



Cite this: *Digital Discovery*, 2026, 5, 4191

MetMINE: a generative large language model-based source-tracked multi-pass framework for metallic materials information extraction

Xin Wang,^a Anshu Raj,^b Matthew Luebke,^c Haiming Wen,^c Shuozhi Xu^{*b} and Kun Lu^{id *a}

Data-driven materials discovery requires large-scale experimental datasets, yet most of the information remains trapped in unstructured literature. Existing extraction efforts often focus on a limited set of features and have not addressed the integrated composition–processing–microstructure–property relationships essential for understanding materials behavior, thereby posing challenges for building comprehensive databases. To address this gap, we propose a multi-pass information extraction pipeline powered by large language models, which captures 47 features spanning composition, processing, microstructure, and properties exclusively from experimentally reported materials. The pipeline, named MetMINE (Metallic Materials INformation Extraction), integrates iterative extraction with source tracking to enhance both accuracy and reliability. Evaluations at the feature level (independent attributes) and tuple level (interdependent features) yielded F1 scores around 0.96. Compared with single-pass extraction without source tracking, MetMINE improved F1 scores of microstructure category by 10.0% (feature level) and 13.7% (tuple level), and reduced missed materials from 49 to 13 out of 396 materials in 100 articles on precipitate-containing multi-principal element alloys (*i.e.*, the miss rate is reduced from 12.4% to 3.3%). The pipeline enables scalable and efficient literature mining, producing databases with high precision, minimal omissions, and zero false positives. These datasets provide trustworthy inputs for machine learning and materials informatics, while the modular design generalizes to diverse material classes, enabling comprehensive materials information extraction.

Received 14th June 2026
Accepted 16th August 2026

DOI: 10.1039/d6dd00374e

rsc.li/digitaldiscovery

1 Introduction

Materials science relies on a comprehensive understanding of processing, microstructure, property, and performance, which together form the well-established materials tetrahedron.^{1–3} Extracting such information from the rapidly growing body of literature is critical to accelerate data-driven discovery. Computational databases such as Materials Project (>200 000 materials), <https://www.materialsatlas.org/> (>130 000 compounds), AFLOW (>3.5 million entries), and NOMAD (>100 million calculations) have structured vast amounts of simulation results.^{4–6} In contrast, while some experimental databases exist (such as HTEM,⁷ RRUFF,⁸ SpringerMaterials), the vast majority of experimental information, including processing histories, microstructures, and properties, remains largely trapped in unstructured text across millions of

publications. Recent analyses reveal that materials science publications grow at 4.10% annually and were doubled within 17.3 years,⁹ creating an ever-expanding yet underutilized repository of knowledge.

Within this broader challenge, and for metallic materials, microstructure information is critical but difficult to extract. Microstructure directly governs macroscopic behavior of metals, from mechanical strength and ductility to thermal and electrical performance.^{10,11} Previous studies have attempted to characterize microstructure using image analysis,^{12–15} but such approaches cannot capture the contextual information available in text, where microstructural details are often reported alongside composition, processing, and properties. Unlike other categories of material information, microstructural details are highly multidimensional, scattered throughout text, and often referenced indirectly through abbreviations or symbols, making it a bottleneck in tracking the complete processing–microstructure–property chain.^{16,17} This, and the fact that microstructures sometimes affect properties of metals more than composition does, highlights why comprehensive information extraction cannot be achieved without microstructure. Without comprehensive information extraction, it is difficult to realize

^aSchool of Library and Information Studies, University of Alabama, Tuscaloosa, 35487, AL, USA. E-mail: klu@ua.edu

^bSchool of Aerospace and Mechanical Engineering, University of Oklahoma, Norman, 73019, OK, USA. E-mail: shuozhixu@ou.edu

^cDepartment of Materials Science and Engineering, Missouri University of Science and Technology, Rolla, 65409, MO, USA

data-driven discovery and accelerated innovation in materials science.¹⁸

Traditional natural language processing (NLP) approaches have made significant contributions to materials information extraction. Rule-based systems such as OSCAR4,¹⁹ ChemicalTagger,²⁰ and ChemDataExtractor²¹ specialized in automated extraction of chemical entities and relationships from scientific literature, establishing important foundations for the field. These tools have been widely adopted and alloy-focused extensions have demonstrated the feasibility of domain-specific extraction.²² Building on these advances, ChemDataExtractor2 expanded capabilities to include quantitative property extraction.²³

In parallel, machine learning approaches introduced semi-supervised and shallow neural models to tackle extraction challenges. Huo *et al.*²⁴ applied topic modeling and Markov chains for synthesis procedure classification, while Yan *et al.*²⁵ employed weak supervision frameworks such as Snorkel to reduce annotation requirements.

These foundational works have paved the way for more comprehensive extraction frameworks, though limitations remain. Rule-based systems struggle with the complexity and variability of scientific language. Machine learning approaches, while more flexible, require extensive manual annotation and continue to struggle with domain-specific terminology and complex multi-dimensional relationships inherent in materials science literature.

A major breakthrough came with the use of distributed representations. Methods such as word2vec,²⁶ doc2vec,²⁷ and BERT (bidirectional encoder representations from transformers)²⁸ enabled models to capture latent semantic relationships in materials text. This paradigm demonstrated success in unsupervised discovery of novel materials from millions of abstracts.²⁹ Later work has extended these approaches to full-text analysis. MatSciBERT achieved state-of-the-art results on materials science NER (named entity recognition) and relation classification tasks.³⁰ MaterialBERT demonstrated effective word embeddings for materials clustering and functional group relationships.³¹ And MaterialsBERT developed a comprehensive pipeline for extracting material property records from polymer literature.³² Nevertheless, these models are fundamentally representational, requiring downstream fine-tuning, task-specific annotations, or handcrafted rules to convert embeddings into structured knowledge.

Generative large language models (LLMs) such as GPT (generative pre-trained transformer) differ from the previous models in that they can directly produce structured outputs and capture relationships beyond the sentence level without fine-tuning. Recent studies have demonstrated their potential for automated information extraction in materials science. Zheng *et al.*³³ used a ChatGPT-based chemistry assistant to mine synthesis parameters of metal-organic frameworks. Gupta *et al.*³⁴ combined LLMs with classifiers to curate large polymer property databases. Torres da Silva *et al.*³⁵ structured synthesis protocols of Metal-Organic Frameworks, Covalent Organic Frameworks, and zeolites directly from full texts (in PDF format), while Thway *et al.*³⁶ parsed solid-state synthesis “recipes” from text. Other studies have

emphasized reliability: Dagdelen *et al.*³⁷ fine-tuned LLMs for joint entity-relation extraction, and Polak and Morgan³⁸ developed ChatExtract to reduce hallucinations and improve precision. Wang *et al.*³⁹ found that GPT-4 substantially outperformed traditional methods in band gap extraction tasks, and with prompt engineering, its accuracy could be improved to 93.98%. More recently, Sayeed *et al.*⁴⁰ developed KnowMat, an agentic multi-pass pipeline that combines iterative evaluation-driven re-runs with schema-constrained tool calling to extract structured property data from materials science PDFs, achieving an F1 score of 86.2% on mechanical and thermal property extraction. Sadruddin *et al.*⁴¹ introduced SciKGExtract, an agentic workflow that integrates LLM-based schema-guided extraction with deterministic validation and PubChem-based chemical normalization, demonstrating effective structured knowledge extraction from atomic layer deposition literature. Yet, existing work remains limited to single dimensions, simple composition-property relations, or simple composition-processing correspondence, restricting the understanding of the full composition-processing-microstructure-property relationships.

Recent targeted benchmarking has confirmed this limitation. Chong and Colindres⁴² introduced LitXBench—a framework for evaluating LLM-based extraction of experimental data from materials literature—and released the LitXAlloy dataset, which comprises 1426 manually annotated data points derived from 19 alloy-related papers. Their analysis revealed a specific manifestation of the aforementioned “relational incompleteness” issue: existing workflows often link measurement data solely to chemical composition while overlooking the processing history that uniquely characterizes a material's properties. Consequently, a critical link in the complete CPMP (Composition-Processing-Microstructure-Property) chain—specifically the relationship between processing and properties—is frequently lost. Addressing this challenge requires the extraction process to be firmly grounded in the specific textual evidence defining each material; this necessity drove the development of the “source-tracking” method presented in this study.

This study proposes a multi-pass extraction pipeline, named MetMINE (Metallic Materials INFORMATION Extraction), to capture the hierarchical nature of materials information across the complete composition-processing-microstructure-property chain. While the well-established materials tetrahedron is typically defined by processing, structure, properties, and performance, we hereby exclude performance while adding composition. Performance often depends on system-level factors beyond intrinsic materials descriptors, whereas composition plays a foundational role in governing processing routes and determining structure-property relations, making it more suitable for the subsequent data-driven discovery. We target 47 distinct features spanning these four categories: 14 composition features, 12 processing parameters, 11 microstructure attributes, and 10 material properties. Fig. 1 shows what features are extracted and how prompts are designed for each category. In this paper, our framework is applied to 100 journal articles on the topic of “precipitate-containing multi-principal element alloys,” because these materials contain

