Camille Demers, University of Montreal, Montreal, QC, Canada

APPLYING LLMS AND SEMANTIC TECHNOLOGIES FOR DATA EXTRACTION IN LITERATURE REVIEWS: A PILOT STUDY IN LIS (Work in progress)

Abstract

This pilot study evaluates the capabilities of two LLMs, Mistral Small 3.1 and GPT-4o mini, in performing ontology-based data extraction to support literature reviews in library and information science (LIS). A sample of four published systematic reviews was selected as ground truth data. The open-access publications included in these reviews (n = 47) were collected as inputs for the models to perform semantic information extraction, using classes from the Document Components Ontology (DoCO). These preliminary findings highlight the opportunities and challenges of using AI and semantic technologies to streamline literature reviews in the social sciences.

Introduction

The rise of generative AI has opened avenues for research and education. The multiplication of applications based on large language models (LLMs) (e.g., Consensus, PaperDigest, PDFChat) suggests a paradigm shift in academia. Meanwhile, the rapid pace of contemporary scholarly production drives the need for solutions to manage an ever-growing body of publications that surpasses manual processing (Hong et al., 2021). In this context, initiatives like the Open Research Knowledge Graph (Jaradeh et al., 2019) or SemOpenAlex (Färber et al., 2023) leverage semantic technologies to promote alternative approaches to traditional document-based scholarly dissemination, which hinders access to scientific knowledge.

Given the known drawbacks of LLMs (e.g., hallucinations, lack of explainability, biases) (Hadi et al., 2023), this new technological ecosystem raises questions about maintaining scientific rigor and research quality standards while prioritizing processing efficiency. In this regard, this work seeks to explore methodological frameworks to evaluate AI and semantic-based tools supporting scientific literature reviews, focusing on data extraction. As this is a work in progress, this paper presents a preliminary study based on a small sample of publications included in published systematic reviews in Library and Information Science.

Background

Literature Review Automation

Automating literature reviews has gained significant interest over the past decade, driven by the rapid growth of scholarly production, with the overwhelming volume of publications across multiple fields making the process of knowledge synthesis increasingly challenging (van Dinter et al., 2021). Advances in natural language processing (NLP), machine learning (ML), and artificial intelligence (AI) have led to the development of tools to automate various stages of the literature review process (van Dinter et al., 2021; Bolanos et al., 2024).

While many steps in the review process are workload-intensive, studies on automating literature reviews have predominantly focused on the screening stage, leaving other crucial steps, like data extraction, underexplored (Affengruber et al., 2024). Additionally, tools designed for data extraction often target domain-specific information, with many focusing on biomedical elements (Legate et al., 2024). This focus has led to an overrepresentation of health-related fields at the expense of other disciplines, like social sciences, which face challenges in transferring methodologies designed for clinical data extraction (Legate et al., 2024; Wagner et al., 2022).

Scientific Information Extraction

Automatically extracting information from scientific publications is commonly referred to as scientific Information Extraction (IE). This task involves the identification of various concepts or semantic entities within textual documents produced in the context of scientific activity (Augenstein et al., 2017). With the increasing availability of scientific literature in digital formats, text mining and NLP tools can be employed to extract key facts and represent scientific knowledge in a more structured and accessible format (Hong et al., 2021). Such tools could serve various purposes for automated data extraction in literature reviews (Bolanos et al., 2024).

Traditional scientific IE models train supervised ML classifiers on datasets that include manually annotated examples of information to extract (Hong et al., 2021). Successful IE models have combined language models with deep neural networks, such as BERT variants (e.g., SciBERT [Beltagy et al., 2019], RoBERTa [Liu et al., 2019]), Convolutional Neural Networks, and Long Short-Term Memory Networks (Wang et al., 2022). Recently, LLMs have shown high potential for both domain-specific and domain-independent IE, enabling flexible data extraction without relying on annotated corpora (Dagdelen et al., 2024).

Semantic Technologies for Literature Reviews

Linked data, ontology models and knowledge graph technologies have gained attention for their ability to structure scientific knowledge. These tools have multiple applications, including citation networks (e.g., OpenCitations [Peroni & Shotton, 2020]), bibliographic data (e.g., BiBO [D'Arcus & Giasson, 2016]), and semantic representations of research papers (e.g., ORKG [Jaradeh et al., 2019], DoCO [Constantin et al., 2016]).

