Reconstructive citation context analysis using large language models

A roadmap

Arno Simons, Hiba Arnaout, and Iryna Gurevych

1. Introduction

Citations are not merely formal pointers to prior work. Within the history, philosophy, and sociology of science (HPSS), citations are examined through a reconstructive lens—that is, an effort to recover how scientific claims, alliances, and boundaries are actively built through acts of citing, rather than merely reflected in them. Patterns of citation and uptake, in turn, offer a window onto the dynamics of knowledge production and circulation, from the formation of scientific facts to the negotiation of epistemic authority and the construction of disciplinary identities (e.g., Cozzens, 1985; Gilbert, 1977; Latour, 1987; Luukkonen, 1997; Myers, 1990; Small, 1980; 2004; Swales, 1986; 1990; White; 2004).

This reconstructive approach to studying citation practice contrasts with the evaluative logic that dominates much of scientometrics, where citations are primarily treated as countable indicators of research quality or influence. Citation context analysis (CCA) emerged in the 1970s within this evaluative tradition as a means of distinguishing "countworthy" from marginal citations in order to refine citation-based measures (Chubin and Moitra, 1975; Moravcsik and Murugesan, 1975). Although HPSS scholars have since promoted CCA for reconstructive purposes, the evaluative framing has remained the dominant influence on how CCA is conceived, operationalized, and applied (cf. Bornmann and Daniel, 2008).

Building on this evaluative orientation, computational approaches to CCA have developed considerably over the past twenty-five years, evolving from rule-based grammars and keyword matching to machine learning, deep learning, and transformer-based models (for recent overviews, see Iqbal et al., 2021; Kunnath et al., 2021; Tahamtam and Bornmann 2019; Yousif et al., 2019; Zhang et al., 2023). For most of this time, the central goal has been to classify citations into predefined categories—typically functional types or sentiment polarities—to refine citation metrics or enable large-scale scientometric studies. While this classificatory focus has long depended on subtasks such as citation

span detection and reference scope resolution, only recently has the field begun to explore generative tasks such as citation text generation (Funkquist et al., 2023) or citation impact summarization (Arnaout et al., 2025).[1] Yet these developments have largely extended the evaluative logic of citation counting into more sophisticated computational forms.

Although such classificatory and generative capabilities hold clear potential for reconstructive analysis of citation practice, their uptake within this field has remained negligible. Two factors help explain this gap. First, computational CCA has remained closely tied to evaluative scientometric goals, such as improving research assessment, which do not align with HPSS's reconstructive priorities. Second, until the recent emergence of large language models (LLMs), these methods lacked interpretative depth: they could assign categories but rarely captured the rhetorical nuance, contextual variability, and intertextual relations that reconstructive inquiry emphasizes.

LLMs open new ground for connecting computational and reconstructive CCA. Their capacity to process language in context, adapt across domains, and respond flexibly to interpretive prompts makes it possible to automate aspects of analysis that once depended entirely on human close reading. This broadens computational CCA toward reconstructive goals—tracing changes in citation framing over time, comparing representations across disciplines, mapping conceptual and argumentative linkages, clustering related framings, and generating interpretative readings of how a work's meaning develops through uptake. Yet these same capacities pose risks: LLMs have limitations and they cannot and should not replace human interpretation.

This chapter offers a roadmap for integrating LLMs into reconstructive CCA in ways that align with HPSS commitments. The aim is to show how HPSS researchers can harness the new capacities of LLMs for CCA without sacrificing methodological rigor, reflexive skepticism, or sensitivity to the complex social life of citations. For a related discussion of LLMs in CCA, see Liesegang and Gläser (2026).

This chapter proceeds as follows. We begin by clarifying a reconstructive conception of CCA in HPSS and by contrasting it with evaluative scientometrics. We then develop a pluralist account of citation practices and aims, followed by an analysis of cross-contextual dynamics in which citation meanings emerge across citing, cited, and uptake contexts. Building on these foundations, we translate reconstructive aims into LLM-based workflows, moving from context detection to context classification, then to citation–concept and citation–argument mining, and to similarity modeling and clustering of citation framings. Finally, we explore interpretative readings with LLMs, including chained and time-aware designs that generate multiple plausible interpretations and make their evidentiary basis visible. A concluding discussion synthesizes methodological lessons, addresses limits and risks, and sets out practical guidance on tooling, datasets, reproducibility, and reflexive use, with an emphasis on keeping human interpretation central.

1 Prior approaches to citation-based document summarization used extractive rather than truly generative methods (e.g., Karimi et al., 2018; Mao et al., 2022).

2. Reconstructive CCA

At the heart of the HPSS tradition lies a commitment to the *reconstruction* of scientific practice—an effort to understand how science is done in practice, how knowledge is shaped, and how epistemic authority is negotiated (e.g., Pickering, 1992). This reconstructive orientation stands in marked contrast to the *evaluative* perspective that underpins much of conventional citation analysis, where citations are primarily treated as proxies for research quality, influence, or impact (Cole and Cole, 1971; Kaplan, 1965). While the evaluative tradition promotes CCA primarily as a means of distinguishing "count-worthy" from marginal references—and, by extension, high-quality from lower-quality papers (e.g., Donner et al., 2025)—the reconstructive perspective views citations as socially situated acts through which scientists shape arguments, establish epistemic positioning, build alliances, and demarcate boundaries (e.g., Cozzens, 1985; Gilbert, 1977; Latour, 1987; Luukkonen, 1997; Myers, 1990; Small, 1980; 2004; Swales, 1986; 1990; White; 2004). Studying these practices offers an entry point into the meaning-making processes of science, revealing how claims are justified, allegiances formed, and disputes settled.

Interpretation is therefore central to this reconstructive project. Understanding citation practices requires attention to nuance, ambiguity, and the contingencies of scholarly communication. Predefined taxonomies and the surface-level cues on which their annotation and classification often depend are valuable starting points for analyzing citation contexts, yet they capture only part of what citations do. Because such approaches typically overlook the broader discursive and historical context—both within the citing text and across the network of citing and cited works—they struggle with cases where meaning is implied rather than stated, where hedging conceals stance, or where subtle rhetorical nuances alter interpretation entirely (cf. Gilbert, 1977; Lyu et al., 2021; Kunnath et al., 2021). As Cronin (1998:48) observed, "a full understanding of why A cites B requires a multi-layered explanation and, ideally, thick description of the process—and the politics of the process". Addressing this complexity requires methods attentive both to the wider contexts in which citations operate and to the implicit cues through which meaning is negotiated.

Traditionally, such methods were qualitative and included close reading, discourse analysis, and historically informed contextualization (Gilbert, 1977; Latour, 1987; Myers, 1990; Swales; 1986; 1990; White; 2004). They remain indispensable for reconstructive CCA, since they enable analysts to grasp the subtleties of argumentation, disciplinary conventions, and historical positioning that shape how citations function. The recent emergence of LLMs introduces new opportunities in this regard. While the aim is rarely to replace qualitative analysis wholesale, in some contexts LLMs may automate portions of the interpretive process—identifying recurrent rhetorical patterns, flagging unusual framings, or generating candidate interpretations for further scrutiny. In many other cases, their value lies in supporting qualitative inquiry within mixed-method designs, where computational breadth and human interpretive depth can be combined (see Hill, 2026; Scharnhorst, 2026; Simons et al., 2026). Used in this way, LLMs can help preserve the contextual richness that reconstructive CCA demands, while extending its reach to larger and more diverse corpora.

