International Journal of English Literature and Social Sciences
Vol-11, Issue-4; Jul-Aug, 2026

Peer-Reviewed Journal
Journal Home Page Available: https://ijels.com/
Journal DOI: 10.22161/ijels

Large Language Models for Citation Context and Cited Content Recognition: From Boundary Detection to Evidence Grounding

Shuqiao Yang, Xiaolei Ma*, Ping He

[1]Basic Education College, National University of Defense Technology, Changsha, China
Email: 1127598981@qq.com, maxiaolei@nudt.edu.cn, peace94@126.com

Received: 22 Jul 2026; Received in revised form: 18 Aug 2026; Accepted: 21 Aug 2026; Available online: 27 Aug 2026

Abstract— *Citation analysis has traditionally been organized around citation intent, function, stance, citation context analysis, and cited text span identification. Recent large language models (LLMs) have been applied to these tasks through prompting, in-context learning, fine-tuning, annotation assistance, and retrieval-augmented generation. Yet it remains unclear whether these studies also develop a more explicit account of citation recognition. This survey reviews LLM-based citation analysis through a recognition-centered lens. We define citation recognition as the identification of the textual and semantic units needed for citation understanding, especially citation context, citation content, and cited content. We summarize pre-LLM assumptions about fixed context windows, interpretative labels, and cited text span grounding, and then synthesize five LLM-based method paradigms: prompting and fine-tuning, LLM-assisted annotation, long-context recognition, retrieval-augmented grounding, and multi-dimensional frameworks. We argue that the next stage of LLM-based citation recognition should give closer attention to boundary quality, content consistency, evidence faithfulness, and reproducibility.*

Keywords— ***citation analysis, citation context, cited content grounding, large language models, scientific text mining***

I. INTRODUCTION

Citation is a fundamental mechanism through which scholarly texts connect claims, methods, evidence, concepts, and research communities [29, 11, 13]. A citation is not merely a bibliographic link: it appears in a local textual environment, refers to some part of a cited work, and supports a scholarly act such as providing background, adopting a method, comparing results, acknowledging prior work, or challenging a claim. Understanding a citation therefore requires more than identifying the referenced paper. It requires recognizing the textual and semantic units that make the citation meaningful.

Computational citation analysis has traditionally been organized around tasks such as citation intent classification, function or purpose detection, stance and sentiment analysis, citation context analysis, reference scope detection, and cited text span identification [14, 4, 5, 22, 28, 27]. These tasks ask different questions: why a work is cited, how it is evaluated, where the relevant citing context lies, and what content of the cited work is invoked. Despite their differences, they share a common prerequisite. Before a citation can be interpreted, evaluated, or categorized, a system must recognize the relevant citing-side context, the semantic content expressed in that context, and, when available, the corresponding evidence in the cited document.

This survey uses **citation recognition** as an organizing lens for reviewing LLM-based citation analysis. We do not treat citation recognition as a single established benchmark task. Rather, we use it to denote the process of identifying the units needed for citation understanding. Three targets are central: **citation context recognition**, which identifies the relevant citing-side span; **citation content**

recognition, which identifies what the citing text says about the cited work; and **cited content recognition** or grounding, which identifies the invoked content or evidence span in the cited work. We formalize the relation among these targets in Section 2. Interpretative labels such as intent, function, stance, sentiment, and polarity remain important, but we treat them as downstream analytical dimensions that depend on recognition while also providing signals for recognizing context, content, and evidence.

Large language models (LLMs) make this recognition problem newly important. Recent studies have applied prompting, in-context learning, fine-tuning, annotation assistance, and retrieval-augmented workflows to citation-related tasks [3, 21, 17, 16, 15, 24, 10, 30, 19]. These methods improve flexibility and reduce dependence on task-specific architectures or large annotated datasets. They also create new opportunities to connect citing-side contexts with cited-side evidence. However, it remains unclear how far current LLM-based work has developed the ability to recognize citation context, citation content, and cited evidence in an explicit and verifiable way.

