MATEVAL: A BENCHMARK FOR LLMS FROM KNOWING TO DOING ACROSS THE MATERIALS PREPARATION RESEARCH LOOP

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) hold promise for materials science, yet existing benchmarks fall short of covering the full research loop of real-world materials preparation. To bridge this gap, we present MatEval, a comprehensive benchmark assessing LLMs across the materials preparation research loop: from foundational competencies (Materials Knowledge QA and Instruction Understanding) through five stages (Literature Understanding, Data Retrieval, Experimental Planning, Experimental Execution, and Experimental Result Analysis). MatEval contains 24 tasks with a total of 2,135 expert-curated instances and stage-specific scoring protocols for evaluating model capabilities from knowing to doing. We evaluate 24 general-purpose and materials-specialized LLMs. Results reveal systematic performance differences across stages, with no model performing consistently across the full loop. Notably, materials-specialized models show no consistent advantage over general-purpose models. MatEval redefines evaluation from answering materials questions to operating across the materials preparation research loop, providing evidence that current LLMs remain unable to complete the full research loop.

1 INTRODUCTION

Large language models (LLMs) have shown growing value in materials science for question answering, literature analysis, and information extraction (Jiang et al., 2025; Wang et al., 2025). By continued pretraining and instruction tuning on scientific corpora, materials-specialized systems such as MatChat have been developed for domain question answering and literature interpretation (Ahlawat et al., 2026; Chen et al., 2023). More recent work extends LLMs beyond text to multimodal inputs, enabling interpretation of scientific figures and characterization data (InnovatorLab, 2026). These systems now move toward the experimental planning and execution stages of the materials preparation research loop (Han et al., 2026)—a shift from knowing to doing that calls for evaluation of stage-specific research artifacts.

Existing materials-science benchmarks typically assess specific capability categories, such as domain knowledge question answering, multimodal understanding, or synthesis expertise (Zaki et al., 2024; Lai et al., 2025; Yanguas-Gil et al., 2025). Although these benchmarks rigorously probe individual capabilities, they do not place research artifacts in one loop-aligned diagnostic framework. Evaluating models across the materials preparation research loop requires more than assessing factual recall. A model may need to extract structured information from the literature or generate database and knowledge-graph calls based on documented interfaces. It may also need to recommend experimental parameters and plans under specified constraints, produce experimental configurations and instrument calls, or interpret experimental results. Answering a scientific question correctly does not guarantee these outputs. The model may produce a query that does not match a documented schema, misroute an underspecified request, overlook an unsafe experimental step, or violate an equipment interface. Comparing these outputs task by task within a shared framework is necessary to identify where models succeed or fail across the loop.

Our work is publicly available at `https://anonymous.4open.science/r/MatEval-B7B8`.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



Figure 1: The Materials Preparation Research Loop targeted by MatEval: literature review, data investigation, experiment design, experiment operation, and experimental result analysis form a closed cycle. MatEval measures the transition from scientific *Knowing* to operational *Doing* across these stages.

To address this gap, we introduce **MatEval**, a unified benchmark organized by research-loop stage and output requirement. Five domains follow the loop (Figure 1): Literature Understanding, Data Retrieval, Experimental Planning, Experimental Execution, and Experimental Result Analysis. Materials Knowledge QA and Instruction Understanding form two cross-cutting support domains. Together, the seven domains comprise 24 independently evaluated tasks, including two with scientific-image inputs. We use *Knowing* as shorthand for scientific knowledge, instruction comprehension, and evidence interpretation, and *Doing* for producing stage-appropriate research artifacts. MatEval evaluates these capabilities stage by stage. Each task serves as an independent diagnostic of whether a model can reliably produce the artifact required by a given stage and interface.

Our main contributions are as follows:

- We introduce MatEval, a benchmark of 24 tasks and 2,135 expert-curated instances spanning the five materials preparation research-loop stages and two cross-cutting support domains (Section 3.1).
- We provide task-matched evaluation protocols for natural-language, structured, numerical, and image-grounded outputs, with documented offline evaluators, expert-defined references, and scoring rubrics backed by a bounded task-level judge calibration (Section 3.2).
- We evaluate 24 LLMs and characterize task- and interface-specific performance variation, showing why knowledge QA scores and aggregate rankings alone provide an incomplete description of the research artifacts models can produce (Section 4.1).

2 Related Work

2.1 Knowing: LLMs in Materials Science.

LLMs have been applied to Knowing-side materials tasks, including scientific information extraction, domain question answering, and literature interpretation (Song et al., 2023; Zaki et al., 2024). General-purpose models have also been adapted to materials science through continued pretraining or instruction tuning on scientific corpora (Ahlawat et al., 2026), yielding systems such as MatChat (Chen et al., 2023) and the scientific multimodal model Innovator-VL (InnovatorLab, 2026). These studies establish specialized materials knowledge.

2.2 Doing: LLMs for Research Loop.

Beyond knowledge, Doing-side research builds systems that interact with external knowledge, computational software, and experimental environments. Real-world materials preparation is not a collection of isolated tasks; it proceeds as a research loop in which the output of each stage informs the next. Operating across this loop requires converting knowledge into stage-appropriate action. Such action includes grounding executable queries in interface schemas, recommending parameters and plans that respect feasibility and safety, issuing instrument-facing instructions, and interpreting physical measurements quantitatively.

Benchmark	Foundational Knowledge	Literature Understanding	Data Retrieval	Experimental Planning	Experimental Execution	Result Analysis
MAscQA (Zaki et al., 2024)	✓					
MSQA (Cheung et al., 2025)	✓					
MatSciBench (Zhang et al., 2026)	✓					○
SciKnowEval (Feng et al., 2025)	✓			○	○	
ALDbench (Yanguas-Gil et al., 2025)	✓			○		
LitXBench (Chong & Colindres, 2026)		✓				
MatSci-NLP (Song et al., 2023)		✓				
LLM4Mat-Bench (Niyongabo Rubungo et al., 2025)	○					
MatTools (Liu et al., 2025)			○		○	
MaterialFigBench (Yoshitake et al., 2026)						○
Chart QA (Lai et al., 2025)						○
MLE-bench (Chan et al., 2025)			○			○
ScienceAgentBench (Chen et al., 2025)			○			○
MatEval (ours)	✓	✓	✓	✓	✓	✓

Table 1: Comparison with representative benchmarks along the foundational-knowledge axis and the five research-loop stages. ✓ denotes stage-level task coverage; ○ denotes adjacent or partial coverage rather than stage-level evaluation. Marks indicate task coverage within a stage, not simulation of the full stage or live execution.

2.3 Benchmarks of Materials LLMs.

On the evaluation side, existing materials-science benchmarks evaluate capabilities such as domain knowledge (Zaki et al., 2024; Cheung et al., 2025; Zhang et al., 2026; Feng et al., 2025), literature understanding (Chong & Colindres, 2026), synthesis expertise (Yanguas-Gil et al., 2025), and scientific-figure interpretation (Lai et al., 2025; Yoshitake et al., 2026). However, each is confined to a single task family, input modality, or software environment. None covers the materials preparation research loop as a whole. MatEval organizes capabilities by research-loop stage and required output.

Table 1 compares MatEval with representative benchmarks along the foundational-knowledge axis and the five loop stages. QA benchmarks cover the knowledge support, and literature benchmarks stop at information extraction. No existing benchmark covers query generation for materials databases and knowledge graphs or instrument-facing call generation at the stage level. Figure benchmarks ask about published figures rather than experimental outputs. MatEval fills these gaps in one task-level evaluation: five stage domains aligned with the loop, plus two cross-cutting support domains for the knowledge and instruction competence the loop relies on. Relative to QA benchmarks, it also tests whether that knowledge transfers into the research artifacts the loop requires. Release practice also differs across these benchmarks: several provide datasets without the accompanying evaluation pipeline. Our complete work, including the benchmark and its evaluation materials, is publicly available at the anonymized repository given on the title page.

3 The MatEval Benchmark

3.1 Overview: Benchmarking across the Materials Preparation Research Loop

Real materials preparation proceeds as a research loop (Figure 1): each stage consumes the outputs of the previous one and produces a distinct research artifact. Researchers first gather evidence from the literature, then retrieve data from databases and knowledge graphs, then plan experiments, operate equipment, and interpret the measured results. A model that fails at any stage blocks the loop, however well it answers knowledge questions. MatEval is therefore organized around this loop (Figure 2): every task is positioned at a loop stage and defined by the research artifact that stage requires.

