Supporting citation context analysis with large language models raises questions that should have been asked 40 years ago

Lydia Liesegang and Jochen Gläser

1. Introduction

The aim of this paper is to explore opportunities to use large language models (LLMs) to improve the validity and reliability of citation context analysis (CCA) as a sociological method.[1] CCA analyses the ways in which cited references are used in citing texts (cf. Simons et al., 2026). It is a powerful method for the study of knowledge flows between texts (Aman and Gläser, 2025: 169–170, Anderson et al., 2023: 77–87), the field-specific gradual acceptance of knowledge claims over time (Cozzens, 1985) or the influence of fraudulent publications on later research (Schneider et al., 2020). It is therefore not surprising that the question of how it could be supported by using LLMs has arisen.

From a sociology of science perspective, the process of integrating LLMs into CCA is an instance of methods becoming routinised, explicated, standardised and in some cases engineered and automatised. These processes have been described for molecular-biological methods by Joan Fujimura, who traced the routinisation and standardisation of for molecular-biological methods through the development of manuals and the standardisation of reagents, instruments, specialised tests and software (Fujimura, 1996: 68–115). Alberto Cambrosio and Peter Keating described the "objectification" of the production of monoclonal antibodies, which was considered as "art" and "magic" in the beginning, through methods sections in experimental articles, experimental protocols in manuscript form, and articles dedicated to the method (Cambrosio and Keating, 1988). In both cases, the authors trace the development from an intuitive approach heavily dependent on tacit knowledge toward an explicit reproducible description of the algorithm

1 We are grateful to Grit Laudel and Arno Simons for comments that helped to significantly improve this chapter. We also acknowledge the contributions by Elaheh Sadat Ahmadi, whose Essay on the comparison of content analysis and citation context analysis (Ahmadi 2025) provided us with a new perspective on the literature on CCA.

underlying the method that will produce results which are expected to be reproducible in any laboratory around the world.

This process – and the possibility of its reversal – can be better understood by using two concepts introduced by the historian of science Hans-Jörg Rheinberger, who distinguished two kinds of objects which are used in research processes. Epistemic objects are objects about which researchers produce knowledge. They are the things researchers want to know more about, and which are considered as ill understood and still taking shape. To investigate epistemic objects, researchers use technical objects, which are things they believe to understand well enough to use them unquestioned (Rheinberger, 1997, 2016). Rheinberger points out that when laboratory work takes unexpected turns, technical objects may be re-questioned and turned back into epistemic objects until their behaviour is well understood again.

We believe that this is what should be happening with CCA. The formalisation and partial automatisation of CCA through attempts to support it with LLMs highlights general problems of the application of CCA. By requiring the explication and formalisation of methodological steps that humans may (and often do) execute subconsciously, the use of LLMs for CCA makes a critical revision of fundamental assumptions possible and necessary. We believe that discussions about LLM use in CCA point out methodological weaknesses in the latter rather than the former. Since the potential of LLM-based methods and other developments of artificial intelligence cannot be currently assessed, we focus on the methodological challenges such methods must meet. In particular we challenge three assumptions underlying any CCA:

A: The assumption that the citation context contains information about the cited source.
B: The assumption that the coding scheme does not need to provide for missing values.
C: The assumption that a syntactically delineated text provides sufficient information about the statements using the cited source.

In the remainder of the paper, we will begin by describing the method of CCA. We will then turn to the three assumptions and demonstrate for each that it is a taken-for-granted assumption of CCA, provide counter examples of citation contexts that violate this assumption, and suggest how the use of LLMs would need to be modified to accommodate cases that violate the assumption. We conclude with two general warnings about using current AIs to support CCA and the validation of that use.

2. Citation context analysis as a sociological method

CCA was originally developed with the intention to improve citation analysis because citation counts were considered to be an insufficient measure of a cited paper's quality (Moravcsik and Murugesan, 1975). Distinguishing between important and unimportant citations and between positive and negative citations remains the main motivation of CCA (Teufel et al., 2006; Zhang et al., 2013; Nishikawa and Koshiba, 2024). So far, none of these CCAs appear to have modified a citation-based evaluation, and no proposal of how

CCA could be integrated in evaluations utilising citation counts has been put forward. Thus, this tradition of CCA, which forms the mainstream, appears to still lack a use case – in the sense of being applied in a practical evaluation of research – after 50 years.

A second, non-mainstream tradition of CCA applies it to answer theory-driven questions about knowledge flows and the transformation of knowledge claims. Important examples include Cozzens (1985), who uses CCA to trace the reception of knowledge claims in biology and in the sociology of science, Amsterdamska and Leydesdorff (1989), who use CCA to distinguish multiple knowledge claims per citation and how they are transformed to fulfil different functions in the citing papers of biochemistry and the use of CCA to identify the impact of fraudulent research papers on subsequent research, and applications (Avenell et al. 2019; Schneider et al., 2020). Anderson and Lemken (2023) suggest using CCA for conducting literature reviews that are guided by a research question, which is also a use of CCA for answering theory-driven questions. This research tradition tends to use more complex coding schemes to answer theory-driven research questions. While it is marginal to CCA, it has demonstrated the potential of CCA to contribute to a wide range of theory-driven research in both qualitative and quantitative science studies.

The following description of CCA as a social-scientific method is applicable to both evaluative and analytical CCA because both need to comply with methodological principles of the social sciences. Discussing the steps of a CCA is useful because it highlights how many highly consequential methodological decisions remain implicit in most CCA applications. It is also a necessary step to evaluate current LLM applications to CCA and inform the discussion of a broader application.