This uncertainty motivates the central question of this survey: **to what extent do LLMs contribute to citation recognition?** We address this question through three research questions: **RQ1:** how have citation context and cited content been operationalized before LLMs? **RQ2:** what exactly do LLMs contribute to citation recognition? **RQ3:** what unresolved problems remain in LLM-based citation recognition?

Existing surveys do not fully answer these questions. Earlier surveys on citation context analysis and NLP-driven citation analysis summarize pre-LLM tasks, techniques, and resources [11, 13], while recent surveys on LLMs and citation analysis cover broader topics such as citation recommendation, citation generation, citation-enhanced language modeling, and citation networks [31]. In contrast, this survey focuses specifically on how LLMs inherit, revise, or expose assumptions about citation context boundaries, cited content grounding, and interpretative label spaces.

The contributions of this survey are fourfold. First, we propose a recognition-centered lens for reviewing LLM-based citation analysis, distinguishing citation context, citation content, and cited content grounding. Second, we synthesize LLM contributions along an ability progression from flexible label recognition to boundary awareness, evidence grounding, and multi-dimensional interpretation. Third, we analyze how datasets and evaluation protocols shape what counts as successful recognition. Fourth, we identify unresolved challenges and future directions, including context-boundary instability, intent-content conflation, hallucinated cited content, weak evidence-level evaluation, prompt sensitivity, and human-LLM annotation reliability.

II. CITATION RECOGNITION: SCOPE AND PRE-LLM ASSUMPTIONS

LLM-based citation recognition does not start from zero. It inherits pre-LLM concepts, datasets, and modeling assumptions. This section therefore focuses not on the entire history of citation analysis, but on the assumptions that LLM-based work reproduces, revises, or must still overcome.

2.1 Recognition Targets and Their Relation

We distinguish three recognition targets, summarized in Figure 1. **Citation context recognition** identifies the citing-side span relevant to a target citation. This span may be a citation sentence, a neighboring window, a paragraph, an implicit context without a citation marker, or a narrower reference scope in a sentence containing multiple citations. **Citation content recognition** identifies the semantic content expressed by the citing-side span, such as a method, dataset, theory, result, limitation, comparison, or criticism attributed to or associated with the cited work. **Cited content recognition**, also called **cited content grounding**, identifies the evidence span in the cited document that supports or corresponds to that recognized content. In short, context recognition asks where to read in the citing document, citation content recognition asks what is being said there, and cited content grounding asks where the corresponding evidence lies in the cited document.

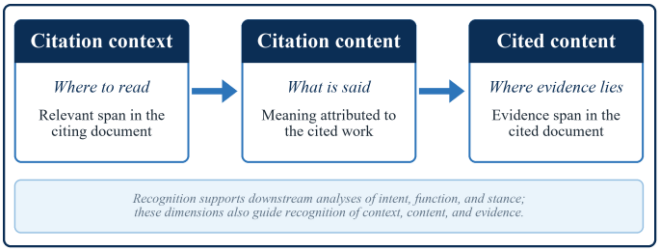

Fig. 1: A recognition-centered view of citation analysis.

These targets are related but not reducible to one another. In many workflows, context recognition provides the textual basis for content recognition, because the model must know which citing-side span to read before deciding what the citation says. Yet the relation is not a simple pipeline. A preliminary content judgment may revise the context boundary, and cited-side evidence may reveal that the chosen context is too narrow, too broad, or attached to the wrong reference. Citation recognition is

therefore an iterative relation among boundary detection, content typing, and evidence grounding. Interpretative labels such as intent, function, and stance become useful signals inside this relation rather than substitutes for it.

2.2 Pre-LLM Assumptions

Pre-LLM work contributed three assumptions that continue to shape LLM-based citation studies. First, citation context was often approximated by **fixed or weakly modeled boundaries**. Many systems used the citation sentence or a fixed window around the marker, although survey and framework work distinguished citing sentence, citation context, implicit context, and reference scope [11, 13, 29]. This established the importance of boundary selection, but many computational systems still assumed the input context rather than learning it.