3. A pluralist understanding of citation practices and their study

A reconstructive approach to CCA recognizes that citations rarely serve a single, uniform purpose. Decades of empirical research have shown—contrary to the sharp contrasts drawn in earlier debates over normative versus persuasive theories of citation (e.g., Cronin, 1981; Gilbert, 1977; Kaplan, 1965; Latour, 1987; Luukkonen, 1997; Merton, 1973; Small, 2008)—that citations are multifunctional acts, carrying overlapping epistemic, functional, social, rhetorical, and even emotional values (for overviews, see Bornmann and Daniel, 2008; Lyu et al., 2021; Tahamtan and Bornmann, 2018; 2019). A single citation context can do more than one thing at once, sometimes acknowledging intellectual debt while also bolstering a claim's credibility, signaling alignment with a tradition, or positioning one line of work as more authoritative than alternatives. Norm-following and strategic persuasion often reinforce one another, since citing "properly" can both satisfy expectations of credit and simultaneously frame what the cited work is taken to mean, thereby strengthening a claim and securing position within a community (see Small, 2008: 76). More overtly selective moves such as emphasis, reinterpretation, or omission are part of the same adaptable repertoire, whose specific configuration depends on the discipline, genre, audience, and rhetorical context in which they occur.

Variation also arises when reference practices cross the boundaries of scientific discourse. The reconstructive questions that motivate CCA have also been pursued in other domains, sometimes explicitly under the banner of citation analysis and sometimes through adjacent traditions such as intertextuality studies and discourse analysis. Examples include public policy (Simons, 2016, 2015; Steiner-Khamsi, 2022), law (Bodström, 2023; Feneberg et al., 2022), journalism (e.g., Fleerackers et al., 2021; Simons and Schniedermann, 2023), and social media (Barber, 2025), where conventions of referencing and communicative aims differ markedly from those in science. These studies also show that scientific writing is likewise embedded in these wider networks: research articles cite legal decisions, government reports, news stories, and other non-scientific sources, while science itself is routinely cited in policy documents, advocacy materials, and journalism. Such cross-domain exchanges can reshape the meaning of a citation, as works migrate between communicative contexts with distinct rhetorical norms, audiences, and institutional logics.

Pluralism applies not only to citation practices but also to the aims that guide CCA in HPSS. Analysts may be concerned with mapping functional roles, assessing evaluative sentiment, identifying framings of uncertainty, tracing the uptake of particular concepts, or reconstructing the role of citations in relation to specific methods. The same citation context may yield different insights depending on which of these dimensions is under scrutiny, and some may remain latent or ambiguous unless read in light of a specific research question. Because reconstructive aims vary, coding schemes and computational frameworks should be treated as methodological choices rather than fixed standards—adapted, combined, or newly created as the interpretive task demands. At the same time, computational approaches should remain open to plurality, allowing more than one plausible reading to coexist and serving as instruments for exploring interpretive variation rather than enforcing categorical closure.

4. Cross-contextual dynamics of citation meaning

Citation meaning is not fixed within the boundaries of a single sentence or paper; it is emergent across multiple contexts. Interpretation is therefore not only contextual but inherently inter-contextual. To understand what a citation does, we must attend not only to its immediate rhetorical setting but also to the broader discursive web in which it is embedded. At the most basic level, a citation involves two distinct textual contexts: the passage in the citing document that frames the citation, and the cited source itself. These two contexts are rarely neutral mirrors of each other. A citing document may emphasize or downplay particular aspects of the cited work, point to central claims or peripheral remarks, or even misrepresent the original content—sometimes intentionally, sometimes not (Latour, 1987; Small, 2004). In some cases, citations may point to fictional, inaccessible, or misattributed sources.

Beyond this dyadic relationship, citation meanings and the claims they support evolve through their uptake in subsequent literature (Latour, 1987). Later authors often orient their own framings of a cited work to how that work has already been cited in earlier papers. Citation functions and framings shift over time as debates evolve, norms change, and epistemic communities reorganize. Such variation is both historical and concurrent: different communities may interpret the same work in distinct ways even within the same period, while framings also change across time as ideas circulate and are recontextualized. The same work may be cited as dissenting in one community and supportive in another, and over time a contribution that was once peripheral may come to occupy a canonical place in the literature. Patterns of uptake across fields can lead to divergent, even contradictory, interpretations of the same work. Over time, certain framings may stabilize or become canonical (Cozzens, 1985; Gilbert, 1977; Small, 2004). In such cases, we can observe the emergence of typical citation contexts, recurring rhetorical patterns or interpretive framings that circulate through a community and help define its shared assumptions, allowing the cited work to function as what Small (1978) calls a "concept symbol".

These shifts in citation meaning can be understood along two intersecting dimensions proposed by Small (2004): *literalness*, the extent to which a citation accurately reflects the content or intent of the cited work, and *consensus*, the degree of alignment among citing authors about that work's significance. Together, they describe how meanings move between fidelity and reinterpretation, stability and variation. The same source may be read and positioned in markedly different ways across disciplinary boundaries or over time, reflecting divergent epistemic agendas and interpretive norms.

Inter-contextuality also extends beyond individual citations to the relational context in which works are cited together. Co-citation analysis (Small, 1973) has long been used to detect attributed alignments or conceptual proximity within the citing literature, especially at aggregated levels. But as Small (1980) argued early on, co-citation patterns could be significantly enriched by analyzing the contexts in which co-cited documents appear. This would allow for a more textually rich and discursively grounded form of "co-citation context analysis", something well suited to computational approaches using LLMs.

Taken together, these forms of inter-contextuality call for methodological approaches that move beyond sentence- or document-level analysis and take seriously

the distributed, evolving, and socially constructed nature of citation meaning. They resonate with semiotic traditions in discourse analysis (Foucault 1970; 1982; Keller 2024), actor-network theory (Callon et al., 1986a; Latour 1987), and theories of intertextuality (Kristeva 1980; Allen 2011), all of which emphasize that meaning emerges relationally, through positioning, uptake, and circulation across networks of discourse and practice.

5. Translating reconstructive CCA into LLM-based workflows

To translate the interpretive goals of reconstructive CCA into computational practice, this section focuses on a set of tasks that are both relevant for reconstructive inquiry and feasible within current NLP capabilities. For each task, we first outline its potential use cases for HPSS—what interpretive questions it might help to pursue—before turning to how the task has been formalized in computational CCA and how such methods might be adapted, extended, or critically reassessed for reconstructive purposes.

In computational CCA, and natural language processing (NLP) more generally, a "task" is a specific, well-defined problem of textual analysis or generation that an algorithm is designed to solve. In the context of CCA, this might include identifying the span of text in which a citation occurs, linking a citation to a particular contribution in the cited work, or generating a summary of how a citation has been taken up in the citing literature over time. Such tasks are formalized to enable large-scale automation, traditionally through supervised or unsupervised learning, and more recently through semi-supervised or prompt-based approaches. While these formalizations open important opportunities, they also shape, and sometimes limit, the ways in which reconstructive aims can be pursued computationally.

The recent rise of transformer-based language models (Vaswani et al., 2017) has transformed how NLP tasks are approached. Within this architecture, encoder models such as BERT (Devlin et al., 2018) read words in relation to both what comes before and after them. This bidirectional analysis makes them especially effective for structural tasks, such as identifying the function or tone of a citation in its immediate context. Decoder models such as GPT (Radford et al., 2018), by contrast, generate text step by step, predicting what comes next based on the preceding sequence. Although primarily generative, these models can also perform analytic tasks like classification or extraction when fine-tuned or guided by carefully designed prompts. More recent systems, such as OpenAI's GPT-5.2, Google's Gemini 3, Anthropic's Claude 4.5, or Meta's LLaMA 4, build on this transformer foundation, combining scale and adaptability with user-friendly chatbot interfaces and multimodal capabilities (text, code, images, and beyond). These advances support applications ranging from structural analysis to open-ended generation with minimal additional training. For reconstructive CCA, this flexibility opens possibilities not only for automating classification but also for pursuing more exploratory and interpretive forms of analysis (see also Simons et al., 2026).