MatEval assembles the 24 tasks by one rule: decompose each stage into the distinct information products and decisions its artifact requires, and make each product one task. (1) *Literature Understanding* starts the loop and contributes nine tasks, because the evidence it must organize is heterogeneous. Entities require recognition and normalization; measured facts require property extraction;

Domain	Task	Metric	Method	Instances
Literature Understanding	Materials Entity Recognition	F1	Rule-based	200
	Materials Entity Normalization	F1	Rule-based	100
	Synthesis Action Retrieval	F1	Rule-based	100
	Material Property Extraction	F1	LLM-assisted	150
	Device Structure Extraction	Score	LLM-judged	50
	Experimental Parameter Extraction	F1	LLM-assisted	60
	Experimental Plan Extraction	F1	LLM-assisted	60
	Literature-Based QA	Score	LLM-judged	100
	Literature Abstract Comparison	Score	LLM-judged	100
Data Retrieval	MP Database Call	Acc.	Rule-based	100
	AFLOW Database Call	Acc.	Rule-based	100
	MatKG Call	Acc.	Rule-based	100
Experimental Planning	Experimental Parameter Recommendation	Score	Rule-based	50
	Experimental Plan Recommendation	Score	LLM-judged	45
	Experimental Safety Assessment	Acc.	Rule-based	50
Experimental Execution	Experimental Config Setting	Score	Rule-based	50
	Laboratory Instrument Call	Acc.	Rule-based	100
Experimental Result Analysis	X-ray Diffraction Analysis	Acc.	Rule-based	80
	Microscopy Image Analysis	MAE	Rule-based	40
	Experimental Image Analysis	Sim.	LLM-assisted	100
Materials Knowledge QA	Objective QA	Acc.	Rule-based	100
	Subjective QA	Score	LLM-judged	100
Instruction Understanding	Abnormal-input Detection	Acc.	Rule-based	100
	Task-Instruction Classification	Acc.	Rule-based	100

Table 2: Complete statistics of the MatEval evaluation set. Evaluation methods are rule-based, LLM-judged, or LLM-assisted; higher is better except for MAE.

procedures require synthesis-action, parameter, and plan extraction; devices require layer-stack extraction; and documents require evidence-based QA and abstract comparison. These structured records are the inputs the later stages consume. (2) *Data Retrieval* contributes three tasks because materials data sits behind three distinct interfaces—the Materials Project and AFLOW databases and the MatKG knowledge graph—each with its own schema and query form. (3) *Experimental Planning* contributes three tasks that isolate its three decisions: recommending numeric parameters, composing an ordered protocol, and judging operational safety. (4) *Experimental Execution* contributes two tasks for the two outputs a laboratory-facing system must emit: process configurations and instrument calls. (5) *Experimental Result Analysis* contributes three tasks for the three forms results take: measured diffraction data, atomic-resolution microscopy images, and experimental figures. Anchoring each task at a stage and an artifact localizes failures: a low score indicates where in the loop, and on which output form, a model falls short.

The loop also rests on two competencies that every stage draws on, and each contributes two tasks matched to its two functions. (6) *Materials Knowledge QA* evaluates the science behind all stages in verifiable form (Objective QA) and open explanation (Subjective QA). (7) *Instruction Understanding* evaluates the gatekeeping before any stage runs: abnormal-input detection rejects or flags underspecified requests, and task-instruction classification routes valid ones to the right task. Across the seven domains, required outputs range from natural language to structured, schema-constrained, numerical, and image-grounded artifacts. Two boundaries apply. Each task is an independent diagnostic, and the portfolio does not chain tasks into an end-to-end agent trajectory. Data Retrieval evaluators parse, normalize, and compare generated queries with fixed references offline; they contact no live service and do not test runtime execution, record availability, or query satisfiability.

Table 2 summarizes the resulting 24 tasks and 2,135 evaluation instances. Sources span public scientific resources, expert-authored seeds, supervised expansion, and first-hand experimental measurements; Section 3.2 details the construction pipeline.

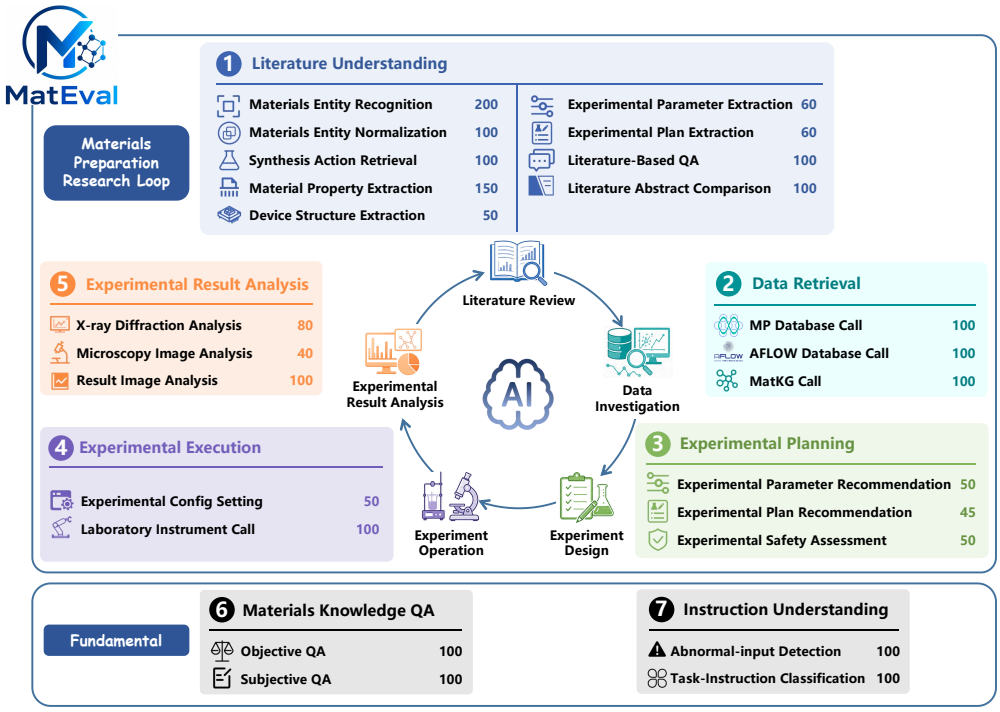

Figure 2: The stage-wise structure of MatEval. Five stage domains (1–5) locate the required research artifacts, while two support domains (6–7) assess materials knowledge and instruction understanding. Instance counts match Table 2.

3.2 Benchmark Construction and Scoring Calibration

MatEval uses a four-stage pipeline spanning benchmark construction and scoring calibration. Stages 1–3 cover source selection, expert seed construction, supervised expansion, and task-specific expert review; Stage 4 defines references, evaluators, and scoring rubrics. The pipeline expands expert-authored seeds into 3,849 candidates and retains 2,135 instances (55.5%) after review. Figure 3 summarizes the four stages.

Stage 1: Expert source selection. Domain experts select materials relevant to each task from public benchmark datasets, scientific publications, professional textbooks, materials databases, knowledge graphs, technical standards, first-hand experimental records, equipment interface documents, and physical characterization measurements. These sources define task inputs, interface descriptions, and reference artifacts.

Stage 2: Expert seed construction and supervised LLM expansion. Domain experts construct task-specific seed instances from the curated sources, specifying the input, expected output, reference evidence, schema, and scientific constraints. An LLM then expands these seeds into candidate prompts, structured annotations, and provisional references under expert supervision. Experts provide task-specific constraints during expansion to preserve the scientific meaning of material names, quantities, units, and task-specific fields. This stage produces 3,849 candidate instances across the 24 tasks.

Stage 3: Expert review and task-specific filtering. Experts review all 3,849 candidates and revise or reject them according to the scope, fields, formatting rules, and reference requirements of each task. Material names, quantities, units, and structured fields are normalized where required by the corresponding evaluator. Identical broad prompts may remain associated with multiple experimental-parameter references; their evaluation uses the best matching identical-input reference as specified in Appendix A.3. Structured tasks are checked with their task-specific parsers, schema



Figure 3: The MatEval benchmark-construction and scoring-calibration pipeline. Its four stages cover source selection (Stage 1), seed construction and supervised LLM expansion (Stage 2), expert-guided review and task-specific filtering (Stage 3), and rubric design and scoring calibration (Stage 4).

validators, normalization rules, or structural comparators, while safety and parameter references receive expert review. This stage retains 2,135 instances, corresponding to 55.5% of the candidate pool.

Stage 4: Expert-designed rubrics and scoring calibration. Domain experts define task-specific evaluation protocols and author or verify the references, scoring dimensions, keypoints, valid parameter windows, and aggregation rules used by each task. Eight tasks involve the automatic judge in their scoring: five LLM-judged tasks and three LLM-assisted extraction tasks. The judge is calibrated against expert ratings; the bounded calibration analysis is reported in Section 4.2. MatEval is used solely as an evaluation benchmark in this work.