Second, citation understanding was frequently operationalized through **interpretative labels**. Function, purpose, centrality, stance, and polarity became practical annotation and classification targets. ACL-ARC-style function annotation and SciCite made large-scale modeling possible [14, 4], but they also encouraged a view of citation recognition as label prediction over predefined contexts. Neural and pre-trained models improved performance within this formulation [4, 23, 25], while leaving boundary selection and evidence grounding largely outside the benchmark.

Third, cited content grounding emerged through **cited text span identification**. Cohan et al. formulated citation text and cited span matching as a search problem; later work treated cited text span identification as imbalanced classification and ensemble learning [5, 22, 28]. This line shifted attention from the citing sentence to the cited document and showed that citation understanding often requires cross-document alignment. LLMs inherit all three assumptions: given contexts, label spaces, and cited-span grounding tasks. The key question is therefore whether, and how, LLMs contribute to citation recognition itself.

III. LLM-BASED METHOD PARADIGMS FOR CITATION RECOGNITION

LLM-based citation recognition has not yet formed a mature family of task-specific architectures comparable to fields with many dedicated neural models. The current literature is better understood through five method paradigms: **prompting and fine-tuning**, **LLM-assisted annotation**, **long-context use**, **retrieval-augmented grounding**, and **multi-dimensional frameworks**. These paradigms show how LLMs contribute to recognition and where context, content, and evidence remain under-specified.

3.1 Prompting and Fine-Tuning for Label-Oriented Recognition

The most visible LLM-era method is the reformulation of citation-related recognition as **prompting or fine-tuning** over natural-language task descriptions. Instead of designing a task-specific architecture, researchers can describe the label scheme in natural language, provide examples, fine-tune an open model, or ask a general LLM to assign an intent or function label. This follows the broader prompting paradigm in NLP [21] and uses few-shot capabilities associated with large language models [3]. In this paradigm, the typical input is a citation sentence or supplied citation context plus a task instruction; the output is a label, verbalized category, or short rationale.

Existing work supports this paradigm. **CitePrompt** uses prompts and label verbalizers to elicit citation intent in zero-shot and few-shot settings [17]. **Prompting studies** compare how prompt design, context representation, and tuning strategy affect citation classification outputs [16]. Experiments with **open LLMs** further compare in-context learning and fine-tuning across shot settings [15]. These studies are methodologically important because they show that LLMs can adapt to inherited citation categories with less task-specific engineering.

The limitation is that many studies still evaluate a pre-existing label over a given context. The model is usually not asked to decide where the relevant citation context begins or which cited-document passage supports its output. If the dataset uses a fixed citation sentence, the LLM learns inside that boundary. If a label scheme conflates method-related intent with method-related cited content, the LLM may reproduce the conflation. Thus, prompting and fine-tuning improve access to inherited labels, but they do not by themselves make recognition boundary-aware or evidence-grounded.

3.2 LLM-Assisted Annotation for Recognition Decisions

A second method paradigm treats **LLM-assisted annotation** as a human-LLM workflow rather than autonomous recognition. In this setting, the model may propose labels, rationales, or candidate contexts, while human annotators validate, revise, or adjudicate the output. Nishikawa and Koshiba apply LLMs to citation context analysis and show that LLM-generated annotations can be useful but are not reliable enough to replace human annotation [24]. More general annotation frameworks such as **AnnoLLM** and **LLMaAA** similarly emphasize guided annotation, active selection, and human-LLM workflows [10, 30].

For citation recognition, annotation assistance is important because it can make boundary and content disagreement visible. If annotators disagree about whether the relevant context is a sentence, neighboring window, or reference scope, the disagreement itself reveals the difficulty of recognition. The remaining challenge is that LLM-assisted annotation requires explicit guidelines, records of human-LLM disagreement, and adjudication procedures.

