



LLM-driven materials knowledge extraction: multimodal parsing, ontology, and agentic systems

Shuai Yang^{1,2,*} , Yimeng Wang¹, Qiong Tu¹, Nian Ran^{3,*}, Ying Lu^{1,2,*}

Keywords:

Large language models, materials informatics, knowledge extraction, multimodal document parsing, ontology engineering, knowledge graphs, AI agents, retrieval-augmented generation

Citation:

Yang, S.; Wang, Y.; Tu, Q.; Ran, N.; Lu, Y. LLM-driven materials knowledge extraction: multimodal parsing, ontology, and agentic systems. *J. Mater. Inf.* 2026, 6, 39.

<https://dx.doi.org/10.20517/jmi.2026.38>

Received: 8 Jun 2026

First Decision: 6 Jul 2026

Revised: 11 Jul 2026

Accepted: 28 Jul 2026

Published: 31 Jul 2026

Academic Editor:

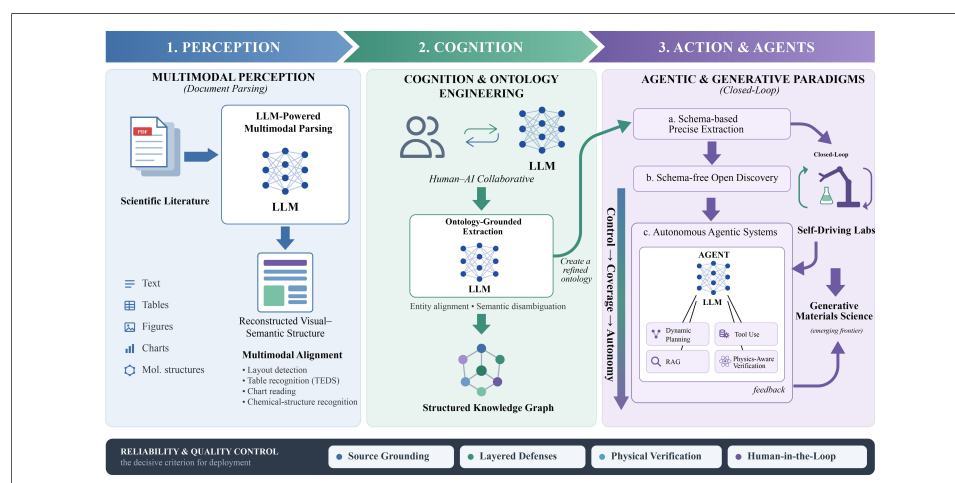
Hao Li

Copy Editor:

Pei-Yun Wang

Production Editor:

Pei-Yun Wang



Abstract

Extracting reliable knowledge from unstructured materials literature remains a central bottleneck for data-driven and AI-enabled materials discovery. Large language models (LLMs) are reshaping this task by integrating multimodal document parsing, ontology-guided semantic grounding, structured extraction, and agentic verification into increasingly unified workflows. This review analyzes these developments through a Perception–Cognition–Action lens. At the perception layer, we examine how scientific document parsers, multimodal LLMs, table and chart readers, and optical chemical-structure-recognition systems convert visually rich papers into computable evidence. At the cognition layer, we discuss how ontologies and knowledge graphs constrain LLM outputs, support entity alignment, and reduce semantic ambiguity. At the action layer, we compare schema-based extraction, schema-free discovery, and agentic extraction as a control–coverage–autonomy spectrum rather than a simple succession of tools. We further argue that reliability is the decisive criterion for large-scale deployment, and synthesize failure modes, layered defenses, and evaluation protocols that connect source grounding,

¹National Science Library (Chengdu), Chinese Academy of Sciences, Chengdu 610299, Sichuan, China.

²Department of Information Resources Management, School of Economics and Management, University of Chinese Academy of Sciences, Beijing 100049, China.

³State Key Laboratory of High Performance Ceramics, Shanghai Institute of Ceramics, Chinese Academy of Sciences, Shanghai 200050, China.

*Correspondence to: Prof. Shuai Yang, Prof. Ying Lu, National Science Library (Chengdu), Chinese Academy of Sciences, Chengdu 610299, Sichuan, China. E-mail: yangs@clas.ac.cn; luy@clas.ac.cn; Prof. Nian Ran, State Key Laboratory of High Performance Ceramics, Shanghai Institute of Ceramics, Chinese Academy of Sciences, Shanghai 200050, China. E-mail: rannian@mails.sic.ac.cn

ontology constraints, physical verification, and human-in-the-loop review. By distinguishing demonstrated extraction capabilities from more speculative AI-scientist and self-driving-laboratory visions, this review provides a comparative and risk-aware account of how LLM-driven systems can produce evidence-linked, physically meaningful, and reusable materials knowledge.

INTRODUCTION

Materials science has entered a data-intensive era, yet the availability of high-quality, comprehensive materials data remains a fundamental bottleneck. Structured resources such as the Materials Project and OQMD provide valuable crystal-structure and thermodynamic data, but synthesis pathways, processing conditions, microstructures, characterization results, and conditional property relationships are still largely embedded in unstructured articles, patents, reports, tables, figures, and supplementary files^[1,2]. Transforming this heterogeneous literature into reliable, computable knowledge is therefore a prerequisite for scalable materials informatics.

Earlier extraction systems made important progress through rule-based chemistry toolkits, sequence-labeling models, and BERT-style discriminative pipelines. Chemistry-aware systems such as OSCAR4, ChemicalTagger, ChemDataExtractor, and ChemDataExtractor 2.0 introduced domain-specific tokenization, chemical named-entity recognition, grammar-based parsing, and materials-oriented property extraction^[3-6]. These systems remain valuable because they are inspectable and can be highly accurate for narrow tasks, yet they are brittle when documents vary in layout, terminology, and reporting style. Dataset-driven deep learning improved generalization for named-entity recognition, relation extraction, layout analysis, and table recognition, but it still often relied on fragmented modules whose errors propagated downstream. More importantly, both rule-based and discriminative pipelines have difficulty preserving non-local processing-structure-property (PSP) linkages when evidence is distributed across text, tables, figures, captions, and supplementary information^[2,7].

A bibliometric analysis of materials information extraction research from 2002 to 2026 [Figure 1] shows both rapid growth and a post-ChatGPT methodological shift toward generative AI. The corpus was retrieved on 26 January 2026 using a reproducible Web of Science topic-search strategy (Web of Science Core Collection, Clarivate; <https://www.webofscience.com>) that combines materials-related terms with information-extraction and AI-methodology terms; the complete query and analysis scripts are provided in the project repository. Records assigned to 2026 represent an incomplete publication year and are shown for transparency; complete-year trend comparisons are therefore restricted to 2002-2025. Papers from 2021-2025 account for 73.6% of the analyzed corpus, and the annual count increased from 70 in 2022 to 227 in 2025. The methodological composition also changed sharply: large language model (LLM)-based extraction rose from a marginal presence in 2021 to nearly half of the methods by 2025. Recent LLM-based systems report F1 values of roughly 0.82-0.97 across tasks ranging from host-dopant relation extraction to materials table mining, often with only hundreds, or even tens, of annotated examples^[8-10].

This shift should not be interpreted as a simple replacement of earlier tools by larger models. LLMs introduce a different interface to extraction: they can transform unstructured passages into structured JSON records, reason over nested conditions, ask follow-up questions, and call external tools^[11]. At the same time, they introduce new risks that are less visible in deterministic pipelines. A model may produce a syntactically valid record that is unsupported by the paper, attach a correct value to the wrong sample, normalize a chemical formula in a chemically plausible but false way, or fabricate identifiers such as CAS numbers and DOIs. Thus, the central question is no longer only whether LLMs can extract more information, but whether

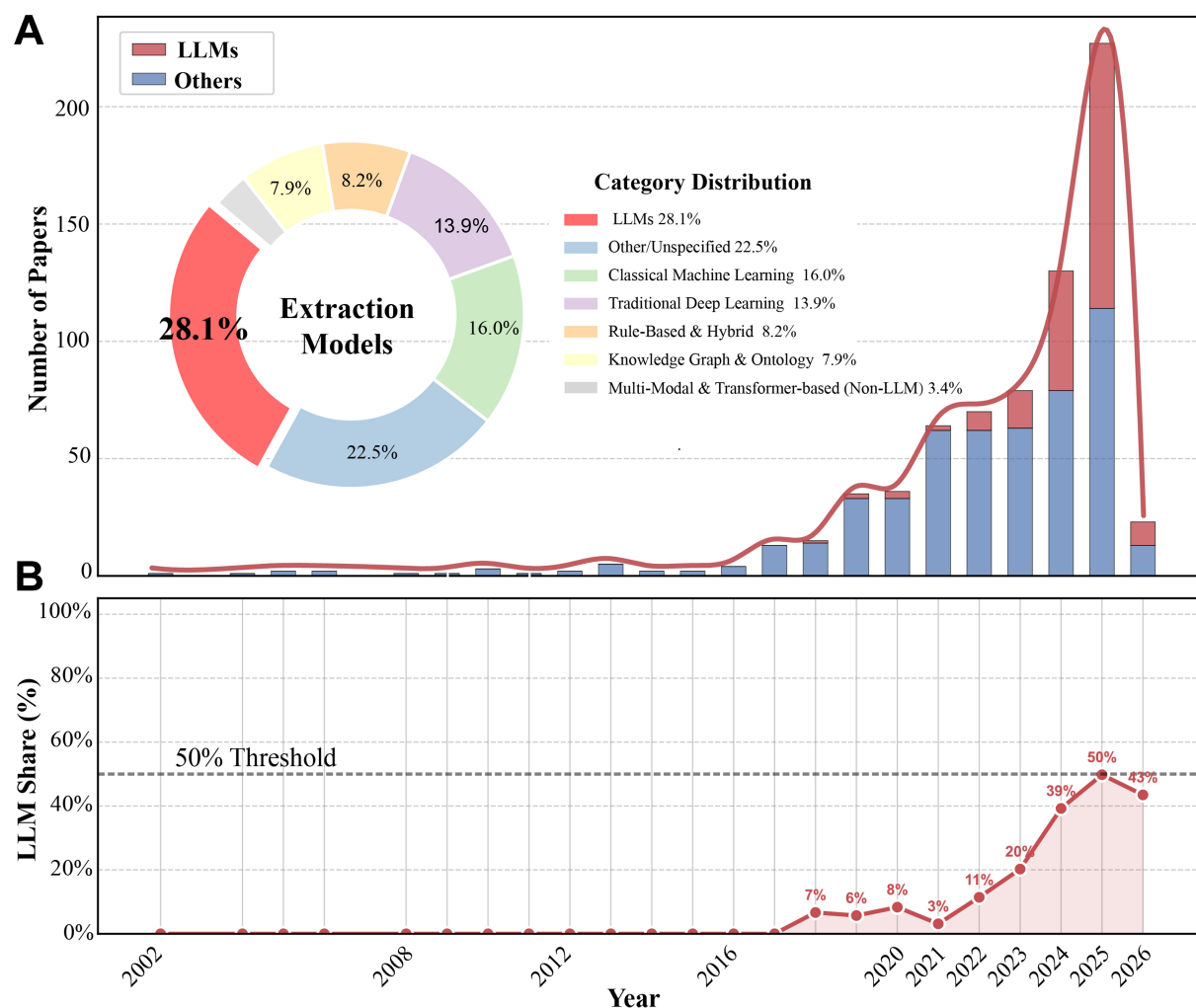


Figure 1. Bibliometric analysis of materials information extraction (2002–2026): the paradigm shift toward generative AI. (A) Annual publication counts by assigned publication year (2002–2026). Papers published in 2021–2025 account for 73.6% of the total. Post-ChatGPT, annual publications surged threefold (from 70 in 2022 to 227 in 2025). The 2026 bar represents an incomplete publication year at the retrieval cutoff (26 January 2026) and is excluded from complete-year trend comparisons. The red curve is a cubic-spline smoothing of total annual publication counts and is shown only as a visual guide, not as a statistical model. Inset: pie chart showing the proportion of the entire retrieved corpus (assigned publication years 2002–2026) that utilizes LLMs; (B) Evolution of extraction methodologies. The share of LLM-based methods rose sharply from 3% in 2021 to nearly 50% in 2025, indicating a rapid transition to the generative AI paradigm. Search queries, data, and analysis scripts are available at https://github.com/sShuaiYang/Al4Mat_KnowExtracti on. AI: Artificial intelligence; LLMs: large language models.

they can extract information that remains faithful to the source, semantically well grounded, and physically meaningful^[12,13].

This review uses the Perception–Cognition–Action (PCA) structure as an analytical lens rather than as a rigid processing pipeline [Figure 2]. Perception covers multimodal scientific document parsing; cognition covers ontology and knowledge-graph grounding; and action covers structured extraction, open discovery, and agentic workflows. The contribution of this lens is not the terminology itself, but the three analyses layered onto it. First, we treat reliability as a first-class dimension by linking each stage to characteristic failure modes and verification mechanisms. Second, we introduce an operational taxonomy [Table 1] that compares systems across modality, supervision, output representation, grounding, verification, autonomy, and evaluation protocol^[8–10,14–21]. Third, we interpret the field’s trajectory as a transfer of control from expert-designed schemas toward bounded autonomy, while emphasizing that domain knowledge must re-enter the workflow through ontologies, retrieval, tools, physical constraints, and human review.

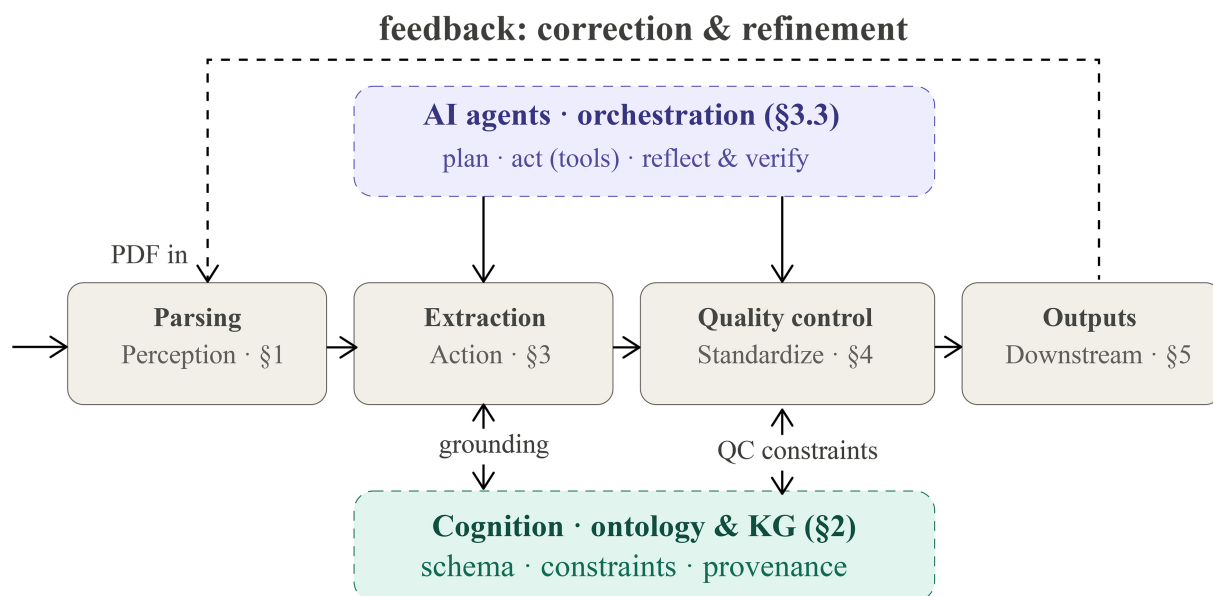


Figure 2. Materials knowledge extraction as a PCA framework. Perception parses raw PDF literature; cognition supplies ontology and knowledge-graph grounding; action converts evidence into structured records through schema-based, schema-free, or agentic extraction. Quality control connects all stages through source provenance, ontology constraints, physical checks, and feedback loops. PCA: Perception–Cognition–Action; AI: artificial intelligence; QC: quality control; KG: knowledge graph.

