Benchmarking Large Language Models on Reference Extraction and Parsing in the Social Sciences and Humanities

Yurui Zhu[1,*], Giovanni Colavizza[1] and Matteo Romanello[1]

[1]Odoma LLC

Abstract

Bibliographic reference extraction and parsing are foundational for citation indexing, linking, and downstream scholarly knowledge-graph construction. However, most established evaluations focus on clean, English, end-of-document bibliographies, and therefore underrepresent the Social Sciences and Humanities (SSH), where citations are frequently multilingual, embedded in footnotes, abbreviated, and shaped by heterogeneous historical conventions. We present a unified benchmark that targets these SSH-realistic conditions across three complementary datasets: CEX (English journal articles spanning multiple disciplines), EXCITE (German/English documents with end-section, footnote-only, and mixed regimes), and LinkedBooks (humanities references with strong stylistic variation and multilinguality). We evaluate three tasks of increasing difficulty—reference extraction, reference parsing, and end-to-end document parsing—under a schema-constrained setup that enables direct comparison between a strong supervised pipeline baseline (GROBID) and contemporary LLMs (DeepSeek-V3.1, Mistral-Small-3.2-24B, Gemma-3-27B-it, and Qwen3-VL (4B–32B variants)). Across datasets, extraction largely saturates beyond a moderate capability threshold, while parsing and end-to-end parsing remain the primary bottlenecks due to structured-output brittleness under noisy layouts. We further show that lightweight LoRA adaptation yields consistent gains—especially on SSH-heavy benchmarks—and that segmentation/pipelining can substantially improve robustness. Finally, we argue for hybrid deployment via routing: leveraging GROBID for well-structured, in-distribution PDFs while escalating multilingual and footnote-heavy documents to task-adapted LLMs.

Keywords

Reference extraction, Reference parsing, Benchmarking, Large language models

1. Introduction

Reliable citation processing is a prerequisite for large-scale scholarly infrastructure, as it facilitates the creation of curated bibliographies, citation graphs and citation indexes. In practice, reference extraction and parsing remain fragile and challenging in Social Sciences and Humanities (SSH) literature, where citations often appear in footnotes, rely on abbreviations and ellipsis, mix languages within a document, and follow style conventions that vary across periods, venues, and national traditions. These characteristics hinder both (i) *reference extraction* (detecting and delimiting reference spans within full text) and (ii) *reference parsing* (recovering structured fields such as author, title, venue/publisher, and date). In summary, providing robust and scalable citation processing of SSH literature is a major technical obstacle to overcome for building an SSH citation index [1], which constitutes the backdrop and ultimate objective of the work presented in this paper.

A second, closely related issue is evaluation mismatch. Widely used benchmarks often center on clean, English end-of-section bibliographies, which can inflate perceived robustness and hide SSH-specific failure modes. Meanwhile, SSH-oriented datasets are rarer and often cover only parts of the pipeline (e.g., parsing from gold strings rather than end-to-end processing from PDFs). Consequently, it remains difficult to assess real-world progress—both for supervised pipelines and LLM-based systems—under

SCOLIA '26: Second International Workshop on Scholarly Information Access (SCOLIA), April 2, 2026, Delft, The Netherlands

*Corresponding author.

yurui.zhu@odoma.ch (Y. Zhu); giovanni.colavizza@odoma.ch (G. Colavizza); matteo.romanello@odoma.ch (M. Romanello)

0009-0002-1429-1139 (Y. Zhu); 0000-0002-9806-084X (G. Colavizza); 0000-0002-7406-6286 (M. Romanello)

the regimes most relevant to SSH digitization and citation indexing.

In this paper, we address this gap with a benchmark designed around SSH-realistic conditions and operational constraints, and use it to assess whether contemporary LLMs are sufficiently reliable for reference extraction and parsing. We evaluate three complementary datasets: CEX, EXCITE, and LinkedBooks, capturing (i) controlled English journal references across disciplines, (ii) German/English documents with explicit citation-placement regimes (end-section, footnote-only, mixed), and (iii) humanities references with high stylistic and multilingual variability. We benchmark three tasks of increasing difficulty: reference extraction, reference parsing, and end-to-end document parsing. For comparability across heterogeneous systems, we use a schema-constrained evaluation: LLMs must produce strict JSON outputs, and GROBID predictions are mapped to the same target schema.

Our empirical focus is twofold. First, we provide a transparent, side-by-side comparison between a strong supervised baseline (GROBID) and several contemporary LLMs, including DeepSeek-V3.1, Mistral-Small-3.2-24B, Gemma-3-27B-it, and the Qwen3-VL family. Second, we study two strategies to boost performance in practice: (i) parameter-efficient LoRA adaptation for SSH-style parsing and (ii) document segmentation/pipelining strategies that trade global context for shorter, verifiable outputs. The resulting findings suggest a pragmatic hybrid direction: GROBID remains highly effective when its operating assumptions hold, but LLMs—especially when adapted—offer stronger robustness under distribution shift (multilinguality and footnote-heavy layouts). This naturally motivates *routing* policies that allocate easy, in-distribution documents to GROBID and escalate high-risk cases to adapted LLMs.

Contributions.

- We introduce a unified SSH-focused benchmark spanning CEX, EXCITE, and LinkedBooks, and covering extraction, parsing, and end-to-end document parsing under a shared schema and evaluation protocol.
- We benchmark GROBID against a diverse set of LLMs (DeepSeek-V3.1, Mistral-Small-3.2-24B, Gemma-3-27B-it, Qwen3-VL), highlighting where LLMs are competitive or superior under SSH-typical distribution shifts.
- We quantify the impact of two practical levers—LoRA adaptation and segmentation/pipelining—showing consistent improvements in robustness and end-to-end utility.
- We synthesize these results into actionable guidance for hybrid deployment via routing, balancing throughput and reliability in large SSH collections.

Paper outline. Section 2 situates our work in prior datasets and tool chains; Section 3 describes the three datasets; Section 4 details tasks, systems, and evaluation; Section 5 reports benchmarking results and error analyses; Sections 5.2 and 5.3 study LoRA and segmentation; and Section 6 concludes with implications for robust SSH citation processing. Our code is available at: Github

2. Related Work

Research on bibliographic reference extraction and parsing has long been shaped by two practical constraints: the limited availability of high-quality gold standards and a persistent mismatch between *laboratory-style* citation data (clean end-section bibliographies in English) and the heterogeneity of real scholarly documents, particularly in the Social Sciences and Humanities (SSH). This section reviews the datasets most commonly used to develop and benchmark reference processing systems, alongside representative toolchains ranging from supervised pipelines to recent LLM-based approaches. This context motivates our benchmark design, which explicitly targets multilinguality, footnote-heavy layouts, and long-form documents—settings in which traditional systems are known to degrade.

Existing datasets vary widely in authenticity, scale, and coverage. EXCITE [2] is notable for providing an SSH-oriented benchmark with a strong German component and an explicit mixture of end-section, footnote-only, and mixed citation regimes. Beyond reference strings, it includes intermediate artifacts

such as page-layout exports and corrected XML, which enable evaluation of end-to-end robustness rather than string-level parsing alone. GEOcite [3] similarly contributes German-language scientific publications with extracted reference strings and field-level segmentation annotations, mainly supporting supervised training for metadata extraction. In addition, CEX [4, 5] offers carefully aligned PDFs and Text Encoding Initiative (TEI) annotations compliant with the GROBID citation schema, yielding a controlled multi-disciplinary English benchmark where citation conventions differ across 27 SCImago subject categories. Complementary corpora leverage paired full-text markup: the XML Markup Evaluation Corpus [6] and Open Research Europe (ORE) [7] provide Journal Article Tag Suite (JATS)-aligned representations that facilitate large-scale harvesting of reference lists with structured metadata. At the other extreme, GIANT [8] provides massive synthetic supervision by rendering Crossref-derived records into reference strings under many Citation Style Language (CSL) styles. While this scale is valuable for training data-hungry models and broad style coverage, the clean generation process underrepresents the noise, abbreviation practices, multilingual variation, and idiosyncratic source types that are common in SSH references. Finally, LinkedBooks [9] stands out as a humanities-first resource, offering tens of thousands of manually annotated references spanning footnotes and bibliographies, including primary sources and abbreviated forms, with fine-grained field markup; its stylistic diversity and historical variability make it particularly suitable for stress-testing parsers under SSH-typical conditions.

