FLOWINTENTBENCH: CAN LLM AGENTS ANALYZE FLOW FIELDS TO ANSWER SCIENTIFIC QUESTIONS?

Anonymous authors
Paper under double-blind review

ABSTRACT

Answering scientific questions about numerical flow fields requires choosing an appropriate analysis and determining what the resulting evidence supports. Evaluating LLM agents on this task is challenging because different valid analyses can yield different findings from the same data. We introduce FlowIntentBench, a benchmark of 24 scientific question families across 15 flow-field datasets. Each family is instantiated in four conditions that vary which analysis decisions the question specifies and how findings are to be selected and reported, and is paired with a set of accepted operationalizations (definitions, measurements, and decision rules) together with the findings each yields. Responses are scored on the validity of the stated analysis, on recall of the findings it requires, and on the consistency between findings and operationalization. Across nine tool-using LLM agents, every model scores lower when all principal analysis decisions are left to it. Mean validity falls from 0.846 to 0.742 and recall from 0.910 to 0.863. Our results show that current agents perform reliably when an analysis is prescribed, but struggle when they must determine how to analyze the scientific question.

1 INTRODUCTION

Large language models (LLMs) have extended the use of AI agents to scientific research, where answering a question requires domain knowledge, analysis of scientific data, and conclusions supported by evidence. Flow-field analysis provides a setting for studying these capabilities. Its numerical data describe physical quantities and spatial structures, while questions about those structures often leave their definitions and measurements unspecified. An agent must therefore determine how to analyze the phenomenon of interest and establish what the resulting evidence supports.

Reliable evaluation is needed to assess these capabilities. In flow analysis, FlowNL and FlowLLM support natural-language interaction with flow data (Huang et al., 2023; Li et al., 2025), while SciVisAgentBench evaluates agents on scientific analysis and visualization tasks (Ai et al., 2026). The need to determine how a scientific question should be analyzed connects flow analysis to broader agent benchmarks such as BLADE and HeurekaBench, which evaluate agents on open-ended scientific questions requiring them to develop and execute analyses of the provided data (Gu et al., 2024; Panigrahi et al., 2026). However, systematic evaluation on comparable flow-analysis tasks with different levels of guidance remains limited, leaving unclear how agent performance changes when questions leave more analysis decisions open.

Asking an LLM agent to answer scientific questions about numerical flow fields, as illustrated in Figure 1, raises three distinct challenges. First, a plausible final answer can conceal an inappropriate analysis or unsupported claims, so evaluation must distinguish the validity of the operationalization from the evidential support for the reported scientific findings. Second, underspecified questions leave key definitions, measurements, and decision rules unresolved, making it difficult for an agent to determine which analytical choices appropriately address the scientific target. Third, even when the scientific target and scope are clear, open-ended analysis can admit multiple valid operationalizations and different defensible findings, making a single reference answer insufficient for judging all valid responses. A meaningful benchmark must therefore evaluate the scientific validity of the analysis and the support and sufficiency of its findings while accommodating multiple valid analysis paths.

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.

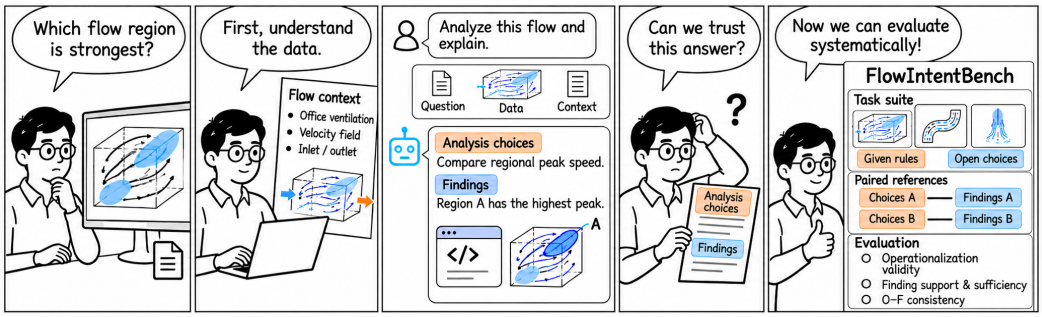


Figure 1: The real-world example. FlowIntentBench helps scientists and researchers evaluate whether an LLM agent's analysis appropriately addresses a scientific question about numerical flow fields and whether its findings are supported by that analysis and sufficient to answer the question.

We introduce FlowIntentBench, a benchmark of 24 scientific question families across 15 flow-field datasets. Each family contains four scientific questions that share the data, scientific target, and scope, yielding 96 cases in total. Three questions progressively leave more analysis decisions to the agent, ranging from a prescribed analysis to one the agent must formulate, while keeping the types of required findings fixed. The fourth retains the prescribed analysis but requires the agent to select findings that jointly provide a sufficient answer. Together, these questions vary the responsibilities agents take on in determining an analysis and selecting the findings to report.

FlowIntentBench pairs each accepted operationalization with findings from its executed reference analysis, preserving the link between analytical choices and the results they support. Evaluation assesses operationalization validity and finding recall and precision, then checks whether reported findings are supported under the agent's declared analysis. We evaluate nine tool-using LLM agents with three runs on each of the 96 questions. When agents formulate analyses under task constraints instead of following prescribed analyses, mean validity falls from 0.846 to 0.742 and required-finding recall from 0.910 to 0.863. These declines identify the transition from following an analysis to defining one as a challenge, reflected in both analysis validity and coverage of required findings. The key contributions of this paper are:

- We introduce FlowIntentBench, with 96 cases in 24 scientific question families across 15 flow-field datasets, to evaluate agents under different analysis and reporting requirements for the same scientific targets.
- We develop an evaluation framework that pairs accepted operationalizations with executed reference findings to assess analysis validity, finding recall and precision, and consistency between findings and the stated operationalization.

2 Related Work

Our work relates to two lines of research. The first, language-guided flow analysis, connects user intent with flow analysis and visualization through natural language. The second, benchmarks for scientific data analysis, evaluates agents' analysis choices and scientific findings. We discuss these two directions in turn.

Language-Guided Flow Analysis. Natural-language interfaces connect user requests to data operations and visual representations. NL4DV identifies attributes and analysis tasks in queries over tabular data and generates visualization specifications (Narechania et al., 2021). In flow analysis, FlowNL provides a declarative language for selecting and combining structures using field values, flow patterns, spatial regions, and connectivity (Huang et al., 2023). FlowLLM supports multi-turn queries, allowing users to refine their requests during exploration (Li et al., 2025). Scientific visualization agents extend this interaction to workflow execution: ChatVis generates Python scripts and revises them using execution errors (Mallick et al., 2024), while ParaView-MCP uses direct API calls and visual feedback to adjust visualization parameters (Liu et al., 2025). Across these systems, the central task is to translate and refine user requests through executable analysis and visualization operations.

Evaluating these workflows requires a clear account of what counts as a correct result. SciVisAgentBench combines expert-written rubrics, reference artifacts, and deterministic checks where available. Its current task design requires a single explicit target outcome and excludes cases with multiple equally valid outputs (Ai et al., 2026). This makes outcome comparison well defined, but leaves out flow questions for which different valid definitions and measurements lead to different findings. Such questions require evaluation to account for both the selected analysis and the findings it supports.

Benchmarks for Scientific Data Analysis. Broader agent benchmarks cover data analysis and modeling in DSBench (Jing et al., 2025), data science code generation in DA-Code (Huang et al., 2024b), computational reproducibility in CORE-Bench (Siegel et al., 2025), and machine learning experimentation in MLAgentBench (Huang et al., 2024a). Scientific benchmarks assess analysis outputs against reference programs, hypotheses, or findings. ScienceAgentBench comprises 102 tasks drawn from 44 peer-reviewed publications across four disciplines and evaluates self-contained Python programs and their execution results (Chen et al., 2025). DiscoveryBench matches the contexts, variables, and relationships in generated and reference hypotheses (Majumder et al., 2025). HeurekaBench compares answers with atomic reference facts derived from reproduced scientific findings, allowing additional findings that do not contradict them (Panigrahi et al., 2026). Assessing whether a finding follows from the agent's stated operationalization requires matching it to evidence produced under that analysis.

BLADE comprises 12 datasets with corresponding research questions from the scientific literature and evaluates multiple justified analyses through reference decisions on conceptual variables, data transformations, and statistical models. Its transformation checks assess equivalence through output values and data-flow structure (Gu et al., 2024). Operationalization-aware data discovery (OADD) connects research questions to measurement choices and supporting data columns (Chen & Jagadish, 2026). FlowIntentBench evaluates both agreement with reference findings and support under the answer's declared operationalization, using paired analyses and findings to assess the claims in a natural-language answer. Its task conditions vary the analysis decisions supplied by the question and the findings the answer must establish.

3 FLOWINTENTBENCH CONSTRUCTION

FlowIntentBench evaluates whether agents select appropriate analyses and report findings supported by them. We first present the benchmark workflow and formulate the task, then describe how we prepare data and physical context, construct related questions, and build ground truth from reviewed and executed reference analyses.

3.1 TASK FORMULATION

FlowIntentBench evaluates how agents turn scientific questions about flow fields into analyses and supported findings. Figure 2 presents the overall workflow. Given a scientific question, 3D field data, and physical context, the agent specifies the definitions, measurements, and decision rules needed to answer the question and reports the resulting findings. Ground truth pairs accepted reference analyses with findings supported by their execution. Evaluation covers three aspects: **operationalization validity**, which assesses analysis choices against task constraints; **scientific findings**, which assesses the support and completeness of reported claims; and **O–F consistency**, which checks whether the findings follow from the stated operationalization.

Each case provides (Q, D, C): a scientific question, 3D field data, and physical context. An operationalization $O = (o_1, \ldots, o_m)$ specifies the definitions, measurements, and decision rules used to answer Q, with each decision prescribed by the question or selected by the agent under the task constraints. The agent returns a natural-language answer stating its operationalization $\hat{O}$ and findings $\hat{\mathcal{F}}$. Each finding is a claim about a quantity or property of D, with a specified scope and measurement definition. Reported findings may be correct, incorrect, or unverified.

Valid operationalizations can yield different findings for the same scientific target and scope. Ground truth comprises K reviewed branches, $GT = \{(O_b, \mathcal{F}_b)\}_{b=1}^{K}$, each pairing an accepted operationalization O_b with findings $\mathcal{F}_b$ supported by executing its reference analysis on D. This finite set may omit valid operationalizations, and findings without sufficient evidence remain unverified. Refer-

ence analyses are reviewed and frozen before evaluation. Section 3.3 describes how related cases vary the supplied analysis decisions and finding requirements.

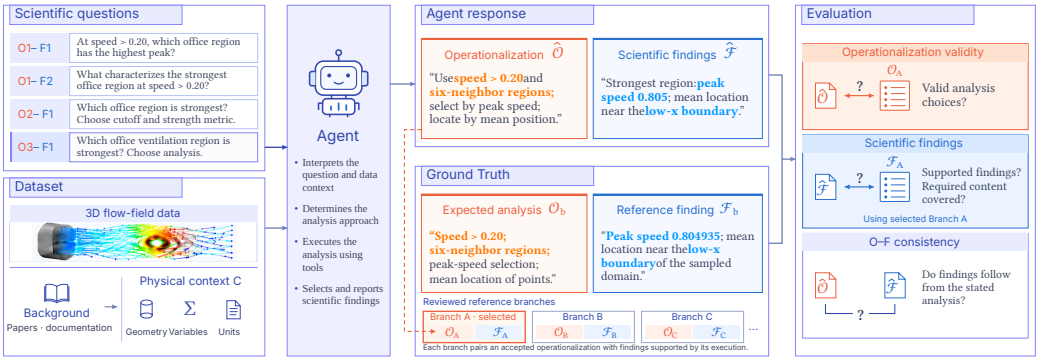


Figure 2: FlowIntentBench workflow. A scientific question, flow-field data, and physical context are provided to the agent, which reports an operationalization and scientific findings. Ground truth pairs accepted operationalizations with supported findings in reference branches. Evaluation assesses operationalization validity, selects a reference branch to assess scientific findings, and checks O–F consistency between the stated analysis and reported findings. Branch A illustrates the selected reference branch. Orange denotes operationalizations and blue denotes scientific findings.

3.2 Dataset and Context Preparation

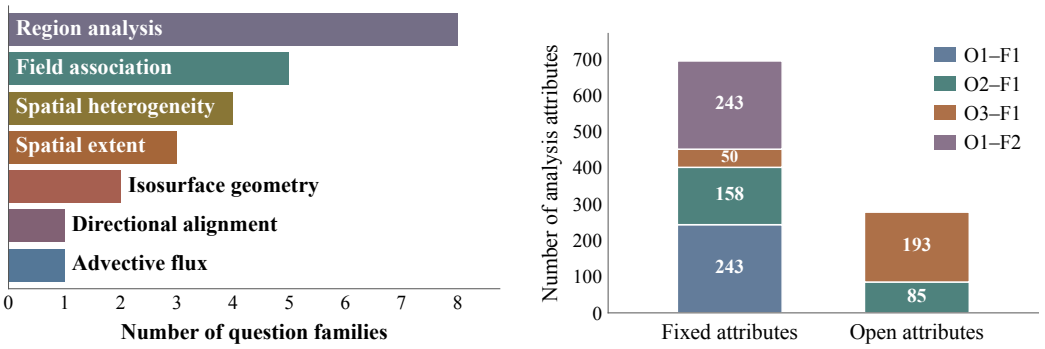
We select 15 three-dimensional field inputs, termed datasets, from the VTK example database and supplementary public scientific sources. Selection requires interpretable fields and physical context, explicit operationalization decisions for a scientific target, and an executable reference analysis. The datasets span aerodynamic, combustion, ventilation, biomedical, industrial, and astrophysical settings on structured and unstructured meshes. Each case supplies one snapshot, loaded through VTK with documented reader or conversion settings. Appendices B.1 and B.2 detail the selection criteria and dataset inventory.

Physical context C is drawn from source-linked documentation and studies, covering documented facts about the study objects, geometry, boundary and operating conditions, and variable meanings relevant to the question. Available fields, point or cell association, and known units connect this context to D. Reader-derived quantities follow documented calculations, while task-specific measurements belong to the operationalization. Reference analyses and findings are reserved for evaluation.

Figure 3(a) summarizes the 24 scientific question families across seven task types: region analysis (8), field association (5), spatial heterogeneity (4), spatial extent (3), isosurface geometry (2), directional alignment (1), and advective flux (1). Families instantiate these analysis types in different physical settings and data representations. For example, high-speed-region analysis appears in office ventilation, carotid flow, and fire flow, with structured grids in Office and Carotid and an unstructured mesh in FireFlow. The shared target requires defining a region and measuring its strength, while its implementation depends on the available fields and mesh connectivity. Field-association families similarly apply association measures to different physical relationships, including density–vertical-velocity and pressure–speed relationships. Each family contributes four task conditions.

3.3 Case Construction

For each scientific question family, we define a target from the available fields and physical context, then specify its scope, principal operationalization dimensions, and finding roles. A role identifies required content, such as a region's location or strength. Each family provides a common scientific target for comparing performance under different analysis and reporting requirements. The four conditions in Table 1 hold the data, target, and scope fixed, yielding 96 cases across 24 families. O1–F1 through O3–F1 progressively delegate analysis decisions to the agent while retaining the required finding roles. O1–F2 retains the prescribed analysis and asks the agent to select findings that satisfy



(a) Scientific task distribution

(b) Decision responsibility distribution

Figure 3: Scientific task coverage and analysis responsibilities. (a) Distribution of 24 question families across seven task types. (b) Question-specified and agent-chosen analysis attributes, with 243 attributes across 24 cases per condition. O1–F1 and O1–F2 share analysis attributes but differ in finding requirements (Table 1).

the family's sufficiency requirements. These comparisons examine how performance changes as agents take responsibility for defining the analysis and determining what evidence constitutes an answer.

Figure 3(b) quantifies this variation across 243 analysis attributes per condition. Open attributes increase from 0 in O1 to 85 in O2 and 193 in O3. O3 retains 50 fixed attributes required by the target, scope, or requested outputs. Appendix A.2 gives the attribute assignments.

Table 1: Four task conditions within each scientific question family. In O1–F2, the selected role combination must include all mandatory roles.

Condition	Operationalization	Finding requirements
O1–F1	All principal decisions specified	Specified roles
O2–F1	1–2 principal decisions left open	Same as O1–F1
O3–F1	Agent-formulated under task constraints	Same as O1–F1
O1–F2	Same as O1–F1	Agent selects sufficient roles

We select the open O2 dimensions before question generation. An LLM writes questions from the family specifications. Metadata validation and LLM review check that original and revised wording preserves the target, scope, finding requirements, and allocation of analysis decisions. Appendices C–E provide case examples, paired questions, and construction details.

3.4 Ground Truth Construction

We construct ground truth by reviewing the scientific validity of operationalizations and executing accepted analyses to obtain supported reference findings. For O1, we review the prescribed operationalization; for O2–F1 and O3–F1, we derive candidates from each family's specification. An LLM checks their scientific appropriateness, data availability, and consistency with the target and scope against source evidence and dataset context. Each operationalization must specify all principal decisions and respect those fixed by the question. Accepted analyses are executed using the data and reader conventions supplied to the agent. Each reference branch records the operationalization, supported findings, computed evidence, and comparison policies specifying the required reporting resolution. These policies are assigned during reference construction and frozen before answer evaluation. Appendix F details the reference records and verification policies.

For F2, we define mandatory finding roles and acceptable role combinations for each family. Each combination includes all mandatory roles and is reviewed for sufficient coverage of the scientific target and support under the specified analysis. A complete F2 answer establishes at least one

acceptable combination through supported findings. Finding requirements and reference evidence are frozen before evaluation and remain fixed across all answers in a comparison.

4 Evaluation Protocol

We first describe how an agent's stated operationalization and scientific findings are extracted and verified using task constraints and reference evidence. We then define metrics for operationalization validity, finding precision and recall, and O–F consistency, together with rules for score availability and aggregation.

4.1 Evaluation Procedure

Evaluation begins by identifying the stated operationalization and scientific findings in an agent's natural-language answer. An LLM evaluator extracts these elements and links them to the corresponding passages in the answer. Findings from the main analysis and explicitly declared supplementary analyses are placed in separate groups, with repeated claims counted once. Analysis choices are identified from the agent's description, rather than inferred from agreement with reference values.

The evaluator assesses fixed operationalization attributes against the question's constraints and open attributes against scientific appropriateness for the target and available data. An open decision can be scientifically appropriate even when no stored branch implements it.

For each finding group, the evaluator maps findings to candidate reference statements of the same quantity, scope, and statistic. Numerical claims are checked programmatically against frozen reference values and units at the reporting resolution specified for each finding (Appendix F.3). Compatibility with the group's declared analysis decisions is checked separately, distinguishing support for a result from consistency with the method attributed to it. Appendix G gives the implementation, reviewer instructions, and a worked check.

4.2 Evaluation Metrics

We quantify the three evaluation aspects using five metrics. The operationalization score S_O measures overall analysis validity, while the unresolved-choice score (URS) focuses on dimensions designated for agent decisions. Finding recall R_F measures coverage of required content, and precision P_F measures support for reported findings. The consistency score C_{OF} measures whether findings are supported under the agent's declared operationalization.

Operationalization scores and recall evaluate the main analysis. Precision and consistency also include task-relevant supplementary findings, which cannot satisfy missing requirements in the main answer.

