QCalEval: Benchmarking Vision-Language Models for Quantum Calibration Plot Understanding

Shuxiang Cao[1], **Zijian Zhang**[1,2,13], **Abhishek Agarwal**[6], **Grace Bratrud**[10,9], **Daniel C. Cole**[7], **Alejandro Gómez Frieiro**[3], **Elena O. Glen**[11], **Hao Hsu**[3], **Gang Huang**[4], **Raymond Jow**[5], **Greshma Shaji**[3], **Tom Lubowe**[1], **Ligeng Zhu**[1], **Luis Mantilla Calderón**[1,2,13], **Nicola Pancotti**[1], **Joel Pendleton**[5], **Brandon Severin**[5], **Charles Etienne Staub**[8], **Sara Sussman**[9], **Antti Vepsäläinen**[3], **Neel Rajeshbhai Vora**[4], **Yilun Xu**[4], **Varinia Bernales**[2], **Daniel Bowring**[9], **Elica Kyoseva**[1], **Ivan Rungger**[6,12], **Giulia Semeghini**[8], **Sam Stanwyck**[1], **Timothy Costa**[1], **Alán Aspuru-Guzik**[1,2,13], **Krysta Svore**[1]

[1]NVIDIA, [2]University of Toronto, [3]IQM Quantum Computers, [4]Lawrence Berkeley National Laboratory, [5]Conductor Quantum, [6]National Physical Laboratory, [7]Infleqtion, [8]Harvard University, [9]Fermi National Accelerator Laboratory, [10]Northwestern University, [11]EeroQ Corporation, [12]Royal Holloway University of London, [13]Vector Institute for Artificial Intelligence

Quantum computing calibration depends on interpreting experimental data, and calibration plots provide the most universal human-readable representation for this task, yet no systematic evaluation exists of how well vision-language models (VLMs) interpret them. We introduce **QCalEval**, the first VLM benchmark for quantum calibration plots: 243 samples across 87 scenario types from 22 experiment families, spanning superconducting qubits and neutral atoms, evaluated on six question types in both zero-shot and in-context learning settings. The best general-purpose zero-shot model reaches a mean score of 72.3, and many open-weight models degrade under multi-image in-context learning, whereas frontier closed models improve substantially. A supervised fine-tuning ablation at the 9-billion-parameter scale shows that SFT improves zero-shot performance but cannot close the multimodal in-context learning gap. As a reference case study, we release **NVIDIA Ising Calibration 1**, an open-weight model based on Qwen3.5-35B-A3B that reaches 74.7 zero-shot average score.

1. Introduction

Quantum computing systems require continuous calibration to characterize and maintain their operating parameters, as quantum states are sensitive to environmental perturbations. Key calibration targets, including transition frequencies, pulse amplitudes, readout settings, trapping conditions, and couplings, vary by platform and drift over time due to environmental fluctuations and hardware instabilities [1, 2, 3]. As systems scale to hundreds of qubits and beyond, the calibration burden grows combinatorially: each qubit requires dozens of characterization experiments, and the results of one calibration step can invalidate others, creating complex dependency chains [4, 5]. Similarly, results from holistic benchmarking of quantum computers can produce large amounts of data, making it challenging to interpret dependencies across metrics and to analyze the effects of different sources of error [6]. Automating calibration analysis, from interpreting results to reasoning about next steps, is therefore essential to scale calibration and tuning workflows reliably, yet common approaches often still rely on manual or semi-manual calibration tasks. The standard approach to analyzing calibration data is *parametric model fitting*, where an assumed functional form (e.g., a decaying sinusoid for Rabi oscillations) is fit to the measured data. Practitioners assess reliability using goodness-of-fit measures such as R^2 or χ^2, residual structure, parameter uncertainty, and platform-specific heuristics before feeding extracted parameters back into the system.

Recent work has moved toward *agentic* calibration in response to the combinatorial scaling of qubit calibration tasks and need to move beyond manual calibration, where AI agents autonomously orchestrate multi-step calibration workflows. Some of us, in Cao et al. [7], introduced *k-agents*, a system where LLM-based agents decide which experiments to run, interpret results, and adapt the calibration strategy in real time. A key bottleneck for agentic systems such as *k-agents* is *interpreting calibration plots*: the agent must examine an experimental result (typically a plot) to determine whether the experiment succeeded, what went wrong if it failed, and what to do next.

Corresponding author: Shuxiang Cao, shuxiangc@nvidia.com

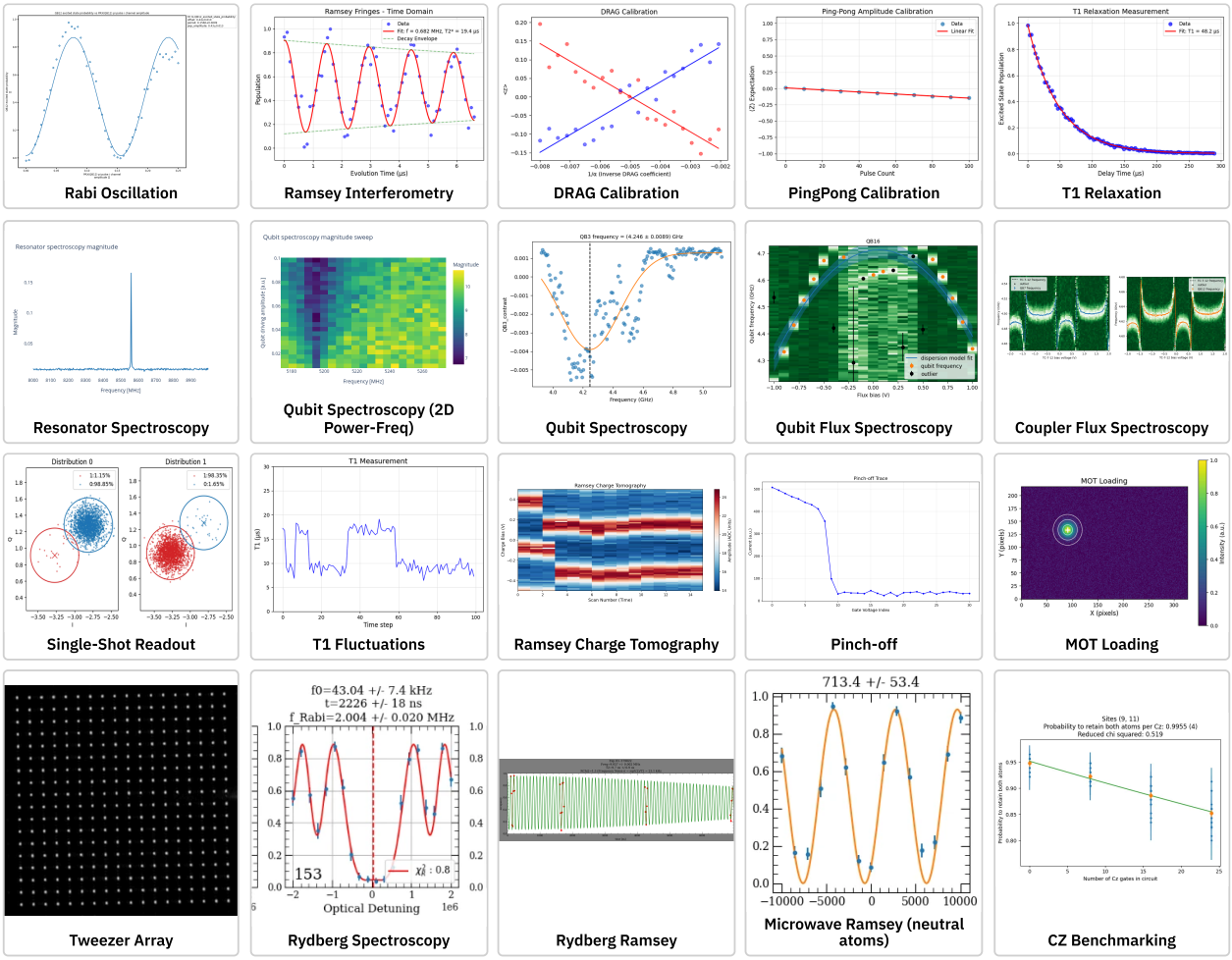


Figure 1 | Representative calibration plots from QCalEval. The benchmark is visually heterogeneous: it includes 1D line traces with oscillations and decays, 2D spectroscopy maps with ridges and hotspots, scatter or histogram-style readout diagnostics, and image-like measurements of spatial structure. Unlike natural-image benchmarks, the key information is carried by scientific geometry rather than object identity: peak locations, fringe spacing, contrast, linewidth, clustering, and the presence or absence of fitted structure determine whether an experiment is reliable or unreliable.

Calibration plots are the *universal human-readable representation* of calibration results. Regardless of hardware platform, software stack, or underlying raw measurement format, any experiment must ultimately be rendered in visual form. This motivates the use of vision-language models (VLMs) as the "eyes" of calibration agents, since a single vision-based model can interpret any calibration experiment without platform-specific integration. This approach is *complementary* to parametric fitting, which assumes a correct model and fails silently when violated [8], whereas a VLM trained on diverse failure modes can catch failures that no single parametric model anticipates.

Despite this potential, no systematic evaluation exists for how well VLMs handle quantum calibration plots. VLMs have demonstrated remarkable capabilities in natural image understanding [9, 10, 11], and multimodal in-context learning (MM-ICL) enables adaptation to new tasks through demonstration examples [12]. However, interpreting calibration plots requires both identifying domain-specific geometric features (oscillation frequencies, decay rates, peak positions) and mapping them to operational statuses, demanding precise visual perception alongside expert domain knowledge. Whether current VLMs can reliably interpret these plots, and whether in-context demonstrations help or hinder, remain open questions.

