Extended abstract:

Evaluating LLM-based metadata extraction for CRIS with a multi-criteria framework

1. Introduction and Motivation

This work builds on research and hands-on implementation in DSpace-based CRIS environments, that included designing, building, and maintaining metadata workflows. Metadata extraction from research documents remains a persistent bottleneck. Librarians and repository administrators are required to manually create or verify records for publications, projects, and persons, and these records feed directly into institutional systems that depend on high data quality. Metadata underpins discovery and retrieval, as well as repository interoperability, including exposure via OAI-PMH and aggregation by OpenAIRE. FAIR-aligned practice presupposes consistent, well-formed metadata. Citation linking through Crossref and DataCite can break when metadata is incomplete or inconsistent. Likewise, digital preservation frameworks such as OAIS and PREMIS rely on metadata to maintain essential context over time. When metadata is incomplete or incorrect, records may be dropped from aggregators, publication to researcher links can break, and institutional reporting becomes unreliable.

Large language models can automate selected components of this work. Contemporary models can process a research article and generate structured metadata, including title, authors, affiliations, funding information, and keywords, within a single prompting step. When integrated with validation, normalization, and external registry lookups, they can form the basis of a viable extraction pipeline. Model performance, however, is not adequately represented by a single aggregate score. It depends on the model family and configuration, prompting strategy, document language, academic domain, media type, and the availability of tool-supported registry queries. In CRIS settings, performance must also be judged in terms of operational reliability. Attributes such as author affiliations, funding identifiers, and ORCID links have institutional implications, and errors can propagate into reporting, assessment, and compliance processes. For this reason, an evaluation framework should record provenance, represent uncertainty explicitly, maintain auditable decision logs, and flag higher-risk fields for targeted human review.

I propose two contributions. First, an LLM-based metadata extraction workflow for CRIS, implemented and tested in a working DSpace environment. The model performs the primary extraction by reading documents, identifying fields, and producing structured output, while validation, normalization, and linking steps are used to detect and correct errors. Second, a CRIS-oriented evaluation framework that compares multiple models across key experimental factors, measuring how closely extracted values align with authoritative reference sources using fuzzy similarity methods such as normalized word overlap, substring containment, and name-normalized Jaccard similarity for author fields. The evaluation is organized around three questions: (1) which metadata fields are reliably extractable from PDF text alone, and which require external sources regardless of the model? (2) how do model family, language, and domain interact to affect extraction quality? (3) where does repair prompting improve schema compliance, and where does it introduce new errors?

2. Related Work

Automated metadata extraction from research documents has evolved through three broad methodological phases. Early rule-based systems and classical machine learning approaches, including CRF and SVM, reported F-measures above 85–90 percent on standard bibliographic fields, but they relied on manual feature engineering and were brittle when confronted with unfamiliar layouts or publishing conventions (Kovacevic et al., 2011). Dedicated extraction pipelines such as GROBID (Lopez, 2009) combine rule-based and CRF components to parse structured bibliographic headers from PDFs with high accuracy on well-formatted documents, though they are limited to predefined fields and do not generalize to subject metadata or multilingual content. Subsequent transformer-based named entity recognition models, such as SciBERT and other domain-adapted BERT variants, improved contextual extraction, yet they remain constrained by long-document handling and typically produce labeled text spans rather than structured records suitable for CERIF-aligned workflows. More recently, zero-shot prompting with large language models has demonstrated strong performance without task-specific training. Turner et al. (2025) found that GPT-4 achieved agreement comparable to trained human annotators on neuroimaging papers (0.91–0.97), although limitations persist, including hallucinations, imperfect schema adherence, and unwarranted inference beyond the source text.