Regarding literature reviews, the use of semantic technologies has been motivated by three main factors: (1) facilitated access to content, (2) standardization of content representation, (3) alignment with FAIR data principles.

(1) **Facilitated access to scientific content:** Linked data and ontology-based content structuration enable a more granular, condensed representation of key research elements. This approach makes the content of documents more explicit, reducing the need to read extensive text to access the knowledge it contains (Mitchell & Mavergames, 2019).

(2) **Standardization of content representation**: Ontology modeling of knowledge reduces the variability in representing scientific content by providing terminological control, grouping terms based on lexical relations like synonymy and hyperonymy. The use of semantic web schemas further promotes practices that rely on standardized models that can be reused (Ali & Gravino, 2018; Mitchell & Mavergames, 2019).

(3) **Alignment with FAIR data principles:** The above-mentioned motivations for using semantic technologies in literature reviews converge on the idea of a better alignment of scientific publications with the FAIR data principles. This includes content accessibility, standardization and potential for reuse, but also interoperability, which is achieved using formal description languages (e.g. RDF) that are machine-readable (Oelen et al., 2020).

Methodology

Ontology model

The first stage of the workflow involved selecting an ontology model to structure the data extracted from the included publications. The Document Components Ontology (DoCO) (Constantin et al., 2016) was chosen for the purpose of this work. DoCO is part of the Semantic Publishing and Referencing Ontologies (SPAR) initiative (Peroni, 2014), which comprises a collection of ontologies for representing the publishing domain. This model was selected for its simplicity and domain independence, making it reusable across a broad range of disciplines. Parts of the model have also been used in other studies for semantic representation of scientific publications (Oelen et al., 2021). The use of this model builds on the rationale from these works.

DoCO is modeled to represent the main sections of a research work, including classes for structural elements (e.g., Paragraph, Section) and rhetorical elements (e.g., Data, Methods). For this study, only a few core classes were selected to represent the data elements to be extracted in the context of a literature review, namely : *deo:ProblemStatement, deo:Data, deo:Methods, deo:Results, deo:Evaluation*.

Data collection and preprocessing

Figure 1 presents an overview of the data collection and preprocessing workflow, which is described in detail in the following sections.

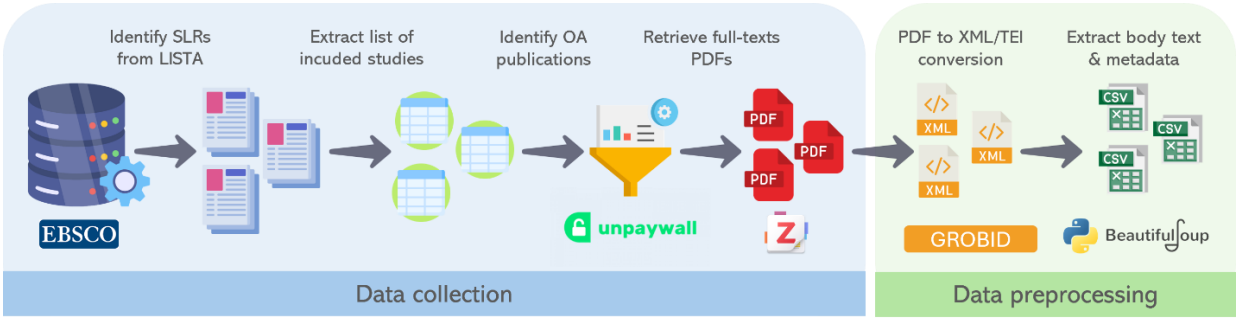

Figure 1 Data collection and preprocessing workflow[1]. SLRs: Systematic Literature Reviews, OA: Open access, TEI: Text Encoding Initiative.

Selection of existing literature reviews

A sample of manually conducted literature reviews was selected to create a ground truth dataset for evaluating the LLMs' information extraction capabilities. The objective is for the LLMs to replicate the manual data extraction process conducted in these reviews, following an approach similar to that of Oelen et al. (2021).