On the methods side, supervised pipelines remain the dominant baseline for reference extraction and parsing. GROBID [10, 11] is widely adopted for PDF-to-TEI conversion and frequently used as a comparative standard; recent benchmarking confirms its strong performance alongside systems such as AnyStyle on heterogeneous evaluation suites [12]. Nevertheless, multiple studies report brittleness under distribution shift, especially for multilingual content, complex page geometry, multi-column layouts, and footnote-embedded citations, where segmentation failures and missed boundaries become common [13, 14]. Motivated by these limitations, recent work has begun to apply LLMs directly to citation parsing. Sarin and Alperin [15] construct paired datasets from JATS sources and show that small open-weight LLMs can achieve strong field-level accuracy in few-shot settings, often surpassing previously reported results on similar tasks. However, their studies largely focus on English, DOI-rich journal references and do not systematically evaluate SSH-specific regimes (footnotes, long monographs, mixed-language corpora), nor do they provide transparent, side-by-side comparisons against strong GROBID configurations on the same inputs.

Overall, a consistent gap remains: evaluation conditions are frequently misaligned with SSH reality, where citations are embedded in footnotes, abbreviated, multilingual, and shaped by diverse historical conventions. Our benchmarking work is designed to address this gap by jointly testing extraction, parsing, and end-to-end behavior on CEX, EXCITE, and LinkedBooks, while retaining GROBID as a strong supervised baseline and extending comparison to contemporary LLMs and parameter-efficient adaptation.

3. Datasets

For evaluation, we select three complementary datasets: the **CEX project Goldstandard**, the **EXparser Gold Standard** from the EXCITE project, and **LinkedBooks**. They differ in domain, language distribution, and citation style (end-of-document bibliographies vs. footnotes), enabling robust benchmarking across document- and reference-level tasks. To avoid skewing the dataset towards simple, non-SSH citation styles, we deliberately exclude additional datasets with overlapping scope. Table 1 summarize their main characteristics.

3.1. CEX

The **CEX project Goldstandard** (hereafter CEX) [5] is a curated and corrected expansion of Cioffi [4], released as paired PDFs and TEI-encoded reference annotations compliant with the GROBID citation model. It contains 112 English papers (1,581 pages; 5,160 references) spanning 27 general SCImago

Table 1
Overview of benchmark datasets. LinkedBooks is released as reference-level records; document counts are not directly comparable.

Dataset	# Documents	# References	Content
CEX	112	5,160	English-language journal articles from 27 SCImago disciplines
EXCITE	351	10,171	Majority German; rich mix of end-section, footnote-only, and mixed citation layouts
LinkedBooks	N/A	1,194	Monographs and articles; multilingual references: IT (72.3%), EN (18.3%), DE (4.3%), FR (3.9%), ES (0.6%), PT (0.5%), NL (0.1%)

disciplines, where reference formatting and citation conventions vary substantially across fields. We use the full set as a controlled benchmark to probe cross-domain generalization.

3.2. EXCITE

The **EXparser Gold Standard** (hereafter EXCITE) was created within the EXCITE project [2], a reference extraction and citation matching toolchain for German-language social-science publications, including sociology and migration studies, political science, health and social policy, and economics. EXCITE is designed to support layout analysis, reference identification, and segmentation/metadata extraction. It comprises 351 SSH papers (8,041 pages; 10,171 references): 251 German documents (219 with end-section references, 12 with mixed end-section and footnotes, and 20 footnote-only) and 100 English documents with standard end-section bibliographies. Beyond layout variation, the footnote-heavy classes (2 and 3) introduce **abbreviated back-references** typical of German SSH writing: Ebd. (ebenda, equivalent to ibid.) and a.a.O. (am angegebenen Ort, equivalent to op. cit.). The gold standard retains these strings verbatim without resolving them to their canonical referents.

The gold standard provides multiple aligned representations per paper (PDF plus intermediate layout/annotation files), enabling evaluation of extraction robustness across layout and citation-placement conditions. We use the full gold standard in our experiments.

3.3. LinkedBooks

LinkedBooks [9] is a humanities-oriented citation mining project centered on the historiography of Venice. It provides 40k+ manually annotated references extracted from both bibliographies and footnotes, with fine-grained field segmentation and substantial variation in humanities citation style, including primary vs. secondary references and abbreviated forms. The corpus is released with predefined splits (train ≈35k, validation ≈2k, test ≈2k). For our benchmark, we evaluate on the full test split but restrict to *primary* citations only. This yields 1,194 references, predominantly in Italian and English (72.3% IT, 18.3% EN), with smaller portions of German, French, Spanish, Portuguese, and Dutch. This multilingual and stylistically heterogeneous subset is well suited to stress-testing reference parsing in history- and humanities-style citation practices.

4. Setup

4.1. Tasks

We evaluate all systems on three tasks of increasing complexity: **reference extraction**, **reference parsing**, and **end-to-end document parsing**. This setup separates span detection from structured metadata prediction and assesses their combined behavior in realistic pipelines.

Reference extraction. A document-level task: given a document (or segment), systems must detect and delimit all bibliographic references—whether in bibliographies or footnotes—and return them as

plain-text spans without internal structure.

Reference parsing. A reference-level task with gold reference strings as input. The goal is to convert each string into a structured record (authors, title, year, venue/publisher, identifiers), isolating fine-grained semantic segmentation from extraction errors.

End-to-end document parsing. A pipeline task that jointly performs extraction and parsing from raw documents, producing structured reference records. Performance here reflects cumulative errors and thus approximates practical usability for citation indexing and linking.

4.2. Systems and infrastructure

We compare GROBID with several LLMs under a unified, schema-constrained setup.

GROBID baseline. We run GROBID with the **Deep Learning sequence labelling model** released by the CEX project [16]. The model is trained on citation data closely related to, but distinct from, our CEX gold standard (Section 3.1). It thus provides a strong supervised baseline for English scientific articles and has a distributional advantage on CEX; we therefore emphasize cross-dataset performance on EXCITE and LinkedBooks to assess robustness on multilingual, heterogeneous SSH material. For the reference parsing task, we use `/api/processCitationList` endpoint on gold reference strings directly; for the end-to-end parsing, we use `/api/processFulltextDocument` endpoint on PDF documents.

LLM models. Our pool includes open-weight models that can be self-hosted and one model we access via API for convenience. DeepSeek-V3.1 [17] represents a very-large, state-of-the-art open-source LLM; although open weights permit self-hosting, we evaluate it through the public API in this study. Mistral-Small-3.2-24B-Instruct-2506 (hereafter Mistral-Small-3.2-24B) [18] and Gemma-3-27B-it [19] cover a practical range for local inference. The Qwen3-VL family (4B, 8B, 30B-A3B, 32B) [20] enables controlled scaling comparisons within one architecture. We recognize that certain documents within our datasets are publicly accessible and may be included in the pretraining corpora of the assessed models. Nonetheless, there is no indication that any of the chosen LLMs were trained on the specific document-reference pairs utilized in our benchmark; the datasets also encompass closed-access publications that the models could not have accessed; furthermore, SSH citation processing is an under-represented domain in LLM training, thereby diminishing the likelihood of systematic bias.