Operationalization validity. Each of the m applicable principal dimensions receives a score $s_j \in [0, 1]$, averaging normalized atomic-attribute grades (Appendix A.1). The operationalization score is $S_O = \frac{1}{m}\sum_{j=1}^{m} s_j$. URS, applicable to O2 and O3, averages complete dimension scores over the condition-designated set J_{open}, including retained fixed constraints: $\text{URS} = \frac{1}{|J_{\text{open}}|}\sum_{j\in J_{\text{open}}} s_j$.

Scientific findings. For reference branch b with findings $\mathcal{F}_b$, let $N_{\text{matched},b}$ and $N_{\text{core},b}$ count matched and required core findings. For F2, let $\mathcal{A}$ contain the acceptable role combinations and $\hat{A}_b$ the roles supported by the main answer under b. Recall is

$$R_{F,b} = \begin{cases} N_{\text{matched},b}/N_{\text{core},b}, & \text{F1}, \\ \max\limits_{A_k\in\mathcal{A}} \dfrac{|A_k \cap \hat{A}_b|}{|A_k|}, & \text{F2}, \end{cases} \qquad R_F = \max_b R_{F,b}. \tag{1}$$

F2 reaches full recall when an acceptable combination, including all mandatory roles, is fully covered.

For each analysis group, we select the branch with highest recall, breaking ties by precision. One branch applies to the group as a whole, preventing combinations of incompatible branches. Let

$\hat{\mathcal{F}}_{\text{all}}$ contain all distinct task-relevant findings, retaining their group attribution. Precision is $P_F = N_{\text{supported}}/|\hat{\mathcal{F}}_{\text{all}}|$, where $N_{\text{supported}}$ counts findings supported by their groups' selected branches or compatible supplementary evidence.

O–F consistency. O–F consistency checks whether reported findings are supported under the agent's declared analysis. We first collect findings from groups whose principal analysis decisions are all identifiable into $\hat{\mathcal{F}}_{\text{bound}}$. Among these findings, $N_{\text{consistent}}$ counts those supported by frozen evidence compatible with the group's declared analysis. The consistency score is therefore $C_{OF} = N_{\text{consistent}}/|\hat{\mathcal{F}}_{\text{bound}}|$. Method-binding coverage, $|\hat{\mathcal{F}}_{\text{bound}}|/|\hat{\mathcal{F}}_{\text{all}}|$, measures how much of the answer can enter this check: the fraction of all findings linked to an identifiable declared analysis.

Unverified or inapplicable scores are recorded as N/A and excluded from the corresponding means. Confirmed violations and missing required items receive zero for those items. Eligible cases receive equal weight within each condition and run, with applicable condition means weighted equally for overall summaries. We report means and sample standard deviations across three runs. Appendix A details grading, boundary cases, branch diagnostics, and aggregation.

5 Experiments

5.1 Experimental Setup

We evaluate nine LLM agents on the 96 cases from Section 3.3, with 24 cases per task condition. The underlying models are GPT-6 Astra, GPT-5.6 Sol, GPT-5.6 Luna, GPT-5.6 Terra, Claude Fable 5.1, Claude Sonnet 5, DeepSeek-V4.1-Flash, Qwen3.8-Max, and Gemini 3.8 Flash. Each agent is assigned three runs per case, yielding 2,592 planned trials.

All 2,592 main-experiment runs use the network-isolated `flow-python-v1` runtime on Linux with Python 3.12 and frozen package versions. All agents use a common Python tool with NumPy, SciPy, Matplotlib, and VTK. Inputs comprise (Q, D, C) and data-reading metadata, excluding reference analyses and ground truth. Each trial has an independent runtime and workspace, with Python state retained between calls. Agents can iteratively execute code and inspect textual outputs and errors. Generated plots are saved but are not returned as image inputs. Instructions prohibit network use and host-filesystem searches. Agents are instructed to resolve unspecified analysis decisions and report operationalizations and scientific findings separately in the final natural-language answer. Appendices B.5 and H document the tool interface, execution controls, package versions, run configurations, and resource accounting.

Answers are evaluated against frozen reference evidence using the protocol in Section 4. We report performance by task condition, distinguishing answer completion and answer-level scoring coverage from quality scores.

5.2 Benchmark Results

Table 2 summarizes overall scores from 2,541 evaluation reports, covering 98.0% of the 2,592 completed runs. Appendix H.4 provides per-model scoring coverage.

Models show different strengths across the evaluation aspects. Claude Fable 5.1 achieves the highest operationalization score, URS, and finding recall, while GPT-6 Astra leads in finding precision and O–F consistency. Claude Fable 5.1's higher recall (0.9333 versus 0.8574) accompanies lower precision than GPT-6 Astra's (0.9022 versus 0.9463), separating required-content coverage from support for reported findings. DeepSeek-V4.1-Flash combines recall of 0.8902 with an operationalization score of 0.6509. Its comparatively high recall and low analysis score illustrate why finding coverage alone gives an incomplete account of performance.

The O1–O3 comparison highlights the difficulty of formulating an analysis from a scientific question (Table 3). Mean operationalization quality decreases from 0.8456 under O1–F1 to 0.8204 under O2–F1 and 0.7421 under O3–F1, a total decline of 10.35 percentage points. Over the same comparison, recall, precision, and consistency decrease by 4.69, 4.70, and 4.59 percentage points. These declines indicate difficulties both in determining an appropriate analysis and in reporting supported

Table 2: Overall results over three runs, with models ordered by descending mean S_O. Scores show mean ± sample standard deviation of run summaries, with conditions weighted equally and URS restricted to O2–F1 and O3–F1. Higher is better (↑). Shading scales linearly within each column from white (minimum mean) to peach (maximum mean). Bold marks the highest mean per metric. Answer-level scoring coverage is reported in Appendix H.4, Table 26.

Model	S_O ↑	URS ↑	R_F ↑	P_F ↑	C_{OF} ↑
Claude Fable 5.1	**0.9463** ± 0.004	**0.8275** ± 0.011	**0.9333** ± 0.007	0.9022 ± 0.007	0.9341 ± 0.009
GPT-6 Astra	0.9271 ± 0.010	0.8024 ± 0.011	0.8574 ± 0.004	**0.9463** ± 0.006	**0.9683** ± 0.008
GPT-5.6 Sol	0.8835 ± 0.009	0.7872 ± 0.007	0.8756 ± 0.003	0.8714 ± 0.011	0.9231 ± 0.003
Qwen3.8-Max	0.8341 ± 0.007	0.7639 ± 0.013	0.8562 ± 0.004	0.9054 ± 0.002	0.8766 ± 0.007
GPT-5.6 Terra	0.8244 ± 0.013	0.6905 ± 0.025	0.8934 ± 0.023	0.8876 ± 0.016	0.7982 ± 0.020
Gemini 3.8 Flash	0.7809 ± 0.014	0.7001 ± 0.026	0.9163 ± 0.012	0.8502 ± 0.014	0.7693 ± 0.013
Claude Sonnet 5	0.7265 ± 0.009	0.6336 ± 0.016	0.8635 ± 0.019	0.7908 ± 0.005	0.7089 ± 0.007
GPT-5.6 Luna	0.6908 ± 0.023	0.5834 ± 0.007	0.8839 ± 0.027	0.7309 ± 0.009	0.6592 ± 0.006
DeepSeek-V4.1-Flash	0.6509 ± 0.016	0.7401 ± 0.021	0.8902 ± 0.012	0.8825 ± 0.007	0.7784 ± 0.015

findings consistent with that analysis. The URS values of 0.7087 under O2 and 0.7421 under O3 summarize different dimension sets: selected open dimensions in O2 and all principal dimensions in O3 (Appendix A.3).

Table 3: Condition-level results, averaging each model's three-run condition mean equally across the nine models. URS applies to O2–F1 and O3–F1 and equals S_O in O3–F1, where its dimension set contains all principal dimensions.

Condition	S_O	URS	R_F	P_F	C_{OF}
O1–F1	0.8456	N/A	0.9100	0.8878	0.8476
O2–F1	0.8204	0.7087	0.8922	0.8693	0.8301
O3–F1	0.7421	0.7421	0.8631	0.8408	0.8017
O1–F2	0.8205	N/A	0.8768	0.8542	0.8166

Finding selection presents a further challenge under a prescribed analysis. Under O1–F2, mean recall falls from 0.9100 to 0.8768 and precision from 0.8878 to 0.8542 relative to O1–F1. Operationalization quality and consistency also decrease, by 2.51 and 3.10 percentage points. Thus, lower performance extends beyond finding coverage to the support and consistency of the reported claims, despite the prescribed operationalization. F2 combines greater responsibility for finding selection with different authored sufficiency requirements, so this contrast assesses both aspects of the reporting condition.

Mean elapsed time per completed run ranges from 195.06 to 797.79 seconds across models. It includes Python execution and request-related waiting, and excludes scientific scoring and earlier unfinished attempts. Appendix H.3 provides runtime and cumulative token statistics.

5.3 User Study

Evaluator selection and expert agreement. We compared nine candidate evaluators by their Pearson correlation with expert scores and reported cost per answer (Figure 4). GPT-6 Astra achieved the highest observed correlation ($r = 0.89$), followed by Claude Fable 5.1 ($r = 0.85$), at similar reported costs of 11.68 and 11.50 per answer, respectively. The remaining evaluators had correlations between 0.61 and 0.77. We selected GPT-6 Astra with expert agreement as the primary criterion. Appendix I documents human review of individual scoring judgments and the resulting corrections.

Participant performance with interactive support. Fifteen doctoral students with backgrounds in fluid mechanics or scientific visualization each completed 28 randomly sampled benchmark cases,

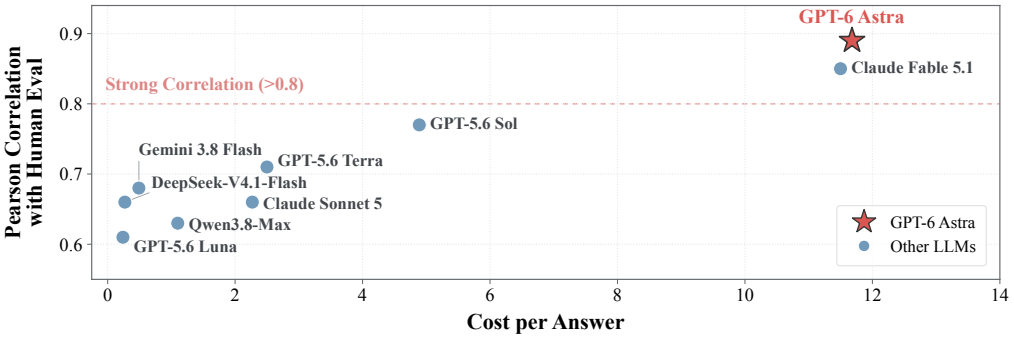


Figure 4: Pearson correlation with expert scores versus reported cost per answer for nine candidate evaluators. The star marks the selected evaluator, GPT-6 Astra; the dashed line marks $r = 0.8$.

seven per condition. Participants used an interactive 3D interface to select formulas, adjust parameters, and identify findings. The interface provided the same formula library and controls across O1–O3, without preselected or recommended analyses, and assembled editable answers from participants' choices and computed quantities without access to ground truth. Answers were scored with the same evaluator as model answers. Appendix J details participant training, the interface, task assignment, and scoring.

With participants weighted equally, mean operationalization quality decreased from 0.959 under O1–F1 to 0.847 under O3–F1, while finding recall decreased from 0.932 to 0.893. Under O1–F2, recall and precision decreased to 0.882 and 0.893, respectively, from 0.932 and 0.948 under O1–F1. All 15 participants showed decreases in each of these four comparisons. The comparisons are descriptive: sampled case sets can differ across conditions, and F2 changes both finding-selection responsibility and authored sufficiency requirements. Table 30 in Appendix J.4 reports the full results.

6 Conclusion and Limitations

FlowIntentBench evaluates agents' analysis choices and findings across 24 question families and 15 datasets. Four conditions vary analysis guidance and reporting requirements for each scientific target. Evaluation pairs accepted operationalizations with executed reference findings. Across nine models, mean operationalization quality and finding recall declined from O1 to O3 and were also lower under O1–F2 than O1–F1. Model rankings differed across analysis validity, finding recall and precision, and O–F consistency. Participants using interactive analysis support showed similar trends, with lower operationalization quality under O3 and lower recall and precision under F2. Together, these results identify the transition from following a prescribed analysis to defining an appropriate analysis as a challenge for the evaluated agents, and motivate assessing analysis validity alongside finding support and O–F consistency.

Limitations. FlowIntentBench covers spatial analysis of individual flow-field snapshots. Its 24 question families reuse analysis patterns across datasets and concentrate on region analysis and field association. Directional alignment and advective flux each have one family, providing limited evidence of generalization across physical settings. Temporal analysis is outside the current task set. Semantic judgments use GPT-6 Astra, which is also an evaluated agent. The source-model-stratified blinded audit in Appendix I checks these judgments, with coverage limited to 96 answers from three models. Finite reference branches leave some alternative analyses unverified. Metric means exclude unverified and inapplicable scores and can therefore summarize different subsets of answers. The participant study uses sampled cases and an interface with formula selection, automated computation, and 3D visualization, limiting direct comparison with agents using Python tools and text-only feedback. Future extensions include additional alignment and flux tasks, temporal questions, vortex identification, separation-line extraction, and streamline analysis, together with expanded executed reference coverage.

AI USE STATEMENT

We used LLMs for dataset context and question generation, review of questions and candidate operationalizations, semantic answer evaluation, and benchmark testing (Sections 3 and 4). Reference findings come from executed analyses, with programmatic numerical verification during scoring. AI assistants also supported manuscript drafting and revision, experimental reporting, result interpretation, aggregation and table-generation scripts, and pipeline figure editing. We take responsibility for the final text, claims, and artifacts. Appendix I documents human review of sampled answers.

ETHICS STATEMENT

The institutional ethics committee approved the user study (Section 5.3). All 15 participants took part voluntarily without compensation and consented to participation and data use. Response data were anonymized, and results are reported as aggregate statistics.

REPRODUCIBILITY STATEMENT

Appendices B.2, B.5, and E document dataset sources, input conventions, the execution interface, and case construction. Appendices F and G describe reference records and evaluation procedures. Sections 5.2 and 5.3 specify result aggregation, N/A handling, and the participant protocol. Appendices H and K.3 describe run configurations and the organization of inputs, references, and evaluation outputs.

REFERENCES

Kuangshi Ai, Haichao Miao, Kaiyuan Tang, Nathaniel Gorski, Jianxin Sun, Guoxi Liu, Helgi I Ingolfsson, David Lenz, Hanqi Guo, Hongfeng Yu, et al. Scivisagentbench: A benchmark for evaluating scientific data analysis and visualization agents. *arXiv preprint arXiv:2603.29139*, 2026.

B. Burkhart, S. M. Appel, S. Bialy, J. Cho, A. J. Christensen, D. Collins, Christoph Federrath, D. B. Fielding, D. Finkbeiner, A. S. Hill, et al. The catalogue for astrophysical turbulence simulations (CATS). *The Astrophysical Journal*, 905(1):14, 2020.

William H. Cabot and Andrew W. Cook. Reynolds number effects on Rayleigh–Taylor instability with possible implications for type Ia supernovae. *Nature Physics*, 2(8):562–568, 2006.

Houming Chen and HV Jagadish. From research questions to columns: Operationalization-aware data discovery. *arXiv preprint arXiv:2608.04536*, 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 *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=6z4YKr0GK6.

Drummond B. Fielding, Eve C. Ostriker, Greg L. Bryan, and Adam S. Jermyn. Multiphase gas and the fractal nature of radiative turbulent mixing layers. *The Astrophysical Journal Letters*, 894(2): L24, 2020.

Oleg Grauberger and Grigorios Tzionis. Metadata design optimisation dataset. Zenodo, version 1.0, 2025. URL https://zenodo.org/records/15720265.

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.

Pooria Hadikhani. Investigating bubble-induced overpotential, current non-uniformity, and bubble distribution in flow-based water electrolyzers: A numerical study. Zenodo, version v2, 2024. URL https://zenodo.org/records/14935434.

Jieying Huang, Yang Xi, Junnan Hu, and Jun Tao. Flownl: Asking the flow data in natural languages. *IEEE Transactions on Visualization and Computer Graphics*, 29(1):1200–1210, 2023. doi: 10.1109/TVCG.2022.3209453.

Qian Huang, Jian Vora, Percy Liang, and Jure Leskovec. MLAgentBench: Evaluating language agents on machine learning experimentation. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 20271–20309. PMLR, 2024a. URL https://proceedings.mlr.press/v235/huang24y.html.

Yiming Huang, Jianwen Luo, Yan Yu, Yitong Zhang, Fangyu Lei, Yifan Wei, Shizhu He, Lifu Huang, Xiao Liu, Jun Zhao, and Kang Liu. DA-Code: Agent data science code generation benchmark for large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 13487–13521. Association for Computational Linguistics, 2024b. doi: 10.18653/v1/2024.emnlp-main.748. URL https://aclanthology.org/2024.emnlp-main.748/.

Liqiang Jing, Zhehui Huang, Xiaoyang Wang, Wenlin Yao, Wenhao Yu, Kaixin Ma, Hongming Zhang, Xinya Du, and Dong Yu. DSBench: How far are data science agents from becoming data science experts? In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=DSsSPr0RZJ.

Zilin Li, Weihan Zhang, and Jun Tao. Flowllm: Large language model driven flow visualization. *Visual Informatics*, 9(3):100241, 2025. doi: 10.1016/j.visinf.2025.100241.

Shusen Liu, Haichao Miao, and Peer-Timo Bremer. Paraview-mcp: An autonomous visualization agent with direct tool use. In *2025 IEEE visualization and visual analytics (VIS)*, pp. 61–65. IEEE, 2025.

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 *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=vyflgpwfJW.

Tanwi Mallick, Orcun Yildiz, David Lenz, and Tom Peterka. Chatvis: Automating scientific visualization with a large language model. In *SC24-W: workshops of the international conference for high performance computing, networking, storage and analysis*, pp. 49–55. IEEE, 2024.

Arpit Narechania, Arjun Srinivasan, and John Stasko. Nl4dv: A toolkit for generating analytic specifications for data visualization from natural language queries. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):369–379, 2021. doi: 10.1109/TVCG.2020.3030378.

Ruben Ohana, Michael McCabe, Lucas Meyer, Rudy Morel, Fruzsina Agocs, Miguel Beneitez, Marsha Berger, Blakesley Burkhart, Stuart Dalziel, Drummond Fielding, et al. The well: A large-scale collection of diverse physics simulations for machine learning. *Advances in Neural Information Processing Systems*, 37:44989–45037, 2024. URL https://polymathic-ai.org/the_well/.

Siba Smarak Panigrahi, Jovana Videnović, and Maria Brbic. HeurekaBench: A benchmarking framework for AI co-scientist. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=Y7xCdFuFw7.

Zachary S. Siegel, Sayash Kapoor, Nitya Nadgir, Benedikt Stroebl, and Arvind Narayanan. CORE-Bench: Fostering the credibility of published research through a computational reproducibility agent benchmark. *Transactions on Machine Learning Research*, 2025. URL https://openreview.net/forum?id=BsMMc4MEGS.

Fernando Zigunov. Volumetric velocity/pressure measurements of a supersonic double-fin flow at M=2 and alpha=10deg. Zenodo, version v1, 2024. URL https://zenodo.org/records/10697026.

