Examining Accuracy in AI-Powered Metadata Extraction for Engineering Education Research

Mr. Joshua Owusu Ansah, Arizona State University

Joshua Owusu Ansah is a researcher and entrepreneur pursuing a PhD in Engineering Education Systems and Design at Arizona State University. His work explores AI-powered metadata extraction and VR-based learning tools to improve research accessibility and laboratory experiences in underfunded universities, bridging technology, education, and social impact.

Dr. Brooke Charae Coley, Arizona State University, Polytechnic Campus

Dr. Brooke Coley, an Assistant Professor of Engineering at Arizona State University, is a pioneering force in disrupting the status quo of engineering to create a more equitable and inclusive field where all individuals can thrive. As the Founding Executive Director of the Center for Research Advancing Racial Equity, Justice, and Sociotechnical Innovation Centered in Engineering (RARE JUSTICE), Dr. Coley leads transformative efforts to challenge systemic barriers and promote equity in academia. Her research focuses on amplifying the lived experiences of racially minoritized scholars, dismantling anti-Blackness in STEM, graduate student education, and fostering awareness of, and ultimately, accountability for, the lived realities of individuals navigating STEM through immersive virtual reality experiences. Collaborating with mental health experts, she also is intentional to integrate a head-on focus on the implications for wellness and wholeness in academic environments. Dr. Coley's transparent and culturally responsive approaches, coupled with her dedication and fortitude, have positioned her as a recognized leader in the field. Since 2017, she has secured millions of dollars in grant funding from the National Science Foundation, employing critical qualitative and arts-based methodologies in her work. She received the Wickenden Award and Betty Vetter Award in 2024 and was named a Virtual Visiting Scholar by the ARC Network in 2023. Launching from the Ph.D. in Bioengineering from the University of Pittsburgh oriented to the challenges of navigating STEM as an underrepresented and minoritized scholar, she continues to lead change and advocate for institutional transformation and accountability through novel applications and approaches.

Monica Lynn Miles, University at Buffalo, The State University of New York

Monica L. Miles, Ph.D. is an early career Assistant Professor of Engineering Education at the University at Buffalo in the School of Engineering and applied sciences. Dr. Miles considers herself a scholar-mother-activist-entrepreneur where all her identities work in harmony as she reshapes her community. She is a critical scholar who seeks transformative solutions to cultivate liberated and environmentally just environments for Black people, and other minoritized individuals. She believes in fostering racial solidarity and finding her own path in the movement.

Examining Accuracy in AI-Powered Metadata Extraction for Engineering Education Research

Abstract

Background: Engineering libraries are drowning in paperwork. Every time a researcher publishes a paper, someone has to record key details about it: who wrote it, where they work, and what the paper is about. As the volume of engineering education research grows, doing this by hand is no longer realistic. Many libraries are turning to automated tools to do this job, but there has been little to no studies that test how well these tools actually work or where they fail.

Purpose: This study asks a simple but urgent question: which automated tool does the best job of pulling key information from research papers, and can a smarter combination of tools do better than any single one alone?

Method: We tested four different automated approaches on 50 peer-reviewed engineering education papers, comparing each tool's output against information that was verified by humans. The four tools ranged from a free, widely used academic document reader, to a rule-based system, to a modern AI assistant, to a custom hybrid approach that assigns each task to whichever tool handles it best.

Results: The hybrid approach, AlphaMind-Extrator, outperformed all others. More importantly, the study uncovered a critical blind spot: one of the most popular free tools used in libraries today gets author affiliations, which university or department an author belongs to, wrong more than 96% of the time, while appearing reliable on everything else. The hybrid approach fixed this failure, achieving 86% accuracy on affiliations at a cost of less than two cents for 50 papers.

Conclusions: No single automated tool does everything well. The safest and most cost-effective approach is to match each task to the tool best suited for it, and to test tools field by field before trusting them with your data.

Glossary of technical terms

To support the adoption of these workflows by library professionals, the following terms and metrics used in this study are defined below.

pipeline architectures

A pipeline, in this context, is a complete automated workflow: a sequence of steps a computer follows to take a PDF as input and produce structured information (like author names and affiliations) as output. Think of it like an assembly line: raw material goes in, finished product comes out.

- GROBID (GeneRation Of Bibliographic Data): This is a machine-learning library designed for parsing and restructuring raw PDF documents into structured XML formats. In plain terms, GROBID reads a research paper PDF and identifies where the

title, authors, and other key information are located, based on how the document is laid out on the page.

- Hybrid Pipeline (AlphaMind-Extrator): A system that triages tasks, using fast, low-cost tools (GROBID) for simple fields like titles, and sophisticated tools (LLMs) only for complex fields like affiliations. Rather than applying one tool to everything, this approach assigns each task to whichever tool does it best.

performance metrics

When evaluating any automated system, we need numbers to tell us how well it worked. The following three metrics are the standard way researchers measure this.

- Precision: The accuracy of the results [5]. It measures what percentage of the extracted metadata was actually correct. High precision means fewer false positives, that is, fewer errors or hallucinations, which are cases where the system produces wrong or made-up information.
- Recall: The completeness of the results [5]. It measures what percentage of the total available metadata the system actually found. A system with high recall misses very little.
- F1 Score: The harmonic mean of Precision and Recall [5]. It provides a single balanced score to compare overall pipeline performance. A score of 1.0 is perfect; 0.0 means complete failure.
- Macro F1: Averages the F1 scores of all fields equally (e.g., Title is treated as just as important as Affiliation). This is useful for spotting hidden weak spots: a system could look strong overall but be failing on one field.
- Micro F1: Calculates the score based on the total number of correct instances across all papers. This reflects the overall volume of work done correctly.
- Coverage: The percentage of the total dataset for which the pipeline generated any output at all. A pipeline that crashes or fails to process a paper scores that paper as zero, which lowers coverage.
- Gold Standard: A ground truth dataset verified by a human expert, used to grade the AI's performance. It is the answer key against which all automated outputs are compared.

extraction concepts

- Field-Selective Extraction: The process of choosing a specific tool for a specific metadata field (e.g., only using the LLM for the Affiliation field). Rather than one-size-fits-all automation, this approach matches tools to tasks based on evidence.
- JSON (JavaScript Object Notation): A standardized, machine-readable text format used to organize data (e.g., {"Author": "Smith", "Year": "2026"}). When the AI model is asked to extract metadata, it is instructed to return the results in this structured format so they can be processed automatically by other systems.
- Schema-Guided Extraction: Providing the AI with a strict blueprint or template to ensure the output matches the library's database requirements. Instead of letting the AI respond however it likes, it is given a specific structure it must follow.