5.1 Citation context detection

Identifying the citation context is more than a preprocessing step. It defines the discursive space in which a citation acquires meaning, where an author positions a source, frames a problem, or signals alignment or distance. Reliable context detection is therefore essential for interpretive analysis: it determines what parts of a text become available for reading, comparison, and interpretation. At the same time, deciding where a citation's context begins and ends is itself an interpretive act, shaped by assumptions about what counts as relevant textual evidence (see also Liesegang and Gläser, 2026). For HPSS, the challenge is not simply to locate citation markers but to capture the textual boundaries within which a citation performs rhetorical or epistemic work.

In computational terms, citation context detection is often reduced to span extraction, that is, identifying a window of text around a citation marker like "Smith et al., 2012", typically the sentence in which it appears. While convenient for large-scale evaluative analysis, this approach risks missing the discursive work that citations perform across broader stretches of text. A meaningful citation context may begin earlier, for example when a problem is first framed or a method introduced, and it may extend across several sentences or even sections. From an interpretive standpoint, the relevant context is defined not by proximity but by semantic relevance: those portions of text that articulate the author's stance toward the cited work, the role it plays in the argument, and the assumptions it mobilizes. Capturing this requires more flexible models of rhetorical segmentation.

Operationally, two intertwined components are involved. First, all citation mentions must be detected (e.g., each occurrence of "Smith et al., 2012"). This is relatively tractable in contemporary scientific writing that follows standardized formats and can be addressed using LLMs (Sarin and Alperin, 2025) or lighter parsers such as Grobid[2]. Greater challenges arise in older or less structured texts, especially when citations appear in footnotes (see Boulanger, 2026). In such cases, markers may be ambiguous ("Ibid.", "op. cit.", "supra note 4") and require cross-footnote tracking or layout-aware models. Here, multi-modal approaches combining linguistic and visual features have proved promising (Khalighinejad et al., 2024; Mishra et al., 2024; Yu et al., 2020).

Second, the model must identify all textual spans that refer to the cited work, including indirect or anaphoric ones (e.g., "this study," "they") and multi-source cases where only part of a sentence pertains to a given citation ("Several studies have explored this issue …"). These challenges are addressed under coreference and anaphora resolution (Jha et al., 2017; Lauscher et al., 2021; Luan et al., 2018) and reference-scope resolution (Abu-Jbara and Radev, 2012; Jha et al., 2017), which together delineate the dynamic boundaries of citation contexts (Kunnath et al., 2021; 2022; Yousif et al., 2019). Khan et al. (2025) recently proposed a refined approach targeting what they call the "reference text": the minimal fragment that explicitly describes the cited work.

Recent advances in generative frontier models have begun to integrate tasks that earlier approaches treated separately. Large, instruction-tuned LLMs can increasingly

2 https://github.com/kermittz/grobid

detect citation contexts holistically, identifying relevant spans that extend beyond sentence boundaries and resolving anaphoric or coreferential links in a single pass. When combined with retrieval-augmented or document-grounded prompting (see below), such models can approximate the interpretive judgment needed to delineate where a citation's meaningful context begins and ends, bringing computational methods closer to the reconstructive aims of HPSS.

In sum, detecting citation context is not merely a matter of locating where a citation occurs but of reconstructing the discursive space in which it acquires meaning. The boundaries chosen at this stage shape what can later be classified, compared, or re-read: if too narrow, the citation may appear perfunctory or detached; if too broad, its rhetorical function may blur depending on the classifier used. Context detection is therefore a hermeneutic act, one that decides which traces of discourse count as evidence for interpretation. For HPSS researchers, it is vital to remain aware of how computational methods impose such boundaries, since every segmentation encodes assumptions about where meaning begins and ends. As we turn to the next task, context classification, these issues become even more pronounced, because any attempt to interpret a citation's function or stance depends on first defining what counts as its meaningful textual frame.

5.2 Citation context classification

For reconstructive CCA, citation classification provides a means of making interpretive judgments—such as assessments of citation intent—explicit, inspectable, and systematically comparable. By assigning structured labels to how citations function, researchers can make visible patterns in how knowledge is framed, circulated, and contested across texts. Within HPSS, such classification can serve several interpretive aims, e.g., to compare rhetorical functions across disciplines, to map shifts in evaluative stance over time, or to identify cases that invite deeper qualitative reading. Used reflexively, classification becomes a way of examining how different rhetorical or evaluative functions of citation are patterned within and across scholarly communities, while keeping in view that these functions are themselves interpretive constructions.

In computational CCA, classification is the central analytic task. Technically, it means that a piece of text—the citation context—is given to a model, which then assigns it to one or more predefined categories. Put simply: context in, label(s) out. Early efforts centered on the citer's intent, asking whether a reference was substantive or perfunctory, supportive or critical (Chubin and Moitra, 1975; Moravcsik and Murugesan, 1975; for an overview and recent synthesis, see Lyu et al., 2021). Over time, this classificatory agenda expanded to encompass other dimensions (for an overview, see Kunnath et al., 2021), including sentiment toward the cited work (Yousif et al., 2019b), hedging (Di Marco et al., 2006; Mercer et al., 2004), and perceived importance (Pride and Knoth, 2017; Valenzuela et al., 2015). These are interpretive questions, deeply embedded in disciplinary norms, genre conventions, and authorial intent. Computational models must translate them into predefined categories to operate at scale, but such formalization can only approximate the variability that gives citation meaning its depth. Inevitably, there are trade-offs between the consistency gained through standardization and the nuance that interpretive reading demands.

For HPSS researchers, the first question is whether classification is the right approach at all, and if so, which aspects of citation practice merit operationalization. Is the goal to distinguish rhetorical functions, detect evaluative stances, track framings of uncertainty, or gauge perceived importance? Each calls for different categories and annotation schemes. Another consideration is granularity: some schemes use broad categories such as *support* or *contrast*, which scale well but risk flattening nuance; others use finer distinctions, such as *methodological use*, *background* or *perfunctory*, that preserve interpretive depth but are harder to apply consistently. Scholars must also decide whether classes are mutually exclusive, hierarchical, or overlapping, since a single citation may perform multiple functions. The MultiCite dataset (Lauscher et al., 2021), for example, allows multi-label annotation to reflect this complexity.

Reflecting on such design choices is crucial because all classifiers depend on training and evaluation datasets whose annotation guidelines codify particular interpretive assumptions about what citations mean and how those meanings can be operationalized. Popular datasets illustrate this clearly. ACL-ARC (Jurgens et al., 2018) encouraged document-level interpretation, inviting domain expert annotators to consider broader textual evidence beyond the citing sentence. SciCite (Cohan et al., 2019) uses minimally trained crowdworkers with few constraints on inference, making labels dependent on broader interpretive judgments. ACT (Pride and Knoth, 2020) asked cited authors to label their own citations, often relying on memory of their motivations rather than on the explicit citation text. MultiCite (Lauscher et al., 2021) builds on the ACL-ARC scheme but extends it to multi-sentence and multi-label contexts: annotators identify both explicit and implicit references, select evidentiary spans, and assign one or more intent labels under refined guidelines and adjudication. This balances interpretive inference with explicit evidence requirements and quality control.