APPENDIX GUIDE

The appendix develops three aspects of FlowIntentBench: how scientific questions become analysis tasks, how answers are assessed, and how the experiments support the reported results. Appendix A specifies the scoring rules. Appendices B–F describe the datasets, illustrative cases, task design, and reference analyses. Appendix G follows an answer through evaluation. Appendices H–J present the experimental setup, additional result summaries, human review, and participant study. Appendix K indexes the instructions and supporting materials.

A EVALUATION DETAILS

A.1 GRADED OPERATIONALIZATION QUALITY

Each applicable operationalization dimension d has a predefined set of atomic attributes A_d. For resolved judgments, attribute a receives a grade $g_{da} \in \{0, 1, 2, 3, 4\}$. The dimension score is

$$s_d = \frac{1}{|A_d|} \sum_{a \in A_d} \frac{g_{da}}{4}. \tag{2}$$

Attributes receive equal weight within a dimension, and dimensions receive equal weight in S_O and within the set selected by URS. These graded scores take values in $[0, 1]$ determined by the finite attribute levels. For example, normalized grades $(1, 0.5, 1)$ give $s_d = 5/6$, a quality score rather than a probability of correctness.

The 96 cases use the five dimension types in Table 4, with applicable attributes specified for each task. Each requirement is assigned to one attribute, so a single defect, such as an incorrect isovalue, receives its deduction at that attribute.

Table 4: Atomic-attribute content for the five operationalization dimension types. Task specifications determine the applicable attributes and their fixed or open status.

Dimension	Attribute content
Feature definition	`identity`: physical quantity and object; `domain`: initial analysis population and exclusions; `construction`: field association, transformation, interpolation, or connectivity.
Criterion	`rule`: selection operation; `cutoff`: threshold or generation rule and statistical population; `boundary`: comparison boundaries and tie handling.
Parameter	`value`: value or generation rule; `definition`: field, scale, and necessary conventions; `validity`: scientific validity of the choice.
Property measure	`quantity`: mathematical definition; `parameters`: necessary parameters and operator conventions; `semantics`: sign, dimensions, and physical meaning.
Aggregation or representation	`weights`: weights and aggregation population; `normalization`: denominator and its domain; `representation`: definition of location, width, or other reported properties.

Fixed and open obligations. Fixed attributes are assessed against the question. Open attributes require a clear, scientifically appropriate choice for the target. Equivalent formulations receive equal credit. A fixed condition may be taken directly from the question when it has one clear interpretation and the answer is compatible with it. Missing open choices receive zero. Scientifically appropriate alternatives can receive credit beyond the stored branches. Mixed dimensions retain both kinds of obligation, such as a fixed signed mass-flux quantity with an open calculation order. Grades measure how well the requirement is met. For example, when P90 is prescribed, choosing P80 is assessed as a threshold error, not as $80/90$ credit. A requested justification is assessed separately from method completeness to avoid duplicate deductions.

Missing, ambiguous, and conflicting evidence. A missing required open choice receives zero and remains in the denominator. Concise, unambiguous statements can receive full credit. Ambiguity that affects the analysis receives partial credit, normally at most 0.5 after normalization.

Table 5: Five-level grading rubric for each atomic attribute. Intermediate dimension scores arise by averaging these normalized grades.

g_{da}	$g_{da}/4$	Judgment criterion
4	1.00	All substantive requirements are satisfied by explicit evidence or fixed question conditions with one clear interpretation.
3	0.75	A specific minor defect remains, while the main definition and scientific meaning hold.
2	0.50	Important parts hold, but a consequential ambiguity, omission, or partial error affects the result.
1	0.25	Little usable content remains and the core definition largely fails.
0	0.00	Required content is entirely missing, or the attribute is wholly incorrect, unclear in the answer, or unusable.

When conflicting declarations leave an attribute's operative definition indeterminate, that attribute receives zero; independent correct attributes retain their grades. A minor labeling defect with an identifiable calculation is assessed under the grade-3 criterion in Table 5. Evaluation uses the final declared method, including compatible fixed conditions taken from the question. If execution settings conflict with the declaration, the declaration remains the basis for method–result consistency. Evaluator uncertainty is recorded as `null`, distinct from a demonstrated omission; inapplicability is predetermined by the task card (Appendix A.4).

A.2 Atomic Attribute Inventory and Responsibility Assignment

The task cards assign 972 applicable attributes to 96 cases in 24 scientific question families: 694 fixed and 278 open. Each dimension contains the three attributes in Table 4. Figure 5 shows their distribution, including 24 case–dimension pairs with both fixed and open attributes across 20 cases.

Attribute grading follows the question's target, scope, and stated constraints. An open dimension may retain a fixed field, sign, domain, or requested output, as the flux example in Figure 5 illustrates. Table 6 specifies the assignments used for all compared answers.

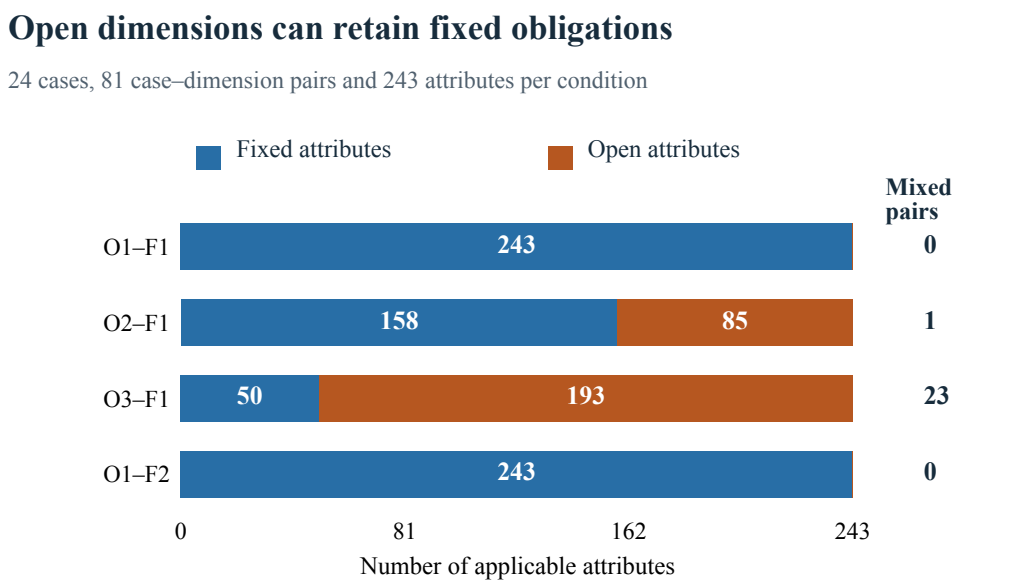


Figure 5: Fixed and open attribute responsibilities across task conditions, reconstructed from all 24 family rows. A mixed case–dimension pair contains both fixed and open attributes. O3 retains 50 fixed attributes despite selecting every applicable dimension for URS. The radiative-flux O2 example shows why retained fixed attributes still contribute to a designated dimension.

Table 6: Atomic responsibilities for all 24 scientific question families. D, C, P, M, and A denote feature definition, criterion, parameter, property measure, and aggregation or representation. F/O marks a fixed/open attribute in the order of Table 4; a dash denotes an inapplicable dimension. O1–F1 and O1–F2 use FFF for every applicable dimension. In O2, the listed dimensions use OOO and all others use FFF, except the property measure for the advective-flux task on Radiative Mixing Layer, which uses FOF in both O2 and O3. The five rightmost columns give the O3 assignments.

Family	O2 set	O3 attribute assignments				
		D	C	P	M	A
AIDEAS: pressure nonuniformity	M	OFO	–	–	OOO	OOO
AIDEAS: viscosity–shear relation	M	OFO	–	–	OOO	OOO
AIDEAS: high-shear region	C	OFO	OOO	–	FFF	FFF
Rayleigh–Taylor: mixing width	M	OFO	–	–	OOO	OOF
Rayleigh–Taylor: density–motion	M	OFO	–	–	OOO	OOO
Rayleigh–Taylor: surface geometry	P	OFO	–	OOO	FFF	FFF
Radiative: density nonuniformity	M	OFO	–	–	OOO	OOO
Radiative: density–pressure	M	OFO	–	–	OOO	OOO
Radiative: advective flux	M	OFO	–	–	FOF	OOO
MHD: density–magnetic relation	M	OFO	–	–	OOO	OOO
MHD: velocity–magnetic alignment	M	OFO	–	–	OOO	OOO
Double Fin: high-pressure region	C	OFO	OOO	–	FFF	FFF
Electrolyzer: speed extent	M	FFO	–	–	OOO	OOF
Electrolyzer Electric: current region	C	OFO	OOO	–	FFF	FFF
Tubes: energy nonuniformity	M	OFO	–	–	OOO	OOO
Tubes: pressure–speed relation	M	OFO	–	–	OOO	OOO
Tubes: speed extent	M	FFO	–	–	OOO	OOF
Blunt Fin: speed region	C, M	OOO	OOO	–	OOO	OOO
Carotid: speed region	C, M	OOO	OOO	–	OOO	OOO
Combustor: density surface	C	OOO	OOO	–	–	OOO
FireFlow: speed region	C, M	FOO	OOO	–	OOO	OOO
Kitchen: concentration comparison	M	OOO	–	–	OOO	OOO
NASA LOx Post: speed region	C, M	FOO	OOO	–	OOO	OOO
Office: ventilation region	C, M	OOO	OOO	–	OOO	OOO

A.3 Dimension-Level URS and Retained Constraints

The term "unresolved choice" refers to the task-condition designation, independently of evaluator uncertainty.

URS averages the complete scores of the condition-designated dimensions J_{open}: none for O1, the one or two listed dimensions for O2, and all applicable dimensions for O3. Attribute labels specify the fixed or open requirement used for grading. Every attribute in a selected dimension contributes to URS, including its fixed constraints.

O3 retains 50 fixed attributes. Requested outputs can even determine an entire dimension: the AIDEAS high-shear task requires a volume-weighted mean and centroid, and the Rayleigh–Taylor surface task requires area and an area-weighted centroid. These dimensions remain in the O3 scoring set. With $\mathcal{D}$ denoting the set of applicable dimensions,

$$\mathrm{URS}_{\mathrm{O3}} = \frac{1}{|\mathcal{D}|} \sum_{d \in \mathcal{D}} s_d = S_O. \tag{3}$$

This equality follows from identical dimension sets and weights, including the retained fixed attributes. O2 and O3 select different sets, as discussed in Section 5.2.

As an illustrative calculation, consider the radiative advective-flux O2 task, where URS includes the property-measure dimension. Its quantity and semantics are fixed, while its parameters are open. Hypothetical grades $(1, 0.5, 1)$ yield $s_d = \mathrm{URS} = 5/6$; grades $(1, 0, 1)$ yield $2/3$.

A.4 Unresolved Attribute Grades and Score Availability

Missing required content is an answer defect, whereas an unavailable judgment reflects insufficient evaluation evidence. Table 7 distinguishes these cases. A method can receive a quality grade from its scientific definition even when its numerical findings lack a matching reference.

Table 7: Attribute assessment states. A missing required choice in the answer is assessed as zero; a missing attribute record in the evaluator output is a structural failure.

State	Grade and treatment
Assessed	Integer grade 0–4, including zero for a demonstrated omission or error. The attribute remains in the predefined denominator.
Unassessable	`null`: evidence is insufficient for a reliable grade. The attribute remains in the denominator, with possible normalized value in $[0, 1]$.
Not assessed	`null`: evaluation has not been performed. The attribute remains in the denominator; a completed point score is unavailable.
Not applicable	No applicable scoring item is created. Exclusion is specified by the task card before grading.
Missing record	A required evaluator entry is absent. Validation rejects the incomplete record instead of inserting zero or `null`.

For a dimension with $n_d = |A_d|$ applicable attributes, let K_d contain assessed attributes and U_d contain attributes explicitly marked unassessable or not assessed. Write $q_{da} = g_{da}/4$ for assessed grades. When U_d is nonempty, the point score s_d remains `null`. Its algebraic bounds are

$$L_d = \frac{\sum_{a \in K_d} q_{da}}{n_d}, \qquad H_d = \frac{\sum_{a \in K_d} q_{da} + |U_d|}{n_d}. \tag{4}$$

Unknown attributes retain the original denominator and receive no assumed grade. The bounds describe possible unresolved scores, not a confidence interval. They coincide with s_d when all attributes are assessed.

The same dimension weights propagate these bounds to S_O and URS. For the nonempty set $T = \mathcal{D}$ or $T = J_{\mathrm{open}}$, respectively, the bounds are

$$[L_T, H_T] = \left[\frac{1}{|T|} \sum_{d \in T} L_d,\ \frac{1}{|T|} \sum_{d \in T} H_d \right]. \tag{5}$$

An unresolved dimension in T makes its aggregate point score unavailable. This includes a fixed attribute inside a dimension selected by URS. Uncertainty outside J_{open} can leave S_O unresolved while URS is resolved; O3 gives identical availability and bounds for both. In the flux example, $(\mathtt{null}, 0.5, 1)$ gives s_d and URS the interval $[0.5, 5/6]$.

Reported tables average resolved, applicable point scores. Each answer retains its predefined attribute denominator, while metric-specific availability determines which answers enter a mean. The intervals above characterize uncertainty within an answer; repeated-run standard deviations characterize variation across runs.

A.5 Judgment States and Requirement Coverage

Scientific finding judgments are supported, contradicted, or unverified. Unverified judgments arise from incomplete reference coverage, uncertain semantic correspondence, or insufficient evidence for an extracted statement. Scientific finding eligibility records whether a statement is a task-relevant empirical claim, independently of its correctness or reference coverage. Clearly ineligible statements are excluded. Demonstrated omissions and contradictions are distinguished from unresolved evaluation judgments.

For each analysis group and reference branch, matching is one-to-one, so repeated statements cannot independently cover several required results. F1 counts matched core reference findings. F2 maps supported findings to their authored roles and applies the acceptable-combination rule in Section 4.2. Supporting findings cover a required role only when that role is explicitly assigned in the reference specification. A numerical claim without a compatible reference does not establish a supported match.

A.6 Branch Selection and Missing Scores

Let $\hat{\mathcal{F}}_g$ contain the distinct task-relevant scientific findings attributed to analysis group g, with $g = 0$ denoting the main analysis and $\hat{\mathcal{F}}_0 = \hat{\mathcal{F}}$ as defined in Section 3.1. We write $\hat{\mathcal{F}}_{\text{all}} = \bigcup_g \hat{\mathcal{F}}_g$, retaining each scientific finding's group attribution.

Within each analysis group, branches are selected by finding recall, with finding precision as the tie-break. The selected branch applies to the group's findings together. Consistency additionally requires support compatible with the group's declared operationalization. Frozen supplementary evidence can support or contradict a claim when its scientific quantity, method, and scope match. A confirmed missing principal decision excludes that group from consistency scoring.

Unverified scores are recorded as N/A and excluded from aggregation for the affected metric. Inapplicable metrics and missing-answer scores are also excluded, with their reasons retained separately. Review-level parsing or extraction failures leave dependent scores unverified. These exclusions operate on metric scores; an answer with an evaluation report can therefore contribute to some metric summaries and not others. Empty finding sets receive precision zero, whereas an empty consistency set makes C_{OF} inapplicable.

A.7 Branch Alignment and Aggregate Reporting

Method-binding coverage is $|\hat{\mathcal{F}}_{\text{bound}}|/|\hat{\mathcal{F}}_{\text{all}}|$, where $\hat{\mathcal{F}}_{\text{bound}} \subseteq \hat{\mathcal{F}}_{\text{all}}$ contains findings from groups whose principal analysis decisions are all identifiable. This diagnostic reports the fraction of findings included in C_{OF} and complements the consistency score by showing how much of an answer can be linked to a declared method (Appendix H.4).

For resolved method matches, let M be the main analysis's branches with the largest number of matching operationalization dimensions, and let W_0 be its best result-matching branches. Branch Alignment is $\mathbf{1}[M \cap W_0 \neq \varnothing]$. This diagnostic uses reference equivalence, whereas S_O measures satisfaction of the task's scientific requirements. A single reference branch, or a tie over all branches on either side, provides no informative method–result comparison. These cases are classified as uninformative alignment. With unresolved matches, we retain possible best branches and report an alignment value only when overlap or non-overlap is established. Established overlap with unresolved ties has a separate diagnostic status.

For each metric and run, eligible cases receive equal weight within a task condition. Overall model summaries weight the applicable condition means equally within each run and report the mean and sample standard deviation across the three run summaries, as in Section 5.2. N/A scores are excluded from the corresponding point-score mean. Metric-specific resolved and applicable counts identify the contributing subsets.

Implementation evidence. Appendix G.4 illustrates a reference coverage boundary through a line-numbered answer, its atomic grades, and finding-level checks. The implementation records and metric computation are described in Appendix G.

B Datasets and Analysis Environment

B.1 Dataset Collection and Selection

The VTK example database provides the initial source of field inputs, supplemented by public scientific data with complementary analysis settings. The unit of selection is a field input with a defined spatial domain and set of physical quantities. A mesh and its associated solution files constitute

one input; the selected snapshot defines the temporal scope of its cases. Distinct field systems and analysis meshes from the same release can provide separate inputs, as in the flow and electrical data from the electrolyzer release.

We assess suitability relative to a scientific target. The available fields and documented physical context must support interpretation of the question, and the supplied snapshot must contain the observations needed to answer it. The target must admit explicit operationalization decisions and an executable reference analysis whose findings can be checked under those decisions. These criteria connect data selection to the benchmark's assessment of how agents define an analysis and support its findings.

Selection combines recurring analysis targets with variation in physical quantities, mesh structure, and measurement conventions. Region tasks require a selection criterion, a connectivity rule, and a measure for comparing spatial objects. Association tasks require a sampling population, an association measure, and a weighting convention. Width tasks connect a spatial coordinate to a weighted distribution and a definition of extent. MHD Turbulence and Radiative Mixing Layer add vector alignment and advective flux, respectively, whose definitions depend on vector normalization and the order of products and aggregation. These choices connect the task inventory in Appendix E to the computational conventions in Appendix F.

B.2 Dataset Inventory

Table 8 summarizes the 15 dataset inputs and their contribution to the 24 scientific question families. Dataset names link to the source records or documentation. The collection includes seven datasets documented in VTK examples and the OpenFOAM tubes data distributed through PyVista. The blow-molding, double-fin, and electrolyzer data come from public dataset releases (Grauberger & Tzionis, 2025; Zigunov, 2024; Hadikhani, 2024). The MHD turbulence, radiative mixing layer, and Rayleigh–Taylor inputs are snapshots from The Well (Ohana et al., 2024); its dataset documentation identifies the associated studies as Burkhart et al. (2020), Fielding et al. (2020), and Cabot & Cook (2006), respectively. Electrolyzer and Electrolyzer Electric contain different fields and meshes from the same source release and are counted as separate dataset inputs.

The 15 field inputs comprise nine legacy VTK datasets, three VTU datasets, and three PLOT3D grid/solution pairs. Manifests record files, checksums, sources, and reader settings; case metadata specify the grid, arrays, point or cell association, documented units, and single-snapshot scope. Units absent from the source remain unspecified. These metadata support array interpretation while leaving task-specific analysis choices to the respondent.

B.3 Field Preparation and Reader Settings

Table 9 identifies each array by name and association, including names shared by point and cell arrays in OpenFOAM Tubes. Coordinate axes retain their stored meaning. Manifests distinguish stored quantities from reader-derived quantities; a vector assembled during source conversion is stored in the resulting benchmark input.