3.3 Long-Context Methods for Boundary-Aware Recognition

Long-context use is another method paradigm. Earlier work already showed that citation context is not always identical to the citation sentence; it may include implicit context, neighboring discourse, or reference scope within a multi-citation sentence [11, 13, 29]. Compared with earlier models, LLMs can read longer contexts and use discourse-level information, which creates an opportunity to recognize broader or implicit citation contexts. Methodologically, this paradigm changes the input from a fixed citation sentence to a larger passage, section, or document-level window.

At the same time, current LLM-based work has not yet made boundary recognition a stable output target. In many studies, the citation sentence or context window is still supplied by the dataset or experimenter, so the model interprets a given span rather than selecting the span that should be read. This matters for multi-citation sentences, implicit contexts, and evaluative meanings that appear before or after the cited sentence. If an LLM reads a broad window without identifying which parts belong to the target citation, it may blend evidence across references or rely on irrelevant context.

Thus, the contribution of LLMs is best understood as an enabling condition rather than a solved capability. Their long-context and instruction-following abilities make boundary-aware recognition more feasible, but a stronger citation recognizer should still output the citing-side span that supports its interpretation and, where relevant, the reference scope that links that span to a particular cited work.

3.4 Retrieval-Augmented Grounding across Citing and Cited Documents

The most distinctive method paradigm for cited content recognition is **retrieval-augmented grounding**. A citation is a cross-document relation, and understanding what is cited often requires connecting the citing context with evidence in the cited work. Pre-LLM research formulated this problem through citedspan matching and cited text span identification [5, 22, 28]. LLMs make this line newly important because they can combine retrieval, semantic comparison, long-context reasoning, and natural-language explanation. In this paradigm, the system retrieves candidate cited passages, compares them with the citing context, and uses the evidence to generate or justify a recognition output.

In an evidence-grounded recognizer, the output would be a relation among citing-side context, cited-side evidence, and interpretative dimensions. Current work supports this direction while showing that it is immature. Li et al. construct cited text spans for scientific citation text generation and argue that citation generation should be grounded in the cited paper rather than relying only on abstracts [19]. Work on attributable generation and generative search shows that generated citations do not automatically guarantee evidential support: citations may be present but unsupported, irrelevant, or only partially faithful [9, 20, 12]. For citation recognition, this means that cited content grounding must be evaluated at the evidence level, together with the fluency and plausibility of generated explanations.

3.5 Multi-Dimensional Frameworks

Another LLM-era opportunity is **multi-dimensional recognition**. Traditional citation classification often asks one label to do too much. A label such as "method" may refer to the citing author's purpose, the cited work's content type, or a surface cue in the citation context. The need for separation predates LLMs: citation content analysis argues for syntactic and semantic analysis of citation content [29], citation concept analysis focuses on the concepts made useful by citations [2], and stance research shows that evaluative relations cannot be reduced to simple opinions [27].

In the LLM era, emerging frameworks such as **SOFT** can be read as attempts to operationalize this broader need by separating citation intent from cited content [6]. Because this line is still developing, it should be treated as an emerging attempt rather than a mature benchmark tradition. Its value is conceptual: it shows how LLM-based citation recognition can move from single-label prediction toward structured interpretation. Cited content can guide intent, intent can guide grounding, and stance can guide boundary selection. The challenge is that multi-dimensional annotation requires clearer guidelines, richer datasets, and evaluation of cross-dimensional consistency.

IV. EVALUATION AND REMAINING CHALLENGES

The decomposition of citation understanding into context, interpretation, and evidence suggests a clearer way to read existing evaluation practices. Recognition is not the whole

of citation understanding, but it supplies the observable units on which understanding depends: the citing-side span, the content or interpretation associated with that span, and the cited-side evidence. Existing resources usually emphasize one of these units. Taken together, this literature suggests four evaluation dimensions for LLM-based citation recognition: boundary quality, content and interpretation consistency, grounding faithfulness, and robustness.