We introduce **QCalEval**, the first comprehensive benchmark for VLMs on quantum calibration plots, evaluating models under both zero-shot (no examples provided) and in-context learning (ICL; with demonstration

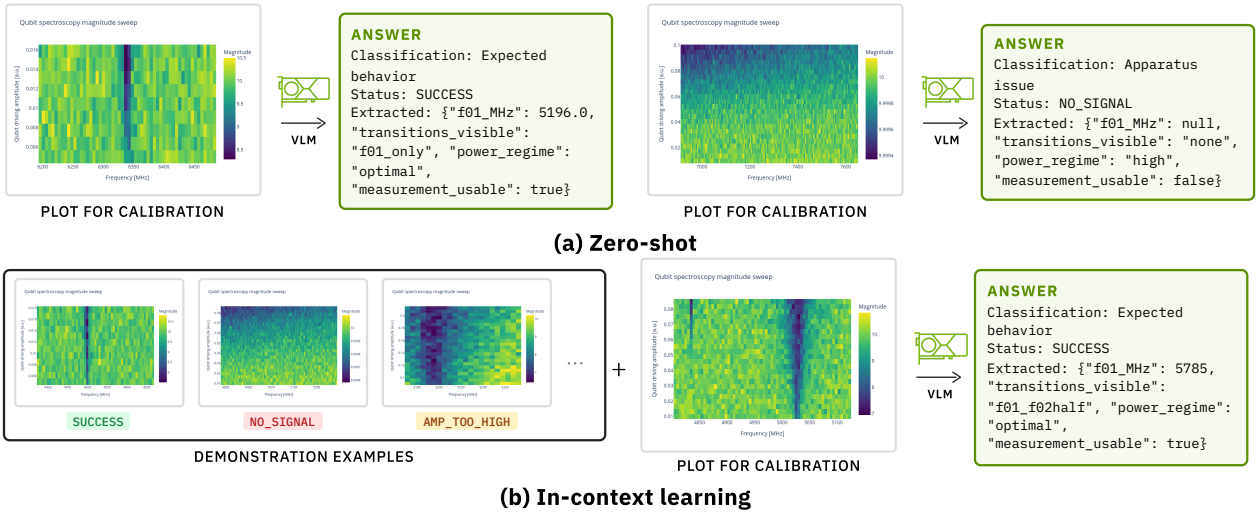


Figure 2 | Task illustration with examples from the qubit spectroscopy experiment. Models are evaluated under both zero-shot (single plot, no demonstrations) and in-context learning (demonstration examples provided) settings. Each evaluation is an independent conversation with no shared context.

examples) conditions (Figure 2). In zero-shot mode, the model receives a single calibration plot and must answer without prior examples; in ICL mode, labeled examples from the same experiment family are provided before the query plot. Our benchmark of 18 VLMs across six question types establishes the first baseline score for this domain: even the best general-purpose zero-shot model achieves a mean score of 72.3. Under in-context learning evaluation, frontier closed models and Gemma 4 improve substantially (up to +29 scores on calibration diagnosis), while many open-weight models degrade with multi-image prompts. A systematic supervised fine-tuning (SFT) ablation (5 recipes at the 9-billion-parameter (9B) scale using Qwen3.5), where training data is formatted either as zero-shot queries (single plot) or ICL queries (with demonstration plots), shows that SFT improves zero-shot performance but cannot close the multimodal in-context learning gap.

Contributions.

1. **QCalEval Benchmark**: The first comprehensive benchmark for VLMs on quantum calibration plots, comprising 243 samples across 87 scenario types from 22 experiment families spanning superconducting qubits and neutral atoms, evaluated on six question types under both zero-shot and in-context learning settings.
2. **Zero-Shot Baseline**: We evaluate 18 VLMs and establish the first score baseline for understanding quantum calibration plots, finding that even the best general-purpose model reaches a mean score of 72.3, revealing this domain as a significant challenge for current VLMs.
3. **MM-ICL Gap**: When provided with labeled visual examples of each scenario type before the query plot, frontier closed models and Gemma improve strongly, while many open-weight models perform *worse* than without any examples, exposing a clear multimodal in-context learning gap.
4. **SFT Ablation**: We systematically evaluate 5 SFT recipes at 9B scale with Qwen3.5, finding that the strongest sequential curriculum in the ablation study is ICL→zero-shot, while no configuration fixes free-text scientific reasoning (Q3) under in-context learning.
5. **NVIDIA Ising Calibration 1**: Guided by the ablation findings, we release an open-weight 35B mixture-of-experts (MoE) model trained with the strongest sequential curriculum identified in that study (two-phase SFT: ICL then zero-shot) as a reference case study.

2. Related Work

Vision-Language Models and Multimodal In-Context Learning. Modern VLMs combine pretrained vision encoders with large language backbones via alignment objectives such as CLIP [13] or modular recipes such as BLIP-2 [14], with instruction tuning converting them into visual assistants [15, 11]. Flamingo [12]

and OpenFlamingo [16] introduced interleaved image-text sequences for few-shot multimodal adaptation, but strong zero-shot performance does not guarantee robust multimodal in-context learning (MM-ICL): VL-ICL Bench [17] and recent analyses [18, 19] show that the effectiveness of in-context learning demonstrations is fragile and highly sensitive to prompt construction.

Chart Understanding and Scientific Figures. Chart reasoning benchmarks evolved from FigureQA [20] and DVQA [21] to PlotQA [22] and ChartQA [23], revealing that chart reasoning depends on OCR, numerical grounding, and structural relations [24]. Chart-specific models include ChartOCR [25], DePlot [26], UniChart [27], ChartLlama [28], ChartInstruct [29], and ChartGemma [30], though evaluations show VLMs remain error-prone on scientific figures [31, 32]. Related scientific-figure resources, such as SciCap [33] and Multimodal ArXiv [34], broaden this line beyond standard charts but still do not target operational diagnosis over calibration plots. These benchmarks focus on general charts rather than expert-oriented diagnosis over quantum calibration plots. Notably, chart-specific pipelines such as DePlot [26] (plot-to-table translation followed by LLM reasoning), UniChart [27] (chart-specific pretraining), and ChartGemma [30] (visual instruction tuning for charts) are designed for data extraction and QA on standard chart types (bar, line, pie). They are not evaluated in our benchmark because quantum calibration requires *domain-specific diagnostic reasoning*, such as determining whether an oscillation frequency is too fast, whether a fit deviation indicates a model failure, or whether a measurement window is sufficient, rather than reading values from axes or answering factual questions about plotted data. Extending these chart-specific approaches to scientific calibration diagnosis is an interesting direction for future work.

Quantum Calibration and Automation. Quantum calibration is an iterative process where practitioners inspect oscillations, decays, and spectroscopy scans to tune device parameters [35, 2], with integrated software stacks for control and analysis [1, 5, 4]. Recent work has begun applying LLMs directly to quantum experiments, including agent-based calibration workflows and instruction-to-experiment translation for superconducting qubit experiments [7, 36], while broader reviews identify calibration and recovery as core automation bottlenecks for scalable quantum computing [37]. Our benchmark evaluates whether VLMs can interpret calibration plots as actionable artifacts for quantum hardware workflows.

Scientific Instrument Agents and Autonomous Laboratories. Across microscopy, beamlines, and autonomous chemistry, LLMs are increasingly used as tool-augmented scientific assistants that translate natural-language goals into scripts, retrieve documentation, orchestrate device APIs, and operate within constrained workflow engines [38, 39, 40, 41, 42, 43, 44]. These systems motivate our focus on execution-adjacent scientific workflows, but they do not benchmark fine-grained visual diagnosis of quantum calibration plots.

3. QCalEval Benchmark

QCalEval evaluates VLM capabilities on quantum calibration plots through six question types, assessed under both *zero-shot* (no demonstrations) and *in-context learning* (with demonstrations) settings.

Benchmark Scope. QCalEval mainly covers **superconducting qubits** and **neutral atoms**, along with emerging platforms, including both shared calibration routines and platform-specific diagnostics. The 22 experiment families span a wide range of visual formats—1D line traces, 2D spectroscopy maps, histograms, and image-like measurements—as illustrated by the representative examples in Figure 1. This mix tests whether VLMs can generalize across familiar calibration patterns as well as domain-specific artifacts and failure modes. The benchmark includes both simulated and real-hardware data provided by multiple partners [45, 46]; Table 1 summarizes the benchmark scale and data-source split, and the full list of experiment families is provided in the appendix. Each experiment family defines multiple scenario types (typically two to seven), such as different success and failure modes; each scenario type contains one or more benchmark samples, where each sample comprises one or more calibration plot images, a scenario type label, and ground-truth answers for all six question types.

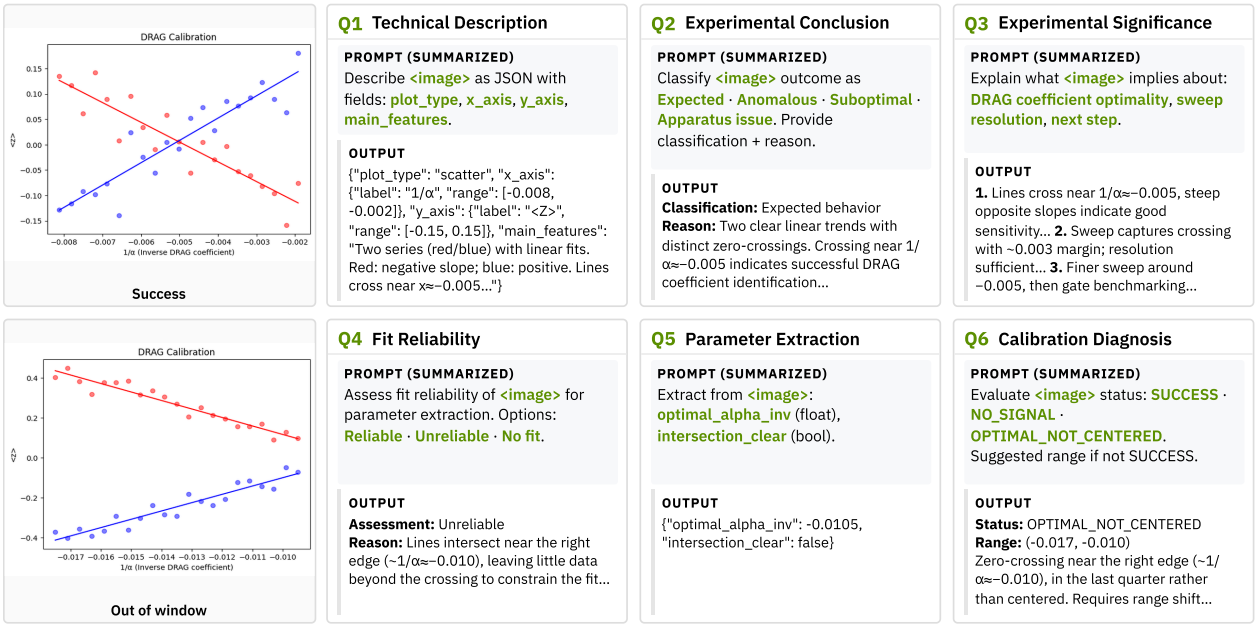


Figure 3 | The six question types in QCalEval, illustrated on a DRAG calibration example. Q1 uses a universal prompt across all families; Q2–Q6 prompts are customized per experiment family with family-specific background context, failure mode definitions, and parameter extraction schemas.