Readers and field decoding. Legacy VTK and VTU inputs use `vtkDataSetReader` and `vtkXMLUnstructuredGridReader`, respectively. The Carotid, Kitchen, and Office manifests enable reading all scalar and vector arrays. The three PLOT3D inputs use `vtkMultiBlockPLOT3DReader` with automatic format detection, scalar function 153, vector function 200, preservation of intermediate functions, and a single-block requirement. Their stored quantities are `Density`, `Momentum`, and `StagnationEnergy`; `VelocityMagnitude` and `Velocity` are reader-derived. The NASA LOx Post reader metadata additionally identify points with `IBlank`> 0 as valid and those with `IBlank`≤ 0 as blanked computational-domain points.

Snapshot selection and conversion. The MHD Turbulence, Radiative Mixing Layer, and Rayleigh–Taylor imports select trajectory index 0 and frame index 60 from their respective source files. These indices identify a snapshot within each source, rather than a common physical time. The conversion preserves the stored coordinate samples in a VTK rectilinear grid. Source fields indexed as (x, y, z), with a final component dimension for vectors, are reordered so that the VTK point index varies fastest in x. The analysis domain spans the stored coordinates without extrapolation to

Table 8: Dataset inventory. S and U denote structured and unstructured grids. Listed fields summarize the available quantities; the case metadata specify exact array names and associations. Velocity marked with † is derived by the PLOT3D reader. Each family contributes one case per task condition.

Dataset	Grid/format	Main fields	Families/ cases
AIDEAS Blow Mold	U/VTU	Velocity, pressure, viscosity, shear rate	3/12
Blunt Fin	S/PLOT3D	Density, momentum, energy, velocity†	1/4
Carotid	S/VTK	Velocity, speed-proportional scalar	1/4
Combustor	S/PLOT3D	Density, momentum, energy, velocity†	1/4
Double Fin	S/VTK	Pressure ratio	1/4
Electrolyzer	S/VTK	Velocity, pressure, volume fraction	1/4
Electrolyzer Electric	U/VTK	Potential, resistance, current density	1/4
FireFlow	U/VTU	Stored arrays: `t`, `uvw`, `u`, `v`, `w`, `mfrac`	1/4
Kitchen	S/VTK	Velocity, pressure, gas concentration	1/4
MHD Turbulence	S/VTK	Density, velocity, magnetic field	2/8
NASA LOx Post	S/PLOT3D	Density, momentum, energy, velocity†	1/4
Office	S/VTK	Velocity, pressure	1/4
OpenFOAM Tubes	U/VTU	Velocity, pressure, turbulence fields	3/12
Radiative Mixing Layer	S/VTK	Density, pressure, velocity	3/12
Rayleigh–Taylor	S/VTK	Density, velocity	3/12
Total			24/96

Table 9: Input arrays for all 15 datasets. P and C denote point and cell association. This table lists the fields used by the scientific question families and associated input quantities; Kitchen also supplies auxiliary arrays documented in its metadata. PLOT3D fields are specified below.

Dataset	Array names and association
AIDEAS Blow Mold	P: `Velocity [m/s]`, `Viscosity [Pa s]`, `NormShearRate [1/s]`; C: `Pressure [bar]`.
Blunt Fin	P: PLOT3D fields, including reader-derived `Velocity`.
Carotid	P: `vectors` (velocity), `scalars` (speed-proportional scalar).
Combustor	P: PLOT3D fields; the surface task uses stored `Density`.
Double Fin	P: `P/Pinf`.
Electrolyzer	C: `p`, `vf`, `velocity`.
Electrolyzer Electric	P: `Potential`, `Resistance`, `jx`, `jy`, `jz`, `current_density`.
FireFlow	P: `t`, `uvw`, `u`, `v`, `w`, `mfrac`; speed-region tasks use `uvw`.
Kitchen	P: `velocity`, `p1`, `u1`, `v1`, `w1`, `ke`, `ep`, `h1`, and concentration arrays `c1`, `c11`–`c17`.
MHD Turbulence	P: `density`, `velocity`, `magnetic_field`.
NASA LOx Post	P: PLOT3D fields, with reader-supplied `IBlank` validity information.
Office	P: `vectors` (velocity), `scalars` (pressure).
OpenFOAM Tubes	P and C: `U`, `k`, `nut`, `omega`, `p`.
Radiative Mixing Layer	P: `density`, `pressure`, `velocity`.
Rayleigh–Taylor	P: `density`, `velocity`.

unstored boundary faces. The provenance records retain the source asset, selected time, conversion metadata, and output checksum.

The Electrolyzer import selects the `01-02-Q300I45R202NoSurfactant` source subset. It checks the XMF dimensions, origin, and spacing for agreement across fields, reads the cell arrays with their declared byte order, and uses the recorded little-endian assumption where byte order is absent. Components `vx`, `vy`, and `vz` are assembled into stored cell `velocity` without interpolation. The electrical input uses `potsolver_0094.vtk` and combines point arrays `jx`, `jy`, and `jz` into `current_density`, retaining the electrical mesh. The OpenFOAM import specification selects time 1000 and the internal mesh for the VTU input. Task-specific finite-value selection, conversion to cell quantities, and weighting are part of the reference operationalization described in Appendix F.2.

B.4 Physical Context and Source Binding

Physical context describes the objects, boundaries, directions, and operating conditions relevant to the question. Each fact is linked to a location in the source documentation. For Office, the context identifies the physical setting, the office object, the ventilation inlet, and the outlet from the VTK example description and its Figure 9-47. The model receives this compact physical description together with the question and data metadata; reference analyses and findings are reserved for evaluation. Appendix K.2 indexes a complete input example.

B.5 Analysis Environment

Each trial has a fresh workspace and a persistent Python interpreter. The tool accepts code and returns execution status, standard output/error, and exceptions as text, allowing revision after ordinary errors. NumPy, SciPy, VTK, and Matplotlib support analysis and plot generation; saved plot pixels are not returned to the model. Inputs comprise (Q, D, C), geometry where applicable, reader metadata, and runtime capabilities, excluding reference analyses and ground truth. Instructions prohibit network use and host-filesystem searches. The main-experiment runtime and package versions appear in Appendix H.2.

Each model turn permits up to eight tool calls. Execution feedback is limited to 16,000 characters per call and 48,000 characters per batch, with truncation recorded in the feedback. Interaction continues until a final answer, the configured case time limit, or a terminating execution failure. The experiment records completion status separately from answer quality. Runtime specifications and run configurations are described in Appendix H.

C Illustrative Test Cases

The four cases illustrate how the benchmark separates analysis choices from their findings: executing a specified analysis in Carotid, choosing a strength measure in Office, choosing a density level in Combustor, and selecting sufficient evidence in Kitchen.

The illustrative responses in this appendix come from development runs separate from the 2,592 main-experiment runs. Unless identified otherwise, the comparisons use development trial 1 for Claude Fable 5.1 (`claude-fable-5-1`), GPT-5.6 Luna, and GPT-5.6 Terra. The collaboration-based interface allowed 900 seconds and at most 60 Python executions per case, with `max` reasoning for Claude Fable 5.1 and `xhigh` for GPT-5.6 Luna and GPT-5.6 Terra; network isolation was inactive. Additional GPT-5.6 Sol responses have separately stated configurations.

Models received single-snapshot data, physical context, and reader metadata, and returned natural-language operationalization and finding sections. We render their stated analyses on the original data to show the spatial consequences; model execution used numerical data and text feedback. Quantities with undocumented units retain their stored scales. In these examples, q90 denotes the 90th percentile, and six-neighbor connectivity joins points one index step apart along any grid axis.

C.1 CAROTID: EXECUTING A SPECIFIED ANALYSIS (O1–F1)

The input represents blood flow in human carotid arteries on a $76\times49\times45$ structured grid. Velocity is the point-associated array `vectors`; the separate speed-proportional scalar is not used to compute velocity magnitude. The context identifies the physical setting without supplying anatomical region labels. The complete question is:

> Investigate the high-speed flow regions in the carotid artery flow field. Calculate speed from the Euclidean magnitude of the velocity vectors, define high-speed locations using the linearly interpolated 90th percentile or higher of the non-zero speeds, and group locations connected through immediate grid neighbors along one grid axis into contiguous regions. Identify the region with the greatest peak speed, then report its strength as that peak speed and its location as the mean spatial location of the retained high-speed locations within it.

This question specifies all four principal decisions: six-neighbor connectivity, the non-zero-speed percentile criterion, peak-speed strength, and the arithmetic mean location of the selected region. Let $s_i = \|\mathbf{v}_i\|_2$, let t be the linearly interpolated 90th percentile of the non-zero speeds, and let R range over connected components of $\{i : s_i \geq t\}$. The selected region maximizes $\max_{i\in R} s_i$; its representative location is $|R|^{-1}\sum_{i\in R}\mathbf{x}_i$.

Executing this analysis gives $t = 0.039422780215$, 16,757 retained points, and 642 connected regions. The selected region contains 13,735 points. Its peak speed is 22.694927991, and its mean location is $(137.808445577, 94.491445213, 15.773061522)$. The peak itself occurs at $(143, 102, 28)$, which is a different location descriptor.

Claude Fable 5.1, GPT-5.6 Luna, and GPT-5.6 Terra all state the prescribed analysis and select the 13,735-point region. They report peak strengths of 22.695, 22.6949272, and 22.694928, respectively, and mean locations agreeing with the reference at their reported precision. The three answers agree on the specified analysis and distinguish the region's mean location from its peak-speed point.

The declared analysis binds both required findings to the stored branch, and both match within their respective strength and spatial tolerances of 0.05. URS is inapplicable to this fully specified question.

C.2 OFFICE: CHOOSING A STRENGTH MEASURE (O2–F1)

The Office input is a $21 \times 20 \times 20$ structured-grid snapshot with point velocity `vectors` and a separate pressure scalar. Its context describes ventilation in a small office and identifies inlet and outlet boundary types. The O2–F1 question asks for the strongest high-speed region, leaving its retention criterion and strength measure open while specifying six-neighbor connectivity and mean retained-point location. Its complete wording appears in Figure 8.

Table 10 gives the three stored reference branches. The first two select different winners from the same retained points because they maximize different strength measures; their spatial consequences appear in Figure 8. The third branch also changes the threshold population and value.

Table 10: Office reference branches. Every branch uses six-neighbor connectivity and the arithmetic mean location of retained points. The non-zero-speed q90 uses linear interpolation.

Branch	Retention rule	Strength	Mean location (x, y, z)
A	$s > 0.20$	Peak: 0.804935	$(0.068125, 2.343750, 2.389375)$
B	$s > 0.20$	Mean: 0.357731	$(4.500000, 2.431250, 2.173750)$
C	$s \geq q_{0.9}(s \mid s > 0)$	Mean: 0.291934	$(4.497742, 2.421505, 2.112043)$

Figure 6 shows the three development responses. Claude Fable 5.1 maximizes peak speed and selects the region near the low-x boundary; GPT-5.6 Luna maximizes mean speed, while GPT-5.6 Terra maximizes the sum of retained-point speeds, and both select the region near the high-x boundary. Their retention rules yield the same 420-point mask in this snapshot. The global mean plus two standard deviations (0.168569914) and the all-point q95 (0.168570906) are numerically distinct,

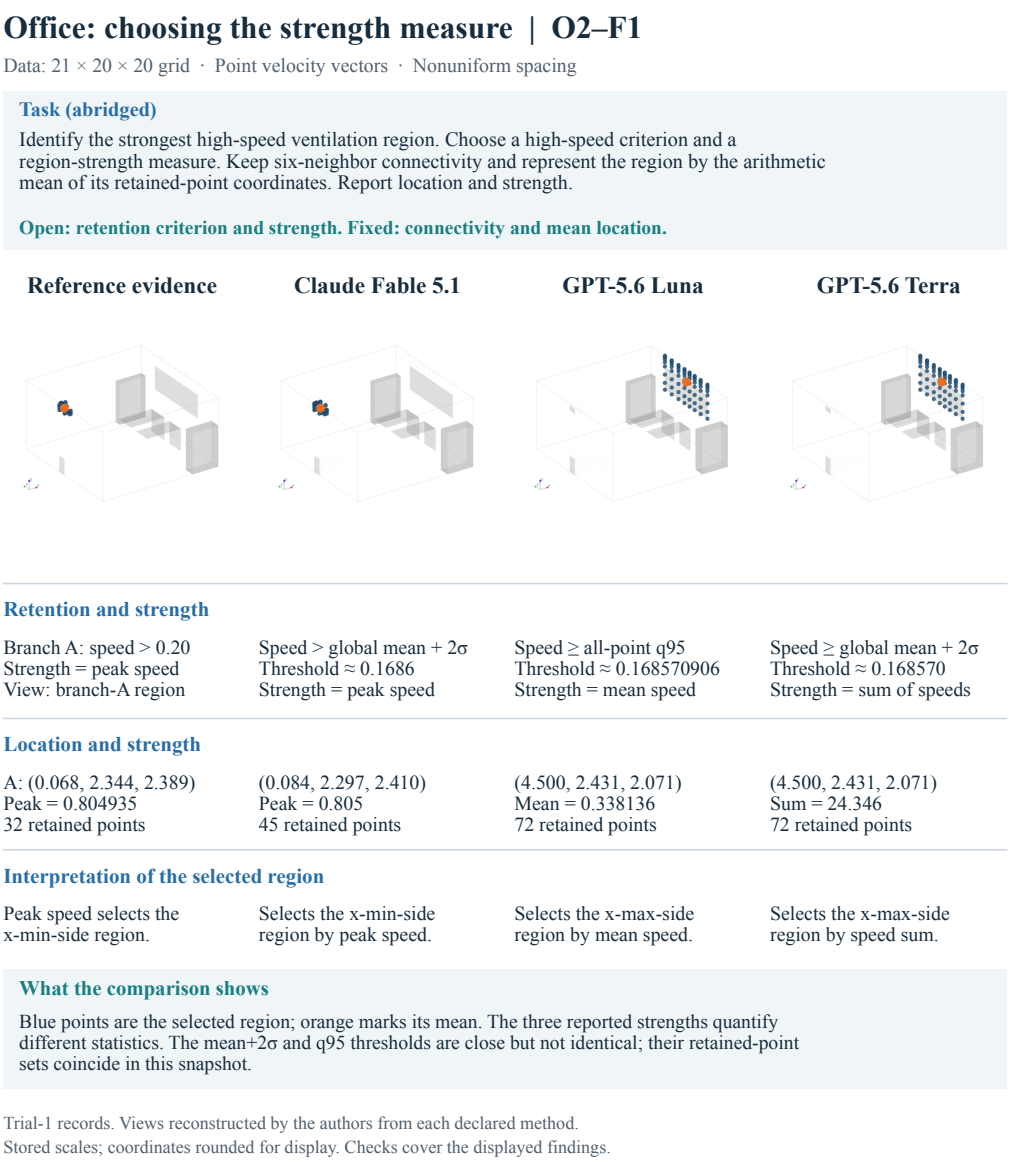


Figure 6: Office O2–F1: reported strength measures select different regions. Blue points show the selected region and orange marks its mean location; gray geometry provides context from the official VTK Office example. All panels share a camera. Reference branch A is one stored analysis. The three model panels reproduce their stated primary analyses on the original data. Peak speed, mean speed, and a sum of point speeds are different statistics, so their magnitudes are not directly comparable.

contrary to Claude Fable 5.1's statement that they coincide exactly. GPT-5.6 Terra's sum measures cumulative speed over retained samples. Its equal point weights make the measure sensitive to sampling density on this nonuniform grid.

A separate development response from GPT-5.6 Sol (`xhigh`, 1,800-second budget) chooses the 90th percentile of *all* point speeds and ranks components by mean retained speed. It reports a 93-point region near the high-x boundary, with mean location $(4.498, 2.422, 2.112)$ and mean speed 0.292.

The displayed location and strength match branch C within its spatial tolerance of 0.05 and strength tolerance of 0.0005. The answer uses a different percentile population: with 239 zero-speed points in the dataset, the all-point q90 is 0.118925337257, whereas the non-zero-speed q90 is 0.120388175490. Direct execution of both rules selects the same winning region in this snapshot. The example therefore combines matching findings with distinct operationalizations. Method assessment uses the declared all-point criterion, and consistency uses evidence from that analysis.

C.3 Combustor: Choosing a Density Level (O3–F1)

The input describes flow in an annular combustor on a $57 \times 33 \times 25$ structured grid. The PLOT3D grid and solution files provide stored `Density`, along with other flow quantities; the reader configuration distinguishes these stored quantities from derived velocity. This task illustrates how the choice of density level affects the selected surface and its findings. The complete question in the 96-case input is:

> Within the annular-combustor flow field, identify and characterize one connected constant-Density isosurface component. Choose and explain how the components are formed, a finite stored-Density isovalue rule, a rule for selecting one component, and how to represent its location. Report the selected component's representative location, defining density value, and surface area.

The scientific target is a connected constant-density surface component. The model specifies how to form components, choose the density level, select a component, and represent its location. These choices constitute the three principal dimensions of feature definition, criterion, and aggregation or representation. The required findings remain location, defining density, and surface area.

The two stored branches use connected triangulated isosurfaces. Branch A extracts the finite-density q90 surface, selects the component with the largest surface area, and reports its area-weighted centroid. Branch B uses a level equal to 0.9 times q90, selects by axis-aligned bounding-box volume, and reports the bounding-box midpoint. The factor scales the q90 density value; it does not change the percentile rank to 81. Table 11 keeps each level, location, and area attached to the analysis that produced it.

Table 11: Combustor stored reference branches. Level, location, and area remain attached to the analysis that produced them. Area uses the squared stored coordinate scale. Model-selected levels and findings appear in Figure 7.

Analysis	Density level	Representative location (x, y, z)	Area
A: q90, largest area	0.414666533	$(5.186063, 0.249285, 30.535099)$	260.783939
B: $0.9\times$q90, largest box	0.373199880	$(8.319478, 0.000000, 29.834877)$	322.281228

Figure 7 compares the surfaces obtained from the model responses. Claude Fable 5.1 chooses the midpoint of the minimum and maximum density; GPT-5.6 Luna and GPT-5.6 Terra choose the median. Both median definitions select the same value for the 47,025 samples, yielding three components and a selected surface with 20,100 triangles. Thus the two median panels show agreement, while the mid-range panel demonstrates a different level choice. All three select by triangle area and report an area-weighted centroid.

The reproduced mid-range and median surfaces illustrate analyses beyond stored branches A and B. Their findings remain unverified against those branches because the defining density levels differ. Method-specific evidence is incorporated through the reference-extension procedure in Appendix F.