Introduction and problem statement

Engineering libraries operate at the intersection of subject expertise and metadata stewardship. In addition to traditional collection development and instruction roles, engineering librarians frequently support institutional repository workflows, faculty profile systems, research impact reporting, and bibliometric analyses; these services rely on structured metadata extracted from scholarly publications, including accurate author identification and institutional affiliations [1],[2]. When metadata are incomplete or inconsistent, downstream services such as authority control, institutional analytics, and collaboration mapping degrade in reliability [1],[2].

Prior work has established strong approaches for bibliographic parsing from scholarly PDFs, including machine learning driven systems such as GROBID that are designed to extract and structure bibliographic metadata fields from article PDFs [3]. Meanwhile, large language models have expanded the design space for flexible information extraction, including schema guided extraction into structured formats, but their cost and runtime can be prohibitive when used for all fields [4]. Despite these advances, prior studies have largely emphasized the application and potential of these systems, with limited systematic, field-level evaluation of their performance. Existing literature in library cataloging and metadata research highlights gaps in empirical performance benchmarks and calls for structured evaluation frameworks to guide implementation in practice [10]. As a result, libraries risk deploying automation strategies that demonstrate strong aggregate performance while producing critical errors in high-impact fields such as author affiliations.

This paper addresses a practical operational question for engineering librarians and repository managers: how should automated metadata extraction tools be evaluated before integration into library workflows? Although tools exist for automated parsing of scholarly PDFs, their performance varies across metadata fields and document layouts. Without field-level validation, libraries risk deploying systems that perform well on visible fields such as titles and author names while failing on high-impact elements such as affiliations, which are central to institutional reporting and bibliometric services.

Research questions

This study is organized around three primary research questions:

RQ1: Which pipeline yields the highest accuracy for extracting key metadata fields from engineering education PDFs, as measured by precision, recall, and F1 against a human coded gold standard?

RQ2: How do pipeline strengths differ across structured fields (for example, title, author identification, author position, total number of authors, abstract) versus affiliation extraction?

RQ3: Does a field selective hybrid design (AlphaMind-Extrator) improve end to end performance while maintaining coverage and reducing time and cost relative to an LLM only pipeline?

Contributions

This study contributes:

1. A benchmark comparison of four pipelines for engineering education metadata extraction against a human coded gold standard of 50 papers.
2. Field level evidence showing where each approach performs best, with a clear failure mode for affiliation extraction in GROBID only workflows.
3. A field selective hybrid design, AlphaMind-Extrator, that preserves structured field strengths of GROBID while using an LLM only where needed, improving end to end macro and micro performance while reducing time and cost relative to LLM only extraction.

Related work and conceptual framing

Metadata quality and engineering library workflows

Metadata quality is central to repository services, including discovery, provenance, and authority control. Studies of metadata quality control practices in digital repositories emphasize accuracy and consistency as recurring concerns and motivate systematic quality assurance mechanisms [1]. In this framing, automated extraction systems should be evaluated not only by aggregate accuracy but also by field specific reliability because downstream tasks often fail when a single field (for example, affiliation) is incorrect.

Library and information science scholarship published in the last several years, particularly studies from 2024 to 2025, has examined artificial intelligence applications in metadata creation, enrichment, interoperability, and access within academic libraries, highlighting both the opportunities for workflow automation and the need for careful evaluation before operational adoption [9],[10]. Scoping reviews of AI's role in metadata contexts highlight the promise of automation for scaling descriptive practices, while also underscoring persistent issues related to quality control, transparency, and responsible evaluation before operational adoption [9]. Literature on AI in library cataloging critically identifies patterns in AI application, notes gaps in empirical evidence for performance benchmarks, and calls for systematic evaluation frameworks to support metadata stewardship in library environments [10]. By comparing extraction pipelines and reporting field-level outcomes, the present study contributes grounded evidence to this broader conversation about responsible AI adoption in library cataloging and metadata workflows.

Automated extraction from scholarly PDFs

GROBID is a widely used system for extracting bibliographic data from scholarly PDFs, representing an established approach for structured metadata extraction at scale [3]. However, scholarly PDFs vary in layout and formatting across publishers. For example, journals differ in how author affiliations are presented, including superscript numbering systems, symbol

based mappings, multi line institutional addresses, and placement of affiliations in footnotes or separate blocks. These variations in document structure and affiliation formatting can produce systematic extraction failures that are not visible when evaluating only a subset of easier fields.

LLMs for information extraction and structured outputs

Recent scholarship shows that LLMs can be adapted to structured extraction by specifying schemas, including JavaScript Object Notation (JSON) style outputs, and that instruction-following models can support on-demand extraction tasks [4]. While these approaches offer strong performance across diverse and unstructured metadata fields, they introduce higher computational cost, latency, and operational overhead. In contrast, traditional machine learning systems such as GROBID provide efficient, low-cost extraction but exhibit systematic limitations in complex fields such as affiliations. This creates a practical tradeoff between accuracy, cost, and field-level reliability. To address this, we evaluate whether traditional ML, LLM-based, or field-selective hybrid approaches provide the most cost-effective and accurate solution for engineering library workflows.

Evaluation framing

To evaluate the performance of each pipeline, we employ Precision, Recall, and the F1 Score, standard metrics in information retrieval that offer specific insights into metadata quality [5]. In a library context, Precision reflects the trustworthiness of the data (minimizing the need for manual corrections), while Recall indicates the completeness of the repository (minimizing missing records).

We report these metrics using both Macro and Micro aggregation to provide a multi-dimensional view of system reliability:

- Macro-averaging treats every metadata field (e.g., Title, Affiliation, DOI) as equally important. This is critical for detecting blind spots in a pipeline: a system that excels at titles but consistently fails at institutional affiliations could score well on average while failing on the field that matters most for your workflow.
- Micro-averaging aggregates all individual extraction instances into a single score. This reflects the overall workload for a cataloger, representing the cumulative accuracy across the entire dataset regardless of field type [5].

