Utilizing LLMs to Evaluate the Argument Quality of Triples in SemMedDB for Enhanced Understanding of Disease Mechanisms.

Shuang Wang, MS[1], Yang Zhang, MD, Ph. D[2], Jian Du*, Ph. D[1]
[1]National Institute of Health Data Science, Peking University, Beijing, China; [2]Peking University First Hospital, Beijing, China

Abstract

The Semantic MEDLINE Database (SemMedDB) has limited performance in identifying entities and relations, while also neglects variations in argument quality, especially persuasive strength across different sentences. The present study aims to utilize large language models (LLMs) to evaluate the contextual argument quality of triples in SemMedDB to improve the understanding of disease mechanisms. Using argument mining methods, we first design a quality evaluation framework across four major dimensions, triples' accuracy, triple-sentence correlation, research object, and evidence cogency, to evaluate the argument quality of the triple-based claim according to their contextual sentences. Then we choose a sample of 66 triple-sentence pairs for repeated annotations and framework optimization. As a result, the predicted performances of GPT-3.5 and GPT-4 are excellent with an accuracy up to 0.90 in the complex cogency evaluation task. The tentative case evaluating whether there exists an association between gestational diabetes and periodontitis reveals accurate predictions (GPT-4, accuracy, 0.88). LLMs-enabled argument quality evaluation is promising for evidence integration in understanding disease mechanisms, especially how evidence in two stances with varying levels of cogency evolves over time.

Introduction

The scientific literature is rich with information across diverse fields concerning potential disease mechanisms. However, identifying and prioritizing mechanisms with more promising and reliable for further analytical evaluation is a great challenge for medical researchers, with the increasing growth of biomedical literature and related datasets (1). Constructing knowledge graphs by extracting concepts and their relations from scientific texts offers an effective solution to this challenge. For example, the Semantic MEDLINE Database (SemMedDB) is a representative repository of subject-predicate-object triples extracted from the entire set of PubMed titles and abstracts (2). SemMedDB has been utilized by manually selecting causal predicates such as CAUSES, PREVENTS, and DISRUPTS to identify intermediates (3-5) and common causes, i.e., confounders (6, 7) between the investigated exposure-and-outcome variables. While widely used for constructing biomedical knowledge graphs, it demonstrated limited performance with precision of 0.69, recall of 0.42, and an F1 score of 0.52 in a relaxed evaluation (8). This would highly influence the quality of its downstream task. For instance, since the concepts extracted are overly general, it can result in many apparent contradictions that are not truly contradictory (9). Additionally, existing natural language processing (NLP) techniques for biomedical entity and relations extraction frequently neglect the quality of arguments, especially the varying degrees of strength across different sentences where claims are expressed. While seven factuality values (fact, probable, possible, doubtful, counterfact, uncommitted, and conditional) were annotated for the extracted triples, representing the real-world nature of biomedical knowledge as hypotheses, speculations, or opinions rather than explicit facts (10), this only captures one facet of argument quality.

In this study, we try to introduce the concept of argument quality to enhance the accuracy and persuasive strength of extracted triples in SemMedDB, for a better understanding of disease mechanisms and evidence-based clinical decision-making. Argumentation plays a fundamental role in complex areas like healthcare, where it directly affects decision-making related to human lives. Argument mining, a method in computational linguistics, has been a significant technique in evidence-based decision-making to use the best evidence to improve the quality of medical research as well as clinical practice (11, 12). Recently, an impressive development in the field of argument mining is Project Debater (13, 14), in which the training corpus curated by IBM is organized under a simple claim-evidence structure for automatic detection of claims and evidence in the context of Wikipedia controversial topics (15). Project Debater focuses on social issues, and the basic structure of a claim is "Topic-Action" or "Topic-View". For example, a policy motion like "Preschool should be subsidized", where the topic is "Preschool", and the action is "be subsidized"; a legitimate analysis motion like "AI brings more harm than good",

where the topic is "AI", and the view is "brings more harm than good". In this study, we mainly focus on scientific claims in biomedical fields. Compared with societal claims, scientific claims mainly represented as a Cause-Effect structure, such as *Cancer_COMPLICATES_Covid-19*, i.e., "Cancer patients are susceptible to COVID-19", and *Hydroxychloroquine_NEGTREATS_Covid-19*, i.e., "hydroxychloroquine should not be used for patients with COVID-19 of any severity". Here we consider an extracted triple as a claim, its multiple sources of supporting sentences as the evidence set. This is a simplified model of argument mining similar with the Project Debater.