Recent studies indicate that these limitations can be reduced by embedding the model within a workflow that enforces structure and applies systematic validation. In CLINES (Yang et al., 2025), an LLM paired with post-extraction validation outperformed a single-prompt GPT-4 baseline by 0.21–0.38 F1 on clinical entity extraction and preserved accuracy on long documents where transformer-based baselines deteriorated. MOLE (Alyafeai et al., 2025) combined schema-driven prompting with validation to extract more than 30 fields from scientific papers. PARSE (Shrimal et al., 2025) further demonstrated that jointly optimizing the output schema and model behavior can reduce extraction errors by up to 65 percent. Despite these advances, the existing literature has not yet provided a unified evaluation that tests extraction performance across multiple languages and academic domains within a single experimental setting.

3. Proposed Approach

The workflow is centered on the validated extraction from PDF, it produces structured JSON that covers formal metadata (title, authors, DOI, journal, year, publisher, language, pages, ISSN) and subject metadata (abstract, keywords, domain), together with per-field confidence scores. Surrounding this core are lightweight components that handle the remaining steps: (1) PDF text extraction at ingest; (2) schema and cross-field validation of the model output, including type checks, format enforcement, and consistency rules; (3) normalization and linking against external registries such as ORCID, ROR, Crossref, and funder databases; (4) repair through targeted re-prompting, where validation failures are fed back to the model for correction; and (5) provenance logging that records source snippets, model version, and any tool calls.

The workflow has been prototyped and tested against records in our DSpace-CRIS instance. Pipeline output is compared with existing curated metadata, and no records are modified until this comparison is complete.

To compare how different models perform on the same extraction task, I developed a validation system that runs multiple times over a shared document set and evaluates their outputs against reference sources. The process begins with OpenAlex. Records are selected using filters for language, domain (via primary_topic.domain.id), and baseline constraints, including articles only and the availability of both a PDF and an abstract. For each record, the system retrieves the best landing-page URL and downloads the PDF. From the landing page it extracts HTML metatags, JSON-LD structured data, and heuristically identified keywords. This yields two reference sources per record: selected fields from the OpenAlex record and the collected landing-page metadata.

Each downloaded PDF is then processed by each LLM which generates structured metadata (including title, authors, abstract, DOI, journal, keywords, and related fields) along with per-field confidence scores. For eleven metadata fields (Title, Authors, Year, DOI, Journal, Publisher, Language, Keywords, Pages, ISSN, Abstract), the system computes similarity between the extracted value and each reference source using fuzzy matching. Identifier fields (DOI, ISSN, Year) are compared using exact match after normalization, while text fields use normalized word overlap and substring containment, author fields use name-normalized Jaccard similarity, and keyword fields use set-based Jaccard similarity. Field-level relevance is taken as the maximum score across the available reference sources, so a field is not penalized when one source is incomplete or noisy. Record-level relevance is computed as the arithmetic mean of these field-level maxima across all fields with comparable data, producing a single score in the range 0.0 to 1.0.

