arxiv.org*, Accessed: Aug. 23, 2025. [Online]. Available: https://arxiv.org/abs/2009.06401

11. D. Wadden *et al.*, “Fact or fiction: Verifying scientific claims,” *EMNLP 2020 - 2020 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, pp. 7534–7550, 2020, doi: 10.18653/v1/2020.emnlp-main.609.

12. A. Sathe *et al*, “Automated fact-checking of claims from Wikipedia,” *aclanthology.orgA Sathe, S Ather, TM Le, N Perry, J ParkProceedings of the Twelfth Language Resources and Evaluation Conference, 2020•aclanthology.org*, pp. 11–16, 2020, Accessed: Mar. 31, 2025. [Online]. Available: https://aclanthology.org/2020.lrec-1.849/

13. T. Chakrabarty *et al*, “COVID-fact: Fact extraction and verification of real-world claims on COVID-19 pandemic,” *arxiv.orgA Saakyan, T Chakrabarty, S MuresanarXiv preprint arXiv:2106.03794, 2021•arxiv.org*, doi: 10.1101/2020.02.28.20029272v2.

14. A. Karamolegkou, “Argument Mining: Claim Annotation, Identification, Verification,” 2021, Accessed: Apr. 18, 2025. [Online]. Available: https://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-448855

15. J. Nørregaard, L. Derczynski, “DanFEVER: claim verification dataset for Danish,” *aclanthology.orgJ Nørregaard, L DerczynskiProceedings of the 23rd Nordic conference on computational, 2021•aclanthology.org*, Accessed: Aug. 23, 2025. [Online]. Available: https://aclanthology.org/2021.nodalida-main.47/

16. Y. Jiang *et al*, “HoVer: A dataset for many-hop fact extraction and claim verification,” *arxiv.orgY Jiang, S Bordia, Z Zhong, C Dognin, M Singh, M BansalarXiv preprint arXiv:2011.03088, 2020•arxiv.org*, Accessed: Aug. 23, 2025. [Online]. Available: https://arxiv.org/abs/2011.03088

17. M. Sarrouti, A. Ben Abacha, Y. Mrabet, and D. Demner-Fushman, “Evidence-based fact-checking of health-related claims,” *aclanthology.orgM Sarrouti, AB Abacha, Y M’rabet, D Demner-FushmanFindings of the association for computational linguistics: EMNLP 2021, 2021•aclanthology.org*, pp. 3499–3512, Accessed: Feb. 26, 2025. [Online]. Available: https://aclanthology.org/2021.findings-emnlp.297/

18. K. Khan *et al*, “WatClaimCheck: A new dataset for claim entailment and inference,” *aclanthology.orgK Khan, R Wang, P PoupartProceedings of the 60th Annual Meeting of the Association for, 2022•aclanthology.org*, vol. 1, pp. 1293–1304, Accessed: Apr. 01, 2025. [Online]. Available: https://aclanthology.org/2022.acl-long.92/

19. J. Vladika, P. Schneider, and F. Matthes, “HealthFC: A Dataset of Health Claims for Evidence-Based Medical Fact-Checking,” 2023, Accessed: Apr. 18, 2025. [Online]. Available: http://arxiv.org/abs/2309.08503

20. N. Ne¸set *et al.*, “Multi2Claim: Generating scientific claims from multi-choice questions for scientific fact-checking,” *aclanthology.orgN Tan, T Nguyen, J Bensemann, A Peng, Q Bao, Y Chen, M Gahegan, MJ WitbrockProceedings of the 17th Conference of the European Chapter of the, 2023•aclanthology.org*, pp. 2652–2664, Accessed: Apr. 18, 2025. [Online]. Available: https://aclanthology.org/2023.eacl-main.194/?trk=public_post_comment-text

21. Y. Cao *et al.*, “Can large language models detect misinformation in scientific news reporting?,” *arxiv.orgY Cao, AM Nair, E Eyimife, NJ Soofi, KP Subbalakshmi, JR Wullert II, C Basu, D ShallcrossarXiv preprint arXiv:2402.14268, 2024•arxiv.org*, Accessed: Aug. 25, 2025. [Online]. Available: https://arxiv.org/abs/2402.14268

22. T. Diggelmann *et al*, “Climate-fever: A dataset for verification of real-world climate claims,” *arxiv.org*, Accessed: Apr. 01, 2025. [Online]. Available: https://arxiv.org/abs/2012.00614

23. C. Zuo *et al*, “From Claim to Evidence: Verifying Chinese Health Claims with Medical Literature,” *SpringerC Zuo, Y Liu, C Wang, R BanerjeeCCF International Conference on Natural Language Processing and Chinese Computing, 2024•Springer*, vol. 15362 LNAI, pp. 171–183, 2025, doi: 10.1007/978-981-97-9440-9_14.