In essence, argument mining is the process of identifying the components and structure of an argument and evaluating the quality of the argumentation in texts (16). A common argument structure is consisted of two components, premise and conclusion. The argument quality depends on two aspects, the truthfulness of argument components and the validity of the argument structure. However, not all arguments hold equal strength or persuasion (17). Various dimensions, such as cogency, reasonableness, and effectiveness, have been proposed to assess argumentative quality. Moreover, evaluating argument quality presents a formidable challenge. For instance, determining logical cogency requires analyzing the acceptability, relevance, and collective sufficiency of premises, demanding close reading and logical reasoning. On the other hand, evaluating rhetorical persuasiveness involves subjective judgments of emotional appeal, style, and credibility, which can vary among different audiences.

Large Language Models (LLMs) have demonstrated substantial potential across diverse applications in the medical field. We agree with such viewpoint that, it is more appropriate to employ LLMs as scientific reasoning engines, rather than knowledge bases since it often fall short when it comes to factual accuracy of the generated scientific knowledge (18). We assume LLMs would perform well in such a complex task of argument quality analysis, given the multitude of logical reasoning processes involved, including discerning premises and conclusions, assessing the varying strength of evidence across different sentences, and ensuring strong correlation in extracted triplets within sentences, thus minimizing information loss. Thus, the present study aims to utilize LLMs to evaluate the argument quality of triples in SemMedDB to improve the understanding of disease mechanisms.

Methods

In general, our argument mining framework contains two main parts, the identification of the argument component and its persuasive strength. We choose gestational diabetes (GDM) as a case study. First, we come up with an initial evaluation framework and then repeatedly annotate 66 triple-source sentence pairs of the SemMedDB database for at least five rounds to constantly adjust and optimize our evaluation framework. We utilize LLMs to conduct the automatic evaluation for feasibility and performance evaluation. Second, we conduct a potential application summary of downstream tasks with the evaluation method. Third, we use an observed association between GDM and periodontitis as a case to interpret the argument quality evaluation results. The project workflow is shown below (Figure 1).

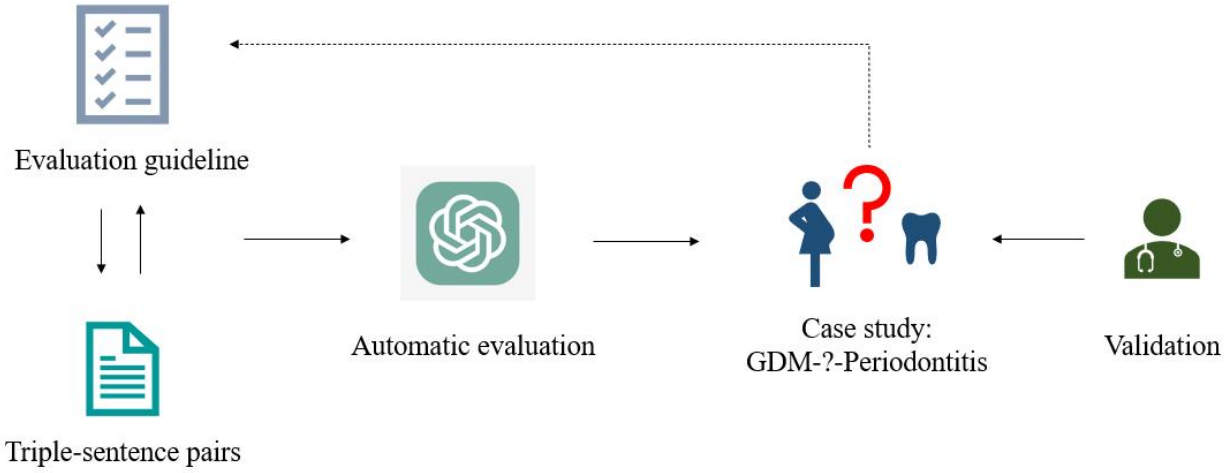

Figure 1. Project workflow

1. The evaluation framework of argument quality

1) Identify argument component

To identify the argument component of a claim, we first examine whether the claim is accurately extracted and then measure the claim (triple) - context (source sentence) correlation to classify it as a premise or a conclusion.

Accuracy We measure the triple accuracy from the accuracy of the head and tail concepts as well as the predicate direction. As for triple concepts, we measure whether the triple concept pair (subject and object) is correctly extracted from the source sentence. Regarding the predicate, we focus on the accuracy of the directional aspect of the predicate rather than the precision of the selection of the predicate, e.g., between CAUSES and PREDISPOSES, despite the subtle differences they have in the same direction. For example, when a sentence indicates *A does not cause B*, but the extracted triple is *A-causes-B*, it results in a misdirected predicate; A neutral direction may be extracted from the conditional statement rather than an assertive expression.