Prompts, input representation, and decoding. All LLMs use a few-shot prompt with a fixed JSON schema (authors, full_title, year, venue/publisher, identifiers, etc.)[1]. Prompts include brief instructions, an inline schema block, and 1–2 style-diverse examples. Outputs must be strict JSON (no prose). For document-level tasks (reference extraction and end-to-end parsing), we first convert each input PDF into a text-only Markdown representation (including footnote text when present) using the `marker` PDF-to-text pipeline,[2] selected on the basis of an internal evaluation balancing extraction accuracy and processing speed, and provide the markdown text as the model input. GROBID, by contrast, ingests PDF files directly and performs its own layout analysis, so no intermediate conversion is applied. For reference parsing, the model input is the gold reference string and the PDF-to-Markdown step is bypassed. We use provider-recommended inference settings for each model (temperature and top_p), and we do not impose additional output-length caps beyond the serving backend's defaults/limits. Generations are validated as JSON; if an output is invalid (e.g., malformed/truncated JSON or empty), we perform a single retry with identical settings, after which the instance is counted as a failure.

[1] The prompts we use can be found at: https://github.com/odoma-ch/ssh-citation-index/tree/main/prompts
[2] https://github.com/datalab-to/marker

Infrastructure. Self-hosted models are served with vLLM on 2×A100 NVIDIA GPUs; DeepSeek-V3.1 is accessed via its public API; GROBID runs on a 24 GB MIG slice. We report accuracy metrics alongside wall-clock runtime and failures (invalid JSON, empty outputs, truncation).

4.3. Evaluation protocol

We evaluate three tasks: reference extraction, reference parsing, and end-to-end document parsing. All tasks use the same soft one-to-one matching: each gold reference is paired with the most similar prediction using a **fuzzy Levenshtein similarity score** over the reference string and key fields. Prior to scoring, we case-fold and normalize whitespace/punctuation; names are canonicalized in an order-insensitive manner (e.g., "Surname, First M." ≈ "First M. Surname"). We do not apply hard similarity thresholds; **similarity directly determines partial credit in precision and recall.** We evaluate against **verbatim gold standard strings** in their original document order, without resolving abbreviated back-references (e.g., Ebd., a.a.O.). Downstream resolution is left to dedicated entity-linking components.

For **reference extraction** (string level), we match predicted reference strings to gold strings and report Precision, Recall, and F1. Extra duplicates count as false positives; unmatched gold references count as false negatives.

For **reference parsing** and **end-to-end document parsing** (field level), we score structured fields for each matched reference under the same rule. We report micro-F1 (pooled over records and fields), macro-F1 (averaged over records), and per-field precision/recall/F1 for key fields (Title, Authors, Year), with additional fields in detailed tables.

Invalid, empty, or ill-formed JSON yields no valid items (TP = 0) and contributes to errors through matching. We additionally track failure counts and runtime as operational indicators.

5. Experiments and Results

5.1. Off-the-shelf LLM baselines

Our first experiment benchmarks off-the-shelf LLMs against the GROBID baseline in a single-call, few-shot setting (Section 4), without task-specific fine-tuning. We evaluate reference extraction, reference parsing, and end-to-end parsing on CEX, EXCITE, and LinkedBooks. Table 2 reports the main results. We do not report a standalone GROBID score for reference extraction, as GROBID is designed as a pipeline and does not provide an extraction-only endpoint that isolates bibliography/footnote segmentation from downstream parsing.

Task-wise performance. Across tasks, the LLMs and GROBID exhibit distinct performance regimes. For **reference extraction**, the larger LLMs largely saturate the problem on both CEX and EXCITE: once the model has sufficient capacity, span identification is highly reliable and invalid outputs are rare. Differences between top models are small, suggesting limited sensitivity to modeling choices.

For **reference parsing**, outcomes depend more on domain match and supervision. On CEX, GROBID leads, consistent with its in-domain training advantage and its role as a strong supervised baseline. On EXCITE and LinkedBooks, however, several LLMs match or surpass GROBID, indicating better cross-style generalization and greater robustness under distribution shift.

End-to-end results mirror this trade-off. GROBID performs best on CEX where its advantage compounds across stages, while on EXCITE the strongest LLMs overtake the GROBID by retaining more stable parsing under non-standard layouts and multilingual references.

Scaling effects. After a moderate capacity threshold, model size has limited impact on **extraction**, but plays a much larger role in **parsing** and **end-to-end** parsing, where systems must maintain structured, well-formed outputs over long and noisy inputs. This pattern holds both *within* and *across* model

Table 2
Benchmark results for reference extraction and parsing (single-call / few-shot). We report Precision (P), Recall (R), Micro F1, and Macro F1. Fail/Error counts include invalid/empty outputs for LLMs and pipeline failures for GROBID. Best Micro F1 per dataset within each task block is shown in bold; second-best is underlined.

