University of Tehran
Central Library
and Documentation Center

Academic Librarianship and Information Research

Online ISSN: 2783 - 4638

Homepage: https://jlib.ut.ac.ir/

The Role of Large Language Models in Identifying Logical Fallacies: A Step towards Improving Accuracy and Transparency in the Peer Review Process

Seyed Mohammad Ali Musavian[1], Amir Manian[2], Lotfollah Nabavi[3], and Babak Sohrabi Yourtchi[4]✉

1. Department of Information Technology Management, Faculty of Technology and Industrial Management, College of Management, University of Tehran, Tehran, Iran. Email: ali.musavian@ut.ac.ir
2. Department of Information Technology Management, Faculty of Technology and Industrial Management, College of Management, University of Tehran, Tehran, Iran. Email: amanian@ut.ac.ir
3. Department of Philosophy and Logic, Faculty of Humanities, Tarbiat Modares University, Tehran, Iran. Email: nabavi_l@modares.ac.ir
4. *Corresponding Author*, Department of Information Technology Management, Faculty of Technology and Industrial Management, College of Management, University of Tehran, Tehran, Iran. Email: bsohrabi@ut.ac.ir

Article Info

Article type:
Research Article

Article history:
Received 5 August 2024
Received in revised form 20 September 2024
Accepted 22 September 2024
Available online 25 September 2024

Keywords:
large language models,
fallacies,
peer review,
LLM

ABSTRACT

Objective: This study investigates the role of large language models (LLMs) in detecting logical fallacies during the peer-review process, aiming to improve the accuracy, transparency, and reliability of scientific publications. Additionally, the research evaluates the potential of LLMs to reduce the workload on human reviewers and standardize evaluation practices.

Method: The research involved a series of experiments designed to evaluate the ability of advanced language models, such as ChatGPT (versions 4 and o1), to identify and classify logical fallacies, solve reasoning problems, and analyze academic texts of varying lengths and complexities. Standard datasets, including the ElecDeb2060 dataset and logic questions from the Iranian Ph.D. Entrance Exam, were used. Classical machine learning models, including Support Vector Machine (SVM) and Random Forest, were employed as baseline comparisons. Advanced optimization techniques and zero-shot learning approaches were applied to prepare the language models for the analyses.

Results: The results demonstrated the exceptional performance of advanced language models, particularly ChatGPT o1, which achieved 98.1% accuracy in detecting logical fallacies and 100% accuracy in solving logic problems from the Ph.D. Entrance Exam. In contrast, classical machine learning models, such as SVM and Random Forest, recorded significantly lower accuracies of 48% and 49%, respectively. Other advanced models, such as Mistral and LLama, exhibited moderate performances, with accuracies ranging from 76% to 78.5% in identifying logical fallacies. For longer and more complex texts, ChatGPT o1 maintained 100% accuracy in identifying and naming fallacies, while other models demonstrated reduced capabilities, with accuracies below 50%.

In addition to their accuracy, the advanced LLMs displayed a remarkable ability to analyze complex arguments, identify subtle logical errors, and provide structured feedback. These features highlight their potential for improving both the efficiency and the quality of the peer-review process by reducing human error and offering detailed, objective evaluations.

Conclusion: Large language models, particularly ChatGPT o1, have shown substantial potential to redefine traditional peer-review practices. These models can enhance the speed, precision, and transparency of evaluations, thereby supporting the publication of high-quality research articles. By identifying logical fallacies and cognitive biases, they offer structured feedback that aids authors in refining their work and ensures the integrity of scientific literature. However, human reviewers remain essential as final arbiters in the process, ensuring a balanced integration of AI's analytical capabilities with human expertise. This synergy can pave the way for a more robust, efficient, and transparent peer-review system, fostering progress in scientific research.

Cite this article: Musavian, S.M.A., Manian, A., Nabavi, L., & Sohrabi Yourtchi, B. (2024). The role of large language models in identifying logical fallacies: A step towards improving accuracy and transparency in the peer review process. *Academic Librarianship and Information Research*, *58* (3), 1-20. http//doi.org/10.22059/JLIB.2025.387796.1767

Publisher: University of Tehran.
DOI: http//doi.org/10.22059/JLIB.2025.387796.1767

Introduction

The peer-review process is a cornerstone of scientific research, designed to ensure the quality, rigor, and integrity of academic publications. However, this process often suffers from limitations such as human error, cognitive biases, and inconsistency among reviewers. Logical fallacies—errors in reasoning that undermine the validity of arguments—frequently go unnoticed, compromising the quality of scientific discourse. This study investigates the role of Large Language Models (LLMs) in automating the detection of logical fallacies during the peer-review process, aiming to enhance the accuracy, transparency, and reliability of evaluations. Additionally, the study explores the potential of LLMs to reduce the workload of human reviewers and standardize assessment practices.

Method

This research adopts a design science research methodology to develop, evaluate, and refine AI-based tools tailored for logical fallacy detection. The design science framework by Peffers et al. (2007) provides the structural foundation, encompassing problem identification, objective definition, artifact design, evaluation, and dissemination. This methodological approach ensures that the study systematically addresses the challenges posed by logical fallacies in peer review and delivers actionable insights.

The experimental design is structured into three phases. The first phase evaluates the ability of LLMs to identify and classify logical fallacies using a standard dataset, ElecDeb2060, which comprises fallacies from U.S. presidential debates spanning 60 years. The second phase focuses on solving logic problems from the Iranian Ph.D. Entrance Exam to test the critical reasoning capabilities of the models. The third phase examines the models' performance in analyzing academic texts of varying lengths and complexities, designed to simulate real-world peer-review scenarios. A variety of models were employed, including state-of-the-art LLMs such as ChatGPT-4 and o1, as well as traditional machine learning models like Support Vector Machines (SVM) and Random Forest, which serve as baselines. These models underwent optimization and zero-shot learning to enhance their ability to analyze text without prior domain-specific training. Performance metrics such as accuracy, F1 scores, and qualitative feedback were used to compare the models.

Results

The results of the experiments highlight the transformative potential of LLMs in detecting logical fallacies and supporting peer review. In the first phase, ChatGPT o1 achieved an impressive 98.1% accuracy in identifying logical fallacies in the ElecDeb2060 dataset, significantly outperforming classical machine learning models such as SVM and Random Forest, which achieved 48% and 49% accuracy, respectively. Other advanced models, including Mistral and LLama, demonstrated moderate performance with accuracies ranging between 76% and 78.5%. In the second phase, ChatGPT o1 excelled in solving logic problems from the Iranian Ph.D. Entrance Exam, achieving 100% accuracy. This performance not only surpassed the baseline models but also demonstrated the model's ability to reason critically and provide

structured explanations for its answers. Traditional machine learning models, by contrast, struggled with the complexity of these problems, highlighting the limitations of conventional approaches in handling nuanced logical reasoning tasks.

The third phase involved analyzing texts of varying lengths and complexities. ChatGPT o1 maintained 100% accuracy in identifying and naming fallacies, even in long and intricate academic texts. Competing models showed diminished performance, with accuracies dropping below 50% for identifying specific types of fallacies. Nonetheless, models like Mistral and LLama showed strengths in identifying the presence and location of fallacies within texts, achieving accuracies of 97.1% and 95.4%, respectively. This suggests that while advanced LLMs excel in detailed categorization, other models still offer value in broader analytical tasks.

Beyond quantitative results, the study revealed qualitative insights into the capabilities of LLMs. ChatGPT o1 demonstrated an exceptional ability to analyze complex arguments, identify subtle logical errors, and provide structured, actionable feedback. This highlights its utility not only as a diagnostic tool but also as an assistant in crafting constructive peer-review reports. These capabilities position LLMs as critical assets in enhancing the consistency and depth of the peer-review process.

Discussion

The findings underscore the limitations of traditional machine learning methods, which lack the capacity to handle the linguistic and logical nuances inherent in academic texts. By contrast, advanced LLMs, particularly ChatGPT o1, have redefined expectations for text analysis and reasoning in peer review. These models leverage advanced architectures, such as transformer-based networks, to analyze contextual relationships within text, enabling them to detect logical fallacies with remarkable precision.

The success of ChatGPT o1 in identifying and categorizing logical fallacies across diverse contexts underscores its potential to address key challenges in peer review. Logical fallacies often go unnoticed due to their subtle nature, yet their presence can significantly undermine the validity of scientific arguments. By automating their detection, LLMs can help reviewers focus on higher-order evaluation tasks, such as assessing methodological rigor and theoretical contributions. However, the integration of LLMs into peer review is not without challenges. Issues such as algorithmic bias, ethical considerations, and the need for transparency in AI decision-making must be addressed. The study highlights the importance of explainable AI (XAI) principles in ensuring that LLMs provide clear, interpretable justifications for their analyses. For instance, ChatGPT o1 not only identifies fallacies but also explains its reasoning, making it a transparent and trustworthy tool for reviewers.

Conclusion

This study provides compelling evidence for the transformative potential of LLMs, particularly ChatGPT o1, in enhancing the efficiency, accuracy, and transparency of the peer-review process. By automating the detection of logical fallacies, these models reduce the cognitive load on reviewers, mitigate the risk of human error, and promote consistency in

evaluations. The results demonstrate that LLMs outperform traditional methods by a substantial margin, offering unprecedented capabilities in logical reasoning and text analysis.

The implications of these findings extend beyond peer review to broader domains of academic publishing and scientific communication. The integration of LLMs into workflows has the potential to standardize practices, ensure methodological rigor, and uphold the integrity of scientific literature. Moreover, the use of LLMs can democratize access to high-quality reviews, enabling researchers from diverse backgrounds to benefit from rigorous, objective evaluations.

Nonetheless, human reviewers remain indispensable in the peer-review process. The study emphasizes the importance of a synergistic approach, combining the analytical power of LLMs with the contextual understanding and domain expertise of human reviewers. This collaboration can foster a more robust and equitable review system, paving the way for scientific advancements that are both rigorous and impactful.

Future research should focus on addressing the limitations identified in this study, such as the need for domain-specific fine-tuning and the mitigation of potential biases in AI models. Additionally, the development of frameworks for ethical AI deployment in peer review will be critical in ensuring that these technologies are used responsibly and effectively. By continuing to refine and integrate LLMs into scientific workflows, the academic community can unlock new possibilities for innovation, collaboration, and knowledge dissemination.

Author Contributions

All authors contributed equally to the conceptualization of the article and writing of the original and subsequent drafts.

Data Availability Statement

Data available on request from the authors.

Acknowledgements

The authors would like to thank anonymous reviewers.