Claim-context correlation The degree of correlation measures whether the triple is connected to the main topic or intention of the original sentence as a conclusion (major correlation) or a triple involved in the sentence as an implied premise (non-major correlation). Also, if a triple is the sufficient and necessary premise of the argumentative sentence, we consider it as a major correlation.

2) Measure argument cogency

In the present study, we use evidence cogency to measure the argument's persuasive strength. Cogency is defined as the logical strength of evidence in the source sentence to conclude the claim. We classify cogency into three categories, 1) lack of supporting information, 2) evidence based on background knowledge or prior experience, and 3) evidence based on authors' experimental findings.

3) Identify conditional information

A notable concern with the triples stored in SemMedDB, extracted by SemRep, is their flat structure, which represents the smallest unit of factual information without representing detailed conditional information and hierarchical structures, especially the complex phenotypes and influential populations of diseases. As we choose GDM as a case study, identifying the influenced object, especially the maternal or fetus, is an essential problem. Extracting and representing the conditional information of triples is a highly complex task, given that conditions can arise from various sources. Presently, there is ongoing studies focused on the extraction of such conditional information in biomedical literature (19) and clinical trials (20). To examine the potential of LLMs in supplementing conditional information, here, we specify the conditional information as the influenced object of the given triple and utilize LLMs to extract such information as a supplement for further refined data integration.

Table 1. An example of the argument quality evaluation.

Triple	Gestational Diabetes-PREDISPOSES-Cardiovascular Diseases
Sentence (PMID: 32739399)	Early pregnancy metabolites predict gestational diabetes mellitus: implications for fetal programming. BACKGROUND: Aberrant fetal programming in gestational diabetes (GDM) appears to increase the risk for obesity, type 2 diabetes, and cardiovascular disease.
Accuracy	Triple concepts correct: Yes Predicate direction: Yes
Claim-context correlation	Yes
Cogency	Evidence cogency: Background knowledge Cue words: BACKGROUND
Conditional information	Influenced object: Fetus
Explanation	The triple relation is involved in the sentence. However, the cue word "BACKGROUND" and the present tense of "appears to" indicate the association is based on prior researchers' results or regular background knowledge rather than the present study's experimental results. Also, 'IMPLICATIONS FOR FETAL PROGRAMMING' indicates the influenced object is fetal than mothers.

2. Large language models (LLMs)

We choose GPT-3.5-turbo and GPT-4 as our examination models (21). Firstly, we use a general standard prompt containing questions and explanations for all measures and the required json output format. In particular, in prompt engineering, we compare the performance results of each model both with zero-shot and few-shot. As the common few-shot cases are below ten, we choose 9-shot for a comparison. However, evidence cogency is essentially a complex task for models. According to the interaction with ChatGPT-4, we find that it is hard to distinguish background experience knowledge from research experimental findings completely. Models often mix methodology statements with new findings, which suggests the need for an optimized detailed prompt. Thus, we then optimize the prompt with detailed explanations and differential definitions of easy-confused situations and re-evaluate this task separately for the performance examination. Finally, we use accuracy and macro F1-core as two performance metrics to examine models' performances.

3. Case study

To test the feasibility of downstream tasks of the argument quality evaluation, we choose the evidence integration of the association between GDM and periodontitis as a case study. We searched all SemMedDB triples and corresponding source sentences of GDM (N = 42,835) and periodontitis (N = 98,244) and finally locate all triples (N = 9) with GDM as the subject and periodontitis as the object for a tentative case. We evaluate the argument quality of all 9 triple-sentence pairs using GPT-4 with the framework and conduct an association stance analysis with evaluation results.

Results

1. Automatic evaluation with LLMs

To identify the optimal evaluation prompt format and model, we compare the performance of different prompts in LLMs. In general, the overall performance of the two models in all evaluation tasks is great and GPT-4 outperforms GPT-3.5-turbo as expected (Table 2). Specifically, as the difficulty of the evaluation task increases, the performance of models decreases gradually; models perform better with few-shot than zero-shot except on a few occasions.

First, we observe special cases where zero-shot exceeds few-shot in the accuracy evaluation task of triple concepts in two models and in the claim-context correlation task in GPT-3.5. As concept correctness is the simplest evaluation task with an F1 score up to 0.96 in GPT-4, the pre-trained base models might be sufficient and examples may introduce extra noise to slightly reduce the results. For claim-context correlation, the abnormal situation of GPT-3.5 may be due to its slightly inferior capacity in understanding complex texts of explanations of examples than GPT-4. Second, models exhibit great performances in the text generation task of identifying research objects, after few-shot in two models. With zero-shot, two models present two exceptional results, both yielding 0 correct extraction. This outcome arises because the models incorrectly identify the literal object of a triple as the designated research object, though we explain in the prompt that the research object should be population or other species like rat. Few-shot fine-tune models to comprehensively understand the task.