24. D. Stammbach *et al*, “The Choice of Knowledge Base in Automated Claim Checking,” Nov. 2021, Accessed: Aug. 24, 2025. [Online]. Available: https://arxiv.org/pdf/2111.07795

25. L. Pan *et al*, “Investigating zero-and few-shot generalization in fact verification,” *arxiv.orgL Pan, Y Zhang, MY KanarXiv preprint arXiv:2309.09444, 2023•arxiv.org*, Accessed: Apr. 18, 2025. [Online]. Available: https://arxiv.org/abs/2309.09444

26. A. Wührl *et al*, “What Makes Medical Claims (Un)Verifiable? Analyzing Entity and Relation Properties for Fact Verification,” *EACL 2024 - 18th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference*, vol. 1, pp. 2046–2058, 2024.

27. Y. Zhang *et al*, “‘Who said it, and Why?’ Provenance for Natural Language Claims,” *aclanthology.orgY Zhang, Z Ives, D RothProceedings of the 58th Annual Meeting of the Association for, 2020•aclanthology.org*, pp. 4416–4426, Accessed: Apr. 01, 2025. [Online]. Available: https://aclanthology.org/2020.acl-main.406/

28. S. Si *et al*, “A new approach to claim check-worthiness prediction and claim verification,” *aclanthology.orgS Si, A Datta, SK NaskarProceedings of the 17th International Conference on Natural Language, 2020•aclanthology.org*, pp. 155–160, Accessed: Mar. 31, 2025. [Online]. Available: https://aclanthology.org/2020.icon-main.20/

29. F. Yang *et al*, “Claim verification under positive unlabeled learning,” *ieeexplore.ieee.org*, 2020, Accessed: Feb. 26, 2025. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9381336/

30. A. Hatua *et al*, “On the Feasibility of Using GANs for Claim Verification-Experiments and Analysis.,” *par.nsf.govA Hatua, AM Mukherjee, R VermaIn proceedings of the 2021 workshop on Reducing Online Misinformation, 2021•par.nsf.gov*, 2021, Accessed: Apr. 01, 2025. [Online]. Available: https://par.nsf.gov/servlets/purl/10292222

31. D. Wright *et al.*, “Generating scientific claims for zero-shot scientific fact checking,” *arxiv.orgD Wright, D Wadden, K Lo, B Kuehl, A Cohan, I Augenstein, LL WangarXiv preprint arXiv:2203.12990, 2022•arxiv.org*, Accessed: Apr. 18, 2025. [Online]. Available: https://arxiv.org/abs/2203.12990

32. F. Zeng *et al*, “Prompt to be consistent is better than self-consistent? few-shot and zero-shot fact verification with pre-trained language models,” *arxiv.org*, Accessed: Aug. 24, 2025. [Online]. Available: https://arxiv.org/abs/2306.02569

33. V. Chadalapaka *et al*, “Low-Shot Learning for Fictional Claim Verification,” *arxiv.orgV Chadalapaka, D Nguyen, JW Choi, S Joshi, M RostamiarXiv preprint arXiv:2304.02769, 2023•arxiv.org*, Accessed: Aug. 24, 2025. [Online]. Available: https://arxiv.org/abs/2304.02769

34. X. Zeng, “Few-shot claim verification for automated fact checking,” 2024, Accessed: Aug. 25, 2025. [Online]. Available: https://qmro.qmul.ac.uk/xmlui/handle/123456789/97718

35. X. Li *et al*, “Minimal evidence group identification for claim verification,” *arxiv.orgX Li, S Chen, R Kapadia, J Ouyang, F ZhangarXiv preprint arXiv:2404.15588, 2024•arxiv.org*, Accessed: Aug. 25, 2025. [Online]. Available: https://arxiv.org/abs/2404.15588

36. D. Stammbach *et al*, “Team DOMLIN: Exploiting evidence enhancement for the FEVER shared task,” *aclanthology.orgD Stammbach, G NeumannProceedings of the Second Workshop on Fact Extraction and, 2019•aclanthology.org*, pp. 105–109, 2019, Accessed: Apr. 01, 2025. [Online]. Available: https://aclanthology.org/D19-6616/

37. W. Gao *et al.*, “Sentence-level evidence embedding for claim verification with hierarchical attention networks,” pp. 2561–2571, 2019, Accessed: Mar. 31, 2025. [Online]. Available: https://ink.library.smu.edu.sg/sis_research/4557/

38. J. Zhou *et al.*, “Gear: Graph-based evidence aggregating and reasoning for fact verification,” *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, pp. 892–901, 2020, doi: 10.18653/v1/p19-1085.