In a CCA, four preparatory and four analytical steps can be identified (Table 1). These steps are preceded by the formulation of a research question, which defines the purpose of the CCA – the knowledge that is supposed to contribute to answer it – and informs all the steps of the CCA, particularly developing the coding scheme, defining the unit of analysis and building the corpora. Given the importance of the research question, its neglect in the many attempts to apply or develop the method further is both surprising and regrettable. In traditional method development (Zhang et al., 2013) as well as in the development of LLM-based methods (Jurgens et al., 2018; Kunnath et al., 2022; Nishikawa and Koshiba, 2024), some (potential) purposes of CCA are mentioned as legitimisation for developing CCA but these purposes do not appear to inform method development. In particular, generic coding schemes like the one introduced by Jurgens et al. (2018), which are not grounded in theories of knowledge production, are used across method developments without justification (Pride and Knoth, 2020; Kunnath et al., 2022). This shifts the epistemic function of categories from operating as a site of theoretical synthesis to operating in a technically efficient manner (Ahmadi, 2025: 16).

Prep. 1: Coding schemes constructed by mainstream evaluative CCA have become more sophisticated over time but essentially still aim at capturing substantial versus perfunctory use and positive versus negative evaluation of the cited source (see section 3 for examples of coding schemes). Analytical coding schemes are more complex but also often less strictly defined. The crucial aspect of this preparatory step is that the coding scheme – and its built-in assumptions – decides what information can be collected with

the CCA from the analysed texts. One of these assumptions is that every citation context contains information about the citation (see section 3), another one is that each citation context contains all the information that is to be collected with the coding scheme. In other words, CCA often assumes that there are no missing values. We will discuss this problem in section 4.

Table 1: Main steps of quantitative content analysis (steps for which LLM support is currently considered are highlighted)

Step		**Content**
Formulating research question		**Define knowledge gap and key concepts**
Prep. 1	Constructing coding scheme	Operationalise the key concepts
Prep. 2	Defining the unit of analysis	Decide which text linked to a citation contains the necessary information
Prep. 3	Building the corpus (cited papers)	Determine which citations are relevant for answering the research question
Prep. 4	Building the corpus (citing papers)	Collect documents which contain the necessary information
1	Locating citations	Find relevant citations in the documents
2	Delineating unit of analysis	Identify text to be coded for each citation
3	Coding unit of analysis	Interpret each text and allocate results to coding scheme
4	Integrate empirical evidence	Collate and comparatively analyse the coding results
Answering research question		Use empirical evidence to formulate new knowledge claims

Prep. 2: An important step of CCA that has been extensively discussed in evaluative CCA is the definition of the unit of analysis. The discussion usually focuses on the size of a syntactically constructed context, i.e. on how many sentences or paragraphs before and after a citation should be included to enable an accurate interpretation. Since delineating the unit of analysis is a key step for which LLMs are currently used, we discuss it in section 5.

Prep. 3/4: Building the corpus also crucially shapes the answer to the research question because the information that can be collected depends on which papers are analysed. The corpus consists of the cited papers, about which information is collected with the CCA

(Prep. 3) and the citing papers, that contain the information in their citation contexts (Prep. 4).

Steps 1–3: Once the preparatory steps have been completed, the actual CCA can be conducted in four steps. Steps 1 to 3 are repeated for every instance: A citation of a source from the source corpus must be located, and the unit of analysis must be delineated and interpreted, i.e. values according to the coding scheme must be assigned.

Step 4: Step 4 collates the interpretations. In the case of evaluative CCA, this usually amounts to a descriptive statistic, which in some cases is correlated with properties of cited or citing documents. In analytical CCA, more complex conclusions are drawn.

Once the CCA is completed, its results are used to answer the research question, often in combination with other information collected in a research project.

3. The citation context may not contain information about the citation

One of the premises of CCA is that the citation context that is interpreted by humans or categorised by LLMs contains information about the cited source. If this is not the case, the interpretation/automated coding of the citation context may return an incorrect result. This situation occurred in an application of a generative LLM for CCA by Nishikawa and Koshiba. The generative LLM was instructed to consider the paragraph including the citation and to categorise the positive, negative or neutral assessment of the cited source with the following instruction:

Figure 1: Prompt used for coding the citation purpose (Source: Fig. 1 of Nishikawa and Koshiba, 2024: 6759)

Prompt for citation purpose (Simple):

```
Classify the type of purpose for which the author of the
following text is citing the target. The type of citation
purpose category is Background, Comparison, Criticize,
Evidence, or Use. The label of the target literature is
Target: onwards, up to a new line. The thesis is the text
after Text:.  Just report the classification result.

Target: {Labels of cited papers in the text to be annotated}
Text:
{A paragraph to be annotated}
```

This did not work in a case where the cited sources (Latour, 2004 and Lupton and Mather, 1997) provided only background information and the criticism in the paragraph following the citation was directed at *a situation described in the cited source* rather than the cited source itself.

> This use of technical devices in an attempt to suppress political debates has been widely documented elsewhere (e.g., Latour, 2004; Lupton and Mather, 1997). At an extreme this use of GIS and mapping reinforces and aggravates existing divides and inequalities. (Nishikawa and Koshiba, 2024: 6769)

Since the LLM was instructed to classify the paragraph after the cited sources, it applied the negative assessment in this paragraph to the cited source. A correct classification requires the identification of the *argument* made in this paragraph rather than the *text* presented there. In Nishikawa's and Koshiba's coding scheme, the use of the cited sources for this argument could be categorised as Background, Evidence or Use.