Third, as for the hardest task of evidence cogency evaluation, we surprisingly find great growths in performance from zero-shot, few-shot, to optimized prompt with few-shot in each model. In the beginning, with the general prompt question, two models perform very badly and GPT-3.5 is even close to random results (Accuracy, GPT-3.5, 0.33; GPT-4, 0.51). Though the performance metric scores have boosted a lot, it still shows a great distance with the best accuracy of 0.67 in GPT-4 compared to other tasks. After optimizing the prompt with detailed explanations and differential definitions of easy-confused situations, the performances of each model are further improved (GPT-4 accuracy, 0.90). This highlights the need for providing a detailed prompt and utilizing a strong base model in evidence cogency evaluation.

Table 2. Performance of automatic argument quality evaluation using LLMs (triple-sentence pairs, N = 57)

Evaluation task	Prompt & K-shots	GPT-3.5-turbo		GPT-4	
		Accuracy	F1 score	Accuracy	F1 score
Concept (0, 1)	Zero-shot	0.86	0.92	0.93	0.96
	Few-shot	0.82	0.90	0.91	0.95
Relation stance* (0, 1, 2)	Zero-shot	0.86	0.88	0.88	0.89
	Few-shot	0.91	0.92	0.95	0.95
Claim-context correlation (0, 1)	Zero-shot	0.70	0.80	0.75	0.84
	Few-shot	0.65	0.77	0.81	0.88
Research object	Zero-shot	0	/	0	/
	Few-shot	0.82	/	0.95	/
Evidence cogency* (0, 1, 2)	Zero-shot	0.33	0.29	0.51	0.51
	Few-shot	0.60	0.60	0.67	0.65
	Optimized prompt & Few-shot	0.79	0.80	0.90	0.90

Note. * indicates all evaluation tasks with multi-classes; the rest are binary tasks. For all muti-class tasks, we use a weighted F1 score for evaluation. Though the research object is the text generative task, to conduct the performance comparison accurately, we manually compare true results with predictive results to measure the accuracy rather than using machine translation evaluation metrics like METEOR or BLEU. Due to the various types of research objects, we do not calculate the F1 score for this task.

2. Applications of the evaluation framework in argument mining

After annotating the sample sentences for multiple rounds, based on our quality evaluation framework, we summarize a general downstream task categorization of application scenarios in argument mining (Fig 2). Based on concept correctness and predicate direction, this assessment initially evaluates the accuracy of triple claims, excluding those incorrectly extracted claims from entering to next tasks. Next, with all correctly extracted claims, the claim-context correlation classifies a triple as an explicitly conclusive claim or an implicit premise in context. Regarding an explicit conclusive triple, based on the evidence cogency of all source sentences, evidence integration analysis can be conducted. For a given pair of exposure-effect (subject-object) of interest, by integrating all conclusive triple-context information, one can delineate a structure with all supporting and attacking arguments, along with the corresponding measurable evidence cogency to get the whole picture of the association's strength. This offers significant insights for the identification and understanding of the current status concerning disease mechanisms, especially in rare diseases with limited or conflicting evidence in place.