The reviews were collected from the Library, Information Science & Technology Abstracts (LISTA) database in January 2025. Only systematic reviews adhering to the PRISMA guidelines (Moher et al., 2009) were included. This criterion was applied to ensure that a table of study characteristics would be included in the reviews (which is recommended by PRISMA), as well as to increase methodological homogeneity between the reviews. Articles were retrieved using a search query for titles containing "a systematic review". Reviews from non-LIS fields (e.g., medical informatics, education) were excluded. Search results were sorted in reverse chronological order, and the first 30 articles were selected for screening. PRISMA compliance was verified through full-text screening, checking for explicit mentions of PRISMA.

Inclusion criteria were applied to refine the list of candidate articles, resulting in a final selection of 4 reviews, presented in the Results section. These criteria included: having at least two authors (n = 28); being written in English (n = 27); being conducted according to PRISMA guidelines (as stated by the authors) (n = 16); including 20–45 studies in the review (n = 7); presenting a table of included studies characteristics (n = 4).

[1] Credits for icons: www.flaticon.com (Vectorslab, Freepik, Flat Icons, kliwir art, Smashicons, surang), https://github.com/Impactstory/unpaywall/blob/master/extension/img/icon-128.png, https://github.com/zotero/zotero/blob/main/app/linux/icons/icon128.png

Collection of included publications from selected reviews

The list of all the studies included in the selected systematic reviews was retrieved either from the full texts of each review or from supplementary materials. We used Unpaywall's REST API[2] to identify Open Access (OA) studies from each review. Only OA studies were retained for subsequent stages, to avoid exposing work under proprietary licenses to the LLMs.

To enable the comparison between manual extraction and LLM-based extraction, the full texts of all OA studies included in the selected reviews were collected. Data collection was conducted semi-automatically using the *Find full-text* function of Zotero to retrieve the PDF of the included studies. Articles unavailable via Zotero were retrieved manually. The publications in PDF format were converted to XML/TEI using GROBID (Lopez, 2009). The XML files were parsed using the Python library BeautifulSoup (Richardson, 2007) to extract the text body and metadata.

Automated Data Extraction and Semantic Structuring with LLMs

Figure 2 presents an overview of the automated data extraction and evaluation workflow, which is described in the following sections.

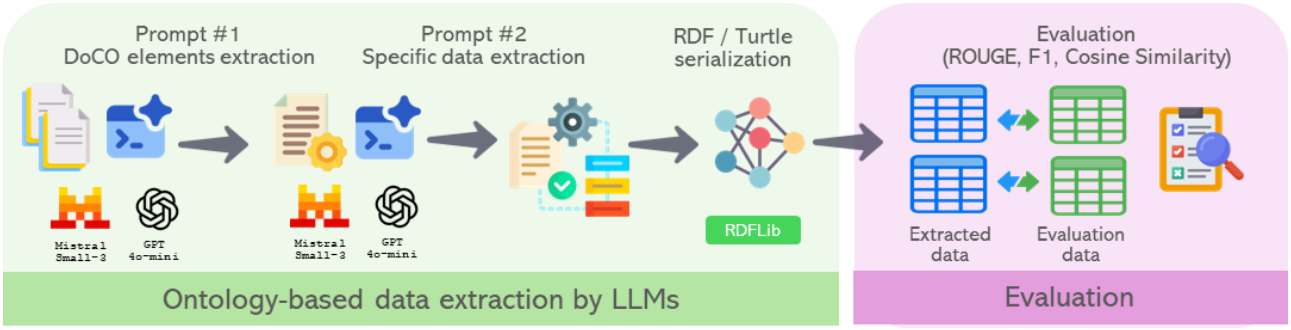

Figure 2 LLM-based data extraction workflow[3]. RDF: Resource Description Framework.

Models' selection

Given the preliminary nature of this study, we selected lightweight versions of two prominent LLMs currently available —one open-source, and the other proprietary: (1) Mistral Small 3.1 (MistralAI); (2) GPT-4o mini (OpenAI). These models both accept context inputs of up to 128k tokens, which suffice to process to full text of most scientific publications at a very low cost.

Prompt development

Two prompts were developed based on the instructions used in five previous studies that explored LLM-based information extraction (Datta et al., 2025; Foppiano et al., 2024; Gartlehner et al., 2024; Khan et al., 2024; Schmidt et al., 2024). The final versions of the prompts are presented in Appendix A and B.