This example could be dismissed as a case where the prescribed citation context (paragraph) did not match the actual citation context (the sentence ending with the citation). However, the problem could easily occur with a citing sentence, too. In the following example, the sentence citing Greenland 1990 and Greenland and Poole 2013 criticises frequentist analyses of observational data:

> In this respect, frequentist analyses of observational data seems to depend on unlikely assumptions that too often turn out to be so wrong as to deliver unreliable inferences, and hairsplitting interpretations of p values becomes even more problematic (Greenland, 1990; Greenland and Poole, 2013). (Schneider, 2015: 419)

Without analysing the cited sources, it is impossible to tell if they are examples of "hairsplitting interpretations of p values" or criticise such hairsplitting interpretations. Thus, it is not even clear to human readers if the text contains an assessment of the cited source or not. An LLM-based CCA working with linguistic markers would be likely to code this example as a negative citation of Greenland 1990 and Greenland and Poole 2013 if the prompt used by Nishikawa and Koshiba (2024) would be applied.

The general situation in which such a transference of negative evaluations to the cited source can be described as follows: If the cited source is used as a source of information about a phenomenon that is not produced by the authors of the cited publication but just reported by them, and if the phenomenon is criticised by the citing paper, this criticism is likely to be ascribed to the cited source. Human coders are likely to recognise the problem during coding and to adapt their interpretation, while it seems currently impossible to design a rule for an LLM that could take this situation into account.

4. The coding scheme must provide for missing values

Another assumption that often appears to inform the construction of coding schemes for CCA is that it matches the information contained by each citation context perfectly. In other words, CCA often assumes that there are no missing values in citation contexts and applies coding schemes that provide no option for handling such cases, as in the following example:

> The scheme which we devised to classify the reasons for citation of these pre-1930 papers is as follows:
>
> A: Historical background
> B: Description of other relevant work
> C: Supplying information or data, other than for comparison
> D: Supplying information or data for comparison
> E: Use of theoretical equation
> F: Use of methodology
> G: Theory or method not applicable or not the best one
>
> (Oppenheim and Renn, 1978: 226)

This coding scheme (and several others listed by Liu, 1993) is neither exhaustive, nor has it an option for missing values and unexpected information. Some of the coding schemes used in LLM-based CCA appear to follow the same tradition. For example, the coding scheme used by Nishikawa and Koshiba (s. Fig. 1) uses Background, Evidence, Comparison, Criticise or Use. Similarly, the highly popular coding scheme introduced by Jurgens et al. (2018) which has also been used by Pride and Knoth (2020) and Kunnath et al. (2022), does not appear to be exhaustive and has no room for missing values (Figure 2).

Figure 2: Coding Scheme used in experiments with LLMs (source: Jurgens et al., 2018: 393)

Class	Description	Example
BACKGROUND	*P* provides relevant information for this domain.	This is often referred to as incorporating deterministic closure (Dörre, 1993).
MOTIVATION	*P* illustrates need for data, goals, methods, etc.	As shown in Meurers (1994), this is a well-motivated convention [...]
USES	Uses data, methods, etc., from *P* .	The head words can be automatically extracted [...] in the manner described by Magerman (1994).
EXTENSION	Extends *P*'s data, methods, etc.	[...] we improve a two-dimensional multimodal version of LDA (Andrews et al, 2009) [...]
COMPARISON OR CONTRAST	Expresses similarity/differences to *P*.	Other approaches use less deep linguistic resources (e.g., POS-tags Stymne (2008)) [...]
FUTURE	*P* is a potential avenue for future work.	[...] but we plan to do so in the near future using the algorithm of Littlestone and Warmuth (1992).

The application is liable to creating situations in which coders a) may subconsciously invent information that is not in the citation context to resolve the conflict or, if they recognise the conflict, b) decide to ignore available information if it does not fit any of the options provided by the coding scheme, or c) expand the coding scheme and start all over again. Not all of these possibilities are open to both human coders and LLM, since only the former are able to deviate from the given instructions if they cannot fulfil their objective, and choose one of the options b) or c).

An LLM on the other hand must follow the given steps and find a matching value in the coding scheme for every citation context. Thus, it is forced to choose b) and produce a 'finding' that is not actually there. To avoid this, the applied coding scheme and the given prompt should be expanded to include an option for missing values.

Unfortunately, LLMs have a built-in tendency to please their human masters (sycophancy, Sharma et al., 2025) and are known to notoriously 'retrieve' non-existent data in

different contexts even when given the opportunity to report a missing value. This fosters the possibility that fictional data are distorting the CCA and should be considered when applying LLMs to CCA.

5. Syntactically delineated text is unlikely to provide accurate information about its semantics

The question of what text surrounding a citation should be analysed as citation context is a long-standing concern of CCA. Zhang et al. phrase this question as follows:

> In the context of citing behavior, … we need to make decisions regarding the following: … Should a single sentence, or a cluster of sentences in which a reference is mentioned, be selected as the unit of analysis? (Zhang et al., 2013: 1494).

In their review of NLP and machine learning approaches, Iqbal et al. review experiments with different numbers of sentences before and after a citation and conclude that "the use of four sentences (the citing sentence, one sentence before the citing sentence, and two sentences following the citing sentence) … is now considered as a quasi-norm for in-text citation analyses studies (Teufel et al., 2006)." (Iqbal et al., 2021: 6561)

These discussions of units of analysis for CCA point to a fundamental problem. As illustrated by Zhang et al. (2013) and Iqbal et al. (2021), traditional CCA implicitly assumes that a number of syntactic units can be selected as citation context. LLM-supported CCA initially used the same approach. However, the use and the assessment of a cited source occur in a scientific statement or argument, i.e. in a semantic unit. Any fixed syntactically delineated citation context is unlikely to coincide with the argument from which the purpose and the meaning of a citation's use arise. Most syntactic delineations of citation contexts including Nishikawa's and Koshiba's instruction of their generative LLM (see above, Figure 1) would fail in the following example from one of our own CCAs (Figure 3).

