

LLM supported reference string recognition and annotation – pilot study with educational research and social science publications

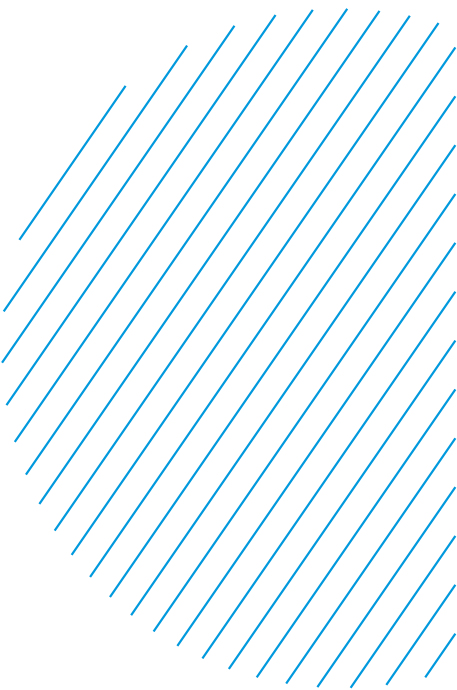
Niklas Kerkfeld; Tamara Heck; Muhammad Ahsan Shahid; Ezgi Tugyan;
Anele Schmidt; Verena Weimer; Ana Schenk; Thomas Oerder; Philipp
Mayr; Christoph Schindler

CiteX 2026, Frankfurt



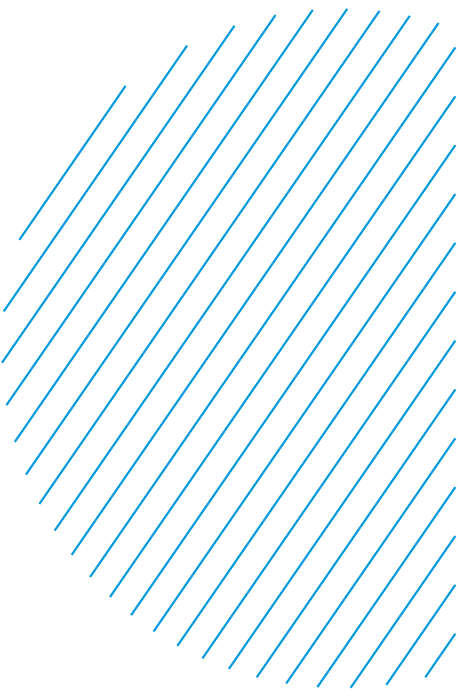
Agenda

- Motivation and related work
- Objectives and research questions
- Methodology and datasets
- Results and discussion
- Lessons learned



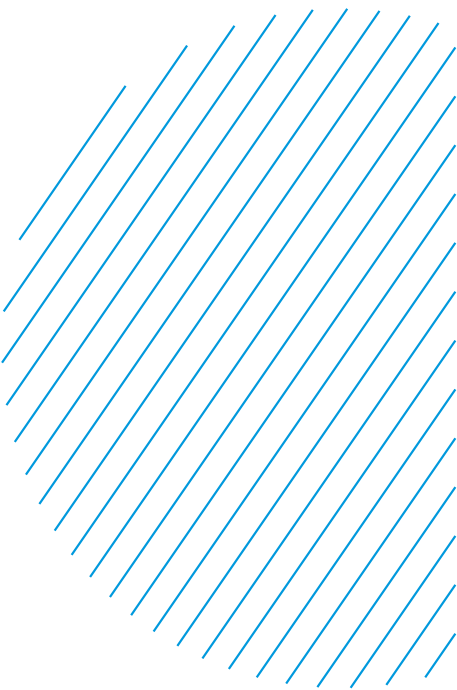
Motivation

- Bibliographic reference extraction and annotation are essential for open research information infrastructures
- Research publications contain:
 - Heterogeneous document types
 - Diverse and non-standard citation styles
- Manual annotation is:
 - Time-consuming
 - Costly
 - Difficult to scale
- Recent **large language models (LLMs)** show potential for:
 - Extracting references
 - Parsing bibliographic metadata
 - Recognizing citation styles



