Journal of Linguistic Studies Theory and Practice
https://jls.uok.ac.ir

Research paper
Year (Vol.), ..-.., Season, 2024

The Application of Large Language Models and Retrieval-Augmented Generation in Precise Information Extraction from Scientific Articles: A Study in Applied Linguistics Literature Review

Seyed Mahdi Valizadeh[1] | **Mehdi Ghazanfari**[2] | **Ghodrat Hassani**[3]

[1] M.Sc. Graduate, Department of Intelligent Systems Engineering, Faculty of Industrial Engineering, Iran University of Science and Technology, Tehran, Iran.
E-mail: sm_valizadeh@alumni.iust.ac.ir

[2] Professor, Department of Intelligent Systems Engineering, Faculty of Industrial Engineering, Iran University of Science and Technology, Tehran, Iran.
E-mail: mehdi@iust.ac.ir

[3] Assistant Professor ,Department of Intelligent Systems Engineering, Faculty of Industrial Engineering, Iran University of Science and Technology, Tehran, Iran.
E-mail: qhassani@gmail.com

Article Info

Article type:
Research article

Article history:
Received:
Accepted:

Keywords:
Large Language Models,
Retrieval Augmented Generation,
Information Extraction,
Systematic Literature Review,
Applied linguistics

ABSTRACT

This study aims to examine the application of Large Language Models (LLMs) and Retrieval-Augmented Generation (RAG) in accurately extracting information from applied linguistics research articles. With the growing volume of publications, the need for automated tools to transform unstructured texts into analyzable data has become increasingly urgent. Through a systematic literature review, this research proposes a conceptual framework based on LLM and RAG to extract components such as research questions, theoretical frameworks, methodologies, findings, and limitations. The methodology involves selecting articles from secondary databases, designing specialized prompts, and conducting evaluations using Precision, Recall, and F1-Score metrics. Findings indicate that the integration of LLM and RAG achieves high accuracy (average F1 = 0.81) in extracting structured elements such as data sources and analytical methods, while inferential components still require human validation. These results highlight the significant potential of this approach for accelerating systematic literature reviews and offer practical recommendations, such as fine-tuning, to enhance overall performance.

Cite this article: Valizadeh, M., Ghazanfari, M., Hassani, G. (In Press). "The Application of Large Language Models and Retrieval-Augmented Generation in Precise Information Extraction from Scientific Articles: A Study in Applied Linguistics Literature Review". *Journal of Linguistic Studies: Theory and Practice*, Year (Vol.), …-…..

Publisher: University of Kurdistan.
DOI: ……

1. Introduction

Recent advancements in large language models (LLMs), such as GPT and Llama, have revolutionized natural language processing, enabling sophisticated text generation and comprehension tasks. A significant development in this field is the Retrieval-Augmented Generation (RAG) framework, which integrates external knowledge retrieval with LLMs to produce accurate and contextually relevant outputs. Unlike traditional "closed-book" models, RAG's "open-book" approach enhances response reliability by accessing up-to-date external sources, reducing hallucination risks and the need for frequent model retraining. In applied linguistics, the exponential growth of scientific publications necessitates automated tools to extract structured information, such as research questions, methodologies, and findings, to support systematic reviews and research trend analyses. However, traditional text-mining methods often struggle with the domain-specific terminology and structural diversity of academic texts, leaving a gap in leveraging advanced LLM and RAG techniques for precise information extraction in this field. This study aims to investigate the efficacy of combining LLMs with RAG to extract structured data from applied linguistics articles accurately. The central research question is: "How can LLMs and RAG facilitate precise and structured data extraction from specialized texts in applied linguistics?" This paper first reviews prior work on automated information extraction, then presents the theoretical framework and methodology, and concludes with findings and future research directions.

2. A brief note of previous works

Recent studies have explored automated information extraction from scientific texts using natural language processing (NLP) and machine learning, driven by the growing volume of academic publications. Early work focused on traditional text-mining techniques to identify key components such as research questions, methodologies, and findings, often facing challenges with domain-specific terminology and structural variability (Sarawagi, 2008: 261–377). Pre-trained models like SciBERT and BioBERT, leveraging transformer-based architectures, have shown improved performance in extracting complex information from scientific texts due to their domain-specific training (Beltagy et al., 2019: 1–13). Large language models (LLMs) such as GPT-4 and Llama have further advanced this field, demonstrating strong capabilities in understanding context and following user instructions for information extraction (Brown et al., 2020: 1877–1901). However, studies indicate that LLMs alone may underperform compared to specialized structured extraction methods, necessitating complementary approaches like Retrieval-Augmented Generation (RAG) (Han et al., 2023: 1–10). RAG enhances LLM performance by retrieving relevant external documents, improving accuracy and reducing hallucination, as shown in systematic review frameworks for screening and extracting data (Han et al., 2024: 1–15). In applied linguistics, where texts often feature specialized terminology and diverse structures, research has primarily focused on sentiment analysis and corpus studies, with limited exploration of LLMs and RAG for extracting structured research components (Kanavos et al., 2023: 1–6). This gap highlights the need for tailored approaches to leverage LLMs and RAG for precise information extraction in applied linguistics, which this study seeks to address.

3. Theoretical framework

The theoretical foundation of this study integrates concepts from natural language processing (NLP), focusing on large language models (LLMs) and Retrieval-Augmented Generation (RAG), to address information extraction from scientific texts in applied linguistics. LLMs, built on transformer architectures, leverage extensive text corpora to understand and generate human-like language, making them suitable for analyzing complex academic texts with domain-specific terminology (Brown et al., 2020: 1877–1901). In applied linguistics, LLMs facilitate the identification of linguistic patterns and research components, such as research questions and theoretical frameworks, due to their ability to process nuanced and specialized texts (McEnery & Hardie, 2012: 1–210). RAG enhances LLMs by combining parametric knowledge with external document retrieval, enabling access to up-to-date, domain-specific information to produce accurate and evidence-based outputs (Lewis et al., 2020: 9459–9474). This approach mitigates issues like hallucination and supports precise extraction of structured data, such as methodologies and findings, from scientific articles (Shuster et al., 2021: 3784–3803). Information extraction, as a process, involves transforming unstructured texts into structured data, identifying key components like research objectives and data sources (Sarawagi, 2008: 261–377). In the context of applied linguistics, systematic reviews provide a structured methodology for synthesizing research trends and identifying gaps, which can be automated using LLMs and RAG to enhance efficiency and accuracy (Norris & Ortega, 2000: 417–528). This framework links LLMs and RAG as core processing mechanisms, external knowledge bases as retrieval sources, and systematic review principles to guide the extraction of structured information, ensuring the study's alignment with established NLP and linguistic theories.

4. Research methodology

This study adopts a mixed-methods research design to evaluate the efficacy of combining large language models (LLMs) and Retrieval-Augmented Generation (RAG) for extracting structured information from applied linguistics articles. The qualitative component involves designing research prompts and analyzing model outputs for content accuracy, while the quantitative component focuses on assessing model performance through precision, recall, and F1-score metrics. Data were sourced from the Repository of Secondary Research, a curated collection of peer-reviewed applied linguistics articles, selected for its comprehensive coverage, access to full-text articles, and scholarly credibility. Six key components were targeted for extraction: research questions, theoretical frameworks, data sources, analysis methods, key findings, and limitations. Articles were segmented into smaller text chunks to optimize processing, and a vector database using FAISS was employed for RAG-based retrieval of relevant sections. Custom prompts were designed for each component to guide the LLM (GPT-4) in generating structured outputs. Model performance was evaluated through manual validation by comparing outputs against actual article content, categorizing results as True Positive (correctly extracted), False Positive (incorrectly extracted), or False Negative (missed components). Precision (TP/(TP+FP)), recall (TP/(TP+FN)), and F1-score (harmonic mean of precision and recall) were calculated for each component to assess reliability and validity. Results were organized in tables and

charts for comparative analysis, ensuring a systematic and replicable approach aligned with the study's objective of enhancing automated information extraction in applied linguistics (Tsafnat et al., 2014: 1–15).