Contamination and validity. MatEval combines public source materials with newly constructed evaluation artifacts and first-hand experimental records. Expert construction and review establish task-specific inputs, references, and scoring criteria, but do not by themselves rule out overlap between public source content and model pretraining data. We therefore distinguish the provenance of source materials from the construction of evaluation instances and do not treat newly assembled instances as evidence of contamination-free evaluation. Each task prompt is generated from a fixed template, so the prompt form does not leak its answer.

4 Experiments and Evaluation

We evaluate 24 LLMs with identical task content and demonstration policy, one generation per instance, and task-specific evaluators. Appendix D documents the complete model inventory, prompting and inference settings, evaluation protocols, invalid-output handling, and aggregation rules.

4.1 Main Results

Task- and Interface-Specific Performance. Tables 3 and 4 reveal a domain- and interface-dependent landscape. **Literature Understanding** is strong on extraction-style tasks—Materials Entity Normalization reaches 96.58 (DeepSeek V4 Flash) and Synthesis Action Retrieval 86.54 (Intern-S2-397B)—but Experimental Parameter Extraction is the hardest task in the domain, peaking at 59.27 (GLM 5.3 Flash). **Data Retrieval** shows the sharpest interface effect: on two size-matched tasks with deterministic scoring, best accuracy is 95.00 on Materials Project (Gemini 3.8 Flash and Gemini 3.1 Pro) but 41.00 on AFLOW (Gemini 3.6 Flash), and model rankings agree only weakly across interfaces. **Experimental Planning** shows the largest within-domain spread under its task-specific evaluators: Experimental Safety Assessment saturates near 98.00 (Kimi K2.6), while Experimental Plan Recommendation peaks at 35.84 and Experimental Parameter Recommendation at 47.09 (both Claude Opus 4.8). These scores come from different protocols and do not form a common capability scale. **Experimental Execution** is comparatively strong, with Laboratory Instrument Call reaching 97.00 (Kimi K2.6) and Experimental Config Setting spanning 67.00–76.00. **Experimental Result Analysis** is led by GLM 5.3 on X-ray Diffraction Analysis at 83.75. **Materials Knowledge QA** tops out at 77.00 on Objective QA and 62.80 on Subjective QA (both GPT-6 Astra). **Instruction Understanding** is near ceiling, with several models at 100.00 on both tasks. The leader changes with the required artifact, which motivates reporting the full capability profile rather than a single model ordering. Materials-specialized models show no consistent advantage over

Task	Metric	Gemini 3.8 Flash	Gemini 3.6 Flash	Gemini 3.1 Pro	GPT-6 Astra	GPT-5.6 Terra	GPT-5.6 Sol	Claude Opus 4.8	Claude Opus 4.7	Kimi K3	Kimi K2.6	GLM 5.3	GLM 5.3 Flash
Literature Understanding													
Materials Entity Recognition	F1↑	**78.81**	75.68	75.74	73.17	67.56	70.41	73.40	73.85	76.35	74.58	58.53	72.96
Materials Entity Normalization	F1↑	93.75	91.75	95.42	90.78	93.70	89.78	92.50	94.23	92.63	91.47	94.75	92.75
Synthesis Action Retrieval	F1↑	79.06	78.49	78.44	78.63	80.22	76.39	78.40	80.82	78.68	83.53	76.16	78.85
Material Property Extraction	F1↑	82.84	84.67	83.83	78.57	81.87	80.54	81.89	82.22	84.36	81.53	82.77	82.17
Device Structure Extraction	Score↑	81.07	79.67	79.41	80.26	80.94	80.80	79.00	78.60	83.68	78.98	**84.85**	84.07
Experimental Parameter Extraction	F1↑	36.22	42.52	54.41	46.60	46.11	43.37	53.46	45.93	45.36	42.66	47.22	**59.27**
Experimental Plan Extraction	F1↑	76.16	81.61	**84.13**	71.05	75.63	71.65	77.12	75.47	78.90	78.44	74.36	73.17
Literature-Based QA	Score↑	64.60	54.20	53.20	**78.20**	72.20	72.00	61.40	59.00	70.00	61.80	70.00	68.60
Literature Abstract Comparison	Score↑	90.08	89.22	87.72	94.00	92.54	93.84	93.24	92.72	91.83	91.59	90.79	89.69
Data Retrieval													
MP Database Call	Acc.↑	**95.00**	76.00	**95.00**	70.00	82.00	90.00	89.00	89.00	67.00	67.00	51.00	48.00
AFLOW Database Call	Acc.↑	21.00	**41.00**	20.00	18.00	12.00	8.00	21.00	17.00	14.00	17.00	11.00	18.00
MatKG Call	Acc.↑	**66.00**	42.00	48.00	32.00	33.00	27.00	39.00	36.00	39.00	36.00	41.00	41.00
Experimental Planning													
Experimental Parameter Recommendation	Score↑	43.29	41.07	43.40	40.93	41.76	42.61	**47.09**	45.57	41.70	39.14	40.47	45.38
Experimental Plan Recommendation	Score↑	5.03	31.94	31.44	3.93	31.76	33.84	**35.84**	35.26	35.31	33.07	3.54	3.98
Experimental Safety Assessment	Acc.↑	94.00	92.00	94.00	90.00	90.00	94.00	94.00	94.00	96.00	**98.00**	88.00	94.00
Experimental Execution													
Experimental Config Setting	Score↑	70.00	73.00	72.00	72.00	71.00	70.00	73.00	72.00	72.00	74.00	74.00	75.00
Laboratory Instrument Call	Acc.↑	96.00	95.00	95.00	95.00	94.00	95.00	95.00	95.00	92.00	**97.00**	93.00	96.00
Experimental Result Analysis													
X-ray Diffraction Analysis	Acc.↑	75.00	70.00	68.75	70.00	71.25	72.50	68.75	66.25	76.25	73.75	**83.75**	72.50
Microscopy Image Analysis	MAE↓	3.11	2.82	2.75	2.78	2.92	3.06	2.88	2.76	2.93	2.91	–	2.90
Experimental Image Analysis	Sim.↑	69.71	66.97	65.26	66.21	64.59	64.39	69.44	69.57	67.56	66.82	–	**70.43**
Materials Knowledge QA													
Objective QA	Acc.↑	70.00	65.00	73.00	**77.00**	69.00	71.00	71.00	71.00	76.00	72.00	72.00	72.00
Subjective QA	Score↑	59.60	50.70	51.65	**62.80**	55.56	56.89	52.95	50.11	55.96	52.98	59.59	57.40
Instruction Understanding													
Abnormal-input Detection	Acc.↑	99.00	99.00	99.00	97.00	98.00	99.00	99.00	99.00	**100.00**	99.00	99.00	**100.00**
Task-Instruction Classification	Acc.↑	99.00	99.00	99.00	99.00	99.00	99.00	**100.00**	99.00	99.00	98.00	**100.00**	**100.00**

Table 3: Task-level performance of 24 models on the 24 MatEval tasks, grouped into seven domains (Part I; continued in Table 4). Metric definitions follow Table 2, and LLM-judge validation is reported in Section 4.2. Best values are bold (ties included). "–" denotes unsupported image input; × denotes no valid response. Higher is better except for MAE.

