MATMICROBENCH: TOWARDS REALISTIC AGENTIC MATERIALS MICROSTRUCTURE CHARACTERIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Real-world microstructural characterization requires a continuous analysis process that *observes* relevant scientific evidence, *models* analysis rules as executable procedures, *measures* quantitative characteristics, and *reasons* from the resulting evidence and measurements to scientific conclusions. Existing benchmarks, however, largely focus on knowledge-centric or single-step tasks and provide limited coverage of such end-to-end analysis in **realistic experimental settings**. We introduce MatMicroBench, a benchmark toward realistic agentic materials microstructural characterization, where agents operate over experimentally acquired data and complete long-horizon scientific analysis with multiple interdependent outputs. Built from controlled laboratory experiments, MatMicroBench contains 64 instances based on SEM and EBSD data and spans four complementary tasks: static defect-state analysis, temporal defect-evolution analysis, inter-grain structure analysis, and intra-grain orientation-variation analysis. Each task requires agents to jointly produce spatial evidence, executable code, numerical results, and scientific conclusions, enabling **capability-resolved evaluation** of *Observe*, *Model*, *Measure*, and *Reason*. We evaluate 15 frontier models under a fixed execution harness and further compare six representative models across multiple agent harnesses. Under the fixed harness, even the strongest models achieve only around 53 points overall, and different models lead on different tasks, indicating substantial room for improving end-to-end microstructural analysis. Capability analysis further reveals task-specific bottlenecks, while strong code execution does not necessarily translate into accurate scientific outputs. These findings expose distinct failure modes across the characterization process and establish MatMicroBench as a testbed for developing more reliable agents for real-world materials analysis.

1 INTRODUCTION

Microstructural characterization is fundamental to materials research, providing evidence that links microstructure to material behavior (Chan et al., 2020). In practice, it involves a continuous workflow of identifying relevant evidence, conducting quantitative and computational analyses, and integrating the results into scientific conclusions (DeCost et al., 2019). Recent advances in multimodal models and agents offer new opportunities for scientific data analysis and tool use (M. Bran et al., 2024; Chen et al., 2026), motivating rigorous evaluation of their performance on end-to-end characterization workflows in realistic materials research settings (Chen et al., 2025).

Existing materials benchmarks cover property prediction (Dunn et al., 2020), multimodal understanding (Alampara et al., 2025; Lai et al., 2025), microscopy automation (Mandal et al., 2025), and computational reproducibility (Huang et al., 2026). However, two gaps remain for microstructural characterization (Figure 1). First, benchmark datasets provide limited coverage of directly acquired microstructural observations and their experimental context. Second, few benchmarks assess both end-to-end performance and the capabilities required at each stage.

We introduce MatMicroBench, a benchmark toward realistic agentic **mat**erials **micro**structural characterization, where agents operate on experimentally acquired data and complete open-ended, multi-step analysis workflows in real-world settings beyond predefined question-answer formats. We construct the benchmark in four stages. First, we define four core capabilities based on real microstructural analysis workflows: *Observe* for identifying and localizing relevant evidence, *Model*

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.

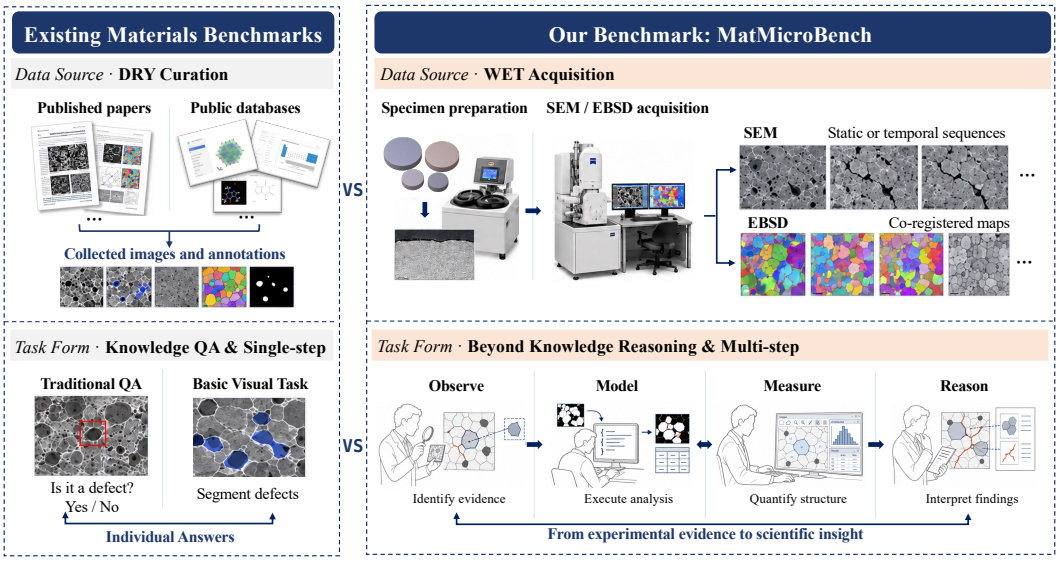


Figure 1: **Comparison of existing materials benchmarks and MatMicroBench.** Existing benchmarks primarily rely on curated or public data and focus on knowledge-centric or isolated tasks. MatMicroBench uses experimentally acquired data to evaluate long-horizon microstructural characterization from scientific evidence to quantitative analysis and scientific conclusions.

for translating analysis rules into executable procedures, *Measure* for quantifying microstructural characteristics, and *Reason* for integrating evidence and measurements into scientific conclusions. Second, we instantiate these capabilities in four complementary tasks using scanning electron microscopy (SEM) and electron backscatter diffraction (EBSD), which provide morphological and crystallographic evidence, respectively (Mohammed & Abdullah, 2018; Humphreys, 2001; Peng et al., 2025). These tasks comprise SEM Static Defect-State Analysis (SEM-S), SEM Temporal Defect-Evolution Analysis (SEM-T), EBSD Inter-Grain Structure Analysis (EBSD-S), and EBSD Intra-Grain Orientation-Variation Analysis (EBSD-O). Third, we collect SEM and EBSD data through controlled laboratory experiments and construct 64 benchmark instances. Finally, we evaluate each capability through its corresponding artifact: spatial evidence for *Observe*, executable code for *Model*, numerical results for *Measure*, and scientific conclusions for *Reason*. We assess these artifacts against expert-validated references using task-specific metrics.

We evaluate 15 frontier models under a fixed Terminus-2 harness and compare six representative models across Terminus-2, OpenCode, and model-specific default harnesses. We find that current agents remain limited in end-to-end microstructural characterization, with the highest *Overall* score reaching 53.38 under Terminus-2 and no model leading all four tasks. Our capability analysis reveals task-specific bottlenecks and shows that strong performance in code generation and execution does not necessarily translate into accurate scientific outputs. We further find that harness effects vary by model and task, while higher cost does not consistently yield better scientific performance, highlighting the importance of selecting models and harnesses jointly. Together, these findings highlight the value of MatMicroBench in identifying capability bottlenecks and guiding the development of more reliable agents for microstructural characterization. Our contributions are threefold:

- **[A workflow-guided capability framework.]** We define four capabilities for microstructural characterization, *Observe*, *Model*, *Measure*, and *Reason*, and evaluate them through corresponding artifacts, enabling both end-to-end assessment and capability-level diagnosis.
- **[An experiment-grounded characterization benchmark.]** We introduce MatMicroBench, comprising 64 instances across four complementary SEM and EBSD tasks built from controlled laboratory experiments. We provide expert-validated references and task-specific evaluators to assess continuous microstructural analysis workflows.
- **[Systematic evaluation and diagnostic insights.]** We evaluate 15 frontier models under Terminus-2 and compare six representative models across three harnesses. We identify task-specific capability bottlenecks, reveal a gap between executable-code performance and scientific accuracy, and characterize performance-cost trade-offs across model-harness configurations.

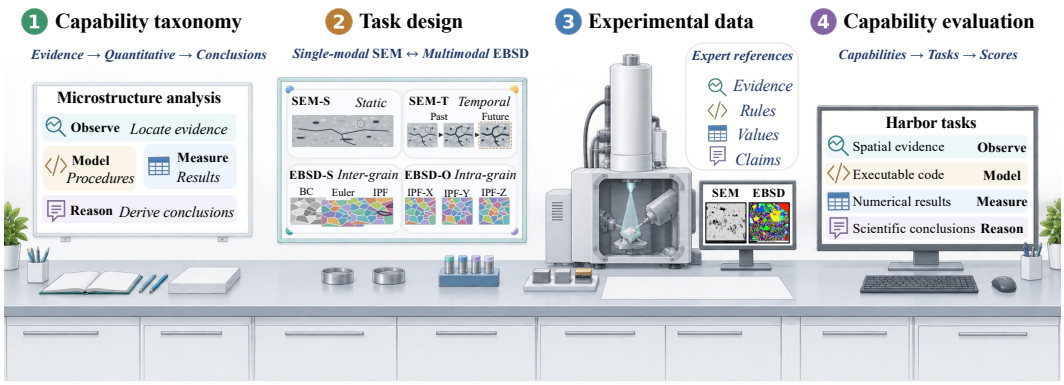


Figure 2: **Overview of benchmark construction and evaluation.** We define four core capabilities, *Observe*, *Model*, *Measure*, and *Reason*, and design complementary SEM and EBSD tasks around them using experimental data and expert-validated references. We evaluate each capability through spatial evidence, numerical results, executable code, and scientific conclusions.

2 Related Work

Materials Science Benchmarks. Materials science benchmarks have evolved from property prediction to domain reasoning, multimodal characterization, and workflow-level scientific tasks. Matbench and LLM4Mat-Bench standardize materials property prediction using predefined inputs and targets from established materials databases (Dunn et al., 2020; Niyongabo Rubungo et al., 2025). MaCBench manually curates multimodal question-answering tasks across the scientific workflow, whereas MatCha constructs characterization questions from published figures and open microscopy datasets (Alampara et al., 2025; Lai et al., 2025). AFMBench evaluates expert-designed tasks on physical AFM hardware, while AutoMat derives computational reproducibility tasks from published materials studies (Mandal et al., 2025; Huang et al., 2026). These benchmarks substantially broaden the realism of materials evaluation but mainly target property prediction, knowledge QA, image interpretation, instrument operation, or computational reproduction. They do not evaluate experimentally grounded, multi-step materials analysis that requires identifying evidence, quantifying it, implementing analysis procedures, and integrating intermediate outputs into scientific conclusions.

Agent Benchmarks for Scientific Tasks. Agent benchmarks for scientific tasks increasingly evaluate executable workflows rather than final answers alone. ScienceAgentBench assesses generated programs, execution results, and cost on tasks derived from peer-reviewed studies (Chen et al., 2025). DiscoveryBench and BLADE evaluate multi-step data-driven discovery with alternative analytical paths, while CORE-Bench focuses on computational reproducibility (Majumder et al., 2025; Gu et al., 2024; Siegel et al., 2024). BixBench and BiomniBench further introduce real biological data and process-level evaluation (Mitchener et al., 2025; Qu et al., 2026). These studies establish a broader evaluation paradigm centered on multi-step execution, intermediate artifacts, scientific validity, robustness, and efficiency. However, comparable agent benchmarks remain limited in materials science, particularly for characterization tasks, motivating the need for this work.