Table 1. An operational taxonomy of representative LLM-based materials extraction systems across seven design dimensions

System (Ref.)	Input modality	Supervision	Output representation	Grounding	Verification	Autonomy	Evaluation
ChatExtract ^[10]	Text	ZS/FS prompting	Triples (material/value/unit)	Parametric	Self-consistency, multi-turn follow-up	Single-pass (iterative)	Exact-match P/R/F1
LLM-NERRE ^[8]	Text	FT (-hundreds)	Fixed schema (JSON)	Parametric	HITL annotation	Single-pass	Exact + manual P/R/F1
MaTableGPT ^[9]	Tables	FS/FT	Fixed schema (JSON)	Parametric	Follow-up hallucination filtering	Single-pass	Structure + value F1
MatSKRAFT ^[14]	Tables	Trained (graph, distant)	Typed KG	Scientific constraints	Constraint checks	Pipeline	F1 + throughput
Schema-Miner ^[15]	Text	Schema induction + expert	Induced → ontology-aligned schema	Reference ontology	HITL + ontology alignment	Multi-agent	Coverage vs. handcrafted
nanoMINER ^[16]	Multimodal (text + figures)	Prompting (ReAct)	Fixed schema	Tool/visual (detector)	Cross-modal alignment	Multi-agent	P/R vs. baselines
DIVE ^[17]	Figures (multimodal)	Prompting	Structured DB	Domain/visual	Multi-agent cross-check	Multi-agent	Accuracy/coverage gains
ChemCrow ^[18]	Text	Prompting	Action/answer	External tools/DBs	Tool execution	Agentic (tool use)	Expert assessment
LLaMP ^[19]	Text	Prompting + RAG	Structured answers	Retrieval (Materials Project)	Source grounding	Agentic	Fidelity/hallucination
Eunomia ^[20]	Text	Prompting	Extracted facts	Source documents	Multi-step evidence validation	Agentic	Validity vs. evidence
Wikontic ^[21]	Text	Prompting + dense retrieval	Ontology-aligned KG	Ontology + Wikidata	Stoichiometry-aware dedup	Pipeline (agentic alignment)	Alignment accuracy

LLM: Large language model; ZS: zero-shot; FS: few-shot; P/R/F1: precision/recall/F1; FT: fine-tuned; HITL: human-in-the-loop; KG: knowledge graph; DB: database; RAG: retrieval-augmented generation.

This review therefore treats LLM-based extraction primarily as a reliability problem, rather than a capability problem: how raw scientific evidence is parsed, semantically constrained, verified, and made reusable. Accordingly, we first examine multimodal parsing, then analyze LLM-enabled ontology engineering, compare schema-based, schema-free, and agentic extraction paradigms, and finally synthesize quality-control and evaluation protocols. Because the review spans ontology-grounded extractors, tool-augmented agents, and constrained knowledge-graph pipelines, Table 1 locates representative systems within a shared design space rather than treating them as mutually exclusive categories. A system may be simultaneously ontology-grounded, tool-augmented, and knowledge-graph-producing; what distinguishes it is the combination of input modality, output representation, grounding mechanism, verification burden, and autonomy level. This operational view supports a comparative account of when a given approach is reliable, scalable, and scientifically appropriate. Given the field's rapid development, we also include selected recent preprints that introduce technically relevant systems not yet represented in peer-reviewed literature; these works are discussed as emerging evidence rather than established consensus.

MULTIMODAL SCIENTIFIC DOCUMENT PARSING

As the Perception layer of materials-intelligence systems, multimodal scientific document parsing transforms visually rich literature into computable evidence [Figure 2]. This task goes far beyond optical character recognition (OCR). A scientific article is a visual-semantic object in which text blocks, section headings, tables, equations, plots, captions, chemical structures, and supplementary references jointly encode the experimental record. In materials science, this structure is especially consequential because a reported property value is rarely self-contained: it often depends on synthesis history, composition, atmosphere, temperature, time, instrument settings, calibration procedures, and sample labels introduced elsewhere in the paper. Failure to preserve these links during parsing can render a correct numerical value unusable or misleading^[22].

The objective of the perception layer is therefore not simply to digitize text, but to reduce the entropy of the document while preserving PSP linkages. For example, a battery paper may report calcination conditions in the experimental section, electrochemical cycling data in a table, and capacity retention in a multi-panel figure. A catalysis paper may encode catalyst composition in a table header, reaction conditions in a caption, and activity trends in a plot. A polymer or organic semiconductor paper may communicate the relevant molecular entity as a graphical structure rather than a string. These examples illustrate why materials parsing requires layout recovery, modality-specific extraction, cross-modal alignment, and provenance tracking as coupled subtasks rather than independent preprocessing steps^[7,22].

Evolution of technical paradigms

Early efforts in scientific document parsing were dominated by modular pipeline architectures, which decomposed the task into PDF text extraction, OCR, layout detection, table recognition, figure extraction, and rule-based or statistical entity extraction. In chemistry and materials science, OSCAR4 provided chemical tokenization and named-entity recognition, ChemicalTagger parsed experimental language into phrase-level semantics, and ChemDataExtractor coupled text and table parsing to populate chemical and materials records^[3-6]. These systems established the importance of domain-specific grammar and dictionaries, but their staged architecture made them vulnerable to error propagation. A missed caption, a split table cell, or an incorrectly recognized chemical formula can propagate into the final database without an explicit mechanism for later correction^[23].

Between 2019 and 2022, the field entered a dataset-driven deep-learning era for layout and table understanding [Figure 3A]. PubLayNet and DocBank enabled layout detectors to recognize regions such as text, titles, lists, figures, and tables at scale^[24,25], while SciTSR, PubTabNet, and PubTables-1M supported

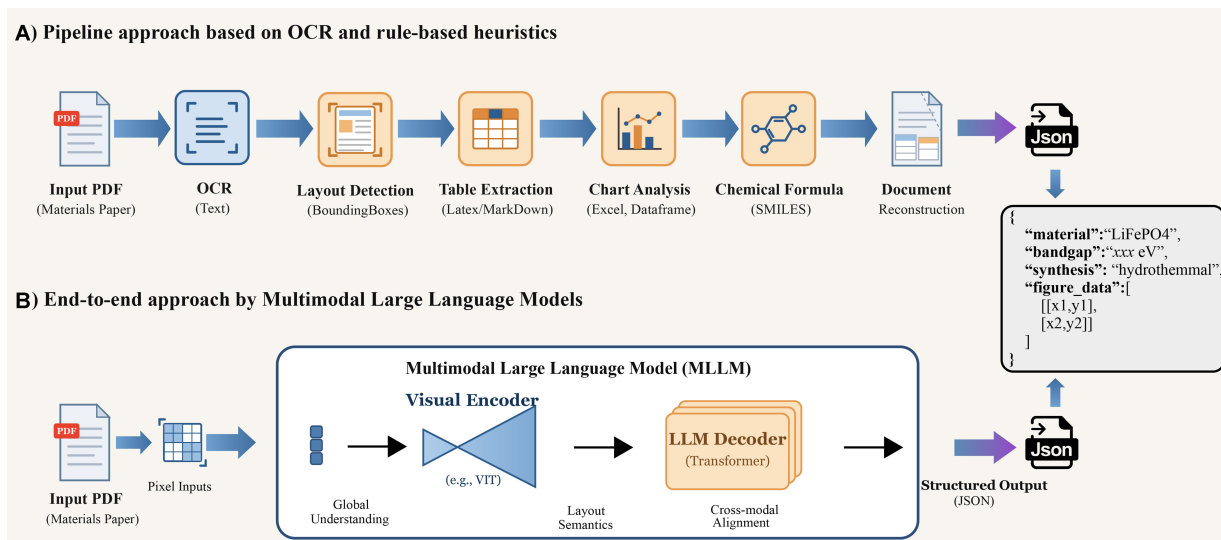


Figure 3. Two document-parsing paradigms for materials-science papers. (A) Modular OCR- and rule-based pipelines separate layout, table, chart, and chemical-structure extraction; (B) Unified end-to-end multimodal LLM parsing directly maps page images or parsed blocks to machine-readable structured output. The two routes differ in controllability, error propagation, cost, and traceability. OCR: Optical character recognition; LLM: large language model.

table-structure recognition using metrics such as tree-edit-distance similarity (TEDS) and cell-level F_1 [26–28]. These datasets improved modular evaluation and made the perception layer more reproducible, but they did not solve the scientific grounding problem. Recognizing a table cell is different from understanding that the value belongs to a specific alloy composition under a specific annealing schedule. Similarly, recognizing a plotted curve is different from binding it to the correct legend entry and sample code.

From 2022 onward, two converging research directions reshaped the technical landscape. The first involved document-specific multimodal Transformers, exemplified by LayoutLMv3, which unified text, layout, and visual features through joint pre-training objectives [29] [Figure 3B]. The second pursued OCR-free, end-to-end image-to-structure generation, as demonstrated by Donut and Pix2Struct, which directly map document images to structured representations [30,31]. Since 2024, document-focused vision-language models (VLMs) such as DocLayLLM and DocVLM have further targeted the ‘seeing but not reading closely’ limitation of general vision-language models by improving structural resolution and efficiency for text-dense pages [32–34]. These advances move the field from loosely connected modules toward end-to-end perception systems, but they also make faithfulness evaluation more important because generative parsers can produce plausible structure even when the source page is ambiguous.

Beyond representative academic models, these paradigms have rapidly translated into a diverse engineering ecosystem. Rather than enumerating incommensurable implementations spanning research models, commercial application programming interfaces (APIs), OCR engines, and format-conversion utilities, we adopt a functional taxonomy that groups tools by their architectural trade-offs. (1) Modular pipeline tools (e.g., MinerU, Marker, Unstructured.io) prioritize throughput and deployability, serving as the workhorse for large-scale ingestion, but inherit the error propagation and weak cross-modal integration of staged architectures; (2) General-purpose vision-language models (e.g., GPT, Qwen-VL, Gemini, InternVL families) offer strong zero-shot layout and table understanding across document types, at the cost of higher inference expense, limited fine-grained numerical fidelity, and weaker traceability; (3) Domain-specialized document VLMs (e.g., dedicated scientific OCR/parsing models) trade generality for higher structural resolution on text-dense pages. The optimal choice is dictated by the target balance among accuracy, cost,

controllability, and traceability rather than by raw capability alone. A categorized inventory is maintained in the project repository (https://github.com/sShuaiYang/AI4Mat_KnowExtraction). Across all categories, challenges of system stability, interpretability, result traceability, and standardized evaluation remain open^[22].

In materials informatics, parsing should be evaluated by how well the evidence structure is preserved, not merely visual reconstruction quality. A parser that perfectly reads a table but loses multi-row header semantics may fail for catalysis or battery data; a chart reader that extracts a curve accurately but ignores the caption may bind data to the wrong sample; and an optical chemical structure recognition (OCSR) model outputting a plausible SMILES string must still be verified against the source image and downstream chemical validity. These considerations motivate the modality-specific discussion in Section “Key subtasks and domain-specific practices in materials science” and the reliability framework in Section “Reliability, Quality Control, and Evaluation”.

Key subtasks and domain-specific practices in materials science

Multimodal parsing can be organized around five interdependent subtasks. Layout understanding provides the spatial and logical scaffold for identifying sections, captions, tables, figures, equations, and supplementary links, enabled by large-scale datasets such as PubLayNet^[24] and DocBank^[25]. Table parsing recovers nested headers, merged cells, units, and condition-dependent columns; materials-focused approaches such as MaTableGPT^[9] report F1 up to 96.8% on water-splitting catalysis tables, while graph-based frameworks like MatSKRAFT^[14] encode scientific constraints for high-throughput processing. Chart and curve parsing extracts quantitative trends from phase diagrams, thermal-analysis curves, and electrochemical impedance spectra^[35]; general tools such as DePlot^[36] convert plots to approximate tables, but materials-specific benchmarks like MatViX^[37] evaluate curve similarity and hierarchical alignment rather than plain text overlap. Recent advances include PlotExtract^[38], which recovers two-axis plot data with over 90% precision/recall and ~5% positional error via chain-of-thought prompting, and DIVE^[17], a multi-agent workflow that improves figure-based extraction accuracy and coverage by 10%–15% over commercial models and > 30% over open-source baselines in solid-state hydrogen-storage literature. OCSR converts molecular depictions into machine-readable SMILES or InChI; recent models such as MolParser^[39], GTR-CoT^[40], and MolSight^[41] incorporate graph-based visual reasoning for reliable end-to-end conversion^[42]. Cross-modal alignment binds these elements to each other and to surrounding text so that extracted values remain scientifically interpretable, as benchmarked by DocGenome^[43], MMDocIR^[44], and M4DocBench^[45].

Materials literature makes these subtasks unusually demanding because scientific meaning is distributed rather than localized. A conductivity value in a plot may depend on a synthesis route described in the Methods section and on a composition code defined only in a table or caption. Figure–caption alignment is therefore a critical bottleneck^[46,47]: captions define sample identities, axis semantics, units, and measurement conditions, while in-text references (e.g., “Figure 3B”) connect the visual element to the surrounding argument. Pipeline-based workflows such as that of Srór *et al.*^[48] and multi-agent readers such as DIVE^[17] explicitly exploit caption grounding to mitigate misattribution errors, where a correctly read value is assigned to the wrong material or condition. Integrated pipelines such as KnowMat^[49] further combine specialized parsers, cross-modal verification, and standardized ingestion workflows for end-to-end materials knowledge extraction.

Parser faithfulness should be evaluated with task-matched, modality-specific metrics rather than a single aggregate score: layout detection with mAP/IoU^[24,25]; table recognition with TEDS and cell-level F1^[26–28]; OCR-free conversion with edit-distance or character accuracy; and chart extraction with curve similarity or tolerance-based positional error. Reporting modality-specific results is essential - Srór *et al.* report F1 above 99% for text and tables but only 85.8% for figures on the same 9,600-entry catalytic-performance corpus^[48],

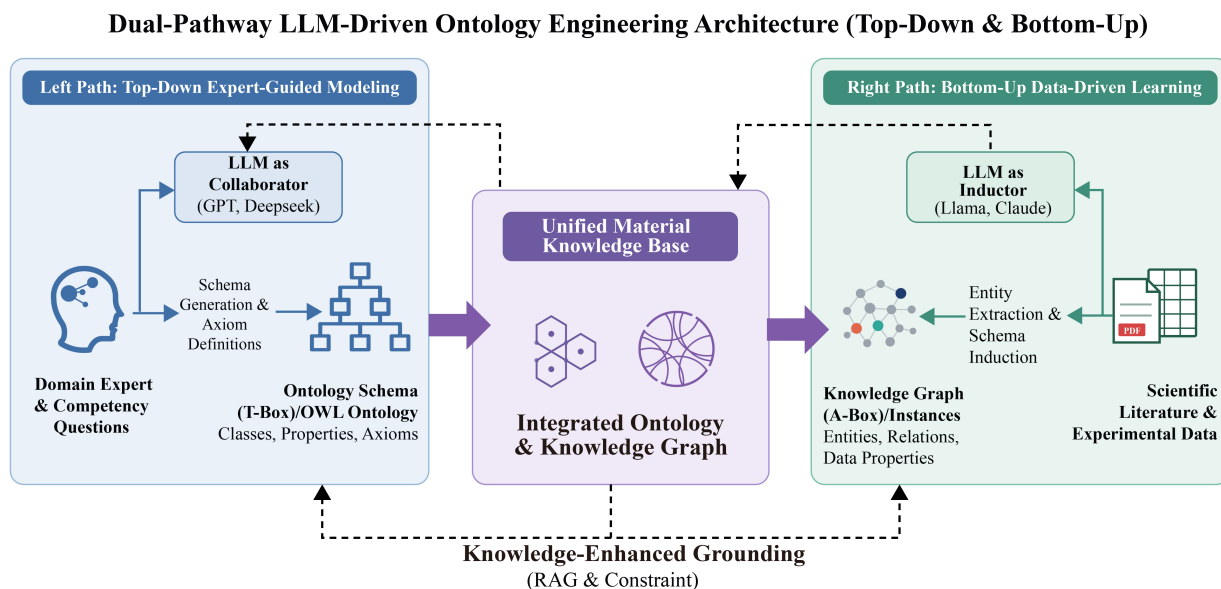


Figure 4. Dual-pathway LLM-driven ontology engineering. In the top-down route, experts use LLMs as drafting assistants under competency questions and reasoner validation. In the bottom-up route, LLMs induce candidate schemas and populate knowledge graphs from literature. The central reliability principle is mutual constraint: LLMs propose candidates, while ontologies, source evidence, formal validation, and expert review determine admission. LLM: Large language model; RAG: retrieval-augmented generation.

quantifying a persistent accuracy gap between visual and textual modalities. Because generative parsers are stochastic, these metrics should be paired with procedural safeguards: figure pre-extraction, source prioritization across modalities, repeatability checks (~87% run-to-run agreement^[48]), and DOI/page-level provenance for every parsed value. These parser-level checks feed into the downstream content-level quality-control mechanisms detailed in Section “Reliability, Quality Control, and Evaluation”.