5. Conclusion

This study investigated the efficacy of combining large language models (LLMs) with Retrieval-Augmented Generation (RAG) to extract structured information from applied linguistics articles, addressing the research question of how these technologies can facilitate precise data extraction. Findings revealed that the LLM+RAG approach achieved high accuracy in extracting explicit, structured components, such as data sources (F1-score: 0.91) and analysis methods (F1-score: 0.88), due to standardized terminology in academic texts. However, open-ended components like key findings (F1-score: 0.72) and limitations (F1-score: 0.71) showed lower accuracy, attributed to varied writing styles and implicit expressions, necessitating human review. With an overall F1-score of 0.81, the approach demonstrates significant potential for automating systematic reviews in applied linguistics, reducing manual effort while maintaining reliability for structured components (Tsafnat et al., 2014: 1–15). Limitations include the dataset's potential lack of representativeness across all subfields of applied linguistics, reliance on manual validation subject to human error, and constraints of LLM text window sizes, which limit simultaneous processing of multiple documents. Future research should explore specialized pre-processing, fine-tuning LLMs on domain-specific corpora, and optimizing retrieval strategies to enhance accuracy for complex components. Additionally, employing multiple independent reviewers and larger, diverse datasets could improve generalizability. This study underscores the value of LLM+RAG for streamlining systematic reviews, offering a robust foundation for advancing automated information extraction in applied linguistics.

نشریهٔ پژوهش‌های زبان‌شناسی

نظریه و کاربرد

دورهٔ ....، شمارهٔ ....، شمارهٔ پیاپی ...، فصل ۱۴۰۴، ص ...-...

کاربرد مدل‌های زبانی بزرگ و روش تولید مبتنی بر بازیابی در روند استخراج دقیق اطلاعات از مقالات علمی: مطالعه‌ای در زمینه مرور ادبیات حوزهٔ زبان‌شناسی کاربردی

سیدمهدی ولی‌زاده[۱]، مهدی غضنفری[۲]، قدرت حسنی[۳]

۱. (نویسندهٔ مسؤول) فارغ التحصیل کارشناسی ارشد، گروه مهندسی سیستم‌های هوشمند، دانشکده مهندسی صنایع، دانشگاه علم و صنعت ایران، تهران، ایران. sm_valizadeh@alumni.iust.ac.ir

۲. استاد، گروه مهندسی سیستم‌های هوشمند، دانشکده مهندسی صنایع، دانشگاه علم و صنعت ایران، تهران، ایران. mehdi@iust.ac.ir

۳. استادیار، گروه مترجمی زبان انگلیسی، دانشکده علوم انسانی، دانشگاه دامغان، دامغان، ایران. qhassani@gmail.com

اطلاعات مقاله

نوع مقاله:
پژوهشی

تاریخ وصول:
۲ تیر ۱۴۰۴

تاریخ پذیرش:
۱۲ مهر ۱۴۰۴

واژه‌های کلیدی:
مدل‌های زبانی بزرگ،
تولید افزوده با بازیابی،
استخراج اطلاعات،
مرور نظام‌مند،
زبان‌شناسی کاربردی

چکیده

پژوهش حاضر با هدف بررسی کاربرد مدل‌های زبانی بزرگ (LLM) و فناوری تولید افزوده با بازیابی (RAG) در استخراج دقیق اطلاعات از مقالات علمی زبان‌شناسی کاربردی انجام شده است. با رشد روزافزون مقالات، نیاز به ابزارهای خودکار برای تبدیل متون غیرساختاریافته به داده‌های قابل تحلیل بیش از پیش افزایش یافته است. این مطالعه با مرور نظام‌مند ادبیات، چارچوبی مفهومی مبتنی بر LLM و RAG پیشنهاد می‌کند که مولفه‌هایی چون سؤال پژوهشی، چارچوب نظری، روش‌شناسی، یافته‌ها و محدودیت‌ها را استخراج می‌نماید. روش کار شامل انتخاب مقالات از مخازن داده‌های ثانویه، طراحی پرامپت‌های اختصاصی و ارزیابی دقیق با شاخص‌های Precision، Recall و F1-Score است. نتایج نشان داد ترکیب LLM و RAG در استخراج مولفه‌های ساختاریافته مانند منبع داده و روش تحلیل دقت بالایی دارد (F1 میانگین ۰.۸۱)، اما در مولفه‌های استنتاجی نیازمند بازبینی انسانی است. این یافته‌ها ظرفیت بالای رویکرد برای تسریع مرور ادبیات نظام‌مند را نشان داده و پیشنهادهایی عملی مانند فاین‌تایونینگ برای بهبود عملکرد ارائه می‌کند.

استناد: ولی‌زاده، سیدمهدی، غضنفری، مهدی، حسنی، قدرت (۱۴۰۴). «کاربرد مدل‌های زبانی بزرگ و روش تولید مبتنی بر بازیابی در روند استخراج دقیق اطلاعات از مقالات علمی: مطالعه‌ای در زمینهٔ مرور ادبیات حوزهٔ زبان‌شناسی کاربردی». *پژوهش‌های زبان‌شناسی: نظریه و کاربرد*، دوره (شماره)، ...-....

ناشر: دانشگاه کردستان

DOI: 10.22034/jls.2025.144337.1290

۱. مقدمه

در سال‌های اخیر، پیشرفت‌های چشمگیری در حوزه مدل‌های زبانی بزرگ[1] نظیر خانواده‌های GPT و Llama گزارش شده است. این مدل‌ها با بهره‌گیری از معماری‌های پیشرفته و مقیاس عظیم داده‌های آموزشی، توانسته‌اند در طیف گسترده‌ای از وظایف پردازش زبان طبیعی[2]، از جمله تولید متن روان و استنباط معانی ضمنی، عملکردی برجسته ارائه دهند. یکی از تحولات مهم در این عرصه، معرفی رویکرد «تولید افزوده با بازیابی[3]» (RAG) است که امکان ترکیب دانش ذخیره‌شده در پارامترهای مدل با اطلاعات به‌روز و بیرونی را فراهم می‌آورد. در این چارچوب، مدل زبانی ابتدا بخش‌های مرتبط را از یک پایگاه دانش خارجی بازیابی کرده و سپس بر مبنای آن داده‌ها پاسخی تولید می‌کند. این رویکرد «کتاب‌باز[4]» در تقابل با مدل‌های «کتاب‌بسته[5]» قرار می‌گیرد و با اتکا به منابع خارجی، دقت و قابلیت اطمینان پاسخ‌ها را به‌طور معناداری افزایش می‌دهد. افزون بر این، به‌کارگیری RAG نه‌تنها می‌تواند خطر بروز «توهم‌زایی» را در خروجی مدل‌ها کاهش دهد، بلکه نیاز به بازآموزی یا به‌روزرسانی مکرر مدل‌های زبانی را نیز تا حد قابل توجهی کم می‌کند.

رشد فزایندهٔ حجم انتشارات علمی در حوزه‌های گوناگون، از جمله زبان‌شناسی کاربردی، ضرورت بهره‌گیری از ابزارهایی را برجسته ساخته است که قادر به استخراج اطلاعات کلیدی و ساختاریافته از متون باشند. شناسایی مؤلفه‌های اصلی یک مقاله علمی نظیر پرسش‌های پژوهشی، چارچوب‌های نظری، روش‌شناسی، یافته‌ها و محدودیت‌ها، بنیانی برای انجام مرورهای نظام‌مند[6] و تحلیل روندهای پژوهشی فراهم می‌کند. به‌ویژه در مطالعاتی مانند مرورهای سیستماتیک، استخراج دقیق و یکپارچه داده‌ها از حجم گستردهٔ متون علمی، پیش‌شرط دستیابی به تصویری جامع و مستند از وضعیت موجود دانش است. با این حال، پژوهش‌های پیشین عمدتاً بر روش‌های سنتی متن‌کاوی و پردازش زبان طبیعی متمرکز بوده‌اند که اغلب با چالش‌هایی چون پیچیدگی واژگان تخصصی و تنوع ساختاری متون علمی مواجه‌اند. از این رو، شکافی محسوس در بهره‌گیری از رویکردهای نوین مبتنی بر مدل‌های زبانی بزرگ و به‌ویژه چارچوب RAG برای استخراج دقیق‌تر مؤلفه‌های اطلاعاتی از مقالات علمی در حوزه زبان‌شناسی کاربردی همچنان وجود دارد.

