ADJUVANTCHART: BENCHMARKING MULTI-PANEL RELATIONAL UNDERSTANDING IN REAL-WORLD SCIENTIFIC CHARTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Scientific chart understanding requires models not only to read numerical values but also to identify experimental conditions and integrate evidence across panels to draw conclusions. Existing benchmarks focus primarily on visual information extraction, numerical reasoning, and single-turn scientific question answering. They place less emphasis on how models organize and use chart evidence in real experimental contexts. We introduce AdjuvantChart, a scientific chart understanding benchmark built from published adjuvant research papers. Guided by domain experts' needs in literature analysis, we design three tasks: structured data extraction, experimental relation verification, and cross-panel composition. For the latter two tasks, we construct question families grounded in traceable atomic facts, comprising local, compositional, and condition-variation questions. We evaluate both answers and supporting evidence to assess whether models consistently use the same underlying chart evidence across related questions. AdjuvantChart contains 600 high-quality composite figures from published papers. The experimental relation verification and cross-panel synthesis tasks include 380 question families and 1,134 question-answer pairs. In our evaluation, GPT-5.5 achieves the highest exact-match accuracy of 81.71% on the question-answering tasks, while Gemini-3.7-Flash achieves the highest RMS-F1 score of 54.92% on structured extraction. Joint evaluation of answers and supporting evidence changes the relative ranking of some models, indicating that evaluating final answers alone does not fully capture models' ability to use evidence from scientific charts. AdjuvantChart provides a new testbed for evaluating multimodal models' chart understanding and evidence use in experimental contexts.

1 INTRODUCTION

Researchers often read experimental charts not to obtain an isolated numerical value, but to answer a question under specific experimental conditions (Li et al., 2024). This requires identifying the change observed for a treatment, the control used for comparison, and the measurement time point at which the change occurs. It also requires determining whether different measured variables jointly support the same conclusion. For example, when researchers assess whether a treatment improves two measured outcomes, identifying the tallest bar in each of two panels is not sufficient. These observations must also correspond to the treatment group, comparison target, and measurement conditions specified in the question. Scientific chart understanding therefore involves transforming visual information into scientific evidence. As shown in Figure 1, using charts as scientific evidence requires more than organizing observations into structured records. It also requires verifying the relevant experimental conditions and comparators and integrating evidence across panels according to the constraints specified in the question. This process requires both understanding experimental semantics and reasoning across panels.

Recent advances in large multimodal models have driven substantial progress in chart understanding (Huang et al., 2024). Benchmarks such as ChartQA (Masry et al., 2022) and ChartQAPro (Masry et al., 2025) evaluate how models combine visual information with numerical and logical operations. CharXiv (Wang et al., 2024) and SPIQA (Pramanick et al., 2024) further extend this research to real-world scientific charts and multimodal information access in scientific papers. These studies

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.

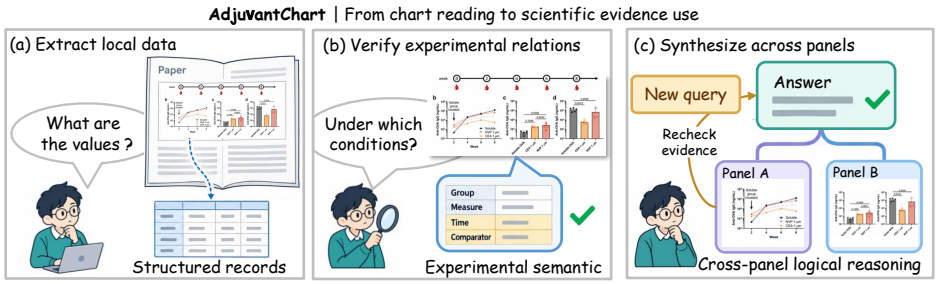

Figure 1: Motivation for AdjuvantChart: from chart reading to scientific evidence use. Scientific chart analysis involves (a) extracting observations into structured records, (b) interpreting observations in terms of their experimental groups, measurements, time points, and comparators, and (c) combining evidence across panels to form supported answers. When query conditions change, the relevant evidence must be re-examined. AdjuvantChart connects these activities to evaluate experimental semantic understanding and cross-panel logical reasoning.

show that charts are not merely visual objects to be recognized but also sources of information for reasoning. However, when compiling scientific evidence, knowing whether a model can answer a question is not enough. The same answer may correspond to different supporting records, and the applicability of the same set of observations may change with the comparison target or query scope. Evaluating scientific chart understanding therefore requires considering both whether answers are correct and how observations are interpreted and used.

Organizing chart observations into evidence for scientific conclusions requires addressing two challenges. The first is *experimental semantic understanding*: values, trends, and statistical annotations are meaningful only within specific experimental relationships. Axes, legends, comparison brackets, and captions must be integrated to identify each observation's group, measurement, time point, and comparison target. The second is *cross-panel logical reasoning*: observations from panels covering different measurements, time points, or experimental subjects must be aligned and combined under query constraints to support valid conclusions. Changed constraints require reassessing evidence rather than carrying over previous conclusions without verification. The challenge is thus not merely handling multiple charts or panels, but organizing their evidence according to experimental semantics.

For diagnostic evaluation, we organize related tasks around shared chart evidence rather than treating questions independently, progressing from basic observation records to cross-panel integration. Models must reassess evidence when comparators, measurement conditions, or filtering scope change, allowing the same atomic facts to serve different roles across questions. Beyond answer correctness, this design distinguishes errors in information extraction, interpretation of experimental relations, evidence composition, and application of query constraints. Existing multi-panel question answering and scientific claim verification emphasize cross-chart reasoning and evidence assessment, respectively. We connect these capabilities within a unified framework where related questions share underlying evidence, enabling analysis of how models read, interpret, and reuse it under different query conditions.

Motivated by these considerations, we introduce AdjuvantChart. The benchmark derives its tasks from domain experts' needs in literature reading, result organization, and evidence verification, and uses real experimental figures from published studies on adjuvants. This domain provides clear group comparisons (Chen et al., 2026), multi-metric measurements, and multi-panel figure settings. The research focus, however, is on the multimodal capabilities required to understand and use chart evidence rather than on additional biological knowledge. All visual materials are directly extracted or cropped from the original papers, preserving their authentic layouts and annotations. Given the figures and necessary context, systems perform structured extraction, experimental relationship verification, and integrated cross-panel analysis, forming a task sequence from reading experimental results to organizing scientific evidence. Our main contributions are as follows:

- We introduce a chart evaluation perspective centered on scientific evidence use. By linking result extraction, relationship verification and integrated analysis, AdjuvantChart evaluates whether chart observations can support verifiable judgments in experimental constraints.

- We construct AdjuvantChart from real scientific figures with shared factual annotations. Traceable facts, executable rules, and linked question families support local, compositional, and condition-variant queries, together with joint evaluation of answers and supporting records.
- We systematically evaluate multiple models across different output levels. Our results reveal distinct performance patterns between comprehensive data recovery and targeted question answering, and show that similar answer accuracy can mask substantial differences in evidence completeness and relative model rankings.

2 Related Work

2.1 General Chart Understanding and Structured Representation

Chart question answering combines visual reading with numerical and logical operations and has become a major paradigm for evaluating chart understanding (Xie et al., 2026). ChartQA (Methani et al., 2020) includes both human-authored questions and questions generated from human-written summaries, emphasizing open-vocabulary answering and joint visual-logical reasoning. ChartQAPro (Masry et al., 2025) further broadens chart sources and question formats to include dashboards, infographics, hypothetical questions, conversational questions, and unanswerable cases, thereby increasing the realism and diversity of evaluation (Zhu et al., 2025). These benchmarks (Li et al., 2026) have substantially expanded the range of tasks and data distributions covered by question answering on charts. However, they generally still use individual question-answer instances as the basic unit of evaluation, focusing primarily on whether models can derive correct answers from charts.

Another line of work evaluates or supports chart understanding through explicit intermediate representations·(Wei et al., 2025). DePlot (Liu et al., 2023) converts charts into linearized tables before passing them to a language model for reasoning, separating data recovery from downstream inference. ChartMimic (Yang et al., 2025) links charts, textual instructions, and executable plotting code, evaluating the combination of visual understanding and code generation through direct reproduction and customized reproduction with new data. Structured extraction in AdjuvantChart also emphasizes reusable representations, but its goal is to preserve experimental results together with their associated conditions and relations as a basis for organizing scientific evidence, rather than to reconstruct the underlying plotting program.

2.2 Scientific Chart Understanding and Evidence Analysis

Scientific chart benchmarks address the more complex layouts, terminology, and semantic relationships found in figures from research papers. SciFIBench (Roberts et al., 2024) evaluates scientific figure interpretation through bidirectional matching between figures and captions, using hard negatives and human verification to improve task discrimination. CharXiv (Wang et al., 2024) constructs descriptive and reasoning questions from real scientific charts, highlighting the challenges posed by naturally occurring research figures. SPIQA (Pramanick et al., 2024) approaches the problem from the perspective of information seeking in scientific papers, incorporating figures, tables, and relevant text into question answering and analyzing model performance under different context conditions and retrieval-augmented settings.

More closely related to our work are studies on cross-panel reasoning (Li et al., 2026) and scientific evidence verification (Borisova et al., 2025). Efat et al. (2026) evaluates panel identification, selection, and joint reasoning over multi-panel scientific figures. ClimateViz (Su et al., 2025) assesses whether chart evidence supports, refutes, or provides insufficient evidence for scientific claims, while also producing structured explanations. AdjuvantChart examines whether models can consistently use the same chart evidence correctly under the constraints imposed by experimental conditions. We connect structured extraction, experimental relation verification, and cross-panel analysis through shared atomic facts. We construct linked question families around these facts to assess whether models can reselect and reorganize evidence when comparators or query conditions change.