Dataset	Model	P	R	Micro F1	Macro F1	# Fail/Errors	Runtime (s)
			Reference Extraction				
CEX	DeepSeek-V3.1	0.9616	0.9196	0.9373	–	0	566.74
	Mistral-Small-3.2-24B	0.9906	0.9915	**0.9910**	–	0	1315.98
	Gemma-3-27B-it	0.7087	0.6153	0.6413	–	0	1889.41
	Qwen3-VL-32B-Instruct	0.9681	0.9703	0.9689	–	0	1826.02
	Qwen3-VL-30B-A3B-Instruct	0.9609	0.9614	0.9611	–	0	656.30
	Qwen3-VL-8B-Instruct	0.9591	0.9559	0.9572	–	0	624.24
	Qwen3-VL-4B-Instruct	0.9652	0.9632	0.9629	–	0	1625.35
EXCITE	DeepSeek-V3.1	0.9472	0.9328	0.9308	–	2	846.02
	Mistral-Small-3.2-24B	0.8956	0.9261	0.8984	–	3	2965.49
	Gemma-3-27B-it	0.7387	0.7573	0.7348	–	2	2202.33
	Qwen3-VL-32B-Instruct	0.9446	0.9421	**0.9367**	–	0	1804.71
	Qwen3-VL-30B-A3B-Instruct	0.9346	0.9344	0.9263	–	0	1486.20
	Qwen3-VL-8B-Instruct	0.9380	0.9361	0.9285	–	2	2355.55
	Qwen3-VL-4B-Instruct	0.9256	0.9298	0.9184	–	5	2242.79
			Reference Parsing				
CEX	DeepSeek-V3.1	0.3966	0.3805	0.3863	0.3779	0	1344.44
	Mistral-Small-3.2-24B	0.7422	0.7401	0.7401	0.7350	0	1569.56
	Gemma-3-27B-it	0.7575	0.7092	0.7233	0.7046	2	4349.06
	Qwen3-VL-32B-Instruct	0.7502	0.7461	0.7475	0.7419	0	1989.12
	Qwen3-VL-30B-A3B-Instruct	0.7336	0.7351	0.7337	0.7271	0	693.55
	Qwen3-VL-8B-Instruct	0.7386	0.7229	0.7293	0.7197	0	762.24
	Qwen3-VL-4B-Instruct	0.7269	0.7115	0.7173	0.7101	1	3870.84
	GROBID	0.9120	0.8081	**0.8560**	0.8457	4	1362
EXCITE	DeepSeek-V3.1	0.8753	0.7838	0.8233	0.8054	12	1945.15
	Mistral-Small-3.2-24B	0.8997	0.8342	**0.8623**	0.8411	2	1409.64
	Gemma-3-27B-it	0.8877	0.7648	0.8167	0.8029	5	3334.53
	Qwen3-VL-32B-Instruct	0.8956	0.8237	0.8550	0.8354	1	2629.65
	Qwen3-VL-30B-A3B-Instruct	0.8659	0.8295	0.8448	0.8255	6	1943.02
	Qwen3-VL-8B-Instruct	0.8870	0.8131	0.8462	0.8293	5	1886.02
	Qwen3-VL-4B-Instruct	0.8562	0.7818	0.8151	0.7986	4	1832.85
	GROBID	0.7688	0.6310	0.6918	0.6853	0	2083
LinkedBooks	DeepSeek-V3.1	0.8714	0.7476	**0.8047**	0.7999	4	305.61
	Mistral-Small-3.2-24B	0.6682	0.8932	0.7645	0.7692	0	227.96
	Gemma-3-27B-it	0.6853	0.5200	0.5907	0.5493	0	405.74
	Qwen3-VL-32B-Instruct	0.6868	0.8956	0.7774	0.7716	0	403
	Qwen3-VL-30B-A3B-Instruct	0.6336	0.8689	0.7328	0.7237	1	349.00
	Qwen3-VL-8B-Instruct	0.6663	0.8577	0.7500	0.7418	0	358.01
	Qwen3-VL-4B-Instruct	0.6382	0.6587	0.6483	0.5968	1	1209.67
	GROBID	0.6615	0.7138	0.6866	0.6517	0	251
			End-to-end Parsing				
CEX	DeepSeek-V3.1	0.4722	0.4421	0.4553	0.4439	3	2392.30
	Mistral-Small-3.2-24B	0.6356	0.6044	0.6076	0.5922	0	3645.20
	Gemma-3-27B-it	0.6452	0.3606	0.4265	0.3577	1	1387.60
	Qwen3-VL-32B-Instruct	0.7325	0.7241	0.7269	0.7182	0	2048.62
	Qwen3-VL-30B-A3B-Instruct	0.7032	0.6972	0.6984	0.6868	6	4472.04
	Qwen3-VL-8B-Instruct	0.6867	0.6769	0.6801	0.6710	1	3108.10
	Qwen3-VL-4B-Instruct	0.6606	0.6472	0.6488	0.6368	7	4497.63
	GROBID	0.8862	0.7850	**0.8314**	0.8090	0	806.28
EXCITE	DeepSeek-V3.1	0.8408	0.6925	0.7518	0.7195	10	2158.85
	Mistral-Small-3.2-24B	0.6149	0.5322	0.5594	0.5339	3	4746.59
	Gemma-3-27B-it	0.7034	0.5097	0.5673	0.5237	3	5113.73
	Qwen3-VL-32B-Instruct	0.8189	0.7051	**0.7524**	0.7278	4	5117.88
	Qwen3-VL-30B-A3B-Instruct	0.7703	0.6739	0.7091	0.6767	7	3923.41
	Qwen3-VL-8B-Instruct	0.8079	0.6975	0.7432	0.7145	10	3367.76
	Qwen3-VL-4B-Instruct	0.7855	0.6822	0.7234	0.6999	34	5126.05
	GROBID	0.6163	0.5117	0.5477	0.5030	41	1813.20

families: smaller variants (e.g., Qwen3-VL-4B-Instruct) exhibit more failures and less stable behavior, while larger models are generally more robust and accurate.

Within Qwen3-VL, moving from Qwen3-VL-4B-Instruct to Qwen3-VL-8B-Instruct yields a clear jump in reliability, and the 30B-A3B and 32B models often approach or exceed the frontier-level performance of DeepSeek-V3.1, matching or outperforming GROBID on heterogeneous SSH datasets. At the same time, these gains are not strictly monotonic with parameter count: Mistral-Small-3.2-24B is frequently among the strongest models on EXCITE parsing and can outperform Qwen3-VL-32B-Instruct in some settings, while Qwen3-VL-4B-Instruct occasionally surpasses larger models such as Gemma-3-27B-it and Mistral-Small-3.2-24B on specific datasets or fields. In practice, scaling laws still provide a useful aggregate trend, but architectural design, instruction tuning, and model "freshness" can be at least as important as raw size when choosing a model for structured citation prediction.

SSH factors The EXCITE breakdown in Table 3 highlights how SSH-specific characteristics, i.e., language and citation placement shape performance. For LLMs, English vs. German bibliographies (Class 1) differ only slightly, whereas GROBID shows a language clearer gap under conventional layouts.

In German, shifting from bibliography-style references to **footnote-only** and **mixed** regimes causes a large drop for all systems, but GROBID deteriorates most sharply. LLMs are consistently more robust in footnote-heavy settings, with DeepSeek-V3.1 exhibiting the smallest placement-induced gaps, consistent with stronger generalization at higher capacity. Mixed regimes remain difficult across the board, likely because they combine segmentation ambiguity with cross-location consistency constraints.

Overall, LLMs are better suited to heterogeneous SSH citation practices than GROBID, but footnote-driven layouts remain the dominant failure mode for all systems.

Table 3
Performance breakdown on the EXCITE dataset by reference class and language. We report Precision (P), Recall (R), and F1 for reference extraction, reference parsing, and end-to-end parsing. Classes correspond to references appearing (1) exclusively in end-of-document bibliographies, (2) exclusively in footnotes, and (3) jointly in footnotes and end-of-document sections.

Task	Class / Lang	DeepSeek-V3.1	Mistral-Small-3.2-24B	Gemma-3-27B-it	Qwen3-VL-32B-Instruct	GROBID
Extraction	Class 1 (EN)	0.97 / 0.98 / 0.98	0.95 / 0.98 / 0.96	0.78 / 0.80 / 0.78	0.98 / 0.98 / 0.98	–
	Class 1 (DE)	0.94 / 0.95 / 0.94	0.88 / 0.94 / 0.90	0.73 / 0.78 / 0.74	0.95 / 0.96 / 0.95	–
	Class 2 (DE)	0.83 / 0.69 / 0.73	0.78 / 0.73 / 0.73	0.65 / 0.49 / 0.54	0.71 / 0.74 / 0.71	–
	Class 3 (DE)	0.81 / 0.46 / 0.57	0.80 / 0.46 / 0.56	0.56 / 0.35 / 0.41	0.70 / 0.36 / 0.46	–
Parsing	Class 1 (EN)	0.88 / 0.82 / 0.85	0.92 / 0.89 / 0.90	0.91 / 0.82 / 0.86	0.92 / 0.89 / 0.90	0.83 / 0.69 / 0.75
	Class 1 (DE)	0.89 / 0.79 / 0.83	0.90 / 0.83 / 0.86	0.89 / 0.76 / 0.82	0.90 / 0.81 / 0.85	0.77 / 0.63 / 0.69
	Class 2 (DE)	0.84 / 0.68 / 0.74	0.84 / 0.68 / 0.74	0.81 / 0.65 / 0.72	0.82 / 0.75 / 0.78	0.48 / 0.41 / 0.44
	Class 3 (DE)	0.70 / 0.50 / 0.57	0.89 / 0.62 / 0.71	0.75 / 0.48 / 0.56	0.79 / 0.57 / 0.65	0.61 / 0.51 / 0.56
End-to-end	Class 1 (EN)	0.89 / 0.78 / 0.83	0.75 / 0.67 / 0.70	0.77 / 0.59 / 0.65	0.89 / 0.79 / 0.83	0.75 / 0.65 / 0.69
	Class 1 (DE)	0.83 / 0.68 / 0.74	0.58 / 0.51 / 0.53	0.69 / 0.52 / 0.57	0.80 / 0.69 / 0.74	0.59 / 0.49 / 0.52
	Class 2 (DE)	0.68 / 0.48 / 0.55	0.47 / 0.32 / 0.37	0.49 / 0.16 / 0.23	0.67 / 0.56 / 0.60	0.22 / 0.14 / 0.16
	Class 3 (DE)	0.78 / 0.42 / 0.51	0.30 / 0.15 / 0.19	0.59 / 0.25 / 0.30	0.70 / 0.32 / 0.43	0.61 / 0.40 / 0.45

Error analysis. Figure 1 provides a reference-level view of failure modes on CEX, complementing the F1-based evaluation in the previous section. For this analysis, we first reconstruct a canonical reference string from each predicted JSON record (by concatenating key fields in a fixed order) and compare it to the gold reference string using the same fuzzy-matching procedure as in Section 4. The per-field similarities are then collapsed into four coarse categories: *Correct* if all fields reach a similarity of at least 0.95; *Minor error* if at least one field falls into the $[0.60, 0.95)$ band while none drops below 0.60; *Major error* if any key field is below 0.60 but the JSON is structurally valid; and *Structural error* if the output cannot be parsed into a reference at all (e.g., malformed or truncated JSON).