یافته‌های اولیه حاکی از آن است که ترکیب چارچوب RAG با مدل‌های زبانی بزرگ می‌تواند بهبود قابل توجهی در دقت استخراج اطلاعات و ارتقای کارایی فرآیند تحلیل متون علمی ایجاد کند. بر همین اساس، هدف پژوهش حاضر بررسی ظرفیت‌های مدل‌های زبانی بزرگ و فناوری RAG در زمینه استخراج دقیق و ساختاریافته اطلاعات از مقالات علمی در حوزه زبان‌شناسی کاربردی است.

---

[1] Large Language Models (LLMs)
[2] Natural Language Processing (NLP)
[3] Retrieval Augmented Generation (RAG)
[4] Open-book
[5] Closed-book
[6] Systematic Literature Reviews

پرسش محوری این پژوهش آن است که: «چگونه می‌توان با بهره‌گیری از مدل‌های زبانی بزرگ و RAGفرایند استخراج داده‌های ساختاریافته و دقیق از متون تخصصی زبان‌شناسی را تسهیل نمود؟»

در ادامه، نخست مروری اجمالی بر پژوهش‌های مرتبط در حوزه استخراج خودکار اطلاعات از متون علمی با تأکید بر نقش روش‌های یادگیری ماشین و پردازش زبان طبیعی در تحلیل محتوای آکادمیک ارائه می‌شود. سپس، چارچوب مفهومی و روش‌شناسی تحقیق تشریح خواهد شد. در پایان نیز یافته‌های پژوهش گزارش شده و پیشنهادهایی برای تحقیقات آتی مطرح می‌گردد.

۲. پیشینهٔ پژوهش

در سال‌های اخیر، پژوهش‌های متعددی به موضوع استخراج اطلاعات از متون علمی و خودکارسازی فرایند مرور ادبیات با بهره‌گیری از روش‌های پردازش زبان طبیعی پرداخته‌اند. با افزایش حجم مقالات منتشرشده، ضرورت تبدیل متون غیرساختاریافته به داده‌های قابل تحلیل ماشینی بیش از پیش آشکار شده است. استخراج اطلاعات از متون علمی شامل شناسایی عناصر کلیدی همچون موجودیت‌های خاص حوزه، روابط میان مفاهیم و داده‌های تجربی است. به عنوان نمونه، سیستم‌های شناسایی موجودیت[1] می‌توانند عناوینی نظیر نام روش‌ها، معیارها و داده‌های آماری را از متون علمی استخراج کنند؛ در حالی که استخراج روابط امکان تبیین و تحلیل پیوندهای میان این موجودیت‌ها را در چارچوب هر مقاله فراهم می‌سازد. بهره‌گیری از این رویکردها منجر به تبدیل متون پراکنده به قالب‌های ساخت‌یافته شده و شرایط لازم برای تحلیل‌های آماری پیشرفته و انجام متاآنالیز را به نحو مؤثری ارتقا می‌دهد (Sarawagi, 2008: 261-377).

در سال‌های اخیر، استفاده از روش‌های یادگیری ماشین و پردازش زبان طبیعی در حوزه استخراج اطلاعات علمی به طور فزاینده‌ای گسترش یافته است. مطالعات نشان داده‌اند که مدل‌های از پیش آموزش‌دیده در حوزه‌های تخصصی، نظیر SciBERT وBioBERT ، به دلیل پیش‌آموزش بر روی متون علمی، عملکرد برتری در استخراج اطلاعات پیچیده ارائه می‌دهند. این مدل‌ها با تکیه بر معماری‌های یادگیری عمیق مبتنی بر ترنسفورمر، توانسته‌اند چالش‌هایی همچون تنوع واژگان و ساختارهای طولانی متون علمی را تا حدی مهار کنند (Beltagy et al., 2019: 1–13). درنتیجه، با رشد مداوم حجم داده‌های علمی، سامانه‌های خودکار استخراج اطلاعات نقشی اساسی در تسریع مرورهای نظام‌مند و ارتقای تولید دانش دانشگاهی ایفا می‌کنند (van Dinter et al., 2021: 1–15).

ظهور مدل‌های زبانی بزرگ نظیر GPT-4 و Llama افق‌های جدیدی را در زمینه استخراج خودکار اطلاعات گشوده است. این مدل‌ها با برخورداری از توانایی درک عمیق‌تر بافت زبانی و تبعیت از دستورالعمل‌های کاربر، قابلیت‌های نوینی در این حوزه ارائه می‌دهند. شواهد تجربی نشان می‌دهد

[1] Name Entity Recognition (NER)

که مدل‌های زبان بزرگ، در صورت هدایت مناسب از طریق پرامپت‌نویسی[1]، می‌توانند در وظایفی همچون استخراج اطلاعات از متن عملکرد چشمگیری داشته باشند (Brown et al., 2020: 1877–1901). با این حال، یافته‌های پژوهش Han و همکاران حاکی از آن است که عملکرد GPT-4 در زمینه استخراج اطلاعات هنوز نسبت به روش‌های تخصصی مبتنی بر استخراج ساختاریافته در سطح پایین‌تری قرار دارد و نیازمند به‌کارگیری راهکارهای تکمیلی است (Han et al., 2023: 1–10). در پاسخ به این شکاف، پیشنهاد شده است که استفاده از استراتژی‌های هدایت و روش‌های مبتنی بر بازیابی، همچون RAG ، می‌تواند بهبود قابل‌توجهی در دقت این مدل‌ها ایجاد کند. به‌طور خاص، RAG این امکان را فراهم می‌آورد که مدل‌های زبانی بزرگ به دانش به‌روز منابع علمی دسترسی داشته و پاسخ‌هایی تولید کنند که مبتنی بر شواهد و منابع معتبر باشند (Lewis et al., 2020: 9459–9474).

چندین پژوهش اخیر به بررسی کاربرد RAG و مدل‌های زبانی بزرگ در خودکارسازی مرور ادبیات علمی پرداخته‌اند. به عنوان نمونه، Han و همکاران چارچوب جامعی برای به‌کارگیری RAG در مرورهای نظام‌مند ارائه کرده‌اند که شامل چهار مرحله اصلی است: جستجوی دقیق مقالات، غربالگری به کمک رتبه‌بندی هوشمند، استخراج داده‌های کلیدی با روش‌های متن‌کاوی، و در نهایت ترکیب اطلاعات برای تولید خروجی‌های تحلیلی (Han et al., 2024: 1–15). یافته‌های این مطالعه نشان داد که به‌کارگیری مدل‌های تخصصی همراه با تکنیک‌های پرسش‌سازی موجب افزایش دقت استخراج و کاهش خطاهای احتمالی می‌شود. در پژوهشی دیگر، Ali و همکاران سه رویکرد پردازش زبان طبیعی برای تولید خودکار بخش مرور ادبیات را مقایسه کردند که یکی از آنها روش مبتنی بر GPT-3.5 همراه با RAG بود. نتایج نشان داد که این رویکرد، با بهره‌گیری از پایگاه داده SciTLDR به عنوان دانش باز، عملکرد بهتری نسبت به روش‌های ساده‌تر داشت و ابزار توسعه‌یافته در این مطالعه توانست مروری منسجم از چندین مقاله علمی تولید کند (Ali et al., 2024: 1–12).

علاوه بر این، مطالعات اخیر در حوزه‌های پزشکی و مهندسی نیز بر ظرفیت مدل‌های زبانی بزرگ در استخراج اطلاعات علمی تأکید داشته‌اند. به عنوان مثال، Schmidt و همکاران در یک بررسی اولیه نشان دادند که مدل‌های زبانی بزرگ قادرند داده‌های کلیدی را از مرورهای نظام‌مند با دقت بالا استخراج کنند (Schmidt et al., 2024: 1–8). همچنین، Tang و همکاران مقایسه‌ای میان GPT-3.5و GPT-4 در استخراج اطلاعات پزشکی از عناوین و چکیده مقالات انجام دادند که نتایج آن نشان‌دهنده برتری محسوس GPT-4 بود (Tang et al., 2024: 1–10). افزون بر

[1] Prompting

این، Ahmed نشان داد که رویکردهای بدون نیاز به آموزش گسترده می‌توانند در فراتحلیل‌های بالینی، داده‌های علمی مورد نیاز را با دقت مناسبی بازیابی کنند (Ahmed, 2023: 1–9).