Ethical Considerations

The authors avoided data fabrication, falsification, plagiarism, and misconduct.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflict of Interest

The authors declare no conflict of interest.

تحقیقات کتابداری و اطلاع رسانی دانشگاهی
شاپا الکترونیکی: ۴۶۳۸-۲۷۸۳
Homepage: https://jlib.ut.ac.ir/

نقش مدل‌های بزرگ زبانی در شناسایی مغالطات منطقی: گامی به سوی ارتقای دقت و شفافیت در فرایند داوری همتا

سیدمحمدعلی موسویان[۱]، امیر مانیان[۲]، لطف‌اله نبوی[۳]، و بابک سهرابی یورتچی[۴]✉

۱. گروه مدیریت فناوری اطلاعات، دانشکده مدیریت صنعتی و فناوری، دانشکدگان مدیریت، دانشگاه تهران، تهران، ایران. رایانامه: ali.musavian@ut.ac.ir

۲. گروه مدیریت فناوری اطلاعات، دانشکده مدیریت صنعتی و فناوری، دانشکدگان مدیریت، دانشگاه تهران، تهران، ایران. رایانامه: amanian@ut.ac.ir

۳. گروه فلسفه و منطق، دانشکده علوم انسانی، دانشگاه تربیت مدرس، تهران، ایران. رایانامه: nabavi_l@modares.ac.ir

۴. نویسنده مسئول، گروه مدیریت فناوری اطلاعات، دانشکده مدیریت صنعتی و فناوری، دانشکدگان مدیریت، دانشگاه تهران، تهران، ایران. رایانامه: bsohrabi@ut.ac.ir

اطلاعات مقاله

نوع مقاله:
مقاله پژوهشی

تاریخ دریافت: ۱۴۰۳/۰۵/۱۴
تاریخ بازنگری: ۱۴۰۳/۰۶/۲۹
تاریخ پذیرش: ۱۴۰۳/۰۶/۳۱
تاریخ انتشار: ۱۴۰۳/۰۷/۰۳

کلیدواژه‌ها:
مدل‌های بزرگ زبانی،
مغالطات،
داوری همتا

چکیده

هدف: این پژوهش به بررسی نقش مدل‌های بزرگ زبانی در شناسایی مغالطات منطقی در فرایند داوری همتا پرداخته و تأثیر این فناوری‌ها بر بهبود دقت، شفافیت و قابلیت اطمینان مقالات علمی را مورد ارزیابی قرار می‌دهد. همچنین، امکان به‌کارگیری این مدل‌ها برای کاهش بار کاری داوران انسانی و استانداردسازی ارزیابی‌ها بررسی شده است.

روش پژوهش: سه آزمایش متفاوت در پژوهش حاضر طراحی و اجرا شد که شامل شناسایی و طبقه‌بندی مغالطات منطقی، حل مسائل استدلالی و ارزیابی متون علمی با طول و پیچیدگی متغیر بود. از مجموعه داده‌های استاندارد نظیر الک‌دب۶۰-۲۰ و سؤالات بخش منطق استعداد تحصیلی در آزمون دکتری ایران استفاده شد. مدل‌های زبانی پیشرفته مانند چت‌جی‌پی‌تی نسخه‌های ۴o و o1 با روش‌های یادگیری ماشین کلاسیک نظیر ماشین بردار پشتیبان و جنگل تصادفی مقایسه شدند. مدل‌ها با استفاده از روش‌های بهینه‌سازی و یادگیری بدون نمونه برای تحلیل داده‌ها آماده شدند.

یافته‌ها: نتایج آزمایش‌ها نشان داد که چت‌جی‌پی‌تی o1 در شناسایی مغالطات منطقی به دقت ۹۸,۱ درصد و در حل مسائل منطقی آزمون استعداد تحصیلی کنکور دکتری به دقت ۱۰۰ درصد دست یافت. در مقایسه، مدل‌های سنتی یادگیری ماشین، مانند ماشین بردار پشتیبان و جنگل تصادفی، به ترتیب تنها دقت ۴۸ درصد و ۴۹ درصد داشتند. مدل‌های زبانی میسترال و لاما نیز دقتی بین ۷۶ درصد تا ۷۸/۵ درصد در شناسایی مغالطات ارائه کردند. در تحلیل متون طولانی‌تر، چت‌جی‌پی‌تی o1 دقت ۱۰۰ درصد را در شناسایی و نام‌گذاری انواع مغالطات ثبت کرد، در حالی که مدل‌های دیگر توانایی کمتری نشان دادند. همچنین مدل‌های زبانی پیشرفته در تحلیل استدلال‌های پیچیده و ارائه بازخوردهای ساختاریافته بسیار مؤثر بودند.

نتیجه‌گیری: مدل‌های بزرگ زبانی ، به ویژه چت‌جی‌پی‌تیo1، توانایی بالایی در شناسایی و تحلیل مغالطات منطقی و بهبود فرایند داوری همتا دارند. این مدل‌ها با کاهش خطاهای انسانی، افزایش سرعت داوری و ارائه تحلیل‌های دقیق، نقشی کلیدی در بهبود کیفیت مقالات علمی ایفا می‌کنند. به‌کارگیری این فناوری‌ها می‌تواند به تقویت انسجام و شفافیت در فرایندهای علمی منجر شود، هرچند نظارت نهایی داوران انسانی برای ترکیب تخصص انسانی و هوش مصنوعی ضروری است.

استناد: موسویان، سید محمدعلی؛ مانیان، امیر؛ لطف‌اله نبوی، و سهرابی، بابک؛ (۱۴۰۳). نقش مدل‌های بزرگ زبانی در شناسایی مغالطات منطقی. *تحقیقات کتابداری و اطلاع‌رسانی دانشگاهی*، ۵۸ (۳)، ۲۰-۱.

http//doi.org/ 10.22059/JLIB.2025.387796.1767

ناشر: دانشگاه تهران.

مقدمه

استدلال دکارت «می‌اندیشم پس هستم[1]!» که مشهورترین جمله در تاریخ فلسفه است، یکی از بنیانی‌ترین و بحث‌برانگیزترین جملات فلسفه مدرن است که در اثر گرانسنگ وی *تأملاتی در فلسفه اولی* طرح می‌شود. در این استدلال که به عنوان نقطه آغازین معرفت‌شناسی جدید تلقی می‌شود، دکارت در پی آن است تا از طریق شک‌کردن به تمام آنچه که می‌توان به آن شک کرد، به نقطه آغازی یقین‌آور دست یابد (مایلز[2]، ۱۹۹۹؛ مو[3]، ۲۰۰۷). دکارت از این گزاره نتیجه می‌گیرد که در حالی که می‌تواند به هر چیز دیگری شک کند؛ اما به هیچ وجه نمی‌تواند شک کند که خود او به عنوان یک فاعل متفکر وجود دارد. بر اساس این استدلال، اندیشیدن نشانه‌ای از وجود است و از آنجا که او در حال تفکر است، پس بدیهی است که وجود دارد (والاتسوس[4]، ۲۰۲۰). با این حال، استدلال دکارت در طول تاریخ فلسفه با نقدهایی از سوی فیلسوفان مختلف روبه‌رو شده است که به بررسی و ارزیابی بنیادهای منطقی و معرفت‌شناختی آن پرداخته‌اند (ایر[5]، ۱۹۵۳) و چندین مغالطه یا خطای عمده منطقی و فلسفی توسط فیلسوفان مختلف در آن شناسایی شده است:

- مصادره به مطلوب: برتراند راسل معتقد است استدلال کوگیتو حاوی یک فرض پنهان است. این استدلال مفهوم «من» را که به دنبال اثبات وجود آن است در استدلال خود پیش‌فرض می‌گیرد (راسل، ۲۰۰۱).
- مغالطه وحدت آگاهی: نیچه[6] (۱۸۸۶) بر این باور بود که دکارت به اشتباه یک من اندیشنده واحد را فرض می‌کند. در حالی که آگاهی انسان متشکل از انگیزه‌های رقابتی متعددی است که به طور همزمان اتفاق می‌افتند و فرض یک وجود اندیشنده منفرد را از اساس دچار خدشه می‌سازد.
- کانت[7] (۱۷۸۱) بر این باور است که ادراک انسان از مفهوم من و خویشتن، مستلزم ادراک جهان و موجودات دیگر است و برای ادراک من، فرد باید ابتدا وجود جهان را پذیرفته باشد. چیزی که دکارت به اشتباه به دنبال اثبات آن است.

استدلال دکارت در ظاهر سعی دارد به عنوان نخستین بنیان برای یک معرفت‌شناسی یقینی عمل کند؛ اما نقدهای متعدد از سوی فیلسوفان مختلف نشان می‌دهند که در دل این استدلال، مشکلات منطقی و معرفت‌شناختی نهفته است. این نقدها همچنان موضوعی مهم در بحث‌های فلسفی معاصر باقی مانده‌اند و به بررسی مبانی اساسی یقین و معرفت‌شناسی در دنیای معاصر ادامه می‌دهند.

یکی از نکات مهم که از تأمل در جدل میان این فیلسوفان آشکار می‌شود آن است که دکارت با همه نقش عظیمی که در بنیان نهادن دانش و روش‌شناسی مدرن داشته است، خود دچار لغزشی جدی در تأمل منطقی و استدلالی شده است و این پرسش را پیش روی هر متفکری می‌نهد که اگر شخصیتی با قابلیت‌های عقلانی، منطقی و ریاضی دکارت گرفتار مغالطه شده است چگونه می‌توان باور داشت که دیگر متفکران و صاحب‌نظران در پهنه گسترده علوم، گرفتار انواع مغالطات منطقی و خطاهای شناختی نشوند؟

۱. استدلال منطقی

در ساختار پیچیده پژوهش علمی، استدلال منطقی به عنوان مبنایی عمل می‌کند که دانش معتبر بر آن بنا می‌شود. اصول علم ایجاب می‌کنند که تفاسیر و نتیجه‌گیری‌های حاصل از داده‌ها از نظر منطقی صحیح و عاری از مغالطه‌ها و سوگیری‌های شناختی باشد (نبوی، ۱۳۸۴). به گفته ابن سینا در دانش‌نامه علایی، علمی که با منطق سنجیده نشود علم نیست (نبوی، ۱۳۸۶).