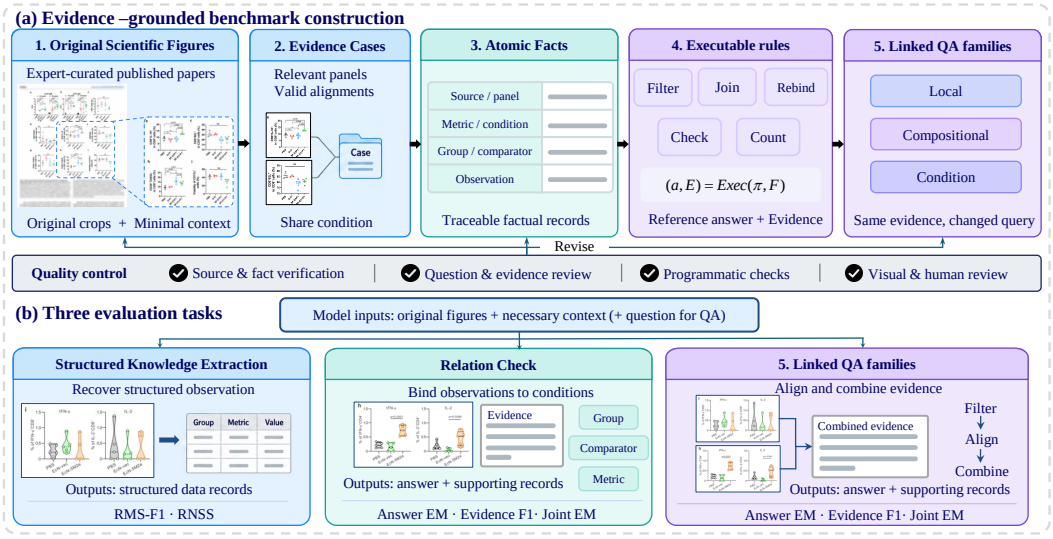

Figure 2: Overview of the AdjuvantChart construction pipeline. Original scientific figures are annotated with traceable facts to construct linked question families, with programmatic checks and human review for quality control.

3 AdjuvantChart

3.1 Task Definition

AdjuvantChart organizes scientific chart analysis into three tasks. Structured data extraction corresponds to result organization, experimental relationship verification to local evidence confirmation, and cross-panel composition to the joint use of multiple results. All three tasks require understanding the experimental context, but models are not required to follow a fixed sequential pipeline. Figure 2 provides an overview of the benchmark construction process and task definitions.

Structured Data Extraction. Given a chart and necessary context, the model recovers categories, series, values, and, when available, units and measurement assignments while preserving their label-value correspondence. The target is information explicitly presented in the figure rather than unobserved experimental data.

Experimental Relationship Verification. Given a query, the model verifies observations and comparison relations under specified experimental conditions, including statistical annotations, effect direction, significance, and measurement coverage for particular groups or time points.

Cross-Panel Composition. The model combines records across panels for joint verification, conditional filtering, cross-metric queries, or set-based judgments. These operations extend the semantics of individual observations to relations among multiple pieces of evidence.

For the QA tasks, the model outputs an answer together with a concise supporting record:

$$(\hat{a}_q, \hat{E}_q) = f_\theta(\mathcal{I}_q, C_q, q), \qquad \mathcal{I}_q = (I_{q,1}, \ldots, I_{q,m_q}), \quad m_q \geq 1. \tag{1}$$

where $\mathcal{I}_q$ contains one or more figure images from the same paper, provided as complete figures or context-preserving crops, and C_q denotes the accompanying text. Each image may contain multiple panels. The model returns an answer $\hat{a}_q$ and structured supporting records $\hat{E}_q$ specifying chart objects, experimental conditions, and comparisons. These records represent evidence, not internal reasoning.

3.2 SOURCE FIGURES AND BASE CASES

We select experimental figures from adjuvant research papers pre-screened by domain experts, prioritizing those with explicit measurement targets, group comparisons, or cross-metric relationships. Each input image is a complete figure or a context-preserving crop, with axes, legends, statistical annotations, and the original layout retained. We also record the source paper, figure identity, and original location to preserve traceability to the source evidence.

Based on these figures, we construct *base cases*. Each base case contains relevant panels, readable records, and permitted alignment relations, and serves as the evidence unit for question construction. Rather than simply grouping panels, a base case specifies why they can jointly support a question, including shared experimental conditions, differing measurements, and alignable entities. Panels may come from the same multi-panel figure or different figures within the same paper, but all connections must be explicitly supported by the publication.

This design preserves the contextual dependencies of real scientific figures. For example, panels may share a legend while using different y-axis measurements, or the same treatment group may appear at different time points. Base cases record both shared attributes and differences, providing constraints for subsequent annotation and question construction.

3.3 EXPERIMENTAL FACT ANNOTATION AND EXECUTABLE QUESTION CONSTRUCTION

Freely generating questions directly from images makes it difficult to maintain consistent experimental conditions across related questions or to determine which observations support a given conclusion. We therefore first annotate atomic facts and then construct questions and reference answers from these facts.

Each atomic fact corresponds to a verifiable chart record and specifies its source location, measurement, experimental conditions, target group, comparison target, and observed content. The observation may represent a numerical value, direction of change, pairwise statistical information, or measurement coverage. Different fact types retain only the fields required for their interpretation. Statistical comparisons must specify the compared group pair, while measurement-coverage records must identify the relevant group and conditions. Coverage records without numerical values are not assigned artificial values.

We explicitly distinguish different evidence states. A comparison explicitly marked as non-significant, a group pair for which no statistical comparison is reported, the absence of a queried measurement from the available input, and a case that cannot be determined from the public evidence represent different conditions. Models should not infer unreported statistical conclusions solely from bar heights or error bars, nor should missing records be interpreted as negative experimental results. These distinctions make insufficient evidence part of the task itself rather than information completed by annotators.

Given a fact set $\mathcal{F}$, each question is associated with a finite executable program π_q, with the reference output defined as

$$(a_q, E_q) = \mathrm{Exec}(\pi_q, \mathcal{F}), \tag{2}$$

where a_q denotes the reference answer and E_q the supporting records. The program comprises operations including filtering, intersection, conditional joins, rebinding comparison targets, statistical lookup, and counting. Execution produces the answer and supporting evidence, establishing an explicit correspondence among query conditions, reference conclusions, and required facts.

Question construction follows five rule types: changing comparators after filtering, querying another endpoint after filtering, modifying directional or significance conditions, verifying group-pair statistics, and checking measurement coverage under specified conditions. These rules recombine existing facts without changing the figures. They describe how facts are queried, whereas family roles describe relationships among questions.

3.4 SHARED-EVIDENCE QUESTION FAMILIES

We organize related questions into families based on shared source figures and factual annotations. Each family contains local and compositional questions, with condition-variant counterparts added

when supported by the source evidence. Local questions verify basic observations or experimental relationships. Compositional questions require multiple records to be combined to reach a judgment, including joint verification and filter-then-query operations with sequential dependencies. Condition-variant questions modify the comparison target, filtering conditions, or query scope to test whether models adjust their evidence selection and interpretation accordingly.

These roles are defined by how evidence is used, without a predefined difficulty ordering. Local questions serve as the basic reference within a family, although their supporting evidence may span multiple panels. Condition variant questions retain the original figures and facts, changing only the query requirements. Such variants may still require composition and need not change the answer. For illustration, suppose two panels report outcomes X and Y under the same conditions. A local question asks whether G_1 significantly exceeds G_0 on X. A compositional question asks whether this holds for both outcomes. A condition variant changes only the comparator for Y from G_0 to G_2.

Question families support joint evaluation of related questions grounded in the same source evidence, assessing whether models can correctly answer all variants. For families in which local questions cover all necessary premises of the compositional questions, we can further analyze compositional performance when the local answers are correct. This examines whether correct verification of basic evidence is accompanied by successful integration of that evidence into a conclusion.

3.5 Quality Control and Dataset Composition

Quality control covers source verification, fact verification, and question validation. We check figure identity and readability, verify experimental conditions, comparison targets, and statistical interpretation, and assess whether questions have unambiguous answers supported by the specified inputs. Programmatic checks verify references and recompute answers. Visual inspection and manual review address factual and semantic errors. Released inputs contain questions, designated images, and necessary context, but exclude fact tables, programs, and reference answers.

Structured extraction contains approximately 200 composite figures. The question-answering component comprises 380 families and 1,134 questions: 380 local, 381 compositional, and 373 condition-variant questions. Extraction and QA are separate task collections; QA questions are not generated from the extraction subset. Family sizes vary because only evidence-supported condition variants are included, and families may contain additional compositional questions. Questions are derived from shared facts rather than assigned one-to-one to figures. A figure may support multiple families; a family may span figures within one paper.

4 Experiments

This section evaluates whether observations from charts can be transformed into verifiable scientific evidence. We first assess whether models can recover structured records that correctly associate data with their experimental context. We then examine whether judgments about experimental relations are supported by complete evidence. Finally, we compare evidence input settings and analyze question families built on shared evidence. These analyses distinguish the demands of recovering information, assessing its applicability under specified conditions, and integrating evidence across panels.

4.1 Experimental Setup

Tasks and inputs. Structured knowledge extraction follows the chart-to-table task defined in the methodology, requiring models to recover categories, data series, numerical values, and their associations from the original figures. For question answering, models must provide both answers and supporting records under the specified experimental conditions. We use autonomous evidence localization as the primary setting, where target panels are not given and models must identify relevant observations based on metrics, groups, time points, and comparators. We additionally evaluate a panel-guided setting that provides relevant panel cues to isolate the effect of localization support. Both settings require understanding the experimental relations underlying the results.

Table 1: Structured knowledge extraction under autonomous localization and panel-guided settings. All scores are percentages. Bold denotes the best reported score within each model group. Dashes indicate unreported results.