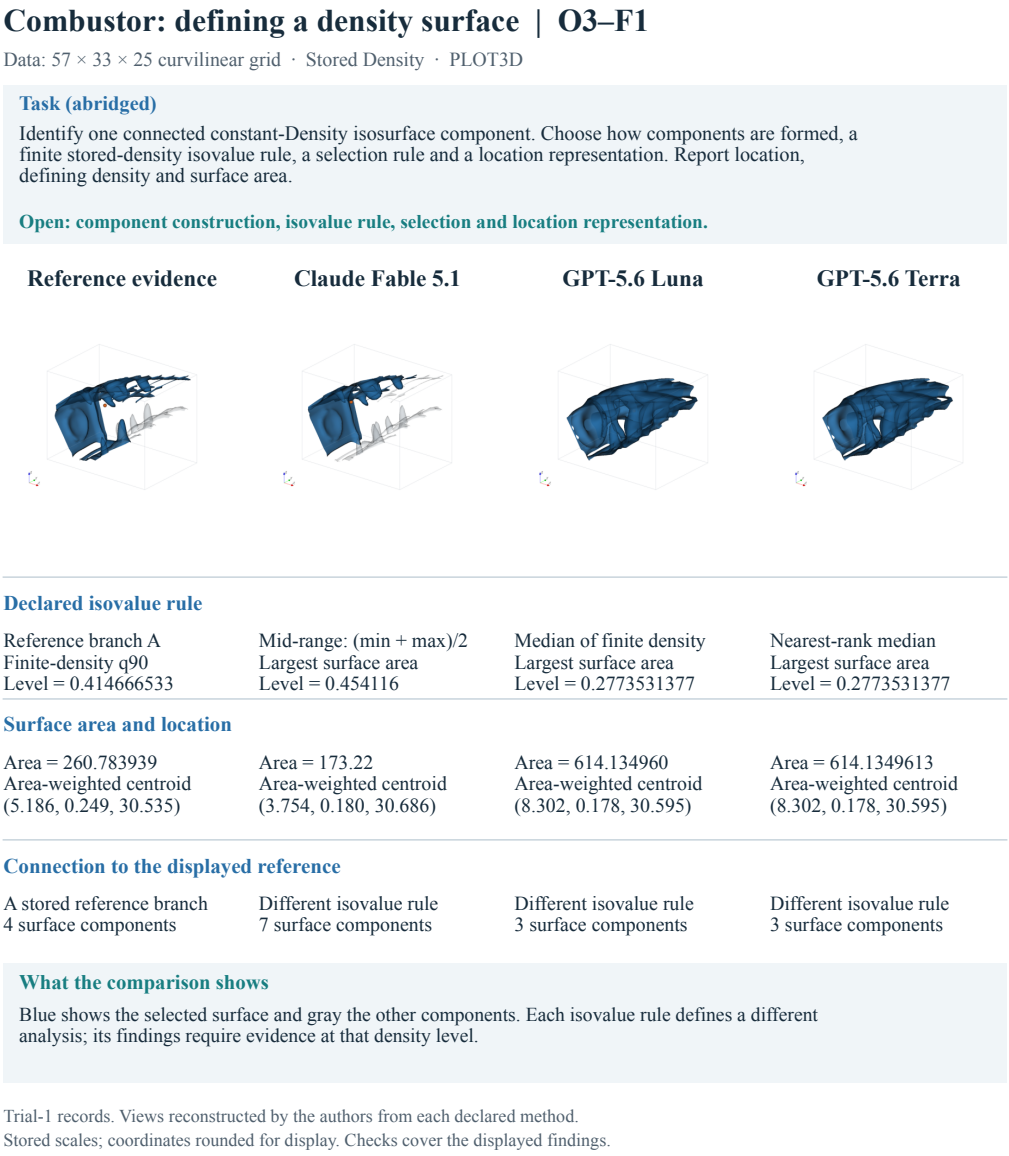


Figure 7: Combustor O3–F1: density-level choices change the selected surface and its findings. Blue shows the largest-area component and gray shows other components, using shared camera and domain bounds. The mid-range and median model results are reproduced analyses outside the displayed q90 reference branch. GPT-5.6 Luna and GPT-5.6 Terra select the same density level on this snapshot.

The stored spatial, density, and area tolerances (0.05, 0.0005, and 0.5) apply to like-defined quantities.

C.4 KITCHEN: SELECTING SUFFICIENT FINDINGS (O1–F2)

The Kitchen input describes airflow around a hot stove on a $28 \times 24 \times 17$ structured grid. The point-associated concentration candidates are `c1` and `c11`–`c17`; velocity and other scalar arrays are also supplied. The complete question is:

> Within the kitchen airflow field, determine which stored gas-concentration field has the greatest spatial heterogeneity. Treat every available stored gas-concentration field as a candidate, compare them using the cell-volume-weighted coefficient of variation, and report which field was selected. Provide evidence from that heterogeneity measure and add one relevant characterization of the result, chosen by you. First take the arithmetic mean of the finite vertex values within each cell before applying the stated measure.

The analysis is specified: form each cell value c_i by averaging its finite vertex values, use positive cell volumes V_i as weights, and select the field with the largest coefficient of variation. In particular,

$$\mu_V = \frac{\sum_i V_i c_i}{\sum_i V_i}, \qquad \sigma_V = \sqrt{\frac{\sum_i V_i (c_i - \mu_V)^2}{\sum_i V_i}}, \qquad \mathrm{CV}_V = \frac{\sigma_V}{|\mu_V|}. \tag{6}$$

All candidate vertex values are finite, and 9,936 positive-volume cells enter the calculation after vertex-to-cell conversion.

Table 12: Kitchen reference evidence. The first two rows compare every candidate under the specified analysis; the remaining rows characterize the selected field. CV is dimensionless. Other statistics use the stored concentration scale.

Field	`c1`	`c11`	`c12`	`c13`	`c14`	`c15`	`c16`	`c17`
CV_V	0.3094	0.1442	**57.2661**	2.0219	5.0342	0.7599	3.3848	2.3047

Selected-field mean: 155.436454	Selected-field standard deviation: 8901.245888
Weighted q10 and q50: 1.0001251	Weighted q90: 1.0002178

Field identity and heterogeneity evidence are mandatory. The answer must also supply one supported characterization: comparison with other fields, informative description of the distribution, distribution shape, spread, or extreme behavior. A supported spread statistic or comparison can satisfy this requirement; field identity and CV alone cannot. Role assignment depends on the statistic's meaning and evidential support.

All three development responses select `c12`, reporting CVs of 57.27 (Claude Fable 5.1), 57.2661402 (GPT-5.6 Luna), and 57.2661 (GPT-5.6 Terra), within the stored CV tolerance of 0.05. Claude Fable 5.1 reports that its CV is approximately 11.4 times that of `c14`. GPT-5.6 Luna reports a volume-weighted q95 of 1.0886434; GPT-5.6 Terra reports a cell-average q95 of 92.159 without explicitly declaring its weighting. Recalculation gives 1.0886434 for a volume-weighted quantile and 92.1593 for an unweighted cell quantile. GPT-5.6 Terra's value matches the unweighted calculation, while its answer leaves that weighting choice unspecified.

Each response supplies an additional characterization. Field selection and CV agree with the reference; support for each tail statistic or physical interpretation is assessed under its stated definition. For the selected-field standard deviation and median, three-significant-figure reporting precision gives absolute tolerances of 5.0 and 0.005, respectively (Equation 8).

D WITHIN-FAMILY TASK CONDITIONS

Figure 8 holds the Office snapshot, physical context, and scientific target fixed across four questions. Its matrix identifies the decisions opened in O2 and O3 and the additional finding required in O1–F2. The rendered branches show how different strength definitions can select different regions

from the same retained points. Office illustrates the condition design; the family-specific attribute assignments are given in Table 6.

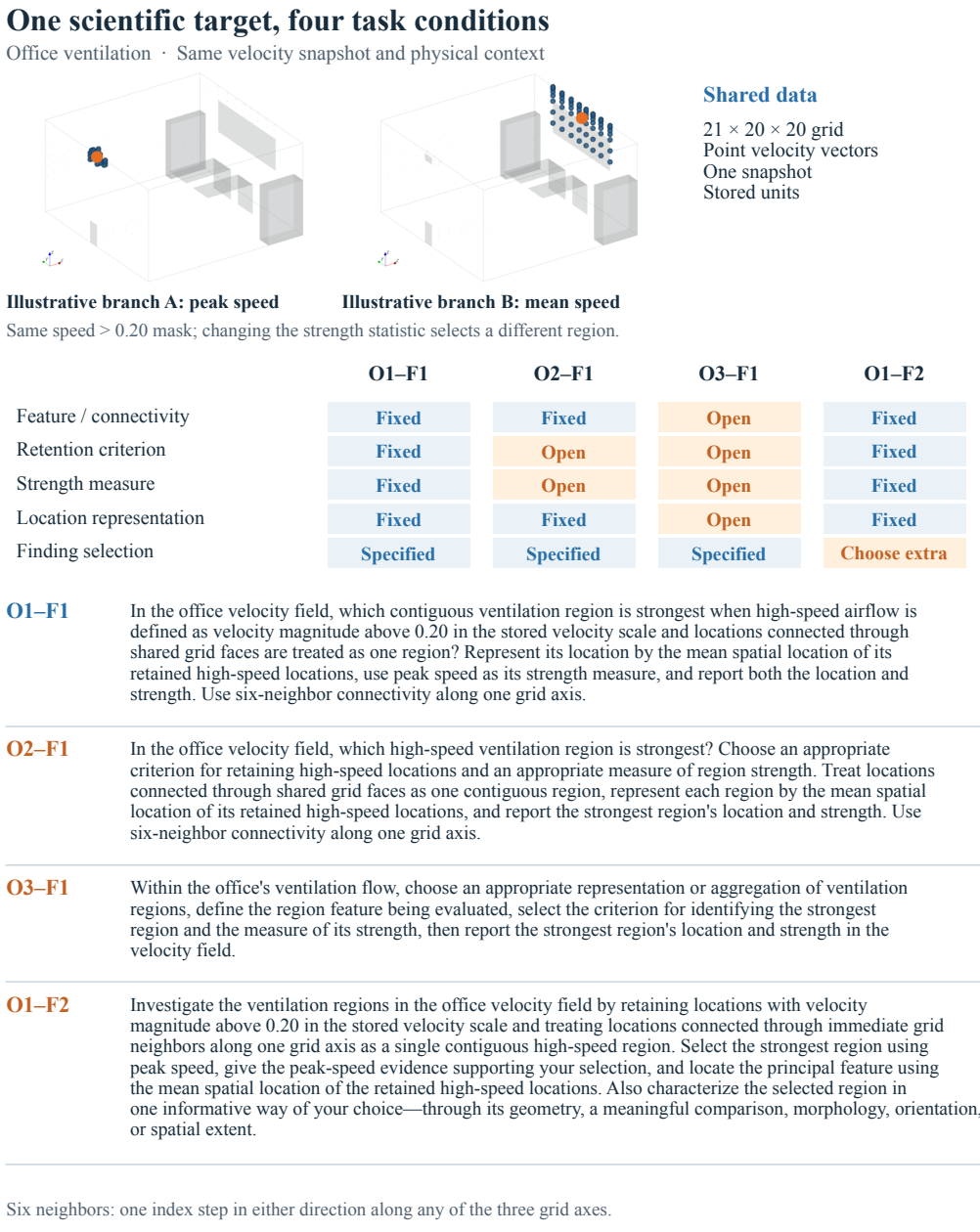


Figure 8: Office paired conditions. The upper views locate two stored reference selections on the same speed-above-0.20 mask. The matrix distinguishes prescribed decisions from choices left to the model. The four complete questions are transcribed from case inputs with typographic normalization. O1–F2 retains the O1 analysis and adds a model-selected characterization to the mandatory location and strength evidence.

E Task Construction and Quality Checks

E.1 From Available Fields to Scientific Targets

A family specification links a scientific target to available quantities, an analysis scope, principal operationalization dimensions, and required finding roles. The scope identifies the snapshot and spatial population to which the question applies. The dimensions describe decisions that can change what is measured: feature definition, property measure, aggregation or representation, and, where relevant, a selection criterion or numerical parameter. For example, a field-association task requires a sampling population, an association measure, and an aggregation rule; a selected-region task additionally needs a selection criterion. The family specification defines the principal dimensions before answer generation.

Task construction uses the documented fields and coordinate meanings. Double Fin supports a question about reconstructed pressure ratio, while the electrolyzer flow and electrical inputs support separate velocity-extent and current-density questions on their respective meshes. The electrolyzer extent question follows the stored y coordinate. All tasks concern spatial quantities and relationships within a single snapshot.

Table 13 lists all 24 scientific question families, the decisions left open in O2, and the F1 reporting obligations. The benchmark contains 24 scientific question families, each instantiated under four conditions, yielding 96 cases with scientific targets paired across conditions. Dataset provenance is given in Appendix B.2.

E.2 Condition Assignment and Wording Review

For O1, the question exposes every principal analysis decision. For O2, the family specification selects one or two principal dimensions to leave open; the remaining constraints are retained explicitly. O3 designates the full principal-dimension set. These designations determine the dimension sets used in URS; the scientific target, scope, and requested outputs can still fix attributes within them (Appendix A.3). The three F1 versions retain the same finding roles. O1–F2 keeps the O1 decisions and replaces a fixed reporting list with the family's sufficient-answer requirements, including any mandatory roles. These requirements are authored before evaluating answers.

Metadata validation checks dimension membership, fixed/open assignments, and preservation of family, dataset, target, and scope. LLM wording review then checks nine properties: scientific meaningfulness, target stability, faithful target wording, preservation of scope, exposure of fixed decisions, preservation of open decisions, absence of unintended analysis obligations, correct finding-selection responsibility, and answerability from the supplied information. Numerical reference verification is performed separately.

Revisions are checked against the original family specification to preserve the target and assigned responsibilities. The resulting `case_input.json` contains the question supplied to the model. Recorded changes and their final review status appear below.

E.3 Family Inventory

The inventory distinguishes the scientific target from the numerical statistic chosen to study it. In the table, "measure" denotes the property measure, "criterion" the selection rule, and "level" the numerical isovalue parameter. Association, heterogeneity, width, flux, and alignment families use feature definition, property measure, and aggregation or representation as their principal dimensions. Selected-region families additionally specify a criterion. The Rayleigh–Taylor surface family adds an isovalue parameter, while Combustor uses feature definition, criterion, and aggregation or representation. Width tasks require width and center under F1; their F2 requirements can retain width as the sufficient core and treat center as supporting information. Requirements are recorded per condition.

Table 13: The 24 scientific question families. O2 designates the listed decisions; O1 specifies the analysis and O3 designates the full principal-dimension set. Retained attribute constraints are detailed in Table 6. F1 requirements apply to O1–F1, O2–F1, and O3–F1.

Dataset	Scientific target	O2 open choice	F1 findings
AIDEAS Blow Mold	Pressure nonuniformity	Measure	Coefficient
AIDEAS Blow Mold	Viscosity–shear-rate association	Measure	Association coefficient
AIDEAS Blow Mold	High shear-rate regions	Criterion	Volume-weighted centroid and regional mean
Rayleigh–Taylor	Intermediate-density vertical extent	Measure	Width and weighted center
Rayleigh–Taylor	Density–vertical-velocity association	Measure	Association coefficient
Rayleigh–Taylor	Middle-density surface geometry	Level	Total area and area-weighted centroid
Radiative Mixing Layer	Density nonuniformity	Measure	Coefficient
Radiative Mixing Layer	Density–pressure association	Measure	Association coefficient
Radiative Mixing Layer	Vertical advective mass-flux density	Measure	Signed domain-average flux density
MHD Turbulence	Density–magnetic-magnitude association	Measure	Association coefficient
MHD Turbulence	Velocity–magnetic-field alignment	Measure	Signed alignment coefficient
Double Fin	High reconstructed pressure-ratio regions	Criterion	Volume-weighted centroid and regional mean
Electrolyzer	Speed extent along stored y	Measure	Width and weighted center
Electrolyzer Electric	High current-density-magnitude regions	Criterion	Volume-weighted centroid and regional mean
OpenFOAM Tubes	Turbulent kinetic energy nonuniformity	Measure	Coefficient
OpenFOAM Tubes	Pressure–speed association	Measure	Association coefficient
OpenFOAM Tubes	Speed extent along stored x	Measure	Width and weighted center
Blunt Fin	Strongest high-speed region	Criterion, measure	Location and strength
Carotid	Strongest high-speed region	Criterion, measure	Location and strength
Combustor	Connected constant-density surface	Criterion	Representative location, density level, and area
FireFlow	Strongest high-speed region	Criterion, measure	Location and strength
Kitchen	Most heterogeneous concentration field	Measure	Field identity and heterogeneity measure
NASA LOx Post	Strongest high-speed region	Criterion, measure	Location and strength
Office	Strongest ventilation region	Criterion, measure	Location and strength

E.4 Scientific and Computational Challenges

The scientific question families vary in the reasoning needed to connect a physical question to a numerical analysis. Five recurring challenges are field derivation, mesh topology, weighting and normalization, numerical sensitivity, and scientific interpretation (Table 14). A domain-wide statistic may have a simple geometric scope but require a precise weighting convention; a regional analysis additionally requires defining and comparing spatial objects.

Within a family, the dataset, scientific target, and scope are shared across conditions. O1 supplies the analysis decisions, O2 leaves selected dimensions open, and O3 requires the respondent to choose all principal dimensions. F2 changes the selection of sufficient findings under the O1 analysis.

Table 14: Recurring analysis challenges and representative scientific question families. These describe the scientific work required by a task; O1–O3 determine which analysis decisions the question supplies.

Challenge	Scientific consequence and examples
Field derivation	Vector magnitudes, point-to-cell conversion, and products define the quantity being analyzed. Radiative advective flux and MHD alignment depend on the order of these operations.
Mesh topology	Connectivity determines which samples belong to one region or surface. Blunt Fin and Carotid use grid adjacency; Combustor and Rayleigh–Taylor operate on triangulated isosurfaces.
Weighting and normalization	Cell volume, sample count, and triangle area define different populations and averages. AIDEAS pressure nonuniformity, Kitchen concentration variation, and surface centroids depend on these choices.
Numerical sensitivity	Quantile populations, threshold boundaries, and distribution tails affect selection and reported statistics. Office speed regions and Kitchen concentration summaries illustrate these effects.
Scientific interpretation	A statistic must support the claimed physical meaning. Density–motion and pressure–speed association describe relationships within a snapshot; electrolyzer width follows the documented coordinate direction.

This design separates the task's underlying scientific demands from the decisions delegated to the respondent. The paired examples in Appendix D and the attribute assignments in Appendix A.2 make that distinction explicit.

E.5 Construction Review Outcomes

The question wording for all 96 cases, 24 cases per condition, passed the nine wording checks described above. Wording review used `gpt-5.6-sol` with `high` reasoning and no tools, and retained a rationale for each question. Scientific suitability and numerical reference verification form the complementary checks described in Appendix F.

Four wording revisions clarified the intended responsibilities (Table 15). FireFlow and NASA LOx Post identify the speed quantity while leaving the intended analysis decisions open. Combustor makes the density-level choice and required surface-area report explicit. Each revision preserves the scientific target.

Table 15: Question revisions that clarify the analysis and reporting responsibilities.

Case condition	Clarification in the revised question
FireFlow O3–F1	Define speed as the Euclidean magnitude of stored `uvw`; leave threshold, connectivity, strength, and location representation open.
NASA LOx Post O3–F1	Identify the reader-derived `Velocity` as the speed source; retain the four open region-analysis decisions.
Combustor O2–F1	Ask for a finite stored-Density isovalue rule, retaining largest-area component selection, area-weighted location, and the area report.
Combustor O3–F1	Expose component construction, density-level choice, component selection, and location representation; retain reports of location, defining density, and area.

The family specifications retain the principal dimensions and sufficient-answer requirements independently of these wording changes. Table 18 completes the family inventory with these requirements. The actual authoring and wording-review instructions are described in Appendix K.1.

F GROUND TRUTH RECORDS AND VERIFICATION

F.1 PAIRING EACH ANALYSIS WITH ITS FINDINGS

A reference branch pairs a complete operationalization with its executed findings and comparison policies (Table 16). Scientific review records whether the analysis suits the target; execution records its numerical outcome. Office branches A and B show why both are needed: changing the strength measure changes the selected region and its findings (Table 10).