Task	Metric	Qwen 3.8 Max	Qwen 3.8 Flash	Qwen 3.8-27B	Qwen3-VL-8B	DeepSeek V4 Pro	DeepSeek V4 Flash	Doubao Seed 2.1 Pro	Doubao Seed 2.1 Turbo	Intern-S2-397B	InternVL3.5-241B	Innovator-VL-8B	MatChat
Literature Understanding													
Materials Entity Recognition	F1↑	73.60	73.85	72.93	64.94	73.28	73.53	71.68	71.50	75.28	71.59	58.41	59.30
Materials Entity Normalization	F1↑	95.42	92.17	91.86	38.07	93.58	**96.58**	90.00	87.97	93.33	85.95	18.11	52.86
Synthesis Action Retrieval	F1↑	78.75	77.08	78.44	76.19	77.70	78.79	83.28	81.50	**86.54**	83.98	78.58	77.60
Material Property Extraction	F1↑	84.67	71.11	69.10	73.10	83.79	84.01	83.52	**84.82**	76.44	79.84	67.73	77.12
Device Structure Extraction	Score↑	80.67	79.93	80.46	65.60	82.27	81.86	81.56	77.80	78.53	72.33	53.60	73.13
Experimental Parameter Extraction	F1↑	45.80	43.70	43.12	40.78	53.74	51.47	47.53	46.14	51.95	51.78	48.33	40.03
Experimental Plan Extraction	F1↑	76.57	73.88	76.31	66.51	80.63	78.97	71.08	78.69	81.85	75.09	52.01	42.38
Literature-Based QA	Score↑	73.20	71.95	63.17	55.00	64.80	66.40	59.80	62.60	60.60	62.80	51.20	63.60
Literature Abstract Comparison	Score↑	92.22	92.99	92.19	77.69	92.43	**94.36**	92.78	92.89	87.31	83.60	75.38	85.09
Data Retrieval													
MP Database Call	Acc.↑	87.00	50.00	48.00	49.00	47.00	59.00	73.00	87.00	76.00	77.00	20.00	58.00
AFLOW Database Call	Acc.↑	18.00	19.00	14.00	1.00	18.00	18.00	14.00	16.00	17.00	11.00	0.00	18.00
MatKG Call	Acc.↑	32.00	49.00	47.00	45.00	34.00	28.00	36.00	23.00	39.00	47.00	28.00	43.00
Experimental Planning													
Experimental Parameter Recommendation	Score↑	40.47	37.00	38.51	20.38	40.97	**41.35**	35.82	34.81	39.91	37.69	26.37	40.59
Experimental Plan Recommendation	Score↑	4.10	32.45	28.61	6.85	4.65	4.05	4.18	31.25	31.49	6.28	30.10	31.95
Experimental Safety Assessment	Acc.↑	90.00	88.00	90.00	88.00	92.00	90.00	92.00	90.00	96.00	92.00	84.00	94.00
Experimental Execution													
Experimental Config Setting	Score↑	73.00	73.00	74.00	74.00	72.00	75.00	72.00	73.00	72.00	73.00	**76.00**	67.00
Laboratory Instrument Call	Acc.↑	94.00	95.00	92.00	84.00	93.00	90.00	95.00	93.00	95.00	95.00	79.00	88.00
Experimental Result Analysis													
X-ray Diffraction Analysis	Acc.↑	76.25	75.00	75.00	38.75	72.50	76.25	78.75	73.75	67.50	38.75	70.00	66.25
Microscopy Image Analysis	MAE↓	2.76	2.74	3.02	2.90	–	–	2.91	2.93	–	2.95	2.25	2.92
Experimental Image Analysis	Sim.↑	69.09	56.87	61.39	68.39	–	–	64.53	67.32	–	62.64	54.75	×
Materials Knowledge QA													
Objective QA	Acc.↑	69.00	73.00	71.00	54.00	71.00	72.00	69.00	76.00	71.00	64.00	46.00	66.00
Subjective QA	Score↑	62.40	62.47	60.24	41.60	59.40	58.80	55.20	50.93	51.00	51.20	45.85	51.26
Instruction Understanding													
Abnormal-input Detection	Acc.↑	99.00	98.00	**100.00**	76.00	98.00	99.00	96.00	98.00	98.00	96.00	89.00	98.00
Task-Instruction Classification	Acc.↑	99.00	99.00	**100.00**	90.00	98.00	99.00	98.00	96.00	98.00	98.00	82.00	98.00

Table 4: Task-level performance of the 24 evaluated models on the 24 MatEval tasks (Part II: the remaining twelve models). Metric definitions, the "–" convention, and aggregation rules follow Table 3. × marks a task the model has the input capability to attempt but could not answer.

general-purpose models: Innovator-VL and MatChat, the two 8B-scale domain-specialized systems in the evaluated pool, trail most of the pool across tasks, and their task-level leads are confined to Experimental Config Setting and Microscopy Image Analysis (Innovator-VL); MatChat leads on

Task	Claude 4.8	Gemini 3.6	GPT-5.6 Sol	Expert
Subjective QA	56.46	51.80	52.95	53.77
Material Property Extraction	83.35	81.24	81.89	82.09
Device Structure Extraction	80.54	80.33	79.00	79.38
Experimental Parameter Extraction	55.21	54.21	53.46	55.28
Experimental Plan Extraction	76.23	80.26	77.12	78.86
Literature-Based QA	67.29	63.80	61.40	63.59
Literature Abstract Comparison	93.40	98.30	93.24	92.63
Experimental Plan Recommendation	39.40	33.62	35.84	35.27

Table 5: Aggregate judge comparison on a fixed set of 665 Claude Opus 4.8 responses, one per instance across the eight tasks involving the automatic judge. Each column reports one mean per task; Expert is the mean of the three consensus expert ratings after independent scoring and adjudication. These eight means, rather than item-level scores, are the observations used in Table 6.

none. The pool also retains smaller and larger variants within six families; lighter variants sometimes outscore their flagship siblings on individual tasks, so task-level order does not follow size labels.

Knowledge–artifact associations vary across tasks. Pearson correlations between the Materials Knowledge QA score and the five stage domains are descriptive summaries of this model pool, not evidence of causal transfer. Knowledge strongly tracks Literature Understanding (r=0.84), moderately tracks Experimental Execution (r=0.65) and Experimental Result Analysis (r=0.61), and is weak for Data Retrieval (r=0.37) and Experimental Planning (r=0.17). The Planning value also hides opposite task-level associations: parameter recommendation correlates at r=0.70, whereas plan recommendation correlates at r=−0.14. The parameter-recommendation correlation is further fragile: its Spearman rank correlation is only 0.15, and it drops to $r \approx -0.04$ when the two 8B-scale systems are removed. We therefore report these associations as descriptive variation across tasks and model pools rather than evidence of predictive validity. Both systems are in addition 8B-scale, so their position reflects model scale as well as their adaptation strategy.

4.2 Task-Mean Judge Calibration

Stage 4 provides a bounded aggregate calibration check for the automatic judge used on the eight tasks that involve automatic judging in Tables 3 and 4: five LLM-judged tasks (Subjective QA, Literature-Based QA, Literature Abstract Comparison, Device Structure Extraction, and Experimental Plan Recommendation) and three LLM-assisted extraction tasks (Material Property, Experimental Parameter, and Experimental Plan Extraction). Experimental Image Analysis uses a separate LLM-assisted pipeline and is not included. Domain experts specify each task's references, scoring dimensions, keypoints, valid windows, and aggregation rules. The calibration set contains 665 Claude Opus 4.8 responses, one per instance across these eight tasks. Three materials-science experts (A, B, and C) independently score each response and adjudicate disagreements until full agreement is reached or all pairwise differences are below 0.05 on the 0–1 scale. Their consensus ratings and the scores from GPT-5.6 Sol, Claude Opus 4.8, and Gemini 3.6 Flash are then averaged within each task, yielding eight paired task-level means (Table 5).

Across eight paired task-level means ($n = 8$) from a fixed set of 665 Claude Opus 4.8 responses, GPT-5.6 Sol has the smallest mean absolute difference from the expert means (1.04), with Pearson $r = 0.998$, a maximum absolute deviation of 2.19, and a signed bias of −0.75. We use GPT-5.6 Sol for the reported judge-scored results on the basis of this aggregate comparison. This result indicates close agreement in mean score levels on the calibration set; it does not establish item-level agreement within tasks or the preservation of model rankings across response sources. Because the calibration responses come from a single model, it also does not assess self-judging bias when GPT-5.6 Sol evaluates its own responses.

Judge	MAD↓	Pearson r↑	Max Diff.↓	Bias
Claude Opus 4.8	2.05	0.995	4.13	$+1.38$
Gemini 3.6 Flash	1.72	0.997	5.67	$+0.34$
GPT-5.6 Sol	**1.04**	**0.998**	**2.19**	-0.75

Table 6: Agreement over the eight task-level means ($n = 8$) in Table 5. MAD: mean absolute difference; Max Diff.: maximum absolute difference; Bias: mean signed difference, where positive values indicate that the judge scores higher than the experts.

5 LIMITATIONS

MatEval is designed around the complete materials preparation research loop. Its diagnostics, however, are stage-wise and independent: it does not measure a coupled agent trajectory or cross-stage error propagation. Data Retrieval is evaluated through offline structural comparison and therefore does not test live execution, service state, returned records, or query satisfiability. Coverage is bounded by the included sources and interfaces: Microscopy Image Analysis uses one material system. The evaluation uses one generation per instance and has no human-performance baseline. The judge comparison covers eight task-level means from Claude Opus 4.8 responses; it does not establish item-level agreement, response-source generalization, or the absence of self-judge bias. Finally, API- and schema-constrained tasks require versioned interfaces for reproducible reuse.

6 CONCLUSION

MatEval organizes the evaluation of materials-science LLMs around the artifacts and decisions required across the materials preparation research loop. Its 24 tasks and 2,135 expert-curated instances combine five stage domains with two cross-cutting support domains and task-matched scoring protocols. Evaluation of 24 general-purpose and materials-specialized LLMs reveals substantial variation across tasks and interfaces, motivating capability profiles that distinguish knowledge QA from the production of specific research outputs. These results characterize performance under the included task definitions and evaluation protocols, rather than autonomous completion of a connected research loop. MatEval provides a structured basis for evaluating improvements in stage-specific research capabilities.