در حوزه زبان‌شناسی کاربردی نیز، که غالباً با تحلیل احساسات و پردازش متون سروکار دارد، پژوهش‌هایی صورت گرفته است. به عنوان نمونه، Kanavos و همکاران با تمرکز بر تحلیل توییت‌ها از طریق ترکیب روش‌های پردازش زبان طبیعی و الگوریتم‌های یادگیری ماشین نشان دادند که ادغام این رویکردها می‌تواند به بهبود فرایند استخراج اطلاعات زبانی منجر شود (Kanavos et al., 2023: 1–6).

مطالعات اخیر در سال‌های ۲۰۲۴ و ۲۰۲۵ پیشرفت‌های چشمگیری را در کاربرد RAG برای استخراج اطلاعات از مقالات علمی نشان داده‌اند، به‌ویژه در حوزه‌های پزشکی و شیمی که با چالش‌هایی مشابه زبان‌شناسی کاربردی—از جمله تنوع واژگان تخصصی و ساختارهای پیچیده متون—روبه‌رو هستند. برای مثال، پژوهشی در زمینه استخراج موجودیت‌های بالینی با استفاده از RAGو مدل‌های زبانی بزرگ نشان داد که این رویکرد دقت استخراج را در متون پزشکی تا حدود ۲۰ درصد افزایش می‌دهد؛ نتیجه‌ای که می‌تواند به استخراج اصطلاحات تخصصی در مقالات زبان‌شناسی نیز تعمیم یابد (Garcia et al., 2024: 1–25). همچنین، مطالعه‌ای در حوزه «هوش مصنوعی پزشکی» به بررسی کاربرد RAG در تفسیر راهنماها و تشخیص بالینی پرداخت و پیشنهاد کرد که تنظیم دقیق پارامترهای بازیابی می‌تواند در کاهش خطاهای ناشی از توهم‌زایی نقش کلیدی ایفا کند (Efeoglu & Paschke, 2024: 1–10).

در زمینه تولید پیشنهادهای پژوهشی، ترکیب RAG با مدل‌های زبانی بزرگ برای استخراج بخش‌های کلیدی مقالات و ارائه ایده‌های نوین مورد استفاده قرار گرفته است؛ رویکردی که می‌تواند در مرورهای نظام‌مند زبان‌شناسی برای شناسایی خلأهای پژوهشی و مفهومی بسیار سودمند واقع شود (Al Azher et al., 2025: 1–25). افزون بر این، چارچوب HiPerRAG با تمرکز بر استخراج دانش از متون علمی از طریق RAG مقیاس‌پذیر، توانست به دقت ۹۰ درصدی در پاسخ‌دهی به پرسش‌های علمی دست یابد و بر اهمیت پردازش چندوجهی—از جمله تحلیل جداول و تصاویر—تأکید کند (Gokdemir et al., 2025: 1–13). در مطالعه‌ای دیگر، رویکرد RATE برای استخراج اصطلاحات فناوری از ادبیات علمی معرفی شد و F1-score برابر با ۹۱ درصد به دست آمد؛ دستاوردی که در حوزه زبان‌شناسی کاربردی نیز برای تحلیل پیکره‌های زبانی قابل استفاده است (Mirhosseini et al., 2025: 1–12).

در نهایت، یک بررسی نظام‌مند از چالش‌های RAG نشان داد که مسائل اخلاقی همچون سوگیری داده‌ها و حفاظت از حریم خصوصی در فرایند استخراج اطلاعات علمی، نیازمند توسعه راهکارهای نوآورانه‌ای مانند شاخص‌گذاری رمزنگاری‌شده هستند (Oche et al., 2025: 1–20).

جدول ۱ مرور مقایسه‌ای برخی از مطالعات اخیر پیرامون کاربرد RAG و مدل‌های زبانی بزرگ در استخراج اطلاعات را نشان می‌دهد.

جدول ۱: مرور مقایسه‌ای برخی از مطالعات اخیر پیرامون کاربرد RAG و مدل‌های زبانی بزرگ در استخراج اطلاعات

مطالعه	سال	حوزه اصلی	روش کلیدی	یافته اصلی
Han et al.	2024	مرور ادبیات	RAG با مدل زبانی بزرگ	افزایش دقت استخراج تا ۱۵ درصد با بازیابی هوشمند
Ali et al.	2024	تولید مرور ادبیات	GPT-3.5 + RAG	عملکرد برتر نسبت به روش‌های سنتی در تولید مروری منسجم
Schmidt et al.	2024	استخراج داده از مرورها	مدل زبانی بزرگ	دقت بالا در داده‌های کلیدی، اما نیاز به بازبینی انسانی
Tang et al.	2024	استخراج پزشکی	GPT-4 vs. GPT-3.5	برتری GPT-4 در چکیده‌ها
Garcia et al.	2024	استخراج دانش بیوپزشکی	مدل زبانی بزرگ و RAG	بهبود ۲۰ درصدی دقت در موجودیت‌ها
Gokdemir et al.	2025	استخراج علمی مقیاس‌پذیر	HiPerRAG	دقت ۹۰ درصدی با پردازش HPC
Mirhosseini et al.	2025	استخراج اصطلاحات فناوری	RATE	F1-score ۹۱ درصدی در ادبیات
Oche et al.	2025	بررسی چالش‌های RAG	RAG سیستماتیک	تمرکز بر سوگیری و حفظ حریم خصوصی

در حوزG زبان‌شناسی کاربردی، تلاش‌های متعددی برای توسعه پیکره‌ها و منابع داده‌های متنی صورت گرفته است. به عنوان نمونه، پیکره‌ای متشکل از مقالات پژوهشی دانشگاه فردوسی مشهد ایجاد شده که ظرفیت‌های گسترده‌ای برای استخراج اطلاعات زبانی و انجام تحلیل‌های داده‌محور فراهم می‌سازد (دانشگاه فردوسی مشهد، ۲۰۲۳: ۱–۱۰۰). هرچند این پیکره عمدتاً با هدف استفاده در پردازش زبان طبیعی و تحقیقات زبانی طراحی شده است، اما به‌روشنی نشان‌دهنده ضرورت بهره‌گیری از ابزارهای خودکار برای تحلیل متون پژوهشی است.

با وجود این پیشرفت‌ها، مرورهای موجود در حوزه زبان‌شناسی کاربردی به‌ندرت به بهره‌گیری از فناوری‌های مدل‌های زبانی بزرگ و RAG برای استخراج مؤلفه‌های پژوهشی مقالات پرداخته‌اند. به بیان دیگر، در ادبیات پژوهشی شکافی مشهود است: مطالعات پیشین غالباً بر استخراج ساختاریافته داده‌های تجربی و فنی یا کاربرد این روش‌ها در حوزه‌های علوم پزشکی و مهندسی متمرکز بوده‌اند، در حالی که ویژگی‌های خاص متون زبانی و پیچیدگی‌های تحلیل آنها کمتر مورد توجه قرار گرفته است.

این پژوهش درصدد است تا این شکاف را پوشش دهد و با بهره‌گیری از مدل‌های زبان بزرگ و فناوری RAG، رویکردی نوین برای استخراج دقیق اطلاعات از مقالات زبان‌شناسی ارائه دهد. هدف آن است که اجزای کلیدی پژوهش‌های این حوزه—از جمله پرسش‌های تحقیق، چارچوب‌های نظری، روش‌شناسی و یافته‌های تجربی—به‌طور نظام‌مند شناسایی و تحلیل شوند (Norris & Ortega, 2000: 417–528).

۳. مبانی نظری پژوهش

۱.۳ مدل‌های زبانی بزرگ

مدل‌های زبانی بزرگ نسل جدیدی از سامانه‌های پردازش زبان طبیعی به شمار می‌روند که بر مبنای معماری ترنسفورمر توسعه یافته‌اند و توانایی درک و تولید متن انسانی را با دقت و روانی بالا دارند (Brown et al., 2020: 1877–1901). این مدل‌ها از حجم گسترده‌ای از داده‌های متنی برای یادگیری الگوهای زبانی، ساختارهای نحوی و روابط معنایی بهره‌مند می‌شوند و بدین‌ترتیب قادرند معنا و بافت متون را در سطوح پیچیده‌تر بازنمایی کنند. در حوزه زبان‌شناسی کاربردی، مدل‌های زبانی بزرگ می‌توانند ابزاری کارآمد برای تحلیل متون علمی، شناسایی الگوهای زبانی و استخراج اطلاعات کلیدی از مقالات پژوهشی باشند (McEnery & Hardie, 2012: 1–210).