3 Benchmark Construction

3.1 Microstructure-Oriented Capability Taxonomy

Inspired by the real workflow of microstructural characterization, we identify four core capabilities underlying the complete analysis process. The process begins by identifying and localizing relevant microstructural features, reflecting *Observe*. These features are then quantified using appropriate computational rules and procedures. We capture the correctness of this process with *Model*, and the correctness of the resulting measurements with *Measure*. Finally, the features and quantitative results are integrated into scientific conclusions, reflecting *Reason*. Together, these four capabilities reflect distinct components of microstructural characterization and are defined as follows.

C1: Observe. The ability to identify, localize, and relate scientifically relevant microstructural evidence from characterization data. Such evidence can include defects, grains, target regions, spatial distributions, and changes across observations.

C2: Model. The ability to formalize scientific analysis rules into correct and executable computational procedures. This includes implementing definitions of quantities, spatial relations, statistical rules, and temporal operations so that the corresponding analysis can be reproduced.

C3: Measure. The ability to produce accurate and comparable quantitative characteristics from identified evidence using the defined analysis procedures. These quantities can include defect area fractions, grain size and shape statistics, population fractions and orientation-related quantities.

C4: Reason. The ability to integrate available evidence, quantitative results, and computational analysis to derive correct scientific conclusions. These conclusions can characterize microstructural state, temporal evolution, spatial heterogeneity, and potential material risks.

Directly assessing these four capabilities is nontrivial since they are abstract constructs within an end-to-end analysis process, so we assess them through four types of artifacts produced during task execution. For *Observe*, we use **spatial evidence**, such as binary masks of defects or grains, which reveals whether the agent correctly identifies relevant structures. For *Model*, we use **executable code** implementing the computational procedures, which reflects whether the underlying analysis rules and operations are correctly formalized. For *Measure*, we use **numerical results**, such as size, area, fraction, and count statistics, which reflect the accuracy of quantitative analysis. For *Reason*, we use **scientific conclusions**, such as material risks and defect evolution patterns, which reflect whether the agent correctly interprets the resulting evidence.

3.2 Scenario-Diverse Task Design

We instantiate the capability taxonomy through four tasks based on two microstructural characterization techniques: scanning electron microscopy (SEM) and electron backscatter diffraction (EBSD). SEM is widely used in materials characterization to reveal microstructural morphology and defects (Mohammed & Abdullah, 2018). EBSD is an important SEM-based technique for obtaining crystallographic information, including crystal orientations, phase identification, and their spatial distributions, by analyzing Kikuchi diffraction patterns (Humphreys, 2001; Sneddon et al., 2016; Peng et al., 2025). We use these complementary techniques to construct two SEM-based single-modal tasks and two EBSD-based multimodal tasks, with the latter requiring integration of multiple co-registered representations from the same field of view.

SEM: Static State to Temporal Evolution.

SEM Static Defect-State Analysis (SEM-S). SEM-S characterizes the defect state from a single SEM observation. The input consists of an SEM image and its observation type. The agent first identifies the valid analysis region and segments pores, cracks, and semi-molten particles. It then calculates the area fraction of each defect type and analyzes defect morphology, spatial relationships, regional material issues, and potential failure risks.

SEM Temporal Defect-Evolution Analysis (SEM-T). SEM-T extends static defect analysis to temporal evolution. The agent receives a sequence of historical SEM observations from the same specimen and field of view and predicts the SEM images at specified future cycles. It then extracts defect regions from the predicted states, quantifies defect fractions and newly formed defects, and analyzes changes in defect morphology, spatial distribution, and dominant damage patterns.

EBSD: Inter-Grain Structure to Intra-Grain Variation.

EBSD Inter-Grain Structure Analysis (EBSD-S). EBSD-S focuses on grain structure analysis using multiple EBSD maps. The input includes band contrast, Euler, and inverse pole figure maps together with sampling metadata. The agent identifies valid complete grains, localizes coarse and elongated grains, and quantifies grain count, size, aspect ratio, principal-axis orientation, and related population statistics. These results are then used to characterize grain morphology, spatial organization, and microstructural heterogeneity.

EBSD Intra-Grain Orientation-Variation Analysis (EBSD-O). EBSD-O focuses on orientation variation within individual grains. The input includes inverse pole figure maps along three sample

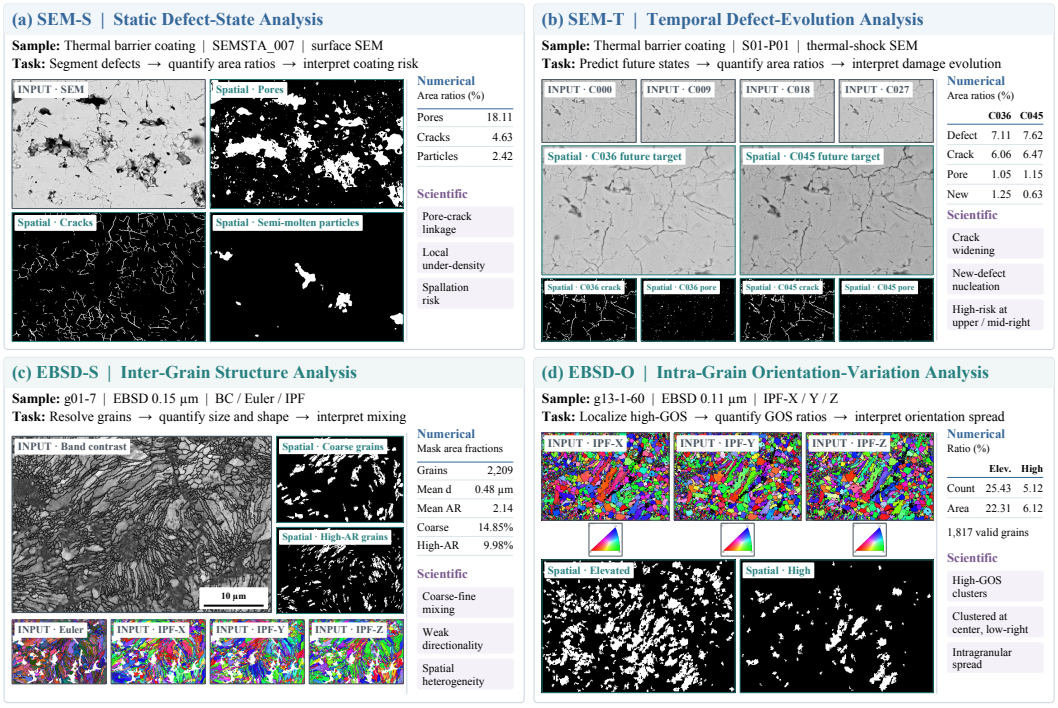


Figure 3: **Illustrative examples of the four benchmark tasks.** We show a simplified view of selected inputs and outputs. Complete task examples are provided in Appendix E.

directions, together with the corresponding color keys and sampling metadata. The agent identifies grains with relatively high intragranular orientation variation, quantifies their count and area fractions, and characterizes their spatial distribution and the resulting microstructural heterogeneity.

3.3 Experiment-Grounded Data Construction

Experiment-grounded data. We construct the data for all four tasks through controlled laboratory experiments. Materials experts prepare 36 specimens spanning different processing conditions and compositions for SEM and EBSD characterization. Among them, 20 specimens are used to construct 20 SEM-S instances, covering diverse morphologies. Another six specimens are used for SEM-T. For each specimen, two fixed locations are repeatedly observed across experimental stages, producing a six-stage early sequence and a ten-stage late sequence at each location. Each specimen-location pair sequence forms one instance, which yields 24 SEM-T instances. The remaining 10 specimens are characterized by EBSD and used to construct both EBSD-S and EBSD-O, yielding 10 instances for each task. This is practical because EBSD acquisition is time-consuming but information-rich, providing multiple co-registered maps from a single measurement. Overall, the benchmark contains 64 instances: 20 SEM-S, 24 SEM-T, 10 EBSD-S, and 10 EBSD-O. Details of the experimental conditions are provided in Appendix A.

Fine-grained expert references. Based on acquired data, materials experts define instance-specific task requirements and annotate the four reference artifact types designed in Section 3.2, with representative examples shown in Figure 3. For **spatial evidence**, experts manually identify and delineate target structures using Label Studio (Tkachenko et al., 2020-2025) as the annotation interface. For **executable code**, experts establish task-specific scientific definitions and calculation rules that the code should implement. For **numerical results**, experts perform the corresponding calculations to obtain reference measurements. For **scientific conclusions**, we first define a task-specific list of essential findings. For example, SEM-S covers defect morphology, spatial relationships, and potential risks, whereas SEM-T covers defect evolution and newly formed damage. Experts then annotate each item in the list separately, resulting in item-level scientific references rather than a single paragraph-level answer. This design enables each required conclusion to be evaluated independently, providing finer-grained supervision than conventional holistic reference answers. Detailed annotation criteria and procedures are provided in Appendix B.

3.4 CAPABILITY-RESOLVED EVALUATION

Executable task formulation. We organize each benchmark instance as an executable agent task using the Harbor format (Harbor Framework Team, 2026a). Each task includes an `instruction.md` describing the scientific objective, inputs, and required outputs in the workspace, a `task.toml` for task configuration, and an `environment/` directory for the execution environment. A skill file provides task-specific skills deemed necessary by domain experts, including relevant definitions and few-shot references for defect identification. Detailed task formulation is provided in Appendix C.1.

Artifact-specific metrics. Each artifact is evaluated against its expert reference using task-specific metrics. **Spatial evidence** is assessed with overlap metrics such as Dice and IoU, or image-similarity metrics. **Executable code** is evaluated by execution validity and calculation consistency. **Numerical results** are evaluated by errors from reference measurements, such as MAE. **Scientific conclusions** are evaluated item by item against the task-specific conclusion list using an LLM-as-a-Judge for correctness and coverage, where the judge receives only the submitted description and reference criteria. Detailed metric definitions are provided in Appendix C.2.