Task Taxonomy and Motivation. We define six question types because quantum calibration assistance is not a single visual question-answering problem, but a pipeline from accurately perceiving a plot to making an operational calibration decision. Each question isolates a different failure mode that would be hidden by a single aggregate score:

- **Q1** (*technical description*): a structured JSON description of the plot type, axes, and salient visual features, isolating visual grounding from domain reasoning.
- **Q2** (*experimental conclusion*): a coarse 4-way outcome classification (`Expected behavior`, `Suboptimal parameters`, `Anomalous behavior`, `Apparatus issue`), testing whether the model can map visual evidence to an experimental interpretation.
- **Q3** (*experimental significance*): experiment-specific scientific analysis of what the observed pattern implies, whether the sweep window or resolution is sufficient, and what next calibration step should follow.
- **Q4** (*fit reliability*): whether a visible fit, if present, is trustworthy for downstream use, forcing a decision over `Reliable`, `Unreliable`, or `No fit`.
- **Q5** (*parameter extraction*): machine-readable extraction of family-specific physical parameters in structured JSON.
- **Q6** (*calibration diagnosis*): the most operational task, assigning a family-specific status code (e.g., `SUCCESS`, `NO_SIGNAL`) and, when needed, providing a corrective range or suggested action.

This decomposition mirrors how experimentalists actually use calibration plots: first identify what is shown, then judge what broad outcome it represents, reason about its scientific meaning, assess whether quantitative readout is justified, extract any usable parameters, and finally decide what action to take. We intentionally include both Q2 and Q6 because they operate at different granularities: Q2 is a shared coarse outcome label across all experiment families, while Q6 is a family-specific actionable diagnosis.

Zero-Shot Evaluation. Each sample is evaluated on all six question types without any demonstration examples (Figure 3): the model receives only the query plot and textual background. For Q2–Q6, prompts include textual domain knowledge, provided as part of the released benchmark for all 22 experiment families: each background describes what the experiment measures, what a successful result looks like, and the set

Table 1 | QCalEval benchmark statistics.

Dataset Statistics		Benchmark Evaluation	
Benchmark samples	243	Zero-shot	243 samples × 6 Qs
Scenario types	87	In-context learning	236 samples × 3 Qs
Experiment families	22		
Unique images	309		

of possible outcome or status labels. This establishes a baseline for how well VLMs understand quantum calibration plots in the absence of visual examples, using only textual context.

In-Context Learning Evaluation. In-context learning provides demonstration examples (previously collected calibration plots with expert-assigned labels) of each scenario type, directly showing the model what each failure mode looks like. These demonstrations are drawn from other samples within the same experiment family in the benchmark, mirroring how practitioners in practice would reference past calibration runs with known outcomes to interpret a new result. Three of the six question types (Q3, Q5, Q6) are additionally evaluated with demonstration examples, each posed as an *independent conversation* with no shared context:

1. **N-way In-Context Analysis (Q3).** The model receives one demonstration per scenario type in the experiment family, each pairing a scenario plot with an expert analysis describing the observed behavior, its diagnostic implications, and recommended next steps. Given a new query plot, the model must produce a similarly structured analysis.
2. **1-Shot Parameter Extraction (Q5).** The model receives a single worked example showing a plot and its extracted parameters as a structured JSON object (e.g., `{"optimal_alpha_inv": -0.005, "intersection_clear": true}`). Given a new query plot, the model must extract analogous parameters.
3. **N-way In-Context Classification (Q6).** The model receives one labeled example per scenario type in the experiment family (e.g., `SUCCESS`, `NO_SIGNAL`), and must classify a new query plot into one of these scenarios.

Scenario types with only one sample are excluded from in-context learning evaluation because they cannot provide support examples without reusing the query itself. We restrict in-context learning evaluation to Q3, Q5, and Q6 because these are the tasks where demonstrations can plausibly transfer structured reasoning patterns, output schemas, or family-specific label vocabularies. By contrast, Q1, Q2, and Q4 already have tightly specified zero-shot output spaces and are less informative probes of MM-ICL. Comparing zero-shot and in-context learning scores on Q3, Q5, and Q6 directly measures whether models can exploit in-context demonstrations.

4. Results and Analysis

We evaluate 18 VLMs spanning frontier closed-source models (GPT-5.4 [47], Gemini 3.1 [48], Claude 4.6 [49]), open-weight models (Qwen3.5 [50], Gemma 4 [51], InternVL3 [52], Kimi-VL [53], MiniCPM-o [54]), and one domain-tuned case-study model **NVIDIA Ising Calibration 1** (Ising-Cal-1) on all six QCalEval question types. Q2, Q4, Q5, and Q6 are scored programmatically; Q1 and Q3 are scored by two independent LLM judges (GPT-5.4 and Gemini 3.1 Pro) and averaged to reduce single-judge bias (details in Appendix B). We report zero-shot and in-context learning (MM-ICL) results. We additionally conduct an N-way scaling study on Q6, varying the number of demonstration examples from 0 to 5 shots across four experiment families with five scenario types each, to probe whether the MM-ICL gap is caused by image overload. Table 2 summarizes the best-performing model for each question type across all evaluation settings; full per-model results appear in Tables 3 and 4.

Table 2 | Best score per question type across all evaluation settings. Q1, Q2, and Q4 are evaluated under zero-shot only; Q3, Q5, and Q6 are evaluated under both zero-shot and ICL. †Domain-tuned case study (Section 5.2).

Task	Best Model	Score
Q1 (Description)	GPT-5.4	90.9
Q2 (Conclusion)	Ising-Cal-1-35B†	67.1
Q3 (Significance)	Claude Opus 4.6 (ICL)	84.7
Q4 (Fit reliability)	Ising-Cal-1-35B†	90.5
Q5 (Param. extraction)	Gemini-3.1-Pro (ICL)	84.5
Q6 (Diagnosis)	Gemini-3.1-Pro (ICL)	89.8

Table 3 | Zero-shot scores across six evaluation axes. Models grouped by access type. **Bold** = best per column. †Domain-tuned case study (Section 5.2).

	Model	Q1	Q2	Q3	Q4	Q5	Q6	Mean
Closed	Gemini-3.1-Pro	88.5	57.2	61.1	84.4	**71.5**	71.2	72.3
	Claude Opus 4.6	90.8	49.0	**65.5**	76.1	64.7	60.5	67.8
	Gemini-3.1-Flash-Lite	89.2	53.5	59.4	82.7	63.8	60.9	68.2
	Claude Sonnet 4.6	89.7	48.6	63.4	76.5	60.4	60.1	66.5
	GPT-5.4	**90.9**	52.7	63.7	54.7	64.3	61.3	64.6
	GPT-5.4-Mini	90.3	39.5	48.3	42.0	62.6	51.4	55.7
	Claude Haiku 4.5	83.4	36.6	40.8	48.6	51.0	42.8	50.5
Open	Gemma-4-31B-IT	85.6	54.3	59.8	82.7	68.3	62.1	68.8
	Qwen3.5-397B-A17B	88.1	42.8	52.0	50.6	62.5	55.6	58.6
	Qwen3.5-27B	87.0	45.7	48.3	56.4	58.7	55.1	58.5
	Qwen3.5-122B-A10B	86.6	44.0	49.0	50.2	61.2	51.9	57.1
	Qwen3.5-35B-A3B	86.8	39.9	45.7	52.7	57.8	50.6	55.5
	Qwen3.5-9B	81.5	37.9	39.5	49.8	57.1	52.3	53.0
	InternVL3-78B	76.3	37.0	34.1	42.8	52.9	45.7	48.2
	MiniCPM-o-4.5	76.7	31.7	29.8	32.5	47.9	48.1	44.5
	InternVL3-38B	79.2	34.6	27.6	33.7	49.2	40.3	44.1
	Kimi-VL-A3B	65.0	34.6	22.1	35.0	38.9	37.4	38.9
	Ising-Cal-1-35B†	87.8	**67.1**	64.7	**90.5**	62.5	**75.3**	**74.7**

4.1. Zero-Shot Performance

Finding 1: Models detect visual features well but lack domain knowledge to interpret them. The best base model (Gemini-3.1-Pro) achieves a mean score of 72.3; the domain-tuned Ising-Cal-1 reaches 74.7. Gemini-3.1-Pro leads on parameter extraction (Q5: 71.5) while Ising-Cal-1 leads on outcome classification (Q2: 67.1), fit assessment (Q4: 90.5), and calibration diagnosis (Q6: 75.3), reflecting the benefit of domain tuning on tasks that require expert knowledge. Interpreting calibration plots requires two capabilities: detecting relevant visual features and mapping them to domain-specific operational outcomes. Models perform well on feature detection (Q1: 65–91%), but performance drops sharply on tasks requiring domain knowledge: outcome classification (Q2: 32–67%) and calibration diagnosis (Q6: 37–75%). The primary failure mode is optimistic bias: across all models, 60.7% of "Suboptimal parameters" cases are classified as "Expected behavior" — models see the features but default to success without the domain knowledge to interpret them. Difficulty scales with domain knowledge required: families where failure is visually obvious, such as resonator spectroscopy (100%), are solved by all models, while families where failures resemble successful experiments, such as single-shot readout and Ramsey T2*, require domain expertise to identify (Appendix D.1).

Finding 2: Fit assessment exposes a visual judgment gap. Fit assessment (Q4: 33–84%) does not require domain expertise but tests whether models can judge if a fitted curve matches the underlying data. Two failure modes emerge: *false confidence*, where models predict "Reliable" for unreliable fits because the curve visually looks plausible despite poor fit quality (ranging from 6% for Gemini-3.1-Flash-Lite to 74% for

Table 4 | In-context learning (MM-ICL) scores. Q3 uses N-way analysis demos, Q5 uses 1-shot extraction demos, Q6 uses N-way classification demos. Δ = change from zero-shot. green = improvement, red = degradation.

	Model	Q3	Δ	Q5	Δ	Q6	Δ	Mean
	Best zero-shot (base)	*65.5*		*71.5*		*71.2*		
Closed	Gemini-3.1-Pro	81.3	+20.2	**84.5**	+13.0	**89.8**	+18.6	**85.2**
Closed	Claude Opus 4.6	**84.7**	+19.2	81.3	+16.6	89.4	+28.9	85.1
Closed	Claude Sonnet 4.6	77.8	+14.4	71.9	+11.5	78.0	+17.9	75.9
Closed	Gemini-3.1-Flash-Lite	78.5	+19.1	73.6	+9.8	82.2	+21.3	78.1
Closed	GPT-5.4	81.0	+17.3	72.9	+8.6	81.4	+20.1	78.4
Closed	GPT-5.4-Mini	58.8	+10.5	72.7	+10.1	66.9	+15.5	66.1
Closed	Claude Haiku 4.5	66.1	+25.3	58.7	+7.7	73.1	+30.3	66.0
Open	Gemma-4-31B-IT	80.6	+20.8	76.9	+8.6	86.0	+23.9	81.2
Open	InternVL3-38B	56.2	+28.6	59.5	+10.3	55.1	+14.8	56.9
Open	Qwen3.5-27B	41.8	-6.5	71.5	+12.8	45.8	-9.3	53.0
Open	InternVL3-78B	50.5	+16.4	46.2	-6.7	44.3	-1.4	47.0
Open	Qwen3.5-397B-A17B	37.4	-14.6	64.3	+1.8	42.4	-13.2	48.0
Open	Qwen3.5-122B-A10B	36.1	-12.9	62.5	+1.3	35.2	-16.7	44.6
Open	Qwen3.5-35B-A3B	33.4	-12.3	64.4	+6.6	33.9	-16.7	43.9
Open	Qwen3.5-9B	32.8	-6.7	63.0	+5.9	33.9	-18.4	43.2
Open	Kimi-VL-A3B	34.9	+12.8	54.3	+15.4	32.6	-4.8	40.6
Open	MiniCPM-o-4.5	19.3	-10.5	50.5	+2.6	29.2	-18.9	33.0

InternVL3-38B); and *no-fit blindness*, where models label raw data without fitted curves as "Unreliable" rather than "No fit" (ranging from 0% for Ising-Cal-1, Gemma-4-31B, and Gemini-3.1-Pro to 91% for InternVL3-38B). Models that learn to output "No fit" (matching the 31.7% ground-truth rate) perform better than those that almost never predict it.