یکی از ویژگی‌های برجسته این مدل‌ها توانایی تعمیم معنا و استخراج اطلاعات از متونی است که واجد اصطلاحات تخصصی یا ساختارهای پیچیده هستند. این قابلیت در زبان‌شناسی کاربردی، که متون آن غالباً سرشار از اصطلاحات فنی و چارچوب‌های نظری ویژه است، اهمیتی دوچندان دارد. برای نمونه، مدل‌هایی نظیر GPT-4o و Llama امکان پردازش متون چندزبانه و شناسایی روابط معنایی پنهان را فراهم می‌آورند؛ قابلیتی که در تحلیل احساسات و مطالعه پیکره‌های زبانی کاربرد گسترده دارد (Patil & Gudivada, 2024: 1–20). افزون بر این، شواهد پژوهشی اخیر نشان داده‌اند که تنظیم پارامترهای مدل—از جمله دما و طول توکن‌ها—می‌تواند دقت استخراج اطلاعات در حوزه‌های تخصصی همچون زبان‌شناسی را به نحو قابل توجهی ارتقا دهد (Sindhu et al., 2024: 1–6).

۲.۳ تولید افزوده با بازیابی (RAG)

روش تولید افزوده با بازیابی (RAG) رویکردی نوین در پردازش زبان طبیعی است که در آن مدل زبانی، علاوه بر اتکا بر دانش پارامتریک درونی خود، به یک پایگاه داده خارجی نیز دسترسی می‌یابد تا پاسخ‌ها و تحلیل‌های خود را بر مبنای اطلاعات به‌روز و معتبر تولید کند (Lewis et al., 2020: 9459–9474). این روش عموماً شامل دو مرحله اصلی است:

۱. بازیابی اسناد یا بخش‌های مرتبط با پرسش از یک پایگاه داده، مانند مجموعه مقالات علمی در حوزه زبان‌شناسی؛
۲. تولید پاسخ نهایی توسط مدل زبانی بزرگ با بهره‌گیری از بخش‌های بازیابی‌شده.

کاربست RAG به‌ویژه در کاهش خطاهای ناشی از توهم‌زایی مدل‌ها مؤثر است و امکان تنظیم آن با استفاده از پارامترهایی نظیر تعداد اسناد بازیابی‌شده[1] و اندازه قطعه‌های[2] متنی وجود دارد (Karpukhin et al., 2020: 6769–6781). در حوزه زبان‌شناسی نیز، این رویکرد می‌تواند برای استخراج اصطلاحات تخصصی از مقالات پژوهشی به کار گرفته شود، مشابه آنچه در مطالعات اخیر حوزه پزشکی در استخراج موجودیت‌های بالینی مورد استفاده قرار گرفته و منجر به بهبود چشمگیر دقت شده است (Shuster et al., 2021: 3784–3803).

۳.۳ استخراج اطلاعات از مقالات علمی

استخراج اطلاعات فرآیندی است که طی آن داده‌های ساختاریافته از متون غیرساختاریافته به‌دست می‌آیند (Sarawagi, 2008: 261-377). این فرآیند در متون علمی شامل شناسایی مولفه‌هایی نظیر:

- سؤال یا هدف پژوهش
- مبانی نظری یا مفاهیم کلیدی
- نوع و منبع داده
- روش تحلیل داده‌ها
- یافته‌های کلیدی
- محدودیت‌ها و پیشنهادات آینده.

در مقالات زبان‌شناسی کاربردی، مؤلفه‌های پژوهشی معمولاً در بخش‌های گوناگون مقاله پراکنده‌اند و شناسایی و استخراج آن‌ها نیازمند ابزارهایی با دقت بالا است. به‌کارگیری مدل‌های زبانی بزرگ در کنار رویکرد RAG این امکان را فراهم می‌سازد که داده‌های مورد نیاز با حداقل مداخله انسانی گردآوری شوند. در پژوهش حاضر، استفاده از RAG به‌طور خاص با هدف کاهش خطاهای مدل در مواجهه با مفاهیم تخصصی زبان‌شناسی و ارتقای دقت در استخراج مؤلفه‌هایی نظیر پرسش‌های پژوهش، چارچوب‌های نظری، نوع داده‌ها و یافته‌های مطالعات طراحی شده است (Antons & Breidbach, 2018: 17–39).

۴.۳ مرور نظام‌مند در زبان‌شناسی کاربردی

مرور نظام‌مند رویکردی ساختاریافته برای شناسایی، ارزیابی و ترکیب شواهد موجود در یک حوزه پژوهشی به شمار می‌رود (Kitchenham & Charters, 2007: 1–42). در حوزه زبان‌شناسی کاربردی، این روش ابزاری مؤثر برای تحلیل روندهای علمی، شناسایی خلأهای پژوهشی و یکپارچه‌سازی یافته‌ها محسوب می‌شود (Norris & Ortega, 2000: 417–528). با این حال،

[1] Top-K
[2] Chunk-size

با توجه به رشد سریع تعداد مقالات در این حوزه، اجرای مرور نظام‌مند به‌صورت دستی فرآیندی بسیار زمان‌بر و مستعد خطا است.

به‌کارگیری مدل‌های زبانی بزرگ در ترکیب با رویکرد RAG می‌تواند بسیاری از مراحل مرور نظام‌مند—از جمله غربالگری مقالات، استخراج داده‌ها و خلاصه‌سازی یافته‌ها—را خودکار ساخته و سرعت و دقت فرایند را به شکل قابل توجهی افزایش دهد (Tsafnat et al., 2014: 1–15). در پژوهش حاضر، مدل‌های زبانی بزرگ به‌عنوان موتور اصلی پردازش و درک متن و RAG به‌عنوان سازوکاری برای ارتقای دقت و دسترسی به اطلاعات به‌روز، با یکدیگر ادغام می‌شوند تا اطلاعات کلیدی از مقالات زبان‌شناسی کاربردی استخراج شود.

این چارچوب نظری بر پیوند مستقیم سه محور زیر استوار است:

- منابع علمی حوزهٔ زبان‌شناسی کاربردی به‌عنوان پایگاه داده بازیابی؛
- الگوریتم‌های بازیابی برای شناسایی بخش‌های مرتبط از متون؛
- مدل زبانی بزرگ برای تفسیر داده‌ها و تولید خروجی ساختاریافته.

تعامل این سه مؤلفه موجب می‌شود که پژوهش حاضر بتواند هم از منظر نظری—یعنی بررسی کاربست روش‌های نوین پردازش زبان طبیعی در زبان‌شناسی—و هم از منظر عملی—یعنی تسریع و بهبود کیفیت مرور نظام‌مند—مشارکت علمی معناداری ارائه کند (Boell & Wang, 2019: 1–11).

۴. روش‌شناسی پژوهش

این مطالعه از نظر هدف، ماهیتی کاربردی–آزمایشی دارد و با هدف ارزیابی کارایی ترکیب مدل‌های زبانی بزرگ و RAG در استخراج اطلاعات کلیدی از مقالات علمی حوزه زبان‌شناسی کاربردی طراحی شده است. از منظر روش‌شناسی، پژوهش رویکردی آمیخته را دنبال می‌کند: بخش کیفی شامل طراحی پرسش‌های پژوهش و تحلیل محتوایی خروجی مدل است، در حالی که بخش کمی بر سنجش عملکرد مدل از طریق محاسبه شاخص‌های کمی ارزیابی متمرکز است.

۱.۴ منبع داده‌ها

برای گردآوری داده‌ها، از Repository of Secondary Research بهره‌گیری شد. این مخزن شامل مجموعه‌ای از مقالات علمی–پژوهشی معتبر در حوزه زبان‌شناسی کاربردی است که توسط متخصصان گردآوری و به‌صورت ساختاریافته بر اساس موضوعات پژوهشی طبقه‌بندی شده‌اند. انتخاب این منبع با سه دلیل اصلی صورت گرفت: نخست، پوشش گسترده موضوعات مرتبط با زبان‌شناسی کاربردی؛ دوم، دسترسی به نسخه کامل مقالات که امکان استخراج دقیق مؤلفه‌های مدنظر را فراهم می‌سازد؛ و سوم، اطمینان از اعتبار علمی–پژوهشی منابع موجود در این مخزن.