Related Work

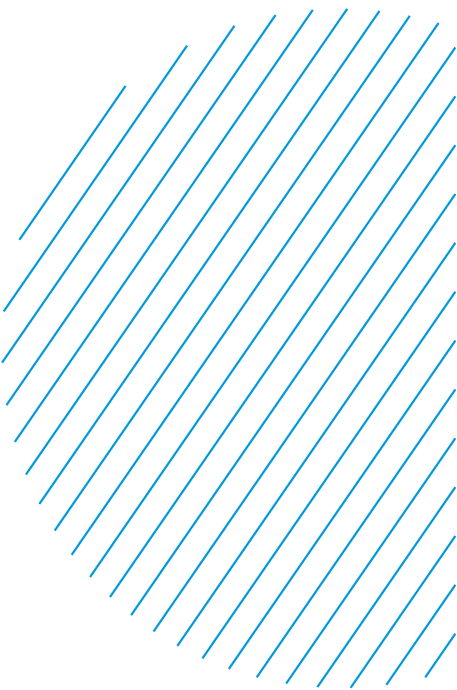
- According to [Backes et al. \(2024\)](#) as part of the OUTCITE project, combinations of tools:
 - GROBID
 - AnyStyleshowed the most robust performance for social science datasets.
- Based on [Zhu et al. \(2026\)](#)
 - Parsing and end-to-end parsing remains challenging
 - GROBID is strong and fast on well-structured PDFs
 - LLMs are more robust against multilingual, footnote-heavy, humanities styles pdfs
 - fine-tuning on Social Sciences and Humanities improves LLM performance
 - segmentation significantly improve LLM robustness and quality
 - Recommends using **GROBID for simpler** and **LLMs for more complex** documents



Research Objectives

This study investigates two questions:

1. How effectively can LLM **extract** references in educational research and social science publications?
2. How accurately can LLM **annotate** reference data from educational research and social science publications across relevant reference fields?



Final prompt

- Develop and iteratively refine prompting strategies based on data subsample
- All tags, concrete example excluded
- Apply best-performing prompt to all datasets

Please process the attached PDF file containing multiple references.

1. Extract ALL references from the PDF.

For each reference

- a) Output the full original reference string exactly as it appears using <refstring>
 - b) Annotate each reference using only the following XML-like tags: <author>, <surname>, <given-names>, <year>, <title>, <source>, <editor>, <volume>, <issue>, <fpage>, <lpage>, <publisher>, <other>.
2. Generate an XML tag only if the corresponding information is explicitly present in the reference.
 3. If a specific bibliographic element (e.g., editor, volume, issue, or pages) is not provided, do not generate the respective tag.
 4. If a reference cannot be clearly identified, mark it as „unidentified“ or leave the row empty.

Output format:

- Produce the result as an XML file structured like an Excel table.
- Each reference must appear in a separate row.
- Start with <refstring> and follow with the annotated elements.

Formatting rules:

<author>: use <surname> and <given-names> as sub-tags, with a comma between them.

<given-names>: always use the full given name as it appears in the reference. Do not abbreviate. (e.g., "Constanze" not "C.")

<editor>: plain string only, no sub-tags. „In:" and „(Hrsg.)" stay outside the tag as plain text. If multiple editors are present, each editor must be annotated in a separate <editor> tag.

<source>: use for any container of the work including: journal names, book titles, edited volumes, reports, websites.

<publisher>: use for the publishing house/institution (e.g. "Springer").

<other>: use for additional information that does not fit into other tags, including:

- place of publication (city)
- institutional details
- series information

If both place and publisher are present: <other>City</other>: <publisher>Publisher</publisher>

Page numbers: always use <fpage> and <lpage> separately.

<doi>: use only if a DOI is explicitly present.

<link>: use only if a URL is explicitly present.