Table 16: Information retained in a reference record and its role in evaluation.

Record component	Contents and purpose
Task specification	Dataset and case identity, target, scope, principal dimensions, and fixed or open constraints.
Analysis branch	Complete decisions, including populations, thresholds, connectivity, measures, and representations where applicable.
Reference finding	Scientific quantity, object and scope, statement, value, unit, importance, and branch association.
Execution evidence	Input and reader provenance, executable analysis definition, and recorded numerical outcomes.
Verification policy	Finding type and frozen reporting precision, spatial resolution, or exact-match rule, with the corresponding numerical tolerance.
F2 requirements	Mandatory roles, acceptable role combinations, and mappings from reference findings to roles.

F.2 DISCRETE CONVENTIONS IN REFERENCE ANALYSES

An executed reference specifies its discrete population and calculation order as part of its operationalization. Table 17 summarizes conventions used by the reference construction routines. The family specification determines which decisions are fixed and which remain open to the respondent. Reference routines provide executable analyses of those specifications; models implement their own choices through Python.

For positive weights w_i and values sorted as $x_{(1)} \le \cdots \le x_{(n)}$, the weighted-quantile routine uses $x_{(k)}$, where k is the smallest index satisfying $\sum_{i=1}^{k} w_{(i)} \ge q \sum_{i=1}^{n} w_{(i)}$. Equal-valued samples retain a stable sorting order. This definition differs from linearly interpolating an unweighted sample quantile. Likewise, $\| \operatorname{mean}_i \mathbf{u}_i \|_2$ and $\operatorname{mean}_i \| \mathbf{u}_i \|_2$ retain their different meanings even when a particular dataset gives similar values.

Width calculations combine cell volume with a declared nonnegative spatial weight. They report either weighted root-mean-square distance from the weighted center or the distance between the weighted 10th and 90th percentiles. Association recipes specify Pearson correlation or correlation of average ranks, together with spatial weighting. Constant-field correlations and zero-support widths are undefined and stop the corresponding reference computation. These conventions identify the analysis behind a finding before any numerical comparison with an answer.

F.3 NUMERICAL COMPARISON AND REPORTING PRECISION

Numerical tolerances define the benchmark's reporting-equivalence resolution: the resolution at which a reported value matches a reference finding of the same quantity and scope. During reference construction, each finding is assigned a typed comparison policy specifying its required reporting resolution. The policy is frozen with the reference package before model answers are evaluated. It describes reporting precision, independently of physical measurement uncertainty.

For a continuous scalar finding f, let y_f be its reference value and p_f its reference reporting precision in significant figures. After resolving quantity identity, scope, and any explicitly supported unit conversion, a reported value $\hat{y}_f$ matches when

$$|\hat{y}_f - y_f| \le \tau_{\mathrm{abs},f}. \tag{7}$$

Table 17: Discrete choices recorded by reference analyses. Each convention applies to the branches that select it. Reader decoding is described separately in Appendix B.3.

Choice	Numerical meaning
Array association	A cell array is used directly; the cell-based recipes convert point arrays using arithmetic vertex averages. Point-based analyses retain their point population.
Vector reduction	Point-speed analyses take magnitudes at points. Cell-based recipes that convert point vectors before taking their magnitude specify a different calculation order.
Spatial weights	Cell-based summaries normalize by the included cell volume. A branch declares whether signed volumes must be positive or their absolute values are used. Zero or non-finite weights are rejected.
Finite population	Recipes with declared finite-cell fields use the intersection of their finite masks. Empty populations and non-finite required quantities produce execution errors.
Quantiles and thresholds	Weighted quantiles select the first value whose cumulative weight reaches the requested fraction. Point-percentile analyses specify the population and interpolation rule. Strict and non-strict threshold comparisons are retained explicitly.
Connectivity and location	Speed-region analyses specify grid or mesh adjacency and the location representation. Structured-grid six-neighbor adjacency and shared-cell point adjacency denote different relations.
Products and aggregation	Advective-flux recipes distinguish forming the pointwise product before cell averaging from multiplying cell-averaged quantities.
Surface representation	Isosurfaces are triangulated. Surface area sums triangle areas; location uses an area-weighted triangle centroid or a bounding-box midpoint, as specified by the branch.

For a nonzero reference value, the absolute tolerance is half a unit in the last significant place retained at precision p_f:

$$\tau_{\mathrm{abs},f} = \tfrac{1}{2}\, 10^{\lfloor \log_{10} |y_f| \rfloor - p_f + 1}, \qquad y_f \neq 0. \tag{8}$$

Each continuous scalar reference finding carries its reporting precision p_f, from which Equation 8 derives the absolute tolerance. This precision is a property of the reference finding and may differ across quantities with different reporting requirements. For example, $p_f = 3$ gives $\tau_{\mathrm{abs},f} = 0.0005$ for the Office strength 0.291934, 0.05 for the Kitchen coefficient of variation 57.2661, and 0.5 for the Combustor surface area 260.783939. Zero-valued scalar references require an absolute reporting resolution specified directly in the frozen policy, since significant-figure precision alone does not define a scale at zero.

Coordinate-valued findings carry a spatial comparison policy with a predefined absolute resolution $\tau_{\mathrm{spatial},f}$ in the stored coordinate scale. Locations use Euclidean distance, $\|\hat{\mathbf{x}}_f - \mathbf{x}_f\|_2 \leq \tau_{\mathrm{spatial},f}$. Coordinate spans use componentwise comparison, $\max_j |\hat{e}_{f,j} - e_{f,j}| \leq \tau_{\mathrm{spatial},f}$, because the components describe extents along separate axes. Discrete identities and counts use exact matching. A coordinate vector is one finding; distinct statistics such as minimum, maximum, and median are separate findings with their own policies.

The frozen policy also governs admissible printed representations. The number of digits in a model answer leaves p_f and the reference tolerance unchanged. If a printed value fails the numerical comparison but its rounding interval overlaps the reference acceptance interval, the check remains unresolved unless the frozen policy explicitly authorizes that representation through a minimum significant-digit count and a maximum rounding radius. Units follow the input documentation and remain unspecified where the source provides none.

F.4 Sufficient Finding Combinations under F2

An F2 requirement record specifies mandatory roles and acceptable combinations of roles. Reference findings are mapped to these roles explicitly. Coverage requires supported findings from a shared branch to satisfy at least one complete combination. A supporting statistic contributes to a required role only when its assigned meaning establishes that role.

For Office, the mandatory roles are defining strength evidence and principal-feature location. One additional role is required from spatial extent, geometric relation, meaningful comparison, morphology, or orientation. The stored extent vector $(0.2400000, 0.5999999, 0.2400000)$ supplies the spatial-extent role. Thus, supported peak strength, mean location, and coordinate spans provide one sufficient combination. The stored region count and retained-point count are assigned supporting-statistic roles; listing them does not automatically establish a geometric or morphological characterization. A proposed characterization outside the listed roles requires scientific review before it counts toward role coverage.

F2 sufficiency is family-specific. For Kitchen, mandatory field identity and heterogeneity evidence are accompanied by a supported characterization, as illustrated in Appendix C.4. Width families can have width as their sufficient core while treating the center coordinate as supporting. The evaluator uses the role requirements defined for each family and condition.

F.5 Freezing and Extending Reference Evidence

An evaluation run binds its questions, reference package, verification policies, scoring implementation, and reviewer configuration before scoring. Reference construction and extension take place separately from answer evaluation. A new branch must specify a complete analysis, provide compatible findings and policies, respect fixed task constraints, and map its required findings to the original requirement structure. Adding supporting evidence does not add new mandatory findings.

A reference extension records the analysis, its scientific rationale, and the associated findings, and is applied uniformly to the compared answers. An appropriate analysis can lie outside the current reference set, leaving its findings unverified until matching evidence is available. Appendix C.3 illustrates this distinction through model-selected density levels.

F.6 Reference Coverage Across Scientific Question Families

Table 18 covers all 24 scientific question families in the 96-case construction manifest. Together with Table 13, it specifies the principal-decision count, the decisions opened in O2, F1 reporting obligations, and F2 sufficiency. Every O1–F1 and O1–F2 case has one stored reference branch. Among the 24 O2–F1 cases, 19 have two branches and five have three; among O3–F1 cases, 19 have two and five have four. The resulting 159 case-specific branch entries comprise 24, 53, 58, and 24 entries in O1–F1, O2–F1, O3–F1, and O1–F2, respectively. Analyses can recur across conditions, so this is a record count rather than a count of distinct scientific methods.

Table 18: Decision counts, F2 sufficient roles, and stored reference branches for the 24 scientific question families. m is the number of principal operationalization dimensions. The last column gives O2–F1/O3–F1 branch counts; both O1 conditions have one branch per family. F1 roles and O2 open decisions are listed in Table 13.

Family	m	F2 sufficient roles	O2/O3
AIDEAS: pressure nonuniformity	3	Coefficient	2/2
AIDEAS: viscosity–shear relation	3	Association coefficient	2/2
AIDEAS: high-shear region	4	Centroid and regional mean	2/2
Rayleigh–Taylor: mixing width	3	Width; center is supporting	2/2
Rayleigh–Taylor: density–motion relation	3	Association coefficient	2/2
Rayleigh–Taylor: isodensity geometry	4	Area and centroid	2/2
Radiative: density nonuniformity	3	Coefficient	2/2
Radiative: density–pressure relation	3	Association coefficient	2/2
Radiative: advective flux	3	Signed flux density	2/2
MHD: density–magnetic relation	3	Association coefficient	2/2
MHD: velocity–magnetic alignment	3	Signed alignment coefficient	2/2

Continued on next page

Table 18 continued

Family	m	F2 sufficient roles	O2/O3
Double Fin: high-pressure region	4	Centroid and regional mean	2/2
Electrolyzer: speed extent	3	Width; center is supporting	2/2
Electrolyzer Electric: current region	4	Centroid and regional mean	2/2
Tubes: energy nonuniformity	3	Coefficient	2/2
Tubes: pressure–speed relation	3	Association coefficient	2/2
Tubes: speed extent	3	Width; center is supporting	2/2
Blunt Fin: speed region	4	Strength, location, and a region characterization	3/4
Carotid: speed region	4	Strength, location, and a region characterization	3/4
Combustor: density surface	3	Density level and a surface characterization	2/2
FireFlow: speed region	4	Strength, location, and a region characterization	3/4
Kitchen: concentration comparison	3	Field identity, heterogeneity evidence, and a field characterization	2/2
NASA LOx Post: speed region	4	Strength, location, and a region characterization	3/4
Office: ventilation region	4	Strength, location, and a region characterization	3/4

For the five speed-region families, the additional region characterization can establish spatial extent, a geometric relation, a meaningful comparison, morphology, or orientation. Combustor pairs the defining density level with surface extent, spatial location, or shape characterization, all attributed to the selected component. Kitchen pairs field identity and heterogeneity evidence with a comparison to other fields or an informative description of the distribution, including shape, spread, or extreme behavior. These are the sufficient role combinations defined during construction. Numerical reference coverage is recorded separately for each role and branch.

Finding labels identify the content available in a reference, while role contracts specify what constitutes a sufficient answer. Office O1–F2 stores strength, location, and extent as core findings, with region and retained-point counts as supporting statistics. Width families treat width as core and center as supporting; Kitchen stores two core and four supporting findings. For F2, supported findings from one branch must cover an acceptable role combination. This permits different sufficient answers while preserving their association with a common analysis.

G Evaluation Procedure and Worked Example

G.1 Answer Interpretation and Evidence Matching

Evaluation combines interpretation of the written answer with numerical comparison against reference analyses. The reviewer identifies the stated decisions and scientific findings, citing the answer lines that support each extraction. It assesses atomic operationalization attributes using the five-level rubric in Appendix A.1 and identifies candidate reference findings with the same scientific quantity, object, and scope. Numerical checks and final metric computation follow the rules in Appendix A.

Each answer has one main analysis. Explicitly declared alternatives form supplementary analysis groups, with findings attributed to the method that produced them. A group can inherit a decision when the answer explicitly states that it is shared. Method extraction follows the declared choices and compatible fixed question conditions. Table 19 summarizes the information connecting the answer to its scores.

The reviewer considers all distinct task-relevant findings, including additional claims and errors. Repeated statements are counted once; coordinates form one vector finding, while different statistics form separate findings. Printed values, signs, and units retain their association with the original text. This preserves the distinction between choosing an appropriate analysis, reporting a supported result, and attributing that result to a compatible method.

Table 19: Information linking an answer to its evaluation. Source passages establish what the answer claims; reference evidence establishes support for those claims.

Evaluation object	Information and role
Analysis group	Main or supplementary purpose, declared decisions, and supporting answer passages.
Atomic attribute	Dimension, fixed or open requirement, grade or unresolved status, source passage, and grading rationale.
Scientific finding	Quantity, object, scope, printed value and unit, analysis group, and source passage.
Reference comparison	Candidate branch and finding, semantic correspondence, numerical comparison, and compatibility with the declared method.
Requirement coverage	Matched core findings for F1; supported roles and acceptable sufficient combinations for F2.
Score summary	Metric value and availability, contributing finding counts, and method-binding coverage, associated with the case, model, and run.

G.2 Reviewer Protocol and Configuration

We use GPT-6 Astra with `max` reasoning to evaluate the saved answers. The reviewer receives the question, context, data metadata, frozen reference evidence, and line-numbered answer. It extracts analysis decisions and findings and assesses their scientific meaning without tool access; the host applies numerical comparison policies and computes the final metrics. Table 20 summarizes the request settings and division of responsibilities. Because GPT-6 Astra is also among the evaluated agents, Appendix I additionally reports a source-model-stratified analysis of the blinded human audit.

The reference-matching and atomic-review prompts are retained with their schemas in the materials indexed in Appendix K.3. A separate development review reassessed 825 saved answers across 96 cases from three models in anonymous case–trial batches. Its original prompts and execution settings are retained in `supplementary-protocol-details.tex`.

Table 20: Configuration and responsibilities of the benchmark evaluator.

Setting	Configuration
Reviewer model	GPT-6 Astra.
Reasoning effort	`max`.
Maximum output	10,000 tokens.
Request timeout	350 seconds per request.
Prompt length limit	200,000 characters.
Strict JSON Schema	Disabled by default; returned records undergo host-side validation.
Inputs	Question, context, data metadata, frozen references, and the saved answer with line numbers.
Reviewer responsibilities	Extract decisions and findings, grade operationalization attributes, and assess semantic correspondence. No tool access.
Host responsibilities	Validate structured records, check numerical agreement, and compute metrics under the fixed scoring rules.

G.3 Verification and Score Computation

Evaluation records retain the source passage, judgment, and rationale for each applicable item. Atomic grades are integers from 0 to 4, or `null` for unavailable judgments, and are normalized by four. Fixed conditions with a unique interpretation can be inherited from the question; missing required open choices receive zero. A separate explanation item assesses a requested rationale. Claim-level diagnostic grades support review but do not add benchmark metrics.

Before computing scores, the host checks that required items are present once, source passages and quotations are valid, and referenced evidence resolves. A review-task execution permits one retry for transport, parsing, or structural errors. If the retry fails, affected scores remain unavailable. Detailed schemas and execution settings are included in the extended protocol material.

Finding support follows one-to-one matching within a common reference branch. The branch with highest recall is selected for each analysis group, with precision breaking ties; unresolved candidates retain their possible outcomes. Consistency then checks evidence compatible with the group's declared method. Supplementary groups contribute to precision and consistency, while recall concerns the main answer. The resulting scores and availability states are aggregated as described in Appendix H.4.

G.4 Reference Coverage Boundary: Blunt Fin, GPT-5.6 Luna, Trial 2

The O2–F1 response `A6` from `gpt-5.6-luna` illustrates a reference coverage boundary: the answer specifies an excess-speed analysis outside the registered peak/mean reference branches. Its operationalization can be graded, and reconstruction of the declared analysis provides evidence for individual findings. The question fixes six-neighbor connectivity and mean-point location, while leaving the retention criterion and strength measure open:

> Within the blunt-fin flow field, select the strongest high-speed flow region. Choose and state an appropriate criterion for retaining high-speed locations and a property measure for quantifying region strength. Treat locations connected through immediate grid neighbors along one grid axis as one contiguous region, represent the selected region by the mean spatial location of its retained high-speed locations, and report both that location and its strength.

Answer and source binding. The substantive answer is reproduced below with its original source-line identifiers. Blank lines and the operationalization and finding headings are omitted.

> **L3:** I computed velocity as momentum divided by density and used its magnitude (U). The high-speed criterion was (U>U_ref), with (U_ref=2.95000005), the broad free-stream plateau (also the leading reference value in the PLOT3D solution header). I formed face-connected components, allowing neighbors differing by ± 1 in exactly one grid index. Region strength was the integrated excess-speed score ($S = \Sigma_{p \in R}$(U_p−U_ref)), using equal weight per retained grid point. The reported location is the arithmetic mean of the xyz coordinates of the retained points.
>
> **L7:** The strongest region contains 47 retained points and has mean location
>
> **L9–L11:** [$(\bar{x}, \bar{y}, \bar{z}) = (-0.54422, 0.27433, 3.57132)$.]
>
> **L13:** Its strength is (S=5.09276) in stored-speed units per grid point. For intensity context, its mean retained speed is (3.05836) and its peak speed is (3.12293). There were 718 retained points in 77 face-connected regions; the next-highest excess-speed score was 3.13052.

Atomic grades and operationalization scores. Table 21 gives all 12 attribute grades. L3 identifies the field conversion, threshold, connectivity, strength operator, and location operator. The strength formula defines an unnormalized point sum, whereas L13 describes it as a per-point quantity. The semantics attribute therefore receives grade 3. This method-level defect and the support check on the reported strength concern different evaluation targets.

Averaging the recorded attribute grades gives

$$S_O = \frac{1 + 1 + 11/12 + 1}{4} = \frac{47}{48} \approx 0.979167, \qquad \mathrm{URS} = \frac{1 + 11/12}{2} = \frac{23}{24} \approx 0.958333.$$

Criterion and property measure are the two dimensions selected for URS in this O2 case.

The chosen strength $S(R) = \sum_{p \in R}(U_p - t) = |R|(\bar{U}_R - t)$ combines retained-point count and excess speed. It defines cumulative excess speed on the fixed sampling grid, with greater scores for larger point sets at equal mean excess speed. Equal point weights make it sensitive to sampling density; a physical volume integral would instead use spatial volume weights.

Table 21: Blunt Fin attribute grades and dimension averages. F/O denotes fixed/open obligations. Sources refer to the answer above.

Dimension	Attributes	Grades	s_d	Evidence and rationale
Feature definition (F)	identity, domain, construction	4,4,4	1	L3: speed field, grid population, six-neighbor components.
Criterion (O)	rule, cutoff, boundary	4,4,4	1	L3: explicit threshold and strict comparison.
Property measure (O)	quantity, parameters, semantics	4,4,3	$11/12$	L3: specified sum and population. L13: per-point interpretation.
Aggregation/ representation (F)	weights, normalization, representation	4,4,4	1	L3: equal weights, point-count denominator, mean xyz.

Finding extraction. The answer contains nine scientific findings, C1–C9 in Figure 9. C3 (location) and C4 (strength) are the two core findings. The remaining seven comprise six supporting statistics and the empirical plateau assertion C1. The coordinate triple forms one finding. The threshold and operator definitions belong to the operationalization; the header-value attribution identifies input provenance. The physical interpretation of that value is assessed separately as C1.