The arguments that use the reference Yan et al. extend to the following paragraph. At the beginning of this paragraph, the authors of the citing paper report results presented in the cited source and state what the authors of the cited source did not study, which is then used to motivate their own research. Thus, the cited source is used to support two entirely different arguments, one about what is known about linc-ITGB1 and one about what is not known. Most CCAs including the one conducted by the generative AI instructed by Nishikawa and Koshiba would find the former but would (be forced to) ignore the latter.

Figure 3: A citation context from cancer research (source: Li et al. 2017: 3400, solid line shows citing paragraph, dotted line the extension of the argument beyond the paragraph)

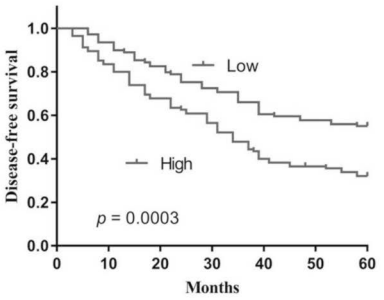


Figure 3. Correlation between linc-ITGB1 expression and DFS time in BC patients.

vention and therapy of breast cancer remain a major public health concern. More studies15,16 have recognized that the appropriate therapies require effective biomarkers to guide. Recently, growing papers reported that changes in expression of lncRNAs can be used as robust and important biomarkers for cancer risk, diagno- sis, and prognosis17-19. Based on previous find- ings, our attention focused on the linc-ITGB1.

Linc-ITGB1 is a newly discovered lncRNA. To our best knowledge, only two papers reported the specific role of linc-ITGB1 in tumors, including gallbladder cancer and BC. Wang et al20 firstly reported that expression of linc-ITGB1 was up-regulated in gallbladder cancer cells. Moreover, they confirmed that linc-ITGB1 was a promoter of gallbladder cancer metastasis due to its regulation of epithelial-to-mesenchymal transition (EMT). Yan et al12 found that linc-ITGB1 was overexpressed in BC tissues and cell lines, and its over-expression significantly inhibits cell proliferation, migration and invasion, while its knockdown exert the contrary role. Furthermore, constant with previous study, they also identified that inc-ITGB1 served as a tumor promoter by affecting EMT. Those findings informed that linc-ITGB1 played an important role in progres- sion and development of gallbladder cancer and BC. However, their research didn't study the clinical significance of linc-ITGB1 in patients.

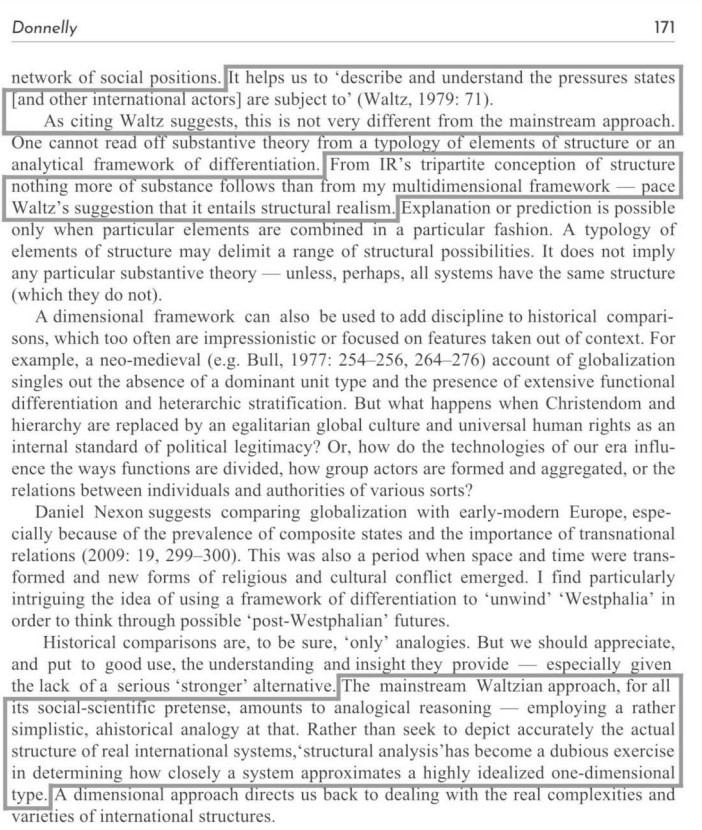
In the present work, we firstly determine the expression levels of linc-ITGB1 in BC and matched normal tissues and observed significantly increased expression of linc-ITGB1 in the BC tissues than in the adjacent normal breast tissues. We further found the expression level of linc-ITGB1 was significantly associated with lymph node metastasis, pathological differentiation and TNM stage, suggesting linc-ITGB1 contributed to progression of this tumor. In addition, Kaplan-Meier analysis demonstrated that high linc-ITGB1 expression level was associated with poorer OS and DFS. Based on these data, we performed multivariate analysis and the results indicated that linc-ITGB1 was an independent prognostic factor for predicting OS and DFS of BC patients. On the basis of the previous reports, we suggested hypothesized that the contribution of high linc-ITGB1 expression to the poor prognosis of BC patients might be caused by its facilitative efficiency on BC cell proliferation and metastasis.

Conclusions

We revealed, for the first time, that linc-ITGB1 may represent a valuable independent prognostic indicator for BC. Further study to discover the molecular mechanism of linc-ITGB1 about tumor migration and invasion is currently in progress.

In other cases, arguments using a cited source may occur completely separated from that source (Figure 4). In an article in the field of international relations, Donnelly (2011) cites a total of five publications by Kenneth Waltz. He refers to publications by Waltz not only by conventionally citing them but also by using his name or labels for his approach like "structural realism" as markers. The sample page shown in Figure 4 begins with a common type of citation context in which a publication by Waltz is used 'substantially' and cited 'positively' according to conventional coding schemes. Further down this page, the name provides the reference for a 'substantial' but 'neutral' use. Still further down and nowhere near a citation of one of Waltz's publications, "Waltzian approach" and "structural analysis" are used as markers for his publications with a thoroughly negative assessment.