Across these datasets, annotators are encouraged, or at least permitted, to make inferences beyond the literal citation wording, meaning that classifiers trained on them inevitably embed interpretive assumptions. A notable exception is Teufel et al. (2006a;b), whose twelve-category scheme was explicitly designed to constrain annotation to surface cues in the citing text itself, avoiding speculation about authorial intent or unstated context. Their framework shows how annotation guidelines can operationalize limits on interpretive scope. Taken together, these examples show how annotation design not only shapes dataset quality but sets the interpretive boundaries of computational CCA, decisions that, for HPSS, are methodological choices about what forms of meaning are rendered legible to machines.

Different technical strategies exist for carrying out classification. Researchers can use off-the-shelf models trained on such datasets or train their own: either encoder-based classifiers such as BERT variants (Beltagy et al., 2019; Lo et al., 2020), or decoder-based classifiers prompted in zero- or few-shot setups (Arnaout et al., 2025; Kunnath et al., 2023; Simons, 2026). The choice depends on resources and the desired balance between interpretive control and scalability. Encoder-based models offer transparency and efficiency but require annotated data and technical expertise. LLM-based approaches reduce those requirements and can flexibly handle multi-label or open-ended schemes, but at the cost of transparency and reproducibility. Hybrid workflows are increasingly viable: LLMs can generate candidate classifications that human experts then audit, correct, or

enrich, balancing scalability with interpretive rigor (Bonet-Jover et al., 2023; Dagdalen et al., 2024; Heseltine and Clemm von Hohenberg, 2024; Törnberg, 2024).

In sum, classification remains a useful but limited tool. Its value for HPSS lies not mainly in replicating scientometric metrics but in enabling structured comparisons of how citations function across texts and contexts, in navigating corpora to identify instructive or atypical cases, or in serving as a starting point for deeper interpretative investigation. In all cases, researchers must treat classification as a methodological choice rather than a default and remain reflexively aware of how their categories, tools, and training data shape the interpretations they produce.

5.3 Citation-entity and -argument mining

For reconstructive CCA, the central question is not only what a citation does rhetorically, but how it mobilizes specific intellectual resources, including ideas, concepts, methods, instruments, data, authorities or arguments, in the fine-grained construction of knowledge. What we might call citation-entity and -argument mining is not yet an established task in computational CCA, but it points toward one: detecting the entities and arguments to which a citation is linked. From an HPSS perspective, this matters because it would make citations legible as connective nodes that bind such elements into local chains of reasoning and persuasion (cf. Callon et al., 1983; 1986a; Gilbert, 1977; Latour, 1987).

Related work in NLP has addressed adjacent problems through entity recognition and relation extraction. Entity recognition identifies relevant spans of text and tags them with semantic categories, for example, marking "citation function classification" as a method and "ACL-ARC" as a dataset. Relation extraction then specifies how these entities are connected, for instance, that ACL-ARC is used to evaluate citation function classification systems. While most existing work on scientific entity–relation extraction (Dagdelen et al., 2024; Dunn et al., 2022; Zhang et al., 2024) and argument mining (Accuosto et al., 2021; Akkasi and Moens, 2021; Binder et al., 2022; Dagdelen et al., 2024; Fergadis et al., 2021; Gorur et al., 2024) does not treat references as first-class analytical units, these pipelines can be adapted for citation-context analysis, either by adding a citation-mention entity type to the prediction space or by incorporating a separate citation-mention detector as an additional stage in the extraction pipeline (see 5.1) and extracting relations that connect a cited work to the entities or arguments it is invoked to support. For example, in a phrase like "ACL-ARC (Jurgens et al., 2018)", the citation mention can be tagged and linked to "ACL-ARC" as the dataset introduced or described in the cited work.

As with classification, researchers could pursue citation–entity or citation–argument mining tasks in several ways. Encoder-based models (e.g., BERT or SciBERT variants) remain efficient and transparent, suitable for targeted extraction tasks where annotation schemes and relation types are well defined (Binder et al., 2022; Fergadis et al., 2021; Zhang et al., 2024). Decoder-based LLMs, by contrast, allow more flexible, exploratory setups: prompted in zero- or few-shot modes, they can identify entities, infer relations, and even propose higher-order interpretive links without additional training, though at the cost of reproducibility and control (Gorur et al., 2024; Zhang et al., 2024). A third strategy, demonstrated by Dunn et al. (2022) and Dagdelen et al. (2024), is to

fine-tune high-capacity decoder-based models such as GPT-3, T5, or LLaMA-2 for joint entity and relation extraction. These systems can achieve clear gains over both encoder-based baselines and zero- or few-shot prompting, yet their computational costs likely remain prohibitive for most HPSS applications.

When expressed as triplets ($entity_1$, relation, $entity_2$), the mined entity relations around citations could be aggregated into knowledge graphs that allow historical analyses (cf. Boulanger, 2026; Schlattmann et al., 2026), and that connect reconstructive CCA to earlier co-word (Callon et al., 1983; 1986b) and co-citation (Small, 1973; 1980) traditions that trace how meanings crystallize through recurrent associations (see also Chavalarias and Cointet, 2013; Roth and Cointet, 2010). Because citations accumulate and transform through time, every link in a knowledge graph should be timestamped, allowing the resulting network to be explored as an evolving structure rather than a static map. Temporal or longitudinal slices of these graphs could then reveal how the conceptual associations around a cited work expand, contract, or reconfigure as that work is cited across disciplines and decades. For example, a paper first cited for the introduction of a new instrument may later be reframed as a theoretical foundation, as well as reappropriated by different research community along the way. Tracking these shifts through dynamic citation–entity or citation–argument networks would enable the reconstruction of the trajectories through which meanings are stabilized, reinterpreted, or forgotten, including when a cited work's recurrent linkages congeal into a relatively standardized "concept symbol" that circulates across communities.

5.4 Citation context similarity and clustering

While citation–concept and –argument mining focuses on tracing explicit argumentative and conceptual linkages, similarity modeling and clustering offer a complementary means of examining how citation meanings converge and diverge across texts. In doing so, they operationalize the cross-contextual dynamics discussed earlier and offer scalable ways to approximate variation along Small's (2004) dimensions of literalness and consensus: the degree to which citations adhere to or reinterpret the cited work, and the extent to which authors converge on shared framings of it.

Embedding-based similarity modeling provides the technical foundation for this approach. By representing citation contexts as high-dimensional vectors whose relative distances approximate semantic relatedness, such methods make it possible to detect patterns of affinity and contrast among citations. Clustering these representations can reveal families of contexts that share rhetorical tone, evaluative stance, or conceptual alignment, as well as variations that signal disagreement, reinterpretation, or drift in meaning over time.

Technically, these approaches rely on language embeddings, i.e. dense vector representations that capture lexical, syntactic, and semantic patterns. Early distributional models such as Word2Vec (Mikolov et al., 2013) assigned each word a single static vector based on co-occurrence statistics, while transformer-based models produce contextualized word embeddings that vary with surrounding text. Aggregating these into sentence- or paragraph-level representations, for example with Sentence Transformers (Reimers

and Gurevych, 2019), enables systematic comparison among citation contexts at multiple scales.

Building on such representations, integrated frameworks like BERTopic (Grootendorst, 2022) combine embedding-based dimensionality reduction, clustering, and keyword extraction to identify coherent clusters of semantically related texts. For reconstructive CCA, these clusters can be interpreted not as objective categories but as families of citation framings, recurring rhetorical or conceptual patterns that indicate degrees of convergence or divergence in how authors mobilize a cited work. Clusters showing consistent evaluative or rhetorical framings may correspond to stabilized interpretive conventions, whereas fragmented or shifting clusters reveal diversity and contestation within the citing community.