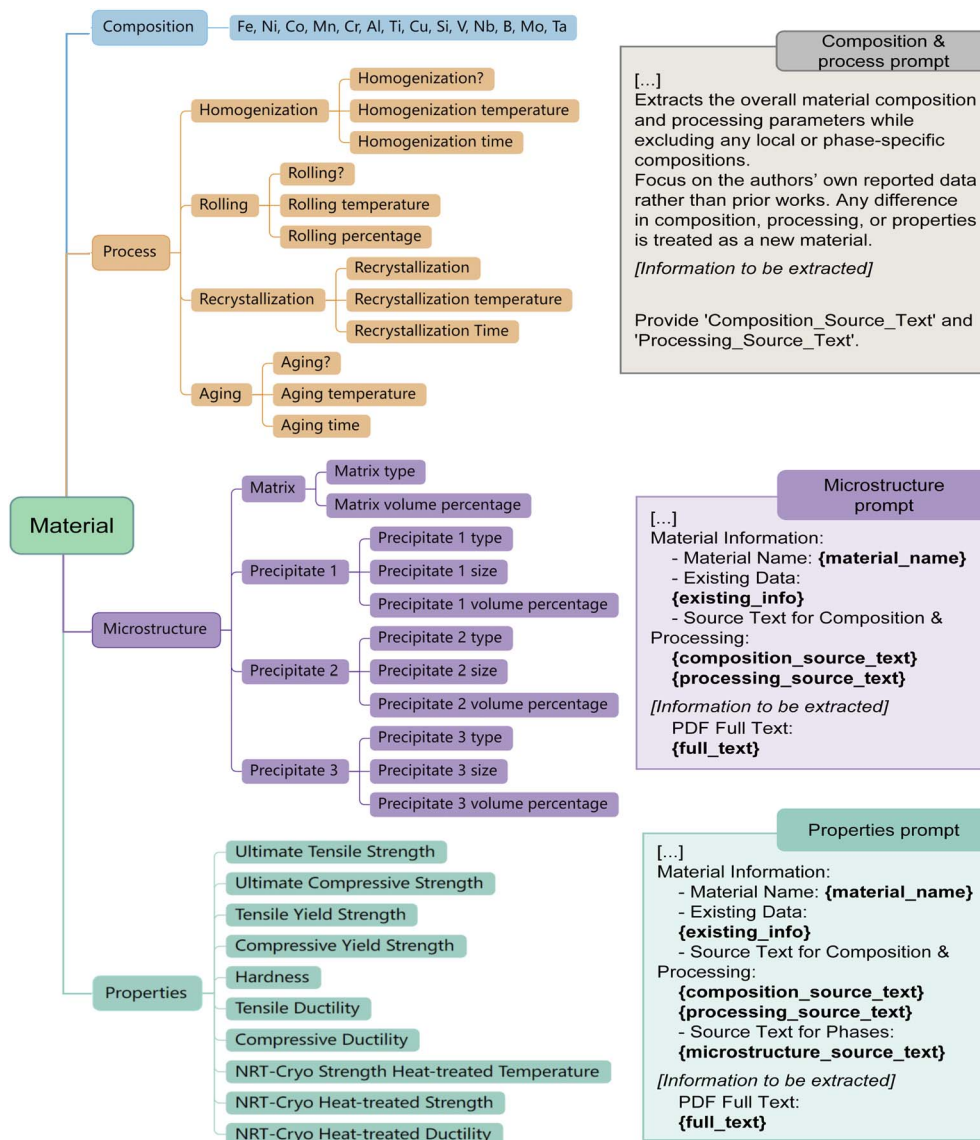


Fig. 1 Extraction framework showing the 47 features extracted across four categories (composition, processing, microstructure, and properties) and the specialized prompt design for each category. Here, "NRT-Cryo" means non-room temperature cryogenic conditions.

rich information on both composition and microstructure. These 100 articles contain original experimental results. While some papers also include modeling or secondary analyses, we instructed GPT to extract only experimental data.

As shown in Fig. 2, MetMINE employs a four-pass approach, with all steps executed using o3-mini via the OpenAI API. Gpt-oss-20B was evaluated as an alternative model. Step 1 extracts all materials along with their chemical compositions and processing parameters from the entire article text. All extracted materials are derived directly from the authors' own experimental results, ensuring that the database reflects only primary research findings. Following this initial extraction, the LLM iteratively processes each identified material: Step 2 extracts microstructure information, including matrix and precipitate types, sizes, and volume fractions, using the composition and processing context from Step 1. Step 3 then extracts mechanical

properties for the same material while leveraging all previously accumulated information. This Step 2–3 sequence repeats for each material identified in Step 1. Only after completing the extraction for all materials does Step 4 perform comprehensive validation using the complete material database and original source texts. This database contains all materials identified in Step 1 along with their extracted composition, processing, phase structure, and property information from Steps 2–3.

A key innovation of our approach is the source tracking mechanism, which requires LLMs to provide an explicit reasoning basis for each extracted value in every pass. This mechanism preserves links between extracted values and their original text sources, reducing reference loss during multi-pass processing and enabling systematic error correction through traceable validation. The source tracking mechanism requires multi-pass processing to maintain contextual links across

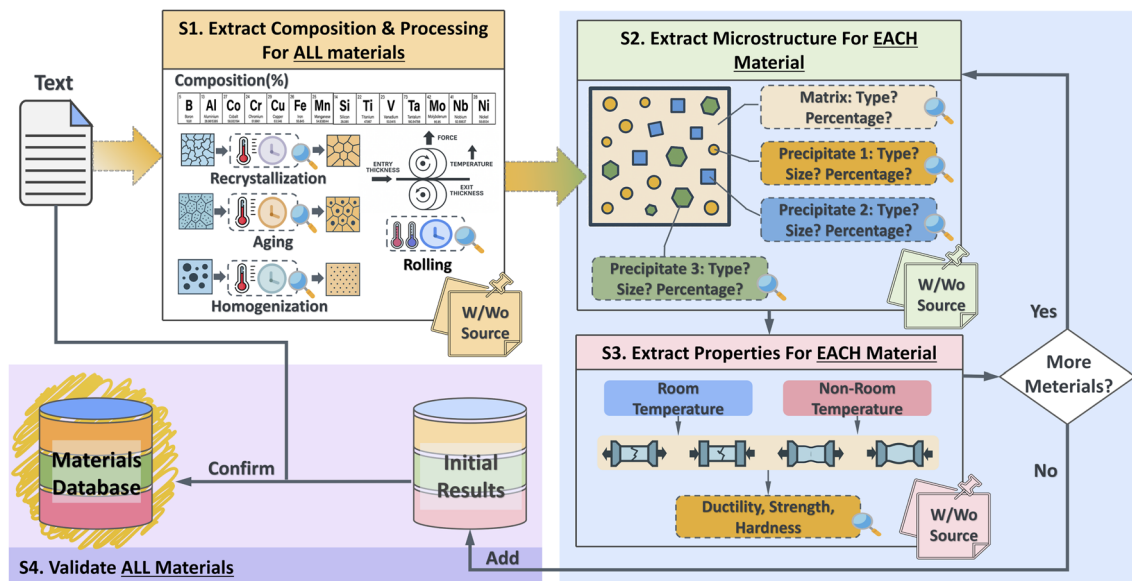


Fig. 2 Architecture of our multi-pass extraction pipelines (with/without source tracking).

different information categories, distinguishing it from the single-pass approach. Thus, to demonstrate the effectiveness of this mechanism and the necessity of multi-pass processing, we compared three pipeline variants: single-pass extraction without source tracking, multi-pass extraction without source tracking, and multi-pass extraction with source tracking (*i.e.*, MetMINE).

These pipelines are evaluated by two complementary approaches: feature-level evaluation that treats each feature independently, while tuple-level evaluation that groups interdependent features and evaluates them holistically (*i.e.*, evaluating interdependent features such as precipitate type, size, and volume fraction as a complete group rather than independently). Using o3-mini model with prompt engineering and confirmation strategies, our approach achieves strong and consistent performance across both feature-level evaluations, which consider independent attributes, and tuple-level evaluations, which assess interdependent features such as precipitate type, size, and volume fraction as a complete group rather than independently. These results confirm the robustness of our extraction methodology. Even in the most challenging microstructure extraction task, MetMINE maintains strong performance across all evaluation metrics. Our work demonstrates consistent accuracy across all feature categories that are critical for materials database construction.

2 Methods

2.1 Dataset construction

100 journal articles related to precipitate-containing multi-principal element alloys were collected for testing the information extraction pipelines. Articles were identified through a systematic keyword search on Web of Science and Scopus, using the terms (“*multi-principal element alloy*” OR “*high-entropy alloy*”) AND (“*precipitate*” OR “*precipitation*”). An initial 100

articles were selected from this pool. Articles were included if they (1) reported original experimental results (excluding purely computational studies and review articles) and (2) contained at least one material with an explicitly stated chemical composition and at least one reported mechanical property. When an article from the initial sample did not meet these two criteria, we randomly picked an article from the search results pool to replace it until we reached 100 articles that met the criteria. While we aimed to cover a representative range of the MPEA precipitation literature, the resulting corpus is not exhaustive. It may also reflect biases inherent to the current publication landscape in this subfield, particularly the overrepresentation of FCC-based alloy systems and English-language publications.

Two materials science experts annotated the dataset, extracting a comprehensive set of 47 features across four distinct categories from 396 materials: composition (14 features), processing parameters (12 features), microstructure (11 features), and material properties (10 features). The four processing categories captured by the pipeline, namely homogenization, rolling, recrystallization, and aging, were selected based on their prevalence in the precipitate-containing MPEA literature represented by our corpus. Rolling was included as one of these four categories because thermo-mechanical deformation *via* rolling is among the most frequently reported processing steps in this subfield. The pipeline is explicitly modular. Additional deformation processes, such as forging, extrusion, cold drawing, or equal-channel angular pressing, can be incorporated by adding corresponding fields and boolean indicators to the step 1 prompt without requiring any structural changes to the pipeline architecture.

2.2 Pre-processing setup

PDF documents were processed using PyMuPDF (fitz) for plain text extraction. No section-level filtering was applied. The full

converted text of each article was provided to the LLM. There are two reasons for adopting this approach. First, o3-mini provides a 200 000-token context window,⁴³ which comfortably accommodates the full converted text of a journal article within a single request, so that no chunk-level retrieval or section filtering is required. Prior work has likewise shown that LLMs can extract structured information directly from unstructured, PDF-derived scientific text without fine-tuning or hand-crafted cleaning rules.^{35,38} Second, and more importantly, the source-tracking mechanism provides a targeted safeguard against the noise that such raw text introduces. By requiring the model to cite the specific sentence(s) supporting each extracted value, erroneous extractions from non-experimental sections, such as reference lists or acknowledgments, are naturally suppressed, since these sections rarely contain reported experimental values in the required form. Rather than removing such content *a priori*, our pipeline tolerates it and filters it at the point of extraction. The strong extraction performance confirms that this minimal-preprocessing design does not compromise accuracy. Although this approach does not preserve table formatting, the numerical data and text content from tables are still extracted and remain interpretable by the LLM. Figures and images were excluded as the o3-mini API only accepts text-based inputs. The extracted content maintained its original formatting as produced by PyMuPDF without additional standardization or character normalization. All o3-mini extractions used the OpenAI model snapshot o3-mini-2025-01-31, with the reasoning_effort parameter configured as high. Output was structured in JSON format to ensure consistent parsing and validation. The experiments were conducted in a Python 3.12.2 environment.

2.3 Four-pass extraction pipeline

Our extraction pipeline employs a systematic four-pass approach designed to leverage the hierarchical nature of material information. Each extraction pass employs a dual structure: predefined fields for standard information categories and corresponding “Additional” fields (Other_Composition, Additional_Processing, Additional_Microstructure, Additional_Properties) to ensure no reported information is lost due to rigid categorization.

Step 1 performs initial material identification and extraction across the entire document. In this pass, we extract all materials mentioned in the article along with their chemical compositions in atomic percentages and processing parameters. Processing information includes homogenization conditions with temperature and time, rolling parameters encompassing temperature and thickness reduction percentage, recrystallization conditions with temperature and time, and aging treatments with their corresponding temperature and duration. Each material is assigned to a unique identifier, and all extracted values are accompanied by their source text from the original document. The extracted information is then converted to JSON format for structured data processing.

Following Step 1, we implement an iterative extraction process for each identified material. For each material, Step 2 first extracts microstructure information. This pass receives the

specific material's composition and processing data from Step 1, along with source texts, extraction guidelines, and the full article text. The extraction targets matrix phase type and volume fraction, along with precipitate phases through structured fields capturing their types, sizes, and volume fractions. Corresponding source texts are recorded for each extracted phase feature, and the microstructure data is converted to JSON format.

Step 3 immediately follows Step 2 for the same material, extracting its mechanical properties. Building upon the accumulated composition, processing, and microstructure information for this specific material, Step 3 captures room temperature mechanical properties, including ultimate tensile and compressive strength, yield strength, hardness, and ductility values. Additionally, non-ambient temperature properties are captured, specifically strength and ductility measurements at elevated or cryogenic temperatures. All property values are linked to their source texts, completing the information extraction for this individual material, with outputs converted to JSON format.