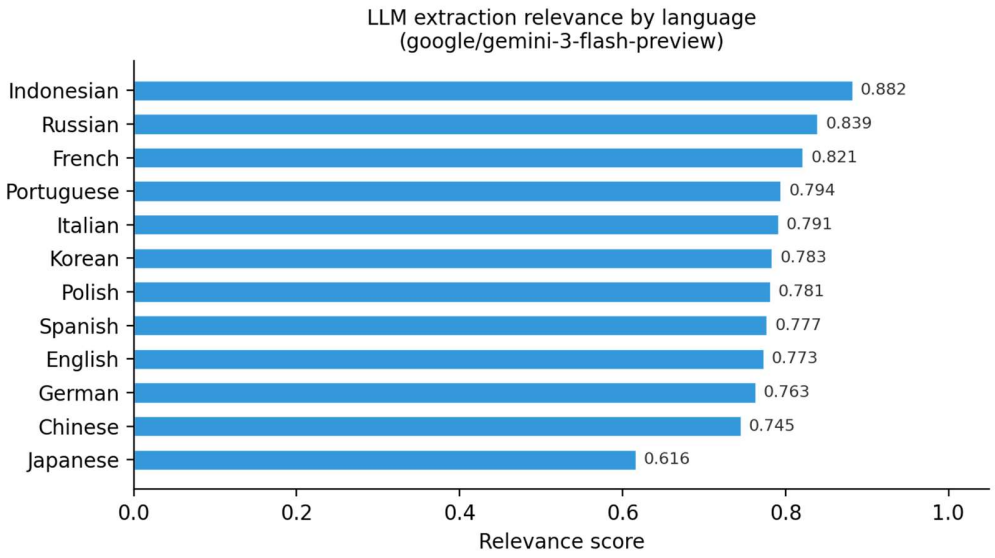

Figure 1. LLM extraction relevance by language (google/gemini-3-flash-preview).

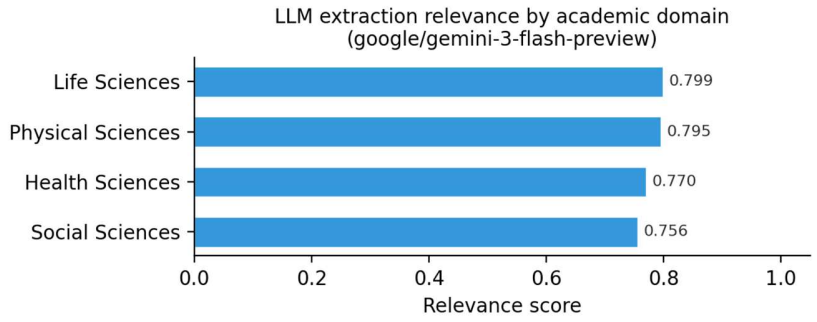

Figure 2. LLM extraction relevance by academic domain (google/gemini-3-flash-preview).

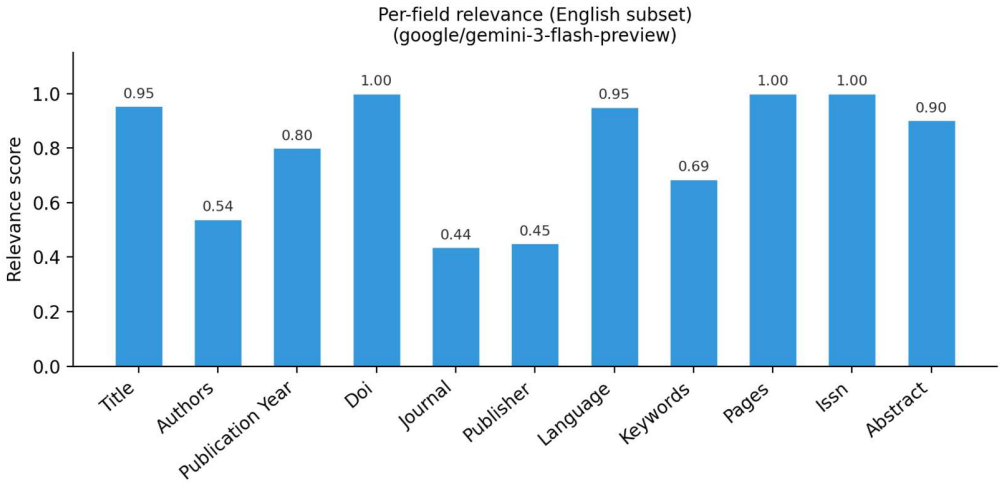

Figure 3. Per-field relevance for the English subset (google/gemini-3-flash-preview).

An important nuance is that these scores conflate two effects: the model's ability to extract metadata and the extent to which the metadata is actually present in the PDF text. Many research papers do not contain all eleven fields in their body text. A DOI may appear only on the landing page rather than in the PDF. Keywords are often omitted. Language is rarely stated explicitly. When a field is missing from the model output, the cause may be extraction failure, but it may also be that the information is not available in the document at all. The scoring approach reflects this directly. Low field scores can indicate a document-level absence rather than a model weakness. Separating these two causes, model limitation versus metadata absence, is therefore part of the evaluation and an important practical result for CRIS workflows. In this setting, the question is not only whether a model can extract a field, but whether that field is extractable from a given document type in the first place.