39. Z. Liu *et al*, “Fine-grained fact verification with kernel graph attention network,” *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 7342–7351, 2020, doi: 10.18653/v1/2020.acl-main.655.

40. A. Soleimani *et al*, “Bert for evidence retrieval and claim verification,” *SpringerA Soleimani, C Monz, M WorringEuropean Conference on Information Retrieval, 2020•Springer*, vol. 12036 LNCS, pp. 359–366, 2020, doi: 10.1007/978-3-030-45442-5_45.

41. L. Pan *et al*, “Zero-shot Fact Verification by Claim Generation,” *ACL-IJCNLP 2021 - 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, vol. 2, pp. 476–483, 2021, doi: 10.18653/v1/2021.acl-short.61.

42. Z. Zhang *et al*, “Abstract, rationale, stance: A joint model for scientific claim verification,” *arxiv.orgZ Zhang, J Li, F Fukumoto, Y YearXiv preprint arXiv:2110.15116, 2021•arxiv.org*, Accessed: Aug. 23, 2025. [Online]. Available: https://arxiv.org/abs/2110.15116

43. M. Sundriyal *et al*, “Document retrieval and claim verification to mitigate COVID-19 misinformation,” *aclanthology.orgM Sundriyal, G Malhotra, MS Akhtar, S Sengupta, A Fano, T ChakrabortyProceedings of the Workshop on Combating Online Hostile Posts in, 2022•aclanthology.org*, pp. 66–74, 2022, Accessed: Feb. 26, 2025. [Online]. Available: https://aclanthology.org/2022.constraint-1.8/

44. P. Atanasova *et al*, “Fact Checking with Insufficient Evidence,” *Trans Assoc Comput Linguist*, vol. 10, pp. 746–763, Jul. 2022, doi: 10.1162/TACL_A_00486/112498.

45. E. Pankovska *et al*, “Suspicious Sentence Detection and Claim Verification in the COVID-19 Domain.,” *dfki.deE Pankovska, K Schulz, G RehmROMCIR@ ECIR, 2022•dfki.de*, 2022, Accessed: Apr. 01, 2025. [Online]. Available: https://www.dfki.de/fileadmin/user_upload/import/12320_paper.pdf

46. M. Mongiovì *et al*, “Graph-Based Retrieval for claim verification Over cross-Document evidence,” *SpringerM Mongiovì, A GangemiInternational Conference on Complex Networks and Their Applications, 2021•Springer*, vol. 1016, pp. 486–495, 2022, doi: 10.1007/978-3-030-93413-2_41.

47. D. Wadden *et al*, “MULTIVERS: Improving scientific claim verification with weak supervision and full-document context,” *NAACL 2022 - 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference*, pp. 175–190, 2022, doi: 10.18653/v1/2022.findings-naacl.6.

48. Z. Zhang, J. Li, and F. Fukumoto, “An Efficient Approach for Improving the Recall of Rough Abstract Retrieval in Scientific Claim Verification,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 14261 LNCS, pp. 63–74, 2023, doi: 10.1007/978-3-031-44198-1_6.

49. S. Churina et al, “Improving evidence retrieval on claim verification pipeline through question enrichment,” *aclanthology.orgS Churina, A Barik, S PhayeProceedings of the Seventh Fact Extraction and VERification Workshop, 2024•aclanthology.org*, pp. 64–70, 2024, Accessed: Aug. 25, 2025. [Online]. Available: https://aclanthology.org/2024.fever-1.6/

50. J. Chen *et al*, “Complex Claim Verification with Evidence Retrieved in the Wild,” *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2024*, vol. 1, pp. 3569–3587, 2024, doi: 10.18653/v1/2024.naacl-long.196.