By using this dual-reporting approach, we ensure that the pipeline's design is not only accurate on average but also robust across specific fields most vital to bibliometric reporting and authority control.

Methods

Research design overview

In this comparative study, we compare multiple extraction pipelines on a shared gold standard dataset and report macro and micro precision, recall, and F1, along with coverage and optional throughput proxies (time and estimated cost where available).

Dataset

The dataset consists of 50 peer reviewed engineering education research papers collected in PDF format, with a link to a folder containing all the papers provided in the appendix. Papers were selected from credible scholarly venues and databases, including the Journal of Engineering Education (JEE), IEEE, and ACM, ensuring consistency in peer review standards and publication quality.

All documents were obtained as publisher version PDFs, rather than author manuscripts or scanned copies. This choice reflects the file types commonly encountered in engineering library workflows and reduces variability introduced by formatting inconsistencies or optical character recognition errors. OCR (Optical Character Recognition) refers to the process of converting scanned images of text into machine-readable text, a potential source of errors that publisher PDFs largely avoid.

Golden standard dataset creation

To establish a benchmark for accuracy, a verified gold standard dataset was constructed. This dataset comprises 50 peer-reviewed engineering education papers representing a diverse cross-section of layout styles from major publishers (e.g., IEEE, JEE, ACM).

Metadata for each paper was extracted directly from publisher-provided PDFs to ensure ground-truth accuracy. A second member of the research team independently verified all extracted metadata against the original PDFs, resolving any discrepancies through consensus to achieve 100% agreement. The extraction process adhered to standard library cataloging principles and prioritized fields critical for institutional repository ingest and bibliometric authority control.

The following metadata fields were included:

- Title: The full title of the scholarly work.
- Target Author: The specific faculty or researcher entity used as the primary search key.
- Found Author Name: The raw string identified by the pipeline, used to evaluate entity matching accuracy against the target author.
- Author's Position and Total Authors: Essential metrics for calculating contribution weight in faculty impact reporting.
- Affiliation: The institutional and departmental data, which serves as the primary pivot for collaboration mapping and institutional analytics. In simpler terms: which university or organization the author belongs to.
- Abstract: The summary text, vital for repository discovery and keyword indexing.

Evaluation was conducted on a per-record basis. To ensure a rigorous test of the pipelines, the gold standard includes cases with complex failure modes, such as multi-institution affiliations and non-standard footnote layouts. Evaluation was conditional on the pipeline's ability to successfully link a PDF to its corresponding gold-standard entry, thereby measuring both extraction accuracy and record-linkage reliability.

Pipelines evaluated

Four pipelines were selected to represent the major approaches used for automated metadata extraction from scholarly PDFs: a structured machine-learning system (GROBID), a classical rule-based NLP/ML pipeline, a large language model (LLM) pipeline, and a field-selective hybrid pipeline. These pipelines were chosen to enable a direct comparison between traditional document-structure methods, lightweight heuristic extraction, and modern LLM-based approaches, while also evaluating whether a hybrid architecture can combine their strengths to improve accuracy, cost, and runtime performance.

GROBID pipeline

The GROBID pipeline uses the GROBID server to extract structured bibliographic metadata from PDFs through a local REST API. Each PDF is sent to the server, which returns machine readable outputs in Text Encoding Initiative (TEI) XML or, when needed, BibTeX format. TEI XML is a standardized format for encoding text in a way that preserves its structure: it can be thought of as a filing system where every piece of information has a labeled slot. The pipeline prioritizes TEI XML because it preserves document structure and enables extraction of fields such as title, authors, affiliations, abstract, journal information, year, and DOI. To improve efficiency, a pre-filtering step scans the first three pages of each PDF for target author names and skips documents without a match. This scope was determined through preliminary analysis, which confirmed that all relevant authorship information was consistently located within the first three pages across the corpus.

GROBID was selected as the primary structured extraction baseline because prior large-scale evaluations of scholarly document parsing systems have shown it to achieve state-of-the-art performance for extracting bibliographic metadata from scientific PDFs, outperforming alternative tools including CERMINE, Science Parse, and Adobe Extract [12]. Its document-structure-aware machine learning architecture makes it effective for stable fields such as titles, authors, and abstracts.

Returned GROBID outputs are parsed using XML and text based routines that map authors to their affiliations through multiple matching strategies, including direct links in the XML, name normalization, and fallback rules when structure is incomplete. If the GROBID server is unavailable or parsing fails, the pipeline falls back to lightweight PDF text extraction using regular expressions to recover core fields. The GROBID pipeline provides high accuracy for structured fields such as titles, authors, and abstracts, runs locally without application programming interface (API) costs, and processes documents efficiently. However, affiliation extraction remains a consistent weakness, motivating the use of hybrid approaches in later pipelines.

Classical NLP/ML pipeline

A classical NLP pipeline was evaluated and reported as NLP/ML. NLP stands for Natural Language Processing: a field of computer science focused on enabling computers to understand and work with human language [7]. This pipeline runs entirely locally, without the use of external APIs, and combines rule-based heuristics, regular expressions, and lightweight natural language processing to extract metadata from PDF documents. In simple terms, it uses a set of hand-crafted rules, such as: if a line contains the word University, it is probably an affiliation, rather than learned intelligence. Such approaches are commonly used in information extraction tasks where structured signals can be inferred from textual patterns and domain-specific rules, particularly when cost, interpretability, and offline execution are important considerations [7].

The workflow begins with a pre-filtering step that scans the first three pages of each PDF for target author name patterns in multiple formats, allowing the system to skip documents that are unlikely to contain a relevant match. Text is extracted using PyMuPDF, with pdfminer used as a fallback when needed. Both are software libraries that convert PDFs into raw text that can then be analyzed.

Metadata extraction is performed using task-specific heuristics. Author names are identified through a combination of byline parsing rules and named entity recognition using spaCy, with results merged and deduplicated. Named Entity Recognition (NER) is a technique that automatically identifies and classifies names of people, organizations, and locations in text. Titles and abstracts are extracted using layout and section-header patterns, while affiliations are inferred using keyword matching and organizational entity detection. The pipeline is fast, cost-free, and capable of offline execution, completing processing for the full dataset in approximately 12 seconds. However, because it relies on heuristic rules and a small general-purpose NLP model, its accuracy is lower than the GROBID and LLM pipelines, particularly for affiliation extraction. Further details on the heuristics are provided in Appendix A.

