# Using reasoning for citation context classification

Birger Larsen[1,*,†], Roman Jurowetzki[2,*,†]

[1] Aalborg University, Dept. of Communication and Psychology, A. C. Meyers Vænge 15, 2450 Copenhagen C, Denmark
[2] Aalborg University Business School, Fibigerstræde 11, 9220 Aalborg East, Denmark

## Abstract

While some success has been achieved in the classification of citation contexts by using Large Language Models (LLMs) [1], significant improvements are still needed to achieve sufficient accuracy for large scale real world application. We propose to use recent LLM models with AI reasoning capabilities (like QwQ, DeepSeek R1 and Gemini Thinking) to study if reasoning processes have the potential to achieve such performance improvements. We outline here how we plan to test this on two tasks: 1) classification of citation contexts, and 2) elimination of low-quality annotations from training sets. Full results will be presented at the 1st Scolia workshop. We also describe our experiences in creating a small annotated citation context dataset, and present some initial results.

## Keywords

Citation Context Classification, Reasoning Models, Data Annotation Cleaning

## 1. Introduction

A citation context can be defined as “that particular passage or statement within the citing document containing the references” [2, p. 288]. Where most work in Bibliometrics and citation analysis has, in one way or another, focused on the *number* of citations received, the analysis of citation contexts are interesting because they may provide insights into *why* a given paper has been cited. With almost all scientific text being produced, published and archived in electronic formats, and with an increasing share of these as Open Access, citation contexts analysis may greatly enhance the understanding scientific discourse, knowledge creation and claims, and the wider impact of science. If citation contexts can be automatically extracted, processed and classified they have great potential for improving access to and exploitation of scientific knowledge from scientific documents with a broad set of potential application, e.g. in research analysis, science of science studies, patent and innovation analysis, information retrieval and information visualization etc.

Where early studies of citer motivations and referencing behavior largely was done manually and qualitatively (see [3] for an overview), the advent of machine learning, large language Models (LLMs) and generative AIs as well as the availability of a sizable proportion of scientific publications in electronic formats, makes it possible to investigate if citation contexts can be automatically extracted and classified - and with what accuracy.

Even with scientific publications in electronic formats extracting citation contexts can be challenging. Publications marked up in structured formats like XML can make identification and extraction of citations contexts relatively straightforward. For instance, citation contexts can be easily identified and extracted from PubMedCentral fulltext articles because of XML tags marking in-text references (and linking them to the bibliography) [4, 5]. Citation contexts can also be extracted from LaTeX documents, albeit with some more difficulty and less accuracy [6]. Extraction from widely used, but unstructured formats like PDF is a bigger challenge, but is possible with combination of several tools and pipelines. For instance, the team at CORE is now able to reliably extract citation contexts as well as salient document features from large corpora of PDFs from diverse sources [7].

*SCOLIA '25: First International Workshop on Scholarly Information Access (SCOLIA), April 10, 2025, Lucca, Italy*

[*] Corresponding author.

[†] These authors contributed equally.

birger@ikp.aau.dk (B. Larsen); roman@business.aau.dk (R. Jurowetzki)
https://vbn.aau.dk/en/persons/birger (B. Larsen); https://https://vbn.aau.dk/en/persons/roman (R. Jurowetzki)
0000-0002-3622-2698 (B. Larsen); 0000-0002-5199-918X (R. Jurowetzki)

Citation context classification using machine learning has mainly relied on supervised methods, using large annotated datasets of citation contexts for fine-tuning [1]. The performance of such models depends mainly of the size of the training set, which has limited research in the area because annotations are cumbersome to generate (ideally annotations should be done by scientists active in the research fields in question). Recently, work has been done using generative models for citation context classification, either by direct zero-shot prompting [8] , by combinations of prompting and LM fine-tuning [1] or by zero to many-shot prompting [9]. Results are promising, but not consistently good enough to e.g. reach the level of human annotators.

Our long term goal is to investigate solutions where citation context classification can be brought to a level of performance that would allow it to be used with confidence in real-world scenarios. This will likely involve a combination of both creating annotated training sets and fine-tuning of existing models on one hand, and on the other investigating where prompting can aid in improving and extending training sets as well as citation context classification, incrementally combining these approaches over time in several rounds to increase performance and scale up the research in citation context classification to process large dirverse corpora of scientific publications reliably.