آلفرد تارسکی[8] (۱۹۴۴)، منطق‌دان برجسته، بر اهمیت اعتبار منطقی به عنوان امری ضروری نه تنها برای ریاضیات، بلکه برای همه رشته‌های علمی که بر آن تکیه دارند، تأکید کرده است. به عقیده تارسکی، دقت منطقی مبنایی را برای وضوح، دقت و انسجام

---

1. Cogito, ergo sum
2. Miles
3. Mo
4. Valatsos
5. Ayer
6. Nietzsche
7. Kant
8. Tarski

در استدلال‌های علمی تشکیل می‌دهد. از نظر تارسکی، قوت یک نظریه علمی در ساختار منطقی آن نهفته است، زیرا هر استدلال معتبر در علم باید مبتنی بر استدلال صحیح باشد تا از ابهام و تفسیر نادرست ذهنی و سلیقه‌ای جلوگیری شود (تارسکی، ۱۹۴۴). مغالطه در تعریف به معنای هرگونه خطا و لغزش فکری است؛ اما در تعریف دقیق منطقی استدلال ظاهرا معتبری است که به منظور نادرست نشان دادن یک اندیشه درست یا درست نمایش دادن اندیشه‌ای باطل به کار گرفته می‌شود (خندان، ۱۳۸۴). مغالطه دارای انواع بسیاری است؛ اما به‌طور کلی می‌توان آن را به دو دسته مغالطات صوری و مادی تقسیم نمود (نبوی، ۱۳۸۴). شناسایی مغالطات منطقی به دلایل متعددی یک کار چالش‌برانگیز است. مغالطات منطقی اغلب شامل نقض ظریف هنجارهای استدلال است و درک و شناسایی آنها مستلزم درکی عمیق‌تر نسبت به درک ما از نحو و ساختار ظاهری زبان است و نیازمند نوعی معناشناسی و سمنتیک سطح بالا است. ماهیت شهودی و ذهنی شناسایی مغالطات حتی می‌تواند منجر به اختلاف نظر بین متخصصان شود و طبقه‌بندی آنها را پیچیده کند. به علاوه، استدلال‌های حاوی مغالطه‌ها اغلب بر دانش ضمنی یا پیش‌زمینه افراد تکیه می‌کنند و برای تشخیص مؤثر آنها ضروری است سازوکارهای خاصی برای آشکار ساختن این دانش ضمنی به‌کار گرفته شوند (سوراتی[1] و دیگران، ۲۰۲۴).

۲. داوری همتا

داوری همتا یک روش اساسی و مهم در پژوهش علمی است که هدف آن بهبود کیفیت تحقیق و ترویج دقت علمی است (الی[2] و دیگران، ۲۰۲۳). شناسایی مغالطات در طول این فرایند به اطمینان از کیفیت و یکپارچگی تحقیقات منتشر شده کمک می‌کند. با شناسایی خطاهای منطقی، استدلال‌های ناقص یا ادعاهای پشتیبانی نشده، داوران می‌توانند بازخورد سازنده‌ای را به نویسندگان ارائه دهند و از انتشار اطلاعات گمراه کننده یا نادرست جلوگیری کنند. اما فرایند داوری همتا خود از سوگیری‌ها و مغالطه‌ها مصون نیست. همان‌طور که اشاره شده است، سوگیری‌هایی مانند سوگیری تأیید و سوگیری اعتماد بیش از حد، ممکن است بر روند داوری نیز تأثیر بگذارد. این امر اهمیت هوشیاری داوران را نه تنها در مورد اشتباهات در متن مقالاتی که ارزیابی می‌کنند بلکه در استدلال و داوری‌های خودشان نیز برجسته می‌کند (حجت، گونلا و کالی[3]، ۲۰۰۳).

در نتیجه، تشخیص مغالطات در داوری همتا برای حفظ اعتبار و پایایی ادبیات علمی ضروری است. این امر به حفظ استانداردهای دقت روش شناختی کمک می‌کند که برای پیشبرد دانش در زمینه‌های مختلف بسیار مهم است (برنارد[4]، ۲۰۲۳). داوران با بررسی انتقادی تحقیقات برای خطاهای منطقی و ادعاهای بدون پشتوانه، به کیفیت کلی و روایی کار علمی منتشر شده کمک می‌کنند.

بااین‌همه، علی‌رغم نقش حیاتی استدلال دقیق در تضمین اعتبار علمی، مغالطات و خطاهای منطقی اغلب در فرایند داوری همتا نادیده گرفته می‌شوند. این غفلت می‌تواند منجر به انتشار مطالعاتی شود که حاوی انواع خطاهای استدلالی هستند و اگر این مطالعات به طور گسترده مورد استناد و پذیرش قرار گیرند، به شکل بالقوه می‌توانند یک حوزه پژوهشی را منحرف کنند (اسمیت و جانسون[5]، ۱۹۹۹؛ شوک و پاوولا[6]، ۲۰۲۱). علاوه‌براین، با افزایش حجم ادبیات علمی، فشار بر داوران، که ممکن است فاقد آموزش تخصصی در تحلیل منطقی یا تشخیص سوگیری شناختی باشند، افزایش می‌یابد. افزون‌براین، ماهیت ذهنی داوری سنتی ناسازگاری‌هایی را در ارزیابی صحت منطقی متن ایجاد می‌کند. بدون بررسی سیستماتیک صحت و دقت منطقی، ادبیات علمی در برابر سوگیری‌ها و خطاهای استدلالی آسیب‌پذیر باقی می‌ماند که می‌تواند یافته‌ها را مخدوش کند و در نتیجه بر سیاست‌گذاری، جهت‌گیری و درک عمومی از علم تأثیر بگذارد (حجت، گونلا و کالی[7]، ۲۰۰۳).

---

1. Sourati
2. Aly
3. Hojat, , Gonnella & Caelleigh
4. Bernard
5. Smith & Johnson
6. Shook & Paavola
7. Hojat, Gonnella & Caelleigh

استریکلند[1] و دیگران (۲۰۲۳) به این مسئله می‌پردازند که چگونه مغالطه‌های منطقی مانند استدلال دوری، وضع تالی و شیءانگاری، اعتبار مدل‌های علمی را در استفاده در درمان بیماری‌ها تضعیف می‌کند و مانع از درمان‌های بالینی می‌شود. این مغالطه‌ها اغلب در گفتمان علمی مورد توجه قرار نمی‌گیرند، زیرا با مفروضات ریشه‌دار یا روش‌شناسی تثبیت‌شده، همسو می‌شوند. داوران ممکن است به دلیل ماهیت ظریف مغالطات، تکیه بر هنجارهای پذیرفته شده یا عدم تسلط کافی بر استدلال انتقادی، این مغالطات را تشخیص ندهند. استریکلند و دیگران (۲۰۲۳) نشان می‌دهند که این خطاهای منطقی می‌توانند در عین حفظ و اشاعه مدل‌های معیوب علمی، توهم درک درست را نیز در متخصان ایجاد کنند.

پربال[2] (۲۰۱۲) چندین دلیل را شناسایی می‌کند که چرا داوران ممکن است موفق به شناسایی مغالطات منطقی در متون علمی نشوند. او تأکید می‌کند که داوران اغلب وقت کافی برای تجزیه و تحلیل کامل پیش‌نویس مقالات ندارند، که منجر به نادیده گرفتن نقص‌های مهمی در متن می‌شود. تخصص محدود داوران در زمینه‌های خاص، توانایی آنها را برای شناسایی خطاهای فنی یا ناسازگاری در تفسیر داده‌ها مختل می‌کند. مسائل سیستمی، مانند فشار برای انتشار سریع و استانداردهای متفاوت مجلات، این مشکلات را تشدید می‌کند و اجازه می‌دهد مطالعات ناقص اجازه انتشار یابند. پربال (۲۰۱۲) همچنین بر تحلیل انتقادی ناکافی یعنی هنگامی که داوران از کنترل یا اعتبارسنجی‌های ضروری غفلت می‌ورزند، تاکید می‌کند. مجموع این عوامل انتشار نتایج غیرقابل اعتماد کمک می‌کنند، که می‌تواند تحقیقات آینده را گمراه کند و اعتبار علمی را به خطر بیندازد (پربال، ۲۰۱۲).

بی‌طرفی داوران همتا می‌تواند توسط عوامل متعددی به خطر افتد. سوگیری تأیید باعث می‌شود که داوران از همسویی مطالعات با نظریه‌های پذیرفته شده حمایت کنند و درعین‌حال مطالعاتی را که نظریه‌های پذیرفته شده را به چالش می‌کشند بی‌اعتبار تلقی کنند (گریمالدو و پالوچی، ۲۰۱۳). مسئله سوگیری انتشار با هدف افزایش احتمال انتشار مقاله باعث مقاومت نویسندگان در برابر انتشار نتایج منفی و انحراف ادبیات علمی به سوی تأیید و انتشار یافته‌های مثبت و موید ادعای نویسندگان است. سوگیری‌های جنسیتی و نژادی نیز بر قضاوت‌ها تأثیر می‌گذارد، که شواهدی متعددی از آن در زمینه‌های مختلف وجود دارد (هلمر و دیگران، ۲۰۱۷). جهت‌گیری ایدئولوژیک و باورهای نظری و عوامل اجتماعی–سیاسی، می‌تواند بر نتایج داوری داوران تأثیر بگذارد، از انتشار برخی مطالعات جلوگیری کند یا منجر به ارزیابی‌های جانب‌دارانه شود. این مشکلات در مجموع بر وجود چالش‌هایی درباره حفظ انصاف، بی‌طرفی و عینیت در فرایند داوری همتا تاکید می‌کنند (حجت، گونلا و کالی، ۲۰۰۳).

ماهونی[3] (۱۹۷۷) نشان می‌دهد که چگونه سوگیری تأیید به نفع نتایجی که با باورهای نظری فرد مطابقت دارند، بر روند داوری همتا تأثیر می‌گذارد. ماهونی کاستی‌های قابل‌توجهی را در آشنایی با قواعد استدلال منطقی در میان دانشمندان، به ویژه عدم توفیق آنها در تشخیص قاعده رفع تالی به عنوان یک شکل معتبر استدلال، نمایان می‌کند. وی تأکید می‌کند که نتایج خلاف اندیشه‌های رایج در مقالات، دارای مفاهیم علمی و منطقی ارزشمندتر و قوی‌تری نسبت به نتایج موید دانسته‌های قبلی است، زیرا این نتایج مستقیماً فرضیه‌ها را به چالش می‌کشند اما شیوه‌های علمی و سیاست‌های انتشار به طور نامتناسبی به نفع یافته‌های تأییدکننده دانش رایج هستند. این «سنت تأیید جزمی»، دقت منطقی مورد نیاز برای پیشرفت علمی را تضعیف می‌کند و منعکس کننده یک سوگیری نظام‌مند در فرهنگ داوری و پژوهش است (ماهونی، ۱۹۷۷).