Figure 1**(a)** shows the resulting distribution over these categories, aggregated across all references. Most open-weight models and GROBID produce a large majority of *Correct* references, with residual mistakes dominated by *Minor* rather than *Major* errors, i.e., by field omissions or local formatting issues instead of completely mismatched records. DeepSeek-V3.1 remains an outlier with a comparatively

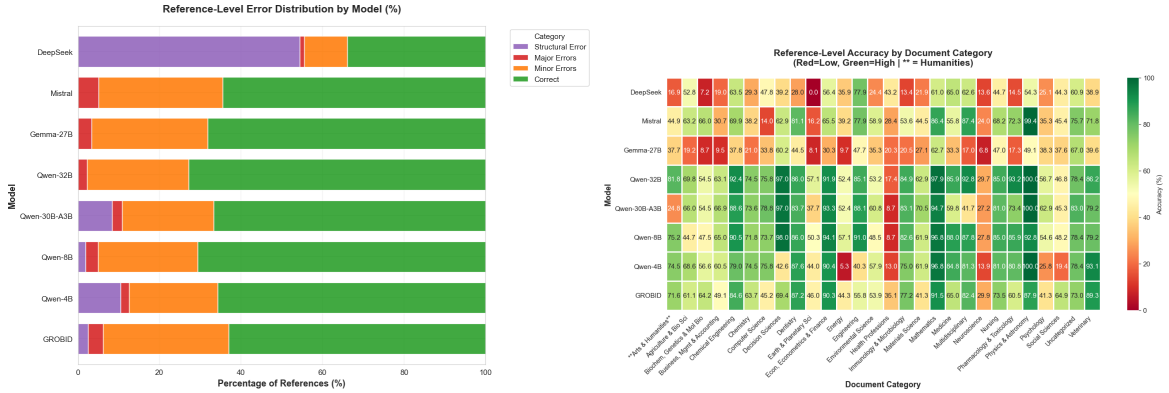

Figure 1: Reference-level error analysis on CEX dataset and end-to-end document parsing task. **(a)** Overall error distribution by model/system, showing percentage of references in each category: structural errors, major errors, minor errors, and correct matches. **(b)** Reference-level accuracy (%) by document category across models; color indicates performance (red = low, green = high). Categories marked with ** denote humanities subjects.

high fraction of *Structural* errors, which we attribute primarily to output-length limits in the API: for documents with long bibliographies, generations are frequently cut mid-output, yielding incomplete JSON that cannot be evaluated. At the same time, the stacked bars make a scaling effect visible within the Qwen3-VL line: the smallest variant Qwen3-VL-4B-Instruct and Qwen3-VL-30B-A3B-Instruct show noticeably more *Structural* failures than its larger counterparts, whereas structural errors almost disappear for Qwen3-VL-32B-Instruct. This pattern supports the main benchmark finding that below a certain capacity threshold models become unstable on long, schema-constrained outputs, even when their token-level accuracy is competitive.

Figure 1**(b)** breaks down the share of *Correct* references (our strict ≥ 0.95 threshold on all fields) by document category. A consistent trend is that all systems achieve higher accuracy on STEM-oriented categories than on SSH/humanities-marked fields (indicated with **), suggesting increased sensitivity to stylistic variation, multilinguality, and heterogeneous citation conventions in the latter. Two categories stand out as difficult across models: *Health Professions* and *Neuroscience*, which show pronounced drops in accuracy. For *Neuroscience*, this correlates with unusually long reference lists in CEX (roughly 75 references per paper, versus an overall average of $\sim$ 41), increasing the risk of truncation and accumulated formatting errors; both categories also contain highly specialised terminology and venue-specific citation patterns that appear less well covered by general instruction-tuned LLMs, leading to field confusions (e.g., identifiers mislabelled as page ranges). These category-level differences should nonetheless be interpreted with caution, as each CEX subject area is represented by only four documents, and some effects may therefore reflect sampling noise rather than stable disciplinary trends.

5.2. LoRA Fine-tuning

We investigate parameter-efficient fine-tuning, more specifically Low-Rank Adaptation (LoRA)[21], to adapt locally deployable LLMs to SSH-typical reference parsing, using two open-weight backbones: Mistral-Small-3.2-24B and Qwen3-VL-8B-Instruct. We train LoRA adapters for **reference parsing** on a supervised dataset [3] we construct from the same three datasets used in our benchmark (CEX, EXCITE, LinkedBooks). Since CEX and EXCITE provide no dedicated training corpora, we sample from their gold references and enforce document-level exclusion from evaluation to prevent leakage. To balance scale across sources, we subsample roughly 10% of the LinkedBooks training/validation material and retain only high-quality primary citations. To better reflect production inputs, we also create "grouped" instances by concatenating references that co-occur in the same gold reference block, apply our conversational templates to these blocks, and supervise the corresponding JSON lists, yielding 1,832 training conversations in total.

[3]Training dataset available at https://huggingface.co/datasets/odoma/reference-parsing-finetuning.

We train with Axolotl[22] on two A100 GPUs, freezing base weights and selecting hyper-parameters via pilot sweeps. Best settings are: Mistral-Small-3.2-24B[4] (r=32, α=64, lr 5×10^{-5}, 2 epochs) and Qwen3-VL-8B-Instruct [5] (r=64, α=128, lr 5×10^{-5}, 3 epochs). To separate parameter adaptation from prompt effects, we evaluate three inference variants per backbone and dataset (Table 4): (i) base model with our detailed few-shot prompt; (ii) LoRA model with the same detailed few-shot prompt; and (iii) LoRA model with the short, zero-shot *training-style* prompt used during fine-tuning.

Table 4
Reference parsing results for LoRA fine-tuning (GROBID as baseline). We compare (i) base models with few-shot prompting, (ii) LoRA-finetuned models with the same prompt, and (iii) LoRA-finetuned models with the zero-shot training-style prompt. Best Micro F1 per dataset and backbone in bold. (ties included).