Two prompts were used successively to instruct the LLMs to: 1) extract information corresponding to the targeted DoCO classes and 2) extract the data elements specific to the

[2] https://unpaywall.org/products/api

[3] Credits for icons : www.flaticons.com (Freepik, fzeetechz, redempticon, Becris, Design Circle, Creatype), https://en.m.wikipedia.org/wiki/File:Mistral_AI_logo_(2025%E2%80%93).svg, https://openai.com/brand/

reference reviews. Both prompts were developed using a subset of publications from one of the reference reviews specifically selected for prompt development. The LLMs were then evaluated on a test set comprising the OA publications included in three LIS systematic reviews. The reviews used for prompt development and test sets are presented in the Results section.

LLMs evaluation

The LLMs were evaluated using two metrics:

1. **ROUGE Score**: This metric measures the overlap of words between automatically generated and manually extracted text. It includes measures of recall, precision, and F-score (Lin, 2004).

2. **Cosine Distance**: This metric calculates the semantic similarity between pairs of phrases using vector embeddings of the sentences (Baeza-Yates et al., 1999). Sentence BERT (SBERT) embeddings (Reimers et Gurevych, 2019) were used in this study.

Results

Selected reviews and included studies

Prompt development set

The systematic review selected for prompt development is outlined in Table 1. This review was selected since only six of its included studies were available in open access.

Table 1 Characteristics of the systematic review used for prompt development

Publication	Title	Included studies (Total)	Included studies (OA)
Lookingbill and Wagner (2025)	The Role of Information and Communication Technologies in Disclosing and Reporting Sexual Assault Among Young Adults: A Systematic Review	23	6

Test set

The systematic reviews selected for the test set are outlined in Table 2. These reviews include studies related to library service platforms, human libraries, and librarians' professional development. Forty-one of the seventy studies included studies were available in open access.

Table 2 Characteristics of the three systematic reviews used in the test set

Publication	Title	Included studies (Total)	Included studies (OA)
Liu and Shao (2024)	A systematic review of library services platforms research and research agenda	22	12
Safdar et al. (2024)	A systematic review of literature on human libraries: Objectives, benefits, and challenges	24	14
Shahzad and Khan (2023)	The relationship between motivational factors and librarians' professional development (PD): A systematic review	24	15
Total		**70**	**41**

LLMs evaluation

Average performance across all studies

Table 3 outlines the average metrics of both models across the OA included studies. The evaluation was restricted to the data elements specific to the reference systematic reviews, as DoCO elements did not always clearly correspond to the review-specific data elements, making it difficult to establish ground truth for DoCO-based extraction.

Table 3. Average evaluation metrics for LLM-based information extraction

Model	ROUGE-Recall	ROUGE-Precision	ROUGE-F1	Cosine similarity
Mistral Small 3.1	0.4260	0.3675	0.3764	0.8619
GPT-4o mini	0.3950	0.3193	0.3349	0.8508

The modest ROUGE scores observed for both models are partially due to the rigidity of this measure, which relies on strict word overlap between the models' predictions and the ground truth, rather than on semantic similarity (Ng et Abrecht, 2015). However, these results varied across data elements. Some data elements have shown to be easier to extract and evaluate – particularly those that can be expressed as short keywords, such as the *Country* where a study was conducted. For this element, GPT-4o mini achieved an average ROUGE-F1 of 0.76, while Mistral Small 3.1 score 0.75.

These findings are further nuanced by cosine similarity measures, which suggest a strong semantic correspondence between model predictions and ground truth (0.8619 for Mistral Small 3.1 and 0.8508 for GPT-4o mini). Cosine similarity scores were also higher for shorter data elements like *Country* (0.9422 for Mistral Small 3.1, 0.9628 for GPT-4o mini), or *Analysis Method* (0.9013 for Mistral Small 3.1, 0.9286 for GPT-4o mini) than for those that are more susceptible to variability in phrasing, such as *Motivational factors toward professional development* (0.7815 for Mistral Small 3.1, 0.7783 for GPT-4o mini).

Given the limitations of the evaluation methods used in this preliminary study, future work will aim to explore the use of manual or qualitative evaluations to better highlight the potential and the challenges of using LLMs for scientific information extraction.