This Step 2–3 sequence repeats for each material identified in Step 1, ensuring that microstructure and property information are extracted with the full context of each material's specific composition and processing history. The iterative approach maintains material-specific context throughout the extraction process, reducing cross-material contamination and improving attribution accuracy.

Step 4 performs confirmation only after all materials have been fully processed through Step 2 and 3. Using the complete material database containing all extracted materials with their composition, processing, microstructure, and property information, along with all source texts and the full article text, the API conducts this final confirmation step to verify extraction accuracy and consistency across the entire dataset. The confirmation process checks phase volume percentages, ensures properties align with their corresponding processing conditions, verifies material uniqueness, and performs overall quality control.

2.4 Prompt engineering strategies

Three core strategies guide our prompt design to ensure accurate and complete extraction.

The first strategy involves the definition of a unique material. Each unique combination of composition and processing conditions constitutes a separate material entry. This prevents conflation of materials with identical compositions but different processing histories.

The second strategy implements controlled flexibility through structured alternative fields. The system includes provisions for “Other_Composition”, “Additional_Processing”, “Additional_Microstructure”, and “Additional_Properties” to ensure comprehensive information capture beyond predefined categories. This design recognizes that materials science literature often reports information that doesn't conform to standard classifications. For composition, while we track 14 primary elements, the “Other_Composition” field captures trace

elements. For microstructure information, we provide three structured fields for primary precipitates with detailed attributes (type, size, volume fraction), but the “Additional_Microstructure” field ensures that papers reporting four, five, or more precipitate types can be fully captured without forcing data into inappropriate categories. Similarly, “Additional_Processing” accommodates novel or non-standard processing steps beyond homogenization, rolling, recrystallization, and aging. “Additional_Properties” captures specialized mechanical or physical properties beyond our core set of strength, hardness, and ductility measurements. This flexible architecture prevents information loss while maintaining structured data organization, achieving both completeness and consistency in extraction.

The third strategy implements source tracking through a three-step mechanism: recording, storage, and forwarding.

2.4.1 Step 1—Recording. Each extraction prompt explicitly instructs the LLM to provide, alongside every extracted value, a companion `Source_Text` field containing the sentence(s) from the original article that directly support that value. These companion fields—`Composition_Source_Text`, `Processing_Source_Text`, and `Phases_Source_Text`—are required outputs of stage 1, 2, and 3, respectively, and are declared in the JSON schema alongside the extracted numeric or categorical values. The recording step makes the model's evidential basis explicit and machine-readable at every pass of the pipeline.

2.4.2 Step 2—Storage. The source text excerpts are embedded within the same structured JSON object as the extracted values, so that every piece of information is co-located with its textual evidence. This design keeps the evidence and the data together throughout the pipeline: passing the JSON forward automatically carries the source text with it. On average, each source text excerpt spans approximately 35–65 words, comprising one to three sentences from the original article.

2.4.3 Step 3—Forwarding. When constructing the prompt for step 2 (microstructure extraction), the pipeline reads `Composition_Source_Text` and `Processing_Source_Text` directly from the stage 1 JSON output and injects them as a dedicated “Source Text for Composition & Processing” section in the stage 2 prompt. Stage 3 (property extraction) receives all three source text fields—composition, processing, and phases—as separate prompt sections.

2.4.4 Step 4—Validation. This step operates on the full material database, including all source texts and the complete article text, enabling cross-referencing across every prior extraction decision.

This cumulative forwarding mechanism ensures that every downstream pass can verify its extractions against the specific sentences in the original article. These sentences provide the evidence that supported all prior decisions. As a result, the system reduces ambiguity in material references and suppresses unsupported inference. The key difference from the multi-pass pipeline without source tracking is that the latter passes only extracted numeric and categorical values between steps. It does not preserve textual anchors from the original article. Consequently, downstream prompts must re-identify

the relevant evidence in each call. This increases the risk of misattribution, especially for materials with similar compositions but different processing histories. A step-by-step worked example illustrating the complete source-tracking mechanism, with actual JSON excerpts and prompt structures, is provided in the SI (Section S1).

2.5 Reasoning configuration

We configured the extraction system with `reasoning_effort=“high”` without restricting the model's reasoning capabilities in prompts. This setting enables the model to invest additional computational resources for deep reasoning, construct complete logical chains, carefully analyze complex causal relationships, and provide well-founded inferences rather than arbitrary guesses when facing ambiguous information.

This approach proves particularly valuable in the materials science literature where certain calculations and inferences are routine. For example, when a paper provides only the chemical formula Fe_2Al_5 without explicit atomic percentages, the model accurately calculates Fe: 28.6 at%, Al: 71.4 at%. Similarly, when papers describe materials containing a matrix phase and two precipitate phases with only partial volume fraction information, such as matrix at 85% and one precipitate at 10%, the model correctly infers that the second precipitate occupies 5% of the volume.

2.6 Evaluation metrics

To evaluate the performance, the ground truth was developed through a multi-expert validation process involving five co-authors. Initially, two domain experts manually extracted material information from the literature to create a preliminary version of the ground truth. To account for human errors, a third co-author with expertise in information extraction reviewed and marked discrepancies between the preliminary ground truth and the LLM-generated output. A fourth domain expert then conducted a review by comparing both versions against the original literature, identifying any missed extractions, erroneous additions, or inconsistencies. The reconciled results were subsequently checked again by the third co-author. Finally, the fourth and fifth experts finalized the ground truth dataset. This multi-pass process ensured the rigor and reliability of the ground truth.

Building on this multi-expert process, we further established rules for ground truth representation and feature value assignment. Our ground truth defines positive cases as values that are explicitly present in the original text or correct values obtained through valid reasoning, while negative cases refer to values that are absent or not mentioned in the text. The confusion matrix classification follows standard definitions:

- True Positives (TP): ground truth and extracted values are present and match.
- True Negatives (TN): both ground truth and extracted values are absent.
- False Negatives (FN): ground truth has a value, but the extraction is empty.

- False Positives (FP): either the ground truth is missing while the extraction provides a value, or both contain values, but the extracted one is incorrect.

In feature value assignment, all elements not reported in the composition are assigned a value of zero. For processing features, we implemented a specific approach to handle missing information: when papers do not mention some processing steps (homogenization, aging, or recrystallization), we set the corresponding boolean indicator to 0 (*e.g.*, Aging? = 0) and assign null values to associated temperature and time. For microstructure features, any precipitate or matrix information not explicitly mentioned in the text is marked as “not reported”, including their corresponding volume fractions and sizes. Similarly, material properties not reported are also marked as “not reported”.

Building on standard classification metrics, our evaluation addresses the challenges of materials information extraction through a two-tiered approach. First, we evaluate performance only for correctly identified materials to provide insight into extraction quality. Then we penalize the system for any missed materials (counting all 47 features as false negatives) or erroneously extracted materials (counting all 47 features as false positives). For processing steps and microstructure information, we implement the two complementary calculation approaches described in Section 2.3: single-feature evaluation and tuple-based evaluation.

To evaluate extraction performance, we use standard classification metrics based on confusion matrix elements. For each evaluation task, we define True Positives (TP) as correctly extracted features or materials that match the ground truth, False Positives (FP) as incorrectly extracted features or materials not present in the ground truth, True Negatives (TN) as correctly identified absent features or materials, and False Negatives (FN) as ground truth features or materials that were missed by the extraction system.

Based on these fundamental counts, the evaluation metrics are calculated as follows:

- Precision measures the proportion of extracted items that are correct:

$$\begin{aligned}\text{Precision} &= \frac{\text{TP}}{\text{TP} + \text{FP}} \\ &= \frac{\text{number of correctly predicted positives}}{\text{total number of extractions}}\end{aligned}$$

- Recall measures the proportion of ground truth items that are successfully extracted:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} = \frac{\text{number of correct extractions}}{\text{total number in ground truth}}$$

- F1 score provides the harmonic mean of precision and recall:

$$\text{F1} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}}$$

- Accuracy measures the overall correctness across all predictions:

$$\begin{aligned}\text{Accuracy} &= \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \\ &= \frac{\text{Number of correctly predicted instances}}{\text{Total number of predictions}}\end{aligned}$$

3 Results

3.1 Overall performance evaluation

Three extraction pipelines are compared to understand the impact of the multi-pass pipeline and source tracking in material information extraction. Fig. 3 illustrates their processes: (1) single-pass extraction without source tracking, (2) multi-pass extraction without source tracking, and (3) multi-pass extraction with source tracking. The fundamental distinction is in how information flows through the system and whether contextual source text is preserved throughout the extraction process.

Table 1 presents a comparison of performance metrics across three pipelines using feature-level and tuple-level evaluation. Unless otherwise specified, all results reported in Sections [3.1]–[3.3] are obtained using o3-mini. Section [3.4] makes a comparison between o3-mini and Gpt-oss-20B under the multi-pass pipeline with source.

Our results demonstrate that the multi-pass pipeline with source tracking achieves the best performance across all metrics in both evaluation approaches. It achieves F1 scores of 0.959 (feature-level) and 0.962 (tuple-level), while the single-pass pipeline shows F1 scores of 0.869 (feature-level) and 0.880 (tuple-level). Compared to our best approach (multi-pass with source tracking), the single-pass pipeline shows performance gaps of 9.0 percentage points in feature-level evaluation and 8.2 percentage points in tuple-level evaluation.

The primary advantage of multi-pass pipeline with source tracking lies in recall performance, where our best approach achieves 0.940 and 0.947 recall in feature-level and tuple-based evaluation, respectively. The other approaches achieve recalls at the range of [0.814, 0.843]. The multi-pass processing without source tracking achieved improvements of 1.5 and 1.7% in F1 score but actually decreased recall performance by 1.3–1.4% compared to single-pass extraction. This modest F1 gain is driven entirely by improved precision (feature-level precision rising from 0.914 to 0.972), which is offset by the slight decline in recall. This result suggests that staged processing primarily trades recall for precision, rather than addressing the fundamental recall limitations of single-pass extraction. However, once source tracking is incorporated, both F1 score and recall improvements become substantial, with F1 score improving by 6.7–7.3% and recall improving by 11.7–12.6%. These huge gaps demonstrate the critical importance of maintaining source context throughout the extraction pipeline.