Dataset	Model variant	P	R	Micro F1	Macro F1	# Fail	Runtime (s)
CEX	Mistral-24B (base, few-shot)	0.7422	0.7401	0.7401	0.7350	0	1569.6
	Mistral-24B (LoRA, few-shot)	0.7470	0.7420	**0.7450**	0.7426	0	1498.0
	Mistral-24B (LoRA, training-style)	0.7486	0.7420	**0.7450**	0.7548	0	1311.0
	Qwen3-VL-8B (base, few-shot)	0.7386	0.7229	0.7293	0.7197	0	762.2
	Qwen3-VL-8B (LoRA, few-shot)	0.7507	0.7398	**0.7445**	0.7475	0	841.6
	Qwen3-VL-8B (LoRA, training-style)	0.7480	0.7265	0.7358	0.7212	0	740.6
	GROBID	0.9120	0.8081	0.8560	0.8457	4	1362
EXCITE	Mistral-24B (base, few-shot)	0.8997	0.8342	0.8623	0.8411	2	1409.6
	Mistral-24B (LoRA, few-shot)	0.9310	0.8470	0.8840	0.8692	2	1543.0
	Mistral-24B (LoRA, training-style)	0.9340	0.8540	**0.8904**	0.8764	2	1523.0
	Qwen3-VL-8B (base, few-shot)	0.8870	0.8131	0.8462	0.8293	5	1886.0
	Qwen3-VL-8B (LoRA, few-shot)	0.9218	0.8039	0.8540	0.8321	0	1557.1
	Qwen3-VL-8B (LoRA, training-style)	0.9237	0.8192	**0.8613**	0.8373	0	1499.0
	GROBID	0.7688	0.6310	0.6918	0.6853	0	2083
LinkedBooks	Mistral-24B (base, few-shot)	0.6549	0.8822	0.7517	0.7473	0	131.7
	Mistral-24B (LoRA, few-shot)	0.7664	0.8879	0.8227	0.8177	0	122.0
	Mistral-24B (LoRA, training-style)	0.7768	0.8825	**0.8293**	0.8214	0	167.1
	Qwen3-VL-8B (base, few-shot)	0.6629	0.8443	0.7427	0.7284	0	62.0
	Qwen3-VL-8B (LoRA, few-shot)	0.7130	0.8936	**0.8211**	0.8296	0	68.5
	Qwen3-VL-8B (LoRA, training-style)	0.7490	0.8847	0.8112	0.8081	0	81.6
	GROBID	0.6615	0.7138	0.6866	0.6517	0	251

Results LoRA consistently improves over the base models across all datasets, with the largest gains on SSH-heavy benchmarks. On CEX, improvements are modest but stable (Mistral: +0.7% relative micro-F1; Qwen: +2.1%). On EXCITE, both models improve more substantially (Mistral: +2.5% with few-shot, +3.3% with training-style prompting; Qwen: +0.9% and +1.8%). The strongest effect is on LinkedBooks, where both backbones gain ≈+10% relative micro-F1 (Mistral: +9.4% and +10.3%; Qwen: +10.6%). Prompt–training alignment yields additional but non-uniform benefits: it helps Mistral across all datasets and Qwen on EXCITE, while Qwen on CEX and LinkedBooks still prefers the richer few-shot prompt, suggesting that detailed instructions remain useful for highly variable humanities references even after adaptation. Operationally, failures remain zero in Table 4, indicating preserved JSON validity; runtime differences mainly track prompt length, with training-style prompting generally reducing latency for Mistral.

Overall, lightweight adaptation with a small SSH supervision set (1,832 instances) substantially improves local reference parsing—especially under multilingual and style-rich distributions—and in the best cases enables more compact, schema-aligned prompting.

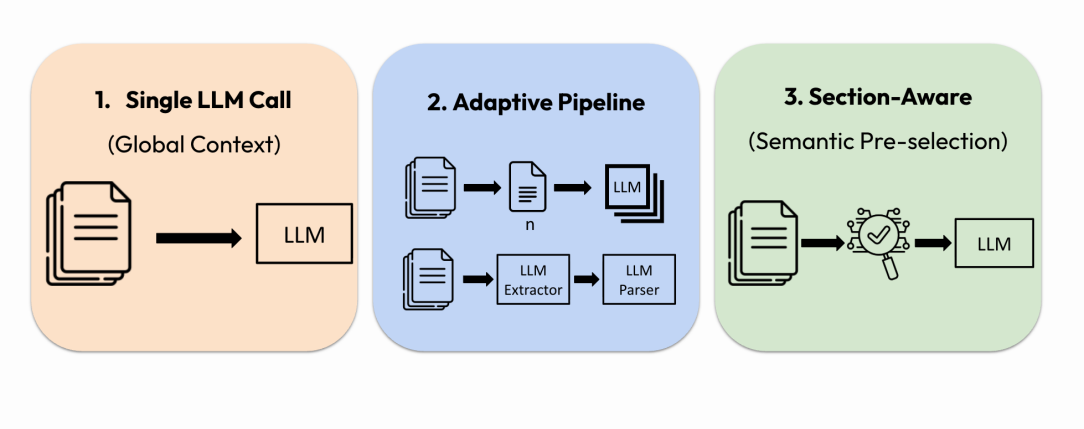

Figure 2: Segmentation strategies for reference extraction and end-to-end parsing.

5.3. Segmentation strategies

Segmentation controls how much text the model sees per request and therefore drives the trade-off between accuracy, robustness, and runtime. Rather than tuning segmentation per document, we treat it as an experimental factor and apply each strategy uniformly on CEX and EXCITE to isolate its effects under SSH-typical conditions (long documents, multilingual passages, and references in both end sections and footnotes). We compare three variants (Figure 2):

- **Single-call (global context)** sends the full document (or one long slice) in a single request when it fits the context window, preserving global coherence but becoming brittle as inputs and JSON outputs grow long.
- **Adaptive pipeline (local context)** reduces task complexity by breaking a single request into multiple smaller ones: for extraction, we apply **per-page segmentation**; for end-to-end parsing, we decompose the task into a sequential **extract** → **parse** pipeline. This keeps structured outputs small and localizes failures.
- **Semantic pre-selection (section-aware)** retrieves likely reference-bearing regions and only sends those chunks to the model, reducing token usage; we use `intfloat/multilingual-e5-large-instruct` for semantic retrieval.

Across variants, we keep the prompt and JSON schema fixed, enforce strict JSON validation, and log wall-clock runtime and invalid/empty outputs as failures.

Results Table 5 summarizes results for Mistral-Small-3.2-24B on CEX and EXCITE. For **reference extraction**, performance is high and differences are small: single-call is consistently best, the per-page adaptive setting is slightly weaker and can be slower on long documents, and semantic pre-selection largely preserves single-call accuracy while reducing runtime (notably on EXCITE).

For **end-to-end parsing**, segmentation has a much larger impact. Single-call is the least reliable, while the two-step extract→parse pipeline performs best on both datasets by keeping structured outputs short and failures local. Semantic pre-selection is fastest and improves over single-call, but can miss scattered or stylistically atypical reference regions (e.g., footnote-heavy pages), limiting recall.

Overall, segmentation is secondary for extraction but critical for end-to-end parsing: routing the model through short, verifiable units (or selectively retrieving reference-bearing regions) is key to robust structured output on SSH documents.

[4]This model is available at: https://huggingface.co/odoma/Mistral-Small-3.2-24B-LoRA-ref-parsing
[5]This model is available at: https://huggingface.co/odoma/Qwen3-VL-8B-LoRA-ref-parsing

Table 5
Segmentation comparison with Mistral-Small-3.2-24B on CEX and EXCITE. Method 2 uses per-page segmentation for extraction and a two-step extract→ parse pipeline for end-to-end parsing. Best values bolded.