In this paper, we propose to investigate if some of the most recently released LLMs with integrated reasoning capabilities can significantly improve citation context classification performance as well as assessing annotation quality. Reasoning models are attractive because they may allow us to get a deeper insight into how the models arrive at their decisions and thus enable us to improve.

Our contributions are:

- We describe our experiences in creating a small annotated citation context dataset, and present some first results of applying it.
- We propose to use reasoning models for citation context classification, and outline several lines of research that might benefit from their application.
- We present some first results of applying reasoning models in assessing annotation quality.

## 2. Methods

Our approach is situated within a broader trend in computational social science where LLMs are increasingly explored as tools for qualitative and quantitative text analysis. LLMs offer the potential to automate and scale complex coding tasks, reduce the labor-intensive nature of manual annotation, and potentially improve consistency and objectivity in social science research [10]. For instance, LLMs have shown promise in assisting deductive coding [11], supporting thematic analysis [12], and providing accurate annotations for complex texts, even outperforming human experts in certain scenarios [13].

We leverage recent advancements in LLMs with integrated reasoning capabilities to address the challenges in citation context classification and annotation quality assessment. We employ a zero-shot classification methodology, directly prompting reasoning-enabled LLMs such as QwQ and DeepSeek R1 to classify citation contexts based on predefined categories. The selection of models like DeepSeek R1 is motivated by their demonstrated expert-level reasoning capabilities on complex tasks as well as the fact that they expose their reasoning traces - unlike OpenAI's o1 and o3 [14, 15].

A key aspect of our methodology is the analysis of the reasoning traces generated by these models, inspired by work on self-reflection in LLM agents [16]. These traces, which detail the step-by-step thought process of the LLM in arriving at a classification decision, offer a unique opportunity for several valuable insights. Firstly, by examining these traces, we can gain a deeper understanding of how the LLM interprets and applies our classification guidelines to the citation contexts. This allows us to iteratively refine and improve these guidelines, making them more precise and less ambiguous for both automated and human annotators. Secondly, the reasoning traces provide a basis for comparison with the cognitive processes of human annotators who previously worked on creating the training dataset. By contrasting the LLM's reasoning with human annotation rationales, we can identify potential discrepancies, biases, or areas of subjective interpretation inherent in the classification task.

**Table 1**
Citation context classification scheme designed by [3] to have non-overlapping classes. The ’Data’ class was added for the present study.

| | |
|---|---|
| 1. Background/Perfunctory | Checks whether the cited article is merely a part of the relevant literature and is not analyzed or compared to other literature |
| 2. Contemporary | Checks whether the given citation is explicitly characterized as “recent” by the author. |
| 3. Contrast/Conflict | Checks whether results or opinions in a given citation show contrast to or conflict with an opinion or result presented by the author of the citing paper. |
| 4. Evaluation | Checks if the results in the cited study are evaluated in the citing study |
| 5. Explanation | Checks whether the cited work helps explain the results or hypotheses in current study |
| 6. Method | Checks whether the given citation is a methodology that was followed in the citing work (with or without modifications) |
| 7. Data | Checks whether the given citation points to a data set from another source that has been used (fully or in part) in the citing article |
| 8. Modality | Checks whether the citing author expresses lack of certainty over a result or opinion in the cited study |
| 9. Similarity/Consistency | Checks whether results or opinions in the given citation are similar or consistent with the given study or another cited work |
| 10. N/A | If none of the above can be assigned |

To quantitatively assess the agreement between the LLM classifications and the existing human annotations, we calculate Inter-Rater Reliability (IRR) scores. Specifically, we plan to use Gwet’s AC1, a robust IRR metric particularly suitable for situations with varying numbers of raters or when dealing with class imbalance, which is often encountered in citation context classification tasks.

In the context of citation context classification, reasoning-enabled LLMs could significantly enhance our ability to process large volumes of scholarly literature, providing valuable insights into the dynamics of scientific knowledge production and dissemination. By focusing on the reasoning process and comparing it to human annotation, we aim to not only improve the accuracy of automated citation context classification but also to gain a deeper understanding of the nuances and challenges inherent in this task, ultimately contributing to more robust and reliable methods for scholarly information access.