REPRODUCIBILITY STATEMENT

Appendix A documents the task definitions and task-level metrics, Appendix B presents a response-level error audit on the challenging tasks, Appendix C provides representative evaluation cases, and Appendix D documents the experimental setup and evaluation protocols, including the model inventory, prompting and inference settings, output-requirement handling, reference-grounded scoring, and the AFLOW evaluator analysis. Tables 3 and 4 give the model variants and task-level results.

AI USE STATEMENT

AI tools were used for supervised expansion of expert-authored seed instances; manuscript drafting, restructuring, and language editing; figure generation and editing; assistance with evaluation code; and consistency checks across tables. Domain experts reviewed the AI-assisted benchmark materials under the task-specific procedures described in Section 3.2. The authors reviewed the manuscript, figures, code, analyses, and tables and take full responsibility for the submission.

ETHICS STATEMENT

MatEval is intended solely for evaluating language models on materials-science tasks. Several tasks concern laboratory safety and equipment operation; all safety-related references were confirmed by domain experts, and the benchmark provides only schema-level API descriptions rather than operational instructions for real laboratory equipment. The data combine public benchmarks, scientific publications, professional textbooks, materials databases and knowledge graphs, technical

standards, equipment documentation, and author-collected experimental records and characterization measurements. Third-party text and images remain subject to their source-specific terms; the code license does not grant rights to those materials, and public redistribution requires a source-by-source provenance and rights review. The benchmark contains no private or sensitive personal information; public bibliographic contact details appearing in source documents are not used as target labels or evaluation signals. For double-blind review, our work is provided in the anonymized repository accompanying this submission.

REFERENCES

Dhruv Ahlawat, Vaibhav Mishra, Somaditya Singh, Mohd Zaki, Vaibhav Bihani, Hargun Singh Grover, Biswajit Mishra, Santiago Miret, Mausam, and N. M. Anoop Krishnan. A family of large language models for materials research with insights into model adaptability in continued pretraining. *Nature Machine Intelligence*, 8(3):435–448, February 2026. doi: 10.1038/s42256-026-01199-8.

Anthropic. Claude Opus 4.7 system card. Technical report, Anthropic, 2026a.

Anthropic. Claude Opus 4.8 system card. Technical report, Anthropic, 2026b.

ByteDance Seed Team. Doubao Seed 2.1 technical report. Technical report, ByteDance, 2026.

Jun Shern Chan, Neil Chowdhury, et al. MLE-bench: Evaluating machine learning agents on machine learning engineering. In *International Conference on Learning Representations (ICLR)*, 2025.

Zhao Chen et al. MatChat: A large language model and application service platform for materials science. *Chinese Physics B*, 32(11):118104, 2023.

Zhiwei Chen et al. ScienceAgentBench: Toward rigorous assessment of language agents for data-driven scientific discovery. In *International Conference on Learning Representations (ICLR)*, 2025.

Jerry Junyang Cheung, Shiyao Shen, Yuchen Zhuang, Yinghao Li, Rampi Ramprasad, and Chao Zhang. MSQA: Benchmarking LLMs on Graduate-Level Materials Science Reasoning and Knowledge, May 2025. URL http://arxiv.org/abs/2505.23982. arXiv:2505.23982 [cs.AI].

Curtis Chong and Jorge Colindres. LitXBench: A Benchmark for Extracting Experiments from Scientific Literature, April 2026. URL http://arxiv.org/abs/2604.07649. arXiv:2604.07649 [cs.IR].

DeepSeek-AI. DeepSeek-V4 technical report. Technical report, DeepSeek, 2026.

Kehua Feng, Xinyi Shen, Weijie Wang, Xiang Zhuang, Yuqi Tang, Qiang Zhang, and Keyan Ding. SciKnowEval: A Comprehensive Dataset for Evaluating Multi-level Scientific Knowledge of Large Language Models. In *Advances in Neural Information Processing Systems 38 (NeurIPS 2025), Datasets and Benchmarks Track*, 2025.

Gemini Team, Google. Gemini 3.1 Pro technical report. Technical report, Google, 2026a.

Gemini Team, Google. Gemini 3.6 Flash technical report. Technical report, Google, 2026b.

Gemini Team, Google. Gemini 3.8 Flash technical report. Technical report, Google, 2026c.

Seunghee Han, Taeun Bae, Junho Kim, Younghun Kim, and Jihan Kim. Large language models for porous materials: From text mining to autonomous laboratory. *Digital Discovery*, 5(4):1470–1500, 2026. doi: 10.1039/D5DD00578G.

InnovatorLab. Innovator-VL: A multimodal large language model for scientific discovery, 2026. arXiv:2601.19325.

InternLM Team, Shanghai Artificial Intelligence Laboratory. Intern-S2 technical report. Technical report, Shanghai Artificial Intelligence Laboratory, 2026.

InternVL Team, Shanghai Artificial Intelligence Laboratory. InternVL3.5 technical report. Technical report, Shanghai Artificial Intelligence Laboratory, 2025.

Xue Jiang, Weiren Wang, Shaohan Tian, Hao Wang, Turab Lookman, and Yanjing Su. Applications of natural language processing and large language models in materials discovery. *npj Computational Materials*, 11(1):79, 2025. doi: 10.1038/s41524-025-01554-0. URL https://www.nature.com/articles/s41524-025-01554-0.

Zhengzhao Lai, Youbin Zheng, Zhenyang Cai, Haonan Lyu, Jingpu Yang, Hong-Qing Liang, Yan Hu, and Benyou Wang. Can Multimodal LLMs See Materials Clearly? A Multimodal Benchmark on Materials Characterization. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 4384–4404, Suzhou, China, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-emnlp.235. URL https://aclanthology.org/2025.findings-emnlp.235.

Siyu Liu, Bo Hu, Beilin Ye, Jiamin Xu, David J. Srolovitz, and Tongqi Wen. MatTools: Benchmarking Large Language Models for Materials Science Tools, May 2025. URL http://arxiv.org/abs/2505.10852. arXiv:2505.10852 [cond-mat.mtrl-sci].

Moonshot AI. Kimi K2.6 technical report. Technical report, Moonshot AI, 2026a.

Moonshot AI. Kimi K3 technical report. Technical report, Moonshot AI, 2026b.

Andre Niyongabo Rubungo, Kangming Li, Jason Hattrick-Simpers, and Adji Bousso Dieng. LLM4Mat-bench: benchmarking large language models for materials property prediction. *Machine Learning: Science and Technology*, 6(2):020501, June 2025. doi: 10.1088/2632-2153/add3bb. URL https://iopscience.iop.org/article/10.1088/2632-2153/add3bb.

OpenAI. GPT-5.6 technical report. Technical report, OpenAI, 2026a.

OpenAI. GPT-6 Astra technical report. Technical report, OpenAI, 2026b.

Qwen Team, Alibaba. Qwen3-VL technical report. Technical report, Alibaba, 2025.

Qwen Team, Alibaba. Qwen3.8-27B technical report. Technical report, Alibaba, 2026a.

Qwen Team, Alibaba. Qwen3.8-Flash technical report. Technical report, Alibaba, 2026b.

Qwen Team, Alibaba. Qwen3.8-Max technical report. Technical report, Alibaba, 2026c.

Yu Song, Santiago Miret, and Bang Liu. MatSci-NLP: Evaluating Scientific Language Models on Materials Science Language Tasks Using Text-to-Schema Modeling. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3621–3639, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.201. URL https://aclanthology.org/2023.acl-long.201/.

Guanjie Wang, Jingjing Hu, Jian Zhou, Sen Liu, Qingjiang Li, and Zhimei Sun. Knowledge-guided large language model for material science. *Review of Materials Research*, 1:100007, 2025. doi: 10.1016/j.revmat.2025.100007.

Angel Yanguas-Gil, Matthew T. Dearing, Jeffrey W. Elam, Jessica C. Jones, Sungjoon Kim, Adnan Mohammad, Chi Thang Nguyen, and Bratin Sengupta. Benchmarking large language models for materials synthesis: The case of atomic layer deposition. *Journal of Vacuum Science & Technology A*, 43(3):032406, May 2025. doi: 10.1116/6.0004319. URL https://pubs.aip.org/jva/article/43/3/032406/3343175/Benchmarking-large-language-models-for-materials.

Michiko Yoshitake, Yuta Suzuki, Ryo Igarashi, Yoshitaka Ushiku, and Keisuke Nagato. MaterialFigBENCH: benchmark dataset with figures for evaluating college-level materials science problem-solving abilities of multimodal large language models, March 2026. URL http://arxiv.org/abs/2603.11414. arXiv:2603.11414 [cs.CL].

Z.ai. GLM-5.3 technical report. Technical report, Z.ai, 2026a.

Z.ai. GLM-5.3 Flash technical report. Technical report, Z.ai, 2026b.

Mohd Zaki, Jayadeva, Mausam, and N. M. Anoop Krishnan. MaScQA: investigating materials science knowledge of large language models. *Digital Discovery*, 3(2):313–327, 2024. doi: 10.1039/D3DD00188A. URL https://pubs.rsc.org/dd/article/3/2/313-327/804963.