Dataset	Method	Extraction					End-to-end parsing				
		P	R	F1	AvgSim	Runtime (s)	P	R	F1	M-F1	Runtime (s)
CEX	1 Single-call	**0.9906**	**0.9915**	**0.9910**	**0.9973**	1315.98	0.6356	0.6044	0.6076	0.5922	3645.20
	2 Adaptive (per-page / two-step)	0.9339	0.8958	0.9021	0.9527	2981.21	**0.7467**	**0.7321**	**0.7364**	**0.7256**	3059.09
	3 Semantic pre-selection	0.9469	0.9442	0.9446	0.9641	**1162.16**	0.7025	0.6550	0.6724	0.6506	**1672.20**
EXCITE	1 Single-call	0.8956	**0.9261**	**0.8984**	**0.9616**	4576.27	0.6149	0.5322	0.5594	0.5339	4746.59
	2 Adaptive (per-page / two-step)	**0.8967**	0.9072	0.8894	0.9553	4377.95	**0.8522**	**0.7913**	**0.8069**	**0.7700**	4960.18
	3 Semantic pre-selection	0.8576	0.8582	0.8444	0.9170	**2904.78**	0.7276	0.6193	0.6565	0.6251	**2702.19**

6. Conclusion and Future Work

This paper benchmarked bibliographic reference processing under SSH-realistic conditions, using CEX, EXCITE, and LinkedBooks across three tasks: reference extraction, reference parsing, and end-to-end document parsing. We compared a strong supervised pipeline baseline (GROBID) against contemporary LLMs under a unified, schema-constrained setup, and further studied practical deployment levers including LoRA adaptation and document segmentation strategies.

Across datasets, extraction is comparatively stable: beyond a moderate capability threshold, most models reach similarly high accuracy and differences are driven more by throughput and failure rates than by peak scores. In contrast, parsing and end-to-end parsing remain the key bottlenecks, as they require consistent structured outputs and long-range robustness under noisy layouts. Here, the results reveal a clear trade-off: GROBID is typically the fastest option and excels on documents close to its training assumptions in the end-to-end document parsing task, but it degrades more strongly under distribution shift, especially for languages beyond English and for references embedded in footnotes or mixed citation regimes, where LLMs, especially after LoRA finetuning, are markedly more robust and often yield higher utility.

These findings suggest **LLM-routing** [23] as a practical way forward to combine speed and robustness at scale. Data produced during evaluation can be used to train a small LLM to function as a router capable of dispatching the input document to the best-performing LLM or system, based on evaluation results (triplets of input document, model name, and F1-score). This approach would allow us to send to GROBID documents that match its operating envelope (e.g., English, well-structured PDFs with end-section bibliographies), while redirecting documents with known risk factors (multilinguality, dense footnotes, complex layouts) to task-adapted LLMs.

Moreover, the stronger generalization of LLMs compared to supervised systems such as GROBID makes them particularly promising for SSH corpora, where complex layouts and multilinguality are the norm rather than the exception. This is especially relevant for highly multilingual collections, where a natural direction is to *decompose* adaptation into modular components: (i) language-specific continuous pretraining for low-represented language (e.g., Latin, Polish), and (ii) task-specific LoRA adapters that capture the structured-output requirements of reference parsing. Composing these adapters separates linguistic coverage from task structure, enabling targeted updates without retraining the full system. Looking ahead, while our LoRA experiments target reference parsing, extending parameter-efficient adaptation to end-to-end document parsing remains an important next step. We will investigate LoRA for the full pipeline—layout analysis, reference extraction, and structured parsing—to improve overall reliability and reduce brittleness under SSH-specific shifts.

Acknowledgments

This work was carried out in the context of the GRAPHIA project and received funding from the European Union (grant ID: 101188018, HORIZON-INFRA-2024-TECH-01 call) and from the Swiss State Secretariat for Education, Research and Innovation (SERI). Views and opinions expressed are however

those of the author(s) only and do not necessarily reflect those of the European Union or the Agency. Neither the European Union nor the granting authority can be held responsible for them.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT in order for grammar and spelling checking, paraphrasing and rewording. After using these tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] G. Colavizza, S. Peroni, M. Romanello, The case for the Humanities Citation Index (HuCI): a citation index by the humanities, for the humanities, International Journal on Digital Libraries 24 (2022) 191–204. URL: https://doi.org/10.1007/s00799-022-00327-0. doi:10.1007/s00799-022-00327-0.

[2] A. Hosseini, B. Ghavimi, Z. Boukhers, P. Mayr, EXCITE – A Toolchain to Extract, Match and Publish Open Literature References, in: 2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL), 2019, pp. 432–433. URL: https://ieeexplore.ieee.org/document/8791194. doi:10.1109/JCDL.2019.00105.

[3] B. Birkeneder, P. Aufenvenne, C. Haase, P. Mayr, M. Steinbrink, Extracting literature references in german speaking geography - the geocite project, in: ULITE@JCDL, 2022. URL: https://api.semanticscholar.org/CorpusID:252599864.

[4] A. Cioffi, Data for Testing and Evaluating References Extraction and Parsing Tools (2022). URL: https://zenodo.org/records/6182066. doi:10.5281/zenodo.6182066, publisher: Zenodo.

[5] O. Pagnotta, CEX Project - Dataset and Gold Standard, 2024. URL: https://zenodo.org/records/10535653. doi:10.5281/zenodo.10535653.

[6] A. Garnett, The XML markup evaluation corpus, 2016. URL: https://pkp.sfu.ca/2016/04/18/the-xml-markup-evaluation-corpus/.

[7] European Commission, Open research europe: Full article corpus, Complete collection of all articles published on Open Research Europe, 2025. URL: https://open-research-europe.ec.europa.eu, accessed 19 March 2025.

[8] M. Grennan, M. Schibel, A. Collins, J. Beel, GIANT: The 1-Billion Annotated Synthetic Bibliographic-Reference-String Dataset for Deep Citation Parsing [Data], 2019. URL: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/LXQXAO. doi:10.7910/DVN/LXQXAO.

[9] G. Colavizza, M. Romanello, Annotated References in the Historiography on Venice: 19th–21st centuries, Journal of Open Humanities Data 3 (2017). URL: https://openhumanitiesdata.metajnl.com/articles/10.5334/johd.9. doi:10.5334/johd.9.

[10] D. Tkaczyk, P. Szostek, M. Fedoryszak, P. J. Dendek, L. Bolikowski, CERMINE: automatic extraction of structured metadata from scientific literature, International Journal on Document Analysis and Recognition (IJDAR) 18 (2015) 317–335. URL: https://doi.org/10.1007/s10032-015-0249-8. doi:10.1007/s10032-015-0249-8.

[11] A. Prasad, C. Si, M.-Y. Kan, Dataset Mention Extraction and Classification, in: V. Nastase, B. Roth, L. Dietz, A. McCallum (Eds.), Proceedings of the Workshop on Extracting Structured Knowledge from Scientific Publications, Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 31–36. URL: https://aclanthology.org/W19-2604/. doi:10.18653/v1/W19-2604.

[12] T. Backes, A. Iurshina, M. A. Shahid, P. Mayr, Comparing free reference extraction pipelines, International Journal on Digital Libraries 25 (2024) 841–853. URL: https://doi.org/10.1007/s00799-024-00404-6. doi:10.1007/s00799-024-00404-6.

[13] B. Ghavimi, W. Otto, P. Mayr, Exmatcher: Combining features based on reference strings and segments to enhance citation matching, 2019. URL: https://arxiv.org/abs/1906.04484. arXiv:1906.04484.

[14] N. S. Adhikari, S. Agarwal, A comparative study of pdf parsing tools across diverse document categories, 2025. URL: https://arxiv.org/abs/2410.09871. arXiv:2410.09871.

[15] P. Sarin, J. P. Alperin, Citation Parsing and Analysis with Language Models, 2025. URL: http://arxiv.org/abs/2505.15948. doi:10.48550/arXiv.2505.15948, arXiv:2505.15948 [cs].

[16] O. Pagnotta, CEX Project - trained GROBID citation models (2024). URL: https://zenodo.org/records/10529709. doi:10.5281/zenodo.10529709, publisher: Zenodo.