LLM-DRIVEN ONTOLOGY ENGINEERING

Ontology engineering forms the Cognition layer of LLM-driven materials knowledge extraction. Parsed outputs are often fragmented and ambiguous: a value such as “500” may denote a calcination temperature, wavelength, pressure, dwell time, or lattice parameter depending on context. Ontologies and knowledge graphs restore meaning by defining classes, relations, constraints, units, and admissible connections among materials, processes, structures, properties, and measurement conditions. This semantic layer converts isolated extracted fields into typed, validated facts, enabling cross-paper integration, interoperability, and machine reasoning^[50–53].

The central principle is mutual constraint [Figure 4]. LLMs do not define ontological truth; they draft candidate concepts, relations, and axioms from literature. Ontologies do not merely store extracted facts; they constrain generation, retrieval, entity alignment, and validation. This distinction avoids a common overclaim in LLM-based ontology engineering. A model may propose that “precursor pulse time” is relevant to atomic layer deposition, but that candidate must be grounded in source text, aligned to an existing concept or explicitly added, checked for logical consistency, and accepted by domain experts before it becomes ontology-level knowledge^[54–59].

In practice, this coupling operates in two directions. In the LLM-to-ontology direction, models propose schema elements from domain corpora and emit draft artifacts in structured formats such as OWL/RDF or JSON-like schema templates. In the ontology-to-LLM direction, three mechanisms dominate: (1) schema-constrained generation, where class, property, and value-type restrictions reject invalid outputs^[51–53]; (2)

ontology-aware retrieval, where class definitions and admissible relations are retrieved as context, reframing extraction as alignment of textual mentions to ontology classes^[60]; and (3) formal validation, where reasoners test consistency and satisfiability, competency questions serve as acceptance tests, and SHACL-like constraints together with unit/type checks validate instance data^[54–56,60–62]. The knowledge graph is the operational bridge: bottom-up extraction populates entities, relations, and provenance, while top-down ontologies constrain which nodes and edges are admissible. Materials knowledge graphs have been constructed for framework materials, multidisciplinary materials science, process chains, and autonomously generated facts, increasingly serving as external memory for retrieval-augmented reasoning^[55,63–65]. However, an LLM-generated graph is not automatically reliable. Entity normalization, stoichiometry checks, synonym resolution, source-span grounding, and expert admission are required to prevent the graph from amplifying extraction errors downstream.

Top-down approach: LLM-empowered collaborative ontology modeling

The top-down route starts with explicit domain requirements, mature scientific abstractions, and competency questions. It is appropriate when a domain already has relatively stable concepts, such as crystallographic descriptors, thermodynamic properties, electrochemical measurement protocols, or standardized process-chain entities. LLMs can accelerate early ontology design by summarizing domain corpora, proposing candidate classes and relations, and translating informal requirements into draft ontology components. This role is best described as semantic drafting rather than autonomous ontology construction: experts begin from a structured candidate rather than a blank schema, while retaining authority over scope, granularity, and logical commitments^[54–56].

Validation is therefore not optional. A practical validation loop should include source-span grounding for each candidate class or relation, competency-question testing, reasoner/SHACL consistency checking, and expert acceptance before ontology admission. This caution is essential because LLMs may fabricate plausible concepts, relations, references, or even DOIs; if admitted, such artifacts can contaminate subsequent extraction as false constraints. Top-down LLM-assisted ontology engineering is strongest when model-generated drafts are coupled with explicit provenance, formal consistency testing, and human-in-the-loop acceptance^[54,56–58]. Materials-domain workflows exemplify this pattern: the Schema-Miner pipeline for atomic layer deposition couples LLM-discovered schemas with ontology grounding and expert refinement^[15,54–56], while retrieval-augmented pipelines build ontologic knowledge graphs from selected papers^[61] or mine decades of literature into ontology-structured triples for expert review^[62].

A further challenge is granularity. Materials concepts can be represented at multiple levels: “thermal treatment”, “calcination”, “annealing under argon”, and “500 °C for 2 h” may all be valid depending on the use case. LLMs can surface these levels from literature, but they cannot decide alone which granularity best serves database integration or downstream modeling. Competency questions provide a practical criterion: if the ontology must answer queries about synthesis temperature, atmosphere, and dwell time, those fields must be explicitly represented; if it only supports high-level process classification, they may remain annotations. Ontology design is therefore a task-specific scientific modeling problem rather than a purely linguistic extraction problem.

Bottom-up approach: from data induction to knowledge enhancement

The bottom-up route addresses the opposite problem: emerging subfields often lack complete schemas, use inconsistent terminology, and report new attributes before formalization. Here, LLMs induce candidate concepts, attributes, and relations directly from a curated literature sample. A representative sample does not require loading an entire corpus into a single context window; documents can be processed iteratively in chunks until schema saturation, when additional documents introduce few new concepts. Schema-Miner

and related workflows illustrate this process: LLMs identify domain attributes, after which human review and ontology grounding consolidate the induced schema^[15,59]. Schema induction is typically performed by prompting general-purpose LLMs, whereas subsequent large-scale named entity recognition and relation extraction are often delegated to smaller fine-tuned models for cost and throughput reasons^[8,10].

A practical workflow comprises five steps: (1) select and sample a representative corpus; (2) propose candidate fields, relations, and reporting patterns; (3) consolidate candidates through synonym merging, unit normalization, and mapping to reference ontologies; (4) populate records or graphs from larger corpora through schema-constrained extraction; and (5) apply provenance, consistency, and expert checks before using the knowledge graph for retrieval or reasoning^[60–62,66]. This sequence clarifies that bottom-up discovery expands coverage but does not remove the need for top-down validation.

Bottom-up extraction also creates knowledge graphs that enhance downstream LLM reasoning through retrieval-augmented generation (RAG), where relevant external evidence is retrieved at inference time and supplied as context^[67]. Ontology-grounded RAG (OG-RAG) goes further by retrieving not only passages but also class definitions, admissible relations, and graph neighborhoods^[60]. This is especially useful in materials science because many questions require both factual evidence and typed scientific meaning: a reported value may be a bandgap rather than an absorption peak, or a temperature may belong to a processing event rather than a measurement condition. Knowledge graphs supply this structure while preserving links to source documents^[61–64]. The same hybrid “knowledge graph (KG)-from-text” paradigm extends beyond materials science; in pharmaceutical engineering, Taskiran *et al.* combine ontology-based information extraction with generative LLMs to construct a literature-derived knowledge graph and expose it through a querying chatbot, using symbolic grounding to improve explainability and reduce hallucination^[68].

Bottom-up induction therefore complements rather than replaces top-down ontology modeling. Induced schemas may contain redundancy, semantic drift, or hallucinated elements, so alignment with reference ontologies, formal consistency checks, source-grounded provenance, and expert review remain mandatory. The most robust workflows are hybrid: bottom-up discovery expands coverage, while top-down ontologies determine admissibility, interoperability, and downstream utility. The field still lacks standardized task definitions, ontology-construction benchmarks, and evaluation protocols, limiting reproducibility^[58]. Closing these gaps is essential for advancing LLM-driven ontology engineering from exploratory schema discovery toward robust, large-scale deployment.

EVOLUTION OF MATERIALS KNOWLEDGE EXTRACTION PARADIGMS

Knowledge extraction is the Action layer of the framework: parsed and semantically grounded evidence is transformed into structured records, summaries, knowledge graphs, or evidence chains that can support downstream reasoning and decision-making. The literature now spans three broad paradigms [Figure 5]. Schema-based extraction uses predefined templates or ontologies to maximize precision. Schema-free extraction induces attributes and relations from the corpus to improve coverage in emerging areas. Agentic extraction introduces planning, retrieval, tool use, and reflection to handle long documents, multimodal evidence, and multi-step validation.

These paradigms should not be read as linear replacements. Representative evidence supporting the comparison is drawn from schema-based, schema-free, and agentic systems^[8–10,15,16,20,59,70,71]. Table 2 summarizes the resulting control-coverage-autonomy spectrum. Schema-based extraction maximizes precision and interpretability when target fields are known. Schema-free extraction broadens coverage when the relevant attributes are not yet fully specified. Agentic extraction increases autonomy and evidence-gathering capacity when records require non-local, cross-modal, or tool-mediated reasoning. In practice,

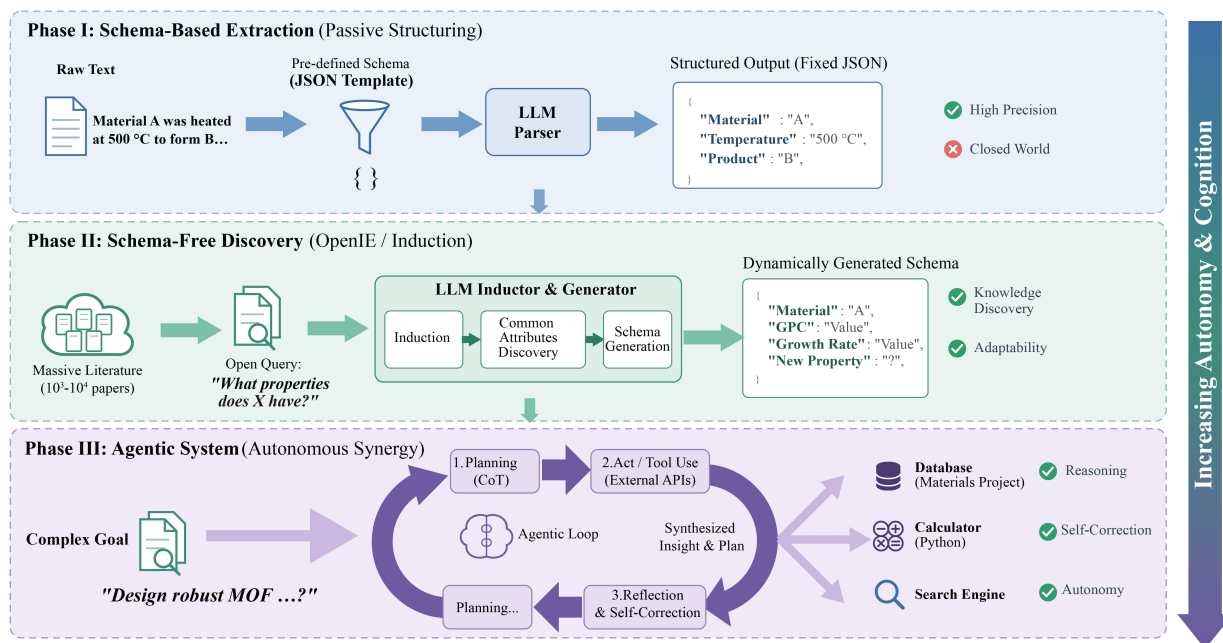


Figure 5. Evolution of LLM-driven knowledge extraction paradigms in materials science. The input “literature corpus” denotes corpus-scale processing (10³-10⁴ papers, yielding $\sim 1.6 \times 10^5$ nodes from a decade of literature^[62]). Across all paradigms, the LLM operates in either (i) prompted, in-context inference without weight updates; or (ii) parameter fine-tuning or domain-adaptive pretraining^[8,69]. The progression from schema-based to schema-free to agentic systems reflects increasing autonomy and decreasing direct human control; in practice, these paradigms are hybridized through schemas, retrieval, tools, and verification loops. LLM: Large language model; MOF: metal-organic framework; APIs: application programming interfaces.

Table 2. Comparative summary of the three materials-extraction paradigms

Dimension	Schema-based	Schema-free	Agentic
Human control/constraint	High (fixed schema)	Medium (induced schema)	Low-variable (goal-driven)
Coverage	Known unknowns	Unknown unknowns	Cross-document/cross-modal evidence
Setup and supervision cost	Low-moderate (schema + few examples)	Moderate (schema induction + curation)	High (orchestration, tools)
Output	Fixed structured records	Open/induced structures, summaries	Verified records + evidence chains
Reasoning	Single-pass mapping	Single-pass induction	Iterative plan-act-reflect
Dominant failure modes	Schema blind spots; context-window pressure	Semantic drift; redundancy	Compounding errors; cost/latency
Verification	Schema constraints	Post hoc structuring	Built-in tool/loop verification
Best-fit scenario	Known fields; high precision at scale	Exploratory/emerging domains	Long, multimodal, non-local evidence
Representative reported performance	Strong with few examples: field-level F1 ≈ 0.82 for host-dopant vs. ≈ 0.50 (seq2rel) / 0.39 (MatBERT) ^[8] ; MaTableGPT up to 96.8% F1, 47,670 data points from 2,406 tables, > 95% with -10 examples ^[9] ; ChatExtract 90.8% P / 87.7% R (bulk modulus), prompt-only ^[10]	Broader coverage and closer alignment with reporting practice than handcrafted schemas (Schema-Miner; inter-model BERTScore ≈ 0.80 -0.81) ^[15,59] ; panoramic summarization preserves sample-procedure-result context ^[70,71]	Higher reliability via verification loops and division of labor: nanoMINER ≈ 0.98 precision on multimodal extraction (obtained on structured nanozyme kinetic parameters; generalization to other domains is not yet established) ^[16] ; Eunomia uses multi-step evidence aggregation and validation ^[20]

Representative performance values were compiled from Refs. ^[8-10,15,16,20,59,70,71].

mature systems increasingly combine the paradigms: induced schemas constrain open discovery, agents call schema-based extractors as tools, and verification loops re-impose control after exploratory extraction. The scientific question is therefore not which paradigm is ‘best’, but how much autonomy is warranted by the evidence structure and how much verification is available.

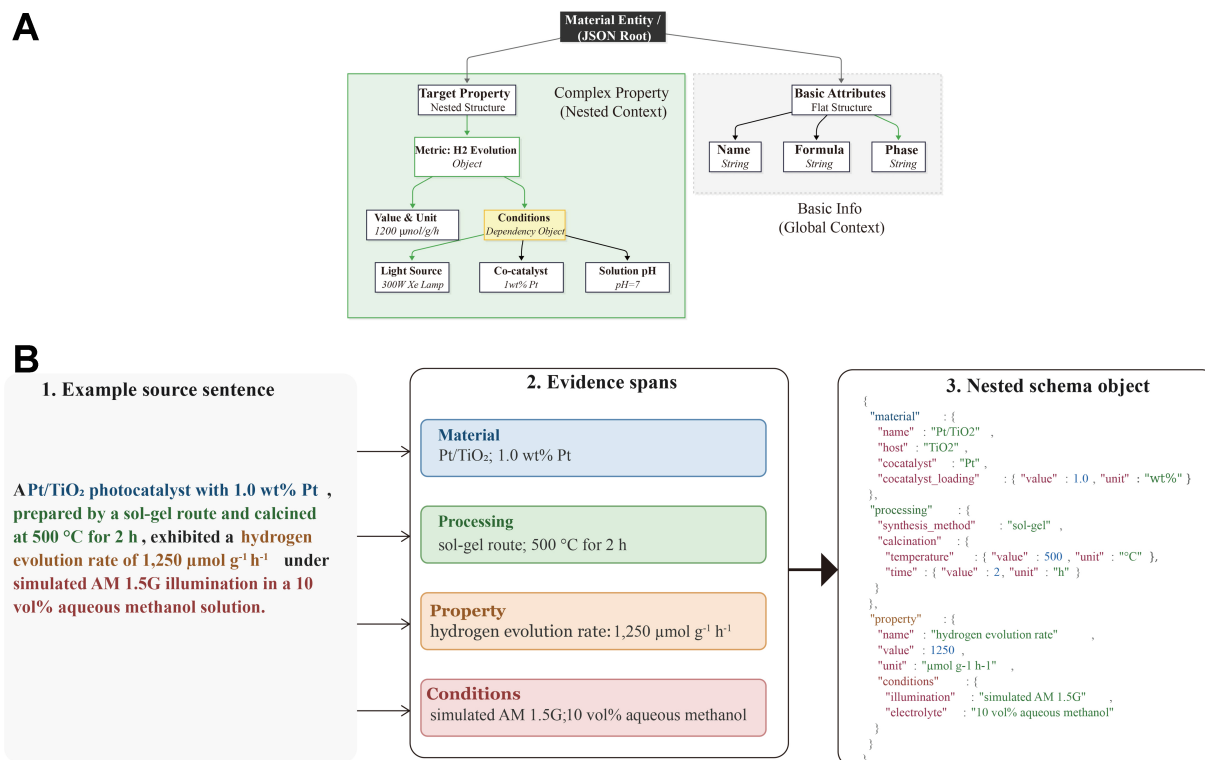


Figure 6. Example of schema-based materials data extraction. (A) Hierarchical schema structure for materials data extraction; (B) A source sentence is mapped into a nested schema that preserves material identity, synthesis and calcination conditions, property value, unit, and measurement conditions. This representation shows why schema-based extraction is useful for conditional materials properties whose meaning depends on experimental context.