Grading rubric. Because artifact content varies across tasks and is evaluated with heterogeneous metrics, raw scores cannot be directly summed or averaged. We therefore define task-specific rubrics with domain experts to map artifact-level metrics to normalized capability scores. For each instance, capability c is scored as $S_c = \sum_{j=1}^{n_c} w_{c,j} s_{c,j}$, where $s_{c,j}$ is the normalized score of item j and $w_{c,j}$ its expert-defined weight, with $\sum_{j=1}^{n_c} w_{c,j} = 1$. The task score is then computed as a weighted combination of the four capability scores. Detailed rubrics and weights are provided in Appendix C.3.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Models and harnesses. We evaluate 15 frontier models on MATMICROBENCH. The evaluated models include Claude Opus 5 and Claude Sonnet 4.6 (Anthropic, 2026b;c), GPT-5.6 Sol, GPT-5.5, and GPT-5.4 (OpenAI, 2026d;c;b), Gemini 3.7 Flash and Gemini 3.1 Pro (Google DeepMind, 2026; Google, 2026a), Kimi K3 (Team et al., 2026), DeepSeek-V4 Pro (Xu et al., 2026), MiniMax M3 (MiniMax, 2026), Grok 4.5 and Grok 4.3 (xAI, 2026b;a), GLM-5.3 (Zeng et al., 2026), Step 3.5 Flash (StepFun Team, 2026), and Qwen3.8 Max (Qwen Team, 2026). All models are evaluated with the Terminus-2 reference harness (Harbor Framework Team, 2026b) for controlled comparison. We further evaluate six representative models with OpenCode (OpenCode, 2026) and their vendor-default harnesses, including Codex for GPT models (OpenAI, 2026a), Claude Code for Claude models (Anthropic, 2026a), and Gemini CLI for Gemini models (Google, 2026b).

Execution and grading. Tasks are executed through Harbor (Harbor Framework Team, 2026a) with task-specific Docker environments, with a maximum execution time of one hour per task. Runs exceeding this limit are terminated as timeouts. Spatial and numerical artifacts are scored by rule-based Python evaluators against ground truth, code artifacts are validated through controlled re-execution, and scientific conclusions are evaluated by GPT-5.4 (OpenAI, 2026b) as an LLM judge. We further validate the consistency of GPT-based evaluation with independent human assessments on a subset of benchmark outputs in Appendix D.4.

4.2 OVERALL PERFORMANCE

Current agents remain far from reliable end-to-end microstructural analysis. Table 1 reports normalized capability scores within each task and Overall performance for 15 models under the fixed Terminus-2 harness, while Figure 5a compares the relative task scores of the six leading models based on task-specific rubrics. Claude Opus 5, GPT 5.6 Sol, and GPT 5.5 achieve the highest Overall scores of 53.38, 53.33, and 52.96, respectively, yet no model exceeds 54 points, indicating considerable headroom for improvement. Performance is also uneven across tasks. Figure 5a reveals distinct task-specific strengths, with Claude Opus 5 performing best on SEM-S and EBSD-S, GPT 5.6 Sol on SEM-T, and GPT 5.5 on EBSD-O, demonstrating that the task design can distinguish complementary model strengths across different microstructural analysis scenarios.

Table 1: **Task- and capability-level performance under Terminus-2.** Capability scores are normalized and directly comparable. Overall is the macro-average across four tasks, with each task following its task-specific rubric. Detailed task-level results are provided in Appendix D.2. Models are ranked by Overall. Bold indicates the best non-saturated value in each column.

	SEM-S				SEM-T				EBSD-S				EBSD-O				
Model	**C1**	**C2**	**C3**	**C4**	**C1**	**C2**	**C3**	**C4**	**C1**	**C2**	**C3**	**C4**	**C1**	**C2**	**C3**	**C4**	**Overall**
Claude Opus 5	**40.80**	100.00	39.82	**71.78**	58.60	100.00	**62.07**	**67.71**	42.36	100.00	31.29	**62.57**	28.59	70.00	49.81	34.29	**53.38**
GPT 5.6 Sol	39.47	100.00	39.93	65.33	**65.67**	100.00	58.91	59.17	40.45	100.00	31.19	58.43	33.61	70.00	50.03	41.71	53.33
GPT 5.5	32.03	100.00	33.88	63.00	64.54	100.00	58.54	57.29	37.97	100.00	30.35	59.71	30.55	80.00	49.75	**54.86**	52.96
Gemini 3.7 Flash	33.22	100.00	**45.26**	53.67	63.21	100.00	59.15	50.42	**47.74**	86.67	**35.41**	43.00	**34.47**	66.67	47.33	30.86	49.04
Kimi K3	36.66	100.00	42.28	62.00	55.55	95.83	58.11	53.54	42.03	100.00	28.77	47.14	24.59	73.33	45.70	38.00	48.82
Claude Sonnet 4.6	31.84	100.00	33.43	56.22	60.26	100.00	51.00	47.19	40.21	96.67	32.29	41.71	24.31	96.67	**52.09**	41.29	48.32
DeepSeek V4 Pro	33.34	100.00	34.22	52.56	60.68	100.00	52.79	47.81	32.62	100.00	27.49	37.29	26.58	66.67	51.56	52.71	47.66
GPT 5.4	22.87	97.14	34.46	51.33	63.04	100.00	52.22	43.96	40.54	90.00	26.47	44.71	21.87	63.33	44.57	53.29	46.68
MiniMax M3	26.98	100.00	33.78	47.44	63.19	99.38	57.51	53.54	32.19	100.00	28.22	39.71	26.47	80.00	49.25	39.71	46.46
Gemini 3.1 Pro	35.03	100.00	40.75	42.56	62.52	100.00	54.15	27.50	35.10	90.00	26.03	37.86	30.70	66.67	50.96	47.29	45.84
Grok 4.5	36.16	100.00	42.14	61.11	63.52	100.00	57.09	55.00	21.52	33.33	13.44	31.43	13.70	33.33	10.38	28.00	39.63
Grok 4.3	21.55	95.00	28.64	38.67	55.57	77.71	40.09	28.12	30.18	76.67	24.10	20.29	25.96	66.67	26.65	33.71	35.85
GLM 5.3	31.26	62.86	30.33	30.78	63.94	100.00	60.57	52.29	32.81	70.00	22.00	45.29	4.75	6.67	5.40	7.71	34.86
Step 3.5 Flash	21.37	95.00	31.62	26.11	39.49	53.02	26.72	18.54	27.53	60.00	18.75	27.71	20.48	40.00	27.66	18.43	28.55
Qwen 3.8 Max	35.74	75.00	33.28	37.67	42.53	70.83	40.87	37.92	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	20.77

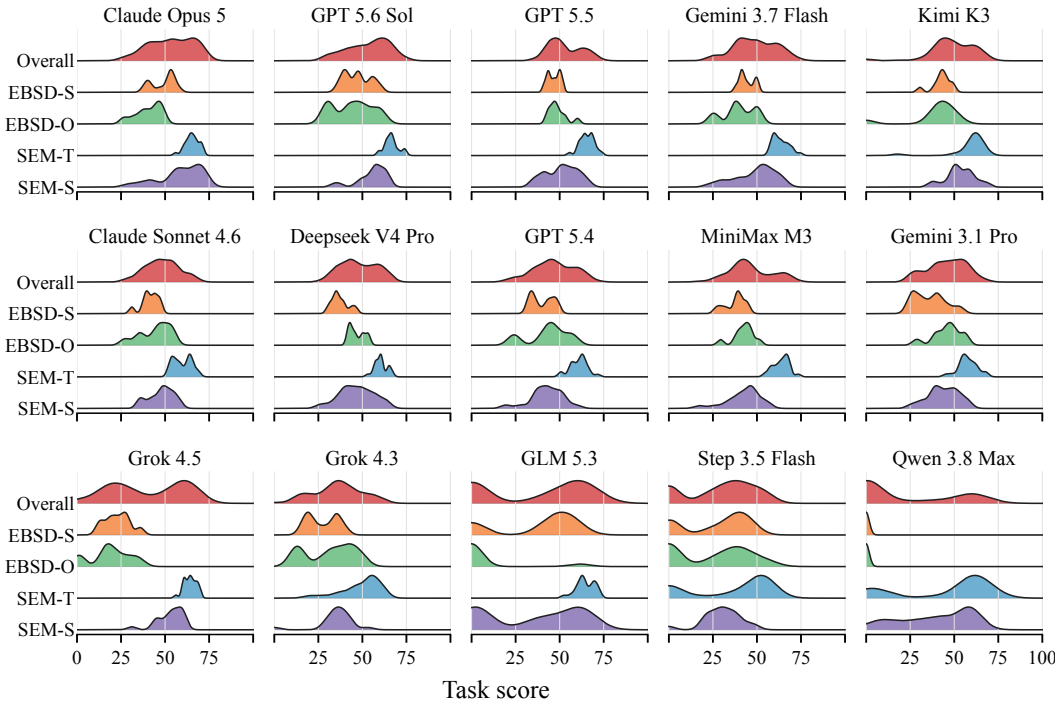


Figure 4: **Distribution of instance-level task scores for 15 models on Terminus-2.** Overall denotes the equally weighted aggregate distribution across the four tasks. The corresponding mean and population standard deviation across task instances are reported in Appendix D.3.

Case-level consistency and execution completion vary substantially across models and tasks. Figure 4 shows distinct case-level performance patterns across tasks. Leading models such as Claude Opus 5, GPT 5.6 Sol, and GPT 5.5 achieve relatively high and concentrated scores on SEM-T, whereas their SEM-S scores are more broadly distributed across specimens. This contrast shows that current agents are not equally reliable across experimental cases, even when their aggregate task scores are competitive. For lower-performing models such as GLM 5.3, Step 3.5 Flash, and Qwen 3.8 Max, we observe a substantial number of cases with scores close to zero. Together with the low non-timeout rates reported for these models in Appendix D.1, this suggests that many failures arise because the agents fail to complete the task within the limited time. For these models, completing the full analysis workflow within the allowed execution time remains a major bottleneck.