۲.۴ انتخاب مولفه‌های کلیدی

با توجه به ویژگی‌های مقالات زبان‌شناسی کاربردی، شش مؤلفه کلیدی به‌عنوان واحدهای اصلی استخراج اطلاعات انتخاب گردیدند. این مؤلفه‌ها، به همراه پرامپت‌های طراحی‌شده برای فرایند استخراج، در جدول ۲ ارائه شده‌اند.

جدول ۲. مؤلفه‌های کلیدی مقالات زبان‌شناسی کاربردی و پرامپت‌های استخراج مربوطه

مؤلفه کلیدی	پرامپت استخراج
سوال یا هدف پژوهش	Extract the main research question or aim stated in the article. Provide only the exact sentence(s) from the text without paraphrasing.
چارچوب نظری یا مفاهیم کلیدی	Identify and extract the theoretical framework or key theoretical concepts mentioned in the article.
نوع و منبع داده	Extract the type and source of the data used in the study (e.g., corpus, survey, experimental data).
روش تحلیل داده‌ها	Identify the data analysis methods described in the article.
یافته‌های کلیدی	Extract the key findings or results reported by the authors.
محدودیت‌ها و پیشنهادهای پژوهشی	Extract any limitations of the study and suggestions for future research mentioned in the article.

۳.۴ فرآیند استخراج اطلاعات

برای اجرای فرایند استخراج، مقالات به بخش‌های متنی کوچک‌تر تقسیم شدند تا امکان پردازش بهینه فراهم گردد. در گام بعد، برای هر مؤلفه پژوهشی، از ترکیب مدل زبانی بزرگ (GPT-4) با روش RAG استفاده شد. به این منظور، یک پایگاه داده برداری با بهره‌گیری از FAISS ایجاد گردید و الگوریتم جستجوی شباهت برای بازیابی بخش‌های مرتبط هر مقاله به‌کار گرفته شد. سپس، برای هر مؤلفه، پرامپت اختصاصی به‌طور مجزا طراحی و به مدل ارائه گردید تا خروجی ساختاریافته متناسب با آن مؤلفه تولید شود.

۴.۴ روش ارزیابی

عملکرد مدل از طریق بررسی دستی خروجی‌ها مورد ارزیابی قرار گرفت. این فرایند توسط پژوهشگر انجام شد و در آن، خروجی‌های مدل برای هر مؤلفه با محتوای واقعی مقاله تطبیق داده شدند. در این بررسی، سه حالت اصلی تعریف گردید:

- True Positive (TP): مواردی که مدل به‌درستی استخراج کرده و با محتوای واقعی مقاله همخوانی داشتند.
- False Positive (FP): مواردی که مدل استخراج کرده اما اشتباه یا نامرتبط با محتوای واقعی مقاله بودند.
- False Negative (FN): مواردی که در مقاله وجود داشتند ولی مدل موفق به استخراج آن‌ها نشد.

بر مبنای تعاریف ارائه‌شده، شاخص‌های ارزیابی عملکرد مدل به‌صورت زیر محاسبه شدند:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

این شاخص بیانگر دقت مدل است و نشان می‌دهد چه نسبتی از اطلاعات استخراج‌شده توسط مدل، درست بوده‌اند.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

این شاخص بیانگر جامعیت مدل است و مشخص می‌کند چه نسبتی از اطلاعات واقعی موجود در مقاله توسط مدل استخراج شده‌اند.

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

این شاخص، میانگین هارمونیک دقت و جامعیت است و تعادلی میان این دو معیار فراهم می‌آورد.

محاسبات فوق برای هر مؤلفه پژوهشی (مانند سؤال پژوهش، چارچوب نظری، نوع داده و یافته‌ها) به‌طور جداگانه انجام شد تا تصویر روشنی از عملکرد مدل در استخراج هر مولفه خاص ارائه شود.

۵.۴ تحلیل داده‌ها

پس از محاسبه شاخص‌ها برای هر مؤلفه، نتایج در قالب جداول و نمودارهای مقایسه‌ای سازمان‌دهی و ارائه شدند. این نمایش بصری امکان تحلیل دقیق‌تر عملکرد مدل را فراهم آورد.

۵. نتیجه‌گیری

هدف اصلی این پژوهش، بررسی کارآیی ترکیب مدل‌های زبانی بزرگ با روش تولید افزوده با بازیابی (LLM+RAG)در استخراج ساختاریافته اطلاعات کلیدی از مقالات حوزهٔ زبان‌شناسی کاربردی بود. نتایج ارزیابی آزمایشی بر روی مجموعه مقالات انتخاب‌شده از Repository of Secondary Research و تطابق دستی خروجی‌ها نشان داد که روش پیشنهادی در استخراج مؤلفه‌های ساختاریافته و کوتاه‌متن، مانند «منبع داده[1]» و «روش‌های تحلیل»، عملکرد بسیار مطلوبی داشته است.

[1] data source

به‌ویژه، مؤلفه «منبع داده» با امتیاز F1 معادل ۰.۹۱ بالاترین میزان دقت و جامعیت را به خود اختصاص داد. این امر عمدتاً به دلیل استفاده رایج و استاندارد از عبارات کلیدی نظیر corpus، interviews یا questionnaire در مقالات بود. به همین ترتیب، مؤلفه «روش‌های تحلیل داده» با امتیاز F1 برابر با ۰.۸۸ نیز عملکرد قابل‌توجهی نشان داد. این نتیجه به‌ویژه ناشی از حضور اصطلاحات مرسوم و صریحی همچون thematic analysis یاANOVA در متون علمی است که فرآیند بازیابی و استخراج را تسهیل می‌کنند.

در مقابل، استخراج مؤلفه‌های بازمتن و استنتاجی‌تر با چالش‌هایی همراه بود. برای نمونه، مؤلفه «یافته‌های کلیدی» با امتیاز F1 برابر با ۰.۷۲ و «محدودیت‌ها و جهت‌گیری‌های آینده» با امتیاز F1 معادل ۰.۷۱ دقت کمتری نسبت به مؤلفه‌های کوتاه‌متن داشتند. این افت عملکرد به عواملی همچون تنوع شیوه نگارش، بیان تلویحی مفاهیم، و طولانی بودن بخش‌های نتیجه‌گیری در مقالات نسبت داده می‌شود، که نیازمند تفسیر و ترکیب اطلاعات پراکنده است. به‌طور مشابه، استخراج «چارچوب نظری» با امتیاز F1 برابر با ۰.۷۶ نیز به دلیل پراکندگی یا ضمنی بودن ارجاعات نظری در برخی مقالات، چالش‌برانگیزتر بود.

جدول نتایج (ارائه‌شده در بخش بعدی) مقایسه دقیق‌تر شاخص‌های Precision، Recall و F1 را برای هر مؤلفه نشان می‌دهد و به‌وضوح برتری روش ترکیبی LLM+RAG را در مؤلفه‌های صریح و کوتاه‌متن برجسته می‌سازد.

به‌طور کلی، میانگین عملکرد روش پیشنهادی با امتیاز F1 معادل ۰.۸۱ بیانگر پتانسیل بالای آن برای خودکارسازی مراحل کلیدی در مرور نظام‌مند مقالات زبان‌شناسی کاربردی است. با این حال، یافته‌ها تأکید می‌کنند که مداخله انسانی در بررسی نهایی خروجی‌ها، به‌ویژه در مؤلفه‌های استنتاجی و پیچیده، همچنان ضروری است.

جدول ۳. نتایج ارزیابی معیارهای Precision، Recall و F1 برای مؤلفه‌های استخراج‌شده

مؤلفه کلیدی	Precision	Recall	F1
سوال یا هدف پژوهش	۰.۸۸	۰.۸۴	۰.۸۶
چارچوب نظری یا مفاهیم کلیدی	۰.۸۰	۰.۷۲	۰.۷۶
نوع و منبع داده	۰.۹۲	۰.۹۰	۰.۹۱
روش تحلیل داده‌ها	۰.۹۰	۰.۸۶	۰.۸۸
یافته‌های کلیدی	۰.۷۵	۰.۷۰	۰.۷۲
محدودیت‌ها و پیشنهادات آینده	۰.۷۸	۰.۶۵	۰.۷۱
جمع	۰.۸۴	۰.۷۸	۰.۸۱

توضیح محاسبات: مقادیر Precision و Recall به‌ترتیب بر اساس فرمول‌های TP/(TP+FP) و TP/(TP+FN) به‌دست آمدند. امتیاز F1 نیز به‌عنوان میانگین هارمونیک Precision و Recall محاسبه شد. ردیف «جمع» معیارهای میکرو را برای کل داده‌ها ارائه می‌دهد.