The evaluation framework is designed to vary multiple experimental factors, including the model backend (model family, context window capacity, and support for structured output), prompting strategy (zero-shot, one-shot, and three-shot, with and without repair prompting), document language, academic domain, and media type (born-digital PDFs, scanned or OCR-derived PDFs,

and complex-layout PDFs). The test corpus spans twelve languages (English, German, Japanese, Spanish, French, Portuguese, Chinese, Korean, Russian, Indonesian, Italian, Polish) and four academic domains (life sciences, social sciences, physical sciences, health sciences), with five publications sampled per language-by-domain cell. The study evaluates nine model backends: Anthropic Claude Sonnet 4.5, DeepSeek V3.2, Google Gemini 3 Flash Preview, Google Gemini 3 Pro Preview, Meta Llama 3.1 8B Instruct, MoonshotAI Kimi K2.5, OpenAI GPT-4o-mini, OpenAI GPT-5 nano, and OpenAI GPT-5.2. Documents are sampled from OpenAlex. Evaluation metrics include field-level relevance against reference sources, schema compliance rate, hallucination rate (fields populated without a corresponding evidence span), and linking accuracy for persistent identifiers. The screening phase is designed to identify large between-model and between-language effects rather than micro-differences; top-performing configurations will advance to a deeper evaluation on a larger document set.

Screening results already indicate meaningful differences between models on higher-variance fields such as author disambiguation and keyword generation. Some models perform strongly on formal bibliographic fields while producing less reliable keywords. Others generate higher-quality subject metadata but struggle with structured identifiers. Equally informative are cases where all models score poorly on the same field for the same document. In practice, this pattern often indicates that the relevant metadata is absent from the PDF rather than reflecting a shared model limitation. Results from a single model (Google Gemini 3 Flash Preview) illustrate the variation across dimensions. Across languages, relevance ranges from 0.616 for Japanese to 0.882 for Indonesian, with most European languages clustering between 0.76 and 0.84 and CJK languages scoring lower. Across domains, the spread is narrower, suggesting that domain has less influence on extraction quality than language. Per-field analysis on the English subset reveals that Title, Language, Abstract, and DOI are extracted reliably, while Authors, Journal, and Publisher score much lower, in part because these fields are often absent from the PDF body text and exist only in landing-page metadata or registry records. These observations directly inform the design and emphasis of the full study.

References

1. Alyafeai, Z., Al-Shaibani, M. S., & Ghanem, B. (2025). MOLE: Metadata Extraction and Validation in Scientific Papers Using LLMs. arXiv:2505.19800.
2. Kovacevic, A., Ivanovic, D., Milosavljevic, B., Konjovic, Z., & Surla, D. (2011). Automatic Extraction of Metadata from Scientific Publications for CRIS Systems. Program, 45(4), 376–396.
3. Lopez, P. (2009). GROBID: Combining Automatic Bibliographic Data Recognition and Term Extraction for Scholarship Publications. Research and Advanced Technology for Digital Libraries, LNCS 5714, 473–474.
4. Shrimal, A., Kumar, S., & Gupta, A. (2025). PARSE: LLM Driven Schema Optimization for Reliable Entity Extraction. arXiv:2510.08623.
5. Turner, M. D., Chakrabarti, C., Jones, T. B., Sheridan, J. F., & Laird, A. R. (2025). Large Language Models Can Extract Metadata for Annotation of Human Neuroimaging Publications. Frontiers in Neuroinformatics, 19, 1609077.
6. Yang, Z., Zhao, Y., Wang, Y., Liu, Z., & Sun, M. (2025). CLINES: Clinical LLM-based Information Extraction and Structuring Agent. medRxiv 2025.12.01.25341355.