	Autonomous Localization		Panel Guidance	
Model	**RMS-F1**	**RNSS**	**RMS-F1**	**RNSS**
Closed-Source Models				
Gemini-3.7-Flash (Google DeepMind, 2026b)	47.33	**75.7**	**54.92**	**87.95**
Gemini-3.6-Flash (Google DeepMind, 2026a)	**47.76**	74.32	52.86	85.98
GPT-5.5 (OpenAI, 2026)	39.27	66.19	43.22	72.51
GPT-5 (OpenAI, 2025)	20.25	69.05	21.64	74.27
Claude-sonnet-5 (Anthropic, 2026c)	43.88	71.02	53.23	86.31
Claude-sonnet-4.6 (Anthropic, 2026b)	37.22	65.51	40.32	80.13
Open-Source Models				
DeepSeek-V4.1-Flash (DeepSeek-AI et al., 2026)	**39.67**	**74.92**	**42.25**	80.01
Qwen3.5-35B-A3B (Qwen Team, 2026d)	33.26	72.55	35.87	78.25
Qwen3.5-27B (Team, 2026)	27.37	62.89	35.12	**80.83**
Qwen3.6-35B-A3B (Qwen Team, 2026c)	31.11	68.13	34.79	76.26
Qwen3.6-27B (Qwen Team, 2026b)	30.06	64.21	32.31	69.33
InternVL3.5-30B-A3B (Wang et al., 2025)	10.23	25.33	21.52	51.57

Data and models. We evaluate 12 models on 200 samples for structured extraction and 15 models on 1,134 questions for question answering, with panel-guided results reported for the same models. Independent runs under the same model name are treated separately. The two tasks are evaluated independently, and structured extraction outputs are not provided to the question-answering models, allowing us to separately assess result organization and evidence-based answering.

For family-level analysis of condition binding and evidence composition, we evaluate GPT-5.5 (OpenAI, 2026), Gemini-3.7-Flash (Google DeepMind, 2026b), Qwen3.5-35B-A3B (Team, 2026), and Qwen3.6-35B-A3B (Qwen Team, 2026d). Using three fixed seeds, we sample 100 families per run, stratified into 35 statistical comparison, 35 measurement coverage, and 30 filtering/join families. After deduplication, this yields 95 families and 286 questions from 79 papers. All models use the same samples, and related variants are answered independently.

Evaluation Metrics We evaluate structured extraction using RMS-F1, which jointly measures matches in table labels and numerical values while allowing row and column permutations that preserve table semantics (Liu et al., 2023). We additionally report RNSS (Masry et al., 2022) as an auxiliary measure of numerical similarity; unlike RMS-F1, RNSS does not preserve complete label-value associations.

For question answering, we report Answer EM, Evidence F1, and Joint EM. Let A_i indicate a correct answer for question i, and B_i indicate a complete match of the required supporting records under the specified equivalence rules:

$$\text{AnswerEM} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}[A_i], \tag{3}$$

$$\text{JointEM} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}[A_i \wedge B_i]. \tag{4}$$

Evidence F1 is computed over complete supporting records and macro-averaged across questions, capturing partial overlap with the reference evidence. Joint EM requires both a correct answer and a complete evidence match. At the family level, we additionally report the proportion of families for which all valid variants achieve Joint EM.

4.2 Structured Knowledge Extraction: From Reading Values to Reusable Records

Organizing scientific results requires preserving not only numerical values but also their associations with categories and measurement series. Table 1 therefore reports both RNSS and RMS-F1. Gemini-

Table 2: Performance comparison under autonomous localization and panel-guided settings.

	Autonomous Localization			Panel Guidance		
Model	**Ans.**	**Evi.**	**Joint**	**Ans.**	**Evi.**	**Joint**
Closed-Source Models						
GPT-5.5 (OpenAI, 2026)	**81.71**	**90.07**	**81.24**	90.44	95.71	90.16
Gemini-3.7-Flash (Google DeepMind, 2026b)	81.06	88.76	80.97	**94.15**	**97.00**	**93.96**
Claude-Opus-5 (Anthropic, 2026a)	80.23	87.10	79.23	92.57	93.76	92.48
Gemini-3.6-Flash (Google DeepMind, 2026a)	79.67	87.80	79.11	87.00	91.82	86.72
GPT-5 (OpenAI, 2025)	63.60	77.10	63.23	67.97	78.67	67.87
Claude-Sonnet-5 (Anthropic, 2026c)	65.23	68.42	64.79	79.29	70.56	63.79
GPT-4o (OpenAI, 2024)	34.17	30.27	18.20	36.03	29.68	17.46
Open-Source Models						
Qwen3.6-27B (Qwen Team, 2026b)	55.52	54.49	46.70	59.33	57.24	50.25
Qwen3.6-35B-A3B (Qwen Team, 2026c)	45.87	42.71	33.33	51.37	50.39	40.28
DeepSeek-V4.1-Flash (DeepSeek-AI et al., 2026)	45.40	65.59	44.66	48.47	66.42	48.28
Qwen3.5-35B-A3B (Team, 2026)	43.73	50.79	39.18	47.22	55.31	43.23
Qwen3.5-27B (Qwen Team, 2026a)	31.29	21.65	16.25	40.25	31.67	27.36
Qwen3.5-9B	23.40	20.37	11.79	25.33	26.99	25.16
InternVL3.5-14B (Wang et al., 2025)	8.82	4.19	1.39	27.28	25.72	25.63
InternVL3.5-30B-A3B	4.09	13.76	3.81	20.65	23.57	19.32

3.7-Flash (Google DeepMind, 2026b) and Claude-sonnet-5 (Anthropic, 2026c) achieve the highest RMS-F1 scores of 54.92% and 53.23%, respectively. Notably, Gemini-3.7-Flash and Qwen3.5-27B (Qwen Team, 2026a) obtain high RNSS scores of 87.95% and 80.83%, while their RMS-F1 scores are 54.92% and 35.12%. This gap shows that recovering similar numerical values does not necessarily preserve the label-value associations required for structured records.

Structured extraction and question answering also produce different model rankings. Gemini-3.7-Flash performs best on record recovery, whereas GPT-5.5 (OpenAI, 2026) achieves the highest Answer EM under autonomous evidence localization (Table 2). Although the two tasks use different metrics and samples, this contrast suggests that targeted question answering cannot substitute for direct evaluation of complete record recovery: answering a query may require only a subset of relevant information, while structured extraction must preserve a broader set of data associations.

4.3 Experimental Relation Verification: Are Correct Answers Fully Supported?

We evaluate both final answers and whether the supporting records correctly match the required entities, conditions, and comparison relations. Table 2 reports results under autonomous evidence localization and panel-guided settings, separating answer correctness from evidence completeness.

Under autonomous localization, GPT-5.5 (OpenAI, 2026) achieves the highest Answer EM and Joint EM at 81.71% and 81.24%, respectively. However, Qwen3.6-35B-A3B (Qwen Team, 2026c) has a higher Answer EM than Qwen3.5-35B-A3B (45.87% vs. 43.73%) but a lower Joint EM (33.33% vs. 39.18%), showing that answer accuracy alone can mask differences in evidence completeness. Panel guidance improves most models, but does not guarantee complete supporting records: GPT-4o (OpenAI, 2024) reaches 36.03% Answer EM but only 17.46% Joint EM. These results motivate evaluating evidence localization and evidence organization separately, rather than relying on final answers alone.

4.4 Evidence Input Comparison: Contributions of Images and Explicit Facts

To separate the effects of visual information and explicit evidence records, we compare three input settings on the same 95 question families: text only, original figures, and reference facts. The reference-fact setting replaces figures with annotated atomic facts and experimental conditions, rather than model-extracted records, so it isolates evidence use from visual extraction.

As shown in Figure 3(a), the original figures substantially outperform text-only inputs, improving Answer EM by 39.90–83.53 percentage points across the four models. With reference facts, GPT-

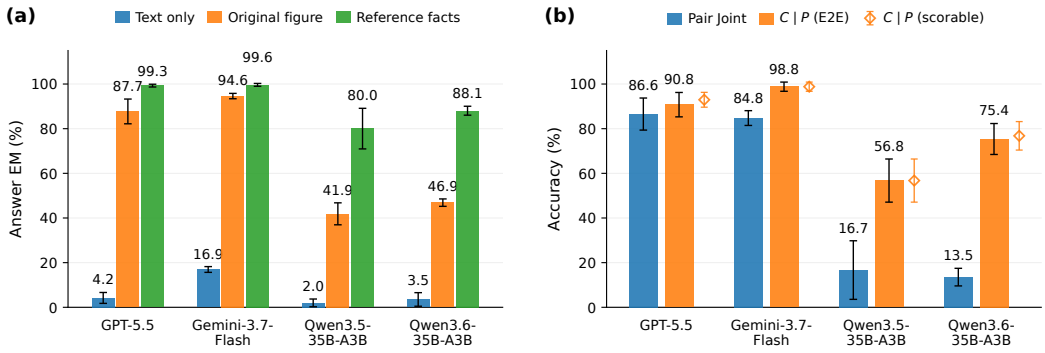

Figure 3: Evidence input comparison and shared-evidence family diagnostics. (a) Answer EM under text-only, original-image, and reference-fact input settings. (b) Paired joint accuracy (Pair Joint) and end-to-end compositional answer accuracy conditional on passing all independent prerequisite tests.

5.5 and Gemini-3.7-Flash reach 99.29% and 99.63% Answer EM, while Qwen3.5-35B-A3B and Qwen3.6-35B-A3B improve by 38.16 and 41.17 points over the original-image setting. Gains also extend to compositional and condition-change questions. These results suggest that visual information recovery and subsequent evidence use are distinct bottlenecks, and that different models are limited by them to different degrees.

4.5 Shared-Evidence Family Analysis: Condition Changes and Evidence Composition

We use shared-evidence question families to assess consistency across related queries. On 58 strictly paired families, we report paired answer accuracy and paired joint accuracy, where both questions must be answered correctly and, for the latter, both supporting records must also be fully matched. On 79 families used for prerequisite diagnosis, we first verify the required basic records and then evaluate the corresponding compositional questions only when all prerequisites are correct; prerequisite answers are not provided to the compositional query.