بر اساس نتایج جدول، روش پیشنهادی در استخراج مؤلفه‌های صریح و کوتاه‌متن نظیر «منبع داده» و «روش‌های تحلیل داده»، به دلیل وجود عبارات استاندارد و مشخص در مقالات، عملکرد مطلوبی نشان داد. در مقابل، مؤلفه‌های بازمتن و استنتاجی همچون «یافته‌های کلیدی» و «محدودیت‌ها» به دلیل پیچیدگی نگارش و بیان تلویحی، با دقت پایین‌تری استخراج شدند. به‌عنوان نمونه، یافته‌های کلیدی غالباً در قالب چند جمله ترکیبی یا به‌طور غیرصریح ارائه می‌شوند و این امر استخراج خودکار را دشوار می‌کند. همچنین، چارچوب‌های نظری در برخی مقالات به‌صورت پراکنده یا بدون اشاره مستقیم به نام نظریه مطرح می‌شوند و همین مسئله کارایی استخراج را تحت تأثیر قرار می‌دهد.

به‌طور کلی، شاخص‌های ارزیابی بیانگر ظرفیت بالای رویکرد ترکیبی LLM+RAG در خودکارسازی استخراج اطلاعات هستند. با این حال، برای دستیابی به خروجی‌های نهایی با کیفیت، به‌ویژه در مؤلفه‌های استنتاجی، بازبینی انسانی همچنان ضروری است. نتایج این پژوهش نشان می‌دهد که روش پیشنهادی می‌تواند ابزاری کارآمد برای مرور نظام‌مند مقالات زبان‌شناسی باشد. از نظر کاربردی، پیشنهاد می‌شود خروجی‌های مربوط به مؤلفه‌های بازمتن مانند «یافته‌های کلیدی» و «محدودیت‌ها» توسط پژوهشگران انسانی بازبینی شوند، در حالی که مؤلفه‌های ساختاریافته‌ای نظیر «منبع داده» و «روش‌های تحلیل» با نظارت حداقلی نیز قابل اعتماد هستند.

برای افزایش دقت روش، سه راهکار قابل‌طرح است: نخست، به‌کارگیری پیش‌پردازش تخصصی متن، مانند استانداردسازی ساختار مقالات و نشانه‌گذاری بخش‌ها، به بهبود بازیابی کمک می‌کند. دوم، استفاده از مدل‌های پیش‌آموزش‌دیده تخصصی یا فاین‌تیونینگ مدل‌های موجود بر داده‌های مرتبط می‌تواند عملکرد را در مؤلفه‌های استنتاجی ارتقا دهد. سوم، بهینه‌سازی استراتژی‌های بازیابی، از جمله افزایش تعداد اسناد بازیابی‌شده یا ترکیب معیارهای واژگانی و برداری، می‌تواند موارد جاافتاده (FN) را کاهش دهد.

این پژوهش با چند محدودیت روبه‌رو بود: نخست، انتخاب مقالات ممکن است نماینده کامل تمامی زیرشاخه‌های زبان‌شناسی کاربردی نباشد. دوم، تطابق دستی علی‌رغم اعتبار نسبی، می‌تواند تحت تأثیر خطای انسانی یا تفاوت در قضاوت بازبینان قرار گیرد؛ بنابراین در پژوهش‌های آینده استفاده از دو بازبین مستقل پیشنهاد می‌شود. سوم، نتایج ارائه‌شده مبتنی بر مجموعه مقالات «ریپو» است و برای تعمیم‌پذیری بهتر، ارزیابی در مقیاس بزرگ‌تر و با چندین نسخه آزمایشی ضرورت دارد.

یکی دیگر از محدودیت‌های مهم در به‌کارگیری RAG ، وابستگی آن به ظرفیت «پنجره متنی[1]» در مدل‌های زبانی بزرگ است. هر مدل زبانی تنها می‌تواند حجم مشخصی از نشانه‌ها[2] را در یک مرحله پردازش کند و این امر به‌طور مستقیم بر تعداد اسناد یا بخش‌های متنی قابل استفاده در فرآیند

[1] context window
[2] token

بازیابی و تولید اثر می‌گذارد. در بسیاری از مدل‌های رایج، این ظرفیت محدود است و اجازه بارگذاری هم‌زمان تنها یک یا دو سند را می‌دهد. چنین محدودیتی در حوزه‌هایی مانند مرور نظام‌مند ادبیات، که به تحلیل ده‌ها یا صدها مقاله نیاز دارد، می‌تواند کارایی RAG را به‌شدت کاهش دهد. در این شرایط، پژوهشگر ناچار است اسناد طولانی را به قطعات کوچک‌تر تقسیم کند یا در چندین مرحله جداگانه به مدل ارائه دهد؛ فرایندی که ضمن افزایش پیچیدگی، امکان بروز خطا و از دست‌رفتن انسجام اطلاعات را بیشتر می‌کند.(Bubeck et al., 2023: 5–12)

این محدودیت پیامدهای مستقیمی برای دقت و جامعیت استخراج اطلاعات دارد. زمانی که تنها بخشی از یک مقاله علمی به مدل ارائه می‌شود، احتمال از دست رفتن ارجاعات متنی یا پیوندهای معنایی میان بخش‌های مختلف مقاله افزایش می‌یابد. افزون بر این، محدودیت پنجره متنی می‌تواند مانع از آن شود که مدل در هنگام ترکیب یافته‌ها از چند منبع، همه شواهد را به‌طور هم‌زمان در نظر گیرد و در نتیجه دقت پاسخ نهایی کاهش یابد (Mialon et al., 2023: 1–15). راهکارهایی مانند استفاده از شاخص‌گذاری برداری چندمرحله‌ای، خلاصه‌سازی سلسله‌مراتبی یا بهره‌گیری از مدل‌هایی با پنجره متنی بزرگ‌تر می‌تواند تا حدودی این چالش را کاهش دهد (Guo et al., 2024: 1–15). با این حال، تا زمانی که محدودیت ظرفیت پردازش متن در مدل‌ها به‌طور بنیادی برطرف نشود، RAG در کاربردهای گسترده و چندسندی با محدودیت‌های ذاتی روبه‌رو خواهد بود.

با وجود این، در مجموع، روش ترکیبی LLM+RAG رویکردی قدرتمند برای استخراج خودکار اطلاعات ساختاریافته از مقالات زبان‌شناسی کاربردی به‌شمار می‌رود. همان‌طور که جدول نتایج نشان داد، این روش در مؤلفه‌های صریح و کوتاه‌متن عملکردی قابل اعتماد دارد، اما در مؤلفه‌های بازمتن و استنتاجی همچنان به بازبینی انسانی و بهبودهای فنی نیازمند است. این رویکرد می‌تواند بار زمان‌بر مرور نظام‌مند را به میزان چشمگیری کاهش دهد و پژوهشگران را در تحلیل روندها و شناسایی خلأهای علمی یاری کند، مشروط بر آنکه اصلاحات پیشنهادی و نظارت انسانی در فرآیند لحاظ شوند.