3.2 Material extraction completeness analysis

The ability to correctly identify and extract all materials is essential for building comprehensive databases and enabling data-driven materials discovery. Among the 396 materials

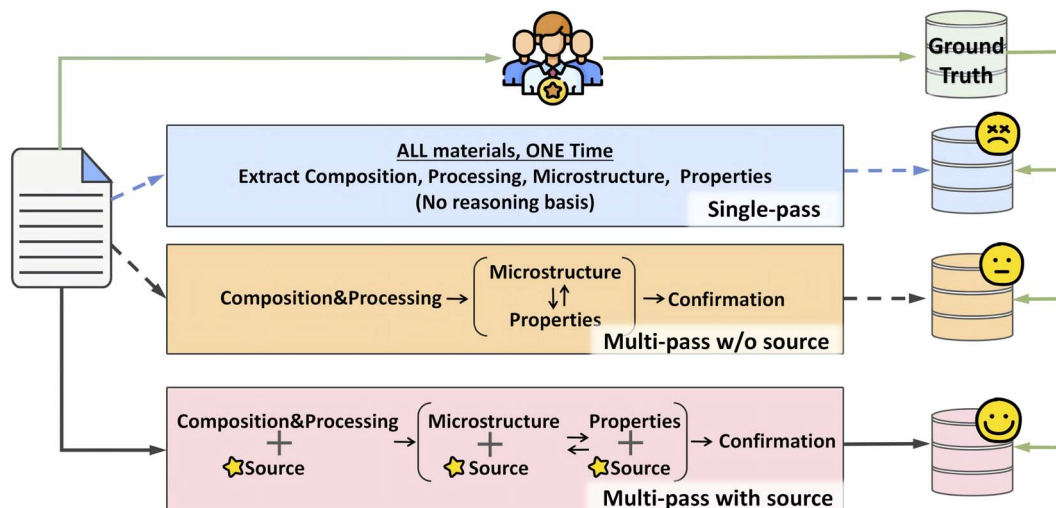


Fig. 3 Processes of three extraction pipelines.

present in our testing corpus, the source-tracking pipeline missed only 13 materials (3.3%), while the multi-pass pipeline without source tracking missed 43 materials (10.8%), and the single-pass extraction both missed 37 materials and incorrectly identified 12 non-existent materials, resulting in 49 total errors (12.4%).

The distribution of missed materials offers insights into the failure modes of each approach. For the multi-pass pipeline with source tracking, the 13 missed materials predominantly fall into two categories:

- Base material overlooked in favor of processed variants (7 cases): these involve as-manufactured materials that serve as the starting point for subsequent processing but lack explicit processing procedure descriptors. The extraction system prioritized materials with clear processing parameters (*e.g.*, “aged at 900 °C for 50 h”) while treating unprocessed base materials as mere reference states rather than independent research objects.
- Materials with implicit processing parameters (6 cases): the system missed materials where processing parameters were implicitly expressed rather than explicitly identified as separate research objects. For example, the text primarily focused on the 2 hours annealed material (“HGS-HP MEA annealed for 2 hours”) as the main research object, while the 30 minutes condition was mentioned only briefly in the context of microstructural evolution analysis. The extraction system failed to

recognize the 30 minutes annealed material as a separate, extractable material due to its implicit presentation, resulting in a missed material that was actually studied experimentally.

The multi-pass pipeline without source tracking resulted in significant omissions when extracting materials. For instance, when the text stated “Unlike the FSP condition, the FSA specimen showed different deformation mechanisms,” the framework without source tracking could not link it to proper material identity. The term “the FSA specimen” lost its specific referential meaning without the contextual framework. This led to the complete omission of the FSP condition from the extracted data. Without source anchoring, ambiguous references and pronouns lost their contextual meaning during validation. As a result, the system conflated materials or eliminated them entirely.

In addition to the limitations mentioned above, the single-pass extraction pipeline also suffers from the drawback of producing redundant information. The 12 non-existent materials were extracted due to misunderstandings of historical references or discussions of hypothetical scenarios rather than actual experimental material. When a paper discussed “Referring to S. Singh’s investigation, the formation of Al–Ni-rich phase ...”, the single-pass system interpreted this as the authors’ own work.

Table 1 Feature-level and tuple-level evaluation performance comparison of three extraction pipelines using o3-mini (With Extra/Missed). Bold values indicate the best performance for each metric

	Feature-level			Tuple-level		
	Single-pass	Multi-pass without source	Multi-pass with source	Single-pass	Multi-pass without source	Multi-pass with source
Precision	0.914	0.972	0.978	0.920	0.971	0.977
Recall	0.828	0.814	0.940	0.843	0.830	0.947
F1 score	0.869	0.886	0.959	0.880	0.895	0.962
Accuracy	0.845	0.864	0.950	0.842	0.860	0.948

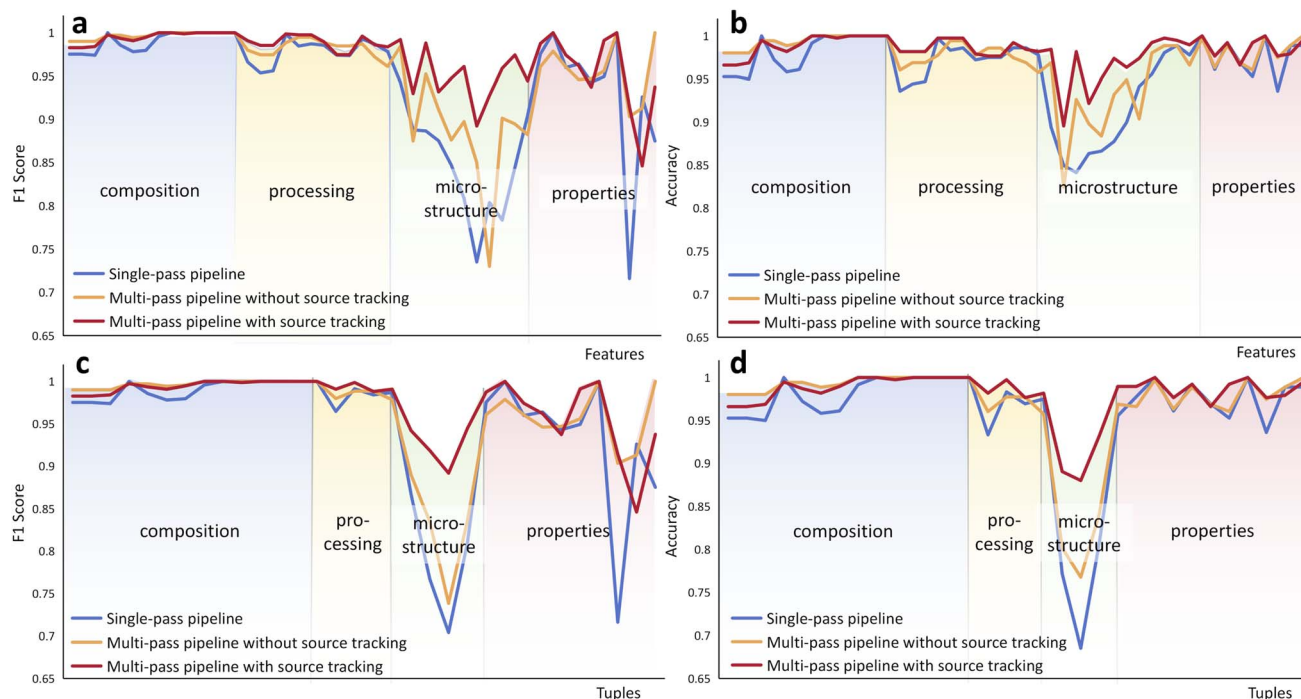


Fig. 4 Performance comparison of three extraction pipelines implemented using the o3-mini model across 47 features organized into composition, processing, microstructure, and properties categories. (a) Feature-level F1 score comparison showing individual feature extraction accuracy for each pipeline across the four categories. (b) Feature-level accuracy comparison demonstrating the proportion of correctly extracted features within each category. (c) Tuple-level F1 score comparison evaluating the extraction of interdependent feature groups, applied specifically to processing and microstructure categories where features exhibit strong relationships. (d) Tuple-level accuracy comparison measuring the correct extraction of complete feature tuples.

3.3 Feature performance analysis

The feature performance analysis focuses on the variations of extraction performance across 47 features in four categories. In this analysis, we excluded cases involving missed or extra materials, as each feature extracted from a missed material is directly treated as a false negative (FN), and each feature from an extra material is treated as a false positive (FP). Therefore, we focused this evaluation on correctly identified materials to ensure an isolated assessment of feature extraction quality.

To assess the average performance, we adopted two complementary evaluation strategies. The first approach, called feature-level evaluation, involved calculating a score for each individual feature, such as matrix phase type or ultimate tensile strength, and then averaging these scores within each of the four categories. This evaluation method treats each feature independently and provides an overall view of the model's performance across different features.

The second approach focused on groupings of interdependent features. This tuple-level evaluation was applied specifically to the processing and microstructure categories, where features exhibit strong interdependencies. For processing, we used tuples such as (aging applied?, time, temperature) or (rolling applied?, thickness reduction, temperature). For microstructure, tuples included (precipitate type, size, volume fraction) or (matrix type, volume fraction). A predicted tuple was counted as a true positive (TP) if it matched a ground truth tuple

exactly. If a tuple was absent from both predictions and ground truth, it was a true negative (TN). A tuple was a false positive (FP) if any component was incorrect, and a false negative (FN) if it lacked one or more components that were present in the ground truth. Notably, composition and properties features were evaluated only using the single-feature approach, as these features are typically reported and interpreted independently. Consequently, the tuple-level and feature-level evaluation results are the same for composition and properties categories.

Fig. 4 presents a comparison of the performance of all 47 features under two evaluation strategies. The results reveal performance patterns across four feature categories. In feature-level evaluation (panels a and b), the multi-pass pipeline with source tracking (red line) consistently maintains high performance across all categories, with most features achieving F1 scores and accuracy above 0.95. The multi-pass pipeline without source tracking (orange line) shows moderate performance with some fluctuations, while the single-pass pipeline (blue line) performs noticeably the worst, especially in the microstructure and property categories, where several features score below 0.75. Tuple-level evaluation (panels c and d) exhibits similar trends, but with slightly lower absolute scores, reflecting the increased difficulty of correctly extracting interdependent feature groups. The source-tracking pipeline shows a significant F1 difference only for two non-room-temperature attribute features; for all other features, it achieves either the best performance or performance very close to the best. Overall, the

Table 2 Precision and recall of different pipelines for composition (o3-mini)

	Precision			Recall		
	Single-pass	Multi-pass without source	Multi-pass with source	Single-pass	Multi-pass without source	Multi-pass with source
Fe	0.974	0.994	0.974	0.977	0.986	0.992
Ni	0.974	0.994	0.974	0.977	0.986	0.992
Co	0.972	0.994	0.976	0.977	0.986	0.992
Mn	1.000	0.994	0.997	1.000	1.000	0.997
Cr	0.972	0.994	0.995	1.000	1.000	0.992
Al	0.969	1.000	0.982	0.989	0.989	1.000
Ti	0.969	1.000	0.990	0.991	0.991	1.000
Cu	1.000	1.000	1.000	0.992	1.000	1.000
Si	1.000	1.000	1.000	1.000	1.000	1.000
V	1.000	1.000	0.997	1.000	1.000	1.000
Nb	1.000	1.000	1.000	1.000	1.000	1.000
B	1.000	1.000	1.000	1.000	1.000	1.000
Mo	1.000	1.000	1.000	1.000	1.000	1.000
Ta	1.000	1.000	1.000	1.000	1.000	1.000
Average	0.988	0.998	0.992	0.993	0.996	0.998

source-tracking pipeline demonstrates strong effectiveness in maintaining contextual relationships among related features. While F1 scores and accuracy provide a balanced perspective on overall performance, detailed precision and recall analysis at the feature level can offer deeper insights into extraction bottlenecks and identify specific failure modes within each category. The following subsections examine precision and recall patterns to better understand the underlying mechanisms driving these performance differences.