Schema-based generative extraction

Schema-based generative extraction is the most controlled LLM extraction paradigm. Domain experts define a target schema, usually as JSON Schema, ontology-aligned fields, or a domain template, and the model maps source evidence into that structure. This strategy is attractive because it transforms open-ended language generation into constrained record population. The model is still generative, but the schema specifies admissible fields, nesting, value types, and sometimes enumerated categories. As shown conceptually in Figure 6, a source sentence can be converted into a nested object that binds material identity, processing conditions, property values, units, and measurement conditions in a single record.

Its main advantage is that hierarchical records can preserve conditional material facts that are difficult to capture with flat entity-relation labels. A catalytic activity, hydrogen-evolution rate, ionic conductivity, or battery capacity is meaningful only together with conditions such as illumination, electrolyte, temperature, current density, cycling rate, atmosphere, or sample history. Traditional sequence labeling can identify entities and relations, but it often fragments this context. Schema-based generation instead represents a conditional property as a nested object, which supports database construction and downstream modeling because values are stored together with the conditions under which they were measured^[8-10].

Reported performance shows why this paradigm has become a foundation for materials data production. In the structured scientific-text extraction study by Dagdelen *et al.*, LLM-NERRE achieved host-dopant F1 of approximately 0.82, compared with roughly 0.50 for seq2rel and 0.39 for a MatBERT-proximity baseline (these figures are drawn from different studies with different corpora, annotation schemes, and matching rules, so they should be read as indicative rather than as a controlled head-to-head comparison), while

Table 3. Summary of representative materials science datasets extracted via LLMs published in 2025–2026

Index	Dataset name	Material type	LLM	DOI
1	AlloyData-2024LT	Low-temperature alloys	GPT-3.5, GLM-4	10.1038/s41597-025-05512-9 ^[72]
2	NEMAD	Magnetic materials	GPT-4o, Gemini 2.0	10.1038/s41467-025-64458-z ^[73]
3	L2M3	Metal-organic framework materials	GPT-3.5-turbo, GPT-4	10.1021/jacs.4c11085 ^[74]
4	LLMs-MatLitRecon	Lithium-ion batteries, hydrogen evolution reactions, etc.	LLaMA3, Gemma, GPT3.5, Phi3	10.1038/s43246-025-01043-3 ^[66]
5	PV600	Perovskites	Mixtral, LLaMA, GPT-4o	10.1038/s41597-025-05637-x ^[75]
6	AuNP-seedmed-recipes	Nanomaterials	Llama-2	10.1039/D4DD00158C ^[76]
7	CATDA	Polymer catalyst materials	GPT-4 Turbo	10.1021/acscatal.5c06431 ^[77]
8	SnTe	Thermoelectric materials	Deepseek-R1	10.1021/acsami.5c16158 ^[78]
9	MatSyn25	Two-dimensional (2D) materials	DeepSeek-R1, GLM-4, Qwen3	10.1021/acsami.5c19849 ^[79]
10	Polymer_Process	Polymers	Llama-2	10.1016/j.polymer.2025.128875 ^[80]
11	Dental_Ceramics	Ceramics	Llama, Qwen, DeepSeek	10.1016/j.identj.2025.109358 ^[81]
12	Text-mined-synthesis	Solid-state synthesis	gpt-4o, Qwen 2.5	10.1038/s41597-025-06222-y ^[82]
13	L2D	Environmental materials	ChatGPT	10.1021/acsestwater.5c00551 ^[83]
14	ChatSSB	Solid-state batteries	GPT-4o, DeepSeek-R1, RAG	10.1021/acsmaterialslett.4c02674 ^[84]
15	OPV	Organic photovoltaic materials	GPT-4	10.1021/acsomega.5c03612 ^[85]
16	HEA	High-entropy alloy catalysts	InternLM 2.5	10.1002/aenm.202500744 ^[86]
17	Stainless_steel	Irradiated materials	GPT, DeepSeek	10.1016/j.jnucmat.2025.156241 ^[87]
18	ALMAP	Aluminum alloys	Llama 3.1	10.1016/j.mtcomm.2025.112843 ^[88]
19	Transition_metal_sulfides	Transition metal sulfides	GPT-3.5, GPT-4.0	10.3390/ma18163793 ^[89]
20	LMExt	Mineral thermodynamic datasets	Gemini-2.5-Flash, GPT-4o	10.1002/aisy.202500857 ^[90]

LLMs: Large language models.

producing parseable structured outputs in about 99% of cases and reaching high manual-scoring precision^[8]. MaTableGPT reached total F1 up to 96.8% for water-splitting catalyst tables and maintained more than 95% F1 with only ten labeled examples, enabling tens of thousands of catalytic-performance records to be extracted from thousands of tables^[9]. ChatExtract similarly demonstrated that prompt-only conversational extraction can reach approximately 90% precision/recall for selected materials properties, especially when multi-turn follow-up is used to reduce missing fields and hallucinated values^[10]. These results justify the use of schema-based LLMs for high-throughput, fixed-target data production.

The recent 2025–2026 wave of LLM-extracted materials datasets further illustrates the breadth of this paradigm. Reported applications span low-temperature alloys, magnetic materials, metal-organic frameworks, perovskites, nanomaterials, polymer catalysts, thermoelectrics, solid-state synthesis, solid-state batteries, organic photovoltaics, high-entropy-alloy catalysts, aluminum alloys, transition-metal sulfides, and mineral thermodynamics^[10,66,72–91]. Table 3 summarizes representative examples. These examples are valuable not only because they produce datasets, but because they expose how extraction tasks differ across materials domains. Metal-organic framework (MOF) extraction often requires topology, linker, metal-node, and adsorption-property fields; battery extraction requires electrode composition, electrolyte, cycling protocol, and capacity-retention context; catalysis extraction requires reaction conditions and normalization basis; synthesis extraction requires multi-step procedural information. A single generic schema is therefore unlikely to serve all subfields.

Schema-based extraction can also integrate limited normalization into the extraction step, such as correcting surface-form OCR variants, canonicalizing simple formulas, or standardizing units under schema

constraints^[91]. This capability should be framed cautiously. LLMs learn statistical regularities from training data; they do not possess guaranteed chemical understanding. Mapping ‘Li Co O2’ to ‘LiCoO2’ may be a safe orthographic normalization, but inferring an oxidation state, assigning a CAS number, or resolving a complex molecular identity without source evidence can produce chemically plausible falsehoods. Such cases exemplify fabrication and misattribution and require external chemical databases, rule-based formula checks, or expert review rather than trust in the model alone.

The limitation of schema-based extraction is the mirror image of its strength. Fixed schemas create blind spots for unexpected properties, mechanisms, or experimental conditions. Deeply nested schemas can impose context-window pressure on long documents, and rigid fields may force uncertain evidence into overconfident categories. Schema-based extraction is therefore best suited to known targets and mature data models, not to exploratory discovery. It remains the most controllable paradigm, but its outputs are only as complete as the schema and only as reliable as the grounding and validation layers that surround it.

Schema-free and open knowledge discovery

Schema-free extraction addresses the opposite need: discovering attributes, relations, and reporting patterns that are not known in advance. It is valuable in emerging or interdisciplinary materials domains where predefined schemas lag behind actual reporting practice. Unlike schema-based extraction, schema-free methods aim at ‘unknown unknowns’: new parameters in synthesis recipes, unexpected structure-property relations, emerging device metrics, or reporting conventions that have not yet been formalized. This paradigm is related to Open Information Extraction (openIE), but materials facts are rarely simple subject-predicate-object triples. They are often high-order events with participants, conditions, temporal steps, measurement settings, and uncertainty^[92,93].

Generic OpenIE is not sufficient for this setting because reducing a materials statement to a flat triple can remove context that determines what material or property is actually being described. Dagdelen *et al.* give a concrete example: an “epitaxial La-doped thin film” of HfZrO₄ is scientifically distinct from a “La-doped thin film” or a “La-doped” HfZrO₄ sample; omitting the epitaxial and thin-film qualifiers changes the represented material state and can change its physical properties^[8]. The same issue arises for quantitative records: water-splitting overpotentials must be associated with the reaction [oxygen evolution reaction (OER) or hydrogen evolution reaction (HER)], electrolyte, and current density^[9], and property records may require temperature, pressure, or processing conditions in addition to a Material-Value-Unit triplet^[10]. Consequently, schema-free discovery should not terminate at unqualified triples but should preserve attributes and conditions in event-centric, frame-based, graph-based, or nested representations. These representations can later be consolidated into schemas or knowledge graphs while retaining the provenance and experimental context required for scientific interpretation.

Schema-Miner illustrates the transition from extraction to schema induction. Rather than requiring experts to specify all fields in advance, the system asks LLMs to analyze a curated corpus, propose candidate attributes such as precursor pulse time or purge duration in the atomic-layer-deposition literature, and then rely on expert review and ontology grounding to remove redundancy and enforce semantic consistency^[15,59]. Related panoramic or free-form summarization workflows, including LLM-assisted Starrydata curation, read full documents and reconstruct associations among samples, procedures, plots, and reported results rather than extracting isolated scalar fields^[70,71]. These systems are closer to scientific reading than to conventional field filling.

The advantage of schema-free extraction is coverage. It can reveal which variables authors actually report, how terminology varies across papers, and which attributes are repeatedly omitted from handcrafted schemas. This makes it useful for building first-generation data models in areas where community standards

are immature. It also provides a feedback mechanism for ontology development: frequently recurring induced attributes can become candidates for formalization, whereas rare or ambiguous attributes may remain annotations. In this sense, schema-free extraction is not the absence of structure but a method for discovering structure from literature.

The trade-off is reduced controllability. Without explicit constraints, outputs may drift semantically across documents, duplicate equivalent concepts under different names, or mix source evidence with model inference. Schema-free extraction therefore requires post hoc structuring: entity alignment, unit normalization, ontology mapping, deduplication, contradiction detection, and provenance audit. Its reliability depends less on the first-pass generation and more on the consolidation pipeline that follows. This is why schema-free and schema-based methods are increasingly hybridized: open induction expands coverage, while schemas and ontologies restore comparability and validation.

A recurring practical difficulty is deciding when an induced schema is stable enough to drive large-scale extraction. In practice, saturation can be approximated by tracking the marginal rate at which new documents introduce previously unseen attributes: once this rate falls below a small threshold across a held-out sample, the schema can be treated as provisionally saturated. Semantic drift - where the same attribute is populated with subtly different meanings across documents or sub-domains - should be monitored explicitly, for example by clustering the values and contexts assigned to each field and flagging fields whose usage diverges over time or across sources. The tension between coverage and consistency is not fully resolvable: broader induction captures more of what authors actually report but admits more redundant or ambiguous fields, whereas tighter constraints improve comparability at the cost of missing emerging attributes. For this reason, induced schemas should be validated before large-scale use through ontology alignment, competency-question testing, expert review of a sampled subset, and consistency checks on populated records, rather than being deployed directly from the induction step.

Agentic systems for materials knowledge extraction

Because this is the most rapidly moving frontier, several of the systems discussed below are reported in recent preprints rather than peer-reviewed venues; we therefore present them as emerging directions whose quantitative claims await independent validation.

Materials knowledge extraction is transitioning from single-pass inference to autonomous agentic systems equipped with planning, reflection, and tool-use capabilities. We define an AI agent operationally: an LLM-based system in a control loop that decides on and performs actions - invoking retrievers, databases, code, or simulators - rather than emitting a single output. The minimal requirement is action: selecting and executing tool calls conditioned on intermediate inference. The defining capability is reflection: evaluating intermediate results, detecting errors or missing evidence, and revising plans accordingly. Single-pass models are therefore not agents; act-only systems qualify as basic agents, whereas the planning-act-reflect loop characterizes mature agentic systems^[94-96].

A note of caution is warranted when interpreting these capabilities. Distinguishing an agent that genuinely revises its plan on the basis of intermediate results from one that merely consumes a longer context window is difficult, and the field currently lacks agreed, principled metrics for making this distinction. Many of the systems discussed below are early prototypes evaluated on narrow tasks, and reported behaviors such as reflection or self-correction are often demonstrated anecdotally rather than measured against controlled baselines. Readers should therefore treat claims of agentic maturity as provisional, and the operational definition above as a functional description rather than as evidence that current systems reliably perform genuine plan revision.

This transition is necessary because materials science literature encodes knowledge through non-local, cross-modal dependencies. A performance metric may depend on synthesis details in supplementary sections, or require correlating text with phase diagrams and characterization plots. Single-pass models fail due to limited context windows, attention dilution, and inability to reconcile distributed evidence. The agentic paradigm introduces a closed-loop reasoning cycle - planning, execution, reflection^[96,97] - that decomposes tasks into subgoals, selectively retrieves additional evidence, and iteratively converges on evidence-supported outcomes. This transforms extraction from passive text processing into active problem-solving resembling human scientific reasoning.

By embedding tool use into this loop, agents extend their effective knowledge horizon. They invoke document retrievers, query structured databases, inspect figures or tables, and perform lightweight computations - mitigating context limitations and grounding knowledge in verifiable sources^[97].

Interactive disambiguation and code-based reasoning

A fundamental limitation of conventional extraction is the assumption that semantic boundaries are fixed. Material concepts are often context-dependent with fuzzy definitions: the distinction between precursor, solvent, or additive varies by synthesis route. Agentic systems address this through interactive disambiguation, flagging ambiguous instances and querying experts or auxiliary resources for resolution. Frameworks such as ClassMATE^[98] exemplify this approach, transforming tacit expertise into explicit, machine-interpretable knowledge.

In parallel, code-guided reasoning strengthens logical rigor by expressing extraction logic in programming constructs - class definitions, type hints, constrained schemas - imposing stricter boundaries than natural-language prompts^[99]. Together, these strategies enable robust handling of conceptual uncertainty in evolving materials domains.

Dynamic knowledge extension and the verification closed loop

Agentic systems extend knowledge dynamically beyond model parameters by invoking external tools - document retrievers, structured databases, simulation utilities - as part of reasoning. This is critical in materials science, where information is distributed across supplementary materials, prior publications, or external databases^[18].

More importantly, tool invocation enables verification loops. Agents generate intermediate hypotheses, verify them against sources or external references, and revise conclusions when inconsistencies are detected. Systems such as Eunomia^[20] and nanoMINER^[16] integrate multi-step evidence aggregation and validation to assess materials modification strategies, ensuring that claims are supported by consistent experimental evidence. LLaMP^[19] demonstrates how retrieval over structured materials resources supports higher-fidelity knowledge extraction. By transforming extraction into hypothesis generation and validation, agentic systems move beyond pattern recognition toward evidence-based scientific reasoning.