As shown in Figure 3(b), GPT-5.5 and Gemini-3.7-Flash achieve paired joint accuracies of 86.55% and 84.76%, while Qwen3.5-35B-A3B and Qwen3.6-35B-A3B reach only 16.68% and 13.52%. The Qwen models sometimes answer both related queries correctly without fully matching the supporting records, showing that answer accuracy alone does not capture consistency of evidence use. Even when all prerequisite records are correct, Qwen3.5-35B-A3B and Qwen3.6-35B-A3B achieve only 56.76% and 75.38% accuracy on compositional questions, whereas Gemini-3.7-Flash reaches 98.81%. Thus, correct recovery of basic evidence does not necessarily guarantee successful composition. Because the qualifying family subsets differ across models, these conditional results are used only for within-model diagnosis rather than cross-model ranking.

5 Conclusion

We introduce AdjuvantChart, a real-world benchmark for scientific chart understanding in adjuvant research. It connects structured extraction, experimental relation verification, and cross-panel synthesis to evaluate experimental semantics and cross-panel reasoning. Using shared-fact question families and joint evaluation of answers and supporting evidence, we examine how models use chart evidence under different query conditions. Experiments show that correct answers do not always come with complete supporting records, and that explicit observations with experimental conditions improve performance. However, some models still struggle with compositional questions even when the required observations are correctly recovered. These results motivate evaluating scientific chart understanding beyond final answers, including data associations, condition binding, and evidence organization, and provide a basis for verifiable multimodal chart analysis.

AI USE STATEMENT

We used generative AI tools to assist with the construction and review of the dataset, refinement of the evaluation design, development of supporting code, and language editing of the manuscript. AI tools also assisted with preliminary illustrations of the benchmark motivation and construction workflow. AI-assisted data construction and code development underwent expert review and revision, and the authors reviewed and revised the manuscript. The experimental chart images in AdjuvantChart were extracted or cropped from published papers and were not generated by AI. The authors take responsibility for the final dataset, code, analyses, and manuscript.

REPRODUCIBILITY STATEMENT

Section 3 and Appendices B–D describe the source figures, atomic-fact annotations, executable question construction, and quality-control procedures. Section 4.1 and Appendix E specify the evaluation inputs, response schemas, generation settings, scoring rules, and handling of failed requests. Appendices E–F document the family-level aggregation procedures and report group-wise diagnostic results. The supplementary protocol files provide complete prompts, evaluation manifests, recorded run configurations, and input and scoring version metadata. Source identifiers support annotation traceability, while reference facts and answers are kept separate from standard model inputs. Exact replication of API-generated responses is subject to provider-side model versions and unspecified service defaults.

REFERENCES

Anthropic. Claude Opus 5 System Card, July 2026a. URL https://www.anthropic.com/claude-opus-5-system-card. Accessed: 2026-09-25.

Anthropic. Claude Sonnet 4.6 System Card, February 2026b. URL https://www.anthropic.com/claude-sonnet-4-6-system-card. Accessed: 2026-09-25.

Anthropic. Claude Sonnet 5 System Card, June 2026c. URL https://www.anthropic.com/claude-sonnet-5-system-card. Accessed: 2026-09-25.

Ekaterina Borisova, Nikolas Rauscher, and Georg Rehm. Scivqa 2025: Overview of the first scientific visual question answering shared task. In *Proceedings of the Fifth Workshop on Scholarly Document Processing (SDP 2025)*, pp. 182–210, 2025.

Yi Chen, Yu Zhang, Jian Xu, Hua Yue, Xinming Wang, Zequan Lyu, Xu-yao Zhang, Wei Wei, and Cheng-Lin Liu. An open-ended benchmark and formal framework for adjuvant research with mllm. In *International Conference on Learning Representations*, volume 2026, pp. 118973–119017, 2026.

DeepSeek-AI et al. DeepSeek-V4.1-Flash: Pushing the limits of KV cache compression, September 2026. URL https://arxiv.org/abs/2609.19969.

Azher Ahmed Efat, Seok Hwan Song, and Wallapak Tavanapong. Beyond single plots: A benchmark for question answering on multi-charts. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 35374–35411, 2026.

Google DeepMind. Gemini 3.6 Flash Model Card, July 2026a. URL https://deepmind.google/models/model-cards/gemini-3-6-flash/. Accessed: 2026-09-25.

Google DeepMind. Gemini 3.7 Flash Model Card, August 2026b. URL https://deepmind.google/models/model-cards/gemini-3-7-flash/. Accessed: 2026-09-25.

Kung-Hsiang Huang, Hou Pong Chan, May Fung, Haoyi Qiu, Mingyang Zhou, Shafiq Joty, Shih-Fu Chang, and Heng Ji. From pixels to insights: A survey on automatic chart understanding in the era of large foundation models. *IEEE Transactions on Knowledge and Data Engineering*, 37(5): 2550–2568, 2024.

Tianjing Li, Julian P. T. Higgins, and Jonathan J. Deeks. Chapter 5: Collecting data. In Julian P. T. Higgins, James Thomas, Jacqueline Chandler, Miranda Cumpston, Tianjing Li, Matthew J. Page, and Vivian A. Welch (eds.), *Cochrane Handbook for Systematic Reviews of Interventions*, chapter 5. Cochrane, version 6.5 edition, 2024. URL https://www.cochrane.org/authors/handbooks-and-manuals/handbook/current/chapter-05. [last updated October 2019; accessed 2026-03-24].

Zhen Li, Duan Li, Yukai Guo, Xinyuan Guo, Bowen Li, Lanxi Xiao, Shenyu Qiao, Jiashu Chen, Zijian Wu, Hui Zhang, et al. Chartgalaxy: A dataset for infographic chart understanding and generation. In *International Conference on Learning Representations*, volume 2026, pp. 52484–52546, 2026.

Fangyu Liu, Julian Eisenschlos, Francesco Piccinno, Syrine Krichene, Chenxi Pang, Kenton Lee, Mandar Joshi, Wenhu Chen, Nigel Collier, and Yasemin Altun. Deplot: One-shot visual language reasoning by plot-to-table translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 10381–10399, 2023.

Ahmed Masry, Jia Qing Tan, Shafiq Joty, Enamul Hoque, et al. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the association for computational linguistics: ACL 2022*, pp. 2263–2279, 2022.

Ahmed Masry, Mohammed Saidul Islam, Mahir Ahmed, Aayush Bajaj, Firoz Kabir, Aaryaman Kartha, Md Tahmid Rahman Laskar, Mizanur Rahman, Shadikur Rahman, Mehrad Shahmohammadi, et al. Chartqapro: A more diverse and challenging benchmark for chart question answering. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 19123–19151, 2025.

Nitesh Methani, Pritha Ganguly, Mitesh M Khapra, and Pratyush Kumar. Plotqa: Reasoning over scientific plots. In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1516–1525. IEEE, 2020.

OpenAI. GPT-4o System Card, August 2024. URL https://openai.com/index/gpt-4o-system-card/. Accessed: 2026-09-25.

OpenAI. GPT-5 System Card, August 2025. URL https://openai.com/index/gpt-5-system-card/. Accessed: 2026-09-25.

OpenAI. GPT-5.5 System Card, April 2026. URL https://openai.com/index/gpt-5-5-system-card/. Accessed: 2026-09-25.

Shraman Pramanick, Rama Chellappa, and Subhashini Venugopalan. Spiqa: A dataset for multimodal question answering on scientific papers. *Advances in Neural Information Processing Systems*, 37:118807–118833, 2024.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026a. URL https://qwen.ai/blog?id=qwen3.5. Accessed: 2026-09-25.

Qwen Team. Qwen3.6-27B: Flagship-level coding in a 27B dense model, April 2026b. URL https://qwen.ai/blog?id=qwen3.6-27b. Accessed: 2026-09-25.

Qwen Team. Qwen3.6-35B-A3B: Agentic coding power, now open to all, April 2026c. URL https://qwen.ai/blog?id=qwen3.6-35b-a3b. Accessed: 2026-09-25.

Qwen Team. Qwen3.6-35B-A3B: Agentic coding power, now open to all, April 2026d. URL https://qwen.ai/blog?id=qwen3.6-35b-a3b.

Jonathan Roberts, Kai Han, Neil Houlsby, and Samuel Albanie. Scifibench: Benchmarking large multimodal models for scientific figure interpretation. *Advances in Neural Information Processing Systems*, 37:18695–18728, 2024.

Ruiran Su, Jiasheng Si, Zhijiang Guo, and Janet Pierrehumbert. Climateviz: A benchmark for statistical reasoning and fact verification on scientific charts. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 23447–23469, 2025.

Qwen Team. Qwen3. 5-omni technical report. *arXiv preprint arXiv:2604.15804*, 2026.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, et al. Charxiv: Charting gaps in realistic chart understanding in multimodal llms. *Advances in Neural Information Processing Systems*, 37:113569–113697, 2024.

Jingxuan Wei, Nan Xu, Junnan Zhu, Gaowei Wu, Qi Chen, Bihui Yu, Lei Wang, et al. Chartmind: A comprehensive benchmark for complex real-world multimodal chart question answering. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 4555–4569, 2025.

Tianchi Xie, Minzhi Lin, Mengchen Liu, Yilin Ye, Changjian Chen, and Shixia Liu. Infochartqa: A benchmark for multimodal question answering on infographic charts. *Advances in Neural Information Processing Systems*, 38, 2026.

Cheng Yang, Chufan Shi, Yaxin Liu, Bo Shui, Junjie Wang, Mohan Jing, Linran Xu, Xinyu Zhu, Siheng Li, Yuxiang Zhang, et al. Chartmimic: Evaluating lmm's cross-modal reasoning capability via chart-to-code generation. In *International Conference on Learning Representations*, volume 2025, pp. 26590–26646, 2025.

Zifeng Zhu, Mengzhao Jia, Zhihan Zhang, Lang Li, and Meng Jiang. Multichartqa: Benchmarking vision-language models on multi-chart problems. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 11341–11359, 2025.