## 3. Dataset

We created a small annotated dataset of citation contexts from PubMedCentral (PMC). PMC is a free full-text archive of biomedical and life sciences journal literature at the U.S. National Institutes of Health’s National Library of Medicine (NIH/NLM)[1]. PMC currently includes more than 10 million full-text articles, stored in a JATS XML format, most of which are available as Open Access.

For annotation of citation contexts, we used a scheme developed specifically for biomedical articles by [3] to be non-overlapping, i.e. only one class is to be used per citation context. Given current interest in the FAIR Guiding Principles for scientific data management and stewardship[2] we added an additional category for ’Data’ - for whether the given citation points to a data set from another source that has been used (fully or in part) in the citing article. The scheme can be seen in Table 1.

We chose a single PMC research article as a starting point, aiming at a recent seed document (from 2020) that had a fair number of citations (41 at the time of sampling), but was not extremely highly cited. We retrieved this document [17] and the ones citing it in PMC. We annotated all citation contexts in the articles directly citing the seed document, and did a stratified sample of other citation contexts in these

[1] https://pmc.ncbi.nlm.nih.gov/about/intro/
[2] https://www.go-fair.org/fair-principles/

citing articles (sampling every nth reference from the reference list to make a total of 15 citation contexts from each citing document). The advantage of this was to sample references across the document, increasing the chance of annotating citations contexts from across each document, while minimizing annotation costs.

We hired two of our MSc students for annotation with a fixed number of hours to complete as many annotations as possible from this sample. As we did not have access to biomedical students, both students were in the field of Information Technology. Annotators worked with the XML fulltext of the citing articles, and extracted: 1) the Xpath to the sentence containing the citation marker, 2) the text of this sentence incl. any XML marking = the citation context, 3) the internal referenceID, 4) Xpath to lowest (sub)section contain the citation context, 5) the title of this section (e.g. Introduction, Discussion, Data Analysis, Statistical Analysis, Strengths and limitations of the data, etc.). In addition, annotators assessed: 8) context citation types (using the scheme discussed above), 9) own confidence in their assessment (High/Medium/Low), and 10) whether in their view additional context was needed for interpretation, i.e. if one, two or more sentences before/after was needed).

A total of 585 annotations were collected. Table 1 shows the distribution of citation contexts types as well as the annotator confidence. As expected from previous research the Background/Perfunctory class has the highest share (50%), followed by Explanation (22%), Similarity/Consistency (7%) and Contrast/Conflict (6%). Other classes were 5% or less, and no contexts were assessed as Contemporary. Overall, assessor confidence in their own judgments were High (87%), with 12% Medium and only 1% Low. In order to highlight the distribution of H/M/L confidence levels, the percentages are tallied for each context type. It can be seen that some of the less used context types has a high share of Medium confidence, e.g. 11 out of 20 Method instances (55%) had Medium confidence - similar levels can be seen for Modality (50%) and Evaluation (46%). Table 3 shows that in 80% of cases no additional context was needed beyond the sentence with the citation marker, and that in 11% one sentence before was needed with two sentences before needed in 4%. In only 2% of cases was one sentence after needed.

The data allows us to study the distribution of citation context types of the seed document: 37 citation contexts citing it were annotated: 16 (43%) were Background/Perfunctory, 12 (32%) were Explanation, 4 (11%) were Similarity/Consistency, 3 (8%) were Contrast/Conflict, and 2 (5%) were Evaluation.

## 4. Preliminary Experiments and Results

To evaluate the feasibility of employing reasoning models for citation context classification, we conducted initial experiments utilizing the DeepSeek R1 (DeepSeek-R1-Distill-Qwen-32B) Large Language Model. We used the 585 citation contexts from our manually annotated dataset, detailed in Section 3. In a zero-shot setting, each citation context, along with its section title and reference ID, was presented to DeepSeek R1 with prompts designed to classify it into one of ten predefined categories (Section 2). The model was instructed to output its classification in JSON format, accompanied by a reasoning trace explaining its decision-making process.

Following classification of all 585 contexts, we quantitatively assessed DeepSeek R1's performance by comparing its classifications to the human annotations in our dataset. We calculated standard evaluation metrics, including Gwet's AC1 inter-rater reliability, accuracy, precision, recall, and F1-score. These metrics provide a comprehensive evaluation of the model's agreement with human annotators and its overall classification effectiveness.