51. X. Zhao *et al*, “PACAR: Automated fact-checking with planning and customized action reasoning using large language models,” *aclanthology.orgX Zhao, L Wang, Z Wang, H Cheng, R Zhang, KF WongProceedings of the 2024 Joint International Conference on, 2024•aclanthology.org*, p. 12564, 1257, Accessed: Feb. 26, 2025. [Online]. Available: https://aclanthology.org/2024.lrec-main.1099/

52. R. Pradeep, X. Ma, R. Nogueira, and J. Lin, “Scientific Claim Verification with VERT5ERINI,” Oct. 2020, Accessed: Aug. 23, 2025. [Online]. Available: https://arxiv.org/pdf/2010.11930

53. M. Schlichtkrull *et al*, “Joint verification and reranking for open fact checking over tables,” *arxiv.orgM Schlichtkrull, V Karpukhin, B Oğuz, M Lewis, W Yih, S RiedelarXiv preprint arXiv:2012.15115, 2020•arxiv.org*, Accessed: Apr. 01, 2025. [Online]. Available: https://arxiv.org/abs/2012.15115

54. L. Hagström *et al*, “A reality check on context utilisation for retrieval-augmented generation,” *arxiv.orgL Hagström, SV Marjanović, H Yu, A Arora, C Lioma, M Maistro, P Atanasova, I AugensteinarXiv preprint arXiv:2412.17031, 2024•arxiv.org*, Accessed: Aug. 25, 2025. [Online]. Available: https://arxiv.org/abs/2412.17031

55. A. Adjali, “Exploring Retrieval Augmented Generation For Real-world Claim Verification,” *aclanthology.orgA OmarProceedings of the Seventh Fact Extraction and VERification Workshop, 2024•aclanthology.org*, pp. 113–117, 2024, Accessed: Apr. 01, 2025. [Online]. Available: https://aclanthology.org/2024.fever-1.13/

56. A. Sriram *et al*, “Contrastive Learning to Improve Retrieval for Real-world Fact Checking,” *arxiv.orgA Sriram, F Xu, E Choi, G DurrettarXiv preprint arXiv:2410.04657, 2024•arxiv.org*, Accessed: Aug. 25, 2025. [Online]. Available: https://arxiv.org/abs/2410.04657

57. J. Yang *et al*, “Take it easy: Label-adaptive self-rationalization for fact verification and explanation generation,” *ieeexplore.ieee.orgJ Yang, A Rocha2024 IEEE International Workshop on Information Forensics and, 2024•ieeexplore.ieee.org*, Accessed: Aug. 25, 2025. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10810687/

58. W. Kao *et al*, “MAGIC: Multi-Argument Generation with Self-Refinement for Domain Generalization in Automatic Fact-Checking,” *aclanthology.orgWY Kao, AZ YenProceedings of the 2024 Joint International Conference on, 2024•aclanthology.org*, p. 10891, 1090, Accessed: Apr. 01, 2025. [Online]. Available: https://aclanthology.org/2024.lrec-main.951/

59. Y. Xia *et al*, “Improving Retrieval Augmented Language Model with Self-Reasoning,” Jul. 2024, Accessed: Feb. 26, 2025. [Online]. Available: http://arxiv.org/abs/2407.19813

60. B. Xu *et al*, “Complex Claim Verification via Human Fact-Checking Imitation with Large Language Models,” *ieeexplore.ieee.orgB Xu, F Sun, X Liu, P Wu, X Wang, L Pan2024 19th International Joint Symposium on Artificial Intelligence, 2024•ieeexplore.ieee.org*, Accessed: Apr. 18, 2025. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10799276/

61. C. Zhao *et al*, “Transformer-xh: Multi-evidence reasoning with extra hop attention,” *openreview.netC Zhao, C Xiong, C Rosset, X Song, P Bennett, S TiwaryInternational*

conference on learning representations, 2020•openreview.net, Accessed: Aug. 23, 2025. [Online]. Available: https://openreview.net/forum?id=r1eIiCNYwS

62. X. Zeng and A. Zubiaga, "QMUL-SDS at SCIVER: Step-by-Step Binary Classification for Scientific Claim Verification," *2nd Workshop on Scholarly Document Processing, SDP 2021 - Proceedings of the Workshop*, pp. 116–123, 2021, doi: 10.18653/v1/2021.sdp-1.15.

63. L. Pan *et al.*, "Fact-Checking Complex Claims with Program-Guided Reasoning," *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, vol. 1, pp. 6981–7004, 2023, doi: 10.18653/v1/2023.acl-long.386.