4.2. In-Context Learning: Can Demonstrations Help?

A summary of results for in-context learning is given in Table 4.

Finding 3: In-context demonstrations improve frontier models and Gemma. Since zero-shot models lack domain-specific knowledge (Finding 1), in-context demonstrations can inject this knowledge directly by providing labeled examples of each scenario type. All seven closed-source models improve on every axis with in-context demonstrations. Claude Opus 4.6 gains +28.9 points on Q6 (60.5→89.4), and Gemini-3.1-Pro gains +18.6 (71.2→89.8). Notably, Gemma-4-31B-IT is the *only open model* we benchmarked that benefits comparably, gaining +23.9 on Q6 (62.1→86.0), matching closed-model behavior while being open-weight. The improvements are largest on Q6 (calibration diagnosis), the most practically relevant task, where labeled examples provide the status vocabulary the model needs.

Finding 4: Multi-image demonstrations degrade open-weight models. In contrast, Qwen3.5 models, MiniCPM-o, and Kimi-VL are *actively harmed* by multi-image demonstrations: Qwen3.5-9B drops 18.4 points on Q6, and Qwen3.5-35B-A3B drops 16.7. The degradation is specific to N-way multi-image prompts: 1-shot parameter extraction (Q5), which uses a single demonstration, improves for nearly all models. This suggests these model families can process a single reference image but fail to relate multiple labeled demonstrations to a query.

Finding 5: N-way scaling shows the MM-ICL gap is not a simple image-overload effect. To determine whether degradation under in-context learning is simply caused by image overload, we conduct N-way scaling experiments (0–5 shots, 1,008 QA pairs; details in Appendix F). Frontier models improve consistently with more demonstrations (Gemini: 44→85; Gemma-4-31B: 49→82), proving the task supports

Table 5 | NVIDIA Ising Calibration 1 scores on QCalEval compared to its base model. **Bold** = best per column.

	Zero-Shot							ICL		
Model	**Q1**	**Q2**	**Q3**	**Q4**	**Q5**	**Q6**	**Avg**	**Q3**	**Q5**	**Q6**
Qwen3.5-35B base	86.8	39.9	45.7	52.7	57.8	50.6	55.5	**33.4**	**64.4**	33.9
Ising-Cal-1	**87.8**	**67.1**	**64.7**	**90.5**	**62.5**	**75.3**	**74.7**	31.2	59.8	**42.4**
Δ	+1.0	+27.2	+19.0	+37.8	+4.7	+24.7	+19.2	-2.2	-4.6	+8.5

multi-image reasoning. In contrast, Qwen3.5, MiniCPM-o, and Kimi-VL all peak at 1-shot then degrade, and larger Qwen models do not recover this behavior (397B peaks at 35 vs. 9B at 44). These results show that the MM-ICL gap is not simply a consequence of too many images or task difficulty: reducing the number of demonstrations does not eliminate the failure pattern in these model families.

Failure patterns. Qualitative analysis (Appendix E) reveals recurring failure modes: (1) *visual similarity confusion*, where models correctly describe plots but misclassify scenarios differing in subtle features like oscillation frequency; (2) *optimistic bias*, where 85% of Q6 errors predict success when ground truth indicates failure; (3) *"No fit" blindness*, where models report "Reliable" for raw data without fitted curves, affecting 38% of samples.

5. SFT Ablation and Case Study

5.1. Ablation Study: What Can Fine-Tuning Improve?

For the ablation study using Qwen3.5 9B models, we use a *train/test-family split*: models are trained on one set of experiment families and evaluated on previously unseen families, rather than on a standard sample-level partition. We generate two SFT datasets by applying QCalEval's question templates to synthetic calibration plots from the train families, paired with ground-truth answers: **Zero-shot** (25.8K zero-shot QA pairs) and **ICL** (12.9K ICL QA pairs), and evaluate five recipes: zero-shot-only, ICL-only, blend, ICL→zero-shot, and zero-shot→ICL. Full results are in Appendix G.

Key findings from the ablation study. We find that: (1) **SFT substantially improves zero-shot scores**: zero-shot SFT improves Q6 from 61.1 to 70.6 (+9.5 points), while sequential curricula raise Q4 from 28.6 to about 60 in the train/test-family ablation. (2) **In-context learning remains difficult after SFT**: different recipes improve different classification axes, but gains are modest and inconsistent across Q5 and Q6. (3) **No recipe improves Q3 under in-context learning**: the best result (24.1) remains below the base score (27.1), suggesting that free-text scientific reasoning may require advances beyond supervised fine-tuning, such as reasoning-oriented training or reinforcement learning from expert feedback. (4) **Sequential training order matters**: among sequential curricula, ICL→zero-shot is the strongest overall in the train/test-family ablation. We hypothesize that training on zero-shot data first teaches the model to rely solely on the query image; subsequent ICL training then fails to override this habit, as the model has already learned it can answer without referencing demonstration images.

5.2. NVIDIA Ising Calibration 1

Guided by the ablation study, we release **NVIDIA Ising Calibration 1**, an open-weight Qwen3.5-35B-A3B (MoE) model [50] trained with the strongest sequential SFT curriculum identified in that study (ICL→zero-shot). Model configuration details are in Appendix H; weights are at `https://huggingface.co/nvidia/Ising-Calibration-1-35B-A3B`.

Table 5 shows that the benchmark-derived SFT recipe substantially improves the base Qwen3.5-35B model in zero-shot evaluation, with the clearest gains on diagnosis, scientific interpretation, and fit-related judgments. The improvement is broad but not uniform: zero-shot plot understanding benefits strongly from

domain tuning, whereas in-context learning remains weak and largely retains the same multi-demonstration failure pattern as the base model. This makes Ising Calibration 1 a stronger reference model for single-plot understanding on QCalEval, but not a solution to the MM-ICL gap.

6. Limitations

Dataset size and coverage. QCalEval contains 243 samples across 22 experiment families, mainly covering superconducting qubits and neutral atoms, with 186 simulated and 57 hardware samples. Future work can expand the benchmark to cover more experiment types, hardware platforms, and modalities. The per-family sample sizes (2–21 samples) limit the statistical power to draw fine-grained per-experiment conclusions.

7. Conclusion

We introduced QCalEval, the first comprehensive benchmark for VLMs on quantum calibration plots, evaluating 18 models across six question types under both zero-shot and in-context learning settings. Our zero-shot evaluation reveals a two-step capability gap: models detect visual features well (Q1: 65–91%) but lack the domain knowledge to map them to operational outcomes (Q2: 32–67%, Q6: 37–75%), with systematic optimistic bias toward predicting success. Fit assessment (Q4) reveals a distinct visual judgment gap, in which weaker models cannot distinguish reliable from unreliable fits. In-context learning, which injects domain knowledge through labeled visual examples, improves frontier models, but many open-weight models degrade under multi-image prompts. A systematic SFT ablation at the 9-billion-parameter scale shows that supervision format is critical: sequential curricula raise Q4 from 28.6 to about 60, and Q6 from 61.1 to 70.6 in zero-shot evaluation, but zero-shot SFT degrades in-context learning performance, and no configuration closes the gap on free-text scientific reasoning (Q3). As a case study, we release NVIDIA Ising Calibration 1, an open model based on Qwen3.5-35B-A3B (MoE) and trained using a two-phase sequential SFT recipe. Reliable calibration plot understanding is a prerequisite for autonomous quantum computing workflows; by releasing QCalEval and Ising Calibration 1 as open resources, we enable the community to benchmark progress and fine-tune models for their specific hardware platforms and experiment types.

Data and Code Availability

The QCalEval benchmark dataset is available at https://huggingface.co/datasets/nvidia/QCalEval. Evaluation scripts are available at https://github.com/nvidia/QCalEval. NVIDIA Ising Calibration 1 model weights are available at https://huggingface.co/nvidia/Ising-Calibration-1-35B-A3B.

Acknowledgments

We thank Shi Xuan Leong for contributions to early discussions, and Lilian Zhong for assistance with data verification. This manuscript has been authored by Fermi Research Alliance, LLC under Contract No. DE-AC02-07CH11359 with the U.S. Department of Energy, Office of Science, Office of High Energy Physics. G. H., N. V. and Y. X. acknowledged the support from the U.S. Department of Energy, Office of Science, National Quantum Information Science Research Centers, Quantum Systems Accelerator (Award No. DESCL0000121), and Advanced Scientific Computing Research Testbeds for Science program under Contract No. DE-AC02-05CH11231. A. G. F. and A. V. acknowledged the support from Business Finland through project CfoQ (787/31/2025). A.A. and I.R. acknowledge the support of the UK government Department for Science, Innovation and Technology through the UK National Quantum Technologies Programme, and of the Engineering and Physical Sciences Research Council (Grant No. EP/Z53318X/1: QCI3 Hub). L.M.C. acknowledges funding from the Novo Nordisk Foundation, Grant number NNF22SA0081175, NNF Quantum Computing Programme. A.A.G. thanks Anders G. Frøseth for his generous support and acknowledges the generous support of Natural Resources Canada and the Canada 150 Research Chairs program. A.A.-G. and V.B. acknowledge the University of Toronto's Acceleration Consortium, which receives funding from the CFREF-2022-00042 Canada First Research Excellence Fund. A.A.-G. acknowledges support from the AI2050 program from the Schmidt Foundation. D.C.C. gratefully acknowledges contributions from Infleqtion team members Garrett Hickman, Kevin Kuper, David Mason, and Peter Mitchell.

References

[1] Nicolas Wittler, Federico Roy, Kevin Pack, Max Werninghaus, Anurag Saha Roy, Daniel J. Egger, Stefan Filipp, Frank K. Wilhelm, and Shai Machnes. Integrated tool set for control, calibration, and characterization of quantum devices applied to superconducting qubits. *Physical Review Applied*, 15(3):034080, March 2021. ISSN 2331-7019. doi: 10.1103/physrevapplied.15.034080. URL http://dx.doi.org/10.1103/PhysRevApplied.15.034080.