Critical Rule:

- a) "In:" and "(Hrsg.)" must be preserved exactly as plain text
- b) Do not remove modify or place them inside any XML tag
- c) If present in <refstring> they must also appear in the structured output

Annotated datasets as gold standard

- EXCITE - SSOAR gold standard (n= 348 publications)
- [peDOCS](#) - Educational research publications (n= 27 publications)
 - Subset resembling disciplines, document types and publications date of peDOCS database
- Both datasets have been manually annotated according to the [EXCITE guideline](#):
 - 13 tags defined
 - `<author><surname> Lopez </surname>, <given-names> P.</given-names>(<year> 2009 </year>). <title> GROBID: Combining Automatic Bibliographic Data Recognition and Term Extraction for Scholarship Publications </title>. In: <editor> Agosti, M. </editor>,<editor>Borbinha, J.</editor>, <editor>Kapidakis, S. </editor>,<editor>Papatheodorou, C. </editor>,<editor>Tsakonas, G.</editor> (eds) <source> Research and Advanced Technology for Digital Libraries. ECDL 2009. Lecture Notes in Computer Science </source>, vol <volume>5714</volume>. <publisher>Springer</publisher>, <others> Berlin </others>,<others> Heidelberg </others>. <doi>https://doi.org/10.1007/978-3-642-04346-8_62 </doi>`
 - S. <fpage>69 </fpage>-<lpage>88</lpage>
 - <issue>, <url>, <identifier>

LLM models

Open WebUI for both datasets

- gpt-5.1 -> no peDOCS data due to copyright
- gpt-oss:120b -> both datasets
- DeepSeek R1 Distill Llama 70B (via Academic Cloud) -> peDOCS data as copyright protected

Different Pdf to text approaches

- Pdf as is, Marker¹, GROBID.txt²

¹ <https://github.com/datalab-to/marker> ² <https://grobid.org/>

Methodology

Processing pipeline for gpt

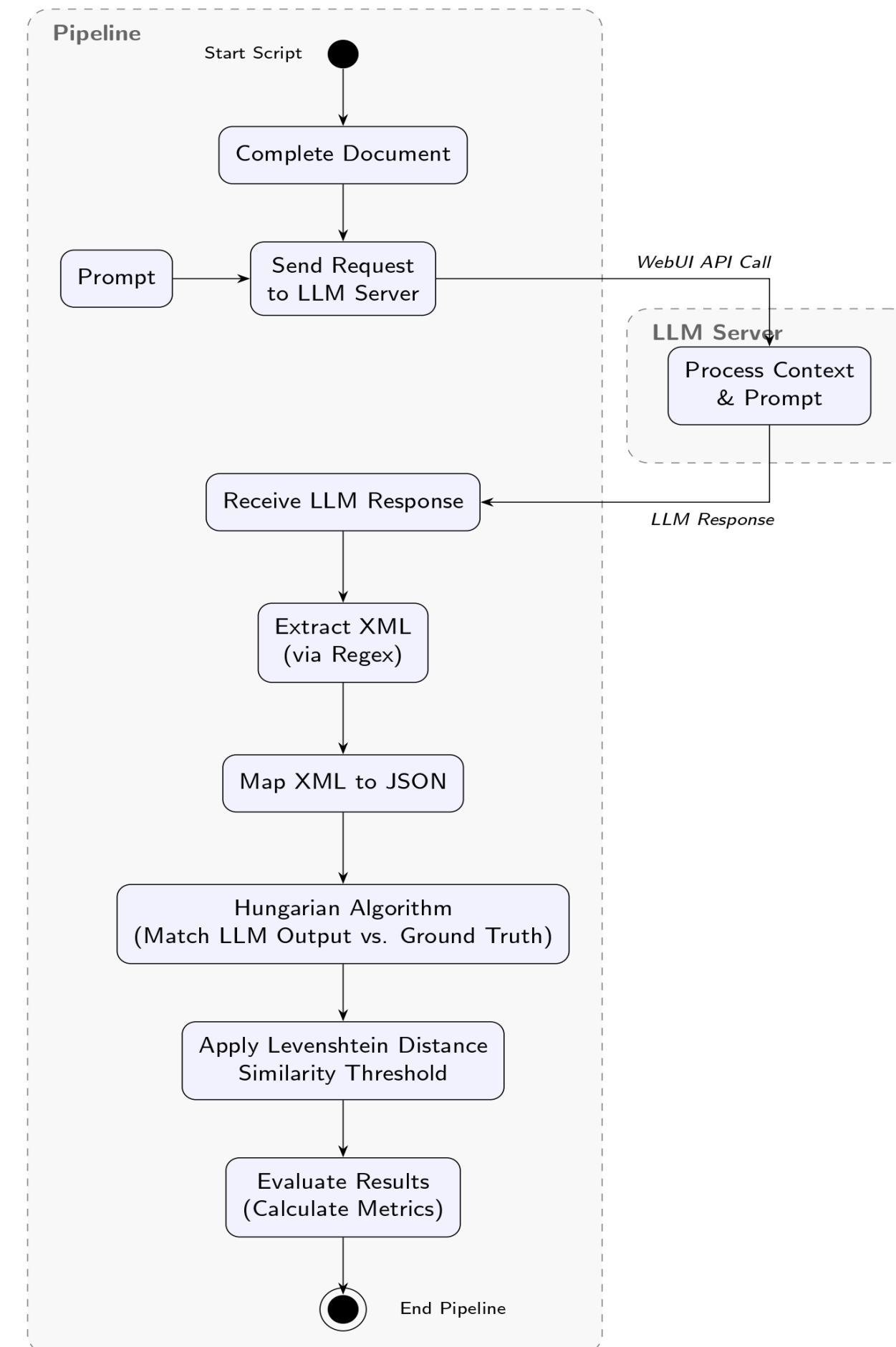
- Complete document is provided as context
- Automated over the WebUI API
- XML extracted using regex