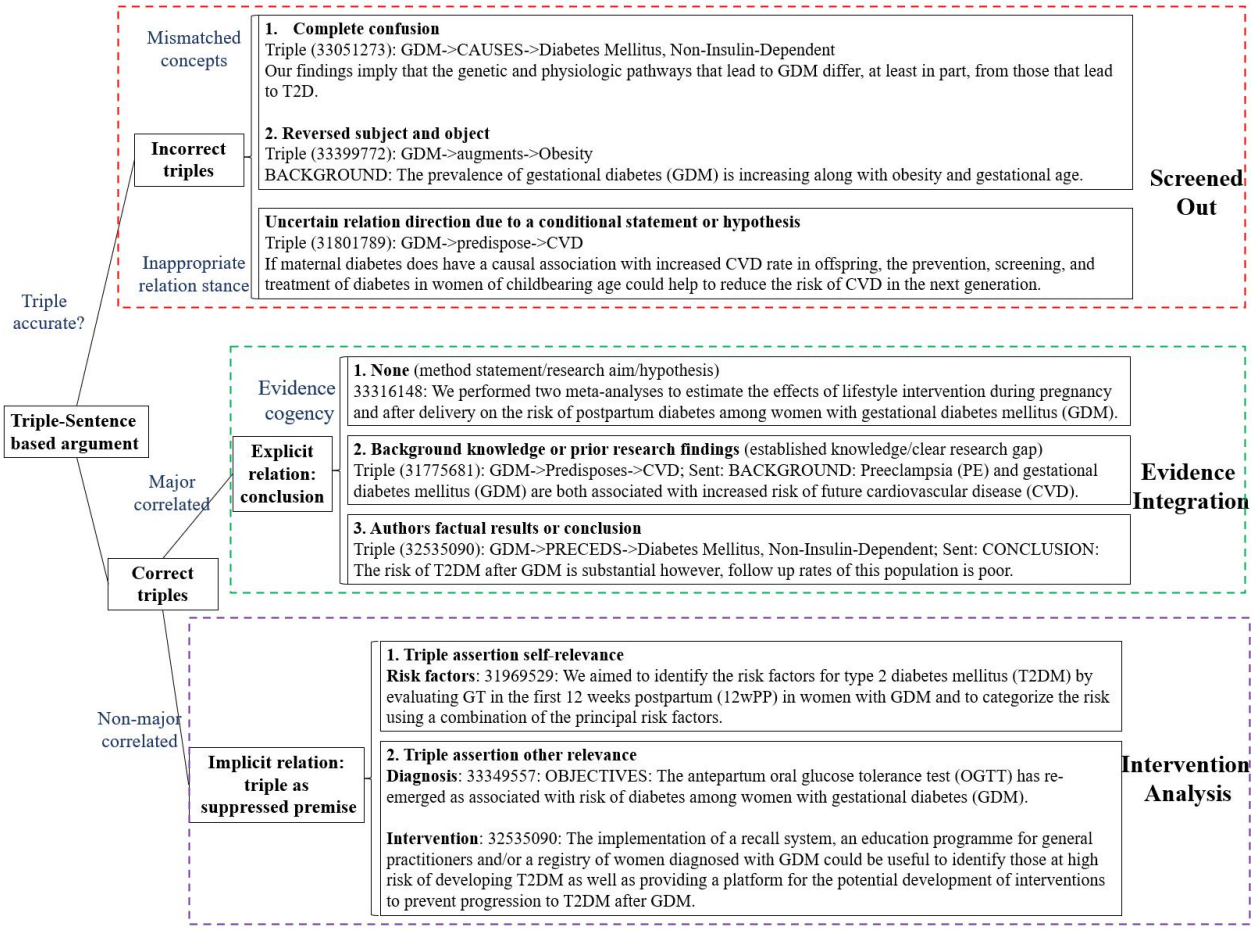

Figure 2. An argument mining pipeline based on argument quality evaluation framework.

When it comes to common diseases, implicit premises within context are crucial. According to our annotation data, we find that when the association of two diseases serves as an implied premise than the major purpose of the argument, the contextual sentence mainly falls into two categories, triple assertion self-relevance (e.g., risk factors) or triple assertion other relevance (e.g., diagnosis and intervention). In the case of common diseases, the association between two diseases might already be well established, eliminating the need for evidence integration for validation. For common diseases, the analysis of risk factors and intervention for prevention is more appropriate. Thus, it is helpful to use the argument quality evaluation workflow to filter out those arguments to supplement explicit triple assertions in diagnosis or intervention.

3. Case study

Finally, to test the evaluation results, we conduct a tentative case study to integrate evidence on the association between GDM and periodontitis. Table 3 exhibits all sentences of the association between GDM (subject) and periodontitis (object) filtered from SemMedDB, along with the corresponding evidence cogency and explanation results using GPT-4. After human curation, we find that all sentences predicted correctly except the #2 resulting an accuracy of 0.88. Though sentence #2 actually looks like a background sentence, it is actually a new finding since it is the title sentence of a meta-analysis of the association between GDM and periodontitis. This suggests that LLMs should be further trained with more information (such as the sentence location) to improve its accuracy again.

Table 3. GPT-4 predictive results of all sentences related to the GDM - periodontitis association