Analysis of the initial classification results revealed an overall accuracy of 0.508. Gwet's AC1 inter-rater reliability was calculated at -0.798, indicating poor agreement beyond chance between the LLM and human annotations across all categories. Examining class distribution, 'Background/Perfunctory' (Class 1) was the most frequent category in human annotations (50.4%), while the LLM distributed its classifications more broadly, notably increasing the proportion of 'Data' (Class 7) and 'Similarity/Consistency' (Class 9) classifications. The weighted-average F1-score, accounting for class imbalance, was 0.517. Per-class metrics showed that 'Background/Perfunctory' (Class 1) achieved a precision of 0.732, recall of 0.583, and an F1-score of 0.649. This higher performance is likely due to the class's dominant

**Table 2**
Distribution of citation contexts types as well as annotator confidence. Percentages for confidence levels are tallied for each context type to highlight differences with each type.

| | | | Confidence | | | | | |
|---|---|---|---|---|---|---|---|---|
| Context type | Sum | (%) | H | | M | | L | |
| 1. Background/Perfunctory | 295 | 50% | 281 | 95% | 13 | 4% | 1 | 0% |
| 2. Contemporary | 0 | | 0 | | 0 | | 0 | |
| 3. Contrast/Conflict | 38 | 6% | 30 | 79% | 8 | 21% | | |
| 4. Evaluation | 28 | 5% | 14 | 50% | 13 | 46% | 1 | 4% |
| 5. Explanation | 126 | 22% | 118 | 94% | 8 | 6% | | |
| 6. Method | 20 | 3% | 9 | 45% | 11 | 55% | | |
| 7. Data | 29 | 5% | 21 | 72% | 8 | 28% | | |
| 8. Modality | 4 | 1% | 2 | 50% | 2 | 50% | | |
| 9. Similarity/Consistency | 43 | 7% | 34 | 79% | 9 | 21% | | |
| 10. N/A | 2 | 0.3 | 0 | | 0 | | 2 | 100% |
| | 585 | 100% | 509 | | 72 | | 4 | |

**Table 3**
Assessor judgment on whether additional context beyond the sentence with the citation marker was needed for the interpretation of citation intent

| More context needed for interpretation? | | |
|---|---|---|
| 1 sentence before needed | 62 | 11% |
| 2 sentences before needed | 22 | 4% |
| 2+ sentences before needed | 14 | 2% |
| No additional context needed | 470 | 80% |
| 1 sentence after needed | 13 | 2% |
| 2 sentences after needed | 2 | 0% |
| 2+ sentences after needed | 2 | 0% |
| | 585 | 100% |

presence in the dataset. However, performance on other categories, particularly less frequent ones, was considerably lower, suggesting room for improvement in aligning LLM classifications with human annotations across the spectrum of citation context categories.

To further understand the discrepancies between human and LLM classifications, we performed a preliminary qualitative analysis on a subset of disagreements. Examining a manual sample of 15 instances, we observed a trend suggesting that the reasoning provided by DeepSeek R1 often indicated classifications that were more contextually accurate than the original human (student) annotations. The LLM demonstrated a nuanced understanding of citation function, frequently capturing subtle contextual cues pointing to categories beyond the often-applied 'Background/Perfunctory' label. The reasoning traces highlighted the model's sensitivity to linguistic markers of contrast, explanation, and consistency, suggesting an ability to discern the rhetorical role of citations in scientific discourse. This initial qualitative exploration suggests that human annotations, while the initial ground truth, could benefit from critical review and revision. The observed discrepancies point to the need to further refine annotation guidelines for greater consistency and to better capture the nuanced functional roles of citations. This iterative process of comparing human and AI classifications holds promise for not only improving automated citation context classification accuracy but also enhancing the quality and reliability of the training data itself.

The detailed results of this evaluation, including specific metric values and further analysis of the model's reasoning traces, will be presented at the SCOLIA'25 workshop. These preliminary experiments provide a crucial initial step in understanding the capabilities of reasoning models for citation context classification and will guide our subsequent research directions.

## Acknowledgments

This research was supported by a seed project grant by *maSSHine*, the Aalborg University computational Social Sciences and Humanities initiative.

## Declaration on Generative AI

During the preparation of this work, the author(s) used the **QwQ-32B reasoning model** to **predict citation context categories**, and for **prompt improvements and instruction fixes** through a self-tuning approach, as well as **Gemma-3-27B** for **citation context classification** using the augmented prompts.