Figure 4: Distribution of arguments independently of citations (source: Donnelly, 2011: 171)



Donnelly 171

network of social positions. It helps us to 'describe and understand the pressures states [and other international actors] are subject to' (Waltz, 1979: 71).

As citing Waltz suggests, this is not very different from the mainstream approach. One cannot read off substantive theory from a typology of elements of structure or an analytical framework of differentiation. From IR's tripartite conception of structure nothing more of substance follows than from my multidimensional framework — pace Waltz's suggestion that it entails structural realism. Explanation or prediction is possible only when particular elements are combined in a particular fashion. A typology of elements of structure may delimit a range of structural possibilities. It does not imply any particular substantive theory — unless, perhaps, all systems have the same structure (which they do not).

A dimensional framework can also be used to add discipline to historical comparisons, which too often are impressionistic or focused on features taken out of context. For example, a neo-medieval (e.g. Bull, 1977: 254–256, 264–276) account of globalization singles out the absence of a dominant unit type and the presence of extensive functional differentiation and heterarchic stratification. But what happens when Christendom and hierarchy are replaced by an egalitarian global culture and universal human rights as an internal standard of political legitimacy? Or, how do the technologies of our era influence the ways functions are divided, how group actors are formed and aggregated, or the relations between individuals and authorities of various sorts?

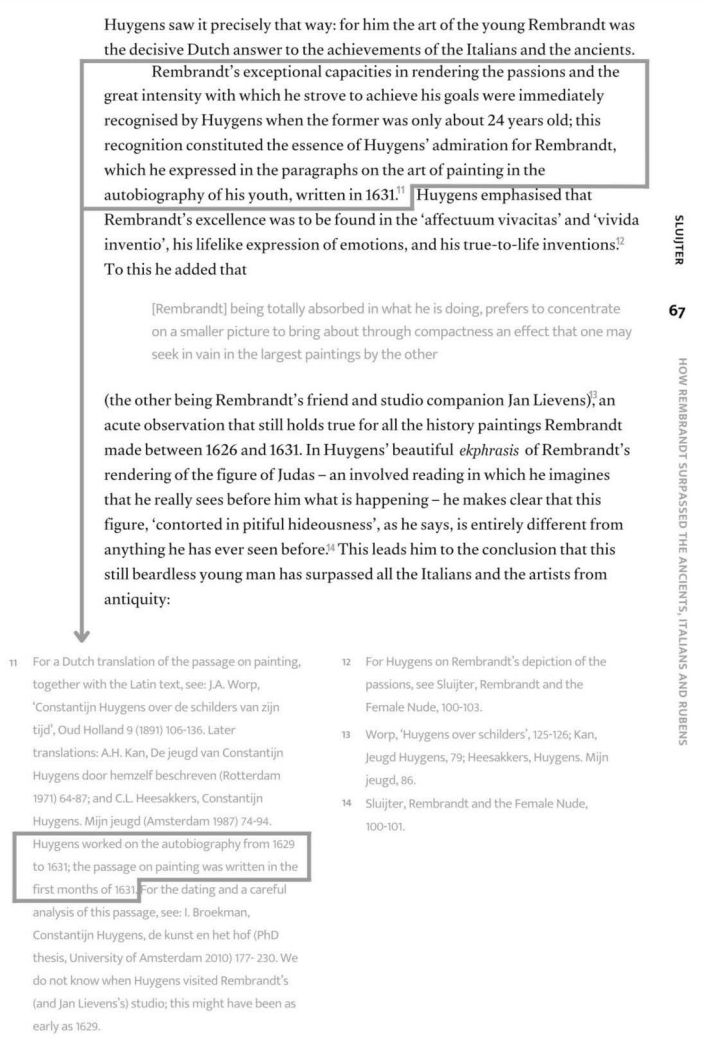
Daniel Nexon suggests comparing globalization with early-modern Europe, especially because of the prevalence of composite states and the importance of transnational relations (2009: 19, 299–300). This was also a period when space and time were transformed and new forms of religious and cultural conflict emerged. I find particularly intriguing the idea of using a framework of differentiation to 'unwind' 'Westphalia' in order to think through possible 'post-Westphalian' futures.

Historical comparisons are, to be sure, 'only' analogies. But we should appreciate, and put to good use, the understanding and insight they provide — especially given the lack of a serious 'stronger' alternative. The mainstream Waltzian approach, for all its social-scientific pretense, amounts to analogical reasoning — employing a rather simplistic, ahistorical analogy at that. Rather than seek to depict accurately the actual structure of real international systems,'structural analysis'has become a dubious exercise in determining how closely a system approximates a highly idealized one-dimensional type. A dimensional approach directs us back to dealing with the real complexities and varieties of international structures.

The cited source is used to support several different arguments. Furthermore, it is clear from phrases like "Waltzian approach" that the cited source is used together with other publications by Waltz referenced in the paper even though no source is provided close to the second and third arguments. In this case, the application of any known rule for syntactically delineating citation contexts would lead to an erroneous interpretation.

There are also field-specific citation practices and publication styles that do not work within narrow parameters, as elements of one argument can be scattered all over the publication, distant paragraphs of other parts of the publication or stay implicit all together. In the following example from the history of arts, the argument using the cited source is partly made in the main text and continuous in the footnote surrounding the cited source (Figure 5).

Figure 5: Distribution of an argument over a page (source: Sluijter, 2014: 67, frames delineate the two parts of the argument)



Huygens saw it precisely that way: for him the art of the young Rembrandt was the decisive Dutch answer to the achievements of the Italians and the ancients.