تشخیص مغالطات در مقالات علمی به دلایل متعددی می‌تواند برای داوران همتا چالش‌برانگیز باشد. داوران همتا اغلب برای شناسایی انواع خطاها و اشتباهات در متون علمی تلاش می‌کنند. بزرگ‌ترین تهدید برای کیفیت انتشار، پذیرش یک پیش‌نویس با کیفیت پایین است (دآندرئا و اودایر[4]، ۲۰۱۷). داوران بسیار تمایل دارند بر محدودیت‌های روش‌شناختی تمرکز کنند در حالی که سایر موضوعات کلیدی مرتبط با روش علمی را که باید وزن بیشتری به آنها داده شود، از دست می‌دهند (سیلز و تاناکا[5]، ۲۰۰۰). این نشان می‌دهد که داوران ممکن است فاقد یک چارچوب جامع برای ارزیابی تمام جنبه‌های یک متن باشند. فرایند داوری همتا

[1]. Strickland
[2]. Perbal
[3]. Mahoney
[4]. D’Andrea & O’Dwyer
[5]. Seals & Tanaka

با چالش‌های متعددی مانند یافتن داوران متعهد، اجتناب از تأخیر در داوری، و ایجاد انگیزه برای داوری با کیفیت بالا مواجه است که ممکن است مشکل تشخیص مغالطه را به شدت افزایش دهد (زهاری و اوسویان، ۲۰۱۵).

۳. مدل‌های بزرگ زبانی

مدل‌های بزرگ زبانی سیستم‌های یادگیری ماشین پیشرفته‌ای هستند که برای درک و تولید متن در زبان طبیعی طراحی شده‌اند. این مدل‌ها از طریق تجزیه و تحلیل مجموعه داده‌های بزرگی عمل می‌کنند که آنها را قادر می‌سازد الگوها را تشخیص دهند و پاسخ‌های متنی منسجمی را ایجاد کنند. مدل‌های بزرگ زبانی نوعی نرم‌افزار هوش مصنوعی هستند که از یادگیری ماشین، به ویژه از شبکه‌های عصبی معروف به ترنسفورمر[1]، برای انجام وظایف مختلف مربوط به پردازش زبان استفاده می‌کنند. این مدل‌ها بر روی حجم وسیعی از داده‌های متنی آموزش دیده‌اند که به آنها امکان می‌دهد تا تفاوت‌های ظریف زبان، دستور زبان، و بافت متن را درک کنند. مدل‌های بزرگ زبانی قادر به انجام طیف گسترده‌ای از کارها مانند ترجمه زبان، خلاصه‌سازی، پاسخگویی به پرسش و حتی تولید محتوای خلاقانه هستند. فن‌آوری زیربنایی آنها از اصول یادگیری عمیق استفاده می‌کند، که این مدل‌ها را قادر می‌سازد تا داده‌ها را تجزیه و تحلیل کنند و بدون دخالت انسان مفاهیم را یاد بگیرند و در نتیجه تعاملات شهودی‌تری با زبان ایجاد کنند.

چنانکه در این مطالعه نشان داده شده است این مدل‌ها به ویژه در نسخه‌های اخیر خود به توانایی‌های خیره‌کننده‌ای در شناسایی فرایند منطقی استدلال در متن، خطاهای منطقی و مغالطات صوری و غیرصوری دست یافته‌اند که آنها را به ابزار مهمی برای استفاده در شناسایی این موارد در متون دانشگاهی تبدیل می‌کند.

۴. هوش مصنوعی تبیین‌پذیر

یکی از مشکلات سنتی در استفاده از مدل‌های یادگیری ماشین و شبکه‌های عصبی، رفتار مبهم آنها در قالب جعبه سیاه بود. کاربران نمی‌توانستند علت پاسخ‌های ارائه شده توسط این مدل‌ها را درک کنند. در حالی که مدل‌های جعبه سیاه مانند شبکه‌های عصبی عمیق می‌توانند نتایج دقیقی تولید کنند؛ اما قادر به توضیح چگونگی رسیدن به آن نتایج نیستند. هوش مصنوعی تبیین‌پذیر یک فرایند برای درک و توضیح تصمیمات مدل‌های هوش مصنوعی است که به انسان‌ها کمک می‌کند رفتار مدل‌ها را بهتر درک و تفسیر کنند. این امر در تضاد با سیستم‌های جعبه سیاه است که در آنها فرایندهای داخلی از دید کاربران پنهان می‌ماند. توانایی مدل‌های بزرگ زبانی در شناسایی و توضیح مغالطات منطقی، آنها را به یک سیستم جعبه سفید تبدیل می‌کند؛ زیرا می‌توانند به طور شفاف فرایند تحلیل و معیارهای تصمیم‌گیری خود را هنگام شناسایی مغالطات منطقی برای مخاطب توضیح دهند (کمبریا و دیگران، ۲۰۲۴).

بیان مسئله

این مطالعه به بررسی این موضوع می‌پردازد که پیشرفت‌های اخیر در هوش مصنوعی، به ویژه مدل‌های بزرگ زبانی[2] با خودکارسازی تشخیص خطاهای منطقی و مغالطات در متون و پیش‌نویس‌های علمی، می‌توانند به طور قابل‌توجهی در راستای دیدگاه تارسکی، کیفیت فرایند داوری همتا را ارتقا دهند. همچنین این مطالعه قابلیت‌های هوش مصنوعی، به ویژه مدل‌های بزرگ زبانی را برای رفع این شکاف‌ها با شناسایی خودکار مغالطه‌های منطقی در متون و پیش‌نویس‌های علمی بررسی می‌کند.

در این مطالعه علاوه بر ارزیابی توانایی مدل‌های بزرگ زبانی در شناسایی مغالطات، توانایی آنها در حل مسائل منطقی و تفکر انتقادی و استدلال انتقادی در آزمون جی‌مت[3] مورد ارزیابی قرار گرفته است به این دلیل که آزمون‌هایی مانند جی‌مت برای تحقیقات آکادمیک به چند دلیل مهم هستند. مهارت‌های تفکر انتقادی برای موفقیت در مطالعات و تحقیقات در سطح تحصیلات تکمیلی ضروری تلقی می‌شوند. هدف این آزمون‌ها ارزیابی توانایی دانشجویان تحصیلات تکمیلی و فارغ التحصیلان دانشگاه‌ها در تجزیه

[1]. Transformer
[2]. Large Language Models
[3]. GMAT

و تحلیل اطلاعات پیچیده، ارزیابی استدلال‌ها و نتیجه‌گیری منطقی است یعنی مهارت‌هایی که در تحقیقات آکادمیک حیاتی هستند (فیلیپس و باند[1]، ۲۰۰۴؛ دانکزاک و دیگران[2]، ۲۰۲۰). اهمیت تفکر انتقادی در آموزش عالی باعث گنجانده شدن آن به عنوان یک ویژگی مهم فارغ التحصیلان در بسیاری از دانشگاه‌ها شده است (دانکزاک و دیگران، ۲۰۲۰). این نشان می‌دهد که این آزمون‌ها می‌توانند به شناسایی دانشجویانی کمک کنند که احتمالاً در امور پژوهشی برتری دارند. این آزمون‌های استاندارد سعی می‌کنند توانایی‌های عمومی تفکر انتقادی را که می‌تواند در زمینه‌های مختلف دانشگاهی از جمله پژوهش و انتشار مقاله به‌کار رود، اندازه‌گیری کنند.

هدف از پژوهش حاضر، نخست تأکید بر اهمیت شناسایی مغالطات در فرایند داوری همتا و دوم معرفی توانایی‌های مدل‌های بزرگ زبانی در شناسایی مغالطات و خطاهای منطقی و امکان استفاده از آنها در فرایند داوری همتا برای کشف این مغالطات است. برای نیل به این مقصود نویسندگان علاوه بر آموزش مدل‌های بزرگ زبانی برای شناسایی این مغالطات، دو مجموعه داده را نیز برای ارزیابی دقت عملکرد این مدل‌ها گرداوری و معرفی نموده‌اند.

پرسش پژوهش ۱: مدل‌های بزرگ زبانی تا چه اندازه می‌توانند مغالطات منطقی را در متون علمی تشخیص دهند و دقت آنها به چه عواملی بستگی دارد؟

پرسش پژوهش ۲: الگوریتم‌های هوش مصنوعی به ویژه مدل‌های بزرگ زبانی تا چه حد می‌توانند برای تشخیص مغالطات منطقی در فرایند داوری همتا به‌کار گرفته شوند و به چه صورت بر دقت و قابلیت اعتماد مقالات علمی تأثیر می‌گذارند؟

پیشینه پژوهش

۱. پیشینه نظری

استفاده از مدل‌های بزرگ زبانی در فرایندهای داوری همتا، به ویژه در تشخیص مغالطه‌های منطقی در مقالات علمی، نشان‌دهنده پیشرفت قابل‌توجهی در حوزه یکپارچگی دانشگاهی و ارتباطات است. این بخش یک زمینه نظری در مورد اینکه چگونه مدل‌های بزرگ زبانی می‌توانند در داوری همتا گنجانده شوند، با تمرکز ویژه بر کاربردهای بالقوه آنها در شناسایی مغالطات ارائه می‌دهد.