Processing pipeline for DeepSeek

- GROBID applied for Pdf to txt
- Each txt files uploaded and processed in user interface

Evaluation metrics

- Mapped XML objects to a JSON data model
- Hungarian Algorithm
- Tolerating small errors
- Precision (P), Recall (R), F1 score



Results

EXCITE	refstr			title			year			author			editor			publ			source			vol			issue			startp			endp		
	P	F1	R	P	F1	R	P	F1	R	P	F1	R	P	F1	R	P	F1	R	P	F1	R	P	F1	R	P	F1	R	P	F1	R	P	F1	R
pdf-gpt-5.1	93	94	96	92	94	96	92	93	94	80	85	93	68	57	50	56	57	58	83	83	83	71	63	57	52	52	53	83	84	84	79	82	84
pdf-gpt-oss:120b	94	93	92	94	92	91	93	92	91	85	86	89	67	59	53	57	56	56	82	80	78	75	65	58	52	50	48	84	80	77	82	79	77
marker.txt-gpt-5.1	94	95	97	93	94	96	93	93	95	80	86	94	68	58	51	58	59	59	82	83	84	71	62	56	50	50	50	85	85	86	80	81	84
marker.txt-gpt-oss:120b	94	95	95	94	93	94	93	93	94	84	87	92	62	54	49	58	57	58	83	81	80	72	64	58	50	49	49	84	83	82	83	82	81
PEDOCS_GOLD	refstr			title			year			author			editor			publ			source			vol			issue			startp			endp		
	P	F1	R	P	F1	R	P	F1	R	P	F1	R	P	F1	R	P	F1	R	P	F1	R	P	F1	R	P	F1	R	P	F1	R	P	F1	R
pdf-gpt-oss:120b	97	93	91	95	91	89	96	91	88	87	84	82	65	61	59	71	66	63	81	73	67	60	56	52	30	29	28	74	74	75	74	74	75
marker.txt-gpt-oss:120b	98	98	98	97	97	97	97	96	96	87	89	91	74	72	70	71	71	71	90	82	76	58	57	56	32	31	32	79	79	80	78	78	79
grobid.txt-gpt-oss:120b	94	93	92	93	92	91	98	94	91	87	85	84	64	60	58	72	70	68	81	76	71	54	52	50	37	35	35	77	78	79	74	73	73
grobid.txt-deepseek-r1	98	95	93	98	95	93	98	92	88	86	86	87	72	64	58	74	74	74	73	72	71	66	60	57	36	34	34	75	74	73	75	74	73

Table 1. LLM models performances for predicting the most important reference fields as annotated in EXCITE (SSOAR Gold Dataset, n=348) and peDOCS (Educational Sciences Dataset, n=27).