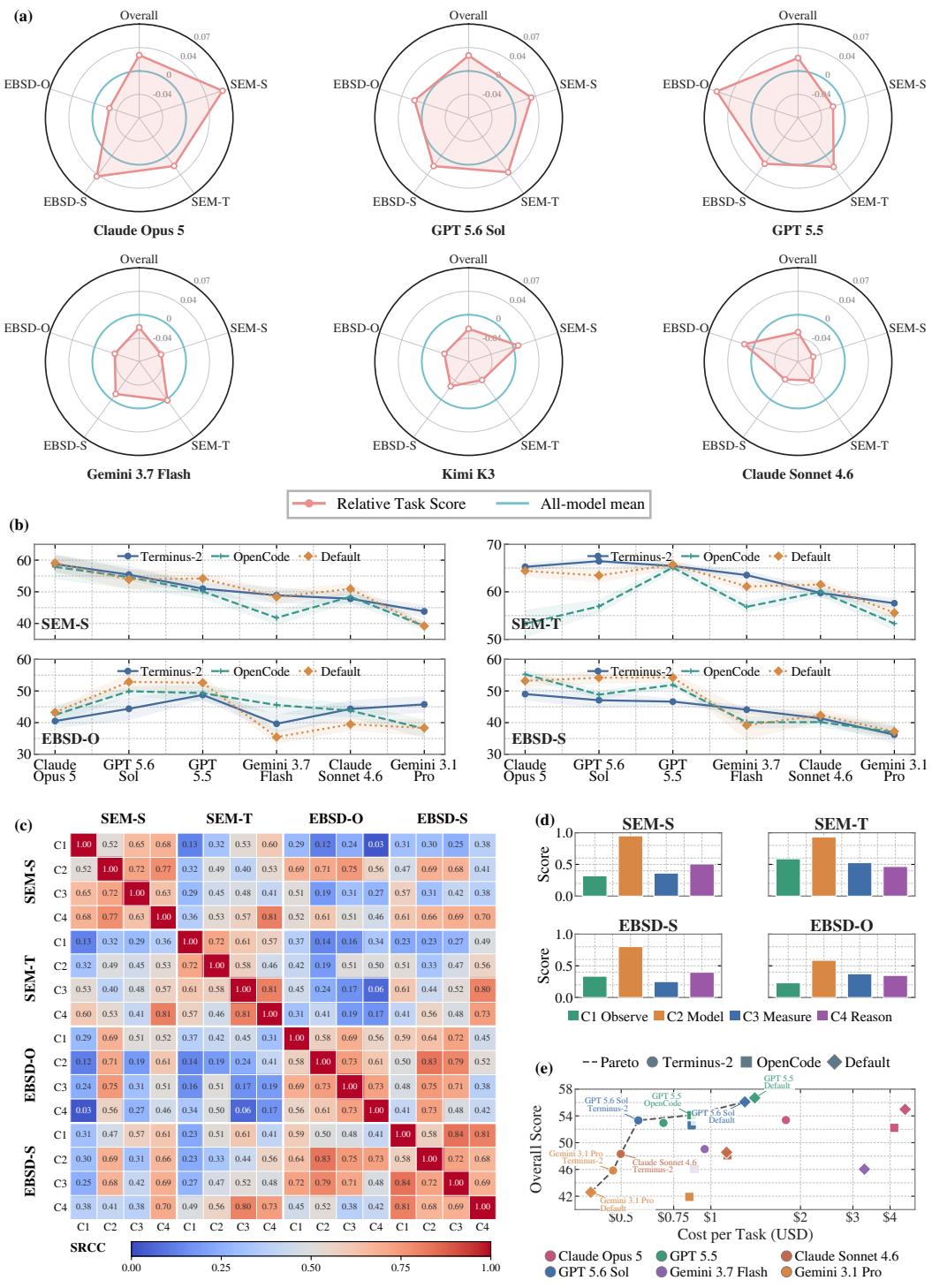


Figure 5: **Multi-view evaluation of agent performance on MatMicroBench. (a)** Relative profiles of six models under Terminus-2, where each axis shows the deviation from the six-model mean and the blue ring marks zero deviation. **(b)** Task-level performance under Terminus-2, OpenCode, and Default. Default denotes Claude Code for Claude models, Codex for GPT models, and Gemini CLI for Gemini models. The width of each shaded band represents the standard error. **(c)** SRCC among the 16 task-capability dimensions across 15 models, where C1–C4 denote Observe, Model, Measure, and Reason. **(d)** Mean capability scores for each task, with C1–C4 arranged along the horizontal axis. **(e)** Overall score versus average cost per task. Colors identify models, markers identify harnesses, and the dashed line shows the Pareto frontier.

4.3 Task and Capability Analysis

Different tasks expose different capability bottlenecks. Figure 5d averages the four capability scores across all 15 models to identify the dominant bottleneck in each task. SEM-S is limited mainly by *Observe* and *Measure*, SEM-T by *Reason*, EBSD-S by *Measure*, and EBSD-O by *Observe*. These differences reflect the dominant challenge of each task: SEM-S requires precise localization and quantification of small defects, SEM-T requires reasoning over temporal evidence, EBSD-S depends on accurate grain-level statistics, and EBSD-O requires identifying subtle orientation differences across co-registered representations. Furthermore, Model scores are consistently higher, partly because executable procedures can be directly validated through code execution and current agents are relatively capable of generating and running analysis scripts. Yet successful execution alone does not guarantee scientific accuracy: errors in *Observe* can propagate to *Measure* and ultimately *Reason*. The central challenge therefore lies in grounding the analysis pipeline in correct material evidence and measurements.

Capability scores provide complementary information within and across tasks. Figure 5c examines rank correlations among the 16 task-capability dimensions across 15 models. Within tasks, some capabilities are strongly correlated: *Observe* and *Measure* reach 0.84 on EBSD-S, while *Measure* and *Reason* reach 0.81 on SEM-T, reflecting the dependence of downstream analysis on accurate evidence and measurements. However, these relationships are not uniform. On SEM-T, *Model* and *Reason* correlate only at 0.46, indicating that stronger computational procedures do not necessarily yield better scientific conclusions. Across tasks, the same capability can capture different challenges: the *Observe* correlation between SEM-S and SEM-T is only 0.13, reflecting the difference between static localization and temporal tracking. These results support separate evaluation of capabilities and tasks, providing a more diagnostic view than a single aggregate score.

4.4 Harness Effects and Cost Efficiency

Performance varies across harnesses for the same model and task. We compare six models across Terminus-2, OpenCode, and their vendor-default harnesses in Figure 5b, keeping the test set and scoring method fixed. For Claude Opus 5, switching from Terminus-2 to OpenCode increases the EBSD-S score by 6.23 points but decreases the SEM-T score by 11.99 points. Default harnesses also offer no consistent advantage over other harnesses. Compared to Terminus-2, Codex improves the *Overall* score of GPT 5.5, whereas Gemini CLI lowers that of Gemini 3.7 Flash. These results suggest that harness selection should be guided by evaluation on the target model and task, with execution cost considered alongside scientific performance.

Higher cost does not consistently yield better scientific performance. We plot the *Overall* score against cost per task for all 18 model-harness configurations in Figure 5e and trace the Pareto frontier to characterize the performance-cost trade-off. GPT 5.5 with Codex scores 56.69 at a substantially lower cost than Claude Opus 5 with Claude Code, which scores 54.98. Claude Opus 5 with OpenCode is also among the most expensive configurations but scores only 52.22, below several cheaper GPT configurations. These comparisons show that higher spending does not necessarily translate into stronger scientific performance. The Pareto frontier further highlights competitive configurations at different budget levels, suggesting that model and harness should be selected jointly according to the desired balance between performance and cost.

5 Conclusion and Limitations

We introduced MatMicroBench, an experiment-grounded benchmark comprising 64 instances across four complementary SEM and EBSD tasks for end-to-end microstructural characterization. Our evaluation of 15 models reveals task-specific capability bottlenecks, a gap between executable-code performance and scientific accuracy, and varying harness effects. These findings highlight the importance of evaluating scientific outputs beyond successful execution and selecting models and harnesses jointly. Future work should strengthen verification and error correction across analysis stages to ensure that measurements and conclusions remain grounded in observed evidence.

Limitations. The acquisition of experimental data and expert-validated references requires substantial time and resources, limiting the current scale and coverage of our benchmark. Extending evaluation to more materials, experimental conditions, and characterization techniques remains important for developing reliable characterization agents.

AI USE STATEMENT

In this work, we used generative AI tools to assist with manuscript drafting and language editing to improve clarity and readability. We also used GPT-5.4 as an LLM-based judge to evaluate scientific-conclusion artifacts produced by the benchmarked agents, following task-specific rubrics and expert-validated reference annotations. We validated this evaluation procedure through comparison with independent human assessments on a subset of benchmark outputs in Appendix D.4. We reviewed and revised all AI-assisted manuscript content and verified the reported analyses and conclusions. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

MatMicroBench supports research on reliable scientific agents using experimentally acquired SEM images and EBSD maps of material specimens. The benchmark data contain no personal information. Materials experts contribute reference annotations and independently assess selected model outputs to validate automated evaluation. Reference artifacts are checked for consistency with experimental evidence, and inferred material issues and risks are stated conditionally.

REPRODUCIBILITY STATEMENT

We support reproducibility through documented data construction, executable task specifications, and explicit evaluation protocols. Section 3.3 and Appendices A and B describe experimental data acquisition and expert reference construction. Section Cand Appendix C specify task packaging, execution constraints, artifact-specific metrics, and score aggregation. Section 4.1 describes the evaluated models, agent harnesses, execution environments, and grading setup. Appendix D provides detailed experimental results and human validation of the LLM-based judge, while Appendix E presents representative task inputs, requirements, and reference artifacts. Complete task-specific rubrics, verification scripts, and scoring configurations are provided with the benchmark.

REFERENCES

Nawaf Alampara, Mara Schilling-Wilhelmi, Martiño Ríos-García, Indrajeet Mandal, Pranav Khetarpal, Hargun Singh Grover, NM Anoop Krishnan, and Kevin Maik Jablonka. Probing the limitations of multimodal language models for chemistry and materials research. *Nature computational science*, 5(10):952–961, 2025.

Anthropic. Claude code documentation, 2026a. URL `https://code.claude.com/docs/en/overview`.

Anthropic. Introducing claude opus 5, 2026b. URL `https://www.anthropic.com/news/claude-opus-5`.

Anthropic. Introducing claude sonnet 4.6, 2026c. URL `https://www.anthropic.com/news/claude-sonnet-4-6`.

Henry Chan, Mathew Cherukara, Troy D Loeffler, Badri Narayanan, and Subramanian KRS Sankaranarayanan. Machine learning enabled autonomous microstructural characterization in 3d samples. *NPJ Computational Materials*, 6(1):1, 2020.

Guangyao Chen, Wenhao Yuan, and Fengqi You. Bridging electron microscopy and materials analysis with an autonomous agentic platform. *Science Advances*, 12(14):eaed0583, 2026.

Ziru Chen, Shijie Chen, Yuting Ning, Qianheng Zhang, Boshi Wang, Botao Yu, Yifei Li, Zeyi Liao, Chen Wei, Zitong Lu, et al. Scienceagentbench: Toward rigorous assessment of language agents for data-driven scientific discovery. In *International Conference on Learning Representations*, volume 2025, pp. 96934–96990, 2025.

Brian L DeCost, Bo Lei, Toby Francis, and Elizabeth A Holm. High throughput quantitative metallography for complex microstructures using deep learning: A case study in ultrahigh carbon steel. *Microscopy and Microanalysis*, 25(1):21–29, 2019.

Alexander Dunn, Qi Wang, Alex Ganose, Daniel Dopp, and Anubhav Jain. Benchmarking materials property prediction methods: the matbench test set and automatminer reference algorithm. *npj Computational Materials*, 6(1):138, 2020.

Google. Gemini 3.1 Pro: A smarter model for your most complex tasks, February 2026a. URL https://blog.google/innovation-and-ai/models-and-research/gemini-models/gemini-3-1-pro/.

Google. Gemini cli, 2026b. URL https://github.com/google-gemini/gemini-cli.

Google DeepMind. Gemini 3.7 flash model card, 2026. URL https://deepmind.google/models/model-cards/gemini-3-7-flash/.

Ken Gu, Ruoxi Shang, Ruien Jiang, Keying Kuang, Richard-John Lin, Donghe Lyu, Yue Mao, Youran Pan, Teng Wu, Jiaqian Yu, et al. Blade: Benchmarking language model agents for data-driven science. In *Findings of the association for computational linguistics: EMNLP 2024*, pp. 13936–13971, 2024.