Distributed cognition through multi-agent collaboration

As tasks scale in complexity, single-agent systems encounter cognitive overload. Multi-agent architectures distribute responsibilities across specialized agents, mirroring human research teams^[15,16,100-103].

Modality division of labor assigns agents by data type. nanoMINER^[16] deploys dedicated visual agents for figure analysis alongside text-focused agents, with a coordinating controller aligning cross-modal outputs. On nanozyme datasets, it attains near-perfect precision (≈ 0.98) for kinetic parameters and surpasses

GPT-4.1 in both precision and recall^[16]. However, because nanozyme kinetic parameters are relatively structured and numerically constrained, this near-perfect precision may not transfer directly to less-structured extraction tasks such as synthesis-recipe mining or polymer characterization. The multi-agent DIVE system for hydrogen-storage materials jointly reads figures and text, reporting extraction gains of ~10%-15% over commercial baselines and ~30% over open-source ones, assembling a database of > 30,000 entries from > 4,000 publications^[17].

Cognitive division of labor assigns agents by reasoning function. In Schema-Miner^[15], distinct agents handle schema induction, ontology alignment, and logical review. This role-based decomposition reduces error accumulation and improves interpretability, marking a shift from monolithic models toward organizational intelligence.

From knowledge extraction to scientific discovery

Agentic systems are expanding beyond extraction toward end-to-end scientific discovery. We deliberately step beyond our core focus to situate extraction within this trajectory, distinguishing demonstrated capabilities - literature mining, retrieval, tool-assisted analysis - from the aspirational vision of autonomous discovery pipelines.

Recent systems illustrate this trend. HoneyComb^[104] focuses on comprehensive retrieval and synthesis to improve factual quality. LLMatDesign^[105] uses agentic reasoning and self-reflection for autonomous materials design. ChatMOF^[106] integrates agents with predictive models for metal-organic frameworks. MatAgent^[107] and MatPilot^[108] emphasize human-agent collaboration, embedding expert feedback into the decision loop. ChemCrow^[18] demonstrates generalizability across chemistry and materials science.

These developments suggest evolution toward AI-scientist prototypes. However, retrieval and knowledge-synthesis components (e.g., HoneyComb^[104], ChemCrow^[18]) are concrete extensions of extraction, whereas fully closed-loop coupling of hypothesis generation, autonomous experimentation, and iterative revision (e.g., LLMatDesign^[105], MatAgent^[107], MatPilot^[108]) remains at an early, largely prototype stage. Addressing risks of hallucination and uncontrolled autonomy requires rigorous quality control and domain-aware constraints, discussed in Section “Reliability, Quality Control, and Evaluation”.

RELIABILITY, QUALITY CONTROL, AND EVALUATION

As LLM-driven extraction scales from case studies to corpus-level database construction, the central bottleneck shifts from how much information can be mined to how reliably extracted records preserve source evidence, semantic identity, and physical validity. This challenge is particularly acute in materials science because a usable record is rarely a standalone string: it typically couples a material name or formula, processing route, measurement condition, property value, unit, and provenance. Errors introduced at any point - layout parsing, entity normalization, relation assignment, unit conversion, or evidence attribution - can corrupt downstream databases and machine-learning models. Generic LLM pipelines are therefore prone not only to ordinary false positives and false negatives, but also to semantic drift, especially when materials aliases, abbreviations, layout noise, and dispersed text-table-figure evidence are involved^[7,109,110]. Reliability must therefore be treated as an integrated design principle across the PCA framework rather than as a final cleaning step.

As summarized in Figure 7, quality assurance in LLM-driven materials extraction can be organized around three linked questions: what fails, how it is controlled, and how it is evaluated. Failure analysis identifies recurrent error modes; layered defenses combine entity standardization, source grounding, ontology/KG

Reliability framework for LLM-driven materials knowledge extraction

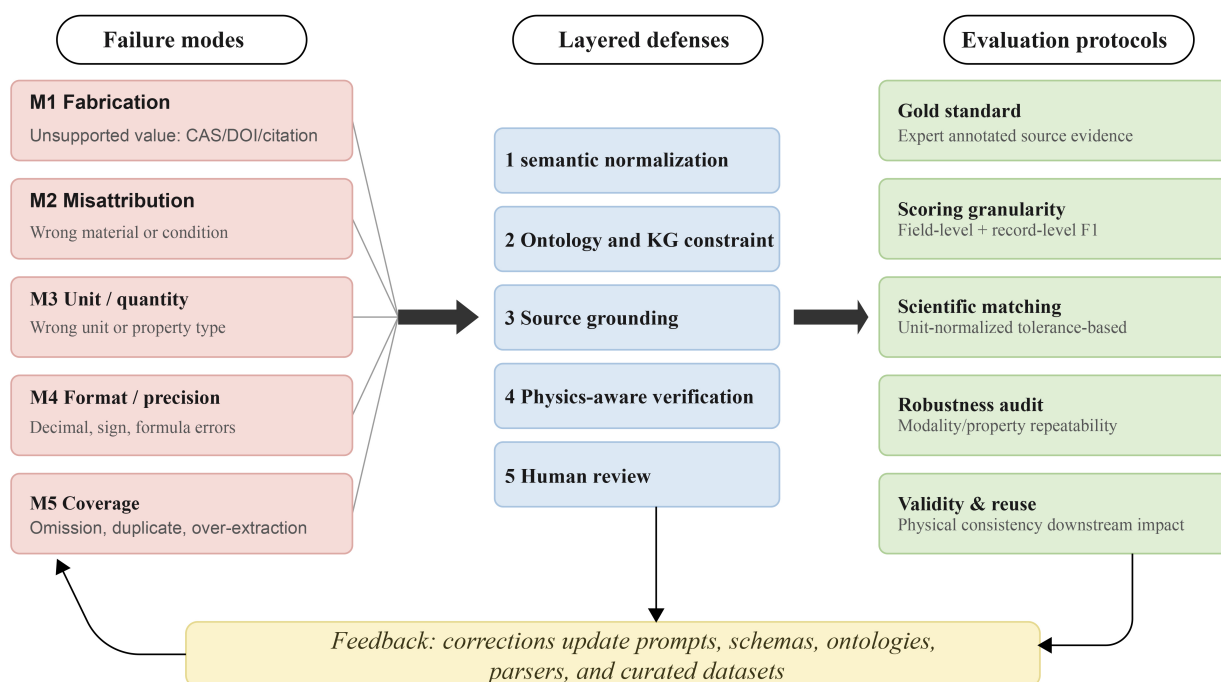


Figure 7. Reliability framework for LLM-driven materials knowledge extraction. The framework links recurrent failure modes to layered defenses and evaluation protocols, emphasizing source grounding, semantic constraints, physical validation, and human review. LLM: Large language model; KG: knowledge graph.

constraints, physical validation, and human review; and evaluation protocols determine whether extracted records are not only textually correct but also scientifically valid.

Failure modes in LLM-based materials extraction

Evidence for these recurring failure modes and their mitigation strategies is drawn from prior work^[67,111–121]. Table 4 summarizes the taxonomy. M1, fabrication, occurs when the model invents a value, entity, relation, CAS number, DOI, or citation not supported by the source. M2, misattribution, occurs when a real value is attached to the wrong material, condition, figure panel, or legend entry. M3, unit and quantity-type error, occurs when the model confuses units, dimensions, or property categories, such as treating a temperature as a time or a gravimetric capacity as an areal capacity. M4, format and precision error, includes misplaced decimals, sign errors, lost significant figures, and malformed chemical strings. M5, coverage error, includes omission, duplication, and over-extraction. These categories are defined at the fact-record level rather than the model level because the downstream question is whether an extracted record is scientifically usable.

These modes often compound across the pipeline^[122]. A figure–caption alignment error in perception becomes a misattributed property in extraction; a normalization error splits equivalent materials into duplicate entities; and an unsupported fact may enter a knowledge graph, only to be retrieved later as apparent evidence. This makes provenance, versioning, and validation integral to the data lifecycle rather than optional post-processing.

The severity of these failures is context-dependent^[2]. A missing optional field may be tolerable for qualitative review, whereas a misattributed measurement condition can invalidate a machine-learning dataset. A unit conversion error corrupts derived descriptors; a fabricated DOI or CAS number misleads curators by appearing authoritative. Materials science is particularly vulnerable because many values are plausible over

Table 4. Taxonomy of failure modes in LLM-based materials knowledge extraction and the quality-control mechanisms that address them

Failure mode	Description	Representative example	Primary mitigation
M1 fabrication	Unsupported value, entity, identifier, relation, or reference is generated	Invented property value; fabricated CAS number or DOI	Evidence anchoring via RAG; NLI entailment; citation-grounded generation ^[67,111–113]
M2 misattribution	Correct value is linked to the wrong material, condition, panel, or legend	Assigning a measured zT to the wrong dopant; wrong curve–legend association	Tuple-level structuring; figure–caption alignment; ontology/KG relation constraints ^[46,47,50,114–117]
M3 unit/quantity-type error	Unit is misread or number is assigned the wrong physical role	Treating 773 K and 500 °C as different; reading “500” as a lattice parameter rather than calcination temperature	Unit normalization; dimensional analysis; ontology type constraints; physics-aware verification ^[118,119]
M4 format/precision error	Quantity is approximately correct but corrupted in precision, formatting, sign, or notation	Decimal-point error; lost exponent in conductivity; malformed formula	Schema-constrained output; repeatability checks; deterministic parsers ^[48,51–53]
M5 coverage error	Relevant facts are omitted, duplicated, or over-extracted	Missing supplementary-table values; duplicated figure/table extraction; extracting cited prior data	Source prioritization; deduplication; provenance tracking; cross-document validation ^[120,121]

LLM: Large language model; RAG: retrieval-augmented generation; KG: knowledge graph.

broad ranges: an incorrect bandgap, modulus, or synthesis temperature may look reasonable unless checked against source evidence and physical constraints. Reliability evaluation should therefore distinguish benign omissions from critical scientific errors, rather than treating all mismatches as equivalent.

Layered defenses: entity standardization, source grounding, ontology constraints, physical checks, and human review

The first defense layer is entity standardization and semantic normalization. LLMs can correct some surface-level variation during extraction, especially when constrained by examples or schema formats^[8]. For instance, LLM-NERRE shows that fine-tuned GPT-3 and Llama-2 models can learn annotator-consistent extraction and normalization patterns from only a few hundred examples, reaching $F1 \approx 0.82$ for host–dopant extraction compared with ≈ 0.39 for a MatBERT-proximity baseline^[8,123]. However, this capability should be carefully delimited. Orthographic normalization is comparatively safe, whereas inferring unstated chemical attributes - oxidation states, identifiers, or registry numbers - is error-prone. More reliable normalization should therefore combine LLM outputs with deterministic or database-backed tools, including OPSIN for name-to-structure conversion, SMILES/InChI representations, STOUT V2.0 for bidirectional name–SMILES translation, ChemProps for polymer-name standardization, and stoichiometry-aware alignment frameworks such as Wikontic^[21,124–126].

The second layer is ontology and KG constraints. Entity normalization answers “what the object is”, whereas ontologies and KGs constrain “which relations are admissible”. Materials ontologies, including upper-level frameworks such as EMMO and domain-specific extensions, provide type systems and relation constraints that reduce semantic drift and support consistency checking^[50,114–117]. In practice, this can be implemented through schema-constrained generation, where JSON Schema or ontology-derived constraints restrict output structure and value types^[51–53], and through ontology-aware retrieval, where textual mentions are aligned to admissible ontology classes. Large-scale KGs such as MatKG^[65], MGED-KG^[127], and FMKG^[62] further convert isolated extractions into connected evidence structures, enabling provenance tracking, cross-document consistency checks, and downstream reasoning.

The third layer is source grounding and retrieval-based verification. Conversational extraction systems such as ChatExtract use multi-turn follow-up prompts and self-consistency checks to reduce mismatches, achieving approximately 90% precision/recall for representative property triplets such as bulk modulus and

metallic-glass critical cooling rate^[110]. Cross-document validation can flag outliers when multiple sources report related facts^[120,121], while NLI-style entailment checks test whether generated claims are supported by cited source text^[111]. RAG further reduces fabrication by conditioning generation on retrieved evidence^[112,113]. In materials informatics, dual-source grounding is especially valuable: unstructured literature retrieval supplies mechanistic and experimental context, whereas structured databases such as the Materials Project provide numerical anchors and reference ranges^[128-130]. Citation-based generation strengthens auditability by requiring each claim to be linked to evidence^[67].

The fourth layer is physics-aware verification, which is the most materials-specific component of quality control. Unit and dimensional analysis can deterministically expose category errors, while enforced unit normalization has been shown to reveal inconsistencies in extracted records^[118]. Derived-property checks can recompute dependent quantities and test internal consistency - for example, recalculating thermoelectric zT from Seebeck coefficient, electrical conductivity, thermal conductivity, and temperature^[131]. Frameworks such as VIRF operationalize this idea through a mentor-apprentice loop, where symbolic or physics-based validators provide corrective feedback when LLM-generated recipes or values violate basic scientific constraints^[119]. Unlike self-consistency or RAG, these checks can provide definite pass/fail signals whenever the extracted fact admits deterministic validation.

The final layer is human-in-the-loop review. Expert review remains essential for ambiguous records, schema induction, ontology extension, and high-impact datasets. The goal is not exhaustive manual reannotation, but targeted curation supported by provenance, confidence scores, detected failure mode, and validator feedback. The most reliable pipelines are therefore hybrid: deterministic checks anchor facts that can be formally validated, probabilistic checks reduce unsupported generation, and human review resolves residual uncertainty^[132].

Evaluation protocols: from accuracy to scientific validity

Reliable extraction also requires transparent evaluation. A useful protocol should define the gold standard, matching rules, scoring level, and stratification strategy. Gold standards are typically constructed through expert annotation; for example, Wang *et al.* use multi-expert validation before benchmarking extraction quality^[133]. Precision, recall, and F1 remain useful, but they must be reported at both field level and tuple/record level. Field-level evaluation scores isolated attributes, whereas tuple-level evaluation requires material identity, conditions, property, value, and unit to be jointly correct. Tuple-level accuracy is more scientifically meaningful because a correct number assigned to the wrong material is unusable.

Matching rules should reflect physical equivalence rather than text overlap. Metrics such as BLEU or ROUGE are poorly suited to materials extraction because “773 K” and “500 °C” are textually different but physically identical^[134]. Evaluation should therefore use normalization-aware matching for units, chemical identifiers, synonyms, and numerical values. For figure-derived data, tolerance-based scoring is essential: Srour *et al.* define correctness using a 2% axis-scale tolerance, while PlotExtract reports roughly 5% positional error with over 90% precision and recall for two-axis plots^[38,48]. These examples show that modality-specific evaluation is necessary. Implementing normalization-aware matching consistently is, however, a substantial engineering undertaking rather than a solved problem. Different physical quantities require different normalization logic: band gaps in eV, synthesis temperatures in K vs. °C, lattice parameters in Å vs. nm, and capacities in mAh g⁻¹ vs. mAh cm⁻² each demand distinct unit systems, reference bases, and equivalence rules. Building and maintaining validators that correctly cover all such quantities and their unit variants across the many subfields of materials science therefore remains an open and labor-intensive challenge.

Results should also be stratified by modality and property type. Srer *et al.* report $F1 > 99\%$ for text and tables but only 85.8% for figures on the same corpus^[48]. Ghosh and Tewari similarly show property-dependent variation, with $F1 \approx 0.91$ for thermoelectric properties but ≈ 0.82 for structural fields^[102]. Because LLM outputs are stochastic, repeatability should also be reported; in one study, approximately 87% of extracted items were identical across five runs, with remaining variance concentrated in figure reading and decimal placement^[48].

Evaluation is now expanding beyond extraction accuracy toward scientific validity. Physical-consistency scoring treats violations of chemical or physical constraints as critical errors. Source-tracked extraction allows each value to be audited against a cited passage, table cell, figure panel, or caption. LLM-as-a-judge methods, such as LeMat-Synth, can help scale rubric-based evaluation, but they should be treated as auxiliary triage tools because the judge may inherit the same biases and hallucination tendencies as the extractor^[135].