3.3.1 Composition and processing features. All three pipelines showed high recall. For composition features, all

three pipelines demonstrated high average recall and precision, as shown in Table 2. The single-pass pipeline achieved a precision of 0.988 and a recall of 0.993, the multi-pass pipeline without source tracking reached a precision of 0.998 and a recall of 0.996, while the multi-pass pipeline with source tracking attained a precision of 0.992 and a recall of 0.998.

This consistently high performance across all pipelines can be attributed to several factors. First, elemental composition is reported using highly standardized notation and terminology. Second, the extraction task operates within a constrained domain: the periodic table contains just over a hundred

Table 3 Processing precision of feature-level and tuple-level under different pipelines (o3-mini). Values are rounded to three decimal places; bold numbers indicate the highest precision before rounding

	Feature level			Tuple level		
	Single-pass	Multi-pass without source	Multi-pass with source	Single-pass	Multi-pass without source	Multi-pass with source
Homogenization						
Homogenization?	0.934	0.977	0.982	0.931	0.977	0.982
Homogenization temperature	0.913	0.964	0.972			
Homogenization time	0.917	0.964	0.972			
Rolling						
Rolling?	0.997	0.994	0.997	0.994	0.994	0.997
Rolling temperature	0.990	0.990	0.995			
Rolling percentage	0.995	0.990	0.995			
Recrystallization						
Recrystallization	0.980	0.994	0.992	0.979	0.994	0.992
Recrystallization temperature	0.971	0.988	0.983			
Recrystallization time	0.971	0.988	0.983			
Aging						
Aging?	0.988	0.994	0.995	0.980	0.977	0.987
Aging temperature	0.990	0.979	0.977			
Aging time	0.974	0.959	0.986			
Average	0.968	0.982	0.986	0.971	0.985	0.989

Table 4 Recall of feature-level and tuple-level under different pipelines for processing (o3-mini)

	Feature level			Tuple level		
	Single-pass	Multi-pass without source	Multi-pass with source	Single-pass	Multi-pass without source	Multi-pass with source
Homogenization						
Homogenization?	1.000	0.983	1.000	1.000	0.982	1.000
Homogenization temperature	1.000	0.986	1.000			
Homogenization time	1.000	0.986	1.000			
Rolling						
Rolling?	1.000	0.983	1.000	0.988	0.983	1.000
Rolling temperature	0.980	1.000	1.000			
Rolling percentage	0.980	1.000	1.000			
Recrystallization						
Recrystallization	0.991	0.983	0.987	0.988	0.983	0.984
Recrystallization temperature	0.977	0.982	0.966			
Recrystallization time	0.976	0.982	0.966			
Aging						
Aging?	0.997	0.980	0.997	0.994	0.980	0.995
Aging temperature	0.984	0.965	0.995			
Aging time	0.984	0.964	0.982			
Average	0.989	0.983	0.991	0.992	0.982	0.995

elements, which significantly reduces ambiguity. Third, composition data typically appears in well-defined sections of papers, often presented in tables or clearly delineated text passages with explicit quantitative values (e.g., $\text{Cr}_{30}\text{Co}_{30}\text{Ni}_{30}\text{-Al}_5\text{Ti}_5$ (at%)). In this scenario, even the simplest single-pass approach can achieve near-perfect performance.

Meanwhile, three pipelines demonstrated high recall and precision in extracting processing features under both evaluation methods, with results ranging from 0.968 to 0.995 (Table 3 and 4). Among them, the multi-pass pipeline with source tracking performed best, achieving a precision of 0.986 and a recall of 0.991 in the feature-level evaluation, and a precision of 0.989 and a recall of 0.995 in the tuple-level evaluation.

Similar to the performance observed in the composition extraction, this performance can be attributed to the standardized and structured language. Terminology for processing steps, such as homogenization and rolling, is consistently used across articles with limited lexical variation. Because LLMs rely on statistical associations between words, this kind of formulaic and repetitive phrasing reduces ambiguity and enhances pattern recognition. In addition, processing descriptions tend to follow predictable grammatical structures and are written in imperative or passive voice. Common formats include sequential procedural descriptions, such as: “After that, the recrystallized samples were further aged at 780 °C for 24 h to introduce precipitates.” They also include conditional statements and parameter descriptions with clearly defined value pairs. These linguistic regularities facilitate automated parsing, as processing steps typically follow a well-structured, logically ordered timeline with explicit quantification.

3.3.2 Microstructure features. Microstructure information extraction presented the biggest challenges in our study due to

its complexity (Table 5 and 6). The single-pass pipeline achieved 0.880 precision and 0.828 recall on microstructure features. Under tuple-level evaluation, the performance further degraded to 0.809 precision and 0.784 recall.

This performance gap compared to other feature categories led us to investigate the underlying causes:

- Cognitive load and task complexity issues: single-pass pipeline requires LLM to complete multiple complex tasks in one processing pass: identifying all materials, extracting 47 features, understanding interdependencies between features, and correctly associating scattered information. As a result, the LLMs experienced cognitive overload.

- Inherent hierarchical structure: microstructure identification often depends on composition and processing conditions. While single-pass pipeline “sees” all information, it may not have sufficiently “understood” compositional and processing information when collecting microstructure information.

- Attention mechanism limitations: information scattered across different sections of papers. Single-pass requires the model to simultaneously track multidimensional information for multiple materials across thousands of words.

Multi-pass pipeline overcomes these problems by first extracting composition and processing, then microstructure, followed by properties, and finally performing a validation of all extracted features against the original text. During the staged extraction process, we provide both the already-extracted features and the original text. This extraction sequence follows the writing order of most material science articles.

This contextual enrichment mechanism slightly improved performance, with feature-level evaluation results showing 0.854 precision and 0.927 recall. Compared to the single-pass pipeline, the multi-pass pipeline without source tracking

Table 5 Precision of feature-level and tuple-level under different pipelines for microstructure (o3-mini)

	Feature level			Tuple level		
	Single-pass	Multi-pass without source	Multi-pass with source	Single-pass	Multi-pass without source	Multi-pass with source
Matrix						
Matrix type	0.923	0.977	0.990	0.854	0.844	0.900
Matrix percentage	0.886	0.830	0.876			
Precipitate 1						
Precipitate 1 type	0.859	0.932	0.980	0.721	0.756	0.866
Precipitate 1 size	0.861	0.860	0.884			
Precipitate 1 percentage	0.847	0.818	0.914			
Precipitate 2						
Precipitate 2 type	0.839	0.881	0.953	0.741	0.672	0.841
Precipitate 2 size	0.794	0.836	0.879			
Precipitate 2 percentage	0.815	0.635	0.878			
Precipitate 3						
Precipitate 3 type	0.906	0.941	0.972	0.919	0.824	0.944
Precipitate 3 size	0.950	0.895	1.000			
Precipitate 3 percentage	1.000	0.789	0.944			
Average	0.880	0.854	0.934	0.809	0.774	0.888

achieved 2.6% lower accuracy and 10.1% higher recall. Specifically, matrix type and each precipitate type showed an average precision improvement of 5.1%, while matrix percentage, precipitate size, and precipitate percentage showed minimal precision improvements or even significant decreases; matrix type and precipitate type showed an average recall improvement of 9.1%, while matrix, precipitate type, size, and percentage displayed an average recall improvement of 10.5%. Tuple-level evaluation results showed 0.774 precision and 0.882 recall.

Compared to the single-pass pipeline, the multi-pass pipeline without source tracking exhibited similar magnitudes of change in overall accuracy, recall, and the accuracy and recall of individual components.

The multi-pass approach reduced the complexity of each task and focused on specific information types at each step, significantly enhancing the model's recall performance. Although significant improvements were achieved in recall, accuracy decreased by approximately 2%. Specifically, matrix

Table 6 Recall of feature-level and tuple-level under different pipelines for microstructure (o3-mini)

	Feature level			Tuple level		
	Single-pass	Multi-pass without source	Multi-pass with source	Single-pass	Multi-pass without source	Multi-pass with source
Matrix						
Matrix type	0.966	0.994	0.997	0.885	0.940	0.991
Matrix percentage	0.886	0.919	0.989			
Precipitate 1						
Precipitate 1 type	0.926	0.977	0.997	0.836	0.928	0.977
Precipitate 1 size	0.892	0.968	0.986			
Precipitate 1 percentage	0.853	0.935	0.983			
Precipitate 2						
Precipitate 2 type	0.803	0.912	0.968	0.692	0.813	0.946
Precipitate 2 size	0.685	0.864	0.906			
Precipitate 2 percentage	0.815	0.870	0.985			
Precipitate 3						
Precipitate 3 type	0.690	0.865	0.946	0.723	0.848	0.944
Precipitate 3 size	0.760	0.895	0.950			
Precipitate 3 percentage	0.826	1.000	0.944			
Average	0.828	0.927	0.968	0.784	0.882	0.965

type and precipitate type showed improved precision and recall with good post-improvement performance. However, matrix percentage, precipitate percentage, and size exhibited minimal precision improvements. This pattern reveals why recall increased while precision slightly decreased. The multi-pass pipeline became more likely to attempt extraction of unreported values, successfully capturing many previously missed true positives (reducing false negatives), but the model also predicted more incorrect “has values,” increasing false positives and slightly decreasing precision.

To address these precision issues while maintaining the improved recall, we implemented source tracking in our multi-pass pipeline. Source tracking enhances attention by improving signal-to-noise ratio through better attention patterns^{44,45} and prevents cascade errors while enabling precise error localization during validation.⁴⁶ This pipeline achieved significant performance improvements in microstructure extraction. In feature-level evaluation, it achieved 0.934 precision and 0.968 recall, improving by 8.0% and 4.1% respectively compared to the multi-pass pipeline without source tracking, and by 5.4% and 14.1% compared to single-pass extraction. In tuple-level evaluation, performance reached 0.888 precision and 0.965 recall, representing improvements of 11.4% and 8.3% over the non-source-tracking multi-pass approach, and 7.9% and 18.1% over single-pass extraction.

A large portion of the errors is still caused by the model's inference attempts. When the model performed reasoning, it sometimes marked outputs as “estimated,” or included intermediate calculation steps in its reasoning process. Following strict evaluation criteria, we counted these incorrectly inferred values as false positives even when explicitly marked as estimates. This is because our basic facts define positive examples as values that are clearly present in the original text or correct values obtained through reasoning. However, the multi-pass pipeline with source tracking already showed a considerable reduction in erroneous or unnecessary inference compared to the other two pipelines.

The multi-pass pipeline with source tracking thus represents an optimal balance. It maintains the high recall achieved

through staged extraction while dramatically improving precision through source anchoring. This approach successfully extracted 96.5% of all true microstructure values in tuple-level evaluation. Even when unnecessary or erroneous inference leads to false positives, these can be filtered out by discarding values marked as inferred.