#	Sentence	Cogency	GPT-4 explanation
1	PMID: 23952077; BACKGROUND: Gestational diabetes mellitus (GDM) and metabolic syndrome have been related to periodontitis.	1	This sentence provides background information that establishes a connection between GDM, metabolic syndrome, and periodontitis based on previous research.
2	PMID: 27449842; Periodontitis may be Associated with Gestational Diabetes Mellitus but not Affirmatively.	1-> 2	The use of "may be Associated" suggests a tentative relationship which is indicative of background 3knowledge or a hypothesized connection rather than definitive new findings.
3	PMID: 23075430; Periodontitis was significantly associated with GDM (odds ratio = 3.00, 95% confidence interval = 1.19 to 7.56).	2	This sentence presents specific statistical results from the authors' research, indicating it is reporting on new findings.
4	PMID: 23075430; CONCLUSIONS: The results suggest that periodontitis is associated with GDM.	2	The word "CONCLUSIONS" and "The results suggest" indicate that the sentence is summarizing the findings of the authors' own research.
5	PMID: 28403353; However, it was not observed impact of periodontitis on the development of type 2 diabetes mellitus among women with previous gestational diabetes.	2	Despite the negative result, this sentence presents a specific finding from the authors' research indicating that they looked for but did not find an impact, which constitutes an author's finding.
6	PMID: 28403353; Conclusions: It was not observed impact of periodontitis on the development of type 2 diabetes among women with previous gestational diabetes.	2	The term "Conclusions" signifies that what follows is a summary of the authors' results. The phrase "It was not observed" presents a specific outcome of the study, which is that the authors did not find a link between periodontitis and the development of type 2 diabetes among women with previous GDM in their research, thus representing the authors' findings.
7	PMID: 34865247; Based on the multiple logistic regression analyses, periodontitis in early pregnancy was associated with GDM, and the three-step regression analyses showed that Porphyromonas gingivalis (P. gingivalis) and the serum TNF-alpha and IL-8 levels played roles in the association between untreated periodontitis and GDM.	2	This sentence provides a detailed account of the results from specific analyses conducted by the authors, which are part of their findings.
8	PMID: 34925709; GDM alone significantly increased the oral microbial diversity (by Shannon index, p = 0.049), and when combined with periodontitis, GDM significantly decreased the intestinal microbial richness (by observed species, p = 0.018) and influenced the structure of intestinal microbial community (by AMOVA, p = 0.043).	2	The inclusion of specific statistical measures and results suggests that this sentence is conveying the authors' own research findings.
9	PMID: 36993820; More well-designed studies differentiating between pregnant women with good oral health and those with periodontitis are needed to ascertain which	1	The recommendation for more research indicates a lack of definitive findings and a call for further investigation, which is not an author's finding but a statement of where the

	differences are due to GDM or periodontitis.		current research stands or what is needed in the future.

Note: in evidence cogency, 1 indicates background knowledge and 2 indicates new empirical findings. The "1-> 2" labeled cogency prediction is a false prediction which should be 2 (new findings).

Next, to summarize all evidence with two stances (association or no-association), we first remove duplicate evidence (#3, 5) from the identical published study and evidence without directly describe the association between two diseases (#6, #8). Then we summarize evidence in two stances, revealing the potential but uncertain correlation with a balanced distribution of evidence in two positions (Table 4). Moreover, considering the time series and the evidence cogency, we plot a scatter chart to reflect how evidence in two stances with different cogency evidence evolve with time (Fig. 3). Interestingly, it starts with a research finding suggesting the existence of the association between two diseases in 2013 and ends with a background statement implying the uncertain status. This further suggests that LLM based quality evaluation using argument mining methods is promising for evidence integration in understanding diseases mechanisms.

Table 4. Distribution of evidence in two stances regarding the association between GDM and periodontitis

GDM-*ASSOCIATED*-Periodontitis		GDM-*NEG_ASSOCIATED*-Periodontitis	
Background knowledge	New findings	Background knowledge	New findings
1	2	1	1