At the benchmark level, the field is developing a layered ecosystem. MatSciRE^[136], SuperMat^[137], and MatSciNLP^[138] evaluate core materials text understanding and extraction; LLM4Mat-Bench^[139] and LLM4Mol^[140] extend evaluation toward crystal, molecular, and predictive reasoning tasks; and multimodal benchmarks such as MatViX^[37], DocGenome^[43], and Macbench^[141] evaluate chart understanding, multimodal document reasoning, and long-context scientific interpretation. Together, these benchmarks mark a shift from string-level correctness toward source-supported and physically valid knowledge extraction. However, standardized task definitions, gold-standard construction procedures, tolerance rules, and reporting conventions remain insufficient. Establishing community-wide evaluation protocols is therefore essential for trustworthy materials databases, knowledge graphs, and agentic discovery systems.

OUTLOOK

LLM-driven materials knowledge extraction is moving from isolated information-retrieval tasks toward reliability-aware materials intelligence^[142]. The field has already shown that LLMs can parse complex scientific documents, induce or populate structured schemas, align extracted facts with ontologies and KGs, and coordinate multi-step extraction workflows. These capabilities, however, should be viewed as an intermediate stage rather than a mature endpoint. The next phase will be shaped by four connected directions: knowledge-native infrastructures, bounded agentic coordination, uncertainty-aware trustworthiness, and evaluation for scientific utility.

First, future systems are likely to evolve from extraction-centric pipelines into knowledge-native infrastructures. Current workflows often treat extraction as a downstream operation applied to static literature. A more robust architecture would instead treat evidence-linked databases, ontologies, and KGs as continuously updated research objects. In such systems, extraction, normalization, ontology alignment, provenance tracking, and reasoning would become mutually constraining components rather than separate modules. This is important because extracted datasets increasingly serve as retrieval corpora, training data, or KG backbones for subsequent LLM systems. Without provenance and version control, errors may be amplified through self-reinforcing knowledge loops, and indiscriminate reuse of model-generated records may accelerate model degradation or model collapse^[143]. Reliable knowledge infrastructures therefore require clear separation of human-curated, machine-extracted, and machine-generated records, together with dataset versioning and evidence-level audit trails.

Second, agentic systems will increasingly function as scientific coordinators rather than standalone extraction tools [Figure 8]. In the near term, their most realistic role is to connect literature-derived priors with simulation, prediction, and experimental-planning workflows under bounded autonomy. For example, an agent may mine prior synthesis conditions, retrieve comparable materials from structured databases,

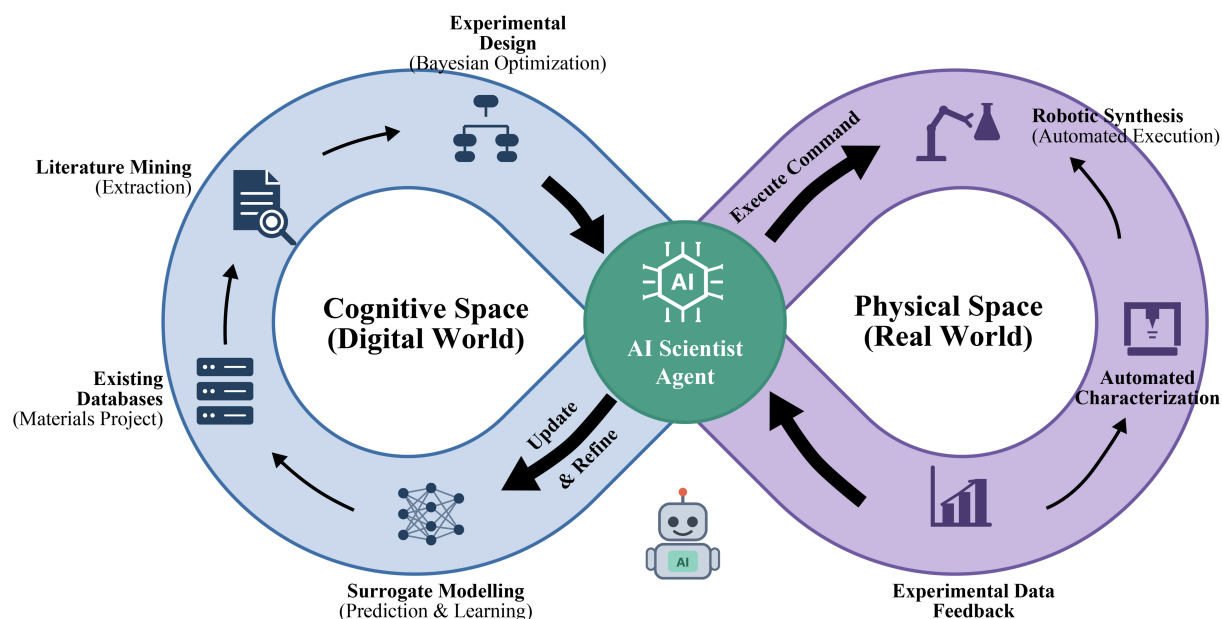


Figure 8. AI Scientist Agent: A closed-loop ecosystem integrating cognitive and physical spaces for autonomous materials discovery. AI: Artificial intelligence.

query predictive models, and suggest candidate experiments for human approval. In the longer term, such systems could support self-driving laboratories (SDLs), where literature mining provides structured prior knowledge for hypothesis generation, candidate selection, and active-learning or Bayesian-optimization loops^[144]. This transition should nevertheless be framed cautiously. Current systems increasingly support literature mining, retrieval, tool-assisted reasoning, and limited planning, but fully autonomous closed-loop discovery - where hypothesis generation, experiment execution, interpretation, and revision proceed without expert intervention - remains largely aspirational^[145]. Human oversight, provenance tracking, tool-level logging, and verifiable reasoning chains will remain necessary for scientific accountability and experimental safety.

Third, trustworthiness and uncertainty awareness will become decisive deployment constraints. As extraction systems begin to influence experimental decisions, a single record or recommendation without confidence is insufficient. The quality-control framework in Section “Reliability, Quality Control, and Evaluation” shows that fabrication, misattribution, unit mistakes, precision loss, and coverage errors have distinct causes and require different defenses. Future systems should therefore report calibrated uncertainty at record, field, and source-evidence levels, and should support selective abstention when confidence is low^[132]. This is especially important for figure-derived values, long-document reasoning, and agentic workflows, where stochastic variation and error propagation are difficult to detect. Uncertainty-aware extraction would also make expert review scalable: high-confidence records could be ingested after deterministic checks, whereas low-confidence or high-impact records could be routed to human curators.

Fourth, evaluation must continue to shift from task accuracy to scientific utility. Field-level F1 remains useful, but it does not fully capture whether an extracted record is source-supported, semantically well typed, physically valid, and reusable in downstream modeling. Future benchmarks should therefore evaluate complete processing–structure–property tuples, multimodal evidence alignment, source-span fidelity, unit and dimensional consistency, and decision relevance. This is particularly important for agentic systems, where the final answer may depend on multiple retrieval, parsing, normalization, and tool-use steps.

Evaluation should also address contamination and feedback effects: if benchmarks, extracted corpora, and model-training data overlap, apparent progress may reflect memorization rather than reliable scientific reasoning. Contamination-aware benchmarks, versioned datasets, and transparent reporting of training and retrieval sources will be essential for fair comparison and reproducibility^[146].

Finally, practical deployment will depend on cost, access, and governance. Corpus-scale mining of full-text papers, figures, tables, and supplementary files is computationally expensive and constrained by copyright, licensing, and publisher access. Large proprietary models offer strong general capabilities but may be costly and difficult to audit, whereas compact domain-specialized models may offer better controllability, lower energy cost, and easier integration into reproducible pipelines. A likely future architecture is therefore hybrid: large models for complex reasoning and schema induction, smaller fine-tuned models for high-throughput extraction, deterministic tools for chemical and physical validation, and human experts for ambiguous or high-value cases. The near-term goal is not full automation for its own sake, but rather evidence-linked and physically validated knowledge infrastructures that make literature-derived materials data reusable at scale.

DECLARATIONS

Acknowledgments

The authors thank Prof. Shujie Tang and Prof. Xiao Kong of the Shanghai Institute of Microsystem and Information Technology, Chinese Academy of Sciences, for valuable discussions and insightful advice.

Authors' contributions

Conception and design of the review: Yang, S.; Ran, N.; Lu, Y.

Manuscript drafting: Yang, S.; Wang, Y.; Ran, N.

Manuscript editing: Yang, S.; Wang, Y.; Tu, Q.

Review and approval of the final manuscript: Yang, S.; Wang, Y.; Tu, Q.; Ran, N.; Lu, Y.

Availability of data and materials

Not applicable.

AI and AI-assisted tools statement

During the preparation of this manuscript, the AI tool DeepSeek (version V3.2, released 2025-12-01) was used solely for language editing. The tool did not influence the study design, data collection, analysis, interpretation, or the scientific content of the work. All authors take full responsibility for the accuracy, integrity, and final content of the manuscript.

Financial support and sponsorship

This work is financially supported by the Innovation Foundation of National Science Library (Chengdu) (E3Z0000903), Natural Science Foundation of Sichuan Province for Distinguished Young Scholars (2023NSFSC1971), Advanced Materials-National Science and Technology Major Project (2025ZD0613500, 2025ZD0619500), National Natural Science Foundation of China (22403103), AI for Science Program, Shanghai Municipal Commission of Economy and Informatization (2025-GZL-RGZN-BTBX-01005), Sponsored by Shanghai Sailing Program (23YF1454900).

Conflicts of interest

All authors declared that there are no conflicts of interest.

Ethical approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Copyright

© The Author(s) 2026.

REFERENCES

1. Agrawal, A.; Choudhary, A. Perspective: Materials informatics and big data: realization of the “fourth paradigm” of science in materials science. *APL. Mater.* **2016**, *4*, 053208. DOI
2. Kononova, O.; He, T.; Huo, H.; Trewartha, A.; Olivetti, E. A.; Ceder, G. Opportunities and challenges of text mining in materials research. *iScience* **2021**, *24*, 102155. DOI PubMed PMC
3. Jessop, D. M.; Adams, S. E.; Willighagen, E. L.; Hawizy, L.; Murray-Rust, P. OSCAR4: a flexible architecture for chemical text-mining. *J. Cheminform.* **2011**, *3*, 41. DOI PubMed PMC
4. Hawizy, L.; Jessop, D. M.; Adams, N.; Murray-Rust, P. ChemicalTagger: a tool for semantic text-mining in chemistry. *J. Cheminform.* **2011**, *3*, 17. DOI PubMed PMC
5. Swain, M. C.; Cole, J. M. ChemDataExtractor: a toolkit for automated extraction of chemical information from the scientific literature. *J. Chem. Inf. Model.* **2016**, *56*, 1894–904. DOI PubMed
6. Mavračić, J.; Court, C. J.; Isazawa, T.; Elliott, S. R.; Cole, J. M. ChemDataExtractor 2.0: autopopulated ontologies for materials science. *J. Chem. Inf. Model.* **2021**, *61*, 4280–9. DOI PubMed
7. Hira, K.; Zaki, M.; Sheth, D.; Mausam; Krishnan, N. M. A. Reconstructing the materials tetrahedron: challenges in materials information extraction. *Digit. Discov.* **2024**, *3*, 1021–37. DOI
8. Dagdelen, J.; Dunn, A.; Lee, S.; et al. Structured information extraction from scientific text with large language models. *Nat. Commun.* **2024**, *15*, 1418. DOI PubMed PMC
9. Yi, G. H.; Choi, J.; Song, H.; et al. MaTableGPT: GPT-based table data extractor from materials science literature. *Adv. Sci.* **2025**, *12*, e2408221. DOI PubMed PMC
10. Polak, M. P.; Morgan, D. Extracting accurate materials data from research papers with conversational language models and prompt engineering. *Nat. Commun.* **2024**, *15*, 1569. DOI PubMed PMC
11. Li, C.; Ran, N.; Liu, J. Agentic material science. *J. Mater. Inf.* **2026**, *6*, 10. DOI
12. Xu, D.; Chen, W.; Peng, W.; et al. Large language models for generative information extraction: a survey. *Front. Comput. Sci.* **2024**, *18*, 40555. DOI
13. Pei, Z.; Yin, J.; Zhang, J. Language models for materials discovery and sustainability: progress, challenges, and opportunities. *Prog. Mater. Sci.* **2025**, *154*, 101495. DOI
14. Hira, K.; Zaki, M.; Mausam; Krishnan, N. M. A. MatSKRAFT: a framework for large-scale materials knowledge extraction from scientific tables. *arXiv* **2025**, arXiv:2509.10448. Available online: <https://doi.org/10.48550/arXiv.2509.10448>. (accessed on 29 Jul 2026).
15. Sadruddin, S.; D’Souza, J.; Poupaki, E.; et al. Schema-Miner Pro: agentic AI for ontology grounding over LLM-discovered scientific schemas in a human-in-the-loop workflow. *Semant. Web.* **2025**. DOI
16. Odobesku, R.; Romanova, K.; Mirzaeva, S.; et al. Agent-based multimodal information extraction for nanomaterials. *npj. Comput. Mater.* **2025**, *11*, 1674. DOI
17. Zhang, D.; Jia, X.; Tran, H. B.; et al. “DIVE” into hydrogen storage materials discovery with AI agents. *Chem. Sci.* **2026**, *17*, 3031–42. DOI PubMed PMC
18. Bran, A. M.; Cox, S.; Schilter, O.; Baldassari, C.; White, A. D.; Schwaller, P. Augmenting large language models with chemistry tools. *Nat. Mach. Intell.* **2024**, *6*, 525–35. DOI PubMed PMC
19. Chiang, Y.; Hsieh, E.; Chou, C. H.; Riebesell, J. LLaMP: large language model made powerful for high-fidelity materials knowledge retrieval. *arXiv* **2024**, arXiv:2401.17244. Available online: <https://doi.org/10.48550/arXiv.2401.17244>. (accessed on 29 Jul 2026).
20. Ansari, M.; Moosavi, S. M. Agent-based learning of materials datasets from the scientific literature. *Digit. Discov.* **2024**, *3*, 2607–17. DOI
21. Chepurova, A.; Bulatov, A.; Burtsev, M.; Kuratov, Y. Wikontic: constructing wikidata-aligned, ontology-aware knowledge graphs with large language models. *arXiv* **2025**, arXiv:2512.00590. Available online: <https://doi.org/10.48550/arXiv.2512.00590>. (accessed on 29 Jul 2026).
22. Zhang, Q.; Wang, B.; Huang, V. S. J.; et al. Document parsing unveiled: techniques, challenges, and prospects for structured information extraction. *arXiv* **2024**, arXiv:2410.21169. Available online: <https://doi.org/10.48550/arXiv.2410.21169>. (accessed on 29 Jul 2026).
23. Van, M. H.; Verma, P.; Zhao, C.; Wu, X. A survey of AI for materials science: foundation models, LLM agents, datasets, and tools. *arXiv* **2025**, arXiv:2506.20743. Available online: <https://doi.org/10.48550/arXiv.2506.20743>. (accessed on 29 Jul 2026).
24. Zhong, X.; Tang, J.; Yepes, A. J. PubLayNet: largest dataset ever for document layout analysis. In *Proceedings of the 2019 International Conference on Document Analysis and Recognition (ICDAR)*, Sydney, Australia. Sep 20–25, 2019. IEEE; 2019. pp. 1015–22. DOI