Within embedding- and LLM-based approaches, empirical work directly addressing Small's dimension of literalness remains limited. Roy and Mercer (2022) offer one of the few attempts to approximate it computationally, aligning each citing sentence with the passage in the cited paper most likely being referenced. Using standard text-embedding similarity, their system searches the cited text for the segment whose representation lies nearest in semantic space. The approach demonstrates how literal alignment might be modeled, but it also assumes that a suitable target passage always exists and that the closest neighbor corresponds to a genuine interpretive link. In practice, this enforces a narrow, high-literalness reading: close proximity can signal adherence to the cited work, yet the method cannot capture diffuse, cross-sentence, or recontextualized framings. From an HPSS perspective, low similarity as well as the absence of a plausible match, may be equally revealing, pointing to interpretive distance, selective citation, or rhetorical repurposing rather than simple misalignment.

As far as we know, Small's second dimension, consensus, has never been explored explicitly in LLM- or embedding-based studies, though clustering methods suggest how it might be approached. Building on the embedding frameworks discussed above, such as BERTopic, researchers can group citation contexts referring to the same work and compare how that work is described across texts, domains, or time periods. Stable clusters with consistent evaluative or rhetorical framings indicate high consensus and shared interpretive conventions, while fragmented or evolving clusters point to pluralism, disagreement, or shifting interpretations. Because these models can also trace topic evolution, they make it possible to reconstruct how consensus around a cited source consolidates, transforms, or dissolves as debates unfold and as the work circulates across disciplinary and institutional boundaries.

A related but distinct line of work examines semantic distance not to gauge consensus or literalness, but to identify conceptual novelty. Shibayama et al. (2021) measure the semantic distance between co-cited references within a paper to assess how conceptually unusual a given combination is. Greater distance signals higher novelty, reflecting the recombination of ideas drawn from distinct domains. Although developed for evaluative purposes, such measures also resonate with Small's framework: large distances between cited works may correspond to low consensus, or to moments where authors purposefully juxtapose disparate sources to create new intellectual linkages. Repurposed for reconstructive inquiry, this approach could help trace how citation practices bridge or

redraw boundaries between fields and how interpretive novelty—rather than evaluative impact—emerges through recombination across time and domains.

Stepping back from these applications, a broader methodological question emerges: What kind of similarity do embedding models actually capture? Their training objectives and corpora shape which relations become salient. Sentence embedders like SPECTER (Cohan et al., 2020) learn similarity from citation links, embedding the assumption that citing papers are conceptually related, even though citations can also mark disagreement or mere acknowledgment. General-purpose models trained on paraphrases instead highlight rhetorical or argumentative affinities. Choosing an embedder is thus not a neutral technical step but a methodological decision that defines what kinds of relations a similarity or clustering analysis can reveal (see Simons et al., 2026).

5.5 Interpretative reading of citation context

While the preceding approaches structure the terrain for interpretation by detecting, classifying, or clustering citation contexts, they stop short of performing interpretation itself. A complementary direction is to make interpretation the object of computation: to use LLMs for generating and interrogating open-ended readings of citation contexts that emulate human interpretive judgment. LLMs can elicit plausible, synthesizing, or diversifying readings, e.g. contextually grounded hypotheses about what a citation is doing or how citation framings evolve across the citing literature. Such uses of LLMs shift computational analysis toward exploratory interpretation, foregrounding plurality and dialogic engagement rather than closure. They also make this form of interpretative reading tractable at scale by generating, comparing, and refining candidate interpretations through interaction with both texts and researchers.

Simons (2026) explores this possibility through an experiment centred on a citation context that served as a touchstone for interpretative debate early on in the development of CCA: footnote 6 in Chubin and Moitra's (1975) foundational paper. Building on Gilbert's (1977) discussion of the same footnote, Simons treats it as a compact example of interpretive instability, one that can be coded as routine on a surface reading but can also support a more pointed rhetorical interpretation on closer inspection (in Gilbert's reading, an implicit priority or credit-related admonishment). His study uses a two-stage prompt-chaining design that retraces this shift in brief: stage one elicits a surface-level classification and sets expectations about the cited sources, while stage two probes that reading by searching for evaluative cues, checking expectations against the relevant texts, and proposing alternative interpretations. By varying prompt complexity and the examples provided, he shows that GPT-5 can generate multiple plausible rereadings, with different settings biasing the model toward certain recurring interpretive themes, without fully determining the readings. More broadly, the experiment underscores that interpretive work, whether human or computational, remains perspectival and highly sensitive to how questions are framed and contexts are read.

A different and temporally-oriented interpretative approach is presented by Arnaout et al. (2025), who use LLMs to reconstruct how a paper's perceived contribution evolves through its citations over time. Their method, developed for "impact summarization", combines fine-grained citation intent classification with time-aware generation, pro-

ducing short narratives that trace how a work's influence shifts from confirmation to critique and refinement. While framed in scientometric terms, this approach effectively performs an interpretative reading across time: it identifies recurrent framings, detects changes in evaluative stance, and reconstructs the temporal trajectory of a paper's uptake. From an HPSS perspective, such diachronic summaries invite reinterpretation not as metrics of impact but as condensed histories of meaning-making within a citation network.

Together these studies show the opportunities that LLMs open for interpretative work with citation contexts. Simons' chained prompting exemplifies multi-reading generation and adversarial re-reading: using models to propose and contest alternative interpretations, and to expose how meaning depends on framing. Arnaout et al. extend this logic in time through temporal re-reading, reconstructing how interpretative framings of a work shift and stabilize across decades. In Small's (2004) terms, each study engages one dimension of cross-textual meaning. Simons' design triangulates the citation context with both the cited and the surrounding citing text, using expectation checks and cue analysis to probe alternative readings. Arnaout et al. trace patterns of consensus across many citing authors and over time. Integrating these dimensions through fuller cross-text triangulation would allow analysts to examine how citation meanings evolve both in relation to their sources and within their collective uptake.

Looking ahead, retrieval-augmented generation (RAG) offers a way to make interpretative workflows more interactive and evidentially grounded. In RAG systems, an external retrieval step draws on a user-supplied corpus, such as the citing and cited papers, to inform the model's responses. Simons (2026) applied this principle through OpenAI's API, uploading both documents so that GPT-5 could ground its readings in retrieved passages, though his approach was not interactive. By contrast, RAG-based interfaces can support live, conversational engagement, allowing HPSS scholars to query how a work is represented across contexts and examine the underlying evidence together with the model. As discussed by Hill (2026) and Scharnhorst et al. (2026), such designs let researchers "chat" with their sources, supporting interpretive inquiry through interactive, evidence-based dialogue. The gain in interpretative depth and reflexivity is greatest when such systems are used interactively, though this inevitably limits scalability, since interpretive dialogue proceeds at the pace of human reflection rather than machine throughput.

6. Discussion

In this chapter, we proposed using LLMs within citation context analysis (CCA) to recover how scientific claims, alliances, and boundaries are actively built through acts of citing rather than merely reflected in them. We argued that interpretation is central to this reconstructive project, along with a pluralistic, inter-contextual, and evolutionary understanding of citation practices. To translate these goals into computational applications, we sketched a roadmap that uses LLMs to delineate and classify citation contexts, and to trace practices and uptake across fields and over time by both mapping links between citations and the concepts or arguments they mobilize and modeling similarity among

citation framings. We also proposed generating scalable, plural, and evidence-grounded interpretative readings of what citations are doing and how those meanings evolve over time.

A key question behind our proposal is whether LLMs can genuinely assist, rather than compromise, the depth of interpretation that reconstructive inquiry requires. This raises a more fundamental issue about the role interpretation should play in reconstructive CCA, and we take the engagement with LLMs as an opportunity to revisit and refine that role. Long before LLMs, White (2004:103) argued that insofar CCA requires "the recovery of implicit meaning", it may "[…]not be delegated to a computer even in principle". Precisely because he foregrounded interpretation, his pragmatic response, shared to a degree by Teufel and colleagues (2006a;b; 2009), was to keep computational schemes anchored in surface, syntactic cues and to reserve interpretive work for human analysts. By contrast, evaluative CCA has often treated classification as readily automatable, assuming that computers can assign meanings without the hermeneutic labor that reconstructive approaches insist upon.