LLM pipeline

An LLM pipeline was evaluated as a standalone metadata extractor using the OpenAI gpt-4o-mini model accessed through the Chat Completions API. GPT-4o-mini is a large language model: the same family of AI technology that powers widely used AI assistants, capable of reading and understanding complex text with human-like flexibility. GPT-4o-mini model was selected because it provides a practical balance between accuracy, cost, and runtime efficiency. Compared to larger models, GPT-4o-mini enables high-quality schema-guided outputs at lower computational cost, making it suitable for large-scale metadata extraction workflows. The pipeline uses a structured, instruction-based prompt that defines the extraction task and requires outputs in a strict JSON format. Each prompt includes targeted text from the PDF, including the author byline section, the first three pages of the document. Model decoding is fully deterministic with a temperature of zero, ensuring consistent outputs across runs. A temperature of zero means the model always produces the same output for the same input, with no randomness, which is important for reproducible research.

To ensure robustness, the pipeline applies multi-stage JSON parsing and validation to recover structured fields even when responses deviate from the expected format. The system does not rely on repeated API retries; instead, it uses a tiered fallback strategy that switches API methods or reverts to rule-based extraction when necessary. This design prioritizes correctness and structured output quality over speed and cost efficiency. As a result, the LLM pipeline achieves strong performance on both structured bibliographic fields and affiliation extraction, but with lower coverage and higher runtime and cost than GROBID and the hybrid AlphaMind-Extrator pipeline.

AlphaMind-Extrator hybrid pipeline

AlphaMind-Extrator is a field-selective hybrid extraction pipeline designed to combine the strengths of GROBID and large language model (LLM) extraction while minimizing their respective weaknesses. Rather than applying one tool to all tasks and accepting its weaknesses, this pipeline uses empirical performance evidence, that is, actual measured results, to decide which tool is responsible for each metadata field. GROBID is used for structured fields where it performs reliably, including title, author identification, author position, total number of authors, and abstract. The LLM is used only for affiliation extraction, which was identified as GROBID's primary failure mode in benchmarking results.

This targeted allocation allows AlphaMind-Extrator to preserve accuracy on well-structured fields while addressing complex affiliation formatting.

To improve efficiency, AlphaMind-Extrator runs the GROBID and LLM components in parallel, simultaneously rather than one after the other, reducing total runtime compared to sequential execution. The LLM is prompted with a minimal, affiliation-only request, reducing token usage and cost. Tokens are the units of text that AI models process; fewer tokens means lower cost. Results from both components are then fused by aligning LLM-extracted affiliations to GROBID-extracted authors using a robust author-matching procedure that accounts for name variations. This design yields the highest overall accuracy across all evaluated pipelines, while remaining faster and less costly than an LLM-only approach and more accurate than any single-method pipeline.

Evaluation metrics

Accuracy is measured using precision, recall, and F1 with macro and micro aggregation. Per-field metrics: TP (True Positive: a correctly extracted piece of information), FP (False Positive: something extracted incorrectly), FN (False Negative: something that existed but was missed) per field; precision = TP/(TP+FP), recall = TP/(TP+FN), F1 = 2 x precision x recall/(precision+recall) [5]. Macro metrics average these per-field values across all fields (mean precision, mean recall, mean F1). Micro metrics aggregate first: sum TP, FP, FN across all fields and all instances, then compute precision, recall, and F1 once on the aggregated totals. Field correctness uses similarity thresholds (0.90-0.93 per field): numeric fields (Author's Position, Total Authors) require exact matches; text fields use normalized text similarity via SequenceMatcher. A matched record with missing/empty predicted values is counted as FN for that field; similarity below threshold is FP; similarity at/above threshold is TP.

Coverage is defined as the proportion of gold-standard papers linked to predictions. A paper is found if record linkage succeeds via either (1) exact File Path match (after normalization and resolution), (2) filename match (basename only, case-insensitive), or (3) Title similarity match (normalized text similarity >= 0.93 threshold) as a fallback. A paper is not found when linkage fails: File Path matching fails, filename matching fails, and Title similarity falls below the 0.93 threshold; in this case the mapping is set to None and all fields are counted as false negatives. Coverage = (number of found papers) / (total gold-standard papers), where found = count of non-None mappings from the linkage process.

Validity, reliability, and trustworthiness considerations

Construct validity: Precision, recall, and F1 score are appropriate measures for evaluating metadata extraction accuracy because they directly capture how often extracted fields match the gold standard. To ensure meaningful interpretation, all pipelines were evaluated using the same field definitions and matching rules for each metadata element.

Internal validity: All extraction pipelines were evaluated on the same set of 50 papers using a single gold standard, which minimizes selection bias and allows for direct comparison across systems. Differences in reported performance therefore reflect true differences in extraction behavior rather than variation in the evaluation dataset. Variations in coverage indicate that some pipelines failed to return outputs for certain papers, which is treated explicitly in the evaluation.

External validity: The dataset consists of 50 peer reviewed engineering education papers drawn from multiple reputable scholarly venues, including ACM, the Journal of Engineering Education (JEE), and IEEE. While this selection captures realistic variation in formatting and publication practices, the limited dataset size means that results should be generalized cautiously to other disciplines, publication years, or PDF types not represented in the sample.

Results

TABLE I: Overall macro and micro performance by pipeline

This table answers RQ1: which pipeline performs best overall? Each row represents one of the four tools tested. Each column represents a different way of measuring performance. Higher numbers are always better, with 1.0 being a perfect score.

Pipeline	Coverage	Macro Prec.	Macro Rec.	Macro F1	Micro Prec.	Micro Rec.	Micro F1
GROBID	0.923	0.833	0.815	0.813	0.833	0.906	0.868
AlphaMind-Extrator	**0.923**	**0.946**	**0.916**	**0.930**	**0.946**	**0.916**	**0.931**
LLM	0.885	0.960	0.880	0.917	0.960	0.880	0.917
NLP/ML	0.500	0.280	0.153	0.196	0.311	0.203	0.246

Table I: Overall Macro and Micro Performance by Pipeline. Scores range from 0 to 1, where 1.0 is perfect. Coverage represents the percentage of the 50 papers processed. Time and cost values represent totals for the full dataset of 50 PDFs.