A Dataset Organization and Provenance

A.1 Units of annotation and evaluation

AdjuvantChart uses figures from published adjuvant research to evaluate how models recover and use scientific chart evidence. Images retain the original visual content, including axes, legends, statistical annotations, and panel layouts. They are extracted from the publications or cropped to a specified region; no experimental chart is redrawn or generated. The source paper, original figure identifier, and image location are retained for tracing annotations back to the publication.

We distinguish six units. A *source figure* is a numbered figure in a paper and may contain several panels. A *panel* is a chart region identified within that figure. An *input image* is a file actually supplied to a model, which may represent a complete figure, a context-preserving crop, a figure part, or a supplementary legend. A *base case* specifies the observations and experimental conditions that can be aligned for question construction. A *question family* groups related queries over shared source evidence, and a *question* is one independently answered evaluation item. These units are not interchangeable: one figure can support several questions, and one question can require several panels or images.

The benchmark contains 600 composite figures overall. The structured extraction component contains approximately 200 composite figures. The question-answering component comprises 380 families and 1,134 questions. The two components are constructed as separate task collections; the extraction figures are not the source pool from which the question-answering component is derived. Structured extraction predictions are not passed to the question-answering models. Consequently, comparisons between the two tasks concern their respective performance profiles, rather than error propagation in a shared extraction–reasoning pipeline.

A.2 From source evidence to question families

A base case records the selected chart regions, entities, measurements, experimental conditions, and permitted alignments. Atomic facts are then annotated within this scope. Each question specifies a query over these facts, and an executable program produces the reference answer and required

Table 3: Question-family composition. Family roles describe relationships among questions and do not define an ordering of difficulty.

Family structure	Families	Local	Compositional	Condition-variant
One question in each role	300	300	300	300
One additional compositional question	37	37	38	37
No valid condition variant	43	43	43	36
Total	380	380	381	373

supporting records. Families preserve the evidence while varying how it is queried. The resulting number of questions depends on the supported queries rather than a fixed number of questions per figure.

As shown in Table 3, 372 families contain one question in each role, one contains an additional compositional question, and seven contain no condition variant. Thus,

$$300 \times 3 + 37 \times 4 + 43 \times 2 = 1{,}134.$$

The corresponding role totals are 380 local, 381 compositional, and 300 condition-variant questions. A local question directly checks a basic observation or relation used as a family reference. A compositional question requests joint checks or dependent operations, such as filtering conditions before querying another endpoint. A condition variant changes a comparator, selection requirement, or query scope while keeping the underlying evidence fixed. A variant need not reverse the answer. These roles also do not determine the number of panels required: a local check can span panels, and a condition variant can itself require composition.

A.3 Input scope and provenance

For a query q, let $\mathcal{I}_q = (I_{q,1}, \ldots, I_{q,m_q})$ denote its designated input images, where $m_q \geq 1$. An image may contain multiple panels, and multiple image files need not represent distinct source figures. In particular, a legend image or another part of the same figure provides context without creating an additional experimental source. When a query uses evidence from different figures in the same paper, the designated source images are supplied together.

The public question record specifies the English query, image list, necessary context, answer schema, and supporting-record requirements. Reference annotations, stored separately from model inputs, retain fact identifiers, source locations, executable programs, and reference outputs. Chinese translations are auxiliary authoring materials and are not additional evaluation items or model inputs. The source and question identifiers support provenance and version tracking; they do not establish that questions sharing a figure or paper are statistically independent.

B Annotation Guidelines and Quality Control

B.1 Atomic facts and condition binding

An atomic fact represents an observation that can be checked against a specified chart region. Its identity depends on the source location, experimental conditions, measurement, and relevant group or group pair. Identical numerical values at different locations are therefore not interchangeable. Table 4 summarizes the main fields. Optional metadata further disambiguate time, dose, units, cell populations, axes, or statistical summaries when required by the source.

For statistical comparisons, annotators bind the comparison bracket or line to its endpoints before reading its label. Direction is recorded from the displayed summary independently of significance: neither a significant p-value nor the vertical order of nearby statistical labels identifies the higher group. Similarly, an annotation for a neighboring group pair, another time point, or an aggregate endpoint cannot substitute for the requested comparison.

Coverage records specify whether the requested measurement is displayed for a particular group under the designated conditions. A group name in a legend is insufficient evidence that its measurement appears on every axis. A false coverage value represents absence of the requested display, not

Table 4: Semantic fields of atomic facts. Comparison and coverage records retain different fields according to the observation being represented.

Field	Interpretation
Source identity and location	Paper, figure or panel, and source-region coordinates for traceability.
Condition	The relevant site, time, population, stimulation, or other experimental condition. The implementation's `site` field is a generic condition key.
Measurement	The requested endpoint and any units needed to distinguish it from other measurements.
Group or group pair	The observed group for coverage, or the two compared groups for statistical records.
Observation	A numerical value, comparison status, direction, or measurement-coverage Boolean, according to the fact type.
Statistical annotation	The printed relation and bound for p, and the literal display marker when requested.
Additional binding information	Axis identifiers, bracket endpoints, and concise source descriptions when available.

a zero measurement, a non-significant effect, or absence of the underlying biological phenomenon. No artificial numerical value is assigned to a coverage-only record.

B.2 Evidence states and statistical conventions

Table 5: Distinct evidence states for comparison records. Missing or insufficient evidence is not converted into a negative experimental conclusion.

State	Definition
`SIG`	The reported value or explicitly defined bound establishes $p < 0.05$.
`NS`	An explicit non-significant annotation or reported value establishes non-significance under the specified criterion.
`NOT_REPORTED`	The relevant measurement is displayed, but the requested comparison is not reported.
`MISSING`	The requested measurement or group–condition combination is absent from the designated scope.
`UNREADABLE`	A required annotation is present but cannot be read reliably.
`UNDETERMINED`	The visible information does not determine the requested status, for example because a symbol is undefined or a bound is inconclusive.

The protocol uses the strict threshold $p < 0.05$. Accordingly, $p = 0.05$ is non-significant, whereas a reported bound $p \leq 0.05$ is insufficient to determine whether the strict criterion holds. A bound $p < 0.1$ is also inconclusive. Star annotations are converted to numerical bounds only when their meanings are supplied in the permitted context or legend. An explicit `ns` supports `NS`, but does not authorize inventing a numerical bound. If neither the figure nor its permitted context supplies that bound, the numerical p field remains `null`.

This separation preserves what the figure establishes. For example, an existing trajectory without a group-pair comparison is `NOT_REPORTED`; a missing requested axis is `MISSING`. Neither can be inferred to be `NS` from nearby error bars or from comparisons reported elsewhere.

B.3 Verification procedures and validation limits

Quality control addresses three aspects: source verification, fact verification, and question validation. Source verification checks figure identity and image readability. Fact verification checks measurement attribution, experimental conditions, comparison endpoints, and statistical meaning. Question validation assesses answerability, determinacy, and evidence sufficiency under the designated English input. Structural checks verify links among images, facts, programs, and questions, and replay the programs to recompute reference outputs.

For structured extraction, returned manual-review decisions distinguish accepted, corrected, deferred, and excluded samples. Recorded corrections are applied before reference targets are restricted to the requested visual scope. Fields outside that scope are not imputed. These review records do not identify reviewers or document independent duplicate annotation, and therefore do not support an inter-annotator agreement estimate.

For question answering, program replay establishes consistency between the annotations and executable rules; it does not establish that the annotations correctly interpret the source figures. Source cross-checking and model-assisted review complement these structural checks. Independent expert adjudication remains incomplete in the archived version, including unresolved source-interpretation cases. The references therefore have not been established as fully expert-validated annotations. Validation results apply to the version that was checked and must be reconsidered when questions or references change.

C Executable Rules and Question Families

C.1 Rule templates and program execution

Rule patterns specify how facts are queried; family roles describe relationships among questions. A family may instantiate several rules, but individual questions execute only their own operations. Composition includes both parallel verification and dependent queries.

Table 6: Rule patterns used to construct questions and reference outputs. Target records are identified in the annotated fact set before execution.

Rule	Required records	Query and output
R1	Source comparisons and an alternative group pair under aligned conditions	Filter conditions by direction and significance, then retrieve the alternative comparison for the selected conditions.
R2	Source measurements and another endpoint with shared condition keys	Select conditions using the source measurements, then retrieve the corresponding target endpoint and, when requested, assess whether all target comparisons are significant.
R3	Directions and statistical states of source comparisons	Apply direction and significance constraints separately; identify qualifying conditions and explicit nonsignificant exceptions.
R4	Comparisons indexed by panel, condition, measurement, and group pair	Read the reported statistical annotation for each requested comparison while preserving its experimental attribution.
R5	Coverage records indexed by panel, condition, measurement, and group	Determine whether each requested measurement is displayed and return the subset with no displayed measurement.

Programs use eight operations: `FILTER_DIRECTION`, `FILTER_SIG`, `INTERSECT`, `JOIN_CONDITION`, `REBIND_PAIR`, `READ_P`, `COUNT`, and `RETURN`. Joining conditions and changing pairs retrieve annotated candidate records using explicit condition keys. The interpreter does not discover visual correspondence.

For selection questions, required evidence contains the requested source records, including rejected candidates, and the target records reached after selection. Unused candidate targets are excluded. The required set is defined by the response protocol and need not be a uniquely minimal proof of the answer. Direct comparison and coverage questions return the records explicitly requested in the query.

For the archived panel-guided construction snapshot, recomputation matches the exported answers, evidence, and required/executed fact identifiers for all 1,134 questions. This verifies agreement between the programs and stored annotations, without establishing source correctness or completion of human review.

C.2 STATISTICAL AND LOGICAL SEMANTICS

Significance uses strict $p < 0.05$, preserving reported inequalities. Thus $p < 0.05$ is `SIG`, while $p \leq 0.05$ alone is `UNDETERMINED`. Exact values at or above 0.05 are `NS` under this benchmark convention. Direction is annotated separately.

A *near miss* satisfies the requested directions but has an explicitly nonsignificant source comparison; unknown status does not suffice. The predicate that all targets are significant is true for a nonempty target set containing only `SIG` records, false when at least one reached target is `NS`, and null otherwise, including for an empty set. This convention avoids treating an empty selection as evidence for a positive conclusion.