## References

[1] S. N. Kunnath, D. Pride, P. Knoth, Prompting strategies for citation classification, in: Proceedings of the 32nd ACM International Conference on Information and Knowledge Management, CIKM '23, Association for Computing Machinery, New York, NY, USA, 2023, p. 1127–1137. URL: https://doi.org/10.1145/3583780.3615018. doi:10.1145/3583780.3615018.

[2] H. Small, Citation context analysis, in: B. J. Dervin, M. J. Voigt (Eds.), Progress in Communication Sciences, volume 3, 1982, pp. 287–310.

[3] S. Agarwal, L. Choubey, H. Yu, Automatically classifying the role of citations in biomedical articles, in: AMIA Annual Symposium Proceedings, 2010, pp. 11–15.

[4] T.-K. Hsiao, V. I. Torvik, OpCitance: Citation contexts identified from the pubmed central open access articles, Scientific Data 10 (2023).

[5] B. Nielsen, S. Skau, F. Meier, B. Larsen, Optimal citation context window sizes for biomedical retrieval, CEUR Workshop Proceedings of the 8th International Workshop on Bibliometric-Enhanced Information Retrieval, BIR 2019 - Cologne, Germany 2345 (2019) 51–63.

[6] A. Dabrowska, B. Larsen, Exploiting citation contexts for physics retrieval, CEUR Workshop Proceedings of the Second Workshop on Bibliometric-enhanced Information Retrieval : co-located with the 37th European Conference on Information Retrieval (ECIR 2015) 1344 (2015) 14–21.

[7] S. Nambanoor Kunnath, V. Stauber, R. Wu, D. Pride, V. Botev, P. Knoth, ACT2: A multi-disciplinary semi-structured dataset for importance and purpose classification of citations, in: N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, S. Piperidis (Eds.), Proceedings of the Thirteenth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2022, pp. 3398–3406. URL: https://aclanthology.org/2022.lrec-1.363/.

[8] K. Nishikawa, H. Koshiba, Exploring the applicability of large language models to citation context analysis, Scientometrics 129 (2024) 6751–6777. URL: https://ideas.repec.org/a/spr/scient/v129y2024i11d10.1007_s11192-024-05142-9.html. doi:10.1007/s11192-024-05142-.

[9] P. Koloveas, S. Chatzopoulos, T. Vergoulis, C. Tryfonopoulos, Can LLMs predict citation intent? an experimental analysis of in-context learning and fine-tuning on open LLMs, 2025. URL: https://arxiv.org/abs/2502.14561. arXiv:2502.14561.

[10] C. Ziems, W. Held, O. Shaikh, J. Chen, Z. Zhang, D. Yang, Can large language models transform computational social science?, Computational Linguistics 50 (2024) 237–291.

[11] R. Chew, J. Bollenbacher, M. Wenger, J. Speer, A. Kim, LLM-assisted content analysis: Using large language models to support deductive coding, arXiv preprint arXiv:2311.18716 (2023).

[12] S.-C. Dai, A. Xiong, L.-W. Ku, LLM-in-the-loop: Leveraging large language model for thematic analysis, arXiv preprint arXiv:2310.15100 (2023).

[13] P. Törnberg, Chatgpt-4 outperforms experts and crowd workers in annotating political twitter messages with zero-shot learning, arXiv preprint arXiv:2304.06588 (2023).

[14] D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma, P. . Wang, X. Bi, DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning, arXiv preprint arXiv:2501.12948 (2025).

[15] D. Rein, B. Hou, A. Stickland, J. Petty, R. Pang, J. Dirani, J. Michael, S. Bowman, GPQA: A graduate-level google-proof Q&A benchmark, arXiv preprint arXiv:2311.12022 (2023).

[16] M. Renze, E. Guven, Self-reflection in LLM agents: Effects on problem-solving performance, arXiv preprint arXiv:2405.06682v1 (2024).

[17] M. Chadeau-Hyam, B. Bodinier, J. Elliott, M. D. Whitaker, I. Tzoulaki, R. Vermeulen, M. Kelly-Irving, C. Delpierre, P. Elliott, Risk factors for positive and negative covid-19 tests: a cautious and in-depth analysis of UK biobank data, International Journal of Epidemiology 49 (2020) 1454–1467. URL: http://dx.doi.org/10.1093/ije/dyaa134. doi:10.1093/ije/dyaa134.