[2] Max Werninghaus, Daniel J Egger, Federico Roy, Shai Machnes, Frank K Wilhelm, and Stefan Filipp. High-speed calibration and characterization of superconducting quantum processors without qubit reset. *PRX Quantum*, 2(2):020324, May 2021. ISSN 2691-3399. doi: 10.1103/PRXQuantum.2.020324. URL https://doi.org/10.1103/PRXQuantum.2.020324.

[3] Abhishek Agarwal, Lachlan P. Lindoy, Deep Lall, Sebastian E. de Graaf, Tobias Lindström, and Ivan Rungger. Fast-tracking and disentangling of qubit noise fluctuations using minimal-data averaging and hierarchical discrete fluctuation auto-segmentation. *arXiv preprint arXiv:2505.23622*, 2025. URL https://arxiv.org/abs/2505.23622.

[4] Andrea Pasquale, Stavros Efthymiou, Sergi Ramos-Calderer, Jadwiga Wilkens, Ingo Roth, and Stefano Carrazza. Towards an open-source framework to perform quantum calibration and characterization. *arXiv preprint arXiv:2303.10397*, 2023. URL https://arxiv.org/abs/2303.10397.

[5] Naoki Kanazawa, Daniel J. Egger, Yael Ben-Haim, Helena Zhang, William E. Shanks, Gadi Aleksandrowicz, and Christopher J. Wood. Qiskit experiments: A python package to characterize and calibrate quantum computers. *Journal of Open Source Software*, 8(84):5329, April 2023. ISSN 2475-9066. doi: 10.21105/joss.05329. URL http://dx.doi.org/10.21105/joss.05329.

[6] Deep Lall, Abhishek Agarwal, Weixi Zhang, Lachlan Lindoy, Tobias Lindström, Stephanie Webster, Simon Hall, Nicholas Chancellor, Petros Wallden, Raul Garcia-Patron, Elham Kashefi, Viv Kendon, Jonathan Pritchard, Alessandro Rossi, Animesh Datta, Theodoros Kapourniotis, Konstantinos Georgopoulos, and Ivan Rungger. A review and collection of metrics and benchmarks for quantum computers: definitions, methodologies and software, 2025. URL https://arxiv.org/abs/2502.06717.

[7] Shuxiang Cao, Zijian Zhang, Mohammed Alghadeer, Simone D. Fasciati, Michele Piscitelli, Mustafa Bakr, Peter Leek, and Alán Aspuru-Guzik. Automating quantum computing laboratory experiments with an agent-based AI framework. *Patterns*, 6(10):101372, October 2025. ISSN 2666-3899. doi: 10.1016/j.patter.2025.101372. URL https://doi.org/10.1016/j.patter.2025.101372.

[8] Abhishek Agarwal, Lachlan P Lindoy, Deep Lall, François Jamet, and Ivan Rungger. Modelling non-Markovian noise in driven superconducting qubits. *Quantum Science and Technology*, 9(3):035017, April 2024. ISSN 2058-9565. doi: 10.1088/2058-9565/ad3d7e. URL https://doi.org/10.1088/2058-9565/ad3d7e.

[9] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023. URL https://arxiv.org/abs/2308.12966.

[10] OpenAI. GPT-4V(ision) System Card. Technical report, OpenAI, September 2023. URL https://cdn.openai.com/papers/GPTV_System_Card.pdf. Published September 25, 2023.

[11] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, pages 34892–34916, 2023. URL https://papers.nips.cc/paper_files/paper/2023/hash/6dcf277ea32ce3288914faf369fe6de0-Abstract-Conference.html.

[12] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikołaj Bińkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. Flamingo: a visual language

model for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 35, pages 23716–23736, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/960a172bc7fbf0177ccccbb411a7d800-Abstract-Conference.html.

[13] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

[14] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR, 2023. URL https://proceedings.mlr.press/v202/li23q.html.

[15] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, pages 49250–49267, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/9a6a435e75419a836fe47ab6793623e6-Abstract-Conference.html.

[16] Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, Jenia Jitsev, Simon Kornblith, Pang Wei Koh, Gabriel Ilharco, Mitchell Wortsman, and Ludwig Schmidt. Openflamingo: An open-source framework for training large autoregressive vision-language models, 2023. URL https://arxiv.org/abs/2308.01390.

[17] Yongshuo Zong, Ondrej Bohdal, and Timothy M. Hospedales. Vl-icl bench: The devil in the details of multimodal in-context learning. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2403.13164.

[18] Sivan Doveh, Shaked Perek, M. Jehanzeb Mirza, Wei Lin, Amit Alfassy, Assaf Arbelle, Shimon Ullman, and Leonid Karlinsky. Towards multimodal in-context learning for vision & language models. In *Computer Vision – ECCV 2024 Workshops*, pages 250–267. Springer Nature Switzerland, 2025. ISBN 9783031938061. doi: 10.1007/978-3-031-93806-1_19. URL https://doi.org/10.1007/978-3-031-93806-1_19.

[19] Yixing Jiang, Jeremy Irvin, Ji Hun Wang, Muhammad Ahmed Chaudhry, Jonathan H. Chen, and Andrew Y. Ng. Many-shot in-context learning in multimodal foundation models. *arXiv preprint arXiv:2405.09798*, 2024. doi: 10.48550/ARXIV.2405.09798. URL https://arxiv.org/abs/2405.09798.

[20] Samira Ebrahimi Kahou, Adam Atkinson, Vincent Michalski, Ákos Kádár, Adam Trischler, and Yoshua Bengio. FigureQA: An annotated figure dataset for visual reasoning. In *International Conference on Learning Representations*, 2018. URL https://arxiv.org/abs/1710.07300. Workshop Track.

[21] Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. Dvqa: Understanding data visualizations via question answering. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5648–5656. IEEE, June 2018. doi: 10.1109/CVPR.2018.00592. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Kafle_DVQA_Understanding_Data_CVPR_2018_paper.html.

[22] Nitesh Methani, Pritha Ganguly, Mitesh M. Khapra, and Pratyush Kumar. Plotqa: Reasoning over scientific plots. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1527–1536, 2020. URL https://openaccess.thecvf.com/content_WACV_2020/html/Methani_PlotQA_Reasoning_over_Scientific_Plots_WACV_2020_paper.html.

[23] Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2263–2279, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.177. URL https://aclanthology.org/2022.findings-acl.177/.

[24] Enamul Hoque, Parsa Kavehzadeh, and Ahmed Masry. Chart question answering: State of the art and future directions. *Computer Graphics Forum*, 41(3):555–572, June 2022. ISSN 1467-8659. doi: 10.1111/cgf.14573. URL https://doi.org/10.1111/cgf.14573.

[25] Junyu Luo, Zekun Li, Jinpeng Wang, and Chin-Yew Lin. Chartocr: Data extraction from charts images via a deep hybrid framework. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1917–1925, 2021. URL https://openaccess.thecvf.com/content/WACV2021/html/Luo_ChartOCR_Data_Extraction_From_Charts_Images_via_a_Deep_Hybrid_WACV_2021_paper.html.

[26] 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 Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10381–10399, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.660. URL https://aclanthology.org/2023.findings-acl.660/.

[27] Ahmed Masry, Parsa Kavehzadeh, Xuan Long Do, Enamul Hoque, and Shafiq Joty. UniChart: A universal vision-language pretrained model for chart comprehension and reasoning. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14662–14684, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.906. URL https://aclanthology.org/2023.emnlp-main.906/.

[28] Yucheng Han, Chi Zhang, Xin Chen, Xu Yang, Zhibin Wang, Gang Yu, Bin Fu, and Hanwang Zhang. Chartllama: A multimodal llm for chart understanding and generation. *arXiv preprint arXiv:2311.16483*, 2023. URL https://arxiv.org/abs/2311.16483.

[29] Ahmed Masry, Mehrad Shahmohammadi, Md Rizwan Parvez, Enamul Hoque, and Shafiq Joty. ChartInstruct: Instruction tuning for chart comprehension and reasoning. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 10387–10409, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.619. URL https://aclanthology.org/2024.findings-acl.619/.

[30] Ahmed Masry, Megh Thakkar, Aayush Bajaj, Aaryaman Kartha, Enamul Hoque, and Shafiq Joty. ChartGemma: Visual instruction-tuning for chart reasoning in the wild. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, Steven Schockaert, Kareem Darwish, and Apoorv Agarwal, editors, *Proceedings of the 31st International Conference on Computational Linguistics: Industry Track*, pages 625–643, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-industry.54/.

[31] Mohammed Saidul Islam, Raian Rahman, Ahmed Masry, Md Tahmid Rahman Laskar, Mir Tafseer Nayeem, and Enamul Hoque. Are large vision language models up to the challenge of chart comprehension and reasoning? an extensive investigation into the capabilities and limitations of lvlms. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3334–3368, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.191. URL https://aclanthology.org/2024.findings-emnlp.191/.

[32] Jonathan Roberts, Kai Han, Neil Houlsby, and Samuel Albanie. Scifibench: Benchmarking large multimodal models for scientific figure interpretation. In *Advances in Neural Information Processing Systems 37*, NeurIPS 2024, pages 18695–18728. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2024. doi: 10.52202/079017-0593. URL https://arxiv.org/abs/2405.08807.

[33] Ting-Yao Hsu, C Lee Giles, and Ting-Hao Huang. SciCap: Generating captions for scientific figures. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3258–3264, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.277. URL https://aclanthology.org/2021.findings-emnlp.277/.

[34] Lei Li, Yuqi Wang, Runxin Xu, Peiyi Wang, Xiachong Feng, Lingpeng Kong, and Qi Liu. Multimodal ArXiv: A dataset for improving scientific comprehension of large vision-language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14369–14387, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.775. URL https://aclanthology.org/2024.acl-long.775/.

[35] M. A. Rol, C. C. Bultink, T. E. O'Brien, S. R. de Jong, L. S. Theis, X. Fu, F. Luthi, R. F. L. Vermeulen, J. C. de Sterke, A. Bruno, D. Deurloo, R. N. Schouten, F. K. Wilhelm, and L. DiCarlo. Restless tuneup of high-fidelity qubit gates. *Physical Review Applied*, 7(4):041001, April 2017. ISSN 2331-7019. doi: 10.1103/PhysRevApplied.7.041001. URL https://doi.org/10.1103/PhysRevApplied.7.041001.

[36] Shiheng Li, Jacob M. Miller, Phoebe J. Lee, Gustav Andersson, Christopher R. Conner, Yash J. Joshi, Bayan Karimi, Amber M. King, Howard L. Malc, Harsh Mishra, Hong Qiao, Minseok Ryu, Xuntao Wu, Siyuan Xing, Haoxiong Yan, Jian Shi, and Andrew N. Cleland. Large language model-assisted superconducting qubit experiments. *arXiv preprint arXiv:2603.08801*, 2026. URL https://arxiv.org/abs/2603.08801.