Harbor Framework Team. Harbor: A framework for evaluating and optimizing agents and models in container environments, 2026a. URL https://doi.org/10.5281/zenodo.20953922.

Harbor Framework Team. Terminus-2, 2026b. URL https://github.com/harbor-framework/harbor/tree/main/src/harbor/agents/terminus_2.

Ziyang Huang, Yi Cao, Ali K Shargh, Jing Luo, Ruidong Mei, Mohd Zaki, Zhan Liu, Wyatt Bunstine, William Jurayj, Somdatta Goswami, et al. Can coding agents reproduce findings in computational materials science? *arXiv preprint arXiv:2605.00803*, 2026.

FJ Humphreys. Review grain and subgrain characterisation by electron backscatter diffraction. *Journal of materials science*, 36(16):3833–3854, 2001.

Zhengzhao Lai, Youbin Zheng, Zhenyang Cai, Haonan Lyu, Jinpu Yang, Hongqing Liang, Yan Hu, and Benyou Wang. Can multimodal llms see materials clearly? a multimodal benchmark on materials characterization. *arXiv preprint arXiv:2509.09307*, 2025.

Andres M. Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D White, and Philippe Schwaller. Augmenting large language models with chemistry tools. *Nature machine intelligence*, 6(5):525–535, 2024.

Bodhisattwa Prasad Majumder, Harshit Surana, Dhruv Agarwal, Bhavana Dalvi Mishra, Abhijeetsingh Meena, Aryan Prakhar, Tirth Vora, Tushar Khot, Ashish Sabharwal, and Peter Clark. Discoverybench: Towards data-driven discovery with large language models. In *International Conference on Learning Representations*, volume 2025, pp. 4556–4579, 2025.

Indrajeet Mandal, Jitendra Soni, Mohd Zaki, Morten M Smedskjaer, Katrin Wondraczek, Lothar Wondraczek, Nitya Nand Gosvami, and NM Anoop Krishnan. Evaluating large language model agents for automation of atomic force microscopy. *Nature Communications*, 16(1):9104, 2025.

MiniMax. Minimax m3: Frontier coding, 1m context, native multimodality, 2026. URL https://www.minimax.io/blog/minimax-m3.

Ludovico Mitchener, Jon M Laurent, Alex Andonian, Benjamin Tenmann, Siddharth Narayanan, Geemi P Wellawatte, Andrew White, Lorenzo Sani, and Samuel G Rodriques. Bixbench: a comprehensive benchmark for llm-based agents in computational biology. *arXiv preprint arXiv:2503.00096*, 2025.

Azad Mohammed and Avin Abdullah. Scanning electron microscopy (sem): A review. In *Proceedings of the 2018 international conference on hydraulics and pneumatics—HERVEX, Băile Govora, Romania*, volume 2018, pp. 7–9, 2018.

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, 2025.

OpenAI. Codex cli, 2026a. URL https://developers.openai.com/codex/cli.

OpenAI. Introducing gpt-5.4, 2026b. URL https://openai.com/index/introducing-gpt-5-4/.

OpenAI. Introducing gpt-5.5, 2026c. URL https://openai.com/index/introducing-gpt-5-5/.

OpenAI. Gpt-5.6: Frontier intelligence that scales with your ambition, 2026d. URL https://openai.com/index/gpt-5-6/.

OpenCode. Opencode: The open source ai coding agent, 2026. URL https://opencode.ai/.

Fan Peng, Xuemei Song, Yiling Huang, Xingyu Jin, Yuqing Jiang, Yue Sun, and Yi Zeng. A new ebsd indexing method with enhanced grain boundary indexing performance using a three-dimensional parameter space. *Ultramicroscopy*, pp. 114227, 2025.

Yuanhao Qu, Yingzhou Lu, Xinming Tu, Serena Zhang, Tianwei She, Alexander Glenn Shaw, Jou-Ho Shih, Bingqing Zhao, Minjie Shen, Haochen Yang, et al. Biomnibench: Process-level evaluation of llm agents for real-world biomedical research. *bioRxiv*, pp. 2026–05, 2026.

Qwen Team. Qwen3.8-Max: A new bar for coding and cowork, August 2026. URL https://qwen.ai/blog?id=qwen3.8.

Zachary S Siegel, Sayash Kapoor, Nitya Nadgir, Benedikt Stroebl, and Arvind Narayanan. Corebench: Fostering the credibility of published research through a computational reproducibility agent benchmark. *arXiv preprint arXiv:2409.11363*, 2024.

Glenn C Sneddon, Patrick W Trimby, and Julie M Cairney. Transmission kikuchi diffraction in a scanning electron microscope: A review. *Materials Science and Engineering: R: Reports*, 110: 1–12, 2016.

StepFun Team. step-3.5-flash. https://static.stepfun.com/blog/step-3.5-flash/, 2026.

Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, Jianfeng Cai, Xinyuan Cai, Peizhou Cao, Yuxuan Cao, Ziwei Chai, Y Charles, et al. Kimi k3: Open frontier intelligence. *arXiv preprint arXiv:2607.24653*, 2026.

Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. Label Studio: Data labeling software, 2020-2025. URL https://github.com/HumanSignal/label-studio. Open source software available from https://github.com/HumanSignal/label-studio.

xAI. Grok 4.3, 2026a. URL https://docs.x.ai/developers/models/grok-4.3.

xAI. Grok 4.5, 2026b. URL https://docs.x.ai/developers/models/grok-4.5.

Anyi Xu, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chenchen Ling, et al. Deepseek-v4: Towards highly efficient million-token context intelligence. *arXiv preprint arXiv:2606.19348*, 2026.

Wufeng Xue, Lei Zhang, Xuanqin Mou, and Alan C Bovik. Gradient magnitude similarity deviation: A highly efficient perceptual image quality index. *IEEE transactions on image processing*, 23(2): 684–695, 2013.

Aohan Zeng, Xin Lv, Zhenyu Hou, Zhengxiao Du, Qinkai Zheng, Bin Chen, Da Yin, Chendi Ge, Chenghua Huang, Chengxing Xie, et al. Glm-5: from vibe coding to agentic engineering. *arXiv preprint arXiv:2602.15763*, 2026.

Lin Zhang, Lei Zhang, Xuanqin Mou, and David Zhang. Fsim: A feature similarity index for image quality assessment. *IEEE transactions on Image Processing*, 20(8):2378–2386, 2011.

A EXPERIMENTAL DATA ACQUISITION

We acquired the SEM images and EBSD maps for MATMICROBENCH using the laboratory instruments shown in Figure 6. Two SEM systems equipped with EBSD detectors provided both microstructural images and crystallographic orientation maps. We used two additional systems for SEM imaging only. All four microscopes were manufactured by Thermo Fisher Scientific (USA). The instrument models and EBSD configurations are specified in the subfigure captions.



(a) Verios G4 UC high-resolution field-emission SEM equipped with a NordlysNano EBSD detector from Oxford Instruments (UK).



(b) Apreo ChemiSEM S equipped with an EBSD system.



(c) Verios G4 UC field-emission SEM used for high-resolution imaging.



(d) Versa 3D focused ion beam/SEM dual-beam system used for SEM imaging.

Figure 6: **Experimental instruments used for benchmark data acquisition.** The upper row shows systems used for both SEM imaging and EBSD acquisition. The lower row shows systems used for SEM imaging only.

B EXPERT ANNOTATION AND REFERENCE CONSTRUCTION

We construct four complementary reference types for each instance, covering spatial evidence, computational rules, numerical measurements, and scientific conclusions. Materials experts define the scientific targets, analysis region, and required outputs. All references follow these shared definitions to connect scientific conclusions with their supporting evidence and calculations.

B.1 SPATIAL EVIDENCE ANNOTATION

Materials experts annotate task-relevant structures using Label Studio (Tkachenko et al., 2020-2025), with experimental images as the primary evidence. Masks preserve the coordinate system and resolution of the corresponding images, and boundaries are verified through overlays.

For **SEM-S**, experts annotate pores, cracks, and semi-molten particles separately based on morphology and surrounding context. For **SEM-T**, crack and pore masks capture defect states at each experimental stage, while experimentally acquired target frames provide references for future-image prediction. For **EBSD-S**, experts identify coarse and elongated grains from co-registered maps and grain measurements. For **EBSD-O**, experts localize grains with elevated intragranular orientation differences using orientation measurements and multi-direction inverse pole figure (IPF) maps. Both EBSD tasks follow the prescribed analysis region and eligible grain population.

B.2 COMPUTATIONAL RULE SPECIFICATION

Materials experts specify the rules that submitted analysis code must implement, including required inputs, eligible structures, spatial calibration, thresholds, statistical operations, and output quantities. SEM rules cover defect areas, mask unions, temporal comparisons, and newly formed damage. EBSD rules cover grain selection, size and shape statistics, major-axis orientation bins, and orientation-related measurements. We explicitly specify the denominators for count and area fractions. These definitions link each numerical output to a reproducible calculation and support evaluation of the underlying computational procedure.

B.3 NUMERICAL REFERENCE CALCULATION

Materials experts calculate numerical references from annotated structures and experimental measurements using the specified rules. **SEM-S** includes area fractions for pores, cracks, and semi-molten particles. **SEM-T** includes joint defect, crack, pore, and newly formed damage area fractions at the required stages. **EBSD-S** includes grain counts, diameter and area statistics, aspect ratios, coarse- and elongated-grain fractions, and major-axis orientation distributions. **EBSD-O** includes orientation statistics and count and area fractions for the selected grain populations. We record each quantity with its unit and analysis scope, using the same regions and population definitions as the spatial references.

B.4 SCIENTIFIC CONCLUSION ANNOTATION

Annotation format and evidence. We record scientific conclusions as task-specific annotation items. Experts inspect the experimental images and associated spatial and numerical references, then complete each item using predefined descriptors or free text.Figure 7 summarizes the supporting evidence and annotation items across the four tasks.

SEM-S. We use SEM images and defect masks to describe pore morphology and location, crack morphology and direction, particle distribution and surrounding defect enrichment, and relationships among defect types. These observations support interpretations of possible microstructural issues, including insufficient particle melting, poor lamellar bonding, or insufficient local densification. We also identify high-risk regions and plausible consequences, such as crack propagation, local spallation, or reduced mechanical performance. All inferred issues and risks are stated conditionally and tied to visible evidence.

SEM-T. We compare initial and final experimental observations from the same specimen and location using the corresponding defect masks and area measurements. Annotation items cover both



Figure 7: **Scientific conclusion annotation across the four tasks.** The SEM-S, SEM-T, EBSD-S, and EBSD-O schemas contain 10, 7, 10, and 8 annotation items, respectively. Each panel links supporting evidence to annotation items and scientific interpretations. Related items are grouped for display but recorded separately.

endpoint morphologies, pore and crack evolution, the dominant mode of damage evolution, high-risk regions, and propagation directions. We distinguish newly appearing defects from the growth or connection of existing defects. Area changes are checked against numerical references, while changes in morphology, branching, and connectivity are checked against images and masks.