RQ1 (overall best): Table I shows clear differences in performance across the evaluated pipelines. The hybrid AlphaMind-Extrator pipeline achieved the strongest and most balanced performance overall, with the highest Macro F1 (0.930) and Micro F1 (0.931). The LLM pipeline demonstrated high precision but slightly lower recall and coverage, while GROBID provided stable performance but lower overall accuracy. In contrast, the classical NLP/ML pipeline performed worse across all metrics, indicating that rule-based approaches alone struggle to extract metadata from scholarly PDFs.

What this means in practice: AlphaMind-Extrator scored 0.930 out of 1.0 overall, meaning it correctly extracted roughly 93 in every 100 pieces of metadata. GROBID scored 0.813, and the AI-only (LLM) approach scored 0.917. The rule-based NLP/ML approach, at 0.196, was too inaccurate to be useful in a real library workflow. The hybrid approach wins not just on accuracy but also on cost and speed, as shown in Table III.

TABLE II: Field level performance by pipeline

This table answers RQ2: do pipelines perform differently depending on which field is being extracted? This is the most important table in the study. An overall score can hide critical failures, and this table reveals exactly where each tool breaks down. Pay particular attention to the highlighted row for Affiliation.

Field	Metric	GROBID	AlphaMind	LLM	NLP/ML	Note
Title	Precision	1.000	1.000	1.000	0.000	

Field	Metric	GROBID	AlphaMind	LLM	NLP/ML	Note
	Recall	0.923	0.923	0.885	0.000	
	F1	0.960	0.960	0.939	0.000	
Target Author	Precision	1.000	1.000	1.000	0.923	
	Recall	0.923	0.923	0.885	0.480	
	F1	0.960	0.960	0.939	0.632	
Found Author Name	Precision	1.000	1.000	1.000	0.923	
	Recall	0.923	0.923	0.885	0.480	
	F1	0.960	0.960	0.939	0.632	
Author Position	Precision	0.979	0.979	1.000	0.077	
	Recall	0.922	0.922	0.885	0.071	
	F1	0.950	0.950	0.939	0.074	
Total Authors	Precision	0.917	0.917	1.000	0.038	
	Recall	0.917	0.917	0.885	0.037	
	F1	0.917	0.917	0.939	0.038	
Affiliation	Precision	0.021	0.812	0.826	0.000	Critical failure in GROBID-only; hybrid resolves this gap
	Recall	0.200	0.907	0.864	0.000	
	F1	0.038	0.857	0.844	0.000	
Abstract	Precision	0.915	0.915	0.891	0.000	
	Recall	0.896	0.896	0.872	0.000	
	F1	0.905	0.905	0.882	0.000	

Table II: Field-Level Performance by Pipeline. F1 scores range from 0 to 1. The highlighted rows show the Affiliation field, the most significant failure mode identified in this study. A score of 0.038 for GROBID on Affiliation means it correctly extracted institutional affiliations less than 4% of the time.

RQ2 (structured vs affiliation): For structured fields (title, author identification and matching, author position, total authors, abstract), GROBID and AlphaMind-Extrator are consistently strongest (F1 approximately 0.905 to 0.960). Affiliation is the main exception: GROBID alone performs poorly (F1 0.038), while AlphaMind-Extrator improves affiliation to 0.857 by delegating affiliation extraction to the LLM and aligning authors. Although GROBID appears strong when looking at the other fields, this sharp drop in affiliation performance illustrates how aggregate performance can conceal critical failures in specific metadata fields that are essential for institutional reporting and bibliometric workflows.

Why this finding matters: A library that deploys GROBID alone, which is a common practice, would be getting titles and author names right 96% of the time while getting institutional affiliations right less than 4% of the time. That means nearly every affiliation in the repository would be wrong or missing. Since affiliations are the foundation of collaboration mapping, institutional reporting, and research impact analytics, this is not a

minor problem; it is a silent data quality crisis. The hybrid approach fixes this at minimal extra cost.

TABLE III: Time and estimated cost by pipeline

This table answers RQ3: does the hybrid design improve efficiency as well as accuracy? For libraries operating under budget and staffing constraints, cost and processing time are real decision factors alongside accuracy.

Pipeline	Time (seconds)	Cost (USD)
GROBID	31.252	0.00000
NLP/ML	11.958	0.00000
LLM	318.826	0.04450
AlphaMind-Extrator	**112.385**	**0.01335**

Table III: Time and Estimated Cost by Pipeline. All figures represent totals for processing the full dataset of 50 PDFs. GROBID and NLP/ML run entirely locally with no API fees. LLM and AlphaMind-Extrator use external AI APIs, which incur a per-use cost.

RQ3 (efficiency): AlphaMind-Extrator reduces time relative to LLM only extraction (112.385 vs 318.826 seconds) and reduces estimated cost (0.01335 vs 0.04450 USD), while improving macro performance. In other words, the hybrid approach is not just more accurate than using AI alone: it is also nearly three times faster and costs about one-third as much.

Putting cost in context: Processing 50 papers with the hybrid approach cost $0.013, less than two cents. Scaling that up, 5,000 papers would cost approximately $1.34. This makes the hybrid approach affordable even for libraries with constrained budgets, particularly compared to manual cataloging or full AI-only extraction.

Statistical interpretation of performance differences

The reported performance values represent descriptive comparisons across pipelines rather than formal inferential statistical tests. Because the evaluation dataset consists of 50 papers and the current reporting aggregates field-level true positives, false positives, and false negatives without retaining per-paper outcome distributions, formal paired significance testing was not conducted in this iteration.

However, the magnitude of observed differences provides evidence of practical significance. For example, the difference in macro F1 between AlphaMind-Extrator (0.930) and the LLM pipeline (0.917) reflects a consistent performance advantage across fields, driven by improvements in affiliation extraction. More substantively, the difference between AlphaMind-Extrator (0.930) and GROBID (0.813) represents a large absolute effect (delta = 0.117), corresponding to a meaningful improvement in correctly extracted metadata instances across the dataset.