[37] Yuri Alexeev, Marwa H. Farag, Taylor L. Patti, Mark E. Wolf, Natalia Ares, Alán Aspuru-Guzik, Simon C. Benjamin, Zhenyu Cai, Shuxiang Cao, Christopher Chamberland, Zohim Chandani, Federico Fedele, Ikko Hamamura, Nicholas Harrigan, Jin-Sung Kim, Elica Kyoseva, Justin G. Lietz, Tom Lubowe, Alexander McCaskey, Roger G. Melko, Kouhei Nakaji, Alberto Peruzzo, Pooja Rao, Bruno Schmitt, Sam Stanwyck, Norm M. Tubman, Hanrui Wang, and Timothy Costa. Artificial intelligence for quantum computing. *Nature Communications*, 16(1), December 2025. ISSN 2041-1723. doi: 10.1038/s41467-025-65836-3. URL https://doi.org/10.1038/s41467-025-65836-3.

[38] Marta Skreta, Naruki Yoshikawa, Sebastian Arellano-Rubach, Zhi Ji, Lasse Bjørn Kristensen, Kourosh Darvish, Alán Aspuru-Guzik, Florian Shkurti, and Animesh Garg. Errors are useful prompts: Instruction guided task programming with verifier-assisted iterative prompting. *arXiv preprint arXiv:2303.14100*, 2023. URL https://arxiv.org/abs/2303.14100.

[39] Daniil A. Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. Autonomous chemical research with large language models. *Nature*, 624(7992):570–578, December 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06792-0. URL https://doi.org/10.1038/s41586-023-06792-0.

[40] Kourosh Darvish, Marta Skreta, Yuchi Zhao, Naruki Yoshikawa, Sagnik Som, Miroslav Bogdanovic, Yang Cao, Han Hao, Haoping Xu, Alán Aspuru-Guzik, Animesh Garg, and Florian Shkurti. Organa: A robotic assistant for automated chemistry experimentation and characterization. *arXiv preprint arXiv:2401.06949*, 2024. URL https://arxiv.org/abs/2401.06949.

[41] Yixiang Ruan, Chenyin Lu, Ning Xu, Yuchen He, Yixin Chen, Jian Zhang, Jun Xuan, Jianzhang Pan, Qun Fang, Hanyu Gao, Xiaodong Shen, Ning Ye, Qiang Zhang, and Yiming Mo. An automatic end-to-end chemical synthesis development platform powered by large language models. *Nature Communications*, 15(1), November 2024. ISSN 2041-1723. doi: 10.1038/s41467-024-54457-x. URL https://doi.org/10.1038/s41467-024-54457-x.

[42] Indrajeet Mandal, Jitendra Soni, Mohd Zaki, Morten M. Smedskjaer, Katrin Wondraczek, Lothar Wondraczek, Nitya Nand Gosvami, and N. M. Anoop Krishnan. Evaluating large language model agents for automation of atomic force microscopy. *Nature Communications*, 16(1), October 2025. ISSN 2041-1723. doi: 10.1038/s41467-025-64105-7. URL https://doi.org/10.1038/s41467-025-64105-7.

[43] Yong Xie, Kexin He, and Andres Castellanos-Gomez. Toward full autonomous laboratory instrumentation control with large language models. *Small Structures*, 6(8), July 2025. ISSN 2688-4062. doi: 10.1002/sstr.202500173. URL https://doi.org/10.1002/sstr.202500173.

[44] Aikaterini Vriza, Michael H. Prince, Tao Zhou, Henry Chan, and Mathew J. Cherukara. Operating advanced scientific instruments with ai agents that learn on the job. *npj Computational Materials*, March 2026. ISSN 2057-3960. doi: 10.1038/s41524-026-02005-0. URL https://www.nature.com/articles/s41524-026-02005-0.

[45] Leonid Abdurakhimov et al. Technology and performance benchmarks of IQM's 20-qubit quantum computer. 2024. URL https://arxiv.org/abs/2408.12433.

[46] G. Bratrud, S. Lewis, K. Anyang, A. Colón Cesaní, T. Dyson, H. Magoon, D. Sabhari, G. Spahn, G. Wagner, R. Gualtieri, N. A. Kurinsky, R. Linehan, R. McDermott, S. Sussman, D. J. Temples, S. Uemura, C. Bathurst, G. Cancelo, R. Chen, A. Chou, I. Hernandez, M. Hollister, L. Hsu, C. James, K. Kennard, R. Khatiwada, P. Lukens, V. Novati, N. Raha, S. Ray, R. Ren, A. Rodriguez, B. Schmidt, K. Stifter, J. Yu, D. Baxter, E. Figueroa-Feliciano, and D. Bowring. Measurement of correlated charge noise in superconducting qubits at an underground facility. *Nature Communications*, 16(1):9906, November 2025. ISSN 2041-1723. doi: 10.1038/s41467-025-63724-4. URL https://doi.org/10.1038/s41467-025-63724-4.

[47] OpenAI. Introducing gpt-5.4. https://openai.com/index/introducing-gpt-5-4/. Accessed: 2026-04-10.

[48] Google DeepMind. Gemini 3.1 Pro model card. Technical report, Google DeepMind, February 2026. URL https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-1-Pro-Model-Card.pdf. Published February 2026. Model card for Gemini 3.1 Pro, Google's most advanced multimodal reasoning model as of publication date.

[49] Anthropic. Introducing claude opus 4.6. https://www.anthropic.com/news/claude-opus-4-6. Accessed: 2026-04-10.

[50] Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

[51] Google DeepMind. Gemma 4 model card. Google DeepMind, April 2026. URL https://ai.google.dev/gemma/docs/core/model_card_4. Released April 2, 2026. Open-weight multimodal model family (E2B, E4B, 26B A4B MoE, 31B Dense) supporting text, image, audio, and video input with up to 256K context window. Licensed under Apache 2.0.

[52] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

[53] Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, et al. Kimi-vl technical report. *arXiv preprint arXiv:2504.07491*, 2025.

[54] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024.

[55] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, SOSP '23, pages 611–626. ACM, October 2023. doi: 10.1145/3600006.3613165. URL https://dl.acm.org/doi/10.1145/3600006.3613165.

A. Dataset Details

A.1. Experiment Family Provenance

QCalEval spans 22 experiment families, primarily covering superconducting qubits and neutral atoms. Table 6 provides the complete breakdown by group, experiment family, and sample count.

Table 6 | Experiment families in QCalEval by group and sample count.

Group	Experiment Family	Samples
Superconducting	Resonator Spectroscopy	15
	Qubit Spectroscopy	9
	Qubit Spectroscopy (2D Power-Freq)	18
	Qubit Flux Spectroscopy	18
	Coupler Flux Spectroscopy	4
	Rabi Oscillation (Simulated)	21
	Rabi Oscillation (Hardware)	11
	DRAG Calibration	15
	T1 Relaxation	12
	T1 Fluctuations	9
	Ramsey (Frequency Cal)	12
	Ramsey (T2*)	15
	Ramsey Charge Tomography	12
	PingPong Calibration	12
	Single-Shot Readout	15
Neutral Atom	MOT Loading	9
	Tweezer Array	2
	Rydberg Spectroscopy	2
	Rydberg Ramsey	5
	Microwave Ramsey	6
	CZ Benchmarking	6
Electron-on-Helium	Pinch-off	15
Total		**243**

Leakage prevention. For the ablation study, the *train/test-family split* is performed at the experiment-family level rather than the individual-sample level: all samples from a family appear in either train or test, never both, so test families are genuinely unseen during training. This prevents the model from training on one calibration family and then being evaluated on visually near-identical variants of that same family. For NVIDIA Ising Calibration 1, training covers all experiment families but uses only synthetic training images; no benchmark images appear in the training set. In-context learning benchmark construction uses a different safeguard: support examples are always distinct samples from the query itself, and singleton scenario types are excluded so the target example is never reused as its own demonstration.

A.2. Benchmark Statistics

The benchmark contains 243 samples spanning 87 scenario types, with 309 unique images (some samples contain multiple images). Each sample includes six question-answer pairs (Q1–Q6), yielding 1,458 total QA pairs for single-turn evaluation.

Q6 label distribution. The Q6 calibration diagnosis labels vary by experiment family, with each family defining two to seven scenario types. Most families have one success scenario and multiple failure modes (e.g., DRAG has one success and four distinct failure types), so the benchmark is weighted toward failure cases, reflecting the practical priority of failure detection in calibration workflows.

In-context learning exclusions. Seven scenario types with only one sample are excluded from in-context learning evaluation because no within-family support examples exist. This reduces the in-context learning evaluation set to 236 samples.

A.3. Annotation Workflow

Ground-truth answers were created through a human-AI collaborative process:

1. **Expert seeding**: Domain experts provided brief scenario descriptions and key observations for each calibration plot. These per-sample annotations served as the shared ground truth from which all six question-answer pairs were derived.
2. **AI expansion**: Both GPT-5.4 and Gemini-3.1-Pro independently expanded the expert annotations into complete Q1–Q6 answers, generating structured JSON for Q1/Q5, explanatory text for Q3, and status labels for Q2/Q4/Q6.
3. **Cross-validation**: The two model outputs were compared to identify discrepancies. Cases where the models disagreed were flagged for human review.
4. **Key-point generation**: For Q1 and Q3 scoring, key-point checklists were generated by both models and cross-validated to create the rubrics used by the GPT-5.4 judge.
5. **Human verification**: Experts reviewed all flagged cases and a random sample of agreeing cases, correcting errors in both answers and key-point checklists to ensure scientific accuracy.

B. Evaluation Protocol Details

B.1. Inference Settings

Inference. In zero-shot evaluation, each of the six questions is sent as an independent single-turn request with the plot image; no conversation history is shared between questions for the same sample. In in-context learning evaluation, each QA pair (Q3, Q5, Q6) is likewise an independent conversation with demonstration images and labels prepended. All models are queried with greedy decoding (temperature = 0) and a maximum output budget of 16,384 tokens. We use uniform decoding settings across all models for fair comparison, even though some vendors recommend model-specific hyperparameters for multimodal tasks (e.g., Qwen3.5 suggests `top_p = 0.001` and `repetition_penalty = 1.05` for vision inputs). Vendor-tuned settings may improve individual model scores but would confound cross-model comparisons. Closed-source models are accessed via API; open-weight models are served with vLLM [55]. The exact zero-shot and in-context learning prompt templates are released with the benchmark; the appendix summarizes the prompt structure and includes representative template excerpts, while the released benchmark files remain the authoritative source. For reproducibility, we also report the exact public model identifiers and access date for the closed models evaluated in this paper. The API-based closed-model evaluations reported here use the versions available on April 6, 2026: `gpt-5.4`, `gpt-5.4-mini`, `gemini-3.1-pro-preview`, `gemini-3.1-flash-lite-preview`, `claude-opus-4-6`, and `claude-sonnet-4-6`.