A instance illustrates this mechanism. For the alloy aged at 800 °C for 16 h, the article reports only TEM-measured precipitate sizes ($\sim 0.30\ \mu\text{m}$ for the Heusler phase, $\sim 35\ \text{nm}$ for the L_{12} phase) and gives no precipitate volume fraction for this specific condition. With source tracking, the model anchors its extraction to this size-only passage and returns “Not specified” for both precipitate volume fractions, matching the ground truth. Without source tracking, the multi-pass pipeline instead retrieves volume fractions from unrelated contexts—a Heusler value measured for a different (1 h) condition and an L_{12} value taken from a theoretical strengthening calculation in the Discussion—producing two false positives. The single-pass pipeline returns a hallucinated round figure for both. A full walkthrough of the prompts, JSON outputs, and source-text strings for this case is given in SI Section S1.4.

3.3.3 Property features. Property extraction achieved strong performance across all pipelines (Table 7). The multi-pass pipeline with source tracking reached 0.941 precision and 0.991 recall. Compared to single-pass extraction, this represents a shift: recall increased significantly by 9.5% (from 0.896 to 0.991) while precision decreased by 3.7% (from 0.941 to 0.978).

Analysis of the 40 property extraction errors from the multi-pass pipeline with source tracking reveals three primary failure modes.

Confusion errors, comprising 62.5% of total errors, predominantly involve attributing properties from one material to another. Three distinct patterns emerged from our analysis. First, academic terminology creates interpretation challenges. When encountering “The hardness value increases from the solutionized condition of $330 \pm 8\ \text{HV}$ to a peak value of $380 \pm 5\ \text{HV}$ after 50 h of aging,” the model failed to recognize “solutionized condition” as the initial state (0 h), instead treating the 330 HV value as belonging to an intermediate aging

Table 7 Precision and recall of different pipelines for property (o3-mini)

	Precision			Recall		
	Single-pass	Multi-pass without source	Multi-pass with source	Single-pass	Multi-pass without source	Multi-pass with source
Ultimate tensile strength	0.976	0.967	0.981	0.976	0.955	0.994
Ultimate compressive strength	1.000	1.000	1.000	1.000	0.957	1.000
Tensile yield strength	0.950	0.947	0.956	0.972	0.976	0.994
Compressive yield strength	1.000	0.972	0.927	0.929	0.921	1.000
Hardness	1.000	0.951	0.951	0.891	0.942	0.924
Tensile ductility	0.958	0.956	0.982	0.941	0.956	1.000
Compressive ductility	1.000	1.000	1.000	1.000	1.000	1.000
NRT-cryo strength heat-treated temperature	0.967	0.955	0.895	0.569	0.857	1.000
NRT-cryo heat-treated strength	1.000	0.955	0.833	0.862	0.875	1.000
NRT-cryo heat-treated ductility	0.933	1.000	0.882	0.824	1.000	1.000
Average	0.978	0.970	0.941	0.896	0.944	0.991

condition. Second, incomplete property reporting across multiple samples led to mis-attribution. When papers focus on specific phenomena such as long-term stability, they may report complete property sets only for key conditions while providing partial data for others. The model, attempting to maintain completeness, occasionally misattributed properties from well-characterized samples to incompletely characterized ones. Third, contextual emphasis created confusion. When a paper stated “large ductility over 22% and good phenomenally strain-hardening capability over 400 MPa can be well retained even at a high yield strength approaching 1.0 GPa,” the model incorrectly extracted 1.0 GPa as the material’s exact yield strength, failing to recognize that the emphasis was on the 22% ductility retention at high strength levels, with the actual yield strength reported elsewhere.

Extra inference errors, comprising 12.5% of all failures, resulted from the LLM inappropriately applying theoretical models rather than limiting itself to data re-expression. A telling example occurred when the text stated “the base strength (*i.e.*, unstrengthened matrix) is 165 MPa and the L_{12} particle shear strengthening contribution is 244.3 MPa.” LLM incorrectly derived a total strength of 409 MPa ($165 + 244$) by applying a linear addition strengthening model, treating this theoretical prediction as an experimental value. While reasonable inferences such as calculating atomic percentages from chemical formulae (*e.g.*, $\text{Fe}_2\text{Al}_5 \rightarrow \text{Fe}: 28.6 \text{ at\%}, \text{Al}: 71.4 \text{ at\%}$) represent legitimate re-expression of experimental data, this case involved unauthorized theoretical modeling that crossed the boundary from data transformation into model prediction. The challenge lies in controlling LLM behavior to distinguish between appropriate format conversions and inappropriate theoretical calculations.

Omission errors constitute 25% of the total and primarily result from indirect property reporting. For example, when papers use expressions like “1200 MPa higher than the ST condition” without explicitly stating the baseline value, the model fails to extract this information, unlike when values are directly stated, such as “ ~ 274 MPa” for yield strength. Models perform unnecessary computations in inference errors and fail to perform necessary computations in omissions. This means that the model lacks the domain knowledge to understand when such calculations are physically meaningful.

3.4 Comparison with an open-source LLM (Gpt-oss-20B)

To assess the performance of our pipeline on open-source models, we evaluated the same multi-pass source-tracking pipeline using Gpt-oss-20B, a 20-billion-parameter open-source language model, as the model in place of o3-mini. All prompt designs, extraction passes, and evaluation protocols were kept identical to enable a fair comparison. All prompt designs, extraction passes, and evaluation protocols were kept identical to enable a controlled comparison. Note that a single fixed prompt design may not suit every model equally well, since different LLMs can favor different prompting conventions. Gpt-oss-20b (20.9B parameters, MXFP4 quantization, model digest 17052f91a42e) was served locally through Ollama 0.14.2, and

Table 8 Performance comparison of gpt-oss-20B and o3-mini, both using the multi-pass pipeline with source tracking. Bold values indicate better performance for each metric

	Feature-level		Tuple-level	
	Gpt-oss-20B	o3-mini	Gpt-oss-20B	o3-mini
Precision	0.843	0.978	0.854	0.977
Recall	0.766	0.940	0.802	0.947
F1 score	0.803	0.959	0.827	0.962
Accuracy	0.780	0.950	0.783	0.948

results are reported at the model’s default reasoning level. The reasoning-effort control is not exposed identically across the two models: for o3-mini it is a first-class API parameter (set to “high”), whereas for gpt-oss-20b it is mediated by the serving harness and chat template. We report results for gpt-oss-20b at its default reasoning level. Because the two models expose reasoning control through different mechanisms, an identical prompting protocol does not guarantee an equivalent reasoning configuration. Accordingly, this comparison should be interpreted as an assessment of the pipeline’s out-of-the-box transferability to an open-weight model under matched prompts, rather than as an upper bound on the achievable performance of gpt-oss-20b.

Table 8 presents the performance of gpt-oss-20B under the same evaluation framework as Table 1. At the feature level, gpt-oss-20B achieved a precision of 0.843 and a recall of 0.766, yielding an F1 score of 0.803 and an accuracy of 0.780. Tuple-level evaluation produced slightly higher precision (0.854) but a similar recall of 0.802, with an F1 of 0.827 and an accuracy of 0.783. These results represent a substantial gap relative to o3-mini under the same pipeline configuration, which achieved feature-level and tuple-level F1 scores of 0.959 and 0.962, respectively—differences of 15.6 and 13.5 percentage points.

The material extraction completeness analysis further highlights this gap. Gpt-oss-20B missed 37 out of 396 materials (9.3%) and incorrectly identified 11 non-existent materials, resulting in 48 total material-level errors. By contrast, o3-mini missed only 13 materials (3.3%) with zero false positive materials. The higher false positive rate under gpt-oss-20b suggests a greater tendency to surface unsupported values. The performance gap likely reflects a combination of factors, including differences in instruction-following behavior and reasoning depth, as well as the reasoning-configuration asymmetry noted above. The multi-pass source-tracking pipeline requires the model to maintain cross-stage contextual links, resolve ambiguous material references, and distinguish between experimentally reported values and inferred or secondary data. These tasks place substantial demands on a model’s reasoning and instruction-following capacity. gpt-oss-20b’s lower recall indicates that, under this configuration, the open-weight model more frequently fails to surface values that are implicitly or indirectly expressed in the source text. These findings indicate that while the multi-pass source-tracking architecture is broadly applicable across LLMs, the extraction quality obtained from

a given model remains sensitive to both its reasoning and instruction-following capacity and its serving configuration.

4 Discussion

4.1 Multi-pass pipeline

The multi-pass architecture's advantages arise from the hierarchical dependencies inherent in materials science. Composition and processing collectively affect microstructures, which in turn play a significant role in governing properties. Our four-pass design mirrors typical materials science paper structure and enables progressive construction of complete material information graphs. The systematic extraction of 47 features across the entire composition–process–microstructure–property chain represents unprecedented comprehensiveness. This approach preserves not just individual features but their critical interconnections.

Building on this architectural foundation, we next introduce the role of source tracking in ensuring reliability. Source tracking introduces a cascade error correction mechanism essential for maintaining extraction fidelity. By preserving the original text context, the source text serves as an anchor and avoids reference loss. When encountering ambiguous references like “the FSA specimen,” the associated source text maintains proper material identification across extraction passes. During final validation, the system cross-references all extracted information against original sources, enabling identification and correction of propagated errors. This retrospective validation capability explains our achievement of zero false positive materials while maintaining exceptional precision.

4.2 Performance and advantages

The integration of source tracking with multi-pass extraction represents a fundamental advance in automated scientific information extraction. The reduction in missed materials from 43 (10.8%) to 13 (3.3%) indicates the positive impact of source tracking. The impact is most pronounced in microstructure extraction, where tuple-level evaluation showed precision improving from 0.809 to 0.888 and recall from 0.784 to 0.965. Composition and processing extraction achieved over 98% accuracy across all pipelines. This high performance highlights LLMs' proficiency with standardized scientific language. The success can be attributed to the formulaic nature of chemical notation and processing descriptions, which follow predictable patterns that align well with LLM training. Source tracking's effectiveness further shows sophisticated contextual understanding—the model correctly resolves material references, interprets abbreviations, and maintains entity consistency when provided with anchoring text.

Our evaluation reveals both capabilities and limitations of current LLMs in materials science. For example, when processing microstructure type encoding, the system needed to recognize diverse expressions such as σ phase, Eutectic $\gamma + \gamma'$ phase, ε -phase, and γ' precipitates, then determine whether these correspond to our predefined encoding system (L12, γ'' , B2, *etc.*). This task requires not merely surface text

understanding but mastery of crystallographic nomenclature, microstructure knowledge, and materials science terminology. The model's success in navigating these complexities while maintaining high accuracy reflects deep domain knowledge.

The material-level reliability of our pipeline has direct consequences for downstream applications in materials informatics. The extracted materials data are frequently used as training inputs for machine learning models designed to predict material properties. Omission of a substantial number of materials reduces the amount of available training data, while the inclusion of hallucinated materials introduces noise that can degrade model performance. Prior studies have shown that even 10% label noise can significantly increase the uncertainty and dispersion of model predictions, and when noise levels reach several tens of percent, model performance deteriorates markedly.^{47,48} By maintaining a 3.3% miss rate and zero false positive materials, our pipeline ensures that the resulting database is of sufficient quality to serve as a reliable foundation for materials informatics applications.