Rembrandt's exceptional capacities in rendering the passions and the great intensity with which he strove to achieve his goals were immediately recognised by Huygens when the former was only about 24 years old; this recognition constituted the essence of Huygens' admiration for Rembrandt, which he expressed in the paragraphs on the art of painting in the autobiography of his youth, written in 1631.[11] Huygens emphasised that Rembrandt's excellence was to be found in the 'affectuum vivacitas' and 'vivida inventio', his lifelike expression of emotions, and his true-to-life inventions.[12] To this he added that

> [Rembrandt] being totally absorbed in what he is doing, prefers to concentrate on a smaller picture to bring about through compactness an effect that one may seek in vain in the largest paintings by the other

(the other being Rembrandt's friend and studio companion Jan Lievens);[13] an acute observation that still holds true for all the history paintings Rembrandt made between 1626 and 1631. In Huygens' beautiful *ekphrasis* of Rembrandt's rendering of the figure of Judas – an involved reading in which he imagines that he really sees before him what is happening – he makes clear that this figure, 'contorted in pitiful hideousness', as he says, is entirely different from anything he has ever seen before.[14] This leads him to the conclusion that this still beardless young man has surpassed all the Italians and the artists from antiquity:

SLUIJTER 67 HOW REMBRANDT SURPASSED THE ANCIENTS, ITALIANS AND RUBENS

11 For a Dutch translation of the passage on painting, together with the Latin text, see: J.A. Worp, 'Constantijn Huygens over de schilders van zijn tijd', Oud Holland 9 (1891) 106-136. Later translations: A.H. Kan, De jeugd van Constantijn Huygens door hemzelf beschreven (Rotterdam 1971) 64-87; and C.L. Heesakkers, Constantijn Huygens. Mijn jeugd (Amsterdam 1987) 74-94. Huygens worked on the autobiography from 1629 to 1631; the passage on painting was written in the first months of 1631. For the dating and a careful analysis of this passage, see: I. Broekman, Constantijn Huygens, de kunst en het hof (PhD thesis, University of Amsterdam 2010) 177- 230. We do not know when Huygens visited Rembrandt's (and Jan Lievens's) studio; this might have been as early as 1629.

12 For Huygens on Rembrandt's depiction of the passions, see Sluijter, Rembrandt and the Female Nude, 100-103.

13 Worp, 'Huygens over schilders', 125-126; Kan, Jeugd Huygens, 79; Heesakkers, Huygens. Mijn jeugd, 86.

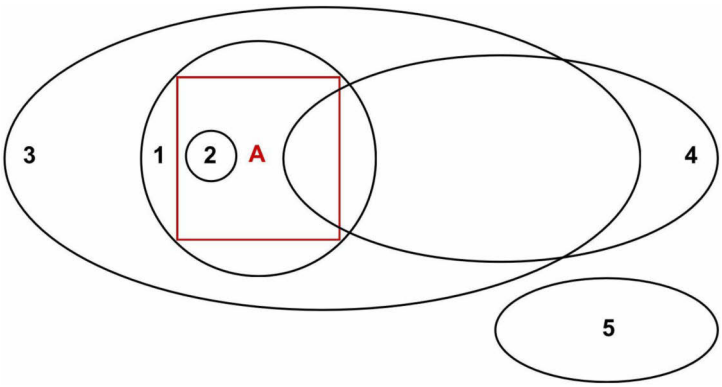
14 Sluijter, Rembrandt and the Female Nude, 100-101.

Thus, while the assumption by Ritchie et al. (2008: 216) "that the words that are used to describe the cited paper will occur close to the citation and that, further away, words are less likely to be about the paper" may be correct in most cases, it is impossible to predict how far away from a cited source arguments using it may occur in any individual case. Our examples illustrate the possibility that the argument or arguments using a cited source do not coincide with the syntactically defined citation context (e.g., the paragraph following or three sentences before and after a citation). If such a mismatch occurs, key information may be excluded from the interpretation because it is located outside the predefined citation context. In other cases of mismatch, the information about the cited

source may be distorted by information from other unrelated arguments that are fully or partly captured by the citation context.

The citation contexts in which citations are used and assessed are the argument(s) using the cited source, i.e. semantic units of analysis. This definition of the unit of analysis is independent from syntactic delineations because arguments have no specific syntactic dimension by which they could be reliably identified. The syntactical delineation of citation contexts that has been dominating both human and LLM-based CCA is unlikely to deliver valid results because mismatches are likely to occur (Figure 6).

Figure 6: Possible relationships between the argument using a cited source (red square) and a syntactically delineated citation context (ellipses)



The relationships between the argument and the citation context can have the following consequences for its representation:

1) **Match:** If the context is coextensive with the argument, the likelihood of a correct interpretation is highest.
2) **Part:** If only part of the argument is captured by the citation context, its interpretation is likely to be incomplete.
3) **Diluted:** If the context exceeds the argument, the argument's interpretation is likely to be diluted by other arguments.
4) **Distorted:** If the context contains part of the argument and a larger amount of unrelated text, its interpretation is likely to be based on a different argument that may even be the opposite of the argument using the citation.
5) **Unrelated:** If the context is not related to the argument at all, its interpretation is highly unlikely to represent the argument using the citation.

While the mismatch between syntactically delineated citation contexts and arguments that use cited sources is a problem of CCA in general rather than LLM-based approaches, it can be overcome much more easily by human coders than by LLMs. Human coders can (and we suspect often do) identify arguments and use them in lieu of the prescribed citation contexts 'on the fly', i.e. when a mismatch becomes obvious in the course of a

CCA. In contrast, AIs must follow the rules and thus cannot adapt to mismatches if the use of a fixed citation context is prescribed.