Field-level analysis further strengthens this interpretation. Affiliation F1 improves from 0.038 in the GROBID-only pipeline to 0.857 in the hybrid AlphaMind-Extrator design, a difference of 0.819. Given that affiliation extraction is central to institutional reporting and bibliometric applications, this magnitude of improvement is operationally consequential even without formal inferential testing.

Future work will incorporate per-paper outcome retention and bootstrap-based confidence interval estimation to formally assess statistical significance across pipelines. Expanding the dataset will also enable paired testing approaches such as McNemar-style comparisons for binary field correctness and confidence interval estimation for F1 scores.

Discussion

What the data show

The results demonstrate that a field-selective hybrid strategy outperforms single-method approaches for metadata extraction in engineering education research. Across both macro and micro metrics, AlphaMind-Extrator achieved the strongest overall performance while maintaining high coverage, confirming that combining extractors based on observed strengths is more effective than relying on any one method alone. In contrast, pipelines that applied a single extraction strategy across all fields exhibited clear performance limitations.

Performance differences across pipelines were field dependent. Structured bibliographic elements such as titles, author identification, author position, total number of authors, and abstracts showed near-ceiling accuracy for both GROBID and AlphaMind-Extrator. Affiliation extraction, however, emerged as a dominant failure mode for GROBID when used alone, despite its strong performance on other fields. This difference arises because structured fields appear in predictable positions and formats across PDFs, whereas affiliation information varies in formatting and layout across journals. While the LLM pipeline performed well on affiliations, it incurred higher time and cost. AlphaMind-Extrator mitigates this tradeoff by limiting LLM usage to affiliation extraction, achieving high accuracy while controlling runtime and expense.

Interpretation for engineering education scholarship and bibliometrics

Reliable metadata underpins engineering education bibliometrics, including analyses of scholarly communities, collaboration networks, and institutional contributions. Prior work has shown how large-scale metadata enables longitudinal mapping of research trajectories in venues such as the Journal of Engineering Education [6]. These analyses, however, depend on accurate author and affiliation information. The present results show that affiliation is the field where unvalidated automation poses the greatest risk: a pipeline may perform well on titles and author names while failing on affiliations.

For engineering librarians and research administrators, this finding has direct operational implications. Deploying automated extraction without field-level validation can introduce quiet errors that propagate through downstream services, particularly authority control, institutional reporting, and impact assessment workflows (e.g., incorrect institutions extracted for authors, partial institutional names captured, or affiliations not extracted at all) [2]. Consistent with prior metadata quality research, these results reinforce the need for transparent benchmarking and targeted validation before deployment, as well as ongoing quality monitoring after automated tools are integrated into production library systems [1].

Why rule-based and structured extractors can outperform here

The observed performance patterns reflect underlying differences in structural regularity across metadata fields. Titles, author lists, and abstracts tend to appear in predictable

locations and formats within scholarly PDFs, making them well suited to systems optimized for document structure, such as GROBID [3]. Affiliation data, by contrast, are often encoded through layout-specific conventions including superscripts, multi-column bylines, footnotes, and many-to-many author-affiliation mappings. This study shows that affiliation features routinely break deterministic parsing strategies: tools following fixed rules fail when they encounter formats they were not designed to handle.

Affiliation extraction is the primary source of error in GROBID-only workflows and the key driver motivating hybridization. By delegating affiliation extraction to a more flexible model while preserving structured extraction for stable fields, AlphaMind-Extrator capitalizes on the complementary strengths of different approaches.

Human-AI collaboration and a threshold of trust

Beyond identifying a top-performing pipeline, this study contributes a workflow-oriented conception of trust in automated scholarship. Rather than treating automation as an all-or-nothing decision, either fully automated or fully manual, the results support a model in which automation is applied selectively based on demonstrated reliability. Fields with stable accuracy can be trusted with minimal oversight, while fields with higher risk, such as affiliations, warrant either targeted automation or focused human validation.

This perspective aligns with emerging work on LLM-based information extraction that emphasizes schema-constrained outputs and task-specific deployment over unconstrained generation [7]. In engineering library practice, such an approach supports effective human-AI collaboration by concentrating validation effort where it is most needed, improving both efficiency and trustworthiness in large-scale metadata workflows.

Implications for engineering librarianship practice

The findings of this study have direct implications for engineering librarians responsible for repository ingest, metadata enrichment, and research analytics. First, the results demonstrate that aggregate accuracy metrics alone are insufficient for workflow decisions. A pipeline may achieve strong overall performance while failing on specific high-impact fields such as affiliations. Engineering librarians evaluating automated tools should therefore examine field-level extraction performance when comparing systems prior to deployment.

Second, the hybrid strategy evaluated here provides a practical model for selective automation. Structured extractors such as GROBID may be appropriate for stable fields including titles, author names, and abstracts. However, affiliation extraction warrants targeted validation or hybridization due to its variability and institutional importance. Libraries implementing automated ingest pipelines may adopt a tiered validation approach in which high-risk fields receive additional verification.

Third, the cost and runtime comparisons illustrate that full LLM-based extraction may not be necessary for most engineering library workflows. A field-selective deployment strategy can reduce operational expense while maintaining or improving accuracy. For engineering librarians operating under constrained budgets and staffing, this evidence supports informed decision-making regarding tool adoption, local infrastructure investment, and validation thresholds.

By situating benchmarking within real engineering library use cases, this study reframes automated metadata extraction not as a purely computational problem but as a workflow design decision that directly affects repository quality, research impact services, and institutional reporting accuracy.

limitations

This study has several limitations that should be considered when interpreting the results.

First, the gold dataset consists of 50 peer reviewed engineering education papers drawn from multiple reputable venues, including ACM, the Journal of Engineering Education (JEE), and IEEE. While this selection captures realistic variation in publication formats and metadata conventions, the dataset size limits generalization to other disciplines, publication years, or PDF types not represented in the sample.

Second, pipeline performance reflects the specific implementations evaluated in this study, including particular configurations of GROBID, spaCy based NLP heuristics, and a single LLM model. Different parameter settings, model versions, or document preprocessing choices may yield different performance profiles, especially for fields such as affiliation that are sensitive to layout variation.