Beyond aggregate accuracy, these metrics also reveal how faithfully the extracted data can help uncover the composition–processing–microstructure–property (CPMP) relationships that motivate this work. Low recall means that some reported composition, processing, microstructure, and property information is missed in the extracted data. This may lead to downstream machine learning models missing the opportunity to identify meaningful patterns in CPMP relationships. Low precision means that wrong information is extracted, leading to incorrect relationships that can suggest a non-existent composition–processing–structure–property correlation and misdirect alloy design. Our tuple-level evaluation targets exactly this by verifying that interdependent descriptors remain linked to their parent alloy, so the tuple-level gap reflects a more reliable basis for designing precipitate-strengthened alloys.

4.3 Limitations and challenges

Despite the high extraction accuracy achieved by the proposed framework, several limitations remain and highlight directions for future improvement.

The most significant challenge concerns inappropriate inference when the model encounters incomplete information. For example, when a paper stated that “precipitate A comprises 20% volume fraction,” the model sometimes assumed the matrix occupied the remaining 80% and ignored the possibility of additional unreported phases. Similar reasoning errors appeared in several microstructure extraction tasks. When aging data were reported under different conditions, the model occasionally performed unsupported interpolation or extrapolation. For instance, given L12 precipitates are ~ 5 nm at 550 °C/150 h and ~ 75 nm at 700 °C/50 h,” the system could incorrectly infer precipitate sizes under unreported conditions despite the absence of experimental evidence. The model also struggled with mixed-phase descriptions, such as determining whether a reported mixed BCC/B2 phase” should be represented as two distinct phases or as a single mixed-structure precipitate. Additional difficulties arose during hierarchical information

aggregation, where regional measurements were sometimes incorrectly generalized to represent bulk material behavior. Similar issues were observed when converting qualitative descriptions into quantitative values, particularly for semi-quantitative expressions such as “less than 5%” or “approximately mid-to-low 300 MPa.”

Comparable reasoning deficits were observed during property extraction. Approximately 12.5% of extraction errors resulted from inappropriate theoretical modeling rather than extraction of experimentally reported values. In these cases, the model generated predicted values based on theoretical relationships that were never explicitly reported in the source article. These observations reveal a fundamental limitation of LLM-based scientific extraction: although LLMs are highly effective at pattern recognition and information retrieval, they do not inherently distinguish between experimentally measured quantities and theoretically inferred quantities.

Future work should therefore focus on developing more explicit constraint mechanisms. One promising direction is to classify extraction operations into permitted and prohibited categories. Permitted operations may include unit conversion and stoichiometric calculations directly derived from chemical formulae, whereas prohibited operations would include strength prediction, interpolation between experimental conditions, and other forms of model-based inference. These constraints could be enforced through dedicated validation rules rather than relying solely on source-text anchoring. Such an approach may reduce unauthorized reasoning while preserving extraction completeness.

A second limitation concerns the non-deterministic nature of large language models. In this study, each article was processed once per pipeline configuration rather than through multiple repeated extraction runs. To reduce output variability, o3-mini was configured with `reasoning_effort = “high”`, and the structured JSON output format combined with source-tracking requirements constrained the space of valid completions. Nevertheless, some degree of run-to-run variability remains unavoidable. Future studies should investigate consensus-based strategies such as repeated extraction and majority voting to improve robustness and reproducibility.

A third limitation is that the current framework processes only textual content and tables, while information embedded in figures, charts, and micrographs is not utilized. This omission can be significant because materials science figures often contain morphological and spatial information that is difficult to describe completely in text. Recent advances in scientific figure mining have demonstrated the feasibility of extracting structured information directly from images and graphical data. Future versions of the framework could integrate multi-modal extraction capabilities, allowing textual and visual information to be analyzed jointly. Such integration would provide a more comprehensive representation of material systems by combining quantitative measurements from text with microstructural information from figures.

Finally, the current multi-pass pipeline requires substantial computational resources. On average, approximately 21 500 tokens are consumed per paper, including both inputs and

outputs, because each pass receives the complete article text together with previously extracted information. Although this design improves extraction consistency, it also increases computational cost and limits scalability. Future work should explore more efficient processing strategies, including selective context windowing, retrieval-based context selection, and incremental extraction workflows that minimize redundant token usage while preserving extraction accuracy.

Similar reasoning deficits appeared in property extraction, where 12.5% of errors involved inappropriate theoretical modeling rather than limiting extraction to experimental data. As discussed earlier, the extraction system applied theoretical models to derive predicted values that were never experimentally measured. These parallel failures across microstructure and properties extraction reveal a fundamental limitation: while LLMs excel at pattern recognition and information retrieval, they struggle to distinguish between experimental measurements and theoretical predictions in scientific literature. This observation raises the question of whether explicit prompt-level constraints could prevent such unauthorized modeling. In our current implementation, the primary mechanism for constraining extraction behavior is the source-tracking requirement. Each extracted value must be accompanied by a supporting source text passage drawn from the original article. This requirement implicitly discourages unsupported inference because a calculated or modeled value, such as a total strength derived by linearly summing strengthening contributions, is not explicitly stated in the article. However, the source-tracking mechanism alone is insufficient to eliminate this failure mode entirely. During step 4 validation, the model is instructed to check the consistency of phase volume percentages. This instruction can inadvertently encourage the model to infer or adjust values through calculation rather than limiting itself strictly to reported data. These parallel observations point to a fundamental tension in prompt design. Instructions that promote completeness and internal consistency, such as checking that volume fractions sum to 100%, can conflict with instructions that promote strict adherence to experimental reporting. Future work should explore more precise constraint strategies. One possibility is to explicitly classify extraction operations into permitted categories, such as unit conversion and stoichiometry calculation from chemical formulae, and prohibited categories, such as model-based strength prediction and interpolation between experimental conditions. These boundaries could then be enforced as separate validation rules rather than relying solely on source-text anchoring.

5 Conclusion

We have presented a multi-pass, source-tracked pipeline named MetMINE for automated extraction of structured materials information from scientific literature. The target is the complete composition–processing–microstructure–property chain in precipitate-containing MPEAs. By decomposing extraction into sequential passes and anchoring every extracted value to its supporting source text, MetMINE addresses two

fundamental failure modes: material omission and precision loss.

Evaluated on 100 peer-reviewed articles, the pipeline achieves F1 scores of 0.959 (feature-level) and 0.962 (tuple-level), and maintains strong performance on the most challenging microstructure extraction task (F1 = 0.950 feature-level, 0.924 tuple-level). Source tracking reduced the missed-material rate from 12.4% to 3.3% relative to single-pass extraction, with zero false positive materials, producing a database suitable as training input for downstream machine learning and materials informatics models.

Our multi-pass source-tracking approach offers several advantages for material information extraction. Compared with manual extraction, the system significantly reduces processing time while maintaining a high F1 score and accuracy, making large-scale literature mining practically feasible. Our 47-feature framework captures a relatively complete material feature space in a unified structured format, including composition, processing, structure, and property descriptors. The resulting low material miss rate and zero false positives provide a high-quality foundation for downstream machine learning applications.

Although demonstrated on alloy systems, the framework is broadly applicable to other classes of materials. The composition–process–microstructure–property paradigm is common across many materials domains, and the modular feature design enables adaptation through feature modification while preserving the core extraction architecture.

Beyond database construction, the extracted structured information can support knowledge graph development, cross-system knowledge transfer, and automated hypothesis generation. Combined with advances in multimodal extraction, reasoning-constrained validation, and computational efficiency, future generations of such systems may contribute to continuously updated materials knowledge bases and data-driven materials discovery workflows.

Author contributions

S. X. and K. L. conceived the study. X. W. performed pipelines development, testing, and analysis. A. R., M. L., H. W., S. X., and X. W. completed dataset creation and validation. K. L. supervised the research and designed the task evaluation metrics. X. W. wrote the manuscript, and all authors contributed to manuscript revision and editing.

Conflicts of interest

The authors declare no competing interests.

Data availability

All research articles used for evaluating the pipeline in this work were obtained through institutional licensing agreements between the University of Oklahoma and major academic publishers, including Elsevier, Wiley, American Chemical Society (ACS), Springer Nature, Taylor & Francis, among others.

These agreements ensure legitimate access to the scientific literature required for comprehensive evaluation of our pipelines. Regarding the use of these texts with the OpenAI API, under current API usage policies at the time of submission, content submitted *via* the API is not used for model training by default. This design ensures that the submitted material remains separate from the model training pipeline.

The dataset generated from this study has been published as a data descriptor by Raj *et al.*⁴⁹ and is available on Zenodo.⁵⁰ Additional materials supporting this work, including the detailed confusion matrix classifications, code for the three extraction pipelines, and the DOIs of the 100 analyzed articles, are available at https://github.com/kun-ou-projects/4categories_materials_extraction.git.

Supplementary information (SI): a complete example of the information extraction pipeline. See DOI: <https://doi.org/10.1039/d6dd00374e>.

Acknowledgements

Some of the computing for this project was performed at the OU Supercomputing Center for Education & Research (OSCCER). X. W. and A. R. acknowledge support from the Graduate Student Seed Funding Program of the Data Institute for Societal Challenges at the University of Oklahoma. H.W. acknowledges the financial support by the U.S. Nuclear Regulatory Commission Faculty Development Program (award number NRC 31310018M0044). S. X. was supported by a grant from the Research Council of the University of Oklahoma Norman Campus. A. R. and S. X. acknowledge the support of the U.S. National Science Foundation (DMREF-2522655). Financial support was provided by the University of Oklahoma Libraries' Open Access Fund for the article processing charge.

References

- 1 J. J. Pablo, N. E. Jackson, M. A. Webb, L.-Q. Chen, J. E. Moore, D. Morgan, R. Jacobs, T. Pollock, D. G. Schlom, E. S. Toberer, *et al.*, New frontiers for the materials genome initiative, *npj Comput. Mater.*, 2019, **5**(1), 41.
- 2 M. E. Deagen, L. C. Brinson, R. A. Vaia and L. S. Schadler, The materials tetrahedron has a “digital twin”, *MRS Bull.*, 2022, **47**(4), 379–388.
- 3 A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, *et al.*, Commentary: The materials project: A materials genome approach to accelerating materials innovation, *APL Mater.*, 2013, **1**(1), 011002.
- 4 J. Hu, S. Stefanov, Y. Song, S. S. Ome, S.-Y. Louis, E. M. Siriwardane, Y. Zhao and L. Wei, MaterialsAtlas.org: a materials informatics web app platform for materials discovery and survey of state-of-the-art, *npj Comput. Mater.*, 2022, **8**(1), 65.
- 5 L. Sbailò, Á. Fekete, L. M. Ghiringhelli and M. Scheffler, The NOMAD artificial-intelligence toolkit: turning materials-science data into knowledge and understanding, *npj Comput. Mater.*, 2022, **8**(1), 250.