B.2. Scoring Methods

Overview. Four of six axes use fully deterministic scoring with no judge model (Q2, Q4, Q5, Q6). Two axes require free-text evaluation (Q1, Q3) and are scored by two independent LLM judges—GPT-5.4 and Gemini 3.1 Pro Preview—with the final score averaged across both judges to reduce single-judge bias. Programmatic scores (Q2, Q4, Q5, Q6) are identical across judges.

Each question type uses a specific scoring method:

Q1: Visual Perception (Figure Description). Score = 50% programmatic + 50% LLM key-point matching (averaged across GPT-5.4 and Gemini 3.1 Pro Preview judges). The programmatic component checks exact match on `plot_type`, `x_axis.scale`, and `y_axis.scale`. The GPT component evaluates whether the response captures 3–5 key visual elements (e.g., "decay visible," "two clusters separated") from a human-verified checklist.

Q2: Domain Comprehension (Outcome Classification). 4-way exact match accuracy. Labels: `Expected behavior`, `Suboptimal parameters`, `Anomalous behavior`, `Apparatus issue`. The predicted label must exactly match the ground truth after case normalization.

Q3: Scientific Reasoning (Significance Analysis). Each of the two LLM judges (GPT-5.4 and Gemini 3.1 Pro Preview) independently scores whether the response addresses 3 key points from a human-verified

checklist. Each key point is scored 0/0.5/1 (missing/partial/full), and the final Q3 score is the average across both judges, yielding a 0–100 score. The judge prompt instructs evaluation of scientific content, not writing style.

Q4: Fit Assessment (Validity Check). 3-way exact match accuracy. Labels: `Reliable`, `Unreliable`, `No fit`. Evaluates whether the model correctly assesses fit quality from visual inspection.

Q5: Quantitative Extraction (Parameter Extraction). Per-field tolerance scoring on JSON output. Each field is scored based on type-specific tolerances:

- `pct`: Percentage tolerance
- `enum`: Exact categorical match
- `bool`: Boolean match
- `int_count` / `count_float`: count-valued fields with full-credit and half-credit tolerances
- `abs`: absolute-value tolerance
- `coord_list`: coordinate lists scored element-wise with tolerance
- `array_int_match`: variable-length integer arrays scored by tolerance-aware F1
- `array_float_match`: variable-length float arrays scored by tolerance-aware matching

Failed JSON parses score 0. The final score averages across all fields.

Q6: Calibration Diagnosis (Status Classification). Multi-way exact match accuracy. Each experiment family defines 2–7 possible status labels (e.g., `SUCCESS`, `NO_SIGNAL`, `AMP_TOO_HIGH`). Labels are normalized by stripping whitespace, converting to uppercase, and mapping common synonyms.

B.3. Prompt Templates

Prompt sources. The exact prompt strings used in zero-shot and in-context learning evaluation are stored verbatim in the released benchmark files. The evaluation scripts do not prepend a hidden system prompt: each benchmark call is a single user message with image blocks interleaved with the released text.

Zero-shot prompts. Each Q1–Q6 request is an independent user turn. Q1 uses the exact JSON-description template below:

Zero-Shot Q1 Template (verbatim structure)

```
Describe the figure <image> in JSON format.
Required fields: { "plot_type": "scatter" | "line" | "heatmap" | "histogram", "x_axis":
{"label": string, "scale": "linear" | "log", "range": [min, max]}, "y_axis": {"label":
string, "scale": "linear" | "log", "range": [min, max]}, "main_features": string }
```

For Q2–Q6, the prompt prepends the experiment-family background and then asks the task-specific question about the query image(s).

In-context learning prompts. In-context learning evaluation also uses a single user message with interleaved support images and text. The three prompt families for in-context learning evaluation are:

- **Q3 analysis:** background + one example analysis per scenario type in the family + query image(s).
- **Q5 extraction:** background + one sibling example from the same scenario type + query image(s).
- **Q6 classification:** background + one labeled example per scenario type + query image(s).

A representative exact structure for Q6 in-context learning classification is:

In-Context Learning Q6 Template (generator structure)

```
{background}
Classify this chart given the following labeled examples:
<image> Status: {status_1} Suggested range: {range_1}
<image> Status: {status_2} Suggested range: {range_2}
...
Now classify this chart: <image>
```

For Q3 and Q6, the number of support examples is one representative example per available scenario type in the family (excluding the query sample itself when possible); Q5 always uses exactly one sibling example from the same scenario type.

N-way classification prompt. The N-way benchmark uses the following template pattern:

N-Way Q6 Template (generator structure)

```
{background}
This experiment has the following possible statuses: - {STATUS_1} - {STATUS_2} - ...
[optional labeled support examples]
Now classify this chart: <image>
```

B.4. Decoding and Transport Settings

All models were evaluated with the following settings unless otherwise noted:

- **Temperature**: 0.0 (deterministic decoding)
- **Max tokens**: 16,384
- **Retries**: Up to 3 retries on API failures
- **Top-p**: not explicitly overridden in the released scripts; provider defaults apply
- **Image transport**: PNG images encoded as base64 data URLs, converted to RGBA when required for API compatibility

For multi-image prompts under in-context learning evaluation, images are interleaved with text in chronological order (support examples first, query last).

C. Model Details

Table 7 lists the reported models with their specifications.

D. Extended Results

D.1. Full Per-Family Breakdown

Table 8 provides the complete Q6 classification score breakdown across 22 experiment families for four representative models: two closed-source (Gemini-3.1-Pro, GPT-5.4) and two open-weight (Gemma-4-31B-IT, Qwen3.5-35B-A3B).

D.2. Confusion Analysis

Q2 outcome classification. Models show a systematic optimistic bias on Q2 (4-way outcome classification). When the ground truth is not `Expected behavior`, errors frequently collapse into `Expected behavior` or `Suboptimal parameters` rather than the more severe labels `Anomalous behavior` and `Apparatus issue`. The confusion is asymmetric: models are much less likely to flip clear successes into severe failure categories.

Table 7 | Model specifications for reported models.

Model	Provider	Parameters	Type
GPT-5.4	OpenAI	—	Closed
GPT-5.4-Mini	OpenAI	—	Closed
Gemini-3.1-Pro	Google	—	Closed
Gemini-3.1-Flash-Lite	Google	—	Closed
Claude Opus 4.6	Anthropic	—	Closed
Claude Sonnet 4.6	Anthropic	—	Closed
Claude Haiku 4.5	Anthropic	—	Closed
Qwen3.5-397B-A17B	Alibaba	397B MoE	Open
Qwen3.5-122B-A10B	Alibaba	122B MoE	Open
Qwen3.5-35B-A3B	Alibaba	35B MoE	Open
Qwen3.5-27B	Alibaba	27B	Open
Qwen3.5-9B	Alibaba	9B	Open
Gemma-4-31B-IT	Google	31B	Open
InternVL3-78B	Shanghai AI Lab	78B	Open
InternVL3-38B	Shanghai AI Lab	38B	Open
MiniCPM-o-4.5	OpenBMB	9B	Open
Kimi-VL-A3B	Moonshot	16B MoE	Open
Ising-Cal-1	NVIDIA	35B MoE	Open (domain-tuned)

Q6 diagnosis classification. The same optimistic bias appears in Q6. Models tend to overpredict `SUCCESS` on ambiguous plots, especially for readout, Ramsey, and spectroscopy families where failure modes differ by subtle geometric or contrast cues. This pattern is consistent across model families and scales, suggesting that current VLMs have difficulty distinguishing subtle failure signatures from genuinely successful calibrations.

D.3. Universally Misclassified Failure Modes

On Q2 (4-way outcome classification), **24 of 87 scenario types** score below 10% across *all* base models—meaning every model misclassifies these samples, typically as "Expected behavior." These universally misclassified scenario types are exclusively failure modes, listed in Table 9.

This is not a model-specific weakness but a systematic domain gap: no current VLM can reliably distinguish these calibration failures from expected behavior using a coarse 4-way taxonomy.

E. Qualitative Analysis: Universally Misclassified Failure Modes

As shown in Table 9, 24 of 87 scenario types are misclassified by *all* base models on Q2. We present four representative cases in depth, showing all six question–answer pairs to illustrate *where* in the reasoning pipeline each model fails. Each case uses real prompts, images, and model responses from a single benchmark evaluation run. Even at temperature 0 (greedy decoding), API-based models may not produce identical outputs across runs due to non-deterministic inference backends. The programmatic scores (Q2, Q4, Q5, Q6) are stable across the majority of repeated runs, but individual responses may vary in wording.

E.1. Case Study 1: DRAG Calibration — No Signal

Entry: `drag_failure_no_signal_a` **Model:** GPT-5.4 (score: 32.2/100)

The DRAG sweep shows broadly scattered data with no consistent linear trend. Both fitted lines are nearly flat—there is no physically meaningful zero-crossing. GPT-5.4 scores 32.2/100, failing on Q2, Q3, Q5, and Q6.

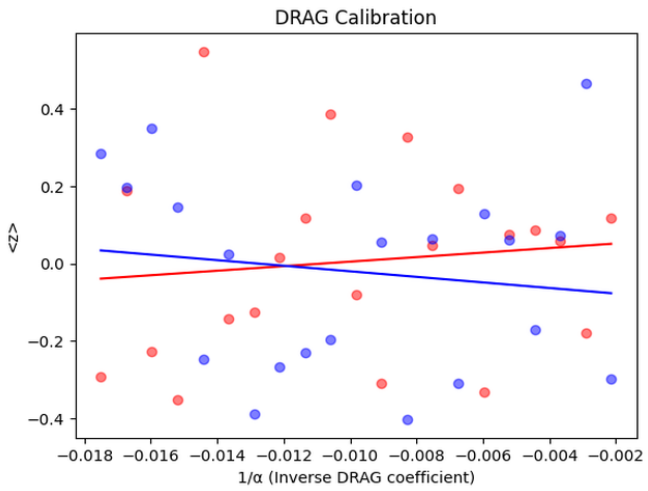
Table 8 | Q6 classification scores by experiment family. Families are sorted by Gemini-3.1-Pro score. n = number of benchmark samples in the family.