64. H. Gong *et al.*, "STRIVE: Structured Reasoning for Self-Improvement in Claim Verification," *arxiv.orgH Gong, J Li, J Wu, Q Liu, S Wu, L WangarXiv preprint arXiv:2502.11959, 2025•arxiv.org*, Accessed: Aug. 25, 2025. [Online]. Available: https://arxiv.org/abs/2502.11959

65. Y. Liu *et al.*, "Bidev: Bilateral defusing verification for complex claim fact-checking," *ojs.aaai.orgY Liu, H Sun, W Guo, X Xiao, C Mao, Z Yu, R YanProceedings of the AAAI Conference on Artificial Intelligence, 2025•ojs.aaai.org*, Accessed: Aug. 25, 2025. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/32034

66. M. Košprdi´košprdi´c *et al*, "Scientific Claim Verification with Fine-Tuned NLI Models," *scitepress.orgM Košprdic, A Ljajic, D Medvecki, B Bašaragin, N Miloševicscitepress.org*, 2024, doi: 10.5220/0012900000003838.

67. D. Stammbach *et al*, "e-fever: Explanations and summaries for automated fact checking," *research-collection.ethz.chD Stammbach, E AshProceedings of the 2020 Truth and Trust Online (TTO 2020), 2020•research-collection.ethz.ch*, Accessed: Aug. 23, 2025. [Online]. Available: https://www.research-collection.ethz.ch/bitstream/handle/20.500.11850/453826/e-Fever.pdf?sequence=1

68. L. Wu *et al*, "Evidence-aware hierarchical interactive attention networks for explainable claim verification," *researchgate.netL Wu, Y Rao, X Yang, W Wang, A NazirProceedings of the twenty-ninth international conference on, 2021•researchgate.net*, Accessed: Apr. 01, 2025. [Online]. Available: https://www.researchgate.net/profile/Lianwei-Wu/publication/342800300_Evidence-Aware_Hierarchical_Interactive_Attention_Networks_for_Explainable_Claim_Verification/links/5f731257a6fdcc008644f739/Evidence-Aware-Hierarchical-Interactive-Attention-Networks-for-Explainable-Claim-Verification.pdf

69. S. Gurrapu *et al*, "Exclaim: Explainable neural claim verification using rationalization," *ieeexplore.ieee.orgS Gurrapu, L Huang, FA Batarseh2022 IEEE 29th Annual Software Technology Conference (STC), 2022•ieeexplore.ieee.org*, Accessed: Apr. 18, 2025. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9951018/

70. S. Liang *et al*, "Explainable Biomedical Claim Verification with Large Language Models," *arxiv.orgS Liang, D SonntagarXiv preprint arXiv:2502.21014, 2025•arxiv.org*, 2025, Accessed: Aug. 25, 2025. [Online]. Available: https://arxiv.org/abs/2502.21014

71. J. Vladika *et al*, "Step-by-Step Fact Verification System for Medical Claims with Explainable Reasoning," *arxiv.orgJ Vladika, I Hacajová, F MatthesarXiv preprint arXiv:2502.14765, 2025•arxiv.org*, Accessed: Aug. 25, 2025. [Online]. Available: https://arxiv.org/abs/2502.14765

72. X. Tan *et al*, "Improving Explainable Fact-Checking with Claim-Evidence Correlations," *aclanthology.orgX Tan, B Zou, A AwProceedings of the 31st International Conference on*

Computational, 2025•aclanthology.org, pp. 1600–1612, Accessed: Aug. 25, 2025. [Online]. Available: https://aclanthology.org/2025.coling-main.108/

A review of automated methods for checking article references

Reza Seidmoradi

Mehdi Naghavi

Master student, Imam Hussein University, Tehran, Iran, reza.seidmoradi@ihu.ac.ir

Assistant Professor, Imam Hussein University, Tehran, Iran, mnaghavai@ihu.ac.ir

Abstract— Validation of scientific claims is one of the key challenges in natural language processing and artificial intelligence, which requires accurate and explainable methods due to the large volume of information and the complexity of data relationships. Recent research has used a variety of approaches, including limited data methods, evidence retrieval and trend enhancement, external search-based, structured reasoning, self-reasoning and self-refining, adversarial learning, natural language inference models, and explainable methods. Also, the generation and use of specialized and bilingual datasets has played an important role in improving the performance of these systems. In this paper, by categorizing and analyzing existing methods and datasets, the strengths and limitations of each approach are examined and areas for improvement for future research are identified.

Keywords: Scientific Claim Verification, Automated Fact-Checking, Evidence Retrieval, Large Language Models, Contrastive Learning, Structured Reasoning, Retrieval-Augmented Generation, Natural Language Inference