Conclusion

This pilot study explored the use of semantic and AI technologies to support ontology-based data extraction for literature reviews in the social sciences. Using lightweight LLMs yielded insightful yet preliminary results in extracting structured data from a sample of 41 studies in LIS. The limited scope of this work raises questions about the generalizability of these methods across diverse fields within social sciences. Future directions include refining the proposed evaluation workflow, exploring visualization functionalities through triple store integration, and incorporating additional language models. Ultimately, this work highlights some of the challenges and opportunities these technologies offer to support knowledge synthesis.

References

Affengruber, L., Maten, M. M. van der, Spiero, I., Nussbaumer-Streit, B., Mahmić-Kaknjo, M., Ellen, M. E., Goossen, K., Kantorova, L., Hooft, L., Riva, N., Poulentzas, G., Lalagkas, P. N., Silva, A. G., Sassano, M., Sfetcu, R., Marqués, M. E., Friessova, T., Baladia, E., Pezzullo, A. M., … Spijker, R. (2024, 12 juillet). An exploration of available methods and tools to improve the efficiency of systematic review production - a scoping review. https://doi.org/10.21203/rs.3.rs-4595777/v1

Ali, A. et Gravino, C. (2018, décembre). *An Ontology-Based Approach to Semi-Automate Systematic Literature Reviews*. 2018 12th International Conference on Open Source Systems and Technologies (ICOSST) (p. 09-16). https://doi.org/10.1109/ICOSST.2018.8632205

Augenstein, I., Das, M., Riedel, S., Vikraman, L. et McCallum, A. (2017, août). *SemEval 2017 Task 10: ScienceIE - Extracting Keyphrases and Relations from Scientific Publications*. S. Bethard, M. Carpuat, M. Apidianaki, S. M. Mohammad, D. Cer et D. Jurgens (dir.), SemEval 2017, Vancouver, Canada (p. 546-555). https://doi.org/10.18653/v1/S17-2091

Baeza-Yates, R., Ribeiro-Neto, B., et others. (1999). *Modern information retrieval* (vol. 463). ACM press New York.

Beltagy, I., Lo, K. et Cohan, A. (2019, novembre). *SciBERT: A Pretrained Language Model for Scientific Text*. K. Inui, J. Jiang, V. Ng et X. Wan (dir.), EMNLP-IJCNLP 2019, Hong Kong, China (p. 3615-3620). https://doi.org/10.18653/v1/D19-1371

Bolanos, F., Salatino, A., Osborne, F. et Motta, E. (2024, 13 février). Artificial Intelligence for Literature Reviews: Opportunities and Challenges. arXiv. https://doi.org/10.48550/arXiv.2402.08565

Constantin, A., Peroni, S., Pettifer, S., Shotton, D. et Vitali, F. (2016). The Document Components Ontology (DoCO). *Semantic Web*, *7*(2), 167-181. https://doi.org/10.3233/SW-150177

Dagdelen, J., Dunn, A., Lee, S., Walker, N., Rosen, A. S., Ceder, G., Persson, K. A. et Jain, A. (2024). Structured information extraction from scientific text with large language models. *Nature Communications*, *15*(1), 1418. https://doi.org/10.1038/s41467-024-45563-x

D'Arcus, B. et Giasson, F. (2016). Bibliographic Ontology (BIBO) in RDF. https://www.dublincore.org/specifications/bibo/bibo/bibo.rdf.xml

Datta, P., Datta, S. et Roy, D. (2025). *RAGing Against the Literature: LLM-Powered Dataset Mention Extraction*. New York, NY, USA. https://doi.org/10.1145/3677389.3702523

Färber, M., Lamprecht, D., Krause, J., Aung, L. et Haase, P. (2023). *SemOpenAlex: The Scientific Landscape in 26 Billion RDF Triples*. T. R. Payne, V. Presutti, G. Qi, M.