Figure 9 compares the reported findings with a reconstruction of the declared analysis. The C4 value agrees with the unnormalized sum, whereas the per-point excess speed is $S/47 \approx 0.108356579$. Its per-point label therefore changes the meaning of the reported quantity. The diagnostic grades for claim quality, attribution, and conditional support are $3, 4, 3$. Full-precision values are retained in `supplementary-protocol-details.tex`.

Numerical reconstruction. The PLOT3D input contains $40 \times 32 \times 32 = 40{,}960$ points. Big-endian float32 density, momentum, and coordinates are promoted to float64 before computing $U = \|\mathrm{Momentum}/\mathrm{Density}\|_2$. The comparison uses the declared decimal $t = 2.95000005$ with strict $>$ and no boundary epsilon. Grid points are indexed with x varying fastest, grouped by six-neighbor connectivity, and ranked by $S(R)$. The header value 2.950000047683716 is not substituted for t.

The VTK cross-check uses automatic format detection, scalar function 153, vector function 200, and preserved intermediate functions. It agrees on the retained mask, winning-component membership, 717 retained points, and 76 components. Its winning sum is 5.092759023520, differing from the direct calculation by 1.78428×10^{-7}. This is an implementation check on the same files; intermediate speed arrays can differ through reader precision.

The reconstruction checks the four five-decimal scalar results with tolerance 5×10^{-6}, counts by exact equality, and location with the existing spatial tolerance 0.05. These checks assess reproduction of the printed values in this diagnostic reconstruction. The registered peak/mean branches use the frozen strength tolerance 0.005, corresponding to three-significant-figure reference reporting precision; their finding scores use the registered comparison policies.

Result branches and method compatibility. All candidate result branches are compared before considering method compatibility for C_{OF}. Table 22 records the core numerical comparisons. A strength claim is compared with the same statistic: C6 with a peak reference, C5 with a mean reference, and C4 with the excess-speed sum.

The peak value is a numerical candidate match, but the registered evidence does not establish that C6 describes the same region as the reference peak. Finding matches require agreement in quantity, object, and scope. Among eligible result matches, the evaluator selects the branch with the highest core recall and breaks ties by precision; method compatibility is assessed separately for C_{OF}.

Metric accounting and evidence status. The A6-alt reconstruction supports C2, C5, C6, and C9. C3, C4, C7, and C8 fail their respective checks, while the physical interpretation in C1 remains unverified. Thus, four of the eight numerical findings pass, including neither core finding.

Blunt Fin: a declared method and its findings

GPT-5.6 Luna · Trial 2 · O2–F1 · Reference coverage boundary

Task and data

40 × 32 × 32 PLOT3D grid. Fix six-neighbor connectivity and mean-point location; choose a high-speed rule and a strength measure. Report location and strength.

Declared analysis (answer L3)

U is the magnitude of momentum / density. Set t = 2.95000005 and retain U > t. Select the six-neighbor region maximizing S = Σ(U − t); report mean point coordinates.

Findings: answer and method-conditioned evidence

Answer	A6-alt evidence	Check
C3 / L7–10: Mean location (−0.54422, 0.27433, 3.57132)	(−0.561799, 0.201613, 3.571317)	**Error 0.074812 Tolerance 0.05**
C4 / L13: Strength 5.09276 “per grid point”	Sum: 5.09275920 Per-point: 0.10835658	**Value agrees; label fails**
C2 / L7: 47 selected points	47	**Pass**
C5, C6 / L13: Mean / peak 3.05836 / 3.12293	3.05835663 / 3.12293204	**Both pass**
C7, C8 / L13: Retained points 718; components 77	717 76	**Both fail**
C9 / L13: Runner-up 3.13052	3.13052169	**Pass**
C1 / L3: Threshold is a free-stream plateau	Physical meaning unresolved	**Unverified**

Recorded grades: S_O = 47/48 ≈ 0.979167; URS = 23/24 ≈ 0.958333.

A6-alt provides diagnostic checks; benchmark finding scores are N/A.

Figure 9: Reference coverage boundary case: Blunt Fin, GPT-5.6 Luna trial 2. Answer values are compared with the A6-alt reconstruction of the declared analysis. Spatial tolerance is 0.05; scalar checks use the five-decimal rounding tolerance 5×10^{-6}, with exact counts. Finding metrics remain N/A under the registered reference evidence. S_O and URS follow the recorded atomic grades.

Table 22: Core numerical comparisons for Blunt Fin. A numerical candidate match also requires quantity and object/scope correspondence. A6-alt is the example's method-conditioned reconstruction.

Branch	Location error	Strength comparison
q90_peak	$2.948214424 > 0.05$	C6 vs. 3.122931938319: error $1.938319 \times 10^{-6} < 0.005$; object correspondence unresolved.
q90_mean	$2.948214424 > 0.05$	C5 vs. 2.922646208401: error $0.135713791599 > 0.005$.
q75_mean	$3.155650715 > 0.05$	C5 vs. 2.698664778055: error $0.359695221945 > 0.005$.
A6-alt	$0.074812075 > 0.05$	C4 value agrees with the sum; its per-point label fails.

The operationalization receives $S_O = 47/48$ and URS $= 23/24$. The registered reference evidence leaves R_F, P_F, and C_{OF} as N/A for this case. This boundary case illustrates the distinction between a specified analysis, evidence for its individual findings, and coverage by the reference set. Reference extensions follow Appendix F.

H Experimental Setup and Additional Results

H.1 Model and Evaluator Settings

The nine evaluated models generate answers through their providers' official APIs in the common analysis environment described in Appendix B.5. Table 23 lists the model names, request identifiers, API providers, snapshot dates, and experimental run dates. Each model uses its highest available reasoning-effort setting, with a 900-second time limit per case attempt and no explicit output-token cap imposed by the benchmark. Saved answers are assessed by GPT-6 Astra using the same metric definitions and frozen reference evidence. The evaluator uses `max` reasoning effort, a 10,000-token output limit, and a 350-second timeout per request; its full configuration is given in Table 20. The generation settings for the illustrative development responses are reported with those examples in Appendix C.

H.2 Runtime and Experiment Versions

All 2,592 main-experiment runs use the `flow-python-v1` network-isolated runtime on Linux with Python 3.12 and frozen package versions. Table 24 lists the analysis dependencies and the additional packages included in the frozen delivery environment. NumPy and SciPy support numerical analysis, VTK reads and processes the flow fields, and Matplotlib supports plotting.

The runtime provides the virtual paths `/case` and `/workspace`, a persistent Python interpreter within a trial, and textual execution feedback. Network access is disabled, and the system instruction prohibits network use and host-filesystem searches.

Each trial configuration identifies the case input, dataset manifest, `flow-python-v1` runtime profile, system instruction, tool specification, and provider/model settings. Answer evaluation separately identifies the reference package, reviewer configuration, and schema. Hardware, runtime fingerprints, backend selection, and reference and evaluation versions identify the execution and scoring environments.

H.3 Runtime and Token Usage

Answer collection and evaluation coverage. Follow-up collection revisited unfinished instances, including interruptions in provider API communication, within the three-run design. Resource statistics use the 288 completed runs per model. Appendix H.4 gives the corresponding scoring coverage.

Resource accounting. Table 25 reports mean resource use per completed run. Elapsed time includes Python execution, request waiting, and within-run retries, and excludes scientific scoring and

Table 23: Model identities and run dates for the nine answer-generating models. All models use their highest available reasoning-effort setting, with a 900-second limit per attempt and no explicit output-token cap imposed by the benchmark. Model order follows Table 2.

Model	Request model ID	Snapshot date	Run dates	Official API provider
Claude Fable 5.1	`claude-fable-5-1`	Not identified	2026-09-13 to 2026-09-16	Anthropic
GPT-6 Astra	`gpt-6-astra`	Not identified	2026-09-22 to 2026-09-23	OpenAI
GPT-5.6 Sol	`gpt-5.6-sol`	Not identified	2026-09-21 to 2026-09-23	OpenAI
Qwen3.8-Max	`qwen3.8-max`	2026-09-02	2026-09-23	Alibaba Cloud[a]
GPT-5.6 Terra	`gpt-5.6-terra`	Not identified	2026-09-13 to 2026-09-16	OpenAI
Gemini 3.8 Flash	`Gemini 3.8 Flash`	Not identified	2026-09-23	Google
Claude Sonnet 5	`claude-sonnet-5`	Not identified	2026-09-13 to 2026-09-16	Anthropic
GPT-5.6 Luna	`gpt-5.6-luna`	Not identified	2026-09-13 to 2026-09-16	OpenAI
DeepSeek-V4.1-Flash	`deepseek-flash`	2026-09-10	2026-09-23	DeepSeek[b]

"Not identified" denotes an unconfirmed snapshot date. Snapshot dates are distinct from training-data or knowledge cutoffs.
[a]Qwen API base URL: `https://ws-nptxbo6gj0b0h4ah.cn-beijing.maas.aliyuncs.com/compatible-mode/v1`.
[b]DeepSeek API base URL: `https://api.deepseek.com`.

Table 24: Software environment for analysis and the additional dependencies in the frozen delivery environment.

Component	Version or setting
Operating system	Linux
Python	3.12
NumPy	1.26.4
SciPy	1.13.1
Matplotlib	3.8.4
VTK	9.6.2
Pydantic	2.5.3
PyYAML	6.0.1
Additional packages in the frozen delivery environment	
h5py	3.16.0
jsonschema	4.26.0
packaging	26.3

earlier unfinished attempts. Input tokens accumulate across requests, including reprocessed history. Output tokens include reasoning tokens once, following each provider's accounting convention.

Table 25: Mean resource use over 288 completed task runs per model. Runtime includes analysis and request waiting; input and output counts accumulate over the run, including reasoning tokens in output. Model order follows Table 2.

Model	Runtime (s)	Input tokens	Output tokens
Claude Fable 5.1	749.87	661,878.62	78,322.75
GPT-6 Astra	565.77	519,825.09	33,250.65
GPT-5.6 Sol	407.19	508,540.15	31,131.86
Qwen3.8-Max	645.06	592,375.84	24,821.65
GPT-5.6 Terra	364.23	501,809.86	26,765.90
Gemini 3.8 Flash	195.06	492,375.84	32,821.65
Claude Sonnet 5	797.79	651,637.98	95,306.43
GPT-5.6 Luna	294.50	477,923.91	23,016.94
DeepSeek-V4.1-Flash	214.61	754,627.68	39,762.29

H.4 Scoring Coverage and Aggregation

Each model completed 288 case runs: three runs for each of the 96 cases. Evaluation reports cover 2,541 of the 2,592 completed answers, or 98.0% (Table 26). This answer-level coverage describes the availability of evaluation reports. Within a report, the availability of a particular metric follows the applicability and evidence rules in Appendix A.

Table 26: Answer-level scoring coverage for the experiments in Table 2. All planned task runs completed. An evaluation report may contain unavailable scores for individual metrics.

Model	Completed	Evaluated	Missing reports	Coverage
Claude Fable 5.1	288	284	4	98.6%
GPT-6 Astra	288	285	3	99.0%
GPT-5.6 Sol	288	283	5	98.3%
Qwen3.8-Max	288	283	5	98.3%
GPT-5.6 Terra	288	282	6	97.9%
Gemini 3.8 Flash	288	281	7	97.6%
Claude Sonnet 5	288	281	7	97.6%
GPT-5.6 Luna	288	280	8	97.2%
DeepSeek-V4.1-Flash	288	282	6	97.9%
Total	2,592	2,541	51	98.0%

Case-level scores are associated with the model, family, condition, and run. For metric k, let $E^{(k)}_{mct}$ be the cases with resolved, applicable scores for model m, condition c, and run t. The condition and run summaries are

$$\bar{s}^{(k)}_{mct} = \frac{1}{|E^{(k)}_{mct}|} \sum_{i \in E^{(k)}_{mct}} s^{(k)}_{mit}, \qquad \bar{s}^{(k)}_{mt} = \frac{1}{|\mathcal{C}_k|} \sum_{c \in \mathcal{C}_k} \bar{s}^{(k)}_{mct}, \tag{9}$$

where $\mathcal{C}_k$ contains all four conditions for S_O, R_F, P_F, and C_{OF}, and O2–F1 and O3–F1 for URS. Table 2 reports the mean and sample standard deviation of the three run summaries, using denominator $3-1$ for the variance. Equal condition weights preserve the task design when metric availability differs across conditions. O1 has no URS requirement, giving 144 planned applicable case–run instances per model for URS and 288 for each of the other metrics before answer-dependent exclusions.

Method-binding coverage describes a separate aspect of evaluation: the fraction of an answer's findings linked to an identifiable declared analysis. It accompanies C_{OF} because a consistency score concerns that bound subset. Answer-level coverage, metric availability, and method-binding coverage therefore refer to answers, case scores, and findings, respectively.

H.5 Performance across Task Conditions

Table 27 summarizes changes from O1–F1 using the condition means in Table 3. Mean S_O is lower by 0.0252 in O2–F1 and by 0.1035 in O3–F1. Recall, precision, and consistency also decrease. Under the specified O1 analysis, F2 yields lower recall and precision. This contrast combines the responsibility for selecting findings with the family's authored sufficiency requirements.

Table 27: Changes relative to O1–F1, calculated from the reported condition means in Table 3. Models receive equal weight. URS is omitted because it is inapplicable to the O1 baseline.

Condition	ΔS_O	ΔR_F	ΔP_F	ΔC_{OF}
O2–F1	-0.0252	-0.0178	-0.0185	-0.0175
O3–F1	-0.1035	-0.0469	-0.0470	-0.0459
O1–F2	-0.0251	-0.0332	-0.0336	-0.0310

The largest aggregate change is in operationalization quality under O3. The examples in Appendix C illustrate decisions behind this challenge: a strength measure can change the winning region, and a density-level rule can change the selected surface.

I Evaluation Reliability and Reference Coverage

I.1 Human Review Protocol and Coverage

Two doctoral students specializing in flow visualization reviewed 96 randomly selected model answers, one per case, using the protocol in Section 4. The sample spans all 24 scientific question families and four conditions, with 24 answers per condition. GPT-5.6 Luna, GPT-5.6 Terra, and GPT-6 Astra each contribute 32 trial-1 answers, eight per condition.

Reviewers received the scientific question, dataset context, ground truth, and complete answer, with model identities hidden during review. The first reviewer assessed every answer. The second reviewed the same answers with access to the first-round comments and produced the final review opinions. This sequential process yielded 192 review entries for 96 distinct answers. Both reviewers made the scientific judgments themselves; a language model assisted with textual summarization.

Review focused on three questions: whether the analysis met fixed constraints and made appropriate open choices, whether the findings had compatible numerical and semantic support, and whether that support matched the declared method. Supplementary claims, units, and physical interpretations were assessed on the same basis. The comments and rationales are retained in `human-review-r01.md` and `human-review-r02.md`.

Evaluator–expert agreement and cost. All nine candidate evaluators assessed the same 96 anonymized model-generated answers reviewed by the two doctoral students. The expert reference consists of the final adjudicated scores from the sequential human review, with each distinct answer serving as one evaluation unit. Expert and model-based evaluations use the same five metrics: S_O, URS, R_F, P_F, and C_{OF}. Pearson correlation is computed separately for each metric using answers with numerical scores from both the human reference and the candidate evaluator. N/A scores are excluded rather than assigned zero, so the number of valid pairs can vary by metric and evaluator. URS applies to O2–F1 and O3–F1. For each evaluator, the five metric-specific correlations receive equal weight in Fisher z space. The summary correlation in Figure 4 is

$$r_{\text{summary}} = \tanh\left(\frac{1}{5}\sum_{m\in\{S_O,\text{URS},R_F,P_F,C_{OF}\}}\operatorname{arctanh}(r_m)\right),$$

where r_m is the Pearson correlation between evaluator and final adjudicated expert scores for metric m. Cost is the mean API cost per evaluated answer in USD.

I.2 Verification Evidence and Review Findings

The two reviewers' sequential assessment of 96 answers identified seven answers for correction of at least one evaluation judgment. The recorded corrections concern operationalization attributes, finding support, or analysis-group attribution. Table 28 summarizes their distribution across answer sources. Judgments with insufficient evidence remain unverified.

Table 28: Judgment corrections recorded in the blinded human audit, stratified by answer source. Counts refer to distinct answers with at least one recorded correction; rates use the audited answers as the denominator. Adverse denotes a recorded change toward a less favorable judgment of the answer.

Answer source	Audited	Corrections	Rate	Adverse	Correction categories
GPT-6 Astra	32	1	3.1%	0	Analysis-group attribution (1)
GPT-5.6 Luna	32	1	3.1%	1	Finding support (1)
GPT-5.6 Terra	32	5	15.6%	0	Finding support (3); operationalization (1); analysis-group attribution (1)
Total	**96**	**7**	**7.3%**	**1**	Finding support (4); analysis-group attribution (2); operationalization (1)

The Astra correction concerns the Blunt Fin answer, where a violation of the task's operationalization constraints had been used to assign zero method–result consistency. The review separates these judgments and calls for verification under the declared method. The Luna correction concerns numerical finding support. The five Terra corrections restore support for alternative calculations or correct an operationalization deduction and a finding's analysis-group attribution.

These cases illustrate the value of evaluating task compliance, numerical support, and method–result consistency separately. A result can be supported under the declared calculation while its operationalization violates a fixed task constraint; conversely, a valid alternative analysis can require evidence beyond the stored reference branches. The framework exposes the affected judgment and its evidence, allowing a correction to be localized while other findings retain their own verification status.

The supplementary machine-readable file `human-review-corrections.csv` maps all 96 answers to their source model, review dispositions, correction category, affected metrics, verification status, and evidence references. Field definitions and aggregation rules accompany it in `human-review-corrections-README.md`, together with the two review records `human-review-r01.md` and `human-review-r02.md`.

Reviewers used stored reference values, comparison tolerances, and finding-level verification records. Five second-round case records additionally draw on targeted recomputation: MHD density–magnetic association under O1–F2, Office under O2–F1, and Rayleigh–Taylor mixing width under O1–F1, O1–F2, and O2–F1. The two O1 mixing-width cases share the same RMS-width and center calculation. Verification scope is recorded per finding, so a supported core result can coexist with unverified supplementary statistics.

Figure 10 presents three cases from the review: a confirmed numerical error, a withdrawn contradiction based on an incompatible quantile convention, and a method-specific consistency assessment. Each row identifies its task, model, trial, and evidence, illustrating how review distinguishes answer errors from evaluation errors.

Two further cases illustrate finding-level verification. For the Electrolyzer Electric current-density region under O1–F1, magnitude-first evidence supports the reported centroid, mean 0.0054037431271, and 49,108 cells. This establishes numerical support under the declared calculation order; its compliance with the task's fixed choices is assessed at the operationalization level. For Office O2–F1, targeted execution confirms 420 retained points, a 72-point winning region, and score 24.3457582074 under the selected analysis. The reported runner-up score of 23.389 remains unverified.



Figure 10: Review-stage evidence illustrating numerical error, a corrected quantile comparison, and separation of task compliance from method–result consistency. The GPT-6 Astra Blunt Fin trial-1 case is distinct from the GPT-5.6 Luna trial-2 reference coverage boundary case.

The review also identified assessment errors in the OpenFOAM Tubes tasks. For pressure–speed association under O1–F1, a feature-definition deduction applied a condition that did not belong to the task. Under O3–F1, a supplementary coefficient of -0.9213 already had a matching equal-cell Pearson reference, while its assessment marked it as contradictory. These cases led to corrections of the affected attribute judgment and the finding's analysis-group support under the scoring rules in Appendix A.