EBSD-S. We jointly inspect co-registered band-contrast, Euler-angle, and IPF-X/Y/Z maps alongside grain masks and statistics. Annotation items cover grain size and distribution, morphology, the elongation criterion, major-axis directionality, coarse- and elongated-grain locations, relationships between these populations, and structural heterogeneity. We ground size descriptions in diameter statistics and image scale, directional descriptions in the prescribed 30° bins, and spatial descriptions in reference masks. Heterogeneity assessments draw on regional differences in grain size, shape, directionality, and population distributions.

EBSD-O. We inspect the three IPF maps and their color keys alongside grain orientation spread (GOS) statistics and masks for elevated- and high-GOS grains. Annotation items cover overall orientation variation and its distribution, high-GOS grain locations and spatial patterns, intragranular IPF color behavior, orientation change, and heterogeneity. Grain selection follows the task-specific GOS percentile criteria. We interpret IPF color variation as indirect evidence alongside orientation measurements and record the evidence linking intragranular variation to regional heterogeneity.

B.5 Consistency Verification

Before finalization, materials experts check annotation completeness and consistency across reference artifacts. We compare masks with experimental images, verify numerical results against the specified calculations, and check scientific conclusions against spatial and numerical evidence and related claims. Unsupported interpretations are revised or removed. We record uncertainty explicitly, distinguishing features that are not observed from those that cannot be determined.

Each finalized record links its instance identifier to the reference artifacts, individual conclusion items, and a readable scientific description, with supplementary notes where needed. This organization preserves correspondence across outputs and supports item-level evaluation while retaining the relationships among the findings.

C Evaluation Protocol

C.1 Executable Task Formulation

We package each benchmark instance as a Harbor task using schema version 1.4 (Harbor Framework Team, 2026a). Each task defines an agent-visible instruction file, scientific inputs, execution environment, required outputs, and a verification procedure.

The `instruction.md` file specifies the task requirements, while `task.toml` defines the execution configuration. Scientific inputs are stored in `environment/`. We use a fixed container with `/tmp_workspace` as the working directory and mount task-specific skills at `/tmp_workspace/skills`. Agents must write all submission artifacts to `/tmp_workspace/results/`.

Each agent run has a 3600 s (one-hour) execution limit, and the verifier has a separate 600 s limit. The verifier is isolated from the agent execution environment and has access to expert references only after the agent finishes. Submitted computational artifacts are executed in a controlled environment for verification.

C.2 Artifact-Specific Evaluation

Each benchmark task evaluates four complementary outputs: spatial evidence, numerical results, executable analysis procedures, and scientific conclusions. These outputs correspond to the four capabilities: *Observe* (C1), *Model* (C2), *Measure* (C3), and *Reason* (C4).

Spatial evidence (C1: Observe). We evaluate submitted spatial outputs against expert references. For binary masks, we use the Dice coefficient and intersection-over-union (IoU),

$$\mathrm{Dice}(P,T) = \frac{2|P \cap T|}{|P| + |T|}, \qquad \mathrm{IoU}(P,T) = \frac{|P \cap T|}{|P \cup T|}, \tag{1}$$

where P and T denote the submitted and reference foreground pixel sets. For SEM-T, we compare predicted images with the corresponding future experimental observations using the Feature Similarity Index (FSIM) and Gradient Magnitude Similarity Deviation (GMSD). FSIM evaluates structural similarity through phase congruency and gradient magnitude, whereas GMSD computes the standard deviation of the pixel-wise gradient-magnitude similarity map (Zhang et al., 2011; Xue et al., 2013). We use FSIM and $1 - \mathrm{GMSD}$ as image-level scores, so that higher values consistently indicate better agreement.

Executable code (C2: Model). We evaluate executable analysis procedures through two aspects: calculation consistency and execution validity. Calculation consistency measures whether reported numerical outputs are consistent with the submitted analysis artifacts. Execution validity evaluates whether the submitted programs implement the required computational procedures under controlled verification cases. These checks evaluate reproducible computation separately from scientific agreement with expert references.

Numerical results (C3: Measure). We compare submitted measurements with expert references using task-appropriate error metrics. Absolute or relative errors are selected according to the physical meaning of each quantity. We convert a nonnegative measurement error e into a normalized score:

$$s(e,\tau) = \frac{\tau}{\tau + e}, \tag{2}$$

where τ is a task-specific tolerance determined before evaluation. The same tolerance settings are applied consistently across all evaluated models and harnesses.

Scientific conclusions (C4: Reason). We evaluate submitted scientific descriptions against expert annotations using task-specific rubrics. An LLM judge assesses whether the conclusions capture the required microstructural findings, agree with the available evidence, and support appropriate scientific interpretations. We instruct the judge to focus on scientific content rather than writing style and to accept semantically equivalent wording. Partial credit is allowed, and unsupported claims reduce credit on the corresponding criteria. The task-specific scientific criteria are provided in the representative benchmark instances.

C.3 Score Aggregation

Within-instance aggregation. Each instance is evaluated with a 100-point rubric distributed across the four capabilities. We first aggregate normalized item scores within each capability:

$$S_c = \sum_j w_{c,j} s_{c,j}, \tag{3}$$

where $s_{c,j} \in [0,1]$ is the normalized score of item j within capability c. The instance score is then computed as:

$$S = \sum_{c=1}^{4} W_c S_c. \tag{4}$$

Let $b_{c,j}$ denote the point budget of item j, and $B_c = \sum_j b_{c,j}$ denote the capability budget. The item and capability weights are defined as

$$w_{c,j} = \frac{b_{c,j}}{B_c}, \qquad W_c = \frac{B_c}{100}. \tag{5}$$

Therefore,

$$\sum_j w_{c,j} = 1, \qquad \sum_c W_c = 1.$$

Task-specific weighting. Capability budgets are determined by materials experts according to the scientific objectives of each characterization task. Different tasks emphasize different capabilities

because their primary scientific targets differ. For example, temporal SEM prediction emphasizes spatial evolution, whereas EBSD characterization requires stronger quantitative measurement.

Benchmark-level reporting. We multiply normalized capability and instance scores by 100 for reporting. We average instance scores within each task category and equally average the four task categories to obtain the *Overall* score. Capability scores are reported separately using the same task-level averaging procedure. The complete task-specific rubrics, verification scripts, and scoring configurations are released with the benchmark.

D Additional Experimental Results

We report supplementary results on execution efficiency, variation across task instances, and selected instance-level runs.

D.1 Execution Efficiency

Tables 2 and 3 report Overall scores and execution statistics for 15 models under Terminus-2 and six models across three harness settings, respectively. We average each metric within each task and then equally across the four tasks, following Appendix C.3. Overall scores retain the task-specific rubric weights.

Time, *Cost*, *Turns*, and *Tokens* denote execution time, monetary cost, interaction turns, and token usage, respectively. We report time in minutes, cost in US dollars, and tokens in thousands. *NTR* denotes the percentage of runs that do not time out under the one-hour execution limit. It records timeout status rather than scientific correctness.

Table 2: **Performance and execution efficiency under Terminus-2.** We report macro-averages across the four tasks. Models are ordered by Overall score, with the highest score in bold.

Model	Overall	Time (min)	Cost ($)	Turns	Tokens (k)	NTR (%)
Claude Opus 5	**53.38**	22.12	1.79	23.75	822.2	100.00
GPT 5.6 Sol	53.33	4.74	0.57	10.00	159.7	100.00
GPT 5.5	52.96	5.05	0.69	10.50	142.0	100.00
Gemini 3.7 Flash	49.04	13.61	0.95	44.00	1272.5	100.00
Kimi K3	48.82	27.83	0.68	22.25	594.5	96.46
Claude Sonnet 4.6	48.32	6.72	0.50	15.75	389.1	100.00
DeepSeek V4 Pro	47.66	16.26	0.42	34.75	1145.8	100.00
GPT 5.4	46.68	3.92	0.40	11.75	289.1	98.96
MiniMax M3	46.46	14.20	0.53	59.00	3174.2	100.00
Gemini 3.1 Pro	45.84	5.26	0.47	10.00	135.7	100.00
Grok 4.5	39.63	15.64	1.29	24.75	1637.6	92.50
Grok 4.3	35.85	2.78	1.00	21.25	786.3	98.75
GLM 5.3	34.86	47.17	1.47	47.25	3266.7	55.21
Step 3.5 Flash	28.55	39.44	0.06	17.25	460.0	68.13
Qwen 3.8 Max	20.77	55.79	0.93	35.25	1506.7	30.42

Table 3: **Performance and execution efficiency across harnesses.** We compare six models under Terminus-2, OpenCode, and their model-specific default harnesses. Bold denotes the highest Overall score within each model.

Model	Harness	Overall	Time (min)	Cost ($)	Turns	Tokens (k)	NTR (%)
Claude Opus 5	Terminus-2	53.38	22.12	1.79	23.75	822.2	100.00
	OpenCode	52.22	29.37	4.13	55.25	3550.2	88.13
	Claude Code	**54.98**	26.38	4.49	55.25	4182.1	98.75
Claude Sonnet 4.6	Terminus-2	48.32	6.72	0.50	15.75	389.1	100.00
	OpenCode	48.11	11.59	1.14	41.00	1694.3	97.50
	Claude Code	**48.57**	9.53	1.13	21.50	1305.0	100.00
Gemini 3.1 Pro	Terminus-2	**45.84**	5.26	0.47	10.00	135.7	100.00
	OpenCode	41.90	5.46	0.85	23.25	426.0	100.00
	Gemini CLI	42.59	5.84	0.39	20.75	410.0	100.00
Gemini 3.7 Flash	Terminus-2	**49.04**	13.61	0.95	44.00	1272.5	100.00
	OpenCode	46.07	9.40	0.88	44.50	1256.1	90.42
	Gemini CLI	46.04	15.55	3.28	93.75	5385.1	92.71
GPT 5.5	Terminus-2	52.96	5.05	0.69	10.50	142.0	100.00
	OpenCode	54.12	6.03	0.86	22.00	558.5	98.96
	Codex	**56.69**	13.48	1.40	32.00	1020.9	100.00
GPT 5.6 Sol	Terminus-2	53.33	4.74	0.57	10.00	159.7	100.00
	OpenCode	52.59	10.55	0.86	25.75	846.0	92.08
	Codex	**56.10**	16.00	1.30	29.75	1192.2	92.92

D.2 Task-Level Performance under Terminus-2

We report task-level performance for all evaluated models under the fixed Terminus-2 harness in Table 4. Unlike the Overall score, which equally averages the four task categories, task-level scores preserve the task-specific rubric weights defined in Appendix C.3. This breakdown shows the performance profile of each model across different microstructural characterization settings. The non-timeout rate (NTR) is reported separately to distinguish successful execution from scientific performance.

Table 4: **Task-level scores and non-timeout rates under Terminus-2.** Task scores retain task-specific rubric weights. The non-timeout rate (NTR) is reported separately to distinguish successful execution from scientific performance. Scores and NTR are reported on a 0–100 scale. Bold marks the highest scientific score in each column.