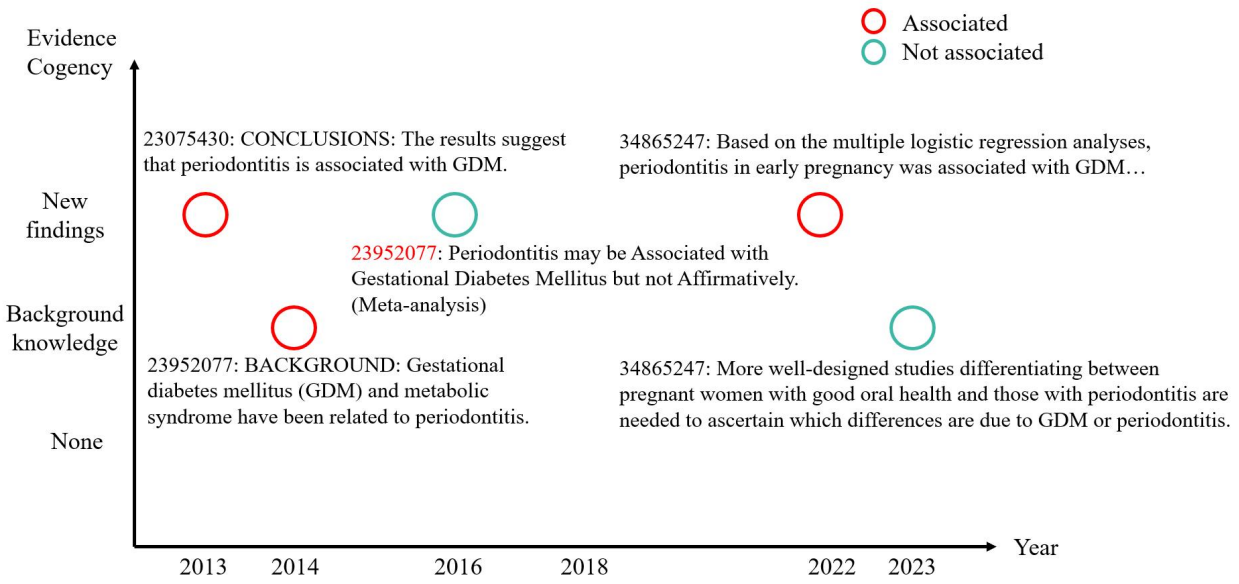

Figure 3. How evidence in two stances with varying levels of cogency evolves over time. Each circle indicates an individual study with the extracted sentence for evaluation. The red circle represents the stance of GDM is associated with periodontitis and green circle represents not associated.

Discussion and Conclusions

The present study presents an LLMs-enabled argument quality evaluation framework to evaluate the accuracy of semantic relations derived from SemMedDB and the evidence cogency of the claim-context argumentation. In general, models perform well in each evaluation task with an accuracy score up to 0.93 of GPT-4 in the simplest concept accuracy task; the complex evidence cogency evaluation necessaries a crafted prompt with clear definitions and attention points with and few-shot. Also, the appropriate adoption of the evaluation process enables two major downstream applications, the evidence integration of unknown association of diseases and the evidence collection for analyzing risk factors and intervention of known association of diseases. Finally, the GDM-periodontitis case

validates the accuracy of the evaluation results using GPT-4 and implies the feasibility as a tentative step for evidence integration.

The performance results reveal referrable method details of prompt engineering in future similar evaluations. In terms of triple's accuracy, models in two simple tasks of concept correctness and predicate stance have already shown brilliant metric scores with zero-shot (GPT-3.5, accuracy, concept 0.86; predicate stance, 0.86), making it possible to deploy other light LLMs to tackle these simple tasks. Especially, few-shot in concept correctness would instead reduce models' performances which probably introduce noises, aligned with another recent findings (22). Though the models' initial performances in the little harder task of claim-context correlation are a bit lower than the last two tasks, few-shot could help slightly improve the evaluation effects. Regarding the text generative task of the research object, though initial evaluation results are limited, the few-shot results are significant and can largely improve models' performances. When it comes to assessing the most challenging and complex evidence cogency, it is crucial to conduct this evaluation separately with a detailed prompt containing differential definitions of easy-confused cases and crafted examples. Advanced LLMs would be more suitable and beneficial in this task.

Currently, though LLMs have great capacities to directly extract entities and relations from contextual information, the limitation of max tokens and the difficulty of entity alignment in specific research fields create barriers. It is difficult to directly use LLMs to generate triples with aligned entities and proper granularity level from millions of unstructured texts and structured data, especially when it comes to constructing a complex network for the mechanical analysis of diseases. Altogether, the present work presents a basic evaluation approach to control the quality of extraction results from the general NLP tools like SemMedDB and SemRep, taking sufficient advantage of the existing automatic tools and LLMs to improve the quality of complex downstream tasks.

Limitations

Though promising results are revealed, limitations exist in the study. First, we only use a small sample size(66 sentences) for an initial analysis. However, the dataset has been annotated repeatedly for at least five rounds to evaluate and optimize the comprehensiveness and appropriateness of the quality evaluation outline and to summarize the annotation pattern and guidelines. Second, limited LLMs are considered and compared in this work. To test the complexity and feasibility of the evaluation process, we choose the art-of-state best models for a test. As they perform very well, it gives us confidence and space to train and fine-tune other open-source light models to conduct these evaluations. Third, as we use the SemMedDB dataset, the contexts are only sentences in title and abstract for an article. Though abstracts are representative covering all sections from background to conclusion, future work will consider including full-text body to further evaluate models and optimize the prompt engineering. Fourth, considering the inherent feature of GDM, we only consider one conditional information (the research object) in our evaluation as an example, for diseases with different traits, one may consider designing a different conditional information extraction schema. Fifth, considering the evidence cogency, we now consider established knowledge and research gap both as background information. However, combined with the predicate direction stance metric, these two types can be separated. As established knowledge will have a certain positive or negative predicate, while unclear knowledge tend to exhibit a neutral predicate stance.