Early debates about the validity of citation interpretation often turned on how much expertise and situated knowledge CCA requires (Cronin, 1981; Edge, 1979; Gilbert, 1977; Peritz, 1983; Swales, 1986; Teufel et al., 2009; White, 2004). Moravcsik and Murugesan (1975:87), for example, criticized sociological readings of physics papers, arguing that sociologists "are not equipped to understand the technical scientific content of the papers they handle". Gilbert (1977:120) similarly argued that an analyst unfamiliar with the Gilbert and Woolgar paper cited in Chubin and Moitra's (1975) "Footnote 6" (see the section on interpretative readings above) was "unlikely […to] have realized that the reference might possibly have been negational". In response, many CCA scholars have advocated interviewing scientists about their citation motivations (Brooks, 1985; Cano, 1989; Cronin, 1981; Vinkler, 1987), a strategy later adopted in computational CCA to create expert-annotated citation-context datasets for training models (see Bornmann and Daniel, 2008; Tahamtan and Bornmann, 2019). Yet citer-motivation surveys introduce their own interpretive risks, including socially desirable responding on tactical motives and the general limitations of retrospective self-report (Lyu et al., 2021). Early on, Mulkay (1974) proposed a middle path: sociologists of science should collaborate with scientists in reconstructive analyses of practice, but without deference. Analysts must be prepared to recognize, challenge, and triangulate the epistemic and institutional biases of their collaborators; otherwise, they risk reproducing participants' retrospective justifications, professional myths, or strategically curated self-descriptions as sociological insight.

Returning to the question posed at the start of this discussion about whether LLMs can assist reconstructive work, the preceding history shows that interpretation has long held a contested, conditional status in CCA, and LLMs add a new layer to that debate. Used carefully, LLMs can widen coverage, bring much greater contextual nuance than earlier computational methods, and even help surface plural readings grounded in textual evidence. White's skepticism about the possibility of automated CCA beyond surface-level syntactic cues has lost its grip since LLMs have shown that the recovery of contextual and implicit meaning—long treated as the exclusive domain of human interpretation—has become, if not unproblematic, then at least computationally plausible. At the same time, LLMs also address the expertise challenge in new ways by approximating field

familiarity through targeted retrieval and modest domain tuning, surfacing terminology, canonical debates, and relevant passages that non-specialists might miss.

The value of using LLMs for reconstructive CCA depends on balancing scalability with interpretive control. That balance means exploiting models for speed and scope without turning interpretation into a black box. Scalability fails when models classify, extract, or "interpret" without exposed evidence, rationale, or reproducible settings. Interpretive control can be gained at three points in the pipeline. At the input stage, tasks should be specified clearly; choices about the scope and length of the citation context can prevent classifications from resting on overly narrow evidence; and prompt design should be tested for nudging effects. At the throughput stage, models should be selected deliberately and matched to the task; when suitable options are lacking, researchers should consider developing small domain datasets for targeted tuning. At the output stage, models can be adjusted to cite the exact passages that support a classification or reading, present plausible alternatives where appropriate, and route difficult cases to human review. Achieving this balance depends on basic LLM literacy, by which we mean enough grasp of model types, prompting, retrieval, and training data to keep outputs grounded, reproducible, and open to challenge (see Simons et al., 2026, as well as the introduction to this volume, Simons et al., 2026).

Ultimately, LLMs cannot and should not replace human interpretation. Reconstructive CCA recognizes that there is no single true reading of a citation or its evolution across contexts. Objectivity can guide inquiry as a regulative ideal, yet it is never fully attainable. This calls for methodological humility, vigilance against overreach, and a commitment to justifying interpretations and making them open to debate. The burden for interpretive judgment remains with researchers. Models can pattern-match and propose, but they cannot bear responsibility.[3]

References

Abu-Jbara A and Radev D (2012) Reference scope identification in citing sentences. In: Proceedings of the 2012 conference of the North American Chapter of the Association for Computational Linguistics: Human language technologies, 2012, pp. 80–90.

Accuosto P, Neves M, Saggion H, et al. (2021) Argumentation mining in scientific literature: from computational linguistics to biomedicine. In: BIR 2021: 11th International Workshop on Bibliometric-enhanced Information Retrieval (eds P Mayr, G Cabanac, and S Verberne), Aachen, 2021, pp. 20–36. CEUR. Available at: http://hdl.handle.net/10230/47600.

Akkasi A and Moens M-F (2021) Causal relationship extraction from biomedical text using deep neural models: A comprehensive survey. Journal of Biomedical Informatics 119: 103820.

3 This chapter was written with support from large language models (LLMs). All model-generated text was reviewed and, where necessary, rewritten by the authors, who remain fully responsible for the final version. For details on the use of LLMs in this volume, see the statement in the volume's introduction.

Allen G (2011) Intertextuality. Second edition. New York: Routledge.
Arnaout H, Sternlicht N, Hope T, et al. (2025) In-depth Research Impact Summarization through Fine-Grained Temporal Citation Analysis. http://arxiv.org/abs/2505.14838.
Barber K (2025) (Don't) click here: Hyperlinks as a quasi-objectification strategy in epistemic legitimisation in extremists' blog posts on sexual violence. Discourse, Context & Media 66: 100912.
Beltagy I, Lo K and Cohan A (2019) SciBERT: A Pretrained Language Model for Scientific Text. http://arxiv.org/abs/1903.10676.
Binder A, Hennig L and Verma B (2022) Full-Text Argumentation Mining on Scientific Publications. In: Proceedings of the first Workshop on Information Extraction from Scientific Publications (eds T Ghosal, S Blanco-Cuaresma, A Accomazzi, et al.), Online, November 2022, pp. 54–66. Association for Computational Linguistics.
Bodström E (2023) Illusions of objectivity: The two functions of country of origin information in asylum assessment. Migration Studies 11(1): 197–217.
Bonet-Jover A, Sepúlveda-Torres R, Saquete E, et al. (2023) Applying Human-in-the-Loop to construct a dataset for determining content reliability to combat fake news. Engineering Applications of Artificial Intelligence 126: 107152.
Bornmann L and Daniel H (2008) What do citation counts measure? A review of studies on citing behavior. Journal of Documentation 64(1): 45–80.
Boulanger C (2026) The potential of LLMs for constructing a socio-legal knowledge graph. In: Simons A, Wüthrich A, Zichert M, et al. (eds) *Understanding Science with Large Language Models? Potentials for the History, Philosophy, and Sociology of Science*. Bielefeld: transcript, part-4.
Brooks TA (1985) Private acts and public objects: An investigation of citer motivations. Journal of the American Society for Information Science 36(4): 223–229.
Callon M, Courtial J-P, Turner WA, et al. (1983) From translations to problematic networks: An introduction to co-word analysis. Social Science Information 22(2): 191–235.
Callon M, Law J and Rip A (eds) (1986a) Mapping the Dynamics of Science and Technology: Sociology of Science in the Real World. London: Palgrave Macmillan.
Callon M, Law J and Rip A (1986b) Qualitative scientometrics. In: Callon M, Law J, and Rip A (eds) Mapping the Dynamics of Science and Technology. Sociology of Science in the Real World. Macmillan, pp. 103–123.
Cano V (1989) Citation behavior: Classification, utility, and location. Journal of the American Society for Information Science 40(4): 284–290.
Chavalarias D and Cointet J-P (2013) Phylomemetic patterns in science evolution—the rise and fall of scientific fields. PloS one 8(2): e54847.
Chubin DE and Moitra SD (1975) Content Analysis of References: Adjunct or Alternative to Citation Counting? Social Studies of Science 5(4): 423–441.
Cohan A, Ammar W, Zuylen M van, et al. (2019) Structural Scaffolds for Citation Intent Classification in Scientific Publications. http://arxiv.org/abs/1904.01608.
Cohan A, Feldman S, Beltagy I, et al. (2020) SPECTER: Document-level Representation Learning using Citation-informed Transformers. http://arxiv.org/abs/2004.07180.
Cole J and Cole S (1971) Measuring the Quality of Sociological Research: Problems in the Use of the "Science Citation Index". The American Sociologist. JSTOR: 23–29.