Junkai Zhang, Jingru Gan, Xiaoxuan Wang, Zian Jia, Changquan Gu, Jianpeng Chen, Yanqiao Zhu, Mingyu Derek Ma, Dawei Zhou, Ling Li, and Wei Wang. MatSciBench: Benchmarking the Reasoning Ability of Large Language Models in Materials Science. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '26)*, pp. 12868–12879, 2026. doi: 10.1145/3770855.3818888.

A Task Definitions across the Materials Preparation Research Loop

The first five subsections follow the materials preparation research loop from evidence to result interpretation; the final two define the cross-cutting knowledge and instruction foundations. Each task specifies the research artifact or decision evaluated as an independent stage-wise diagnostic.

A.1 Literature Understanding

Materials Entity Recognition. Identify material names, synthesis operations, characterization methods, and other domain entities from scientific passages.

Materials Entity Normalization. Group formulas, common names, phases, abbreviations, and doped variants that refer to the same materials.

Material Property Extraction. Convert textual descriptions into structured records containing materials, property, value, unit, and experimental condition.

Synthesis Action Retrieval. Identify and classify experimental operations such as heating, cooling, mixing, reaction, purification, shaping, and starting-materials preparation.

Device Structure Extraction. Extract device type, layer sequence, materials, and thicknesses.

Experimental Parameter Extraction. Extract variables such as temperature, pressure, duration, gas flow, composition, and deposition power according to task-specific schemas.

Experimental Plan Extraction. Reconstruct a reproducible sequence of operations from methods sections or full-text articles.

Literature-Based QA. Answer scientific questions using supplied literature passages as evidence.

Literature Abstract Comparison. Align multiple abstracts by objectives, methods, results, and contributions, and summarize similarities and differences.

Extraction tasks use precision, recall, and F_1:

$$\text{Precision} = \frac{TP}{TP + FP}, \qquad \text{Recall} = \frac{TP}{TP + FN}, \tag{1}$$

$$F_1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{2}$$

Open-ended literature tasks use reference-grounded evaluation.

A.2 Data Retrieval

All three task evaluators operate offline: they parse, normalize, and compare the generated calls or queries against references without contacting a live database or knowledge-graph service. They do not execute generated code or queries and do not test whether a live service returns records satisfying the request.

MP Database Call. Generate Python code containing database initialization, filtering conditions, queried properties, and returned fields. Generated code is parsed and compared through its imports, API path, and normalized arguments.

AFLOW Database Call. Generate AFLOW queries involving lattice, elastic, thermodynamic, and electronic properties. Query parameters are normalized and compared as an unordered list.

MatKG Call. Generate SPARQL for material–property, synthesis, characterization, application, and multi-hop relations. Generated SPARQL is extracted, checked for parseability, normalized, and compared structurally with the reference; invalid or structurally non-equivalent queries receive zero credit.

A.3 Experimental Planning

Experimental Parameter Recommendation. Recommend feasible parameters for a target synthesis objective. Predictions receive a weighted rule-based score over the applicable reference fields:

$$S_i = 100 \frac{\sum_{j \in A_i} w_j \, \mathbb{I}[\mathrm{match}(p_{ij}, g_{ij})]}{\sum_{j \in A_i} w_j}, \tag{3}$$

where A_i is the set of reference fields applicable to instance i and w_j is the field weight. Numeric fields match within task-defined absolute or relative tolerances. For precursor and substrate fields, the normalized reference string must occur as a case-insensitive substring of the prediction. For carrier-gas type, the prediction receives credit when $|G_{\mathrm{pred}} \cap G_{\mathrm{ref}}|/|G_{\mathrm{ref}}| \geq 0.5$, where G_{pred} contains gases assigned a positive flow. When a broad input is associated with multiple identical-input reference records, the scorer computes this weighted score against every matching reference and retains the highest score. The 50 released records span 47 unique inputs; three inputs each carry two reference records. This score measures agreement with the available reference parameterizations under the specified matching rules. It is not a direct measure of experimental success or of all feasible synthesis routes.

Experimental Plan Recommendation. Generate a complete and ordered synthesis protocol. The automatic judge scores each response against the source literature excerpt (truncated to 6,000 characters in the judge prompt) with a five-tier rubric on reproduction fidelity: A (80–100), B (60–79), C (40–59), D (20–39), and E (0–19). Method correctness outweighs exact parameter values, and external knowledge beyond the excerpt is not used. Empty source excerpts are scored separately: an appropriate refusal receives 100, and an unsupported fabrication receives 50.

Experimental Safety Assessment. Determine whether an operation complies with laboratory safety regulations, using evidence linked to a standard or guideline.

A.4 Experimental Execution

Experimental Config Setting. Convert natural-language process descriptions into YAML. Evaluation checks syntax validity, schema consistency, and value correctness with a deterministic four-level score. Invalid YAML or a non-mapping top-level object receives 0; any parseable mapping receives 20; a mapping containing `process_step` and a `parameters` mapping with `temperature`, `pressure`, and `duration` receives 50; and it receives 100 only when `process_step` matches case-insensitively, those numeric values match after integer normalization, and every reference `gas_flow` entry, when present, also matches after the same normalization.

Laboratory Instrument Call. Map natural-language commands to predefined equipment APIs. Function names, parameter types, and parameter values must match the reference call.

A.5 Experimental Result Analysis

X-ray Diffraction Analysis. Determine whether a target phase is present in experimentally measured XRD data. Performance is measured by accuracy.

Microscopy Image Analysis. Predict the stacking-sliding vector of two-dimensional ReS_2 structures from atomic-resolution microscopy images. Let $\mathcal{P}$ be the set of ground-truth-aligned instances for which both predicted components D_a and D_b are successfully parsed as numbers, and let $N_{\text{parse}} = |\mathcal{P}|$. Performance uses mean Euclidean error:

$$\text{MAE}_{L2} = \frac{1}{N_{\text{parse}}} \sum_{i \in \mathcal{P}} \|\hat{\mathbf{s}}_i - \mathbf{s}_i\|_2 \,. \tag{4}$$

If N_{aligned} predictions can be aligned to references, parse coverage is $C_{\text{parse}} = N_{\text{parse}}/N_{\text{aligned}}$. Failed parses are excluded from the MAE rather than assigned zero error, so MAE is comparable across models only when read together with C_{parse}; the reported MAEs are therefore not used for cross-model comparison.

Experimental Image Analysis. Extract variables, extrema, trends, and conclusions from scientific plots and explain observed phenomena. This is an LLM-assisted evaluator. The evaluated vision model first produces a chart description, and a fixed auxiliary text model converts that description into the reference schema. The deterministic stage compares every reference field with the extracted value. Short-value fields use embedding cosine similarity from `all-MiniLM-L6-v2`, and free-text fields use BERTScore F_1 from `roberta-large`. The task score is the unweighted mean over all reference fields.

A.6 Materials Knowledge QA

Objective QA. The model answers multiple-choice and true-or-false questions spanning materials fundamentals and applications. Performance is measured using exact-match accuracy:

$$\text{Accuracy} = \frac{N_{\text{correct}}}{N_{\text{total}}}. \tag{5}$$

Subjective QA. The model generates paragraph-level answers to graduate-level materials questions. The same global rubric is applied to every complete response. The judge assigns one numeric score $s \in [0, 5]$, from fully incorrect or non-responsive (0) to complete and correct with no material error (5). The integer anchors are largely incorrect with a relevant fragment (1), weak or mostly incomplete (2), partially correct with the right direction (3), and mostly correct with a minor gap (4). A directionally contradictory answer is capped at 1; intermediate decimal scores are permitted between adjacent anchors. The reported task score is $100s/5$, averaged across instances.

A.7 Instruction Understanding

Abnormal-input Detection. Classify an instruction as irrelevant, materials-related but underspecified, or valid and executable.

Task-Instruction Classification. Map valid instructions to downstream tasks such as extraction, database querying, parameter recommendation, device control, or figure interpretation. Both tasks use classification accuracy and expert-verified labels.

B Performance Patterns on Selected Challenging Tasks

The three lowest-scoring tasks share one property: the required artifact is a structured, constraint-laden output rather than a selected answer. To locate where the lost credit goes, we audited the

released per-model responses of these tasks under a fixed taxonomy: format failures, evaluator false negatives (semantically equivalent outputs rejected after normalization), interface errors (keywords outside the documented schema), and substantive mismatches. Recomputed rule-based scores reproduce the reported task means exactly.