Third, this study did not evaluate performance across multiple Large Language Models (e.g., Claude, Llama, or Gemini). While the specific precision and recall values may vary depending on the underlying model's reasoning capabilities, the relative performance gains observed in the AlphaMind-Extrator hybrid design are expected to be model-agnostic: the architectural advantage of the hybrid approach should hold regardless of which AI model is used. The core contribution of this work lies in the architectural strategy, that is, delegating complex affiliation fields to an LLM while using GROBID for structured bibliographic data, rather than the specific performance of a single proprietary model. Future research could explore whether open-source models provide a similar accuracy-to-cost ratio for engineering library applications.

Finally, although coverage and aggregation procedures are defined and applied consistently across pipelines, the reported results emphasize comparative performance rather than absolute guarantees of correctness. As with any automated metadata extraction system, downstream use in library or bibliometric workflows should be accompanied by targeted validation for high risk fields.

Conclusion and future work

This study provides field-level empirical evidence that pipeline choice matters for engineering education metadata extraction and that a field-selective hybrid strategy can outperform single-approach pipelines in end-to-end performance. Across all evaluated systems, AlphaMind-Extrator achieved the strongest overall results, combining high accuracy, strong coverage, and reduced time and cost by limiting large language model use to affiliation extraction and fusing results through robust author matching. These findings demonstrate that reliable automation in engineering library workflows is best achieved through evidence-based allocation of extraction tasks, rather than uniform reliance on a single tool.

The results also reinforce a broader implication for automated scholarship: high performance on structured bibliographic fields does not guarantee reliability across all metadata elements. Affiliation extraction represents a persistent risk when automated outputs are used without validation. By identifying and addressing this failure mode, the hybrid approach evaluated here offers a practical model for trustworthy metadata extraction that aligns with the operational needs of engineering librarians and bibliometric researchers.

Future work will extend this benchmarking effort in several directions. First, the dataset will be expanded to include a larger and more diverse sample of engineering education papers, with transparent documentation of sampling decisions to support reproducibility and external validity. Second, future evaluations will incorporate formal gold-standard adjudication procedures, including interrater reliability reporting, to further strengthen methodological rigor. Third, the benchmarking framework and evaluation harness will be released openly to enable consistent comparison across extraction tools, in alignment with Engineering Libraries Division empirical reporting expectations. Finally, deeper error analysis will be conducted to characterize affiliation failure modes across PDF layout types, and the role of keywords in library workflows will be evaluated to determine whether they should be included as a required metadata field in future benchmarks.

References

[1] J. R. Park and Y. Tosaka, “Metadata quality control in digital repositories and collections: Criteria, semantics, and mechanisms,” Cataloging & Classification Quarterly, vol. 48, no. 8, pp. 696-715, 2010, doi: 10.1080/01639374.2010.508711.

[2] N. Palavitsinis, N. Manouselis, and S. Sanchez-Alonso, “Metadata quality in digital repositories: Empirical results from the cross-domain transfer of a quality assurance process,” Journal of the Association for Information Science and Technology, vol. 65, no. 6, pp. 1202-1216, 2014, doi: 10.1002/asi.23045.

[3] P. Lopez, “GROBID: Combining automatic bibliographic data recognition and term extraction for scholarship publications,” in Research and Advanced Technology for Digital Libraries (ECDL 2009), Lecture Notes in Computer Science, Berlin, Germany: Springer, 2009, pp. 473-474, doi: 10.1007/978-3-642-04346-8_62.

[4] J. Dagdelen et al., “Structured information extraction from scientific text with large language models,” Nature Communications, vol. 15, no. 1, p. 1418, 2024, doi: 10.1038/s41467-024-45563-x.

[5] M. Sokolova and G. Lapalme, “A systematic analysis of performance measures for classification tasks,” Information Processing & Management, vol. 45, no. 4, pp. 427-437, 2009, doi: 10.1016/j.ipm.2009.03.002.

[6] S. Qiu and M. Natarajarathinam, “Fifty-three years of the Journal of Engineering Education: A bibliometric overview,” Journal of Engineering Education, vol. 113, no. 4, pp. 767-786, 2024, doi: 10.1002/jee.20547.

[7] Y. Jiao et al., “Instruct and extract: Instruction tuning for on-demand information extraction,” in Proc. 2023 Conf. Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, 2023, pp. 10030-10051, doi: 10.18653/v1/2023.emnlp-main.620.

[8] Kermitt2, “GROBID: A machine learning software for extracting information from scholarly documents,” GitHub repository. [Online]. Available: https://github.com/kermitt2/grobid

[9] S. K. Sukula, R. K. Sharma, and P. Verma, “Metadata Enrichment and Interoperability: Scoping Review of Artificial Intelligence Contexts in Metadata in Academic Libraries Experiences Across the Globe,” International Journal of Research in Library Science, vol. 11, no. 3, 2025.

[10] J. Y. Engel, D. T. Do, B. Salem, and T. A. Cunningham, “Artificial Intelligence in Library Cataloging: A Review of Literature,” Journal of Library Metadata, vol. 25, no. 4, pp. 261-276, 2025, doi: 10.1080/19386389.2025.2526913.

[11] N. T. Camp, J. A. Bengtson, and J. C. Sandstrom, “The citation catastrophe: Propagation of AI-generated counterfeit citations in scholarship,” The Journal of Academic Librarianship, vol. 51, no. 4, Article 103065, Jul. 2025, doi: 10.1016/j.acalib.2025.103065.

[12] N. Meuschke, A. Jagdale, T. Spinde, J. Mitrovic, and B. Gipp, “A Benchmark of PDF Information Extraction Tools using a Multi-Task and Multi-Domain Evaluation Framework for Academic Documents,” presented at iConference 2023, Mar. 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2303.09957

Appendix A: heuristics used in the classical NLP/ML pipeline

This appendix documents the rule-based heuristics, regular expressions, and lightweight NLP components employed in the classical NLP/ML pipeline (reported as NLP/ML in the main text). The purpose of this documentation is to ensure transparency and reproducibility.

A.1 pre-filtering

Scope: Only the first three pages of each PDF are scanned for target-author presence. Documents that do not match are skipped prior to full extraction.