Cronin B (1981) The need for a theory of citing. Journal of Documentation 37(1): 16–24.
Dagdelen J, Dunn A, Lee S, et al. (2024) Structured information extraction from scientific text with large language models. Nature Communications 15(1): 1418.
Devlin J, Chang M-W, Lee K, et al. (2018) BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. http://arxiv.org/abs/1810.04805.
Di Marco C, Kroon FW and Mercer RE (2006) Using Hedges to Classify Citations in Scientific Articles. In: Shanahan JG, Qu Y, and Wiebe J (eds) Computing Attitude and Affect in Text: Theory and Applications. The Information Retrieval Series. Dordrecht: Springer Netherlands, pp. 247–263.
Donner P, Stahlschmidt S, Haunschild R, et al. (2025) Does citation context information enhance the validity of citation analysis for measuring research quality? An empirical comparison of peer assessments and enriched citations. Quantitative Science Studies: 1–36.
Dunn A, Dagdelen J, Walker N, et al. (2022) Structured information extraction from complex scientific text with fine-tuned large language models. http://arxiv.org/abs/2212.05238.
Edge D (1979) Quantitative Measures of Communication in Science: A Critical Review. History of Science 17(2): 102–134.
Feneberg V, Gill N, Hoellerer NIJ, et al. (2022) It's Not What You Know, It's How You Use It: The Application of Country of Origin Information in Judicial Refugee Status Determination Decisions – A Case Study of Germany. International Journal of Refugee Law 34(2): 241–267.
Fergadis A, Pappas D, Karamolegkou A, et al. (2021) Argumentation mining in scientific literature for sustainable development. In: Proceedings of the 8th Workshop on Argument Mining, 2021, pp. 100–111.
Fleerackers A, Riedlinger M, Moorhead L, et al. (2021) Communicating Scientific Uncertainty in an Age of COVID-19: An Investigation into the Use of Preprints by Digital Media Outlets. Health Communication 0(0). Routledge: 1–13.
Foucault M (2002) Archaeology of Knowledge. London: Routledge.
Funkquist M, Kuznetsov I, Hou Y, et al. (2023) CiteBench: A Benchmark for Scientific Citation Text Generation. In: 2023 Conference on Empirical Methods in Natural Language Processing, Singapore, December 2023, pp. 7337–7353. Association for Computational Linguistics.
Gilbert GN (1977) Referencing as persuasion. Social Studies of Science 7(1): 113–122.
Gilbert GN and Woolgar S (1974) Essay Review: The Quantitative Study of Science: an Examination of the Literature. Science Studies 4(3): 279–294.
Gorur D, Rago A and Toni F (2024) Can Large Language Models perform Relation-based Argument Mining? http://arxiv.org/abs/2402.11243.
Grootendorst M (2022) BERTopic: Neural topic modeling with a class-based TF-IDF procedure. http://arxiv.org/abs/2203.05794.
Heseltine M and Clemm von Hohenberg B (2024) Large language models as a substitute for human experts in annotating political text. Research & Politics 11(1).
Hill M (2026) The data interview. Reflexive integration of large language models in qualitative content analysis. In: Simons A, Wüthrich A, Zichert M, et al. (eds) *Understanding*

Science with Large Language Models? Potentials for the History, Philosophy, and Sociology of Science. Bielefeld: transcript, part-5.

Iqbal S, Hassan S-U, Aljohani NR, et al. (2021) A decade of in-text citation analysis based on natural language processing and machine learning techniques: an overview of empirical studies. Scientometrics 126(8): 6551–6599.

Jha R, Jbara A-A, Qazvinian V, et al. (2017) NLP-driven citation analysis for scientometrics. Natural Language Engineering 23(1): 93–130.

Jurgens D, Kumar S, Hoover R, et al. (2018) Measuring the evolution of a scientific field through citation frames. Transactions of the Association for Computational Linguistics 6: 391–406.

Kaplan N (1965) The norms of citation behavior: Prolegomena to the footnote. American Documentation 16(3): 179–184.

Karimi S, Moraes L, Das A, et al. (2018) Citance-based retrieval and summarization using IR and machine learning. Scientometrics 116(2): 1331–1366.

Keller R (2024) The Sociology of Knowledge Approach to Discourse. Springer.

Khalighinejad G, Scott S, Liu O, et al. (2024) MatViX: Multimodal Information Extraction from Visually Rich Articles. http://arxiv.org/abs/2410.20494.

Khan D, Ahmed I, Ullah I, et al. (2025) Finding the reference text in citation contexts using attention model. Service Oriented Computing and Applications 19(1): 45–55.

Kristeva J (1982) Desire in Language: A Semiotic Approach to Literature and Art. Columbia University Press.

Kunnath SN, Herrmannova D, Pride D, et al. (2021) A meta-analysis of semantic classification of citations. Quantitative science studies 2(4): 1170–1215.

Kunnath SN, Pride D and Knoth P (2022) Dynamic Context Extraction for Citation Classification. In: The 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing, Virtual, November 2022.

Kunnath SN, Pride D and Knoth P (2023) Prompting Strategies for Citation Classification. In: Proceedings of the 32nd ACM International Conference on Information and Knowledge Management, Birmingham United Kingdom, 21 October 2023, pp. 1127–1137. ACM.

Latour B (1987) Science in Action: How to Follow Scientists and Engineers through Society. Harvard University Press.

Lauscher A, Ko B, Kuehl B, et al. (2021) MultiCite: Modeling realistic citations requires moving beyond the single-sentence single-label setting. http://arxiv.org/abs/2107.00414.

Liesegang L and Gläser J (2026) Supporting citation context analysis with large language models raises questions that should have been asked 40 years ago. In: Simons A, Wüthrich A, Zichert M, et al. (eds) *Understanding Science with Large Language Models? Potentials for the History, Philosophy, and Sociology of Science*. Bielefeld: transcript, part-6.

Lo K, Wang LL, Neumann M, et al. (2020) S2ORC: The Semantic Scholar Open Research Corpus. http://arxiv.org/abs/1911.02782.

Luan Y, He L, Ostendorf M, et al. (2018) Multi-Task Identification of Entities, Relations, and Coreference for Scientific Knowledge Graph Construction. http://arxiv.org/abs/1808.09602.

Luukkonen T (1997) Why has Latour's theory of citations been ignored by the bibliometric community? Discussion of sociological interpretations of citation analysis. Scientometrics 38(1): 27–37.

Lyu D, Ruan X, Xie J, et al. (2021) The classification of citing motivations: a meta-synthesis. Scientometrics 126(4): 3243–3264.

Mao Y, Zhong M and Han J (2022) CiteSum: Citation Text-guided Scientific Extreme Summarization and Domain Adaptation with Limited Supervision. http://arxiv.org/abs/2205.06207.