- 6 S. Curtarolo, W. Setyawan, G. L. Hart, M. Jahnatek, R. V. Chepulskii, R. H. Taylor, S. Wang, J. Xue, K. Yang, O. Levy, *et al.*, AFLOW: An automatic framework for high-throughput materials discovery, *Comput. Mater. Sci.*, 2012, **58**, 218–226.
- 7 A. Zakutayev, N. Wunder, M. Schwarting, J. D. Perkins, R. White, K. Munch, W. Tumas and C. Phillips, An open experimental database for exploring inorganic materials, *Sci. Data*, 2018, **5**(1), 1–12.
- 8 B. Lafuente, R. T. Downs, H. Yang and N. Stone, *The power of databases: The RRUFF project*, Walter de Gruyter GmbH, Berlin, Boston, 2016, pp. 1–29, DOI: [10.1515/9783110417104-003](https://doi.org/10.1515/9783110417104-003).
- 9 L. Bornmann, R. Haunschild and R. Mutz, Growth rates of modern science: a latent piecewise growth curve approach to model publication numbers from established and new literature databases, *Hum. Soc. Sci. Commun.*, 2021, **8**(1), 1–15.
- 10 J. Schmidt, M. R. Marques, S. Botti and M. A. Marques, Recent advances and applications of machine learning in solid-state materials science, *npj Comput. Mater.*, 2019, **5**(1), 83.
- 11 M. A. Groeber and M. A. Jackson, DREAM. 3D: a digital representation environment for the analysis of microstructure in 3d, *Integr. Mater. Manuf. Innov.*, 2014, **3**(1), 56–72.
- 12 P. Zhao, Y. Wang, B. Jiang, M. Wei, H. Zhang and X. Cheng, A new method for classifying and segmenting material microstructure based on machine learning, *Mater. Des.*, 2023, **227**, 111775.
- 13 B. L. DeCost and E. A. Holm, A computer vision approach for automated analysis and classification of microstructural image data, *Comput. Mater. Sci.*, 2015, **110**, 126–133.
- 14 R. Paudel and M. N. Al-Ameen, Priming through persuasion: Towards secure password behavior, *Proc. ACM Hum.-Comput. Interact.*, 2024, **8**(CSCW1), 1–27.
- 15 M. Calvat, C. Bean, D. Anjaria, H. Park, H. Wang, K. Vecchio and J. Stinville, Learning metal microstructural heterogeneity through spatial mapping of diffraction latent space features, *npj Comput. Mater.*, 2025, **11**(1), 284.
- 16 C. Zhu and M. Seita, Microstructure analysis on complex surfaces enables digital quality control of metal parts, *npj Comput. Mater.*, 2024, **10**(1), 279.
- 17 S. V. Kalinin, B. G. Sumpter and R. K. Archibald, Big-deep-smart data in imaging for guiding materials design, *Nat. Mater.*, 2015, **14**(10), 973–980.
- 18 O. Wodo, S. Broderick and K. Rajan, Microstructural informatics for accelerating the discovery of processing–microstructure–property relationships, *MRS Bull.*, 2016, **41**(8), 603–609.
- 19 D. M. Jessop, S. E. Adams, E. L. Willighagen, L. Hawizy and P. Murray-Rust, OSCAR4: a flexible architecture for chemical text-mining, *J. Cheminf.*, 2011, **3**(1), 41.
- 20 L. Hawizy, D. M. Jessop, N. Adams and P. Murray-Rust, ChemicalTagger: A tool for semantic text-mining in chemistry, *J. Cheminf.*, 2011, **3**(1), 17.
- 21 M. C. Swain and J. M. Cole, ChemDataExtractor: a toolkit for automated extraction of chemical information from the scientific literature, *J. Chem. Inf. Model.*, 2016, **56**(10), 1894–1904.
- 22 O. P. Pfeiffer, H. Liu, L. Montanelli, M. I. Latypov, F. G. Sen, V. Hegadekatte, E. A. Olivetti and E. R. Homer, Aluminum alloy compositions and properties extracted from a corpus of scientific manuscripts and US patents, *Sci. Data*, 2022, **9**(1), 128.
- 23 J. Mavracic, C. J. Court, T. Isazawa, S. R. Elliott and J. M. Cole, ChemDataExtractor 2.0: Autopopulated ontologies for materials science, *J. Chem. Inf. Model.*, 2021, **61**(9), 4280–4289.
- 24 H. Huo, Z. Rong, O. Kononova, W. Sun, T. Botari, T. He, V. Tshitoyan and G. Ceder, Semi-supervised machine-learning classification of materials synthesis procedures, *npj Comput. Mater.*, 2019, **5**(1), 62.
- 25 R. Yan, X. Jiang, W. Wang, D. Dang and Y. Su, Materials information extraction via automatically generated corpus, *Sci. Data*, 2022, **9**(1), 401.
- 26 T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado and J. Dean, Distributed representations of words and phrases and their compositionality, in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., New York, US, 2013, vol. 26, https://proceedings.neurips.cc/paper_files/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf.
- 27 Q. Le and T. Mikolov, Distributed representations of sentences and documents, in *International Conference on Machine Learning*, PMLR, 2014, pp. 1188–1196.
- 28 J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (long and Short Papers)*, 2019, pp. 4171–4186.
- 29 V. Tshitoyan, J. Dagdelen, L. Weston, A. Dunn, Z. Rong, O. Kononova, K. A. Persson, G. Ceder and A. Jain, Unsupervised word embeddings capture latent knowledge from materials science literature, *Nature*, 2019, **571**(7763), 95–98.
- 30 T. Gupta, M. Zaki and N. A. Krishnan, Mausam: MatSciBERT: A materials domain language model for text mining and information extraction, *npj Comput. Mater.*, 2022, **8**(1), 102.
- 31 M. Yoshitake, F. Sato, H. Kawano and H. Teraoka, MaterialBERT for natural language processing of materials science texts, *Sci. Technol. Adv. Mater.: Methods*, 2022, **2**(1), 372–380.
- 32 P. Shetty, A. C. Rajan, C. Kuenneth, S. Gupta, L. P. Panchumarti, L. Holm, C. Zhang and R. Ramprasad, A general-purpose material property data extraction pipeline from large polymer corpora using natural language processing, *npj Comput. Mater.*, 2023, **9**(1), 52.
- 33 Z. Zheng, O. Zhang, C. Borgs, J. T. Chayes and O. M. Yaghi, ChatGPT chemistry assistant for text mining and the

- prediction of mof synthesis, *J. Am. Chem. Soc.*, 2023, **145**(32), 18048–18062.
- 34 S. Gupta, A. Mahmood, P. Shetty, A. Adeboye and R. Ramprasad, Data extraction from polymer literature using large language models, *Commun. Mater.*, 2024, **5**(1), 269.
- 35 V. T. Silva, A. Rademaker, K. Lioni, R. Giro, G. Lima, S. Fiorini, M. Archanjo, B. W. Carvalho, R. Neumann, A. Souza and *et al.*, Automated, LLM enabled extraction of synthesis details for reticular materials from scientific literature, *arXiv*, 2024, preprint, arXiv:2411.03484, DOI: [10.48550/arXiv.2411.03484](https://doi.org/10.48550/arXiv.2411.03484).
- 36 M. Thway, A. K. Low, S. Khetan, H. Dai, J. Recatala-Gomez, A. P. Chen and K. Hippalgaonkar, Harnessing GPT-3.5 for text parsing in solid-state synthesis—case study of ternary chalcogenides, *Digital Discovery*, 2024, **3**(2), 328–336.
- 37 J. Dagdelen, A. Dunn, S. Lee, N. Walker, A. S. Rosen, G. Ceder, K. A. Persson and A. Jain, Structured information extraction from scientific text with large language models, *Nat. Commun.*, 2024, **15**(1), 1418.
- 38 M. P. Polak and D. Morgan, Extracting accurate materials data from research papers with conversational language models and prompt engineering, *Nat. Commun.*, 2024, **15**(1), 1569.
- 39 X. Wang, L. Huang, S. Xu and K. Lu, How does a generative large language model perform on domain-specific information extraction? – a comparison between GPT-4 and a rule-based method on band gap extraction, *J. Chem. Inf. Model.*, 2024, **64**(20), 7895–7904.
- 40 H. M. Sayeed, C. Clark, T. Mohanty and T. D. Sparks, KnowMat: An agentic approach to transforming unstructured materials science literature into structured data, *Integr. Mater. Manuf. Innov.*, 2026, **15**, 465–474.
- 41 S. Sadrudin, E. Poupaki, J. D'Souza, S. Auer, A. Watkins, B. Karasulu, A. J. Mackus and E. Kessels, Designing an agentic ai workflow for structured information extraction from scientific text: A case study on atomic layer deposition of ZnO and IGZO, *J. Vac. Sci. Technol. A*, 2026, **44**(3), 032413.
- 42 C. Chong and J. Colindres: LitXBench: A benchmark for extracting experiments from scientific literature. *arXiv*, 2026, preprint, arXiv:2604.07649, DOI: [10.48550/arXiv.2604.07649](https://doi.org/10.48550/arXiv.2604.07649).
- 43 OpenAI: OpenAI o3-mini. Model snapshot o3-mini-2025-01-31; 200,000-token context window. Accessed: 2026-07-23 (2025). <https://openai.com/index/openai-o3-mini/>.
- 44 A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser and I. Polosukhin, Attention is all you need, In *Proceedings of the 31st International Conference on Neural Information Processing Systems. NIPS'17*, Curran Associates Inc., Red Hook, NY, USA, 2017, pp. 6000–6010.
- 45 C. Wang, X. Lu, S.-K. Ng, B. K. H. Low, Trace: Transformer-based attribution using contrastive embeddings in llms. *arXiv*, 2024, preprint, arXiv:2407.04981, DOI: [10.48550/arXiv.2407.04981](https://doi.org/10.48550/arXiv.2407.04981).
- 46 T. Caselli, P. Vossen, M. Erp, A. Fokkens, F. Ilievski, R. Izquierdo, M. Le, R. Morante and M. Postma, When it's all piling up: investigating error propagation in an NLP pipeline. in *WNACP@ NLDB*, 2015.
- 47 H. Song, M. Kim, D. Park, Y. Shin and J.-G. Lee, Learning from noisy labels with deep neural networks: A survey, *IEEE Transact. Neural Networks Learn. Syst.*, 2022, **34**(11), 8135–8153.
- 48 C. Ding, T. Pereira, R. Xiao, R. J. Lee and X. Hu, Impact of label noise on the learning based models for a binary classification of physiological signal, *Sensors*, 2022, **22**(19), 7166.
- 49 A. Raj, X. Wang, M. Luebbe, H. Wen, K. Lu and S. Xu, A dataset of precipitate-containing multi-principal element alloys, *Data Brief*, 2026, **65**, 112540, DOI: [10.1016/j.dib.2026.112540](https://doi.org/10.1016/j.dib.2026.112540).
- 50 X. Wang, A. Raj, M. Luebbe, H. Wen, S. Xu and K. Lu, *Multi-Category Materials Information Extraction, Composition, Processing, Microstructure, Properties*, Zenodo, 2025, DOI: [10.5281/zenodo.18021833](https://doi.org/10.5281/zenodo.18021833), <https://zenodo.org/records/18021833>.