C.3 A COMPLETE FAMILY WITH A CHANGED COMPARISON PAIR

Family `P05_D01` uses source Figure 1 (DOI: `10.1038/s41467-023-42923-x`). The source labels G1, G2, and G3 denote alum, pSer:alum, and pSer:alum+SMNP, respectively. BGC denotes germinal-center B cells and AUC denotes area under the curve. Table 7 presents the comparison fields of all four facts used by this family under the post-prime condition. Questions below are condensed from the original panel-guided English versions.

Table 7: All comparison records used by family `P05_D01`. The short identifiers are used only to make the example compact.

ID	Source panel	Measurement	Group pair	Annotation
f_1	1d	BGC AUC	G1 vs G3	$p = 0.0009$, SIG
f_2	1f	Env+ BGC AUC	G1 vs G3	$p = 0.0005$, SIG
f_3	1d	BGC AUC	G2 vs G3	$p = 0.0014$, SIG
f_4	1f	Env+ BGC AUC	G2 vs G3	$p = 0.0007$, SIG

Local (ACQ033). Read and classify the G1–G3 p annotations for BGC AUC in panel d and Env+ BGC AUC in panel f. Answer and evidence: f_1, f_2.

Compositional (ACQ034). Read both G1–G3 and G2–G3 comparisons on these axes. Answer and evidence: f_1, f_2, f_3, f_4.

Condition-variant (ACQ035). Change the requested pair to G2–G3. Answer and evidence: f_3, f_4.

All three programs read and return designated records. Composition here is parallel verification. Changing the pair changes the records and p values even though every classification remains `SIG`; meaningful variants need not reverse the answer status.

C.4 A COMPLETE FAMILY WITH SELECTION AND CROSS-PANEL RETRIEVAL

Family `P01_D03` uses source Figure 3 (DOI: `10.1038/s41467-025-59353-6`). Let A, B, and C denote RBD-MSA, BnT-Pfizer, and RBD, respectively. Tables 8–9 present the comparison fields of all twelve facts used by the family. BALF denotes bronchoalveolar lavage fluid, represented by the key `balf`. The supplied context maps the blood-drop and lung icons to serum and BALF, respectively. Binding denotes the plotted endpoint, not its complement, $100 - \mathrm{Binding}$.

Table 8: Source records for family `P01_D03`. All comparisons are A versus B in source panel 3b. Each measurement entry is a separate fact.

Site	IgG annotation	Higher mean	IgA annotation	Higher mean
serum	$p = 0.0010$, SIG	A	$p = 0.0467$, SIG	A
balf	$p = 0.0162$, SIG	A	$p = 0.0027$, SIG	A
nose	$p = 0.0574$, NS	B	$p < 0.0001$, SIG	A
saliva	$p = 0.0104$, SIG	B	$p < 0.0001$, SIG	A

Local (ACQ007). Read the direction, p annotation, and status of A–B comparisons for IgG and IgA at all four sites. Answer and evidence contain the eight source records.

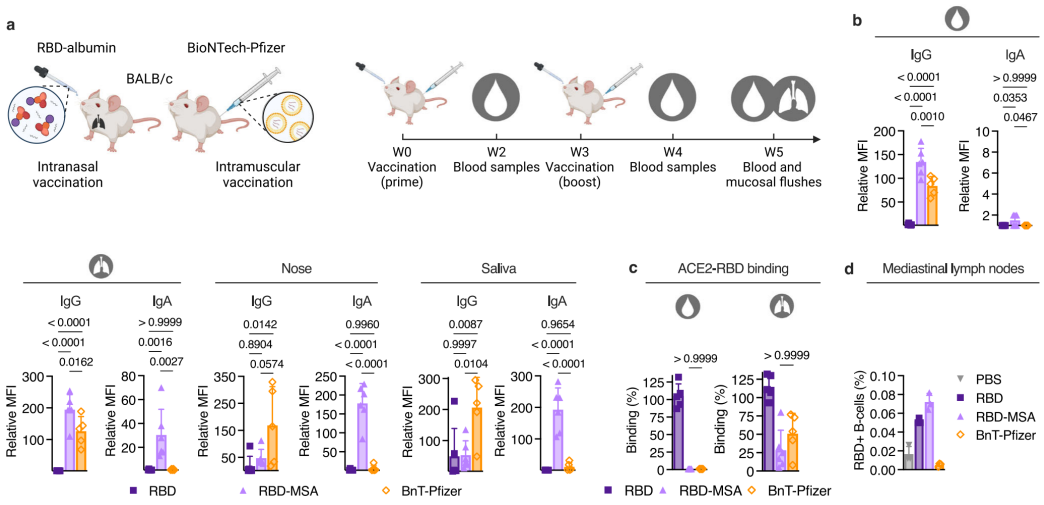

Figure 4: Original input figure for family `P01_D03`, reproduced from source Figure 3 of *An intranasal subunit vaccine induces protective systemic and mucosal antibody immunity against respiratory viruses in mouse models* (DOI: `10.1038/s41467-025-59353-6`). Panel b supplies the selection records; panel c supplies the Binding targets.

Table 9: Candidate target records for family `P01_D03`. A query returns only the targets reached by its selection.

Source panel	Site	Measurement	Group pair	Annotation
3b	nose	IgG	C vs A	$p = 0.8904$, NS
3b	saliva	IgG	C vs A	$p = 0.9997$, NS
3c	serum	Binding	A vs B	$p > 0.9999$, NS
3c	balf	Binding	A vs B	$p > 0.9999$, NS

Compositional (ACQ008). Select sites where B has higher IgG and A has higher IgA, both significant; read their C–A IgG comparisons and report near misses. Only saliva qualifies, with target $p = 0.9997$, `NS`. Nose is a near miss because its IgG source comparison is `NS`. Evidence comprises the eight source records and the saliva target.

Condition-variant (ACQ009). Remove the source-significance requirement. Nose and saliva qualify; target values are $p = 0.8904$ and $p = 0.9997$, both `NS`. Evidence comprises the eight source records and both targets.

Additional composition (ACQ010). Select sites where A has higher IgG and IgA, both significant, then read their A–B Binding comparisons in panel 3c. Serum and balf qualify. Both targets have $p > 0.9999$, `NS`; `all_target_significant` is false. Evidence comprises the eight source records and two Binding targets.

The compositional programs filter and intersect condition sets before changing pairs or joining endpoints. Two distinct compositional queries produce a four-question family. The local question excludes target facts, so its correctness alone cannot isolate a subsequent error as purely compositional.

C.5 Coverage and families without a condition variant

For R5, family `P01_D02` checks group coverage on two axes of the same figure: balf Binding excludes PBS, whereas mediastinal-lymph-node RBD+ B cells include PBS. Absence refers only to the specified displayed measurement, not a zero value or an unperformed experiment.

Family `P04_D01` retains two questions: all five specified variants pass the Day-35 filters and have significant Day-260 targets. No condition variant was retained in this family. Because every candidate passes, this example provides limited evidence about a model's ability to reject candidates. The absence of a retained variant does not imply that every alternative question would be uninformative.

D Evaluation Protocols and Implementation Details

D.1 Inputs and Response Protocol

The supplementary protocol files contain the complete system instructions, question records, representative request payloads, and ordered image identifiers. Run configurations and the archived evaluation manifests are provided separately. The metadata retain the input and scoring versions for each documented collection.

Each question is evaluated in a fresh request containing its English question, answer schema, evidence-record instructions, neutral context, and designated source images. Images are supplied in their listed order, with a source identifier preceding each image. Models receive neither other questions' responses nor external retrieval tools. Reference facts and reference answers are excluded from the standard image-based evaluation.

The response is a JSON object with four fields:

`question_id`, `answer`, `checks`, and `explanation`.

The structure of `answer` is specified separately for each question. The `checks` field contains all requested supporting records, including records needed to reject candidates; reporting only the finally selected entries is insufficient. Comparison records identify the panel, condition, metric, group pair, reported p annotation, and evidence status. Direction and literal significance markers are included only when explicitly requested. Coverage records instead identify the panel, condition, metric, group, and a Boolean `measurement_displayed`. A false coverage value denotes the absence of the requested measurement on the specified axis, rather than a zero value or a non-significant comparison.

The historical QA protocol additionally requires `explanation` to be a JSON-encoded string containing an answerability assessment and brief visual evidence descriptions. These fields are not answer targets, but their placement can affect strict parsing. The diagnostic requests instead use a

common evidence-input supplement without this additional requirement. Thus, the diagnostic image condition and the earlier QA runs differ in their prompting protocol as well as their evaluation scope. The supplementary files include the complete system instructions and representative request payloads.

For structured extraction, the user prompt specifies the target scopes, subplot indices, requested data fields, and statistic types. The target crop is followed by the complete figure, which supplies legends and panel context. Other panels remain outside the requested scope. The system instruction is:

> Extract the requested chart data from the supplied images. Follow the supplied scope and output schema. Return one JSON object without surrounding prose.

The output contains a list of subplots and their data records. Original subplot indices are retained; these indices identify annotation groups and need not correspond one-to-one to physical axes. Instructions prohibit inferring individual identities, raw replicates, sample sizes, or spreadsheet cell references that are not visible in the supplied evidence. The parser accepts a JSON object, including one enclosed by a complete JSON code fence; free-form prose, CSV, and Markdown tables are not accepted as substitutes for the requested structure.

D.2 Evidence Localization and Image Processing

Panel guidance provides explicit panel identifiers in the question, schema, or record descriptions. Autonomous locating retains the source-image identity and semantic conditions, but requires the model to determine the relevant panel from the images. For example, a guided question may distinguish antibody measurements in `F01_c` from a binding measurement in `F01_d`; the corresponding autonomous question identifies `F01` while retaining the measurement names, conditions, and group pairs. Original panel letters remain visible in the image. Neutral context may explain icons or measurement terminology in both settings.