[17] DeepSeek-AI, A. Liu, B. Feng, B. Xue, B. Wang, B. Wu, C. Lu, C. Zhao, C. Deng, C. Zhang, C. Ruan, D. Dai, D. Guo, D. Yang, D. Chen, D. Ji, E. Li, F. Lin, F. Dai, F. Luo, G. Hao, G. Chen, G. Li, H. Zhang, H. Bao, H. Xu, H. Wang, H. Zhang, H. Ding, H. Xin, H. Gao, H. Li, H. Qu, J. L. Cai, J. Liang, J. Guo, J. Ni, J. Li, J. Wang, J. Chen, J. Chen, J. Yuan, J. Qiu, J. Li, J. Song, K. Dong, K. Hu, K. Gao, K. Guan, K. Huang, K. Yu, L. Wang, L. Zhang, L. Xu, L. Xia, L. Zhao, L. Wang, L. Zhang, M. Li, M. Wang, M. Zhang, M. Zhang, M. Tang, M. Li, N. Tian, P. Huang, P. Wang, P. Zhang, Q. Wang, Q. Zhu, Q. Chen, Q. Du, R. J. Chen, R. L. Jin, R. Ge, R. Zhang, R. Pan, R. Wang, R. Xu, R. Zhang, R. Chen, S. S. Li, S. Lu, S. Zhou, S. Chen, S. Wu, S. Ye, S. Ye, S. Ma, S. Wang, S. Zhou, S. Yu, S. Zhou, S. Pan, T. Wang, T. Yun, T. Pei, T. Sun, W. L. Xiao, W. Zeng, W. Zhao, W. An, W. Liu, W. Liang, W. Gao, W. Yu, W. Zhang, X. Q. Li, X. Jin, X. Wang, X. Bi, X. Liu, X. Wang, X. Shen, X. Chen, X. Zhang, X. Chen, X. Nie, X. Sun, X. Wang, X. Cheng, X. Liu, X. Xie, X. Liu, X. Yu, X. Song, X. Shan, X. Zhou, X. Yang, X. Li, X. Su, X. Lin, Y. K. Li, Y. Q. Wang, Y. X. Wei, Y. X. Zhu, Y. Zhang, Y. Xu, Y. Xu, Y. Huang, Y. Li, Y. Zhao, Y. Sun, Y. Li, Y. Wang, Y. Yu, Y. Zheng, Y. Zhang, Y. Shi, Y. Xiong, Y. He, Y. Tang, Y. Piao, Y. Wang, Y. Tan, Y. Ma, Y. Liu, Y. Guo, Y. Wu, Y. Ou, Y. Zhu, Y. Wang, Y. Gong, Y. Zou, Y. He, Y. Zha, Y. Xiong, Y. Ma, Y. Yan, Y. Luo, Y. You, Y. Liu, Y. Zhou, Z. F. Wu, Z. Z. Ren, Z. Ren, Z. Sha, Z. Fu, Z. Xu, Z. Huang, Z. Zhang, Z. Xie, Z. Zhang, Z. Hao, Z. Gou, Z. Ma, Z. Yan, Z. Shao, Z. Xu, Z. Wu, Z. Zhang, Z. Li, Z. Gu, Z. Zhu, Z. Liu, Z. Li, Z. Xie, Z. Song, Z. Gao, Z. Pan, DeepSeek-V3 Technical Report, 2025. URL: http://arxiv.org/abs/2412.19437. doi:10.48550/arXiv.2412.19437, arXiv:2412.19437 [cs].

[18] Mistral AI Team, Mistral small 3, 2025. URL: https://mistral.ai/news/mistral-small-3, news post.

[19] Gemma Team, T. Mesnard, C. Hardin, R. Dadashi, S. Bhupatiraju, S. Pathak, L. Sifre, M. Rivière, M. S. Kale, J. Love, P. Tafti, L. Hussenot, P. G. Sessa, A. Chowdhery, A. Roberts, A. Barua, A. Botev, A. Castro-Ros, A. Slone, A. Héliou, A. Tacchetti, A. Bulanova, A. Paterson, B. Tsai, B. Shahriari, C. L. Lan, C. A. Choquette-Choo, C. Crepy, D. Cer, D. Ippolito, D. Reid, E. Buchatskaya, E. Ni, E. Noland, G. Yan, G. Tucker, G.-C. Muraru, G. Rozhdestvenskiy, H. Michalewski, I. Tenney, I. Grishchenko, J. Austin, J. Keeling, J. Labanowski, J.-B. Lespiau, J. Stanway, J. Brennan, J. Chen, J. Ferret, J. Chiu, J. Mao-Jones, K. Lee, K. Yu, K. Millican, L. L. Sjoesund, L. Lee, L. Dixon, M. Reid, M. Mikuła, M. Wirth, M. Sharman, N. Chinaev, N. Thain, O. Bachem, O. Chang, O. Wahltinez, P. Bailey, P. Michel, P. Yotov, R. Chaabouni, R. Comanescu, R. Jana, R. Anil, R. McIlroy, R. Liu, R. Mullins, S. L. Smith, S. Borgeaud, S. Girgin, S. Douglas, S. Pandya, S. Shakeri, S. De, T. Klimenko, T. Hennigan, V. Feinberg, W. Stokowiec, Y.-h. Chen, Z. Ahmed, Z. Gong, T. Warkentin, L. Peran, M. Giang, C. Farabet, O. Vinyals, J. Dean, K. Kavukcuoglu, D. Hassabis, Z. Ghahramani, D. Eck, J. Barral, F. Pereira, E. Collins, A. Joulin, N. Fiedel, E. Senter, A. Andreev, K. Kenealy, Gemma: Open Models Based on Gemini Research and Technology, 2024. URL: https://arxiv.org/abs/2403.08295. doi:10.48550/ARXIV.2403.08295.

[20] S. Bai, Y. Cai, R. Chen, K. Chen, X. Chen, Z. Cheng, L. Deng, W. Ding, C. Gao, C. Ge, W. Ge, Z. Guo, Q. Huang, J. Huang, F. Huang, B. Hui, S. Jiang, Z. Li, M. Li, M. Li, K. Li, Z. Lin, J. Lin, X. Liu, J. Liu, C. Liu, Y. Liu, D. Liu, S. Liu, D. Lu, R. Luo, C. Lv, R. Men, L. Meng, X. Ren, X. Ren, S. Song, Y. Sun, J. Tang, J. Tu, J. Wan, P. Wang, P. Wang, Q. Wang, Y. Wang, T. Xie, Y. Xu, H. Xu, J. Xu, Z. Yang, M. Yang, J. Yang, A. Yang, B. Yu, F. Zhang, H. Zhang, X. Zhang, B. Zheng, H. Zhong, J. Zhou, F. Zhou, J. Zhou, Y. Zhu, K. Zhu, Qwen3-vl technical report, 2025. URL: https://arxiv.org/abs/2511.21631. arXiv:2511.21631.

[21] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, Lora: Low-rank adaptation of large language models, 2021. URL: https://arxiv.org/abs/2106.09685. arXiv:2106.09685.

[22] Axolotl maintainers and contributors, Axolotl: Open Source LLM Post-Training, 2023. URL: https://github.com/axolotl-ai-cloud/axolotl, original-date: 2023-04-14T04:25:47Z.

[23] A. Cooper, K. Kato, C.-H. Shih, H. Yamane, K. Vinken, K. Takemoto, T. Sunagawa, H.-W. Yeh, J. Yamanaka, I. Mason, X. Boix, Rethinking VLMs and LLMs for image classification, Scientific Reports 15 (2025) 19692. URL: https://www.nature.com/articles/s41598-025-04384-8. doi:10.1038/s41598-025-04384-8.