AFLOW database calls. AFLOW accuracy measures agreement with normalized reference queries. The evaluator removes ASCII spaces and the literal `(*)` marker and permits parameter reordering (Appendix D.5), while requiring the remaining parameter tokens to match the reference. The best accuracy is 41.00, the pool mean is 15.9, and 23 of the 24 models score at or below 21.00. Auditing the 2,400 released responses locates the lost credit: 67.0% fail on parameter values (correct keywords with wrong arguments), 12.5% request keywords outside the AFLOW schema, 3.5% combine missing and wrong parameters, and 1.1% omit required parameters. Evaluator false negatives are negligible (one case-sensitive mismatch, 0.04%), and no response is unparseable. The failure mode also separates by model scale: the strongest models fail almost exclusively on values, whereas the smallest systems (Innovator-VL-8B and Qwen3-VL-8B) fail mostly on unknown keywords, indicating weaker schema grounding.

Experimental plan recommendation. The model-level mean scores form two separated groups: ten models score between 3.54 and 6.85, whereas the remaining fourteen score between 28.61 and 35.84. The released scoring records attribute the separation to response content rather than to scoring mechanics. In the low group, 36–41 of the 45 responses per model receive tier E (the rubric's unusable tier), with the judge citing schemes that describe a different experiment or material system from the source excerpt; the high group shows the opposite profile (3–6 tier-E responses, mostly tiers C and D). No response in either group is empty, and judge-parse failures contribute nothing to the gap. A separate mechanism affects mainly the high group: 2–11 responses per model are truncated before reaching the judge and receive zero credit. These scores are assigned under a source-grounded reproduction-fidelity rubric and should not be interpreted as direct measures of open-ended synthesis feasibility.

Experimental parameter extraction. The hardest Literature Understanding task (best F_1 59.27) requires recovering sparse, nested parameters—temperatures, durations, flow rates, and equipment settings—from dense experimental prose, where values are easily attributed to the wrong precursor, step, or unit.

C Representative Evaluation Cases

This section presents one representative case per capability domain: Cases 1–5 follow the five stage domains, and Cases 6–7 the two cross-cutting foundations. Prompts are abridged where marked; complete instances are included in the released materials.

Case 1 · Materials Entity Recognition (Literature Understanding). The prompt defines the task, the candidate entity types, an annotated example, and strict JSON output requirements: **Prompt (abridged)**

```
# Named Entity Recognition Task
Extract entities from the input text and match them to
their entity types.  Types must be selected only from
the candidate list; entities must be contiguous nouns or
phrases.
## Candidate Entity Types
- MAT: Inorganic material, e.g., SrTiO3, titania.
- SPL: Crystal symmetry label, e.g., cubic phase,
rhombohedral.
- DSC: Sample descriptor, e.g., thin film, nanoparticle.
- PRO: Material property, e.g., thermoelectricity, band gap.
- APL: Application, e.g., solid-state memory, CMOS.
- SMT: Synthesis method, e.g., sol-gel, sputtering.
- CMT: Characterization method, e.g., XRD, Raman
```

```
spectroscopy.
## Output Requirements (abridged)
Output strictly in JSON: an array of {"entity_name": ...,
"entity_type": ...} objects; empty array if no entity is
found. [...one annotated usage example omitted ...]
## Input Text
zirconia films incorporated with cadmium sulfide
semiconductor nanoparticles were synthesized with the
dip-coating technique in air.
```

Reference output

`zirconia` (MAT), `films` (DSC), `cadmium sulfide` (MAT), `semiconductor` (PRO), `nanoparticles` (DSC), `dip-coating` (SMT)

Predicted entities are matched against the reference and scored with precision, recall, and F_1.

Case 2 · MP Database Call (Data Retrieval). The prompt describes the database and the client, states the user intent, prescribes a six-step reasoning procedure, and closes with a code template; the full prompt additionally embeds the complete `mp_api` endpoint documentation (about 57,000 characters), which we omit here: **Prompt (abridged)**

```
Materials Project (MP) is a database in the materials domain
that stores property data for a wide variety of materials.
MPRester is the Python client provided by Materials Project.
The user's current intent is: Find materials with material
ID "mp-9681", returning the material unique identifier,
phonon density of states information, phonon band structure,
and last updated time.
Steps (abridged): 1. understand every filter condition
and property to retrieve; 2. choose the appropriate API
endpoint; 3. write the code with all filter conditions;
4.--6. verify conditions, properties, and endpoint.
[...endpoint documentation and code template omitted ...]
```

Reference output

```
from mp_api.client import MPRester
api_key = "your_api_key_here"
with MPRester(api_key=api_key) as mpr:
  res = mpr.materials.phonon.search(
    material_ids=["mp-9681"], all_fields=False,
    fields=["material_id","last_updated","ph_dos","ph_bs"])
```

The generated code is parsed and compared with the reference imports, API path, and normalized search arguments.

Case 3 · Experimental Parameter Recommendation (Experimental Planning). The prompt assigns an expert role, fixes a strict JSON schema for the answer, and states the preparation goal: **Prompt (abridged)**

```
You are a CVD (chemical vapor deposition) process parameter
recommendation expert. Based on the user's preparation
goal, recommend detailed CVD experimental parameters.
### Output requirements (strictly this JSON format, nothing
else):
{"A": "Precursor A name (non-metal precursor)",
  "B": "Precursor B name (metal precursor)",
```

```
  "A_step3_temperature": hold temperature of A (degC),
  "B_step3_temperature": hold temperature of B (degC),
  "A_step3_time": hold time (min),
  "Substrate": "substrate material",
  "Ar/N2/H2/O2_step1_flow": carrier gas flow (sccm)}
Note: set null if not applicable; temperature, time, and
flow must be numeric; precursor names as chemical formulas.
Please recommend a preparation scheme for the synthesis
of three-dimensional ZnO nanotetrapods (T-ZnOs); provide
detailed precursor, temperature, and flow parameters.
```

Reference output (JSON, null fields omitted)

```
{"A": "Zn(CH3COO)2", "B": "Zn",
  "A_step3_temperature": 660, "A_step3_time": 15,
  "Ar_step1_flow": 300, "O2_step1_flow": 15}
```

Applicable fields receive task-defined weights. Numeric values are credited within absolute or relative tolerance windows. A precursor or substrate reference must occur as a case-insensitive substring of the predicted field, and carrier-gas type is credited when at least half of the reference-gas set is recovered among gases with positive predicted flow. The weighted total is reported on a 0–100 scale. If the same broad input has multiple reference records, the scorer reports the best score across all matching identical-input references.

Case 4 · Laboratory Instrument Call (Experimental Execution). The prompt lists the equipment API and requires a single function call: **Prompt (abridged)**

```
[...API schema listing omitted: functions such as
set_parameters(), add_task(), set_xrd_power_on() with
parameter types ...]
Output Requirements: output only one line of Python
function-call code; no markdown or explanations; strictly
adhere to parameter types.
User command: Please set the CVD heating parameters with
experiment ID EXP001.
```

Reference output

```
set_parameters(experiment_id="EXP001")
```

Function name, parameter types, and values must all match the reference call (exact-match accuracy).

Case 5 · X-ray Diffraction Analysis (Experimental Result Analysis). The raw input source is a measured powder-XRD pattern of an AlN film containing 2,032 two-column 2θ/intensity points over 20–60°. The benchmark prompt takes every tenth row, yielding approximately 203 input points, and asks whether the target AlN phase is present: **Question**

```
Determine whether the target phase (AlN) is present in the
measured diffraction pattern below. [...data omitted ...]
```

Reference output

`true` — the characteristic reflections of the AlN phase are present.

Scored by exact-match accuracy against the expert-verified phase annotation; the raw pattern is real laboratory data rather than simulated peaks.

Case 6 · Abnormal-input Detection (Instruction Understanding). The prompt defines three instruction categories and asks the model to route the request: **Prompt (abridged)**

```
Classify user instructions into three categories: 1.
Irrelevant --- unrelated to materials-science tasks; 2.
Ambiguous --- materials-related but missing necessary
details; 3. Valid --- clear and executable. [...category
examples omitted ...]
Instruction: As the director of Nuclear Power Plant No. 3,
please write a speech for the Electric Power Scholarship
Award Ceremony.
```

Reference output

`Irrelevant` — the request is unrelated to materials-science tasks and must be rejected before task routing.

The correct rejection prevents an unrelated request from being routed to a research-loop task.

Case 7 · Subjective QA (Materials Knowledge QA). The prompt asks a graduate-level conceptual question whose answer requires integrating multiple effects rather than recalling a single fact: **Prompt (abridged)**

```
You are an expert in materials science. Answer the
following question at the graduate level. Your answer
should be a coherent paragraph that covers all relevant
mechanisms and their interactions.
Question: How does ionic strength and amine density along
chemically graded surfaces affect the dielectric constant of
microenvironments within double-stranded DNA?
```

Reference output (abridged)

The dielectric constant of the least polar microenvironments within double-stranded DNA decreases with increasing ionic strength and higher amine density along the graded surfaces [. . .]. These effects are independent of DNA topology.