For the localization comparison reported in the main text, the autonomous and panel-guided settings use the same question instances, source images, reference annotations, answer targets, evidence requirements, and scoring procedure. The experimental difference is whether explicit target-panel cues are supplied. Apart from these cues, the semantic query conditions and required output structure remain unchanged. Model and generation settings are held fixed within each model's comparison.

For scoring autonomous responses, public record identities are mapped to annotated coordinates only when the correspondence is unique. This mapping uses the supplied identifiers and experimental conditions, without selecting matches using reference numerical values or significance labels. Responses from both settings are evaluated against the same canonical reference records.

The evaluation clients transmit the stored image bytes without resizing, cropping, or re-encoding them. Extraction crops are prepared during dataset construction. API requests specify high image detail. This parameter does not establish identical visual tokenization or internal image resolution across model providers.

D.3 Generation Settings

The API request configurations specify maximum completion budgets of 8,192 tokens for QA and 16,384 tokens for extraction, using a Chat Completions-compatible interface. The corresponding API clients do not send a temperature, top-p, stop sequence, or reasoning-effort setting. These runs therefore use unspecified service defaults rather than a uniform greedy-decoding configuration. The recorded local Qwen3.5-9B and InternVL deployments use temperature 0 and seed 0. Their deployment configuration specifies bfloat16 computation, tensor parallelism of one, and a context limit of 32,768 tokens; thinking is explicitly disabled for the local Qwen3.5-9B run.

The diagnostic configuration permits two infrastructure retries with a 120-second timeout; the earlier API QA configuration uses no client retries. Answers are not retried because they are incorrect. Extraction timeouts are 120 seconds for API runs and 1,200 seconds for local InternVL runs. Run-level configurations and provider-returned model identifiers are included in the supplementary metadata. The returned identifiers do not independently verify an immutable model checkpoint, and omitted parameters remain unspecified.

D.4 Question Answering Metrics and Normalization

Let G_q be the reference evidence-record set and P_q the normalized predicted set for question q, including unmatched placeholders for invalid records. A true positive requires a complete normalized record match. A correct number attached to the wrong condition, metric, panel, or comparison pair is not a match. Invalid predicted records contribute false positives rather than being silently discarded. With $\mathrm{TP}_q = |P_q \cap G_q|$, $\mathrm{FP}_q = |P_q \setminus G_q|$, and $\mathrm{FN}_q = |G_q \setminus P_q|$, evidence F1 is

$$F_{1,q} = \frac{2\mathrm{TP}_q}{2\mathrm{TP}_q + \mathrm{FP}_q + \mathrm{FN}_q}. \tag{5}$$

Every question has at least one reference record. Evidence F1 is averaged over questions, rather than computed by pooling all record counts. Answer EM compares the complete normalized answer. Joint EM requires both an exact answer and exact evidence-set agreement:

$$J_q = \mathbf{1}[\hat{a}_q = a_q]\,\mathbf{1}[P_q = G_q]. \tag{6}$$

Consequently, J_q cannot exceed the answer-correctness indicator, and aggregate Joint EM cannot exceed Answer EM when the samples and weights are identical.

Answer lists and evidence lists use order-insensitive set comparison; the order of the two groups in a comparison pair is also ignored. Text normalization includes Unicode normalization, whitespace trimming, selected dash variants, and a fixed set of identifier aliases. It does not perform unrestricted semantic paraphrase matching or unit conversion. Missing fields, explicit null values, and additional optional fields are not automatically equivalent.

Numerical p annotations use exact decimal normalization: 0.050 and 0.05 are equivalent, but no numerical tolerance is applied and comparison operators are preserved. Statistical states follow the conventions in Appendix B.

When a literal significance marker is requested, its star count and comparison operator must be preserved. Complete numerical markers allow an optional p prefix, surrounding whitespace, and exactly equivalent decimal or scientific notation. This normalization does not convert stars into numbers or reconstruct an omitted marker. The scorer's structural-validity flag is distinct from correctness and does not validate every constraint in a full JSON Schema. Joint EM measures answer correctness together with complete submission of the requested evidence. It does not measure internal evidence use directly, and an answer–joint gap may include failures to follow the evidence-output protocol.

D.5 Structured Extraction Metrics

The RMS implementation is adapted to the benchmark's structured JSON representation. Subplots are aligned by their supplied indices. Within each subplot, a cell key combines its row labels and column name, and one-to-one matching is obtained by the Hungarian algorithm. Values cannot be moved across subplots to improve the score.

A cell-key similarity is $1 - d_{\mathrm{lev}}$ when normalized Levenshtein distance $d_{\mathrm{lev}} < 0.5$, and zero otherwise. For a nonzero reference value v, numerical similarity is $1 - |\hat{v} - v|/|v|$ when the relative error is strictly below 0.1, and zero otherwise. A zero reference value matches only zero. The contribution of a matched cell is the product of its key and numerical similarities. If S is the total matched contribution and n_p, n_g are the prediction and reference denominators, respectively, then, for positive denominators,

$$\text{RMS-P} = S/n_p, \qquad \text{RMS-R} = S/n_g, \qquad \text{RMS-F1} = \frac{2S}{n_p + n_g}. \tag{7}$$

Each denominator sums $\max(1, \text{number of cells})$ over subplots present on that side. An explicitly present but empty subplot therefore retains a denominator contribution, while an absent subplot yields no matched contribution. A completely missing prediction receives zero. Metrics are computed per sample and then macro-averaged; the reported F1 is not reconstructed from averaged precision and recall. Missing values are not replaced by zero, and repeated rows retain their multiplicity.

RNSS compares numerical multisets within corresponding subplots, without preserving the label–value association. Missing or additional whole subplots receive zero, and subplot scores are aggregated using the maximum of one, the reference numerical count, and the predicted numerical count

as weights. The implemented numerical distance assigns zero to two zeros and one to a zero reference paired with a nonzero prediction. Within an otherwise nonempty matched subplot, unmatched values do not incur an additional penalty in this RNSS implementation. For example, reference values $[1, 2]$ and prediction $[1]$ can obtain a perfect numerical similarity. RNSS consequently supplements RMS-F1 and cannot establish extraction completeness on its own. Sample identifiers, chart-type labels, titles, and axis-configuration metadata are not independently scored as data cells.

D.6 Response Processing and Failed Requests

For completed evaluation collections, denominators are fixed by the declared sample manifests. API failures, refusals, truncated outputs that cannot be parsed, and missing predictions within the evaluation scope receive zero; the denominator is not reduced to successfully parsed responses. Samples outside the predefined scope are separate from model failures.

Three response-processing variants evaluate the same generated responses. They distinguish strict parsing (V0), extraction of a unique complete JSON object from the final-answer channel (V1), and conservative separation of evaluable answer/evidence fields from explanatory material (V2). None may fill missing fields from the reference answer, repair significance labels, move nested evidence into the required structure, or use hidden reasoning content. Comparisons across V0–V2 therefore measure sensitivity to response processing, not a change in model generation. The diagnostic results below use V2; this designation does not imply that every main-table run was scored with V2.

D.7 Input Controls and Shared-Family Diagnostics

The input comparison keeps the question, context, answer schema, and evidence requirements fixed across Text-only, Original Figure, and Reference Facts. Text-only omits the figures. Original Figure supplies the designated images. Reference Facts replaces the images with case-level records and their condition attribution; unannotated information remains unknown. Reference Facts removes both visual extraction and part of the burden of organizing evidence, so its gain over Original Figure cannot be interpreted as a pure visual-error rate.

Strict pairs are selected by static criteria: the questions share an evidence scope and operation/output structure while changing a specified condition. Pair Answer requires both answers to be correct, and Pair Joint requires both answers and both evidence sets to be exact. When a family has multiple eligible pairs, scores are averaged within the family before aggregation.

The prerequisite diagnostic independently queries all atomic observations required by the associated compositional program. It covers the candidate domain and necessary branches, rather than only the observations used along the reference answer's selected path. Existing local questions are not assumed to cover this entire set. The prerequisite request asks for observations without requesting the final selected set or joint conclusion. Its response is never supplied to the compositional request.

Each eligible family contributes one designated compositional question. Let P_f indicate that every prerequisite record for that question is correct, and C_f that its independently generated compositional answer is correct. For the eligible families $\mathcal{F}_g$ in group g, the conditional score is

$$\mathrm{C} \mid \mathrm{P}_g = \frac{\sum_{f \in \mathcal{F}_g} P_f C_f}{\sum_{f \in \mathcal{F}_g} P_f}. \tag{8}$$

Composition failures remain in the denominator after prerequisites pass. An empty conditioning set is reported as undefined. This diagnostic tests whether successful independent observation checks coincide with successful composition; it does not establish that the same observations were correctly recovered during the compositional response. Because the conditioning set differs across models, the result supports within-model diagnosis rather than direct model ranking.

For the archived input-condition analysis, question scores are averaged within families and then within strata. Statistical comparison, measurement coverage, and filtering/joining receive fixed weights of 0.4, 0.4, and 0.2, respectively, retained after eligibility filtering. This aggregation differs from an unweighted average over questions.

The archived family-level diagnostics use three groups sampled with fixed seeds and prespecified strata. Overlapping groups reuse the corresponding saved responses. For group estimates x_1, x_2, x_3,

we report $\bar{x} = \frac{1}{3}\sum_{g=1}^{3} x_g$ and the sample standard deviation $s = \sqrt{\frac{1}{2}\sum_{g=1}^{3}(x_g - \bar{x})^2}$. This deviation describes differences among overlapping subsets; it is neither a standard error nor variability across independent model generations. The estimates apply to the corresponding eligible subsets.

E Additional Family-Level Diagnostic Results

We report archived family-level results for the 58 eligible families in the strict-pair analysis and the 79 eligible families in the prerequisite analysis. Answers and supporting records are evaluated under the format-normalized procedure (V2). Prerequisite scoring removes a complete enclosing JSON code fence when present and checks the submitted records without adding reference content. Thus, failure of the collection-time strict parser does not necessarily imply failure of the prerequisite-content check. Subset membership is specified in the archived diagnostic manifests. The tables describe these eligible subsets and are not extrapolated to the full question collection.