مدل‌های بزرگ زبانی، توانایی‌های قابل‌توجهی را در تولید منسجم متن نشان داده‌اند. آنها می‌توانند در کارهای مختلف دانشگاهی، از جمله فرایند داوری همتا، پژوهشگران را یاری کنند (یه[3]، ۲۰۲۴). ادغام مدل‌های بزرگ زبانی در داوری همتا به طور بالقوه می‌تواند هم بهره‌وری و هم کیفیت داوری را افزایش دهد. این ادغام می‌تواند با کمک به داوران در ایجاد بازخورد سازنده و ارائه سریع خلاصه مقالات، فرایند داوری را تسهیل کند (حسینی و هورباخ[4]، ۲۰۲۳). کشف مغالطه‌های منطقی در نوشته‌های علمی یک کار ظریف است که نیازمند درک عمیق هم از محتوا و هم ساختار استدلال‌های ارائه شده در مقالات تحقیقاتی است. مدل‌های بزرگ زبانی پتانسیل مقابله با این چالش را با استفاده از تکنیک‌هایی دارند که الگوهای استدلال را به طور سیستماتیک تجزیه و تحلیل می‌کنند (گراد[5]، ۲۰۲۳؛ چوو و دیگران[6]، ۲۰۲۴). تحقیقات نشان داده است که مدل‌های بزرگ زبانی را می‌توان برای تشخیص مغالطات و خطاهای منطقی رایج با استفاده از مجموعه داده‌ها آموزش داد. شناسایی مغالطه‌هایی مانند حمله شخصی و پهلوان پنبه را می‌توان از طریق استفاده از رویکردهای ساختاریافته که روابط سلسله مراتبی بین ادعاها و شواهد پشتیبان آنها را نشان می‌دهند، تسهیل کرد. اثربخشی مدل‌های بزرگ زبانی در تشخیص مغالطات را می‌توان از طریق تکنیک‌های مختلف یادگیری ماشین افزایش داد. مطالعات اخیر بر اهمیت ترکیب یک ساختار منطقی سیستماتیک برای بهبود توانایی مدل‌های بزرگ زبانی در طبقه‌بندی مغالطات تاکید می‌کند. به عنوان مثال، استفاده از یک درخت ساختار منطقی می‌تواند به ردیابی جریان منطقی متن

---

[1]. Phillips & Bond
[2]. Danczak et al.
[3]. Ye
[4]. Hosseini & Hurbach
[5]. Grad
[6]. Chu

کمک کند، در نتیجه شناسایی مواردی که در آن استدلال ناقص و معیوب است، آسان تر می‌شود. استفاده انتقادی از روش‌های آموزشی، از جمله توسعه مجموعه داده‌های نظارت شده که به طور خاص برای تشخیص مغالطه طراحی شده‌اند نیز این قابلیت را افزایش می‌دهد و باعث بهبود قابل‌توجهی در دقت مدل‌های زبانی می‌شود (لیم و پراولت[۱]، ۲۰۲۴).

۲. پیشینه تجربی

اشرفی موغاری[۲] و دیگران (۲۰۲۴) به ارزیابی عملکرد مدل‌های زبانی بزرگ در آزمون جی‌مت[۳] پرداخته‌اند. این آزمون که به عنوان یک معیار کلیدی در پذیرش برنامه‌های تحصیلات تکمیلی کسب‌وکار استفاده می‌شود، توانایی داوطلبان را در استدلال‌های کلامی، ریاضی و نوشتاری تحلیلی می‌سنجد و نشان‌دهنده آمادگی آنها برای محیط‌های آکادمیک چالش‌برانگیز است. در مطالعه اشرفی موغاری و دیگران (۲۰۲۴)، هفت مدل از جمله مدل‌های جی‌پی‌تی-۴ توربو[۴] ، جی‌پی‌تی-۴[۵] ، کلاود[۶] ۲/۱ و جمینای ۱ پرو[۷] مورد بررسی قرار گرفته‌اند. نتایج این مطالعه نشان می‌دهد که جی‌پی‌تی-۴ توربو با دقت ۸۵ درصد در پاسخ به سؤالات آزمون، بالاترین عملکرد را داشته و حتی از میانگین نمرات دانشجویان برتر در مدارس کسب‌وکار پیشی گرفته است. به طور متوسط، مدل‌های زبانی بزرگ در مجموع نمره‌ای معادل ۶۴۳/۵ کسب کرده‌اند که بالاتر از حدود ۶۳ درصد از داوطلبان انسانی است جی‌پی‌تی-۴ توربو که در سال ۲۰۲۳ معرفی شده است، با عملکرد خود در میان ۱ درصد برتر داوطلبان انسانی قرار گرفته است که نشان‌دهنده قابلیت‌های چشمگیر این فناوری است (اشرفی موغاری و دیگران، ۲۰۲۴).

بر اساس این مطالعه در بخش استدلال انتقادی[۸]، مدل‌ها به طور کلی توانستند فرضیات را شناسایی کرده و استدلال‌ها را ارزیابی کنند، برخی مدل‌ها در تحلیل‌های عمیق‌تر و بررسی شرایط فرضی با چالش‌هایی مواجه شدند اما مدل جی‌پی‌تی-۴ توربو نتیجه چشمگیر ۹۶/۳ درصد را در بخش استدلال انتقادی کسب نمود. این مطالعه همچنین نشان داده است که مدل‌های بزرگ زبانی می‌توانند به عنوان ابزارهای کمکی در یادگیری، ارزیابی و تدریس آزمون جی‌مت استفاده شوند. با این حال، استفاده از این فناوری مستلزم چارچوب‌های نظارتی، ملاحظات اخلاقی، و بررسی صحت اطلاعات است (اشرفی موغاری و دیگران، ۲۰۲۴). در نهایت، این پژوهش اهمیت جی‌مت را به عنوان معیاری برای ارزیابی مهارت‌های حیاتی مرتبط با تحصیلات تکمیلی کسب‌وکار برجسته کرده و استفاده مسئولانه از هوش مصنوعی برای بهبود تجربه‌های آموزشی را توصیه می‌کند.

لیم و پراولت (۲۰۲۴) به ارزیابی توانایی یک مدل بزرگ زبانی در شناسایی مغالطات منطقی می‌پردازند. نویسندگان با استفاده از مجموعه داده لاجیک[۹]، عملکرد جی‌پی‌تی-۴ را در تشخیص هفت نوع مغالطه منطقی رایج ارزیابی کردند. نتایج نشان داد که این مدل بزرگ زبانی در کل داده‌ها دقت ۰,۷۹ و در زیرمجموعه‌ای که موارد نامشخص حذف شده بودند، دقت ۹۰ درصد را به دست آورد. مدل در تشخیص مغالطه «حمله شخصی[۱۰]» بهترین عملکرد را با دقت ۹۴ درصد داشت، اما در شناسایی «توسل به احساسات[۱۱]» ضعیف‌تر عمل کرد. نویسندگان بر اهمیت ارزیابی دقیق مدل‌های بزرگ زبانی قبل از استفاده در پروژه‌های تحقیقاتی تأکید کرده و ملاحظاتی را در مورد طراحی فرایند ارزیابی، انتخاب مجموعه داده مناسب و مهندسی پرامپت ارائه دادند. مطالعه ایشان نشان می‌دهد که مدل‌های بزرگ زبانی می‌توانند در شناسایی مغالطات منطقی مؤثر باشند، اما همچنان محدودیت‌هایی دارند که باید در نظر گرفته شود (لیم و پراولت، ۲۰۲۴).

---

[1]. Lim & Perrault
[2]. Ashrafimoghari
[3]. GMAT
[4]. GPT-4 Turbo
[5]. GPT-4
[6]. Claud
[7]. Gemini 1.0 Pro
[8]. Critical Reasoning
[9]. LOGIC
[10]. Ad hominem
[11]. Appeal to Emotion

پن[1] و دیگران (۲۰۲۴) به بررسی توانایی مدل‌های زبانی بزرگ در تشخیص و طبقه‌بندی مغالطات منطقی به صورت یادگیری بدون نمونه[2] می‌پردازد. نویسندگان روش‌های مختلف مهندسی پرامپت شامل روش‌های تک مرحله‌ای و چند مرحله‌ای را برای استخراج مغالطات به وسیله مدل‌های بزرگ زبانی پیشنهاد می‌دهند. نتایج آزمایش‌ها روی مجموعه داده‌های معیار نشان می‌دهد که مدل‌های بزرگ زبانی با پرامپت تک مرحله‌ای می‌توانند عملکردی مشابه یا بهتر از مدل‌های تی۵[3] در برخی مجموعه داده‌های عمومی داشته باشند. به طور خاص، جی‌پی‌تی-۴ توانست در مجموعه داده آرگوتاریو[4] به امتیاز F1 برابر با ۷۸/۹۴ درصد دست یابد که از بهترین نتیجه روش‌های سنتی با ۷۲/۳۸ درصد بهتر است. همچنین روش‌های پرامپت چند مرحله‌ای پیشنهادی توانستند عملکرد را به ویژه برای مدل‌های زبانی کوچک‌تر بهبود دهند. مهمترین دستاورد مقاله نشان دادن پتانسیل مدل‌های بزرگ زبانی به عنوان طبقه‌بندی کننده‌های در یادگیری بدون نمونه از مغالطات است که نیاز به آموزش با داده‌های برچسب‌خورده را برطرف می‌کند و قابلیت تعمیم‌پذیری بهتری نسبت به روش‌های یادگیری نظارت شده دارد (پن و دیگران، ۲۰۲۴).

لی[5] و دیگران (۲۰۲۴) به بررسی و بهبود توانایی مدل‌های زبانی بزرگ در استدلال منطقی از طریق درک مغالطات منطقی می‌پردازد. نویسندگان پنج وظیفه مشخص را در سه بعد شناسایی و طبقه‌بندی، استنتاج و اصلاح خطا تعریف کرده و یک مجموعه داده جدید به نام ال.اف.یو.دی[6] با کمک جی‌پی‌تی-۴ ایجاد کردند که شامل ۴،۰۲۰ نمونه از ۱۲ نوع مغالطه منطقی است. نتایج آزمایش‌ها نشان می‌دهد که جی‌پی‌تی-۴ بهترین عملکرد را در تمام وظایف با دقت بالای ۷۸ درصد داشته است. همچنین بهینه‌سازی[7] مدل‌های زبانی با ال.اف.یو.دی باعث بهبود قابل توجه عملکرد آنها در استدلال منطقی شده است، به طوری که مدل لاما ۱۳ میلیارد پارامتری[8] پس از بهینه‌سازی در مجموعه داده تاکسی ان.ال.آی[9] ۷/۵۳ درصد و در مجموعه داده فولیو[10] ۷/۱۶ درصد بهبود نشان داده است. این پژوهش نشان می‌دهد که درک خطاهای منطقی می‌تواند به طور مؤثری توانایی استدلال منطقی مدل‌های زبانی را افزایش دهد.

روش پژوهش

۱. انتخاب روش پژوهش

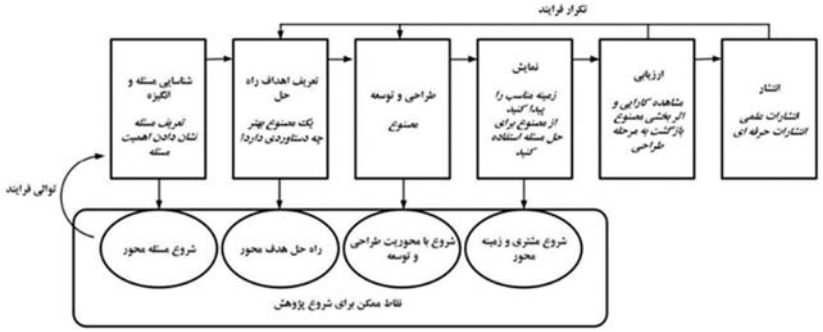
روش پژوهش به‌کاررفته در این مطالعه، روش پژوهش علم طراحی[11] است که در طراحی سیستم‌های اطلاعاتی و سیستم‌های هوش مصنوعی به شکل قابل‌توجهی استفاده می‌شود (آپیولا و سوتینن[12]، ۲۰۲۰). در این روش، هدف پدید آوردن و ارزیابی یک مصنوع یا فرآورده فنی است که یک مشکل اجتماعی، انسانی یا سازمانی را رفع می‌کند.