Mercer RE, Di Marco C and Kroon FW (2004) The Frequency of Hedging Cues in Citation Contexts in Scientific Writing. In: Advances in Artificial Intelligence (eds AY Tawfik and SD Goodwin), Berlin, Heidelberg, 2004, pp. 75–88. Springer.

Merton RK (1973) The Sociology of Science: Theoretical and Empirical Investigations. Chicago: University Of Chicago Press.

Mikolov T, Sutskever I, Chen K, et al. (2013) Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems, 2013, pp. 3111–3119.

Mishra S, Gauquier A and Senellart P (2024) Modular Multimodal Machine Learning for Extraction of Theorems and Proofs in Long Scientific Documents (Extended Version). http://arxiv.org/abs/2307.09047.

Moravcsik MJ and Murugesan P (1975) Some results on the function and quality of citations. Social Studies of Science 5(1): 86–92.

Mulkay MJ (1974) Methodology in the sociology of science: Some reflections on the study of radio astronomy. Social Science Information 13(2): 107–119.

Myers G (1990) Writing Biology: Texts in the Social Construction of Scientific Knowledge. University of Wisconsin Press Madison.

Peritz BC (1983) A classification of citation roles for the social sciences and related fields. Scientometrics 5(5): 303–312.

Pickering A (ed.) (1992) Science as Practice and Culture. University of Chicago Press.

Pride D and Knoth P (2017) Incidental or Influential? – Challenges in Automatically Detecting Citation Importance Using Publication Full Texts. In: Kamps J, Tsakonas G, Manolopoulos Y, et al. (eds) Research and Advanced Technology for Digital Libraries. Lecture Notes in Computer Science. Cham: Springer International Publishing, pp. 572–578.

Pride D and Knoth P (2020) An Authoritative Approach to Citation Classification. In: Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020, Virtual Event China, August 2020, pp. 337–340. ACM.

Radford A, Narasimhan K, Salimans T, et al. (2018) Improving Language Understanding by Generative Pre-Training. OpenAI. Available at: https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf.

Reimers N and Gurevych I (2019) Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. http://arxiv.org/abs/1908.10084.

Roth C and Cointet JP (2010) Social and semantic coevolution in knowledge networks. Social Networks 32(1): 16–29.

Roy SS and Mercer RE (2022) Biocite: a deep learning-based citation linkage framework for biomedical research articles. In: Proceedings of the 21st workshop on biomedical language processing, 2022, pp. 241–251.

Sarin P and Alperin JP (2025) Citation Parsing and Analysis with Language Models. http://arxiv.org/abs/2505.15948.

Scharnhorst A, Yang H, Touber J, et al. (2026) Co-creation of AI technology, empowering curators of cultural heritage information and guarding research commons. In: Simons A, Wüthrich A, Zichert M, et al. (eds) *Understanding Science with Large Language Models? Potentials for the History, Philosophy, and Sociology of Science*. Bielefeld: transcript, part-5.

Schlattmann R, Kaye A and Vogl M (2026) From source to structure. Extracting knowledge graphs with LLMs. In: Simons A, Wüthrich A, Zichert M, et al. (eds) *Understanding Science with Large Language Models? Potentials for the History, Philosophy, and Sociology of Science*. Bielefeld: transcript, part-4.

Shibayama S, Yin D and Matsumoto K (2021) Measuring novelty in science with word embedding. PloS one 16(7): e0254034.

Simons A (2015) Fact-making in permit markets: document networks as infrastructures of emissions trading. In: Voß J-P and Freeman R (eds) Knowing Governance. Palgrave, pp. 177–192.

Simons A (2016) Documented Authority. The Discursive Construction of Emissions Trading in the Expert Literature. Berlin: Technische Universität Berlin. Available at: https://depositonce.tu-berlin.de/bitstreams/56c1cf1d-2a99-4f36-9844-c8ccbd55231d/download.

Simons A (2026) Scaling In, Not Up? Testing Thick Citation Context Analysis with GPT-5 and Fragile Prompts. In: Simons A, Wüthrich A, Zichert M, et al. (eds) *Understanding Science with Large Language Models? Potentials for the History, Philosophy, and Sociology of Science*. Bielefeld: transcript, part-6.

Simons A, Wüthrich A and Zichert M (2026) Doing science studies with large language models. An introduction. In: Simons A, Wüthrich A, Zichert M, et al. (eds) *Understanding Science with Large Language Models? Potentials for the History, Philosophy, and Sociology of Science*. Bielefeld: transcript.

Simons A, Zichert M and Wüthrich A (2026) Large language models for history, philosophy, and sociology of science: Interpretive uses, methodological challenges, and critical perspectives. *Studies in History and Philosophy of Science* 117: 102151. https://doi.org/10.1016/j.shpsa.2026.102151.

Small H (1973) Co-citation in the scientific literature: A new measure of the relationship between two documents. Journal of the American Society for information Science 24(4): 265–269.

Small H (1978) Cited documents as concept symbols. Social studies of science 8(3): 327.

Small H (1980) Co-citation context analysis and the structure of paradigms. Journal of Documentation 36(3): 183–196.

Small H (2004) On the shoulders of Robert Merton: Towards a normative theory of citation. Scientometrics 60(1): 71–79.

Swales J (1986) Citation analysis and discourse analysis. Applied Linguistics 7(1). Oxford University Press: 39–56.

Swales J (1990) Genre Analysis: English in Academic and Research Settings. Cambridge University Press.

Tahamtan I and Bornmann L (2019) What do citation counts measure? An updated review of studies on citations in scientific documents published between 2006 and 2018. Scientometrics 121(3): 1635–1684.

Teufel S, Siddharthan A and Tidhar D (2006a) An annotation scheme for citation function. https://aclanthology.org/W06-1312/.

Teufel S, Siddharthan A and Tidhar D (2006b) Automatic classification of citation function. In: Proceedings of the 2006 conference on empirical methods in natural language processing, 2006, pp. 103–110.

Teufel S, Siddharthan A and Batchelor C (2009) Towards domain-independent argumentative zoning: Evidence from chemistry and computational linguistics. In: Proceedings of the 2009 conference on empirical methods in natural language processing, 2009, pp. 1493–1502.

Törnberg P (2024) Best Practices for Text Annotation with Large Language Models. http://arxiv.org/abs/2402.05129.

Valenzuela M, Ha V and Etzioni O (2015) Identifying meaningful citations. In: Workshops at the twenty-ninth AAAI conference on artificial intelligence, 2015.

Vaswani A, Shazeer N, Parmar N, et al. (2017) Attention is all you need. Advances in neural information processing systems 30.

Vinkler P (1987) A quasi-quantitative citation model. Scientometrics 12(1–2): 47–72.

White HD (2004) Citation analysis and discourse analysis revisited. Applied Linguistics 25(1). Oxford University Press: 89–116.

Yousif A, Niu Z, Tarus JK, et al. (2019) A survey on sentiment analysis of scientific citations. Artificial Intelligence Review 52: 1805–1838.

Yu J, Jiang J, Yang L, et al. (2020) Improving Multimodal Named Entity Recognition via Entity Span Detection with Unified Multimodal Transformer. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (eds D Jurafsky, J Chai, N Schluter, et al.), Online, July 2020, pp. 3342–3352. Association for Computational Linguistics.

Zhang Q, Chen Z, Pan H, et al. (2024) SciER: An Entity and Relation Extraction Dataset for Datasets, Methods, and Tasks in Scientific Documents. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Miami, Florida, USA, 2024, pp. 13083–13100. Association for Computational Linguistics.

Zhang Y, Wang Y, Wang K, et al. (2023) When Large Language Models Meet Citation: A Survey. https://arxiv.org/abs/2309.09727.