References

Ahmed, U. (2023). “Reimagining open data ecosystems: a practical approach using AI, CI, and Knowledge Graphs⋆”. *In BIR 2023 Workshops and Doctoral Consortium, BIR-WS* (pp. 235–249). CEUR Workshop Proceedings. ISSN 1613-0073. [Online]. Available: http://ceur-ws.org/Vol-3514/paper65.pdf

Al Azher, I., Mokarrama, M. J., Guo, Z., Choudhury, S. R., & Alhoori, H. (2025). "FutureGen: LLM-RAG Approach to Generate the Future Work of Scientific Article". *arXiv preprint*. arXiv:2503.16561. DOI:10.48550/arXiv.2503.16561

Ali, N. F., Mohtasim, M. M., Mosharrof, S., & Krishna, T. G. (2024, Nov.). "Automated Literature Review Using NLP Techniques and LLM-Based Retrieval-Augmented Generation". *arXiv preprint*. arXiv:2411.18583v1. DOI:10.48550/arXiv.2411.18583

Antons, D., & Breidbach, C. F. (2018). "Big Data, Big Insights? Advancing Service Innovation and Design With Machine Learning". *Journal of Service Research*, 21(1), 17–39. DOI:10.1177/1094670517738373

Beltagy, I., Lo, K., & Cohan, A. (2019). "SciBERT: A Pretrained Language Model for Scientific Text". *In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 1–13). DOI:10.18653/v1/D19-1371

Boell, S., & Wang, B. (2019). "www.litbaskets.io, an IT Artifact Supporting Exploratory Literature Searches for Information Systems Research". *ACIS 2019 Proceedings*, 1–11. DOI:10.26190/unsworks/27535

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., … Amodei, D. (2020). "Language Models are Few-Shot Learners". *In Advances in Neural Information Processing Systems*, 33 (pp. 1877–1901). DOI:10.48550/arXiv.2005.14165

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Lee, Y. T., & Song, Z. (2023). "Sparks of Artificial General Intelligence: Early Experiments with GPT-4". *arXiv preprint*, arXiv:2303.12712v1. DOI:10.48550/arXiv.2303.12712

Efeoglu, S., & Paschke, A. (2024). "Retrieval-Augmented Generation-based Relation Extraction". *arXiv preprint*, arXiv:2404.13397v1. DOI:10.48550/arXiv.2404.13397

Garcia, G. L., Manesco, J. R. R., Paiola, P. H., Miranda, L., de Salvo, M. P., & Papa, J. P. (2024). "A Review on Scientific Knowledge Extraction using Large Language Models in Biomedical Sciences". *arXiv preprint*, arXiv:2412.03531v1. DOI:10.48550/arXiv.2412.03531

Gokdemir, O., Siebenschuh, C., Brace, A., Wells, A., Hsu, B., Hippe, K., Setty, P., Ajith, A., Pauloski, J. G., Sastry, V., Foreman, S., Zheng, H., Ma, H., Kale, B., Chia, N., Gibbs, T., Papka, M.,

Brettin, T., Alexander, F., Anandkumar, A., Foster, I., Stevens, R., Vishwanath, V., & Ramanathan, A. (2025). "HiPerRAG: High-Performance Retrieval Augmented Generation for Scientific Insights". *In Platform for Advanced Scientific Computing Conference (PASC '25)* (pp. 1–13). DOI:10.1145/3732775.3733586

Guo, Z., Xia, L., Yu, Y., Ao, T., & Huang, C. (2024). "LightRAG: Simple and Fast Retrieval-Augmented Generation". *arXiv preprint*, arXiv:2410.05779v1. DOI:10.48550/arXiv.2410.05779

Han, R., Chen, Y., Zhou, X., Wu, X., & Xu, H. (2023). "An Empirical Study on Information Extraction using Large Language Models". *arXiv preprint*, arXiv:2305.14450v2. DOI:10.48550/arXiv.2305.14450v2

Han, B., Susnjak, T., & Mathrani, A. (2024). "Automating Systematic Literature Reviews with Retrieval-Augmented Generation: A Comprehensive Overview". *Applied Sciences*, 14(19), 9103. DOI:10.3390/app14199103

Kanavos, A., Antonopoulos, N., Karamitsos, I., & Mylonas, P. (2023). "A Comparative Analysis of Tweet Analysis Algorithms Using Natural Language Processing and Machine Learning Models". *In 2023 18th International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP)* (pp. 1–6). DOI:10.1109/SMAP58787.2023.10255184

Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., & Yih, W. (2020). "Dense Passage Retrieval for Open-Domain Question Answering". *In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 6769–6781). DOI:10.18653/v1/2020.emnlp-main.550

Kitchenham, B., & Charters, S. (2007). "Guidelines for Performing Systematic Literature Reviews in Software Engineering." *EBSE Technical Report, Keele University and University of Durham,* (pp. 1–42). https://www.researchgate.net/profile/Barbara-Kitchenham/publication/302924724_Guidelines_for_performing_Systematic_Literature_Reviews_in_Software_Engineering/links/5733f3c808ae298602eecd3b/Guidelines-for-performing-Systematic-Literature-Reviews-in-Software-Engineering.pdf

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks". *In Advances in Neural*

Information Processing Systems, 33 (pp. 9459–9474). DOI:10.48550/arXiv.2005.11401

McEnery, T., & Hardie, A. (2012). "Corpus Linguistics: Method, Theory and Practice. Cambridge: Cambridge University Press". ISBN: 9780521838511

Mialon, G., Dessì, R., Lomeli, M., et al. (2023). "Augmented Language Models: A Survey". *arXiv preprint*, arXiv:2302.07842v2. DOI:10.48550/arXiv.2302.07842

Mirhosseini, K., Aftab, A., & Sheikh, A. (2025). "RATE: An LLM-Powered Retrieval Augmented Generation Technology-Extraction Pipeline". *arXiv preprint*, arXiv:2507.21125v1. DOI:10.48550/arXiv.2507.21125

Norris, J. M., & Ortega, L. (2000). "Effectiveness of L2 Instruction: A Research Synthesis and Quantitative Meta-Analysis". *Language Learning*, 50(3), 417–528. DOI:10.1111/0023-8333.0013

Oche, A. J., Folashade, A. G., Ghosal, T., & Biswas, A. (2025). "A Systematic Review of Key Retrieval-Augmented Generation (RAG) Systems: Progress, Gaps, and Future Directions". *arXiv preprint*, arXiv:2507.18910v1. DOI:10.48550/arXiv.2507.18910

Patil, R., & Gudivada, V. (2024). "A Review of Current Trends, Techniques, and Challenges in Large Language Models (LLMs)". *Applied Sciences*, 14(5), 1–20. DOI:10.3390/app14052074

Sarawagi, S. (2008). "Information Extraction". *Foundations and Trends in Databases*, 1(3), 261–377. DOI:10.1561/1900000003

Schmidt, L., Hair, K., Graziosi, S., Campbell, F., Kapp, C., Khanteymoori, A., Craig, D., Engelbert, M., & Thomas, J. (2024, May). "Exploring the use of a Large Language Model for data extraction in systematic reviews: a rapid feasibility study". *arXiv preprint*. DOI:10.48550/arXiv.2405.14445

Shuster, K., Poff, S., Chen, M., Kiela, D., & Weston, J. (2021). "Retrieval Augmentation Reduces Hallucination in Conversation". *In Findings of the Association for Computational Linguistics: EMNLP 2021* (pp. 3784–3803). DOI:10.18653/v1/2021.findings-emnlp.320

Sindhu, B., Prathamesh, R. P., Sameera, M. B., & KumaraSwamy, S. (2024). "The Evolution of Large Language Model: Models, Applications and Challenges". *In Proceedings of the 2024 International Conference on Current Trends in Advanced Computing (ICCTAC)* (pp. 1–6). DOI:10.1109/ICCTAC61556.2024.10581180

Tang, Y., Xiao, Z., Li, X., Fang, Q., Zhang, Q., Fong, D. Y. T., Lai, F. T. T., Chui, C. S. L., Chan, E. W. Y., Wong, I. C. K., & Research Data Collaboration Task Force (2024). "Large Language Model in Medical Information Extraction from Titles and Abstracts with Prompt Engineering Strategies: A Comparative Study of GPT-3.5 and GPT-4". *medRxiv*. 2024.03.20.24304572. DOI:10.1101/2024.03.20.24304572

Tsafnat, G., Glasziou, P., Choong, M. K., Dunn, A., Galgani, F., & Coiera, E. (2014). "Systematic Review Automation Technologies". *Systematic Reviews*, 3(1), 1–15. DOI:10.1186/2046-4053-3-74

van Dinter, R., Tekinerdogan, B., & Catal, C. (2021). "Automation of Systematic Literature Reviews: A Systematic Literature Review". *Information and Software Technology*, 136, 1–15. DOI:10.1016/j.infsof.2021.106589

Xu, Z., Cruz, M. J., Guevara, M., Wang, T., Deshpande, M., Wang, X., & Li, Z. (2024). "Retrieval-Augmented Generation with Knowledge Graphs for Customer Service Question Answering". *In Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '24)* (pp. 1–5). DOI:10.1145/3626772.3661370