Results

PEDOCS_GOLD	refstr			title			year			author			editor			publ			source			vol			issue			startp			endp		
	P	F1	R	P	F1	R	P	F1	R	P	F1	R	P	F1	R	P	F1	R	P	F1	R	P	F1	R	P	F1	R	P	F1	R			
pdf-gpt-oss:120b	97	93	91	95	91	89	96	91	88	87	84	82	65	61	59	71	66	63	81	73	67	60	56	52	30	29	28	74	74	75	74	74	75
marker.txt-gpt-oss:120b	98	98	98	97	97	97	97	96	96	87	89	91	74	72	70	71	71	71	90	82	76	58	57	56	32	31	32	79	79	80	78	78	79
grobid.txt-gpt-oss:120b	94	93	92	93	92	91	98	94	91	87	85	84	64	60	58	72	70	68	81	76	71	54	52	50	37	35	35	77	78	79	74	73	73
grobid.txt-deepseek-r1	98	95	93	98	95	93	98	92	88	86	86	87	72	64	58	74	74	74	73	72	71	66	60	57	36	34	34	75	74	73	75	74	73

Table 1. LLM models performances for predicting the most important reference fields as annotated in EXCITE (SSOAR Gold Dataset, n=348) and peDOCS (Educational Sciences Dataset, n=27).

PEDOCS_GOLD	refstr			title			year			author			source		
	P	F1	R	P	F1	R	P	F1	R	P	F1	R	P	F1	R
cerm.txt–anyst	96	92	88	<u>93</u>	87	82	92	87	83	82	80	79	71	68	66
cerm.ref–anyst	90	87	85	<u>83</u>	80	78	92	84	79	73	76	79	67	64	63
grob.txt–anyst	<u>99</u>	<u>96</u>	93	91	88	85	<u>99</u>	<u>95</u>	92	80	81	82	<u>77</u>	<u>75</u>	<u>75</u>
grob.ref–anyst	95	95	<u>96</u>	89	88	87	98	<u>95</u>	93	79	81	85	<u>76</u>	<u>74</u>	<u>74</u>
pdftotxt–anyst	95	92	90	88	84	81	97	<u>92</u>	87	79	79	79	<u>77</u>	72	67
cermine	90	87	85	81	77	74	91	83	77	61	64	67	<u>43</u>	40	39
grob.ref–cerm	95	95	<u>96</u>	<u>93</u>	89	86	97	94	92	79	75	72	52	50	48
grobid	95	95	<u>96</u>	<u>91</u>	<u>90</u>	<u>90</u>	97	<u>95</u>	<u>94</u>	<u>85</u>	<u>88</u>	<u>92</u>	75	68	63

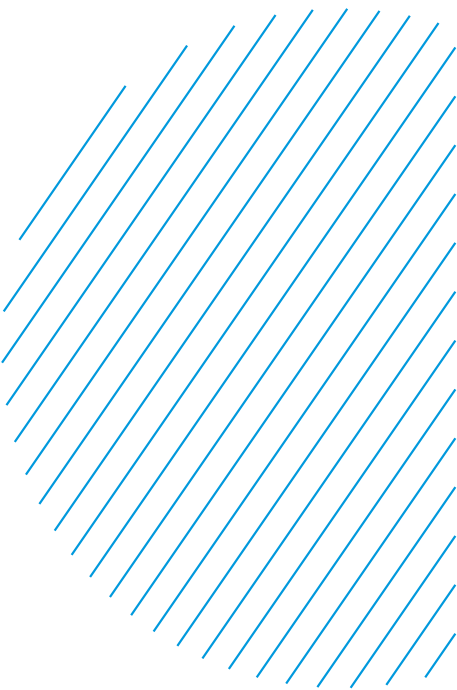
Tests in OFFZIB (presented at RefExtract25, n = 27))

Comparison with Backes et al., specifically with EXCITE

- Our LLMs have better performance in all fields except volume and issue
- Report that using LLMs didn't function in 2024 because of hallucinations

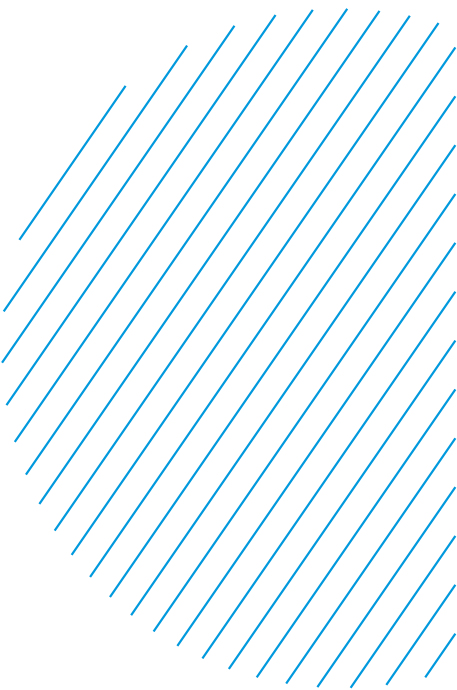
Comparison with Zhu et al., specifically with EXCITE