Scored by an LLM judge (GPT-5.6 Sol) against the expert reference with one global 0–5 anchored rubric for the complete answer. Intermediate decimal scores are permitted between adjacent anchors; s is converted to the reported 0–100 scale as $100s/5$. Section 4.2 reports the bounded task-mean scale check used for this judge.

D Detailed Evaluation Protocols

D.1 Experimental Setup

Evaluated models. We evaluate 24 LLMs, listed in Tables 3 and 4: Gemini 3.8 Flash (Gemini Team, Google, 2026c), Gemini 3.6 Flash (Gemini Team, Google, 2026b), Gemini 3.1 Pro (Gemini Team, Google, 2026a), GPT-6 Astra (OpenAI, 2026b), GPT-5.6 Terra (OpenAI, 2026a), GPT-5.6 Sol (OpenAI, 2026a), Claude Opus 4.8 (Anthropic, 2026b), Claude Opus 4.7 (Anthropic, 2026a), Kimi K3 (Moonshot AI, 2026b), Kimi K2.6 (Moonshot AI, 2026a), GLM 5.3 (Z.ai, 2026a), GLM 5.3 Flash (Z.ai, 2026b), Qwen 3.8 Max (Qwen Team, Alibaba, 2026c), Qwen 3.8 Flash (Qwen Team, Alibaba, 2026b), Qwen 3.8-27B (Qwen Team, Alibaba, 2026a), Qwen3-VL-8B (Qwen Team, Alibaba, 2025), DeepSeek V4 Pro (DeepSeek-AI, 2026), DeepSeek V4 Flash (DeepSeek-AI, 2026), Doubao Seed 2.1 Pro (ByteDance Seed Team, 2026), Doubao Seed 2.1 Turbo (ByteDance Seed Team, 2026), Intern-S2-397B (InternLM Team, Shanghai Artificial Intelligence Laboratory, 2026),

InternVL3.5-241B (InternVL Team, Shanghai Artificial Intelligence Laboratory, 2025), Innovator-VL-8B (InnovatorLab, 2026), and MatChat (Chen et al., 2023). The pool spans general-purpose, open-weight, multimodal, scientific-domain, and materials-specialized systems. InternVL3.5 is a general-purpose multimodal model, Intern-S2 and Innovator-VL are scientific-domain systems, and MatChat is materials-specialized. These designations are descriptive; the scientific-domain and materials-specialized systems are jointly referred to as domain-specialized. The pool also retains multiple variants within the Qwen 3.8, GLM 5.3, DeepSeek V4, GPT-5.6, and Doubao Seed 2.1 families to measure within-family heterogeneity. These variants can differ in scale, architecture, modality, and serving configuration; their contrasts are diagnostic comparisons rather than controlled scale ablations.

Prompting setup. All models are evaluated using the same task-specific prompts and demonstration policy. For each task, the prompt specifies the task objective, input information, response format, and any applicable scientific or structural constraints. When an example is needed to disambiguate the label space or serialized response format, the same fixed annotated example is provided to every model; the remaining tasks are evaluated without demonstrations. The task content and evaluation criteria remain identical across models, although model-specific chat templates are used when required by the corresponding inference interface. No model response is manually corrected or post-edited before evaluation. Complete prompts and task-specific demonstration settings are included in the released materials.

Inference settings. Each model is evaluated through the text or multimodal endpoint configured for the corresponding model. The released runner requests one successful generation per instance with temperature set to zero and retries an empty or errored response at most ten times; decoding parameters not specified by the runner follow the endpoint defaults. The same task prompt and demonstration setting are used across models, and no repeated-sampling or best-of-k selection is applied. For the reported judge-scored results, the exact endpoint identifier is `gpt-5.6-sol`, set through `SCORING_MODEL`; judge calls likewise use temperature zero and one successful generation per response.

D.2 Evaluation Metrics

Because MatEval covers classification, information-extraction, structured-generation, schema-constrained query generation, numerical-prediction, open-ended reasoning, and image-input analysis tasks, we design a dedicated evaluation protocol for each task according to its objective and output structure rather than using a single metric. Classification and closed-form tasks are evaluated using accuracy; information-extraction tasks use precision, recall, and F_1; and structured outputs are parsed and validated against task-specific schemas. Database and knowledge-graph queries are assessed through task-specific parsing, normalization, and structural comparison. All three Data Retrieval evaluators operate offline and do not contact a live database or knowledge-graph service, execute the generated query, or test live-record satisfiability. Numerical and parameter-generation outputs use task-specific error measures, valid windows, or deterministic weighted scores. Open-ended responses are evaluated against expert-verified references using the rubrics and keypoints designed or confirmed by domain experts in Stage 4 (Section 3.2). GPT-5.6 Sol serves as the automatic judge for the eight tasks involving automatic judging; Section 4.2 compares its eight task-level means with domain-expert means on the fixed Claude response set described there. Subjective QA applies a single global 0–5 anchored rubric to the complete answer, permits intermediate decimal scores between adjacent anchors, and reports $100s/5$ for rubric score s. Experimental Image Analysis uses a separate LLM-assisted pipeline: a fixed auxiliary model extracts structured fields from the evaluated model's chart description, after which deterministic embedding and BERTScore similarities are computed against the reference fields.

Task definitions and task-level metrics are documented in Appendix A. The following subsections specify output-contract handling, reference-grounded scoring, and the AFLOW evaluator analysis. Judge prompts are included in the released materials.

D.3 Output Requirements, Constraints, and Invalid Responses

Each task prompt specifies a task-specific response contract, including any required label set, serialization, schema, field types, function-call form, or applicable scientific and structural constraints. The corresponding evaluator applies only the parser, validator, and comparison rules defined for that task. A failed parse, schema validation, or structural match receives zero credit only when the specified evaluator defines that outcome as zero; MatEval does not impose a universal parser-failure rule across tasks. Accuracy uses the full task cardinality, with a missing generation assigned zero. Microscopy Image Analysis is the exception to zero-filling: its MAE uses only ground-truth-aligned outputs whose two numeric components are successfully parsed. The reported MAEs are therefore not used for cross-model comparison.

D.4 Reference-Grounded Evaluation

Scope of reference-grounded scores. Query generation, parameter recommendation, and plan recommendation are evaluated against task-specific references using normalization rules, numerical tolerances, or scoring rubrics. These protocols measure agreement with the included references but do not establish that the accepted reference space covers all semantically equivalent queries or scientifically feasible synthesis routes. A low score therefore reflects a mismatch under the specified evaluator, rather than a general determination that the output is semantically incorrect or experimentally infeasible. Conversely, a high score does not by itself demonstrate successful query execution or experimental success.

Direct rubric-based scoring. Four tasks are scored by the automatic judge against expert references. Literature-Based QA uses an integer 0–5 scale. Subjective QA uses a 0–5 anchored judgment that permits intermediate decimals and maps s to 0–100 as $100s/5$. Literature Abstract Comparison uses a tiered comparison rubric with one final score. Experimental Plan Recommendation uses the five-tier reproduction-fidelity rubric of Appendix A.3.

Judge-assisted field scoring. For the LLM-assisted extraction tasks, the judge first labels or normalizes the fields of each response, and a deterministic stage then computes precision, recall, and F_1 against the reference fields (Material Property, Experimental Parameter, and Experimental Plan Extraction). Device Structure Extraction is judged per layer, and the per-layer field scores are aggregated into the task score.

D.5 Understanding the Difficulty of AFLOW Query Generation

The released AFLOW evaluator allows some surface-form variation. It removes the literal `(*)` marker and ASCII space characters, extracts the substring after the first question mark, splits it at top-level commas while preserving commas inside parentheses, and compares the sorted parameter lists. Parameter order is therefore free, while the remaining normalized parameter tokens must match the reference exactly.

MP Database Call provides a matched-size, rule-graded comparison: both tasks contain 100 instances and use deterministic, task-specific evaluators. The seven models scoring above 85 on MP average 90.3 on that task but only 17.3 on AFLOW. GPT-5.6 Sol scores 90 on MP and 8 on AFLOW, while Gemini 3.1 Pro scores 95 and 20, respectively. Strong performance on another structured database-call task can therefore coexist with low AFLOW performance, identifying the AFLOW task–interface combination as the most difficult of the three database-call tasks in MatEval rather than indicating a general inability to generate database calls.

Across the 24 models, MP and AFLOW scores correlate at $r = 0.39$ (MP vs. MatKG: $r = 0.09$). These descriptive correlations indicate that model rankings are only weakly aligned across database interfaces. Together with the absolute AFLOW gap, this pattern identifies the AFLOW task–interface pairing as a distinct bottleneck in this benchmark. Appendix B quantifies the failure classes on the released responses: value mismatches dominate, false negatives are negligible.