پژوهشگران به فراخور نیاز و نوع مسئله، مدل‌های متنوعی را برای پژوهش علم طراحی پیشنهاد نموده‌اند. در میان این مدل‌ها چند مدل بیشتر مورد اقبال واقع شده‌اند که عبارتند از: مدل نونامیکر و دیگران[13] (۱۹۹۱)، مدل هونر[14] و دیگران (۲۰۰۴)، مدل پفرز و دیگران (۲۰۰۷) و مدل وایشناوی و کوچلر (۲۰۱۲). پژوهشگران در ارائه این مدل‌ها رویکردهای متفاوتی در اتخاذ پارادایم‌های پوزیتیویستی، پراگماتیستی و تفسیری داشته و در آن مدل‌ها برای طراح سیستم در مراحل مختلف پژوهش، نقش‌های مختلفی اعم از طراح، ناظر بی‌طرف یا مفسر تعریف نموده‌اند (وایشناوی و کوچلر[15]، ۲۰۰۴). بر اساس مطالعه ونابل[16] و دیگران (۲۰۱۶) به دلیل

---

1. Pan
2. zero-shot Learning
3. T5
4. Argotario
5. Li
6. LFUD
7. Fine-Tuning
8. LLaMA2-13B
9. TaxiNLI
10. FOLIO
11. Design Scince Research
12. Apiola, M., & Sutinen
13. Nunamaker
14. Hevner
15. Vaishnavi & Kuechler
16. Venable

اینکه مدل پفرز و دیگران بر اساس پارادایم پوزیتیویستی گام‌های شفاف و کاملی ارائه می‌دهد و در هنگام اجرا برخلاف سایر مدل‌ها موجب سردرگمی محقق نمی‌شود، مبنای بسیاری از پژوهش‌های علم طراحی قرار گرفته است، و به‌همین دلیل در پژوهش حاضر نیز به عنوان مدل پژوهش علم طراحی انتخاب شده است. این مدل شامل شش مرحله شناسایی مسئله، تعریف اهداف، طراحی و توسعه، ارزیابی، انتشار، و ارتباط است و برای اجرا چارچوبی جامع و مرحله‌به‌مرحله ارائه می‌نماید. شکل ۱ مدل پفرز و دیگران را نشان می‌دهد.



شکل ۱. مدل پفرز و دیگران (۲۰۰۷) برگرفته از (حاج کاظمی، ۱۴۰۳)

۲. فرایند اجرای پژوهش

۲–۱. روش آماده‌سازی مدل‌ها

در این پژوهش، علاوه بر مدل‌های زبانی، دیبرتا–لارج[1] ۳۰۴ میلیارد پارامتری، لاما–۳٫۲ ۱ میلیارد پارامتری[2]، لاما–۳٫۱ ۸ میلیارد پارامتری[3]، میسترال–۰٫۳ ۷ میلیارد پارامتری[4]، چت‌جی‌پی‌تی 4o و چت‌جی‌پی‌تی o1، از دو مدل یادگیری ماشین کلاسیک شامل ماشین بردار پشتیبان و جنگل تصادفی به عنوان مبنای ارزیابی استفاده کردیم.

قبل از انجام آزمایش‌های اصلی، داده‌های متنی توسط ابزارهای متناسب، به بردارهای عددی تبدیل شدند. مدل‌های زبانی تحت فرایند آماده‌سازی قرار گرفتند که شامل بهینه‌سازی[5] و یادگیری بدون نمونه[6] بود. همچنین برای مدل‌های ماشین بردار پشتیبان [7] و جنگل تصادفی[8] عملکردی از روش‌های استاندارد پردازش متن و استخراج ویژگی استفاده شد.

۲–۲. آزمایش اول: تشخیص مغالطات در مجموعه داده الک‌دب۶۰–۲۰

الک‌دب۶۰–۲۰ [9] یک مجموعه داده شامل مغالطه‌های منطقی دسته‌بندی شده است که در مناظرات انتخاباتی نامزدهای ریاست جمهوری آمریکا از سال ۱۹۶۰ تا ۲۰۲۰ میلادی به‌کار رفته‌اند. این مجموعه داده شامل ۱۷۹۰ عبارت حاوی مغالطه در شش دسته حمله شخصی[10]، توسل به احساسات[11]، توسل به مرجع[12]، شیب لغزنده[13]، علت جعلی[14] و توسل به شعار[15] است (گوفردو[16] و دیگران، ۲۰۲۳).

[1]. Diberta-large
[2]. Llama-3.2-1B
[3]. Llama-3.1-8B
[4]. Mistral-7B-v0.3
[5]. Fine-Tuning
[6]. Zero-Shot Learning
[7]. Support Vector Machine
[8]. Random Forest
[9]. ElecDeb60to20
[10]. Ad Hominem
[11]. Appeal to Emotion
[12]. Appeal to Authority
[13]. Slippery Slope
[14]. False Cause
[15]. *Argument by Slogan*
[16]. Goffredo

پس از آماده‌سازی و بازآموزی مدل بر اساس بخش آموزش مجموعه داده الک‌دب۶۰-۲۰ ، مدل را با ۲۰۰ متن حاوی مغالطه، از بخش آزمون همین مجموعه داده مورد ارزیابی قرار دادیم و با نتایج به‌دست آمده از مقاله گوفردو و دیگران (۲۰۲۳) مقایسه نمودیم. نتایج به دست آمده در جدول ۱ نشان داده شده است. مدل مورد استفاده این مطالعه در این فرایند، دیبرتا-لارج نسخه ۳ با ۳۰۴ میلیون پارامتر است که در ۱۵ دور اجرای مدل روی یک پردازنده گرافیکی ۳۰۹۰ با حافظه گرافیکی ۲۴ گیگابایت بهینه شده است.

۲-۳. آزمایش دوم: حل سؤالات بخش منطق آزمون استعداد تحصیلی دکتری

در این آزمایش، مدل‌ها را با ۳۵ سؤال بخش منطق از آزمون‌های استعداد تحصیلی دکتری سال‌های ۱۳۹۷ تا ۱۴۰۲ ایران مورد ارزیابی قرار دادیم. در این آزمایش، مدل‌ها بدون هرگونه آموزش قبلی و به روش یادگیری بدون نمونه مورد استفاده قرارگرفتند. نتایج در جدول ۲ ارائه شده است.

۲-۴. آزمایش سوم: تشخیص مغالطات در متون با اندازه‌های متفاوت

نویسندگان در این مطالعه، یک مجموعه داده متشکل از ۱۵۰ متن با طول متغیر از ۱ تا ۱۰ پاراگراف که هر یک حاوی یکی از پانزده نوع مغالطه منطقی رایج در متون علمی بودند، ایجاد نمودند. این مجموعه داده برای ارزیابی عملکرد مدل‌های مختلف زبانی در تشخیص و طبقه‌بندی مغالطات منطقی استفاده شد. سپس مجموعه داده به دو بخش آموزش و آزمون با نسبت ۹۰ درصد به ۱۰ درصد تقسیم شد و مدل‌های چت جی‌پی‌تی 4o، چت جی‌پی‌تی o1، لاما-۳.۱-۸ میلیارد پارامتری، میسترال-۰,۳-۷میلیارد پارامتری و جنگل تصادفی برای طبقه‌بندی روی آن اعمال شدند. نتایج در جدول ۳ ارائه شده است.

یافته‌های پژوهش

آزمایش نخست در بخش ۲-۱ روی مجموعه داده الک‌دب۶۰-۲۰ به دقت ۹۱ درصد و امتیاز اف-۱ ۸۶/۳۵ درصد رسید و در این امتیاز همان‌طور که در جدول ۱ مشاهده می‌شود بیش از ۱۲ درصد نسبت مدل‌های قبلی بهبود یافت. این نتایج نشان می‌دهد که مدل‌های پیشرفته‌تر و دارای پارامترهای بیشتر شامل دیبرتا-لارج، لاما ۳ میلیارد پارامتری و به ویژه چت‌جی‌پی‌تی o1، توانایی قابل‌توجهی در تحلیل و تشخیص ظرایف استدلال‌های مغالطه‌آمیز دارند، درحالی‌که مدل‌های پایه که توسط گوفردو و دیگران (۲۰۲۳) استفاده شده بودند، عملکرد ضعیف‌تری از خود نشان دادند.

جدول ۱. نتایج عملکرد مدل‌های یادگیری ماشین و مدل‌های زبانی بر مجموعه داده الک‌دب۶۰-۲۰
(مدل نشان‌گذاری شده با * مدل استفاده شده در این پژوهش است.)

مدل	امتیاز اف-۱ در تشخیص مغالطات
ElectraFTC	٪۴۰,۳۳
BERT + LSTM	٪۴۶,۹۷
BERT + BiLSTM + LSTM (comp. and rel. features)	٪۵۶,۱۴
DistilbertFTC distilbert-base-cased	٪۷۰,۱۰
DebertaFTC microsoft/deberta-base	٪۷۲,۲۲
BertFTC dbmdz/bert-large-cased-finetuned-conll03-english	٪۷۲,۳۷
MultiFusion BERT (comp., rel. and PoS features)	٪۷۳,۹۴
Deberta-v3-large Fine-Tuned *	٪۸۶,۳۵

جدول ۲. نتایج عملکرد مدل‌های یادگیری ماشین و مدل‌های زبانی بر مجموعه داده سوالات استعداد تحصیلی آزمون دکتری

مدل	دقت حل سؤالات منطق
SVM	٪۴۸
Random Forest	٪۴۹
Llama-3.1 8B	٪۴۹/۵
Mistral-0.3 7B	٪۵۶/۱
ChatGPT 4o	٪۶۸
ChatGPT o1	٪۱۰۰

نتایج آزمایش سوم در بخش ۲-۳ که در جدول ۳ ارائه شده است نشان داد که چت جی.پی.تی 01 با استفاده از رویکرد یادگیری بدون نمونه به دقت ۱۰۰ درصد در شناسایی مغالطات دست یافت. مدل‌های لاما[1] و میسترال[2] اگرچه در شناسایی دقیق نوع و نام مغالطه عملکرد بسیار پایین‌تری داشتند (به ترتیب ۴۳ درصد و ۴۶ درصد)، اما در تشخیص وجود مغالطه و موقعیت آن در متن عملکرد قابل‌توجهی (۹۷/۱ درصد و ۹۵/۴ درصد) از خود نشان دادند. مدل جنگل تصادفی که به عنوان مدل پایه استفاده شد، تنها به دقت ۲۲ درصد دست یافت.