For LLM-based methods to avoid this mismatch, the rules must enable them to identify and delineate all kinds of statements in which authors use cited sources. While the likely mismatch between syntactically delineated citation contexts and arguments using the cited paper has not been widely discussed in the context of LLM-based CCA, the problem of using fixed citation contexts has been noted and motivated attempts to use LLMs for a dynamic delineation of citation contexts (e.g., Kunnath et al., 2022). The authors delineated citation contexts based on the similarity between embeddings of sentences in the citing paper and embeddings of title and abstract of the cited paper. They demonstrated that the classification of citation purposes 'improved' when variable citation contexts were used and 'improved' again when non-contiguous variable citation contests were used.

Although the experiments conducted by Kunnath et al. (2022) fall short of a validation of their method, two aspects merit discussion because they are likely to persist with LLM-based dynamic context detection. First, the method is based on the assumption that the text which is semantically most similar to the reference's title is most relevant for the categorisation of a citation context. This assumption, which is not explicated by Kunnath et al., has not undergone any empirical test. It is unlikely to be correct in all cases because a publication can be cited for different knowledge claims, not all of which are equally well represented by its title. Second, the method appears to replace fixed citation contexts with fixed thresholds for semantic similarity. Therefore, the possible mismatches between arguments and citation contexts shown in Figure 4 remain, with the ellipses representing text that is semantically similar to the cited source's title above a fixed threshold.

Another approach to identify the semantically relevant citation context is to train LLMs with a corpus in which text segments with specific semantic functions are annotated and citation used to train LLMs. Jantsch et al. (2025) defined three functions of a citation context, namely the information the citing author references, the relationship between the citing and the cited work, and the reason why a citation is included. They manually annotated 1056 citation contexts and used this data set to train an LLM. While this training of an LLM still exploits semantic similarity (in this case: between the annotated text segments and the model that is trained), the general approach could be used for the identification of scientific statements rather than epistemic functions of word sequences. This would require a training data set on all kinds of scientific statements in a research field that use citation, a task that seems currently impossible.

6. Discussion

We discussed three methodological challenges of CCA which precede the development of LLM-supported CCA but are highlighted by this development because LLM support requires the formulation of explicit rules for the LLM application. The explication of rules, in turn, reveals a gap in the methodological foundations of CCA and enables their discussion. In the context outlined in the introduction, the technical object (traditional CCA) is

turned into an epistemic object again because its future routinisation through LLM support problematises its foundations. CCA as a social-scientific method is in urgent need of methodological qualification. So far, neither the construction of coding schemes nor the delineation of citation contexts have sufficiently considered methodological principles of social science research. The resulting problems remained hidden in traditional CCA because (a) CCA is rarely used to answer research questions and thus can rarely fail, and because (b) human coders are likely to compensate for them 'on the fly', sometimes possibly without even noticing them.

Given that the development of LLM-supported CCA is still in its infancy and LLMs are rapidly evolving, we refrain from predicting how far they will be able to go. However, we would like to draw attention to three problems that need to be addressed as early as possible in the process of developing LLM-based CCA. First, the validation strategies for LLM-based methods need to be revised. If LLM-based methods applying faulty rules are assessed by comparing them to annotated citation contexts that have been produced applying the same faulty rules, the outcome contains little useful information.

Second, it seems doubtful that the development of LLM-supported CCA can productively continue without considering the relevant theoretical and methodological backgrounds in science studies. Developing methods without the input of a concrete, theory-based research question can isolate the methods from the very field they are intended to serve.

Third, there is a danger of stretching the conceptual foundations of LLM use. It is worth noting that the "distributional hypothesis" as formulated by Harris (1954) states that differences in the linguistic meaning of words correlate with differences between the distribution of linguistic entities in their context (Sahlgren, 2008; Lenci 2018: 152–153; Haber and Poesio, 2024; Resnik, 2025: 894). While it seems legitimate to conclude from this hypothesis that correlations between distributions of syntactic units in word contexts reflect semantic similarity, the distributional hypothesis does *not* state that the context of a word or phrase determines its meaning, which is often assumed in the literature (e.g., Andrews et al., 2014: 361; Boutyline and Arseniev-Koehler, 2025: 90). LLM developers and users need to carefully distinguish between what a LLM *does* and the actions they *ascribe* to a LLM based on their own creation of meaning. The delineation of citation contexts is a case in point. A delineation based on semantic similarity might be more useful than fixed citation context windows for some purposes but is still fundamentally different from the identification of a scientific statement that uses a cited source.

7. Conclusions

In this chapter, we highlighted some optimistic assumptions underlying current human-based and LLM-based applications of CCA, and demonstrated that these assumptions may prove wrong in an unknown but potentially large number of cases. Attempts to use LLMs for the support and scaling-up of CCA require explicating and formalising the method and thus reveal weaknesses it has had all along. From this follows that developing LLM support for CCA requires exploiting the possibility suggested by Rheinberger, namely transforming CCA back into an epistemic object by re-opening its conceptual

exploration. With regard to the conceptual foundations of LLM applications, our discussion highlights the difference between LLMs' mimicking of the human practice of using linguistic markers to assign labels to texts and true interpretation, which uses layers of contexts of varying sizes in a holistic act to establish a possible meaning of a text. While there are many cases in which humans use the same approach LLMs use today, and while marker-based labelling and interpretation may produce the same result in even more cases, there is an unknown number of cases in which the correct application of the two approaches would lead to different results.

Given these problems, we would like to invoke Robert Merton's "organised scepticism", which demands a temporary suspension of judgment and the detached scrutiny of beliefs in terms of empirical and logical criteria (Merton, 1973 [1942]: 277). While we do not see evidence that LLMs can solve the methodological problems we discussed, we do not consider this to be impossible. We also believe that experiments with LLMs can be productively used to further elaborate the methodological problems of CCA.