Poveda-Villalón, G. Stoilos, L. Hollink, Z. Kaoudi, G. Cheng et J. Li (dir.), Cham (p. 94-112). https://doi.org/10.1007/978-3-031-47243-5_6

Foppiano, L., Lambard, G., Amagasa, T. et Ishii, M. (2024, 31 décembre). Mining experimental data from materials science literature with large language models: an evaluation study. *Science And Technology of Advanced Materials-Methods*. TAYLOR & FRANCIS LTD. https://doi.org/10.1080/27660400.2024.2356506

Hadi, M. U., Tashi, Q. A., Qureshi, R., Shah, A., Muneer, A., Irfan, M., Zafar, A., Shaikh, M. B., Akhtar, N., Wu, J. et Mirjalili, S. (2023). A Survey on Large Language Models: Applications, Challenges, Limitations, and Practical Usage. https://www.authorea.com/doi/full/10.36227/techrxiv.23589741.v1?commit=b1cb46f5b0f749cf5f2f33806f7c124904c14967

Hong, Z., Ward, L., Chard, K., Blaiszik, B. et Foster, I. (2021). Challenges and Advances in Information Extraction from Scientific Literature: a Review. *JOM*, *73*(11), 3383-3400. https://doi.org/10.1007/s11837-021-04902-9

Gartlehner, G., Kahwati, L., Hilscher, R., Thomas, I., Kugley, S., Crotty, K., Viswanathan, M., Nussbaumer-Streit, B., Booth, G., Erskine, N., et others. (2024). Data extraction for evidence synthesis using a large language model: A proof-of-concept study. *Research Synthesis Methods*.

Jaradeh, M. Y., Oelen, A., Farfar, K. E., Prinz, M., D'Souza, J., Kismihók, G., Stocker, M. et Auer, S. (2019, 23 septembre). *Open Research Knowledge Graph: Next Generation Infrastructure for Semantic Scholarly Knowledge*. New York, NY, USA (p. 243-246). https://doi.org/10.1145/3360901.3364435

Khan, M. A., Ayub, U., Naqvi, S. A. A., Khakwani, K. Z. R., Sipra, Z. B. R., Raina, A., Zou, S., He, H., Hossein, S. A., Hasan, B., Rumble, R. B., Bitterman, D. S., Warner, J. L., Zou, J., Tevaarwerk, A. J., Leventakos, K., Kehl, K. L., Palmer, J. M., Murad, M. H., … Riaz, I. B. (2024). Collaborative Large Language Models for Automated Data Extraction in Living Systematic Reviews. *medRxiv: The Preprint Server for Health Sciences*, 2024.09.20.24314108. https://doi.org/10.1101/2024.09.20.24314108

Legate, A., Nimon, K. et Noblin, A. (2024, 20 juin). (Semi)automated approaches to data extraction for systematic reviews and meta-analyses in social sciences: A living review. F1000Research. https://doi.org/10.12688/f1000research.151493.1

Lin, C.-Y. (2004). *Rouge: A package for automatic evaluation of summaries* (p. 74-81).

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. et Stoyanov, V. (2019, 26 juillet). RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv. https://doi.org/10.48550/arXiv.1907.11692

Lopez, P. (2009). *GROBID: Combining Automatic Bibliographic Data Recognition and Term Extraction for Scholarship Publications*. M. Agosti, J. Borbinha, S. Kapidakis, C. Papatheodorou et G. Tsakonas (dir.), Berlin, Heidelberg (p. 473-474). https://doi.org/10.1007/978-3-642-04346-8_62

Mitchell, A. et Mavergames, C. (2019). Using linked data for evidence synthesis. *Systematic Searching: Practical ideas for improving results*, 171.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G. et PRISMA Group. (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Annals of internal medicine*, *151*(4), 264-269. https://doi.org/10.7326/0003-4819-151-4-200908180-00135

Ng, J.-P. et Abrecht, V. (2015, septembre). *Better Summarization Evaluation with Word Embeddings for ROUGE*. L. Màrquez, C. Callison-Burch et J. Su (dir.), EMNLP 2015, Lisbon, Portugal (p. 1925-1930). https://doi.org/10.18653/v1/D15-1222

Oelen, A., Jaradeh, M. Y., Stocker, M. et Auer, S. (2020, août). *Generate FAIR Literature Surveys with Scholarly Knowledge Graphs*. JCDL '20: The ACM/IEEE Joint Conference on Digital Libraries in 2020, Virtual Event China (p. 97-106). https://doi.org/10.1145/3383583.3398520

Oelen, A., Stocker, M. et Auer, S. (2021, 14 avril). *Crowdsourcing Scholarly Discourse Annotations*. IUI '21: 26th International Conference on Intelligent User Interfaces, College Station TX USA (p. 464-474). https://doi.org/10.1145/3397481.3450685

Peroni, S. (2014). The semantic publishing and referencing ontologies. *Semantic web technologies and legal scholarly publishing*, 121-193.