جدول۳. نتایج عملکرد مدل‌های یادگیری ماشین و مدل‌های زبانی بر مجموعه داده متون حاوی مغالطات با اندازه‌های مختلف

مدل	دقت شناسایی نوع مغالطه	دقت تشخیص وجود و محل مغالطه
ChatGPT 4o	٪۸۹	٪۱۰۰
ChatGPT o1	٪۱۰۰	٪۱۰۰
Llama-3.1 8B	٪۴۳	٪۹۷
Mistral-0.3 7B	٪۴۶	٪۹۵
Random Forest	٪۲۲	-

بحث

دقت آزمایش نخست در بخش ۲-۱ در امتیاز اف-۱ بیش از ۱۲ درصد نسبت به مدل‌های قبلی بهبود یافت. این نتیجه نشان می‌دهد که مدل‌های پیشرفته‌تر و دارای پارامترهای بیشتر توانایی قابل‌توجهی در تحلیل و تشخیص ظرایف استدلال‌های مغالطه‌آمیز دارند، در حالی که مدل‌های پایه که توسط گوفردو و دیگران (۲۰۲۳) استفاده شده بودند، عملکرد ضعیف‌تری از خود نشان دادند. نکته مهم این آزمایش این است که حتی می‌توان با استفاده از ابزارهای ارزان‌قیمت و مقرون‌به‌صرفه برای عموم نیز به نتایج قابل قبولی در شناسایی مغالطات منطقی دست یافت.

نتایج به‌دست آمده از آزمایش دوم در بخش ۲-۲ که در جدول ۲ نمایش داده شده، نشان‌دهنده پیشرفت چشمگیر در توانایی مدل‌های زبانی، به ویژه چت جی.پی.تی o1، در حل مسائل پیچیده منطقی و استدلالی است. این توانایی با توجه به نتایج کسب شده توسط پذیرفته‌شدگان کنکور دکتری دانشگاه‌های ایران در این درس، چشمگیر است. مدل‌های کلاسیک ماشین بردار پشتیبان و جنگل تصادفی عملکردی مشابه با آزمایش اول داشتند، که نشان می‌دهد مدل‌های کلاسیک یادگیری ماشین برای حل مسائل پیچیده منطقی مناسب نیستند.

مجموعه داده جدیدی که در این مطالعه گردآوری شده است و در آزمایش سوم در بخش ۲-۳ شرح داده شده است، از چندین جنبه نسبت به مجموعه داده‌های رقیب برتری دارد. نخست، مجموعه داده این مطالعه شامل ۱۵۰ متن با ۱۵ نوع مغالطه منطقی است و برای ارزیابی مدل‌های زبانی در تشخیص مغالطات در متون علمی طراحی شده است. در مقایسه با سایر مجموعه‌های داده موجود، مانند کوکولوفا[3] شامل ۷،۷۰۶ خبر با تنوع مغالطات (یه[4] و دیگران، ۲۰۲۴)، لاجیک[5] شامل ۲،۴۴۹ نمونه از ۱۳ نوع مغالطه

[1]. LLama 3.2 8B
[2]. Mistral 0.3 7B
[3]. COCOLOFA
[4]. Yeh
[5]. LOGIC

در متون غیررسمی (جین[1] و دیگران، ۲۰۲۲)، و مافالدا[2] شامل یک بنچمارک برای تحقیقات آکادمیک (هلوه[3] و دیگران، ۲۰۲۳)، مجموعه داده این مطالعه از نظر اندازه و حجم متنوع‌تر و متمرکز بر متون علمی است, درحالی که مجموعه داده این مطالعه برای تشخیص خودکار مغالطات در متون علمی مناسب است، مجموعه‌های داده رقیب مانند کوکولوفا برای توسعه مدل‌های تشخیص مغالطات عمومی کاربرد دارند. از نظر تأثیر بر آموزش مدل‌های زبانی، مجموعه داده جدید عملکرد بسیار بالاتری را نشان می‌دهد. مدل چت جی‌پی‌تی o1 با استفاده از رویکرد یادگیری بدون نمونه به دقت ۱۰۰ درصد در شناسایی و نام‌گذاری مغالطات دست یافته است، در حالی که دقت مدل‌های پیشرفته در مجموعه داده‌های رقیب معمولاً کمتر از ۹۰ درصد است . همچنین، مدل‌های لاما و میسترال اگرچه در شناسایی دقیق نوع و نام مغالطه عملکرد پایین‌تری نسبت به مدل o1 داشتند (به ترتیب ۴۳ درصد و ۴۶ درصد)، اما در تشخیص وجود مغالطه و موقعیت آن در متن عملکرد قابل‌توجهی (۹۷/۱ درصد و ۹۵/۴ درصد) از خود نشان دادند که این میزان عملکرد در مجموعه داده‌های رقیب معمولاً کمتر از ۸۰ درصد است.

مقایسه نتایج هر سه آزمایش نشان می‌دهد که مدل‌های پیشرفته‌تر، به ویژه چت جی.پی.تی ۰۴ و چت جی.پی.تی o1، عملکرد بسیار بهتری در هر دو زمینه تشخیص مغالطات و حل مسائل منطقی دارند. به طور خاص، چت جی.پی.تی o1 با دستیابی به دقت ۹۸/۱ درصد در تشخیص مغالطات و ۱۰۰ درصد در حل سؤالات منطق، نشان داد که توانایی فوق‌العاده‌ای در درک و تحلیل استدلال‌های پیچیده دارد. عملکرد ضعیف مدل‌های ماشین بردار پشتیبان و جنگل تصادفی نشان می‌دهد که روش‌های سنتی یادگیری ماشین برای این نوع وظایف پیچیده زبانی کافی نیستند. این تفاوت چشمگیر در عملکرد بین مدل‌های کلاسیک و مدل‌های زبانی پیشرفته، اهمیت استفاده از معماری‌های پیچیده‌تر و روش‌های یادگیری عمیق را در پردازش زبان طبیعی و استدلال منطقی نشان می‌دهد.

نتیجه‌گیری

نتایج این پژوهش به‌وضوح نشان داد که مدل‌های بزرگ زبانی، به ویژه چت‌جی‌پی‌تی‌o1 ، توانایی قابل‌توجهی در شناسایی مغالطات منطقی در متون علمی دارند. در پاسخ به **پرسش اول پژوهش**، مشخص شد که این مدل‌ها، با دستیابی به دقت ۹۸/۱ درصد در تشخیص انواع مغالطات منطقی و عملکرد ۱۰۰ درصد در حل مسائل منطقی، نسبت به روش‌های سنتی یادگیری ماشین مانند ماشین بردار پشتیبان و جنگل تصادفی عملکردی به‌مراتب برتر دارند. این موفقیت به عواملی نظیر معماری پیشرفته مدل‌ها، بهره‌گیری از مجموعه داده‌های گسترده و متنوع، و بهینه‌سازی‌های دقیق مربوط است. همچنین، انطباق مدل‌ها با نوع داده‌ها و نیازهای تحقیقاتی نقش مؤثری در افزایش دقت این مدل‌ها داشته است.

در پاسخ به **پرسش دوم پژوهش**، یافته‌ها نشان دادند که مدل‌های بزرگ زبانی می‌توانند تأثیرات قابل‌توجهی بر بهبود دقت، شفافیت و قابلیت اعتماد فرایند داوری همتا داشته باشند. این مدل‌ها توانستند با شناسایی سریع و دقیق مغالطات منطقی، کاهش خطاهای انسانی، و ارائه بازخوردهای ساختاریافته، کیفیت و سرعت فرایند داوری را ارتقا دهند. علاوه بر این، استفاده از این فناوری‌ها در کاهش فشار کاری بر داوران انسانی، بهبود انسجام در ارزیابی‌ها و استانداردسازی فرایندهای داوری مؤثر بوده است. با ظهور این توانایی و در دسترس بودن آن برای جامعه علمی، از این پس عواملی مانند عدم وجود مهارت کافی در داوران و دشواری شناسایی مغالطات برای عامل انسانی، نمی‌توانند مانعی برای بررسی مغالطات و الزام بر رعایت آنها در هنگام داوری همتا باشند و برای افزایش اعتبار علمی و پیشگیری از انحراف منطقی در انواع پژوهش‌ها لازم است هوش مصنوعی با رعایت معیارها و ضوابطی جدی در حوزه سنجش اعتبار منطقی پژوهش‌ها و مقالات علمی در همه حوزه‌ها و در همه مجلات معتبر علمی جایگزین عامل انسانی شده و به‌کارگرفته شود.

---

[1]. Jin
[2]. MAFALDA
[3]. Helwe

منابع

نبوی، لطف‌اله. (۱۳۸۴). *مبانی منطق و روش شناسی*. تهران: انتشارات دانشگاه تربیت مدرس.

نبوی، لطف‌اله. (۱۳۸۶). *تراز اندیشه*. تهران: انتشارات بصیرت.

خندان، علی‌اصغر. (۱۳۸۴). *منطق کاربردی*. تهران: کتاب طه.

حاج کاظمی، محبوبه (۱۴۰۳). توصیف و طراحی مصنوعات و مکانیزم‌های فرهنگی رسانه‌ای به منظور تغییر فرهنگ سازمانی همسو با ارزش‌های جدید. پایان‌نامه دکتری. دانشگاه تهران.

References

Aly, M., Colunga, E., Crockett, M. J., Goldrick, M., Gomez, P., Kung, F. Y. H., McKee, P. C., Pérez, M., Stilwell, S. M., & Diekman, A. B. (2023). Changing the culture of peer review for a more inclusive and equitable psychological science. *Journal of Experimental Psychology: General*, 152(12), 3546-3565. https://doi.org/10.1037/xge0001461

Apiola, M., & Sutinen, E. (2020). Design science research for learning software engineering and computational thinking: Four cases. Computer Applications in Engineering Education, 29, 101 - 83. https://doi.org/10.1002/cae.22291

Ashrafimoghari, V., Gürkan, N., & Suchow, J. W. (2024). *Evaluating large language models on the GMAT: Implications for the future of business education*. arXiv preprint arXiv:2401.02985.