The more general argument that can be derived from our discussion is that regardless of the outcomes of current uses of LLMs so support interpretive methods in the social sciences, these experiments can lead to an improvement of methods because they force the explication of taken-for-granted assumptions and implicit methodological rules without which a LLM cannot be properly instructed. Attempts to use LLMs thus exercise pressure on social science methodology to improve the understanding of methods and to increase the precision of their description.

References

Ahmadi, ES (2025) Content Analysis and Citation Context Analysis in Comparison: A Methodological Perspective. TU Berlin. https://www.static.tu.berlin/fileadmin/www/10005401/Publikationen_sos/Ahmadi_2025_Measuring_codification.pdf

Aman, V and Gläser J (2025) Investigating Knowledge Flows in Scientific Communities: The Potential of Bibliometric Methods. Minerva 63(1): 155–182.

Amsterdamska, O and Leydesdorff, L (1989) Citations: Indicators of Significance? Scientometrics 15(5-6): 449–471.

Anderson, MH and Lemken RK (2023) Citation Context Analysis as a Method for Conducting Rigorous and Impactful Literature Reviews. Organizational Research Methods 26(1): 77–106.

Andrews, M, Frank, S and Vigliocco, G (2014) Reconciling Embodied and Distributional Accounts of Meaning in Language. Topics in Cognitive Science 6(3): 359–370.

Avenell, A, Stewart, F, Grey, A, et al. (2019) An investigation into the impact and implications of published papers from retracted research: systematic search of affected literature. BMJ Open 9(10): e031909.

Boutyline, A and Arseniev-Koehler, A (2025) Meaning in Hyperspace: Word Embeddings as Tools for Cultural Measurement. Annual Review of Sociology 51(Volume 51, 2025): 89–107.

Cambrosio, A and Keating, P (1988) "Going Monoclonal": Art, Science, and Magic in the Day-to-Day use of Hybridoma Technology. Social Problems 35(3): 244–260.

Cozzens, SE (1985) Comparing the Sciences: Citation Context Analysis of Papers from Neuropharmacology and the Sociology of Science. Social Studies of Science 15(1): 127–153.

Donnelly, J (2011) The differentiation of international societies: An approach to structural international theory. European Journal of International Relations 18(1): 151–176.

Fujimura, JH (1996) Crafting Science: A Sociohistory of the Quest for the Genetics of Cancer. Cambridge, Mass., Harvard University Press.

Haber, J and Poesio, M (2024) Polysemy—Evidence from Linguistics, Behavioral Science, and Contextualized Language Models. Computational Linguistics 50(1): 351–417.

Harris, ZS (1954) Distributional Structure. WORD 10(2-3): 146–162

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.

Jantsch, LM, Koh, D-J, Yoon, S et al. (2025) FineCite: A Novel Approach For Fine-Grained Citation Context Analysis, Vienna, Austria, Association for Computational Linguistics.

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.

Lenci, A (2018) Distributional Models of Word Meaning. Annual Review of Linguistics 4: 151–171.

Li, W-X, Sha, R-L, Bao, J-Q, et al. (2017) Expression of long non-coding RNA linc-ITGB1 in breast cancer and its influence on prognosis and survival. Eur Rev Med Pharmacol Sci 21(15): 3397–3401.

Liu, M (1993) Progress in documentation – The complexities of the citation practice: A review of citation studies. Journal of Documentation 49(4): 370–408.

Merton, RK (1973 [1942]) The Normative Structure of Science. The Sociology of Science. R. K. Merton. Chicago, The University of Chicago Press: 267–278.

Moravcsik, MJ and Murugesan, P (1975) Some Results on the Function and Quality of Citations. Social Studies of Science 5(1): 86–92.

Kunnath, SN, Pride, D and Knoth, P (2022) Dynamic Context Extraction for Citation Classification, Online only, Association for Computational Linguistics.

Nishikawa, K and Koshiba, H (2024) Exploring the applicability of large language models to citation context analysis. Scientometrics 129(11): 6751–6777.

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

Resnik, P (2025) Large Language Models Are Biased Because They Are Large Language Models. Computational Linguistics 51(3): 885–906.

Rheinberger, H-J (1997) Toward a History of Epistemic Things: Synthesizing Proteins in the Test Tube. Stanford, Stanford University Press.

Rheinberger, H-J (2016) On the Possible Transformation and Vanishment of Epistemic Objects. Teorie vědy / Theory of Science; Vol 38, No 3 (2016)DO – 10.46938/tv.2016.364.

Ritchie, A, Robertson, S and Teufel, S (2008) Comparing citation contexts for information retrieval. Proceedings of the 17th ACM conference on Information and knowledge

management. Napa Valley, California, USA, Association for Computing Machinery: 213–222.

Sahlgren, M (2008) The distributional hypothesis. Rivista di Linguistica 20(1): 33–53.

Schneider, JW (2015) Null hypothesis significance tests. A mix-up of two different theories: the basis for widespread confusion and numerous misinterpretations. Scientometrics 102(1): 411–432.

Schneider, J, Ye, D, Hill, AM, et al. (2020) Continued post-retraction citation of a fraudulent clinical trial report, 11 years after it was retracted for falsifying data. Scientometrics 125(3): 2877–2913.

Sharma, M, Tong, M, Korbak, T et al. (2025) Towards Understanding Sycophancy in Language Models. arXiv:2310.13548 [cs.CL].

Simons A, Arnaout H and Gurevych I (2026) Reconstructive citation context analysis using large language models. A roadmap. 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.

Sluijter, EJ (2014) How Rembrandt surpassed the Ancients, Italians and Rubens as the Master of "the Passions of the Soul". Low Countries Historical Review 129(2): 63–89.

Teufel, S, Siddharthan, A and Tidhar, D (2006) Automatic classification of citation function. Proc. 2006 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics: 103–110.

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