Model	SEM Static		SEM Temporal		EBSD Shape		EBSD Orientation	
	Score	NTR	Score	NTR	Score	NTR	Score	NTR
Claude Opus 5	**58.74**	100.00	65.26	100.00	**49.02**	100.00	40.50	100.00
GPT 5.6 Sol	55.42	100.00	**66.45**	100.00	47.08	100.00	44.38	100.00
GPT 5.5	51.00	100.00	65.44	100.00	46.64	100.00	**48.74**	100.00
Gemini 3.7 Flash	48.90	100.00	63.52	100.00	44.10	100.00	39.64	100.00
Kimi K3	53.34	100.00	59.69	95.83	42.66	100.00	39.60	90.00
Claude Sonnet 4.6	47.82	100.00	59.77	100.00	41.35	100.00	44.32	100.00
DeepSeek V4 Pro	46.79	100.00	60.46	100.00	36.50	100.00	46.88	100.00
GPT 5.4	42.62	100.00	60.76	95.83	40.04	100.00	43.30	100.00
MiniMax M3	42.32	100.00	63.74	100.00	37.50	100.00	42.29	100.00
Gemini 3.1 Pro	43.82	100.00	57.59	100.00	36.18	100.00	45.76	100.00
Grok 4.5	52.75	100.00	64.18	100.00	22.87	80.00	18.72	90.00
Grok 4.3	35.46	95.00	49.20	100.00	27.38	100.00	31.36	100.00
GLM 5.3	33.11	55.00	64.55	95.83	35.63	60.00	6.13	10.00
Step 3.5 Flash	30.19	90.00	34.10	62.50	26.47	60.00	23.45	60.00
Qwen 3.8 Max	38.99	55.00	44.10	66.67	0.00	0.00	0.00	0.00

D.3 VARIATION ACROSS TASK INSTANCES

We summarize instance-level score variation for all 15 models under Terminus-2 in Table 5. For each task with n instances and scores x_i on a 0–100 scale, we compute the mean and population standard deviation as

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad \sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}. \tag{6}$$

We use the population denominator n because each task is treated as a finite evaluation set. We compute these statistics from the four-decimal instance scores in the result files. They describe differences across instances rather than variability across repeated executions of the same instance.

Table 5: **Score variation across task instances under Terminus-2.** We report mean scores with population standard deviations shown as subscripts. Scores use a 0–100 scale. The model order follows Table 2.

Model	SEM-S $n = 20$	SEM-T $n = 24$	EBSD-S $n = 10$	EBSD-O $n = 10$
Claude Opus 5	$58.74_{\pm 12.10}$	$65.26_{\pm 3.83}$	$49.02_{\pm 6.40}$	$40.50_{\pm 7.05}$
GPT 5.6 Sol	$55.42_{\pm 8.21}$	$66.45_{\pm 3.88}$	$47.08_{\pm 7.28}$	$44.38_{\pm 10.86}$
GPT 5.5	$51.00_{\pm 8.92}$	$65.44_{\pm 4.06}$	$46.64_{\pm 3.39}$	$48.74_{\pm 4.85}$
Gemini 3.7 Flash	$48.90_{\pm 11.35}$	$63.52_{\pm 4.51}$	$44.10_{\pm 4.04}$	$39.64_{\pm 8.75}$
Kimi K3	$53.34_{\pm 7.90}$	$59.69_{\pm 9.71}$	$42.66_{\pm 5.09}$	$39.60_{\pm 13.94}$
Claude Sonnet 4.6	$47.82_{\pm 6.84}$	$59.77_{\pm 5.09}$	$41.35_{\pm 4.62}$	$44.32_{\pm 8.51}$
DeepSeek V4 Pro	$46.79_{\pm 9.57}$	$60.46_{\pm 3.53}$	$36.50_{\pm 4.92}$	$46.88_{\pm 4.19}$
GPT 5.4	$42.62_{\pm 8.97}$	$60.76_{\pm 4.82}$	$40.04_{\pm 6.05}$	$43.30_{\pm 11.01}$
MiniMax M3	$42.32_{\pm 9.35}$	$63.74_{\pm 4.48}$	$37.51_{\pm 5.61}$	$42.29_{\pm 5.64}$
Gemini 3.1 Pro	$43.82_{\pm 8.51}$	$57.59_{\pm 5.24}$	$36.18_{\pm 8.74}$	$45.76_{\pm 7.76}$
Grok 4.5	$52.75_{\pm 7.33}$	$64.18_{\pm 3.45}$	$22.87_{\pm 6.90}$	$18.72_{\pm 10.82}$
Grok 4.3	$35.46_{\pm 9.84}$	$49.20_{\pm 10.92}$	$27.38_{\pm 8.33}$	$31.36_{\pm 13.40}$
GLM 5.3	$33.11_{\pm 26.63}$	$64.55_{\pm 5.10}$	$35.63_{\pm 23.78}$	$6.13_{\pm 18.39}$
Step 3.5 Flash	$30.19_{\pm 10.52}$	$34.10_{\pm 23.44}$	$26.47_{\pm 18.02}$	$23.45_{\pm 20.50}$
Qwen 3.8 Max	$38.99_{\pm 20.03}$	$44.10_{\pm 26.70}$	$0.00_{\pm 0.00}$	$0.00_{\pm 0.00}$

D.4 Validation of GPT-Based Evaluation

We use GPT-5.4 as an LLM-based judge for evaluating scientific descriptions in MatMicroBench. To examine the reliability of this evaluation protocol, we conduct a human-GPT agreement study on a subset of benchmark outputs. We randomly select four benchmark instances covering SEM-S, SEM-T, EBSD-S, and EBSD-O, and evaluate five representative model outputs for each instance. Three independent human evaluators score each output using the same rubric provided to GPT-5.4. The average human score is used as the reference for comparison.

Table 6 presents the results. GPT-5.4 judgments show close agreement with the human evaluations across different tasks and models. The average deviation from the human average is 2.15 points on the original task-specific scoring scale. The remaining differences are comparable to the variation among human evaluators, particularly for tasks involving scientific interpretation. These results indicate that the structured rubric provides effective guidance for GPT-5.4 to produce consistent evaluations. Based on this validation, we use GPT-5.4 as a scalable judge for the full benchmark evaluation. This provides a consistent and reproducible scoring procedure while reducing the cost of large-scale human assessment.

Table 6: **Human-GPT agreement study.** Three independent human evaluators score sampled model outputs using the same rubric as GPT-5.4. The rightmost column reports the GPT-5.4 score and its deviation from the human average in parentheses.

Task ID	Model	Human 1	Human 2	Human 3	Human Avg	GPT-5.4 Judge
SEM-S-01	Claude Opus 5	37.0	37.5	35.5	36.7	40.0 (+3.3)
	GPT 5.6 Sol	35.0	35.0	30.0	33.3	30.0 (-3.3)
	GPT 5.5	36.0	31.0	30.0	32.3	32.0 (-0.3)
	Kimi K3	32.5	32.0	29.0	31.2	28.0 (-3.2)
	Claude Sonnet 4.6	28.5	26.0	19.0	24.5	25.0 (+0.5)
SEM-T-01	Claude Opus 5	16.0	17.5	14.0	15.8	18.0 (+2.2)
	GPT 5.6 Sol	15.0	17.0	13.5	15.2	14.0 (-1.2)
	GPT 5.5	15.5	15.5	14.5	15.2	17.0 (+1.8)
	Kimi K3	15.5	19.0	14.0	16.2	15.0 (-1.2)
	Claude Sonnet 4.6	11.0	11.5	11.0	11.2	8.0 (-3.2)
EBSD-S-01	Claude Opus 5	28.5	26.5	27.0	27.3	26.0 (-1.3)
	GPT 5.6 Sol	25.5	25.5	23.5	24.8	23.0 (-1.8)
	GPT 5.5	29.5	28.5	25.5	27.8	26.0 (-1.8)
	Kimi K3	21.5	21.5	22.0	21.7	19.0 (-2.7)
	Claude Sonnet 4.6	20.0	20.0	17.0	19.0	16.0 (-3.0)
EBSD-O-08	Claude Opus 5	18.5	15.5	15.0	16.3	13.0 (-3.3)
	GPT 5.6 Sol	21.0	17.5	23.5	20.7	22.0 (+1.3)
	GPT 5.5	27.5	28.0	26.0	27.2	29.0 (+1.8)
	Kimi K3	18.0	16.0	14.0	16.0	13.0 (-3.0)
	Claude Sonnet 4.6	19.5	18.0	15.5	17.7	15.0 (-2.7)

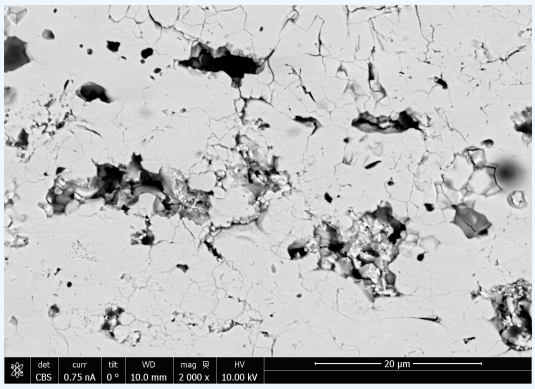
E Representative Benchmark Instances

The following examples illustrate one instance from each task. They summarize the input representations, the requested analysis, and selected expert references for the four artifact types. They are condensed presentations rather than verbatim agent prompts or complete scoring specifications. Reference images and measurements shown below were withheld from the agents. Numerical fractions are on a 0–1 scale and displayed values are rounded. Each *Reason* bullet represents a separately annotated scientific-conclusion item.

SEM-S: Static Defect-State Analysis

Example of a single-image surface observation.

Input and task requirements

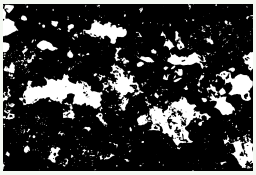


Input SEM image

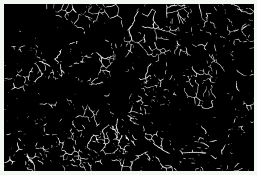
Given this image, its observation type, and task definitions, the agent identifies the valid imaging region and segments pores, cracks, and semi-molten particles. It submits aligned defect masks, area fractions, executable code for the measurements, and an interpretation of defect morphology, spatial relationships, regional issues, and potential risks.

Expert reference: Observe

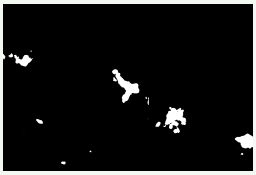
The aligned reference masks delineate the three defect classes within the valid imaging area for this instance.



Pores



Cracks



Semi-molten particles

Expert reference: Model

The reference procedure computes each class area fraction from its mask within the valid imaging region. Submitted code is checked for consistency with the reported fractions and for implementation of the task-defined counting and region conventions.