E.1 Paired Questions under Changed Conditions

A strict pair contains a base question and a variant that changes one specified condition while retaining the underlying figure evidence. Pair Answer requires both answers to be correct. Pair Joint additionally requires complete evidence matches for both questions. Each eligible family contributes one pair, and the three groups cover 58 unique paired families. The pair is evaluated as one unit: averaging the two individual question scores would not measure success on both variants.

Table 10: Strict-pair results across the three diagnostic groups. Group entries are successful pairs / evaluated pairs. The last column is the mean and sample standard deviation of the three group percentages.

Model	Metric	Group 1	Group 2	Group 3	Mean ± SD (%)
GPT-5.5	Pair Answer	18/23	19/21	20/22	86.55 ± 7.18
	Pair Joint	18/23	19/21	20/22	86.55 ± 7.18
Gemini-3.7-Flash	Pair Answer	20/23	17/21	19/22	84.76 ± 3.31
	Pair Joint	20/23	17/21	19/22	84.76 ± 3.31
Qwen3.5-35B-A3B	Pair Answer	4/23	3/21	9/22	24.20 ± 14.56
	Pair Joint	2/23	2/21	7/22	16.68 ± 13.12
Qwen3.6-35B-A3B	Pair Answer	7/23	5/21	4/22	24.14 ± 6.13
	Pair Joint	4/23	2/21	3/22	13.52 ± 3.94

For GPT-5.5 and Gemini-3.7-Flash, Pair Answer and Pair Joint coincide in these groups. For both Qwen models, requiring complete support reduces paired performance. The answer–joint gap is the proportion of evaluated pairs with both answers correct but at least one evidence set incomplete or incorrect.

Because group membership overlaps, adding the three denominators would overcount unique evaluated pairs.

E.2 Prerequisite Checks and Conditional Composition

For each eligible family, a separate prerequisite request tests the atomic records required by a designated compositional question. A family passes only when all required records match. These prerequisite answers are not supplied to the compositional request. Table 11 reports the number of passing families in each group, and Table 12 reports the corresponding outcomes after deduplicating families across groups.

Conditional composition accuracy is computed within each group by dividing the number of correct compositional answers after passing prerequisites by the number of passing families. The reported mean averages the three group percentages. It need not equal the ratio over unique families, because membership overlaps and the numbers of passing families differ.

Table 11: Families passing all prerequisite checks in each diagnostic group. Entries are passing families / eligible families. Group membership can overlap.

Model	Group 1	Group 2	Group 3
GPT-5.5	28/32	30/34	29/33
Gemini-3.7-Flash	28/32	26/34	27/33
Qwen3.5-35B-A3B	16/32	12/34	19/33
Qwen3.6-35B-A3B	18/32	17/34	18/33

Table 12: Prerequisite outcomes over unique families. "Correct" and "wrong" refer to the independently generated compositional answer after a successful prerequisite check. "Failed/invalid" denotes a transport failure or an unscorable answer. Failed or unscorable answers are kept distinct from content-level errors.

Model	Prerequisites passed	Correct	Wrong	Failed/invalid
GPT-5.5	71/79	65	4	2
Gemini-3.7-Flash	69/79	68	1	0
Qwen3.5-35B-A3B	40/79	21	19	0
Qwen3.6-35B-A3B	46/79	34	11	1

Table 13: Compositional answer accuracy conditional on passing all prerequisite checks. Group entries show correct compositional answers / passing families. Failed or unscorable compositional answers remain in the denominator. The last column reports the mean and sample standard deviation of the three group percentages.

Model	Group 1	Group 2	Group 3	Mean $\pm$ SD (%)
GPT-5.5	24/28	27/30	28/29	90.76 ± 5.46
Gemini-3.7-Flash	27/28	26/26	27/27	98.81 ± 2.06
Qwen3.5-35B-A3B	9/16	8/12	9/19	56.76 ± 9.66
Qwen3.6-35B-A3B	15/18	12/17	13/18	75.38 ± 6.93

GPT-5.5 has one transport failure and one unscorable compositional answer after passing prerequisites; Qwen3.6-35B-A3B has one unscorable answer. These cases count as unsuccessful outcomes but are not classified as content-level composition errors.

Passing the prerequisite check does not ensure success on the independently answered compositional question. This result identifies a discrepancy between targeted record recovery and the broader question, but it does not establish that the required records were correctly recovered within the compositional attempt itself. Because the set of passing families differs by model, conditional scores support within-model diagnosis rather than a ranking on a common subset.

F Qualitative Analysis

We examine four responses from the original-figure diagnostic condition without panel guidance. The supplementary case records link each response to its image, request, and corresponding reference version. These purposively selected examples illustrate observable output errors, not their frequency across the benchmark. Source checks were AI-assisted and conducted after examining the references, rather than through blinded expert adjudication. The GPT-5.5 responses carry the provider-returned identifier `gpt-5.5-2026-04-24`; this identifies the logged service response, not an independently verified model checkpoint.

A readable value assigned to the wrong comparison. In ACQ534, GPT-5.5 reports $p = 0.0009$ for G/Cho-LNPs versus G/RM on the tumour $CD3^+CD4^+$ T-cell axis. In the source figure (`P129_F09`, panel b), this annotation belongs to the short bracket connecting G/Rg3-LNPs and G/RM. The requested G/Cho-LNPs–G/RM comparison has no explicit annotation and should therefore be recorded as `NOT_REPORTED`. The model reproduces the printed value but assigns it to different bracket endpoints, and repeats the error in both its answer and supporting records. This example shows that recovering a statistical value does not ensure that it remains associated with the correct comparison. For this panel-b record, the reference used for the run corrects an earlier annotation with the same binding error. This illustration is restricted to that record and does not validate every comparison in the question.

Correct answer content with omitted evidence records. ACQ524 asks whether measurements for `sex` and `Fp03` are displayed on the selection-frequency axis for adjuvanted donors. Qwen3.6-35B-A3B correctly reports that `sex` is displayed and `Fp03` is absent. Its explanation also correctly distinguishes the adjuvanted-donor panel G from the non-adjuvanted-donor panel H, where `Fp03` appears. However, it returns an empty `checks` list, omitting both required evidence records. The original response is enclosed in a Markdown JSON fence and is rejected by strict parsing (V0). Under the format-normalized procedure (V2), the answer is scored as correct, whereas Joint EM is zero. Thus, the answer–evidence discrepancy in this example reflects failure to submit the required evidence records despite correct answer content. It should not be interpreted as a failure to identify the relevant measurements, or as a valid response under the strict output protocol.

Inconsistency between prerequisite and combined requests. For ACQ103, GPT-5.5 correctly reports all four required statistical comparisons in a separate prerequisite request. In particular, it reads $p = 0.0015$ for DLPE-LNV versus DOPE-LNV on the MHC I/SIINFEKL-displaying BMDC axis (`P13_F03`, panel k). When asked to report the combined set of comparisons, it replaces this value with $p < 0.0001$ in both the answer and supporting records, while retaining the other three comparisons correctly. This is a content discrepancy that remains after format normalization. It demonstrates that successful prerequisite checks do not guarantee consistent reporting in a separate combined response. Because the requests differ in wording and output organization, and the combined task consists of parallel comparison checks, the example does not isolate a failure of multi-step reasoning or establish that the premise was correctly recovered during the combined attempt.

An unreported comparison interpreted as non-significant. In ACQ941 (Figure 5), Qwen3.6-35B-A3B assigns `NS` to Ctrl versus Neut on the change-of-RFI axis (`P255_F03`, panel B). The supplied context states that asterisks denote comparisons against Ctrl and that unmarked bars do not supply a p-value. The Neut bar has no asterisk, so the supported record is `NOT_REPORTED`, rather than an explicit non-significant result. The model's final explanation directly makes the un-

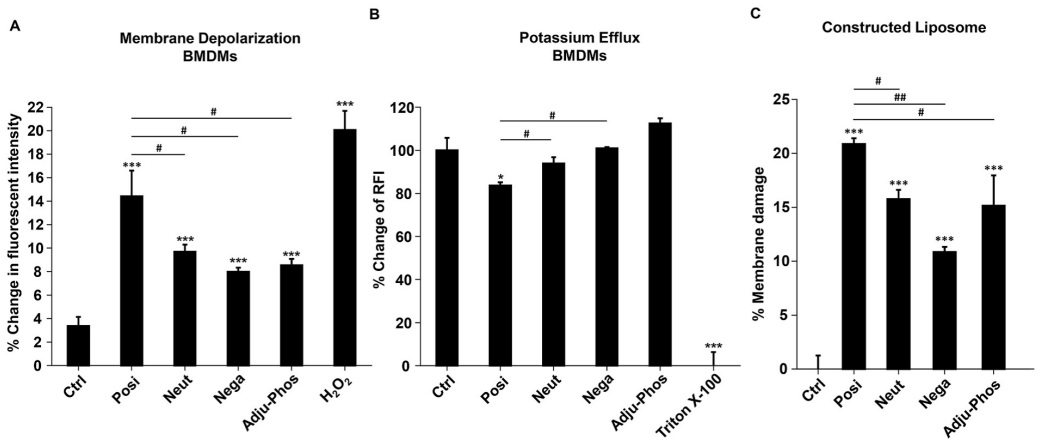

Figure 5: Source figure for ACQ941 (DOI: 10.1016/j.biomaterials.2021.120960). Under the supplied legend convention, asterisks identify comparisons against Ctrl. The unmarked Neut bar in panel B does not establish an explicitly non-significant Ctrl–Neut comparison. The figure is reproduced without altering its annotations.

supported inference from absence of a reported p-value to non-significance. The response also uses panel identifiers that fail the run's identity-matching rules; consequently, its complete evidence-score loss cannot be attributed solely to this semantic error. The source-level discrepancy nonetheless illustrates why absent statistical evidence and an explicitly reported non-significant comparison must remain distinct.