4.1 Boundary Accuracy and Context Stability

The first dimension is boundary accuracy: whether the model identifies the citing-side span relevant to a target citation. Pre-LLM surveys and frameworks already distinguish citation sentence, citation context, implicit context, and reference scope [11, 13, 29]. However, many widely used resources provide the context as input. ACL-ARC-style resources, SciCite, and 3C-style purpose datasets are valuable for comparing citation function or intent systems, but they usually do not ask the model to select the context boundary itself [14, 4, 25]. For LLM-based recognition, this matters because longer input windows do not automatically produce boundary awareness.

Boundary evaluation should therefore include span overlap or boundary F1 for citation contexts, reference-scope accuracy in multi-citation sentences, and stability under changes in prompt wording or context window size. Prompting studies and LLM annotation work show that prompt design, context representation, and annotation setup can affect citation-related outputs [16, 24]. These findings suggest that boundary recognition should be evaluated not only by one selected context, but also by whether the model remains stable when surrounding discourse changes.

4.2 Content and Interpretation Consistency

The second dimension is content and interpretation consistency. Once a context has been identified, the model must recognize what content is being attributed to the cited work and how that content functions in the citing argument. This is where citation content analysis, citation concept analysis, and stance studies remain important: they show that citation meaning may involve syntactic scope, semantic content, useful concepts, and evaluative positioning rather than a single flat dimension [29, 2, 27]. Emerging schemes such as SOFT can be understood as attempts to separate intent from cited content, although this direction is still developing [6].

Evaluation should therefore check whether recognized content and interpretative dimensions remain consistent with the selected context and, when available, the cited evidence. For example, a method-use interpretation should correspond to a context that discusses use, adoption, or comparison, and the cited-side evidence should plausibly contain the invoked method, dataset, or procedure. Human-LLM annotation studies also imply that consistency must be evaluated through guidelines, agreement, disagreement records, and adjudication rather than through a single prompt output [24, 10, 30].

4.3 Grounding Faithfulness

The third dimension is grounding faithfulness: whether the cited-side evidence supports the recognized citation content. Cited text span identification directly addresses this requirement by matching citation text to evidence spans in the cited document [5, 22, 28]. Citation text generation resources such as CiteBench and recent cited-span work further connect citing contexts with cited papers, making evidence grounding central to generation and recognition [7, 19].

LLM-based grounding also inherits risks from attributable generation and generative search. Work on citation generation and verifiability shows that a generated citation can be present while its cited evidence is irrelevant, incomplete, or unsupported [9, 20, 12]. For citation recognition, grounding evaluation should therefore assess support correctness, evidence relevance, and the relation between cited evidence and the recognized content. This is especially important when evidence is distributed across multiple spans or when the citing author paraphrases the cited work.

4.4 Robustness, Bias, and Reproducibility

Finally, recognition-centered evaluation requires robustness and reproducibility checks. LLM outputs may vary with prompt wording, demonstrations, retrieval settings, and decoding choices [16, 24, 32]. Citation bias is another concern: LLMs may reflect and heighten human citation patterns, especially high-citation bias [1]. Although citation bias has been studied more directly in recommendation and generation than in recognition, it matters whenever models interpret or reconstruct citation relations. Reliable citation recognition should therefore report prompts, retrieval settings, evidence sources, and robustness checks. The central gap is that few resources jointly annotate citation context boundaries, citation content, cited evidence, and interpretative dimensions.

Table 1 summarizes the main LLM method paradigms discussed in Section 3 and the recognition gaps that motivate the evaluation dimensions above.

Table 1: LLM method paradigms and remaining recognition gaps

Method paradigm	Recognition target	Remaining gap
Prompting and fine-tuning	Intent/function labels and content cues in given contexts; represented by CitePrompt, prompting studies, and open-LLM adaptation	Context boundaries and cited evidence are usually assumed
LLM-assisted annotation	Context labels, rationales, and annotation decisions; represented by citation-context annotation, AnnoLLM, and LLMaAA	Human validation and disagreement analysis remain necessary
Long-context recognition	Broader citation contexts, implicit context, and reference scope	Long context does not guarantee boundary awareness
Retrieval-augmented grounding	Cited content and evidence spans across citing and cited documents	Evidence support and faithfulness remain immature
Multi-dimensional frameworks	Relations among intent, stance, citation content, and cited content	Needs richer annotation and consistency metrics