I.3 Reference Coverage and Interpretation

The review distinguishes reference coverage from answer correctness. Alternative populations, weighting schemes, connectivity rules, and quantile conventions require evidence for the same quantity and scope. Supplementary counts and spatial or causal interpretations retain their own evidence status even when core findings match. The per-family reference inventory in Appendix F.6 describes the available branches; the review records identify where additional evidence is needed.

J User Study Protocol and Analysis

The study measures participants' selection of analysis definitions, parameters, and findings with an interface that computes results, renders the flow, and assembles an editable answer. This section documents the support available, task exposure, and participant-level aggregation.

J.1 Participants and Training

Fifteen doctoral students with backgrounds in fluid mechanics or scientific visualization participated. None had participated in benchmark construction or seen the benchmark questions or answers before the study. All participants received a 20-minute introduction to the interface and its workflow before an approximately one-hour task session; the task-session duration excludes training. Participation was voluntary and uncompensated, and participants consented to participation and data use. The institutional ethics committee approved the study. Response data were anonymized, with participant identifiers used to associate task records within the study.

Training used synthetic values to practice analysis settings, object selection, view manipulation, and answer review without prescribing a benchmark solution. Practice responses were excluded. Familiarization screenshots are retained in `supplementary-protocol-details.tex`.

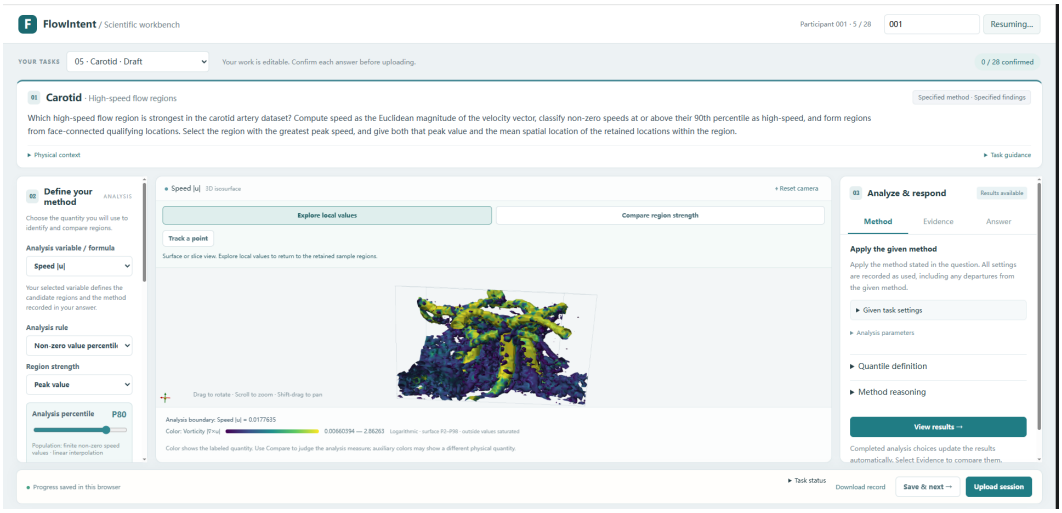
J.2 Interface and Analysis Support

The interface presents the question, physical context, and interactive 3D view. Participants choose formulas and parameters from the same library across O1/O2/O3, with no preselected or recommended scientific analysis. Available operations are defined by the library, and participants can select settings that differ from the prescribed choices. They select objects and evidence, choose F2 characterizations, and edit an answer assembled from their choices and computed quantities without ground truth access. Finding selection determines which results enter the submitted answer.

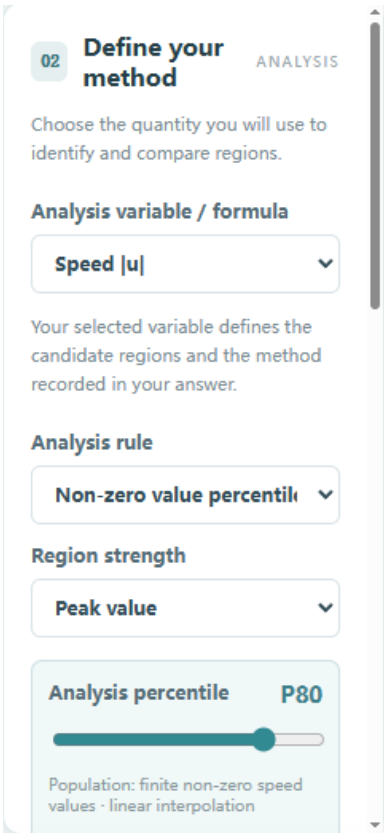
Figure 11 shows a participant-selectable P80 setting for a question that specifies P90. Analysis changes recompute results and clear earlier evidence selections, while display changes leave calculations unchanged. Participants review the resulting values and their edited wording before saving the final answer. Appendix J.5 describes the supported operations.

J.3 Task Assignment and Session Procedure

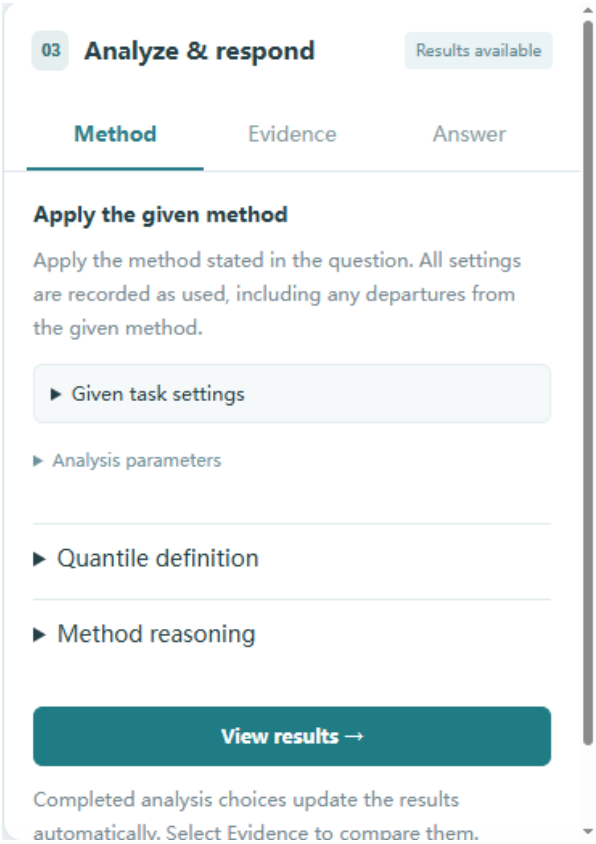
Each participant received 28 distinct cases, sampled as seven per condition, in randomized order, with revisits and revisions allowed. Condition identifiers were hidden. The 420 responses comprise 105 per condition and cover all 96 cases and 24 scientific question families, with each case seen by one to nine participants. Participants encountered 15–19 families, including 6–10 under multiple conditions. Thus allocation is balanced by condition but repeats some scientific targets.



(a) Task workspace: question, visualization, and analysis controls.



(b) Analysis choices and percentile setting.



(c) Method review and access to results.

Figure 11: Task interface illustrated with Carotid. Panel (a) shows the question, visualization, and response workspace; panels (b) and (c) enlarge the analysis settings and method review. The selected 80th-percentile setting differs from the prescribed 90th percentile, illustrating the participant's responsibility for the submitted analysis.

Table 29 reports exposure and time. Accumulated task time is 60.9 ± 2.5 minutes per participant (sample standard deviation), with a range of 57.4–65.3 minutes. It sums the 28 recorded task durations and excludes the 20-minute training.

Table 29: Recorded task coverage and accumulated task time. Every participant answered 28 distinct cases, seven per condition. Repeated families are families encountered under at least two conditions by that participant. Time is the sum of the 28 recorded task durations, excluding training.

Participant	Distinct families	Repeated families	Task time (min)
001	18	9	62.3
002	18	9	60.6
003	19	7	65.3
004	17	8	59.8
005	17	9	64.4
006	15	10	58.2
007	18	10	61.7
008	19	8	64.1
009	19	7	58.5
010	17	9	57.7
011	18	8	59.8
012	18	8	61.0
013	19	8	63.0
014	18	9	57.4
015	19	6	60.2

Participants selected an analysis, inspected its visualization, chose findings, and reviewed or edited the assembled answer before confirming, saving, and uploading it. Retained records link participant and case IDs, presentation order, task time, final text, and evaluation scores.

J.4 Scoring and Participant-Level Results

Final participant answers used the same scorer, reference-package version, evaluator model and configuration, and criteria as model answers, including the graded rubric in Appendix A.1. Each participant's condition score averages the seven applicable case scores equally. Table 30 reports the mean and sample standard deviation of these averages across 15 participants. Scores are available for all applicable participant responses: S_O, R_F, P_F, and C_{OF} for all 420 responses, and URS for the 210 O2/O3 responses. URS is inapplicable to O1. Standard deviations describe variation between participant averages.

Table 30: Participant performance with interactive analysis support. Scores are means $\pm$ sample standard deviations across the 15 participants' condition averages. URS applies to O2–F1 and O3–F1.

Condition	S_O	URS	R_F	P_F	C_{OF}
O1–F1	0.959 ± 0.032	N/A	0.932 ± 0.036	0.948 ± 0.035	0.962 ± 0.028
O2–F1	0.943 ± 0.027	0.858 ± 0.033	0.912 ± 0.040	0.926 ± 0.038	0.940 ± 0.034
O3–F1	0.847 ± 0.034	0.847 ± 0.034	0.893 ± 0.033	0.911 ± 0.035	0.932 ± 0.033
O1–F2	0.914 ± 0.035	N/A	0.882 ± 0.037	0.893 ± 0.038	0.913 ± 0.037

From O1–F1 to O3–F1, participant means decrease by 0.1125 in S_O and 0.0390 in R_F. From O1–F1 to O1–F2, they decrease by 0.0501 in R_F and 0.0546 in P_F. All 15 participants show decreases in these four comparisons. Each contrast compares a participant's condition averages over the assigned cases, which can differ in family composition. As in the model experiment, F2 includes both finding selection and its authored sufficiency requirements.

Participant scores assess the final interface-supported answer, including edited wording. The study characterizes analysis and reporting choices with interactive 3D support, while the model experiment uses Python execution and text-only feedback. Correctness judgments come from the shared eval-

uator, without a separate human correctness assessment. Method-binding coverage links extracted findings to the participant's declared analysis.

J.5 Formula Definitions and Finding Operations

Table 31 illustrates the interface's analysis operations through seven families: five speed-region tasks, Combustor density surfaces, and Kitchen concentration comparison. Their O1 recipes show how field selection, spatial objects, measurements, and findings are connected. Table 13 lists all 24 scientific question families.

Fields, selections, and measurements. Field identity includes both the array name and its point or cell association. For point velocity $\mathbf{u}_i$, speed is $s_i = \|\mathbf{u}_i\|_2$. A percentile selection first computes a linearly interpolated, unweighted quantile Q_q on its stated population, then retains points at or above that threshold. The speed-region percentile uses finite non-zero speeds, whereas the Combustor isovalue uses all finite stored Density values. Absolute thresholds retain the comparison operator specified by the analysis, including the strict inequality in Office.

Connectivity acts on the retained population. Structured-grid six-neighbor connectivity joins retained points separated by one index step along one axis. FireFlow instead uses point adjacency through a shared cell. Surface connectivity operates on the extracted isosurface. For a point region R, the peak and mean speed and the representative location are

$$p(R) = \max_{i \in R} s_i, \qquad \bar{s}(R) = \frac{1}{|R|} \sum_{i \in R} s_i, \qquad \bar{\mathbf{x}}(R) = \frac{1}{|R|} \sum_{i \in R} \mathbf{x}_i. \tag{10}$$

Peak and mean speed define different regional strength measures. The interface lists candidate values under the selected measure, and the participant chooses the object to report.

For a triangulated surface, each triangle contributes area $A_t = \frac{1}{2}\|(\mathbf{b}_t - \mathbf{a}_t) \times (\mathbf{c}_t - \mathbf{a}_t)\|_2$. Surface area is $\sum_t A_t$, and the representative position is the triangle-centroid average weighted by A_t. Kitchen instead forms each cell concentration from its finite vertex values and computes the cell-volume-weighted mean, population standard deviation, and coefficient of variation as specified in Appendix C.4. Its candidate fields are `c1` and `c11`–`c17`. These point, surface, and cell representations have distinct weighting conventions. Coordinate values retain the stored coordinate scale when physical units are unspecified.

Table 31: Example analysis recipes and reporting elements. The recipe column specifies O1 analyses for seven scientific question families; additional characterization is selected by the participant under F2.

Scientific question families	Analysis recipe	Reporting elements
Blunt Fin, Carotid, FireFlow, NASA LOx Post	Point speed; non-zero-speed linear 90th percentile; connected regions; comparison by peak speed; mean retained-point location. FireFlow uses shared-cell point adjacency; the other three use six neighbors.	F1: selected region's location and strength. F2: retain both and add a characterization, such as extent, comparison, or a participant observation.
Office	Point speed strictly above 0.20; six-neighbor regions; comparison by peak speed; mean retained-point location.	F1: location and strength. F2: retain both and add a characterization.
Combustor	Linear 90th percentile of finite point Density; isosurface extraction; connected components; comparison by surface area; area-weighted centroid.	F1: location, density level, and area. F2: identity and density level, with a selected extent, location, or shape characterization.
Kitchen	Finite-vertex cell means for each concentration field; positive-volume weighting; comparison by coefficient of variation.	F1: selected field and heterogeneity value. F2: retain field identity and measure evidence, with an additional characterization.

Finding selection and answer assembly. The finding operations associate a reported quantity with its object or domain, field, measure, and spatial representation. Table 32 summarizes the documented reporting operations for these examples. Object and field comparisons use values of the same measure; a position comparison identifies its coordinate axis. An axis-aligned extent is the coordinate range $\Delta a = \max_{i \in R} a_i - \min_{i \in R} a_i$. Principal orientation requires a separate directional measure. Distribution summaries require the underlying mean, standard deviation, or quantiles, rather than being reconstructed from a single heterogeneity coefficient.

The participant chooses the object and evidence used to assemble the answer. Template slots are populated by the selected results and participant input. Free-text observations retain the participant's wording and are evaluated as part of the final response. The content produced by a reporting operation is assessed against the F2 sufficiency rules in Appendix F.

Table 32: Finding operations and their associated evidence for the example scientific question families. Availability depends on the dataset and object type.

Reporting operation	Associated evidence and interpretation
Metric, location, or level	Numerical value and measure, representative coordinates and representation, or the field and threshold/isovalue defining the object.
Object selection	Participant-selected identity and candidate metrics under the chosen comparison measure.
Spatial extent	Coordinate minima and maxima of the selected object and the corresponding axis spans.
Metric or position comparison	Values for the compared objects or fields, with a common measure or a specified coordinate axis.
Distribution summary	Mean, population standard deviation, or P10–P90 interval, associated with the field, domain, weighting, and quantile definition.
Surface geometry	Triangle area and area-weighted centroid for the selected surface component.
Auxiliary quantity or count	An auxiliary field sampled on the selected object, or a count identifying whether its entities are samples or objects.
Scientific observation	Participant-authored interpretation linked to the selected object or domain and supporting evidence.

Each task offers operations and parameter ranges appropriate to its fields and scientific target. The submitted answer retains the participant's selected analysis and findings.

K Prompts and Reproducibility Materials

K.1 Prompts for Context and Task Construction

Authoring starts from the frozen target, scope, fixed clauses, open dimensions, finding requirements, and dataset context. The template produces two or three phrasings without adding obligations or selecting an optional F2 characterization. A check after rewriting verifies unchanged meaning, followed by the nine final wording checks in Appendix E.5. Construction and evaluation evidence are not exposed as respondent instructions.

The authoring excerpt and complete final wording-review instruction are retained verbatim in `supplementary-protocol-details.tex`. Context preparation and candidate-analysis scientific review are separate stages, each with its own inputs and execution records.

K.2 Model Instructions and Tool Specifications

Figure 12 reproduces the system instruction used in all 2,592 main-experiment runs and summarizes the model-visible packet. These runs use the network-isolated `flow-python-v1` runtime with Linux, Python 3.12, and the frozen package versions listed in Appendix H.2.

The Python tool accepts a required `code` string and no additional arguments, returning textual execution feedback. Metadata in the Scientific question, Flow data, and Case context packet are

Model instructions and supplied information

Main experiment · 2,592 runs · Network-isolated flow-python-v1

System instruction (verbatim)

Answer the supplied scientific question. Analyze the supplied flow data when needed; a persistent Python environment is available. When the question leaves scientific analysis choices unspecified, select appropriate choices yourself. Runtime paths are fixed and non-scientific: read-only case files are under /case, /case/case_input.json contains the model-visible case input, /case/reader_metadata.json contains reader-level decoding metadata when provided, /case/case_files.json lists the bounded visible file set, and /case/runtime_contract.json lists runtime capabilities. NumPy, SciPy, Matplotlib, VTK are available; network access is unavailable. Write scripts and intermediate artifacts only under /workspace. Inspect those paths directly; do not search the host filesystem or use the network. Your final response must be natural language with these two sections and no JSON: '## Operationalization' describing the consequential analytical choices you actually used, followed by '## Finding' describing the result supported by the data.

Model-visible packet

- User message
- Scientific question
- Flow data
- Case context
- Staged files
- Data and declared geometry
- Case input and reader metadata
- Runtime contract and file index

Python tool and information boundary

The tool accepts a required code string and returns textual output or errors. Python state persists within a trial. Reference analyses, ground truth and construction records are excluded from the model-visible files.

All 2,592 main-experiment runs use Linux, Python 3.12, frozen package versions, and the network-isolated flow-python-v1 runtime.

Figure 12: System instruction used in all 2,592 main-experiment runs under the network-isolated flow-python-v1 runtime, with a summary of the model-visible packet and tool boundary. The quoted system text is transcribed from the retained supplementary protocol.

JSON-serialized with sorted keys. The full tool description, JSON schema, Office packet example, and staging specification are retained in `supplementary-protocol-details.tex`.

K.3 Artifact Schemas, Access, and Reproduction Steps

The 96-case manifest connects each question to its family, condition, dataset, context, and references. Reproduction follows these links to prepare the data, collect an answer under the retained model and runtime configuration, and evaluate it against the associated reference package. Table 33 lists the supporting materials. Case, model, run, group, and finding identifiers connect each score to its answer and evidence.

Table 33: Supporting materials connecting task construction, answer evaluation, and reported results.

Material	Contents and use
Tasks and conditions	Case manifest, scientific questions, family specifications, fixed/open attributes, finding requirements, and wording reviews.
Data and context	Numerical inputs, reader metadata, file digests, source-linked physical context, and dataset access and licensing information.
Reference evidence	Complete analysis branches, executed findings, numerical comparison policies, and sufficient-role combinations.
Answers and execution	Answer text, case/model/run identity, completion status, model configuration, `flow-python-v1` runtime profile, and tool trajectory.
Assessment evidence	Atomic grades and rationales, cited answer passages, finding groups, reference matches, role coverage, and selected branches.
Scores and summaries	Case scores and availability; condition and run means, repeated-run standard deviations, and scoring coverage.
Human audit	Two sequential review records and a machine-readable correction mapping with source-model identity, affected judgments, verification status, and evidence references.
Instructions and schemas	Construction and reviewer prompts, model instructions, tool specification, and structured assessment formats.