-
25. Li, M.; Xu, Y.; Cui, L.; et al. DocBank: a benchmark dataset for document layout analysis. *arXiv* **2020**, arXiv:2006.01038. Available online: <https://doi.org/10.48550/arXiv.2006.01038>. (accessed on 29 Jul 2026).
 26. Chi, Z.; Huang, H.; Xu, H. D.; Yu, H.; Yin, W.; Mao, X. L. Complicated table structure recognition. *arXiv* **2019**, arXiv:1908.04729. Available online: <https://doi.org/10.48550/arXiv.1908.04729>. (accessed on 29 Jul 2026).
 27. Zhong, X.; ShafieiBavani, E.; Jimeno Yepes, A. Image-based table recognition: data, model, and evaluation. *arXiv* **2019**, arXiv:1911.10683. Available online: <https://doi.org/10.48550/arXiv.1911.10683>. (accessed on 29 Jul 2026).
 28. Smock, B.; Pesala, R.; Abraham, R. PubTables-1M: towards comprehensive table extraction from unstructured documents. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, USA, Jun 18–24, 2022. IEEE; 2022. pp. 4624–32. DOI
 29. Huang, Y.; Lv, T.; Cui, L.; Lu, Y.; Wei, F. LayoutLMv3: pre-training for document AI with unified text and image masking. *arXiv* **2022**, arXiv:2204.08387. Available online: <https://doi.org/10.48550/arXiv.2204.08387>. (accessed on 29 Jul 2026).
 30. Kim, G.; Hong, T.; Yim, M.; et al. OCR-free document understanding transformer. *arXiv* **2021**, arXiv:2111.15664. Available online: <https://doi.org/10.48550/arXiv.2111.15664>. (accessed on 29 Jul 2026).
 31. Lee, K.; Joshi, M.; Turc, I.; et al. Pix2Struct: screenshot parsing as pretraining for visual language understanding. *arXiv* **2022**, arXiv:2210.03347. Available online: <https://doi.org/10.48550/arXiv.2210.03347>. (accessed on 29 Jul 2026).
 32. Liao, W.; Wang, J.; Li, H.; Wang, C.; Huang, J.; Jin, L. DocLayLLM: an efficient multi-modal extension of large language models for text-rich document understanding. *arXiv* **2024**, arXiv:2408.15045. Available online: <https://doi.org/10.48550/arXiv.2408.15045>. (accessed on 29 Jul 2026).
 33. Shpigel Nacson, M.; Aberdam, A.; Ganz, R.; et al. DocVLM: make your VLM an efficient reader. *arXiv* **2024**, arXiv:2412.08746. Available online: <https://doi.org/10.48550/arXiv.2412.08746>. (accessed on 29 Jul 2026).
 34. Huang, B.; Hua, H.; Yu, Z.; Darrell, T.; Feris, R.; Herzig, R. DAVE: a VLM vision encoder for document understanding and web agents. *arXiv* **2025**, arXiv:2512.17221. Available online: <https://doi.org/10.48550/arXiv.2512.17221>. (accessed on 29 Jul 2026).
 35. Circi, D.; Bradley, M.; Blouir, S.; et al. Information extraction from diverse charts in materials science. In *Proceedings of the LLM for Scientific Discovery: Reasoning, Assistance, and Collaboration*, 2025. <https://openreview.net/forum?id=vj8dqNrZEE>. (accessed on 29 Jul 2026).
 36. Liu, F.; Eisenschlos, J. M.; Piccinno, F.; et al. DePlot: one-shot visual language reasoning by plot-to-table translation. *arXiv* **2022**, arXiv:2212.10505. Available online: <https://doi.org/10.48550/arXiv.2212.10505>. (accessed on 29 Jul 2026).
 37. Khalighinejad, G.; Scott, S.; Liu, O.; et al. MatViX: multimodal information extraction from visually rich articles. *arXiv* **2024**, arXiv:2410.20494. Available online: <https://doi.org/10.48550/arXiv.2410.20494>. (accessed on 29 Jul 2026).
 38. Polak, M. P.; Morgan, D. Leveraging vision capabilities of multimodal LLMs for automated data extraction from plots. *arXiv* **2025**, arXiv:2503.12326. Available online: <https://doi.org/10.48550/arXiv.2503.12326>. (accessed on 29 Jul 2026).
 39. Fang, X.; Wang, J.; Cai, X.; et al. MolParser: end-to-end visual recognition of molecule structures in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Honolulu, USA, Oct 19–25, 2025. IEEE; 2025. pp. 24528–38. DOI
 40. Wang, J.; He, Y.; Yang, H.; et al. GTR-CoT: graph traversal as visual chain of thought for molecular structure recognition. *arXiv* **2025**, arXiv:2506.07553. Available online: <https://doi.org/10.48550/arXiv.2506.07553>. (accessed on 29 Jul 2026).
 41. Zhang, W.; Wang, X.; Feng, B.; Liu, W. MolSight: optical chemical structure recognition with SMILES pretraining, multi-granularity learning and reinforcement learning. *arXiv* **2025**, arXiv:2511.17300. Available online: <https://doi.org/10.48550/arXiv.2511.17300>. (accessed on 29 Jul 2026).
 42. Rajan, K.; Brinkhaus, H. O.; Zielesny, A.; Steinbeck, C. Advancements in hand-drawn chemical structure recognition through an enhanced DECIMER architecture. *J. Cheminform.* **2024**, *16*, 78. DOI PubMed PMC
 43. Xia, R.; Mao, S.; Yan, X.; et al. DocGenome: an open large-scale scientific document benchmark for training and testing multi-modal large language models. *arXiv* **2024**, arXiv:2406.11633. Available online: <https://doi.org/10.48550/arXiv.2406.11633>. (accessed on 29 Jul 2026).
 44. Dong, K.; Chang, Y.; Goh, X. D.; Li, D.; Tang, R.; Liu, Y. MMDocIR: benchmarking multimodal retrieval for long documents. *arXiv* **2025**, arXiv:2501.08828. Available online: <https://doi.org/10.48550/arXiv.2501.08828>. (accessed on 29 Jul 2026).
 45. Dong, K.; Huang, S.; Ye, F.; et al. Doc-Researcher: a unified system for multimodal document parsing and deep research. *arXiv* **2025**, arXiv:2510.21603. Available online: <https://doi.org/10.48550/arXiv.2510.21603>. (accessed on 29 Jul 2026).
 46. Clark, C.; Divvala, S. PDFFigures 2.0: mining figures from research papers. In *Proceedings of the 2016 IEEE/ACM Joint Conference on Digital Libraries (JCDL)*, Newark, USA, Jun 19–23, 2016. IEEE; 2016. pp. 143–52. <https://ieeexplore.ieee.org/document/7559577>. (accessed on 29 Jul 2026).
 47. Cao, S.; Liu, K. Figuring out figures: using textual references to caption scientific figures. *arXiv* **2024**, arXiv:2407.11008. Available online: <https://doi.org/10.48550/arXiv.2407.11008>. (accessed on 29 Jul 2026).

-
48. Sror, G.; Etzyoni, G.; Fire, M.; Berkovich, R.; Rokach, L.; Herskowitz, M. AI-driven methodology for mining scientific literature and extracting databases: a case study on a chemical process. *Energy. AI* **2026**, *24*, 100741. DOI
 49. Sayeed, H. M.; Clark, C.; Sparks, T. D. KnowMat: transforming unstructured material science literature into structured knowledge. 2025. https://www.researchgate.net/publication/392925301_KnowMat_Transforming_Unstructured_Material_Science_Literature_into_Structured_Knowledge. (accessed on 29 Jul 2026).
 50. Bayerlein, B.; Schilling, M.; Birkholz, H.; et al. PMD Core Ontology: achieving semantic interoperability in materials science. *Mater. Design* **2024**, *237*, 112603. DOI
 51. Schilling, M.; Bresch, H.; Bayerlein, B.; Ruehle, B. WeChemSynOntology: semantic modeling of wet chemical syntheses in a self-driving lab for nano-and advanced materials. *Digit. Discov.* **2026**, *5*, 2074-88. DOI
 52. Garcia Trelles, E.; Schweizer, C.; Thomas, A.; von Hartrott, P.; Janka-Ramm, M. Digitalizing material knowledge: a practical framework for ontology-driven knowledge graphs in process chains. *Appl. Sci.* **2024**, *14*, 11683. DOI
 53. Dreger, M.; Malek, K.; Eikerling, M. Large language models for knowledge graph extraction from tables in materials science. *Digit. Discov.* **2025**, *4*, 1221-31. DOI
 54. Saeedizade, M. J.; Blomqvist, E. Navigating ontology development with large language models. In *Proceedings of the The Semantic Web*, 2024. Springer, Cham; 2024. pp. 143-61. DOI
 55. Mateiu, P.; Groza, A. Ontology engineering with large language models. In *Proceedings of the 2023 25th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC)*, Nancy, France. Sep 11-14, 2023. IEEE; 2023. pp. 226-9. DOI
 56. Lippolis, A. S.; Saeedizade, M. J.; Keskisärkkä, R.; Gangemi, A.; Blomqvist, E.; Nuzzolese, A. G. Large language models assisting ontology evaluation. In *The Semantic Web - ISWC 2025*, 2026. Springer, Cham; 2026. pp. 502-20. DOI
 57. García-Fernández, J.; Verhoosel, J.; Ubacht, J.; Bakker, R. M. Ontology engineering with large language models: unveiling the potential of human-LLM collaboration in the ontology extension process. 2025. https://ceur-ws.org/Vol-4020/Paper_ID_8.pdf. (accessed on 29 Jul 2026).
 58. Li, J.; Poveda, M.; Garijo, D. Large language models for ontology engineering: a systematic literature review. *Semant. Web.* **2025**. DOI
 59. Sadruddin, S.; D'Souza, J.; Poupaki, E.; et al. LLMs4SchemaDiscovery: a human-in-the-loop workflow for scientific schema mining with large language models. *arXiv* **2025**, arXiv:2504.00752. Available online: <https://doi.org/10.48550/arXiv.2504.00752>. (accessed on 29 Jul 2026).
 60. Sharma, K.; Kumar, P.; Li, Y. OG-RAG: ontology-grounded retrieval-augmented generation for large language models. *arXiv* **2024**, arXiv:2412.15235. Available online: <https://doi.org/10.48550/arXiv.2412.15235>. (accessed on 29 Jul 2026).
 61. Buehler, M. J. Generative retrieval-augmented ontologic graph and multiagent strategies for interpretive large language model-based materials design. *ACS. Eng. Au.* **2024**, *4*, 241-77. DOI PubMed PMC
 62. Ye, Y.; Ren, J.; Wang, S.; et al. Construction and application of materials knowledge graph in multidisciplinary materials science via large language model. *arXiv* **2024**, arXiv:2404.03080. Available online: <https://doi.org/10.48550/arXiv.2404.03080>. (accessed on 29 Jul 2026).
 63. Bai, X.; He, S.; Li, Y.; et al. Construction of a knowledge graph for framework material enabled by large language models and its application. *npj. Comput. Mater.* **2025**, *11*, 1540. DOI
 64. An, Y.; Greenberg, J.; Uribe-Romo, F. J.; et al. Knowledge graph question answering for materials science (KGQA4MAT). In *Metadata and Semantic Research*, 2023. Springer, Cham; 2024. pp. 18-29. DOI
 65. Venugopal, V.; Olivetti, E. MatKG: an autonomously generated knowledge graph in material science. *Sci. Data.* **2024**, *11*, 217. DOI PubMed PMC
 66. Li, S.; Wei, S.; Huang, C.; Zhang, Y.; Zhang, G.; Sun, S. Extracting and reconstructing knowledge in materials science literature using large language models. *Commun. Mater.* **2026**, *7*, 1043. DOI
 67. Lewis, P.; Perez, E.; Piktus, A.; et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. *arXiv* **2020**, arXiv:2005.11401. Available online: <https://doi.org/10.48550/arXiv.2005.11401>. (accessed on 29 Jul 2026).
 68. Taskiran, N. P.; Tsai, C. J.; Huang, S.; Chakraborty, A.; Venkatasubramanian, V. A knowledge-graph-based pharmaceutical engineering chatbot for drug discovery. *Comput. Chem. Eng.* **2025**, *203*, 109318. DOI
 69. Mishra, V.; Singh, S.; Zaki, M.; et al. LLaMat: large language models for materials science. In *Proceedings of the AI for Accelerated Materials Design-Vienna*, 2024. <https://neurips.cc/virtual/2024/103725>. (accessed on 29 Jul 2026).
 70. Katsura, Y.; Kumagai, M.; Mato, T.; et al. Starrydata: from published plots to shared materials data. *Sci. Technol. Adv. Mater. Methods.* **2025**, *5*, 2506976. DOI
 71. Katsura, Y.; Mato, T.; Takada, Y.; et al. Development of LLM-assisted data curation tools for the Starrydata materials science database. *Sci. Technol. Adv. Mater. Methods.* **2025**, *5*, 2590811. DOI
 72. Tang, H.; Chen, Z.; Yao, X.; Xu, Z. Mechanical performance dataset for alloy with applications at low temperatures. *Sci. Data.* **2025**, *12*, 1235. DOI PubMed PMC

-
73. Itani, S.; Zhang, Y.; Zang, J. The northeast materials database for magnetic materials. *Nat. Commun.* **2025**, *16*, 9415. DOI PubMed PMC
74. Kang, Y.; Lee, W.; Bae, T.; Han, S.; Jang, H.; Kim, J. Harnessing large language models to collect and analyze metal-organic framework property data set. *J. Am. Chem. Soc.* **2025**, *147*, 3943–58. DOI PubMed
75. Sipilä, M.; Mehryary, F.; Pyysalo, S.; Ginter, F.; Todorović, M. Annotated textual dataset PV600 of perovskite bandgaps for information extraction from literature. *Sci. Data.* **2025**, *12*, 1401. DOI PubMed PMC
76. Lee, S.; Cruse, K.; Gleason, S. P.; Alivisatos, A. P.; Ceder, G.; Jain, A. Data-driven analysis of text-mined seed-mediated syntheses of gold nanoparticles. *Digit. Discov.* **2025**, *4*, 93–104. DOI
77. Chen, H.; Liu, H.; Tew, Y.; Ren, X.; Tang, X.; Wang, X. Distilling knowledge from catalysis literature with long-context large language model agents. *ACS. Catal.* **2025**, *15*, 18244–54. DOI
78. Su, H.; Lei, S.; Chen, Y.; Lv, Y.; Zhang, Q. Human-machine collaborative design of SnTe-based thermoelectric materials via a multiagent framework leveraging large language models. *ACS. Appl. Mater. Interfaces.* **2025**, *17*, 63572–84. DOI PubMed
79. Li, C.; Wang, Y.; Wang, Q.; et al. Material Synthesis 2025 (MatSyn25) data set for 2D materials. *ACS. Appl. Mater. Interfaces.* **2026**, *18*, 7092–102. DOI PubMed
80. Chen, Z.; Li, T.; Yang, Y.; et al. Intelligent information extraction pipeline driven by large language model for building polymer processing database. *Polymer* **2025**, *336*, 128875. DOI
81. Zhou, H.; Bai, Y.; Chen, Y.; et al. Large language models and machine learning framework for predicting dental ceramics performance. *Int. Dent. J.* **2026**, *76*, 109358. DOI PubMed PMC
82. Lee, S.; Cruse, K.; Baibakova, V.; Ceder, G.; Jain, A. Text-mined dataset of solid-state syntheses with impurity phases using large language model. *Sci. Data.* **2025**, *12*, 1969. DOI PubMed PMC
83. Cheng, Y.; Zhang, H. L2D: a versatile literature-to-data set pipeline for complex, user-specific data extraction in environmental research. *ACS. EST. Water.* **2025**, *5*, 4897–907. DOI
84. Leng, Y.; Zhong, Y.; Gu, Z.; et al. Intelligent, personalized scientific assistant via large language models for solid-state battery research. *ACS. Mater. Lett.* **2025**, *7*, 1807–16. DOI
85. Hu, H.; Stirrat, H. J.; Alayli, A.; Saeki, A.; Huang, Y. AI-powered workflow for constructing organic materials databases from the literature: integrating large language models. *ACS. Omega.* **2025**, *10*, 49545–56. DOI PubMed PMC
86. Fu, Z.; Huang, P.; Wang, X.; et al. Artificial intelligence-assisted ultrafast high-throughput screening of high-entropy hydrogen evolution reaction catalysts. *Adv. Energy. Mater.* **2025**, *15*, 2500744. DOI
87. Zhou, J.; Wang, W.; Jiang, X.; Bai, B.; Su, Y. LLMs and NLP integration enables ion irradiation-induced swelling prediction in stainless steels. *J. Nucl. Mater.* **2026**, *619*, 156241. DOI
88. Pichlmann, L.; Raffiezadeh, S.; Hofbauer, M.; Ocansey, E. D.; Österreicher, J. A. Predicting mechanical properties in aluminum alloys: a data-driven framework leveraging LLM-based data extraction and physics-based feature engineering. *Mater. Today. Commun.* **2025**, *47*, 112843. DOI
89. Zhao, Z.; Hu, L.; Wang, H. Fine-tuned large language models for high-accuracy prediction of band gap and stability in transition metal sulfides. *Materials* **2025**, *18*, 3793. DOI PubMed PMC
90. Liu, J.; Anderson, H.; Waxman, N. I.; et al. Predicting materials thermodynamics enabled by large language model-driven dataset building and machine learning. *Adv. Intell. Syst.* **2026**, *8*, e202500857. DOI
91. Ozcan, A.; Coudert, F. X.; Rogge, S. M. J.; et al. Artificial intelligence paradigms for next-generation metal-organic framework research. *J. Am. Chem. Soc.* **2025**, *147*, 23367–80. DOI PubMed PMC
92. Mo, B.; Yu, K.; Kazdan, J.; et al. KGGen: extracting knowledge graphs from plain text with language models. *arXiv* **2025**, arXiv:2502.09956. Available online: <https://doi.org/10.48550/arXiv.2502.09956>. (accessed on 29 Jul 2026).
93. Radevski, G.; Gashteovski, K.; Hung, C. C.; Lawrence, C.; Glavaš, G. Linking surface facts to large-scale knowledge graphs. *arXiv* **2023**, arXiv:2310.14909. Available online: <https://doi.org/10.48550/arXiv.2310.14909>. (accessed on 29 Jul 2026).
94. Gridach, M.; Nanavati, J.; El Abidine, K. Z.; Mendes, L.; Mack, C. Agentic AI for scientific discovery: a survey of progress, challenges, and future directions. *arXiv* **2025**, arXiv:2503.08979. Available online: <https://doi.org/10.48550/arXiv.2503.08979>. (accessed on 29 Jul 2026).
95. Oliveira, O. N. Jr.; Christino, L.; Oliveira, M. C. F.; Paulovich, F. V. Artificial intelligence agents for materials sciences. *J. Chem. Inf. Model.* **2023**, *63*, 7605–9. DOI PubMed
96. Wei, J.; Yang, Y.; Zhang, X.; et al. From AI for science to agentic science: a survey on autonomous scientific discovery. *arXiv* **2025**, arXiv:2508.14111. Available online: <https://doi.org/10.48550/arXiv.2508.14111>. (accessed on 29 Jul 2026).
97. Zhang, H.; Li, Y.; Huang, W.; et al. Towards agentic intelligence for materials science. *arXiv* **2026**, arXiv:2602.00169. Available online: <https://doi.org/10.48550/arXiv.2602.00169>. (accessed on 29 Jul 2026).
98. Zhang, X.; Yuan, J.; Zhang, P.; Liu, J.; Li, L. Active knowledge structuring for large language models in materials science text mining. *Trans. Assoc. Comput. Linguist.* **2025**, *13*, 1186–203. DOI