V. CONCLUSION

This survey has examined how LLMs contribute to citation recognition. Pre-LLM work established the main units that still organize the field: citation contexts, interpretative labels, and cited text spans. LLM-based studies inherit these units, but they also reconfigure them through prompting and fine-tuning, LLMassisted annotation, long-context use, retrieval-augmented grounding, and multi-dimensional frameworks. The main contribution of LLMs is therefore not a single new task or model family, but a set of method paradigms that make citation recognition more flexible, scalable, and potentially more connected across citing context, citation content, and cited evidence.

At the same time, this survey shows that flexibility is not the same as explicit recognition. Current LLM-based work still leaves several core problems open: identifying citation boundaries rather than relying on supplied windows, keeping citation content and interpretative dimensions consistent, grounding cited content in evidence, and making outputs robust to prompts, retrieval settings, and citation bias. Future work should therefore evaluate whether LLMs can expose the textual and evidential basis on which citation understanding depends, rather than only producing plausible labels, rationales, or generated citations.

Because LLM-based citation recognition is still emerging, relevant work is dispersed across citation annotation, citation generation, evidence grounding, and verifiability. This survey therefore focuses on studies most directly related to citation context, citation content, and cited evidence, without claiming exhaustive coverage of broader citation analysis, citation recommendation, citation networks, or bibliometric indicators.

REFERENCES

[1] Algaba, A., Mazijn, C., Holst, V., Tori, F., Wenmackers, S., Ginis, V.: Large language models reflect human citation patterns with a heightened citation bias. Findings of NAACL 2025. https://doi.org/10.18653/v1/2025.findings-naacl.381

[2] Bornmann, L., Wray, K.B., Haunschild, R.: Citation concept analysis: A new form of citation analysis revealing the usefulness of concepts for other researchers illustrated by exemplary case studies. Scientometrics 114, 1037-1052 (2018)

[3] Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., et al.: Language models are few-shot learners. In: NeurIPS 2020 (2020)

[4] Cohan, A., Ammar, W., van Zuylen, M., Cady, F.: Structural scaffolds for citation intent classification in scientific publications. In: NAACL-HLT 2019 (2019)

[5] Cohan, A., Soldaini, L., Goharian, N.: Matching citation text and cited spans in biomedical literature: A search-oriented approach. In: NAACL-HLT 2015. https://doi.org/10.3115/v1/N15-1110

[6] Duan, C., Tan, Z.: Semantically orthogonal framework for citation classification: Disentangling intent and content. arXiv:2601.05103 (2026)

[7] Funkquist, M., Kuznetsov, I., Hou, Y., Gurevych, I.: CiteBench: A benchmark for scientific citation text generation. In: EMNLP 2023. https://doi.org/10.18653/v1/2023.emnlp-main.455

[8] Gao, L., Dai, Z., Pasupat, P., et al.: RARR: Researching and revising what language models say, using language models. In: ACL 2023. https://doi.org/10.18653/v1/2023.acl-long.910

[9] Gao, T., Yen, H., Yu, J., Chen, D.: Enabling large language models to generate text with citations. In: EMNLP 2023. https://doi.org/10.18653/v1/2023.emnlp-main.398

[10] He, X., Lin, Z., Gong, Y., Jin, A.-L., Zhang, H., Lin, C., et al.: AnnoLLM: Making large language models to be better crowdsourced annotators. In: NAACL Industry 2024. https://doi.org/10.18653/v1/2024.naacl-industry.15

[11] Hernandez-Alvarez, M., Gomez, J.M.: Survey about citation context analysis: Tasks, techniques, and resources. Natural Language Engineering 22(3), 327-349 (2016)

[12] Huang, C., Wu, Z., Hu, Y., Wang, W.: Training language models to generate text with citations via fine-grained rewards. In: ACL 2024. https://doi.org/10.18653/v1/2024.acl-long.161