Ayer, A. J. (1953). Cogito, Ergo Sum. *Analysis*, 14(2), 27-31. https://doi.org/10.2307/3326309

Bernard, C. (2020). On Fallacies in Neuroscience. *eNeuro*, 7. https://doi.org/10.1523/ENEURO.0491-20.2020

Cambria, E., Malandri, L., Mercorio, F., Nobani, N., & Seveso, A. (2024). XAI meets LLMs: A Survey of the Relation between Explainable AI and Large Language Models. *ArXiv*, abs/2407.15248. https://doi.org/10.48550/arXiv.2407.15248

Chu, Z., Ai, Q., Tu, Y., Li, H. & Liu, Y. (2024). PRE: A peer review based large language model evaluator. *arXiv:2401.15641v2*. https://doi.org/10.48550/arXiv.2401.15641

D'Andrea, R., & O'Dwyer, J. P. (2017). Can editors save peer review from peer reviewers?. *PloS One*, 12(10), e0186111. https://doi.org/10.1371/journal.pone.0186111

Floridi, L. (2009). Logical fallacies as informational shortcuts. *Synthese*, 167, 317-325. https://doi.org/10.1007/s11229-008-9410-y

Garcia, J. A., Rodriguez-Sánchez, R., & Fdez-Valdivia, J. (2020). Confirmatory bias in peer review. *Scientometrics*, *123*, 517–533. https://doi.org/10.1007/s11192-020-03357-0

Goffredo, P., Chaves, M., Villata, S., & Cabrio, E. (2023, December). Argument-based detection and classification of fallacies in political debates. In *EMNLP 2023-Conference on Empirical Methods in Natural Language Processing* (Vol. 2023, pp. 11101-11112). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.684

Goodman, S. (1999). Toward Evidence-Based Medical Statistics. 1: The P Value Fallacy. *Annals of Internal Medicine*, 130, 995-1004. https://doi.org/10.7326/0003-4819-130-12-199906150-00008

Grimaldo, F., & Paolucci, M. (2013). A simulation of disagreement for control of rational cheating in peer review. *Advances in Complex Systems*, *16*, 1350004+. https://doi.org/10.1142/s0219525913500045

Haj Kazemi, M. (2024). *Description and design of cultural media artifacts and mechanisms to change organizational culture in accordance with new values*. Doctoral dissertation, University of Tehran. (in Persian)

Han, S. J., Ransom, K. J., Perfors, A., & Kemp, C. (2024). Inductive reasoning in humans and large language models. *Cognitive Systems Research*, *83*, 101155. https://doi.org/10.1016/j.cogsys.2023.101155

Helmer, M., Schottdorf, M., Neef, A., & Battaglia, D. (2017). Gender bias in scholarly peer review. *ELife*, 6, e21718. https://doi.org/10.7554/eLife.21718

Helwe, C., Calamai, T., Paris, P. H., Clavel, C., & Suchanek, F. (2023). MAFALDA: A benchmark and comprehensive study of fallacy detection and classification. arXiv preprint arXiv:2311.09761.

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). design science in information systems research. *MIS Quarterly*, *28*(1), 75-105. https://doi.org/10.2307/25148625

Hojat, M., Gonnella, J. S., & Caelleigh, A. S. (2003). Impartial judgment by the "gatekeepers" of science: Fallibility and accountability in the peer review process. Advances in Health Sciences Education, *8*(1), 75–96. https://doi.org/10.1023/A:1022670432373

Hosseini, M., & Horbach, S. P. J. M. (2023). Fighting reviewer fatigue or amplifying bias? Considerations and recommendations for use of ChatGPT and other large language models in scholarly peer review. *Research Integrity and Peer Review*, *8* (4). https://doi.org/10.1186/s41073-023-00133-5

Jin, Z., Lalwani, A., Vaidhya, T., Shen, X., Ding, Y., Lyu, Z., Sachan, M., Mihalcea, R., & Scholkopf, B. (2022). Logical fallacy detection. ArXiv, abs/2202.13758. https://doi.org/10.18653/v1/2022.findings-emnlp.532

Jin, Z., Lalwani, A., Vaidhya, T., Shen, X., Ding, Y., Lyu, Z., ... & Schoelkopf, B. (2022). Logical fallacy detection. arXiv preprint arXiv:2202.13758

Kant, I. (1781/1787). Critique of pure reason (N. K. Smith, Trans.). Macmillan. (Original work published 1781/1787). See "*Transcendental Dialectic, Book II*, Chapter I: The Paralogisms of Pure Reason."

Khandan, A. A. (2005). *Applied Logic*. Tehran: Ketab Taha. (in Persian).

Lawson, H. (2006). Breaking the language barrier. *Symbolic Interaction*, *29*, 423-427. https://doi.org/10.1525/SI.2006.29.3.423

Li, Y., Wang, D., Liang, J., Jiang, G., He, Q., Xiao, Y., & Yang, D. (2024). Reason from fallacy: Enhancing large language models' logical reasoning through logical fallacy understanding. arXiv preprint arXiv:2404.04293

Lim, G., & Perrault, S. T. (2023). Evaluation of an LLM in identifying logical fallacies. *CSCW Companion '24: Companion Publication of the 2024 Conference on Computer-Supported Cooperative Work and Social Computing*, pp. 303–308. https://doi.org/10.1145/3678884.3681867

Mahoney, M. J. (1977). Publication prejudices: An experimental study of confirmatory bias in the peer review system. *Cognitive Therapy and Research*, *1*(2), 161–175. https://doi.org/10.1007/BF01173636

Miles, M. (1999). Insight and inference: Descartes's founding principle and modern philosophy. https://doi.org/10.2307/3182574

Mo, W. (2007). Cogito: From Descartes to Sartre. *Frontiers of Philosophy in China*, *2*, 247-264. https://doi.org/10.1007/s11466-007-0016-0

Nabavi, L. (2005). *Fundamentals of logic and methodology*. Tehran: Tarbiat Modares University Press. (in Persian)

Nabavi, L. (2007). *The Balance of Thought*. Tehran: Basirat Publications. (in Persian)

Nietzsche, F. (1886). *Beyond good and evil* (W. Kaufmann, Trans.). Vintage. (Original work published 1886). See Aphorisms 16 and 17.

Oswald, A. (2008). Can we test for bias in scientific peer-review?. *IZA Discussion*, Paper No. 3665, Available at SSRN: https://ssrn.com/abstract=1261450 or http://dx.doi.org/10.2139/ssrn.1261450

Pan, F., Wu, X., Li, Z., & Luu, A. T. (2024). Are LLMs good zero-shot fallacy classifiers?. arXiv preprint arXiv:2410.15050

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2008). A design science research methodology for information systems research. *Journal of Management Information Systems*, *24*(3), 45-77. https://doi.org/10.2753/MIS0742-1222240303

Perbal, B. (2012). Flaws in the peer-reviewing process: A critical look at a recent paper studying the role of CCN3 in renal cell carcinoma. *Journal of Cell Communication and Signaling*, *6*(3), 199–210. https://doi.org/10.1007/s12079-012-0171-1

Peter Grad. (2023). Large language models prove helpful in peer-review process. Phys.org. https://phys.org/news/2023-10-large-language-peer-review.html

Rui Ye. (2024). Are we there yet? Revealing the risks of utilizing large language models in scholarly peer review. arXiv.org. https://arxiv.org/abs/2412.01708v1

Russell, B. (2001). The problems of philosophy. OUP Oxford.

Seals, D. R., & Tanaka, H. (2000). Manuscript peer review: A helpful checklist for students and novice referees. *Advances in Physiology Education*, *23*(1), 52-58. https://doi.org/10.1152/advances.2000.23.1.S52

Shook, J. R., & Paavola, S. (Eds.). (2021). Abduction in cognition and action: Logical reasoning, scientific inquiry, and social practice (Vol. 59). *Springer Nature*. https://doi.org/10.1007/978-3-030-61773-8

Sizo, A., Lino, A., Reis, L., & Rocha, Á. (2019). An overview of assessing the quality of peer review reports of scientific articles. *International Journal of Information Management*, *46*, 286-293. https://doi.org/10.1016/j.ijinfomgt.2018.07.002

Smith, J., & Johnson, R. (1999). Logic of scientific reasoning. Holy Cross College. Retrieved from https://college.holycross.edu/projects/approaches5/PDFs/chap2.pdf

Smith, J., & Jones, A. (2021). Understanding peer review: Challenges and biases. *Journal of Academic Publishing*, 15(3), 45-60. https://doi.org/10.1234/jap.2021.015

Sourati, Z., Ilievski, F., Sandlin, H. Â., & Mermoud, A. (2023). Case-based reasoning with language models for classification of logical fallacies. arXiv preprint, arXiv:2301.11879

Stelmakh, I., Rastogi, C., Liu, R., Chawla, S., Echenique, F., & Shah, N. (2022). Cite-seeing and reviewing: A study on citation bias in peer review. *PLOS One*, *18*. https://doi.org/10.1371/journal.pone.0283980

Strickland, J. C., Stoops, W. W., Banks, M. L., & Gipson, C. D. (2023). Logical fallacies and misinterpretations that hinder progress in translational addiction neuroscience. *Journal of the Experimental Analysis of Behavior*, *117*(3). https://doi.org/10.1002/jeab.757

Takata, N., & Mimura, M. (2022). [The logic of scientific reasoning in peer review process]. *Brain and nerve = Shinkei kenkyu no shinpo*, *74*(4), 335-340. https://doi.org/10.11477/mf.1416202040

Tarski, A. (1994). *Introduction to logic and to the methodology of the deductive sciences* (Vol. 24). Oxford University Press. https://doi.org/10.1093/oso/9780195044720.001.0001

Valatsos, V. (2020). A propositional logic review of Descartes' Phrase "Cogito, Ergo Sum". *viXra*. https://vixra.org/pdf/2007.0024v2.pdf

Venable, J. R., Pries-Heje, J., & Baskerville, R.L. (2017). Choosing a design science research methodology (2017). *ACIS 2017 Proceedings*. 112. https://aisel.aisnet.org/acis2017/112

Vaishnavi, V., & Kuechler, B. (2004). Design Science Research in Information Systems. Association for Information Systems. https://www.researchgate.net/publication/235720414

Yeh, M. H., Wan, R., & Huang, T. H. K. (2024). CoCoLoFa: A dataset of news comments with common logical fallacies written by LLM-Assisted Crowds. arXiv preprint arXiv:2410.03457