-
99. Li, Z.; Zeng, Y.; Zuo, Y.; et al. KnowCoder: coding structured knowledge into LLMs for universal information extraction. *arXiv* **2024**, arXiv:2403.07969. Available online: <https://doi.org/10.48550/arXiv.2403.07969>. (accessed on 29 Jul 2026).
100. Li, B.; Conen, J.; Aller, F. AID-agent: an LLM-agent for advanced extraction and integration of documents. In *Proceedings of the 1st Workshop for Research on Agent Language Models (REALM 2025)*, Vienna, Austria, July, 2025. Association for Computational Linguistics; 2025. pp. 80-8. DOI
101. Hu, B.; Liu, S.; Ye, B.; et al. A multi-agent framework for physical laws discovery. *arXiv* **2024**, arXiv:2411.16416. Available online: <https://doi.org/10.48550/arXiv.2411.16416>. (accessed on 29 Jul 2026).
102. Ghosh, S.; Tewari, A. LLM-based AI agents for automated extraction of material properties and structural features. *Comput. Mater. Sci.* **2026**, *265*, 114521. DOI
103. Chakraborty, A.; Taskiran, N. P.; Kottooru, R.; Mann, V.; Venkatasubramanian, V. Building hybrid AI models in chemical engineering: a tutorial review. *Comput. Chem. Eng.* **2025**, *201*, 109236. DOI
104. Zhang, H.; Song, Y.; Hou, Z.; Miret, S.; Liu, B. HoneyComb: a flexible LLM-based agent system for materials science. *arXiv* **2024**, arXiv:2409.00135. Available online: <https://doi.org/10.48550/arXiv.2409.00135>. (accessed on 29 Jul 2026).
105. Jia, S.; Zhang, C.; Fung, V. LLMatDesign: autonomous materials discovery with large language models. *arXiv* **2024**, arXiv:2406.13163. Available online: <https://doi.org/10.48550/arXiv.2406.13163>. (accessed on 29 Jul 2026).
106. Kang, Y.; Kim, J. ChatMOF: an artificial intelligence system for predicting and generating metal-organic frameworks using large language models. *Nat. Commun.* **2024**, *15*, 4705. DOI PubMed PMC
107. Bazgir, A.; Madugula, R. C. P.; Zhang, Y. MatAgent: a human-in-the-loop multi-agent LLM framework for accelerating the material science discovery cycle. In *Proceedings of the AI for Accelerated Materials Design - ICLR 2025*, 2025. <https://openreview.net/pdf?id=2Nm6Ef4tZD>. (accessed on 29 Jul 2026).
108. Ni, Z.; Li, Y.; Hu, K.; et al. MatPilot: an LLM-enabled AI materials scientist under the framework of human-machine collaboration. *arXiv* **2024**, arXiv:2411.08063. Available online: <https://doi.org/10.48550/arXiv.2411.08063>. (accessed on 29 Jul 2026).
109. Huang, Z.; He, L.; Yang, Y.; et al. Application of machine reading comprehension techniques for named entity recognition in materials science. *J. Cheminform.* **2024**, *16*, 76. DOI PubMed PMC
110. Mysore, S.; Jensen, Z.; Kim, E.; et al. The materials science procedural text corpus: annotating materials synthesis procedures with shallow semantic structures. In *Proceedings of the 13th Linguistic Annotation Workshop*, Florence, Italy, August, 2019. Association for Computational Linguistics; 2019. pp. 56-64. DOI
111. Sriramanan, G.; Bharti, S.; Sadasivan, V. S.; Saha, S.; Kattakinda, P.; Feizi, S. LLM-check: investigating detection of hallucinations in large language models. In *Proceedings of the Advances in Neural Information Processing Systems*, Vancouver, Canada, Dec 10-15, 2024. Neural Information Processing Systems Foundation, Inc. (NeurIPS); 2024. pp. 34188-216. DOI
112. Takahara, W.; Yamaguchi, Y.; Ogano, M.; et al. Materials dual-source knowledge retrieval-augmented generation for local large language models in photocatalysts. *J. Chem. Inf. Model.* **2025**, *65*, 13098-114. DOI PubMed PMC
113. Müller, S. How to crack a SMILES: automatic crosschecked chemical structure resolution across multiple services using MoleculeResolver. *J. Cheminform.* **2025**, *17*, 117. DOI PubMed PMC
114. Norouzi, E.; Waitelonis, J.; Sack, H. The landscape of ontologies in materials science and engineering: a survey and evaluation. *arXiv* **2024**, arXiv:2408.06034. Available online: <https://doi.org/10.48550/arXiv.2408.06034>. (accessed on 29 Jul 2026).
115. Clark, S.; Bleken, F. L.; Stier, S.; et al. Toward a unified description of battery data. *Adv. Energy. Mater.* **2022**, *12*, 2102702. DOI
116. Bayerlein, B. PMDco - platform material digital core ontology. In *Proceedings of the FEMS EUROMAT 2023*, Frankfurt am Main, Germany, Sep 03-07, 2023. 2023. <https://opus4.kobv.de/opus4-bam/frontdoor/index/index/year/2023/docId/58197>. (accessed on 29 Jul 2026).
117. Rajamohan, B. P.; Bradley, A. C. H.; Tran, V. D.; et al. Materials data science ontology(MDS-Onto): unifying domain knowledge in materials and applied data science. *Sci. Data.* **2025**, *12*, 628. DOI PubMed PMC
118. Xu, Y.; Wu, Y. J.; Li, H.; et al. A comprehensive data network for data-driven study of battery materials. *Sci. Technol. Adv. Mater.* **2024**, *25*, 2403328. DOI PubMed PMC
119. Wu, F.; Zheng, X.; Qu, Y.; Wang, Z.; Feng, Z.; Li H. Grounding generative planners in verifiable logic: a hybrid architecture for trustworthy embodied AI. *arXiv* **2026**, arXiv:2602.08373. Available online: <https://doi.org/10.48550/arXiv.2602.08373>. (accessed on 29 Jul 2026).
120. Athar, S.; Mecibah, A.; Jund, P. Tackling dataset curation challenges towards reliable machine learning: a case study on thermoelectric materials. *Mater. Today. Phys.* **2025**, *59*, 101948. DOI
121. Huang, S.; Cole, J. M. A database of battery materials auto-generated using ChemDataExtractor. *Sci. Data.* **2020**, *7*, 260. DOI PubMed PMC
122. Schilling-Wilhelmi, M.; Ríos-García, M.; Shabih, S.; et al. From text to insight: large language models for chemical data extraction. *Chem. Soc. Rev.* **2025**, *54*, 1125-50. DOI

-
123. Trewartha, A.; Walker, N.; Huo, H.; et al. Quantifying the advantage of domain-specific pre-training on named entity recognition tasks in materials science. *Patterns* **2022**, *3*, 100488. DOI PubMed PMC
124. Rajan, K.; Zielesny, A.; Steinbeck, C. STOUT V2.0: SMILES to IUPAC name conversion using transformer models. *J. Cheminform.* **2024**, *16*, 146. DOI PubMed PMC
125. Hu, B.; Lin, A.; Brinson, L. C. ChemProps: a RESTful API enabled database for composite polymer name standardization. *J. Cheminform.* **2021**, *13*, 22. DOI PubMed PMC
126. Marshall, J.; Sanchez, S. L.; Desai, R.; et al. POLARIS: perovskite optimization using LLM-assisted refinement and intelligent screening. *Digit. Discov.* **2026**, *5*, 1765-82. DOI
127. Zhang, Y.; Chen, F.; Liu, Z.; et al. A materials terminology knowledge graph automatically constructed from text corpus. *Sci. Data.* **2024**, *11*, 600. DOI PubMed PMC
128. Na, Y.; Kim, J. J.; Park, C.; et al. Advanced scientific information mining using LLM-driven approaches in layered cathode materials for sodium-ion batteries. *Mater. Adv.* **2025**, *6*, 2543-8. DOI
129. Krotkov, N. A.; Sbytov, D. A.; Chakhoyan, A. A.; et al. Nanostructured material design via a retrieval-augmented generation (RAG) approach: bridging laboratory practice and scientific literature. *J. Chem. Inf. Model.* **2025**, *65*, 11064-78. DOI
130. Reed, S. M. Augmented and programmatically optimized LLM prompts reduce chemical hallucinations. *J. Chem. Inf. Model.* **2025**, *65*, 4274-80. DOI PubMed PMC
131. Cornelio, C.; Ito, T.; Cory-Wright, R.; Dash, S.; Horesh, L. The need for verification in AI-driven scientific discovery. *arXiv* **2025**, arXiv:2509.01398. Available online: <https://doi.org/10.48550/arXiv.2509.01398>. (accessed on 29 Jul 2026).
132. Liu, X.; Chen, T.; Da, L.; Chen, C.; Lin, Z.; Wei, H. Uncertainty quantification and confidence calibration in large language models: a survey. In *Proceedings of the Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, Toronto, Canada, 2025. Association for Computing Machinery; 2025. pp. 6107-17. DOI
133. Wang, X.; Raj, A.; Luebke, M.; Wen, H.; Xu, S.; Lu, K. Reliable end-to-end material information extraction from the literature with source-tracked multi-stage large language models. *arXiv* **2025**, arXiv:2510.05142. Available online: <https://doi.org/10.48550/arXiv.2510.05142>. (accessed on 29 Jul 2026).
134. Blagec, K.; Dorffner, G.; Moradi, M.; Ott, S.; Samwald, M. A global analysis of metrics used for measuring performance in natural language processing. In *Proceedings of NLP Power! The First Workshop on Efficient Benchmarking in NLP*, Dublin, Ireland, May, 2022. Association for Computational Linguistics; 2022. pp. 52-63. DOI
135. Lederbauer, M.; Betala, S.; Li, X.; et al. LeMat-Synth: a multi-modal toolbox to curate broad synthesis procedure databases from scientific literature. *arXiv* **2025**, arXiv:2510.26824. Available online: <https://doi.org/10.48550/arXiv.2510.26824>. (accessed on 29 Jul 2026).
136. Mullick, A.; Ghosh, A.; Chaitanya, G. S.; et al. MatSciRE: leveraging pointer networks to automate entity and relation extraction for material science knowledge-base construction. *Comput. Mater. Sci.* **2024**, *233*, 112659. DOI
137. Foppiano, L.; Dieb, S.; Suzuki, A.; et al. SuperMat: construction of a linked annotated dataset from superconductors-related publications. *Sci. Technol. Adv. Mater. Methods.* **2021**, *1*, 34-44. DOI
138. Song, Y.; Miret, S.; Liu, B. MatSci-NLP: evaluating scientific language models on materials science language tasks using text-to-schema modeling. *arXiv* **2023**, arXiv:2305.08264. Available online: <https://doi.org/10.48550/arXiv.2305.08264>. (accessed on 29 Jul 2026).
139. Niyongabo Rubungo, A.; Li, K.; Hattrick-Simpers, J.; Bouso Dieng, A. LLM4Mat-bench: benchmarking large language models for materials property prediction. *Mach. Learn. Sci. Technol.* **2025**, *6*, 020501. DOI
140. Zhong, Z.; Zhou, K.; Mottin, D. Benchmarking large language models for molecule prediction tasks. *arXiv* **2024**, arXiv:2403.05075. Available online: <https://doi.org/10.48550/arXiv.2403.05075>. (accessed on 29 Jul 2026).
141. Alampara, N.; Mandal, I.; Khetarpal, P.; et al. MacBench: a multimodal chemistry and materials science benchmark. In *Proceedings of the Proceedings of the 38th Conference on Neural Information Processing Systems (NeurIPS 2024)*, 2024. <https://neurips.cc/virtual/2024/103763>. (accessed on 29 Jul 2026).
142. Wang, H.; Fu, T.; Du, Y.; et al. Scientific discovery in the age of artificial intelligence. *Nature* **2023**, *620*, 47-60. DOI PubMed
143. Shumailov, I.; Shumaylov, Z.; Zhao, Y.; Papernot, N.; Anderson, R.; Gal, Y. AI models collapse when trained on recursively generated data. *Nature* **2024**, *631*, 755-9. DOI PubMed PMC
144. Canty, R. B.; Bennett, J. A.; Brown, K. A.; et al. Science acceleration and accessibility with self-driving labs. *Nat. Commun.* **2025**, *16*, 3856. DOI PubMed PMC
145. Miret, S.; Krishnan, N. M. A. Enabling large language models for real-world materials discovery. *Nat. Mach. Intell.* **2025**, *7*, 991-8. DOI
146. Alampara, N.; Aneesh, A.; Rios-García, M.; et al. General-purpose models for the chemical sciences: LLMs and beyond. *Chem. Rev.* **2026**, *126*, 2484-549. DOI PubMed PMC

Disclaimer/Publisher's Note: All statements, opinions, and data contained in this publication are solely those of the individual author(s) and contributor(s) and do not necessarily reflect those of OAE and/or the editor(s). OAE and/or the editor(s) disclaim any responsibility for harm to persons or property resulting from the use of any ideas, methods, instructions, or products mentioned in the content.



© The Author(s) 2026. Open Access This article is licensed under a Creative Commons Attribution 4.0 International License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, sharing, adaptation, distribution and reproduction in any medium or format, for any purpose, even commercially, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.