Expert reference: Measure

Area fraction	Reference value
Pores	0.181
Cracks	0.046
Semi-molten particles	0.024

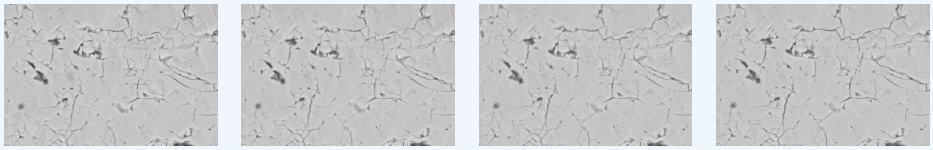
Expert reference: Reason

- **Pore distribution.** Fine dispersed pores coexist with larger irregular, locally clustered, and connected pores.
- **Pore location.** Larger irregular pores appear in localized specimen regions.
- **Crack morphology.** Isolated microcracks and locally connected cracks occur, including cracks along particle boundaries or lamellar interfaces.
- **Crack direction.** No clear overall preferred direction is recorded.
- **Particle distribution.** A small number of semi-molten particles are dispersed through the region.
- **Particle surroundings.** Some particle surroundings show defect enrichment.
- **Defect relationships.** Pores and cracks locally connect, and particles co-occur with pores or cracks.
- **Regional material issue.** The local defect pattern is consistent with insufficient densification.
- **High-risk region and evidence.** Regions containing larger pores are identified as potential concerns based on the observed morphology.
- **Potential failure risk.** The observed defects may contribute to reduced mechanical integrity or local spallation.

SEM-T: Temporal Defect-Evolution Analysis

Example with four observed stages and two held-out target stages.

Input and task requirements

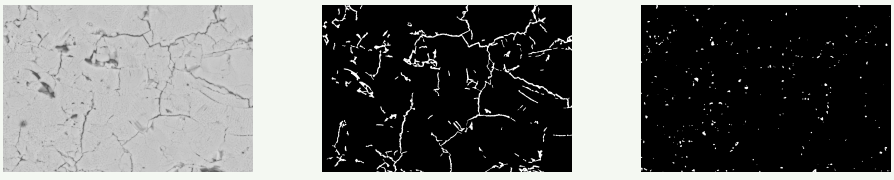


Observed C000 | Observed C009 | Observed C018 | Observed C027

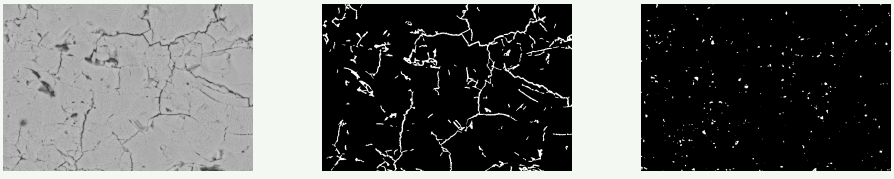
The agent receives repeated SEM observations of the same specimen and field of view. It predicts images at C036 and C045, extracts crack and pore regions from the predicted states, and reports the defect, crack, pore, and newly formed defect fractions for each target. It also describes the overall evolution from the initial observation to the final predicted state, including changes in morphology, spatial distribution, and dominant damage patterns.

Expert reference: Observe

The later experimental observations serve as image references. Corresponding crack and pore masks provide spatial references for the two targets.



C036: reference SEM | C036: cracks | C036: pores



C045: reference SEM | C045: cracks | C045: pores

Expert reference: Model

The defect region is the union of the crack and pore regions. For each target, the newly formed defect fraction counts pixels that become defective relative to the preceding stage. The first predicted stage is compared with the last observed stage; the second is compared with the first predicted stage. The same calculation rules apply to submitted and reference masks.

Expert reference: Measure

Target	R_{defect}	R_{crack}	R_{pore}	R_{new}
C036	0.0711	0.0606	0.0105	0.0125
C045	0.0762	0.0647	0.0115	0.0063

The first three fractions describe each target state; R_{new} describes the change from the preceding stage.

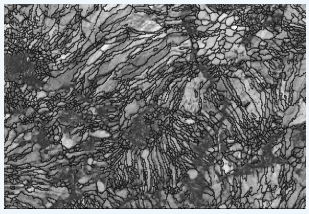
Expert reference: Reason

- **Initial morphology.** The observed C000 image contains a continuous principal crack together with pores.
- **Final morphology.** The held-out C045 image still contains a continuous principal crack, with cracks and pores coexisting.
- **Pore evolution.** Pores remain mostly isolated or locally clustered.
- **Crack evolution.** Some local cracks widen and connect over the sequence.
- **Dominant damage pattern.** Both new defective regions and growth of existing defects contribute to the observed change.
- **High-risk regions.** An upper-central region and a region to the right of center are marked for attention.
- **Crack direction.** Horizontal and oblique components are recorded.

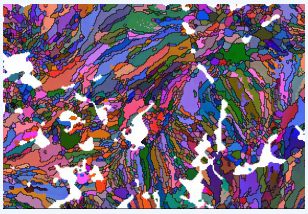
EBSD-S: Inter-Grain Structure Analysis

Example with co-registered EBSD representations of one field of view.

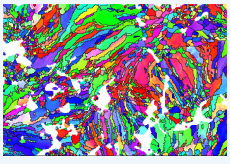
Input and task requirements



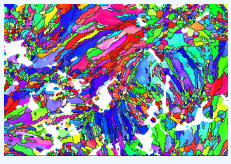
Band-contrast map



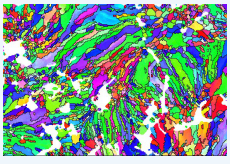
Euler map



IPF-X map



IPF-Y map

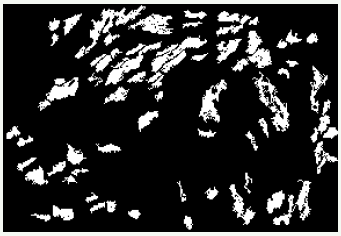


IPF-Z map

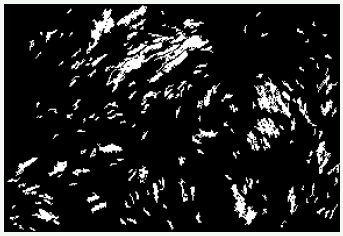
The agent also receives sampling metadata. It identifies valid complete grains, locates coarse and elongated grain regions, and reports grain count, size and area distributions, aspect ratio, and principal-axis orientation. It submits the two region masks, quantitative results, executable calculation code, and an interpretation of morphology, spatial organization, and heterogeneity.

Expert reference: Observe

The two spatial references mark coarse and elongated grains within the valid EBSD region.



Coarse grains



Elongated grains

Expert reference: Model

The reference rules restrict grain statistics to valid complete grains, derive size and shape measures using the sampling scale, and select coarse and elongated grains using the task-defined criteria. Count fractions use the valid complete-grain population; region area fractions use the valid EBSD area. Principal-axis orientation is summarized in angular bins. Submitted code is checked for the required area calculations and consistency with its outputs.

Expert reference: Measure

Selected measurements for this instance are shown below.

Indicator	Reference value
Valid complete grains	2,209
Equivalent diameter: mean / median / P95 (μm)	0.477 / 0.378 / 0.987
Mean aspect ratio	2.139
Coarse-grain count / area fraction	0.048 / 0.148
Elongated-grain count / area fraction	0.123 / 0.100
Dominant principal-axis interval	120°–150°

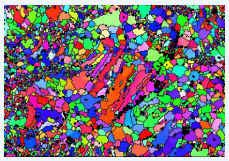
Expert reference: Reason

- **Overall grain size.** The field contains grains over a broad range of sizes.
- **Size distribution.** A pronounced mixture of fine and coarse grains is present.
- **Grain shape.** Grain shapes vary across the field.
- **Principal-axis direction.** The overall directional preference is weak.
- **Coarse-grain location.** Coarse grains are more apparent in the upper region.
- **Elongated-grain location.** Elongated grains occur in the upper and right regions.
- **Spatial relationship.** The coarse and elongated selections partly overlap.
- **Heterogeneity.** The microstructure is spatially heterogeneous.
- **Supporting evidence.** Uneven regional distributions of coarse and elongated grains accompany the measured size and area distributions.

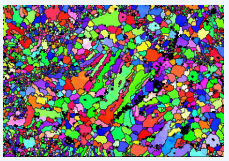
EBSD-O: Intra-Grain Orientation-Variation Analysis

Example with three IPF directions and their corresponding color keys.

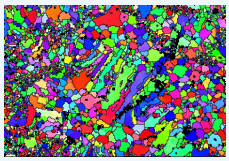
Input and task requirements



IPF-X map

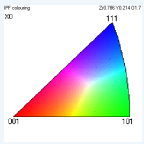


IPF-Y map

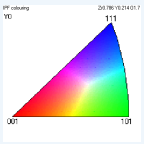


IPF-Z map

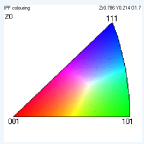
Input and task requirements (continued)



IPF-X color key



IPF-Y color key

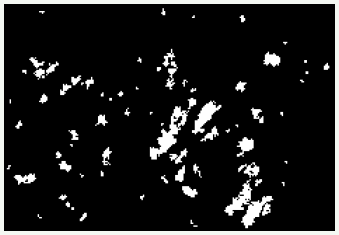


IPF-Z color key

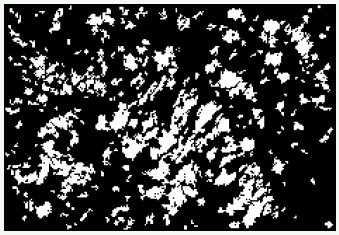
The agent also receives sampling metadata and task definitions. It uses the co-registered IPF representations to identify complete grains with elevated within-grain orientation variation, including a more selective high-variation group. It submits grain-region masks, count and area fractions, executable calculation code, and a description of their spatial distribution and microstructural heterogeneity.

Expert reference: Observe

The expert reference selects grains at two task-defined variation levels. The high-variation selection is contained within the broader elevated-variation selection.



High-variation grains



Elevated-variation grains

Expert reference: Model

The reference procedure applies the task-defined grain selections and computes each count fraction relative to valid complete grains. Each selected-grain area fraction uses the valid EBSD region as its denominator. Submitted code is checked for the required area calculations and consistency between selected regions and reported fractions.

Expert reference: Measure

Indicator	Reference value
Valid complete grains	1,817
High-variation count / area fraction	0.051 / 0.061
Elevated-variation count / area fraction	0.254 / 0.223

Expert reference: Reason

- **Overall variation.** The reference characterizes within-grain orientation variation as relatively high for this field of view.
- **High-variation population.** Only a small fraction of complete grains belongs to the high-variation selection.
- **High-variation location.** Selected grains occur in central and lower-right areas.
- **Spatial pattern.** They form multiple localized clusters.
- **IPF visual evidence.** The maps show noticeable within-grain color changes or gradients in selected regions.
- **Intra-grain variation.** The color changes support a description of variation within individual grains.
- **Heterogeneity.** Orientation variation is unevenly distributed across the field of view.
- **Supporting evidence.** Some high-variation grains form band-like groupings.