[13] Jha, R., Abu-Jbara, A., Qazvinian, V., Radev, D.R.: NLP-driven citation analysis for scientometrics. Natural Language Engineering 23(1), 93-130 (2017)

[14] Jurgens, D., Kumar, S., Hoover, R., McFarland, D., Jurafsky, D.: Citation classification for behavioral analysis of a scientific field. Transactions of the Association for Computational Linguistics 6, 391-407 (2018)

[15] Koloveas, P., Chatzopoulos, S., Vergoulis, T., Tryfonopoulos, C.: Can LLMs predict citation intent? An experimental analysis of in-context learning and fine-tuning on open LLMs. arXiv:2502.14561 (2025)

[16] Kunnath, S.N., Pride, D., Knoth, P.: Prompting strategies for citation classification. In: CIKM 2023 (2023)

[17] Lahiri, A., Sanyal, D.K., Mukherjee, I.: CitePrompt: Using prompts to identify citation intent in scientific papers. In: JCDL 2023 (2023)

[18] Lewis, P., Perez, E., Piktus, A., et al.: Retrieval-augmented generation for knowledge-intensive NLP tasks. In: NeurIPS 2020 (2020)

[19] Li, X., Lee, Y.-H., Ouyang, J.: Cited text spans for scientific citation text generation. In: SDP 2024. https://doi.org/10.18653/v1/2024.sdp-1.9

[20] Liu, N.F., Zhang, T., Liang, P.: Evaluating verifiability in generative search engines. In: Findings of EMNLP 2023. https://doi.org/10.18653/v1/2023.findings-emnlp.467

[21] Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., Neubig, G.: Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. ACM Computing Surveys 55(9), Article 195 (2023)

[22] Ma, X., Xu, J., Zhang, C.: Automatic identification of cited text spans: A multiclassifier approach over imbalanced dataset. Scientometrics 116, 1303-1335 (2018). https://doi.org/10.1007/s11192-018-2754-2

[23] Mercier, D., Rizvi, S.T.R., Rajashekar, V., Dengel, A., Ahmed, S.: ImpactCite: An XLNet-based method for citation impact analysis. arXiv:2005.06611 (2020)

[24] Nishikawa, K., Koshiba, H.: Exploring the applicability of large language models to citation context analysis. Scientometrics (2024). https://doi.org/10.1007/s11192-024-05142-9

[25] Oesterling, A., Ghosal, A., Yu, H., Xin, R., Baig, Y., Semenova, L., Rudin, C.: Multitask learning for citation purpose classification. In: SDP 2021 / 3C Shared Task (2021)

[26] Ouyang, L., Wu, J., Jiang, X., et al.: Training language models to follow instructions with human feedback. arXiv:2203.02155 (2022)

[27] Rosati, D.: Citations are not opinions: A corpus linguistics approach to understanding how citations are made. Quantitative Science Studies 2(2), 581-612 (2021)

[28] Wang, P., Zhou, H., Tang, J., Wang, T., Li, S.: Cited text spans identification with an improved balanced ensemble model. Scientometrics 120, 1111-1145 (2019). https://doi.org/10.1007/s11192-019-03167-z

[29] Zhang, G., Ding, Y., Milojevic, S.: Citation content analysis: A framework for syntactic and semantic analysis of citation content. Journal of the American Society for Information Science and Technology 64(7), 1490-1503 (2013)

[30] Zhang, R., Li, Y., Ma, Y., Zhou, M., Zou, L.: LLMaAA: Making large language models as active annotators. In: Findings of EMNLP 2023. https://doi.org/10.18653/v1/2023.findings-emnlp.872

[31] Zhang, Y., Wang, Y., Wang, K., Sheng, Q.Z., Yao, L., Mahmood, A., Zhang, W.E., Zhao, R.: When large language models meet citation: A survey. arXiv:2309.09727 (2023)

[32] Ziems, C., Held, W., Shaikh, O., Chen, J., Zhang, Z., Yang, D.: Can large language models transform computational social science? Computational Linguistics 50(1) (2024). https://doi.org/10.1162/coli_a_00502