- Comparison difficult as they combine all fields to one score
- Results seem comparable
- LLMs capable of reference extraction and parsing
- Also report failures of LLMs on single pdf files
- Beneficial on complex, less structured documents



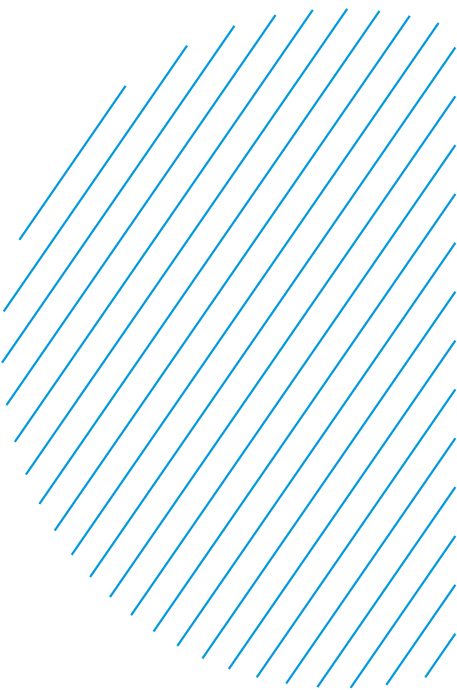
Lessons Learned

- Import issues of Pdfs
 - LLM servers could not process some Pdfs
- Copyright issues
 - German copyright law applies to (almost) all peDOCS documents
- Complex annotation due to special cases, e.g.
 - 2 x year of publication from original and second version (difference can be relevant for historical education)
 - Specific descriptions that do not belong to title or source, like “paper presented at...”, “reports of...”, “diploma thesis at...”
 - Non-standard page information, like „S.16 **und** S. 30.“
 - Abbreviations: Gysi, Jutta. 1990. „Frauen in Partnerschaft und Familie.“ In: Schwarz, Zenner (Hg.): **a.a.O.**, S. 90–119. -> „am angegebenen Ort“: editor manuscript cited separately
- Prompt optimizations
 - Generate an XML tag only if the corresponding information is **explicitly present** in the reference
 - <publisher> and <source> hard to distinguish
 - **Input format** needs to be exactly defined, e.g. Pdf file or text
- Complete Failures
 - Empty responses, responded instructions instead of extracted references, errors in xml



References

- Backes, Tobias et al. (Dec. 2024). “Comparing free reference extraction pipelines.” In: International Journal on Digital Libraries 25.4, pp. 841–853. <https://doi.org/10.1007/s00799-024-00404-6>
- EXCITE project: https://github.com/exciteproject/EXgoldstandard/tree/master/Goldstandard_EXparser
- GROBID (2008–2025). <https://github.com/grobidOrg/grobid>
- Open Citation Data for Educational Research (OFFZIB). <https://doi.org/10.17605/OSF.IO/HCKXY>
- Zhu, Yurui, Giovanni Colavizza, and Matteo Romanello (2026). Benchmarking Large Language Models on Reference Extraction and Parsing in the Social Sciences and Humanities. arXiv: 2603 . 13651 [cs.CL]. url: <https://arxiv.org/abs/2603.13651>



Thank you!

Questions?

Citation suggestion: Kerkfeld, N.; Heck, T.; Shahid, M. A.; Tugyan, E.; Schmidt, A.; Weimer, V; Schenk, A; Oerder, T.; Mayr, P.; Schindler, C. (2026). LLM supported annotation and reference style recognition – pilot study with educational research and social science publications. Presentation at the Workshop on Citation Extraction and Parsing (CiteX), 28-29 May, Frankfurt am Main, Germany.

Slides under: <https://doi.org/10.17605/OSF.IO/HCKXY>



This work is licensed under a [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/). Logos excluded.