Peroni, S. et Shotton, D. (2012). FaBiO and CiTO: Ontologies for describing bibliographic resources and citations. *Journal of Web Semantics*, *17*, 33-43.

Peroni, S. et Shotton, D. (2020). OpenCitations, an infrastructure organization for open scholarship. *Quantitative Science Studies*, *1*(1), 428-444.

Reimers, N. et Gurevych, I. (2019, 27 août). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. arXiv. http://arxiv.org/abs/1908.10084

Sahlab, N., Kahoul, H., Jazdi, N. et Weyrich, M. (2022). A Knowledge Graph-Based Method for Automating Systematic Literature Reviews. *Procedia Computer Science*, *207*, 2814-2822. https://doi.org/10.1016/j.procs.2022.09.339

Schmidt, L., Hair, K., Graziozi, S., Campbell, F., Kapp, C., Khanteymoori, A., Craig, D., Engelbert, M. et Thomas, J. (2024, 23 mai). Exploring the use of a Large Language

Model for data extraction in systematic reviews: a rapid feasibility study. arXiv. https://doi.org/10.48550/arXiv.2405.14445

van Dinter, R., Tekinerdogan, B. et Catal, C. (2021). Automation of Systematic Literature Reviews: A Systematic Literature Review. *Information and Software Technology*, *136*, 106589. https://doi.org/10.1016/j.infsof.2021.106589

Wagner, G., Lukyanenko, R. et Paré, G. (2022). Artificial Intelligence and the Conduct of Literature Reviews. *Journal of Information Technology*, *37*(2), 209-226. https://doi.org/10.1177/02683962211048201

Wang, Y., Zhang, C. et Li, K. (2022). A review on method entities in the academic literature: extraction, evaluation, and application. *Scientometrics*, *127*(5), 2479-2520. https://doi.org/10.1007/s11192-022-04332-7

Appendix A: Prompt used for DoCO elements extraction

The prompts used in this study were developed based on the prompts developed in five previous studies that explored LLM-based information extraction (Datta et al., 2025; Foppiano et al., 2024; Gartlehner et al., 2024; Khan et al., 2024; Schmidt et al., 2024).

Instruction	Prompt section
Role assignement	- You are an expert at extracting semantic information extraction from scientific papers.\n
Data extraction	- You extract data from the paper provided by the user, based on the classes of the Discourse Elements Ontology (DEO) specified below:\n {deo_classes_description}
Expected shape of the output	- The data extracted should be a few keywords only, no full sentences.\n - Return the annotated paper into a valid JSON object with one field for each DEO element.\n
Expected behavior when the info is not available	If the information from a specific data class is not available in the paper, return NA for that element.\n
Avoid modifying / rephrasing the extracted data	Return the data as closely as they appear in the original paper, do not modify the text.\n
Avoid hallucinations	Do not include information outside the given paper. Do not make up an answer if the information is not available.

Appendix B : Prompt used for specific data extraction

The prompts used in this study were developed based on the prompts developed in five previous studies that explored LLM-based information extraction (Datta et al., 2025; Foppiano et al., 2024; Gartlehner et al., 2024; Khan et al., 2024; Schmidt et al., 2024).

Instruction	Prompt section
Role assignement	- You are an expert in data extraction for literature reviews in the social sciences.\n
Data extraction task	- You extract data from the paper provided by the user, based on the data elements specified below:\n {specific_data_elements}
Expected shape of the output	- The data extracted should be a few keywords only, no full sentences.\n - Return the annotated paper into a valid JSON object, with one field for each data element.\n
Expected behavior when the info is not available	If the information from a specific data element is not available in the paper, return NA for that element.\n
Avoid modifying / rephrasing the extracted data	Return the data as closely as they appear in the original paper, do not modify the text\n
Avoid hallucinations	- Do not include information outside the given paper.\n - Do not make up an answer if the information is not available.