Target-Author Matching Criteria
A document is retained if any of the following patterns are detected (case-insensitive):

- Exact substring: The normalized full target name (after stripping accents and collapsing spaces) appears as a substring in the head text.
- First Last: Word boundary + first name + whitespace + last name + optional digit(s) + word boundary (e.g., Brian Nord, Brian Nord1)
- Initial Last: Word boundary + single initial (optional period) + whitespace + last name + optional digit(s) + word boundary (e.g., B. Nord, B Nord)
- Last, First: Word boundary + last name + optional spaces + comma + optional spaces + first name + word boundary
- Last, Initial: Word boundary + last name + optional spaces + comma + optional spaces + single initial (optional period) + word boundary

Name Normalization
For comparison: spaces normalized; accents stripped using Unicode NFKD decomposition (combining characters removed). For parsing: comma form split as surname, given name; otherwise split on whitespace or hyphens (last token treated as surname).

A.2 text extraction and byline block

Primary extraction engine: PyMuPDF (fitz). Fallback engine: pdfminer.

Byline Block Definition
Lines are collected from the start of the document until the earliest occurrence of: an Abstract header (see A.5); a Keywords header (see A.6); the word Introduction; or the first 300 lines.

Normalization
All extracted text is space-normalized (multiple whitespace collapsed to a single space and trimmed).

A.3 author extraction

Authors are produced by merging two sources and deduplicating results (order preserved).

A.3.1 Byline/List Parsing
Splitting: The byline block is split on commas (,) or the word and (case-insensitive). Per chunk processing: Remove superscripts and footnote markers (digits, asterisk, Unicode superscripts, dagger, double dagger, section and paragraph symbols); normalize spaces; strip

leading/trailing commas, colons, semicolons. Acceptance criteria: (1) 2-6 space-separated tokens; (2) contains at least one subpattern matching capital letter followed by lowercase letters (e.g., McDonald). Deduplication: by string identity (first occurrence retained).

A.3.2 Named Entity Recognition (NER)
Model: spaCy en_core_web_sm. Entities collected: all spans labeled PERSON in the byline block or head text (if byline block empty). Filtering: 2-6 space-separated tokens. Deduplication: by string identity (first occurrence retained).

A.3.3 Target-Author Abbreviation
Comma form (Last, First [Middle]): surname = segment before first comma; given names = tokens after comma. Otherwise: split on spaces/hyphens; last token = surname; remaining tokens = given + middle. Abbreviation rule: middle names longer than one character produce single capital letter + period (e.g., Robert T. becomes R.); single-token first last forms left unchanged.

A.4 title extraction

Input: First 40 non-empty lines of head text (first three pages). Skip conditions: lines shorter than 6 characters; lines beginning with Abstract header, Keywords header, or Introduction (case-insensitive). Title Selection Heuristics: candidate line must be (1) <= 180 characters, (2) not end with a period, (3) not resemble an author list (fewer than 2 commas, or does not contain and). Fallback: first non-skipped line among the 40 examined.

A.5 abstract extraction

Abstract headers (case-insensitive): Abstract | Summary | Resume | Zusammenfassung | Resumen. Keywords headers (case-insensitive): Keywords? | Index Terms | Mots-cles | Palabras clave | Schlusselworter (optional hyphen permitted). Extraction strategies: Strategy 1 (primary): match Abstract header, capture text up to (but not including) Keywords header. Strategy 2 (fallback): match Abstract header, capture text up to Introduction or newline + 1. + Introduction. Normalization: collapse whitespace in captured block.

A.6 keywords, year, and publication metadata

Keywords: first occurrence of Keywords header followed by optional colon or hyphen; capture truncated at first newline or 2000 characters. Year: first four-digit year in range 1900-2099 within first 5000 characters. Journal identification: first line (within first 80 lines of first 8000 characters) containing: Journal | Proceedings of | Transactions on | Conference on. Volume/issue patterns: (1) Vol. or Vol + digits, optionally followed by No. + digits; (2) Volume + digits + (Issue or No.) + digits. Pages: pp. or pp + page range (digits + hyphen/en-dash + digits).

A.7 affiliation extraction

Affiliations are collected from: (1) keyword-based lines; (2) NER organizations. Results are cleaned and deduplicated.

A.7.1 Keyword-Based Lines
Search space: byline block and head text (first three pages). Line must contain at least one of: university, department, institute, school, college, centre, center, laboratory, lab, faculty; and

be longer than 8 characters. Cleaning: strip leading superscripts, digits, asterisks, footnote symbols; collapse internal spaces; remove duplicates; keep at most 10 affiliations per document.

A.7.2 NER-Based Organizations
Model: spaCy en_core_web_sm. Entities collected: all spans labeled ORG. Filtering: length > 3 characters after normalization. Merged with keyword-based list (maximum 10 total).

A.7.3 Author-Affiliation Mapping
Superscript indices in the PDF are not used. All extracted affiliations are assigned uniformly to every author in the document.

A.8 email and DOI extraction

Email: standard email regular expression applied to head text then full text; maximum stored: 5 addresses. DOI: first match of 10. + 4-9 digits + / + non-whitespace; trailing punctuation removed.

A.9 target-author robust matching

Parsing rules: comma form: segment before comma = last name; segment after comma = first name. Otherwise: first token = first name; last token = last name. Normalization: lowercase; strip accents (NFKD); remove all characters except letters, digits, hyphen, apostrophe. Match criterion: exact equality of normalized first and last names.

A.10 constants and limits summary

Parameter	Value
Pages scanned	3
Max byline lines	300
Max title candidate lines	40
Max title length	180 characters
Author token count	2-6
PERSON token count	2-6
ORG minimum length	> 3 characters
Max affiliations	10
Max emails	5

Table A.10: Constants and limits used in the Classical NLP/ML pipeline.

A.11 file discovery and concurrency

PDF Discovery: recursive directory walk; filenames ending in .pdf (case-insensitive); deduplicated and sorted. Row Indexing: row index = 1-based position of filename when sorted within its directory. Concurrency: configurable worker count (default = 4) for parallel processing after pre-filtering.

This appendix corresponds to the implementation in NPL_extraction_pipeline.py and may be used to reproduce or audit the classical NLP/ML pipeline behavior.

Appendix B

Source of 50-paper dataset: https://docs.google.com/spreadsheets/d/1QOV45nVEprve0bEpD7JZr_dnw46cTChr/edit

Code source for the project: https://github.com/Joshua-123-ansah/efficient-extraction-pipeline