Acknowledgments

This study was funded by the National Key R&D Program for Young Scientists (Project number 2022YFF0712000 to JD) and the National Natural Science Foundation of China (Project number 72074006 to JD).

References

1. Jin Q, Leaman R, Lu Z. PubMed and beyond: biomedical literature search in the age of artificial intelligence. EBioMedicine. 2024;100:104988.
2. Kilicoglu H, Shin D, Fiszman M, Rosemblat G, Rindflesch TC. SemMedDB: a PubMed-scale repository of biomedical semantic predications. Bioinformatics (Oxford, England). 2012;28(23):3158-60.
3. Elsworth B, Gaunt TR. MELODI Presto: a fast and agile tool to explore semantic triples derived from biomedical literature. Bioinformatics. 2021;37(4):583-5.
4. Liu Y, Elsworth B, Erola P, Haberland V, Hemani G, Lyon M, et al. EpiGraphDB: A database and data mining platform for health data science. Bioinformatics (Oxford, England). 2021;37(9):1304-11.

5. Elsworth B, Dawe K, Vincent EE, Langdon R, Lynch BM, Martin RM, et al. MELODI: Mining Enriched Literature Objects to Derive Intermediates. Int J Epidemiol. 2018;47(2):369-79.
6. Malec SA, Taneja SB, Albert SM, Elizabeth Shaaban C, Karim HT, Levine AS, et al. Causal feature selection using a knowledge graph combining structured knowledge from the biomedical literature and ontologies: a use case studying depression as a risk factor for Alzheimer's disease. Journal of biomedical informatics. 2023:104368.
7. Malec SA, Wei P, Bernstam EV, Boyce RD, Cohen T. Using computable knowledge mined from the literature to elucidate confounders for EHR-based pharmacovigilance. Journal of biomedical informatics. 2021;117:103719.
8. Kilicoglu H, Rosemblat G, Fiszman M, Shin D. Broad-coverage biomedical relation extraction with SemRep. BMC Bioinformatics. 2020;21(1):188.
9. Sosa DN, Altman RB. Contexts and contradictions: a roadmap for computational drug repurposing with knowledge inference. Briefings in bioinformatics. 2022;23(4):bbac268.
10. Kilicoglu H, Rosemblat G, Rindflesch TC. Assigning factuality values to semantic relations extracted from biomedical research literature. PLoS One. 2017;12(7):e0179926.
11. Djulbegovic B, Guyatt GH. Progress in evidence-based medicine: a quarter century on. Lancet. 2017;390(10092):415-23.
12. Atkinson K, Baroni P, Giacomin M, Hunter A, Prakken H, Reed C, et al. Towards Artificial Argumentation. AI Magazine. 2017;38(3):25-36.
13. Slonim N, Bilu Y, Alzate C, Bar-Haim R, Bogin B, Bonin F, et al. An autonomous debating system. Nature. 2021;591(7850):379-84.
14. Lawrence J, Reed C. Argument Mining: A Survey. Computational Linguistics. 2020;45(4):765-818.
15. Aharoni E, Polnarov A, Lavee T, Hershcovich D, Levy R, Rinott R, et al., editors. A Benchmark Dataset for Automatic Detection of Claims and Evidence in the Context of Controversial Topics2014 June; Baltimore, Maryland: Association for Computational Linguistics.
16. Lawrence J, Reed C. Argument mining: A survey. Computational Linguistics. 2019;45(4):765-818.
17. Marro S. Argumentation quality : from general principles to healthcare applications
Qualité de l'argumentation : des principes généraux aux applications dans le domaine de la santé: Université Côte d'Azur; 2023.
18. Truhn D, Reis-Filho JS, Kather JN. Large language models should be used as scientific reasoning engines, not knowledge databases. Nature medicine. 2023;29(12):2983-4.
19. Jiang T, Zeng Q, Zhao T, Qin B, Liu T, Chawla NV, et al. Biomedical Knowledge Graphs Construction From Conditional Statements. IEEE/ACM Transactions on Computational Biology and Bioinformatics. 2021;18(3):823-35.
20. Pengfei Y, Hansi Z, Xing H, Matthew D, Qian L, Shubo T, et al. Towards Formal Computable Representation of Clinical Trial Eligibility Criteria for Alzheimer's Disease. medRxiv. 2022:2022.03.21.22272707.
21. Open A. ChatGPT Overview 2023 [Available from: https://openai.com/chatgpt.
22. Chen G, Cheng L, Luu AT, Bing L. Exploring the Potential of Large Language Models in Computational Argumentation. ArXiv. 2023;abs/2311.09022.