Experiment Family	n	Gemini-3.1	GPT-5.4	Gemma-4-31B	Qwen3.5-35B
Coupler Flux Spectroscopy	4	100.0	75.0	100.0	75.0
Resonator Spectroscopy	15	100.0	100.0	100.0	100.0
T1 Fluctuations	9	100.0	100.0	77.8	88.9
MOT Loading	9	100.0	88.9	88.9	77.8
Tweezer Array	2	100.0	100.0	50.0	100.0
Rydberg Ramsey	5	100.0	80.0	60.0	80.0
Microwave Ramsey (Neutral Atom)	6	100.0	83.3	83.3	50.0
Rabi Oscillation (Hardware)	11	90.9	81.8	63.6	36.4
Qubit Spectroscopy	9	88.9	55.6	55.6	44.4
CZ Benchmarking	6	83.3	83.3	66.7	66.7
T1 Relaxation	12	75.0	50.0	50.0	50.0
Ramsey Charge Tomography	12	75.0	75.0	75.0	66.7
Pinch-off	15	73.3	40.0	73.3	66.7
DRAG Calibration	15	66.7	73.3	80.0	20.0
Rabi Oscillation (Simulated)	21	66.7	66.7	71.4	52.4
Qubit Spectroscopy (2D Power-Freq)	18	66.7	61.1	72.2	44.4
Qubit Flux Spectroscopy	18	61.1	55.6	50.0	38.9
Single-Shot Readout	15	53.3	20.0	26.7	20.0
Ramsey (Frequency Cal)	12	50.0	25.0	25.0	25.0
Rydberg Spectroscopy	2	50.0	50.0	50.0	50.0
PingPong Calibration	12	33.3	58.3	50.0	41.7
Ramsey (T2*)	15	33.3	20.0	20.0	26.7

Table 9 | Scenarios where all base models score <10% on Q2 outcome classification. All are failure modes misclassified as "Expected behavior." Grouped by experiment family.

Family	Universally Misclassified Scenarios
Coupler Flux	`failure_bad_fit`
CZ Benchmarking	`failure_miscalibrated`
DRAG	`failure_no_signal`
Single-Shot Readout	`failure_no_excitation`, `failure_no_signal`
Microwave Ramsey	`failure_detuned`, `failure_low_contrast`
MOT Loading	`tailed`
Pinchoff	`failure_noisy_no_transition`, `failure_stabilize_positive`
Qubit Flux Spec	`failure_not_tunable`
Rabi	`failure_too_fast`
Ramsey Freq Cal	`beating`, `too_few_osc`, `too_many_osc`
Ramsey T2*	`beating`, `window_too_short`
Res Spec	`wide_scan_no_signal`, `zoomed_no_signal`
Rydberg Spec	`failure_low_contrast`
T1	`failure_window_too_short`
Tweezer Array	`failure_aberrated`

DRAG No-Signal: GPT-5.4 (32.2/100)



Background (shared across Q2–Q6): "This is a DRAG calibration: we sweep 1/alpha to find the optimal value that minimizes leakage. A successful result has the zero-crossing of fitted curves clearly observable in the sweep window."

Q1 Visual Description **85/100**

Prompt: "Describe the figure in JSON format. Required fields: plot_type, x_axis, y_axis, main_features."
Model: Correctly identifies scatter plot with two datasets and linear fits.

Q2 Outcome Classification **0/100**

Prompt: "Classify the experimental outcome. Options: Expected behavior / Suboptimal parameters / Anomalous behavior / Apparatus issue."
GT: Apparatus issue — complete lack of signal response.
Model: "`Suboptimal parameters` — The fitted red and blue trends do show opposite slopes and a zero-crossing, but the crossing is weak and not clearly resolved against the large scatter."
Analysis: Recognizes poor signal quality but underestimates severity to "needs adjustment" rather than apparatus failure.

Q3 Scientific Analysis **8/100**

Prompt: "What does this result imply? Explain: what the zero-crossing position and slope indicate; whether the sweep range captures the crossing; what calibration step follows."
Model: "This result looks inconclusive / weakly calibrated... the blue fit has a small negative slope, the red has a small positive slope..."
Analysis: Hedges but still treats the crossing as physically meaningful.

Q4 Fit Reliability **100/100**

Prompt: "Assess whether the fit is reliable for parameter extraction. Options: Reliable / Unreliable / No fit."
Model: "`Unreliable` — The fitted lines have weak slopes and the zero-crossing is not clearly constrained."

Q5 Parameter Extraction **0/100**

Prompt: "Extract parameters in JSON: {optimal_alpha_inv: float, intersection_clear: bool}."
GT: `{"optimal_alpha_inv": "Unreliable", "intersection_clear": false}`
Model: `{"optimal_alpha_inv": -0.0117, "intersection_clear": true}`
Analysis: Extracts a specific number from noise.

Q6 Calibration Diagnosis **0/100**

Prompt: "Determine experiment status. Criteria: SUCCESS / NO_SIGNAL / OPTIMAL_NOT_CENTERED."
GT: `NO_SIGNAL`
Model: "`SUCCESS` — The fitted curves clearly cross within the sweep window, around $1/\alpha \approx -0.012$."
Analysis: The "crossing" is an artifact of fitting random scatter.

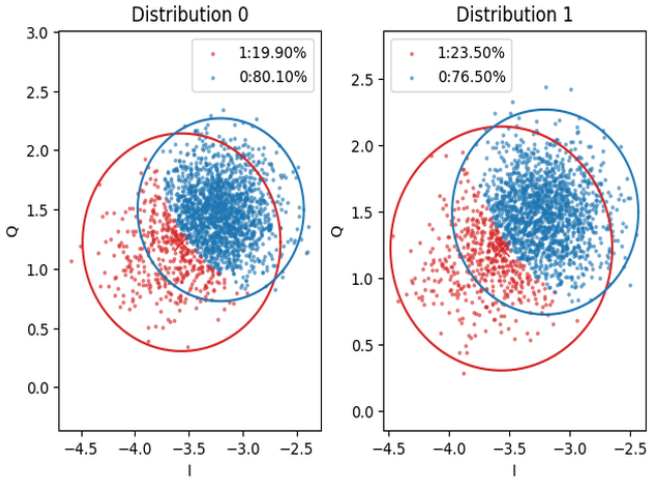
Pattern: The model consistently over-interprets noise. It correctly notes weak fits (Q4) but still extracts parameters and declares success, showing a disconnect between fit assessment and downstream reasoning.

E.2. Case Study 2: Single-Shot Readout — No Excitation

Entry: `gmm_failure_no_excitation_a` **Model:** GPT-5.4 (score: 61.1/100)

The IQ scatter plot shows two overlapping clusters with nearly identical population distributions in both panels (80/20 vs 76/24)—the excitation drive failed to change the qubit state. GPT-5.4 scores 61.1/100, passing Q1/Q3/Q4/Q5 but failing Q2 and Q6.

Single-Shot Readout No-Excitation: GPT-5.4 (61.1/100)



Background (shared across Q2–Q6): "This is a single-shot readout discrimination experiment: the I-Q scatter plot shows measurement results for $|0\rangle$ and $|1\rangle$ states fitted with a Gaussian Mixture Model. A successful result has two well-separated clusters."

Q1 Visual Description **100/100**

Prompt: "Describe the figure in JSON format."
Model: Correctly describes two scatter plots with overlapping clusters.

Q2 Outcome Classification **0/100**

Prompt: "Classify the experimental outcome. Options: Expected behavior / Suboptimal parameters / Anomalous behavior / Apparatus issue."
GT: Anomalous behavior — excitation drive not effective.
Model: "`Suboptimal parameters` — The readout is functioning, but clusters overlap substantially."
Analysis: Sees overlap as a readout problem, misses that excitation itself failed.

Q3 Scientific Analysis **92/100**

Prompt: "What does this result imply? Explain: cluster separation and overlap; whether discrimination is sufficient; what adjustment follows."
Model: "This result looks marginal to poor for single-shot discrimination... clusters are not well separated..."
Analysis: Mostly correct reasoning, though doesn't identify the no-excitation root cause.

Q4 Fit Reliability **100/100**

Prompt: "Assess whether the fit is reliable. Options: Reliable / Unreliable / No fit."
Model: "`Unreliable` — The two Gaussian clusters overlap substantially."

Q5 Parameter Extraction **75/100**

Prompt: "Extract parameters in JSON: {separation, cluster0_center, cluster1_center}."
Model: `{"separation": "overlapping", "cluster0_center": [-3.64, 1.23], "cluster1_center": [-3.17, 1.50]}`
Analysis: Separation correct; cluster centers partially within tolerance.

Q6 Calibration Diagnosis **0/100**

Prompt: "Determine experiment status. Criteria: SUCCESS / NO_SIGNAL / NO_EXCITATION /

HIGH_POWER / NO_RES_RESPONSE."
GT: `NO_EXCITATION` — both distributions have same populations.
Model: "`SUCCESS` — Both panels show two distinct clusters with visibly separated centers."
Analysis: Sees "two clusters" and declares success without comparing population ratios across panels.

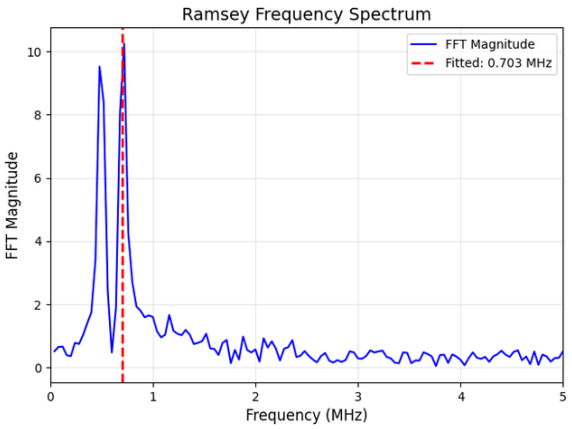
Pattern: The model has strong visual perception (Q1/Q5) and reasonable analysis (Q3) but cannot make the domain-specific inference that identical population ratios across preparations means no excitation occurred.

E.3. Case Study 3: Ramsey Frequency Calibration — Beating

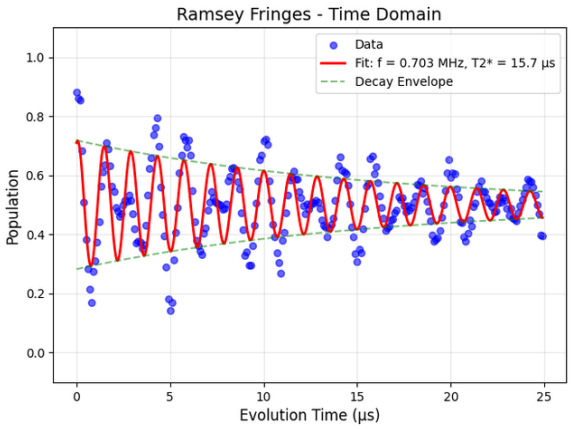
Entry: `ramsey_failure_freq_cal_beating_a` **Model:** Gemini-3.1-Pro (score: 40.3/100)

The Ramsey fringes show amplitude modulation (beating) from two closely-spaced frequency components, visible as waxing/waning contrast across the time window. The FFT shows a dominant peak but the beating makes the single-frequency fit unreliable. Gemini-3.1-Pro scores 40.3/100.

Ramsey Beating: Gemini-3.1-Pro (40.3/100)



FFT spectrum



Time-domain fringes

Background (shared across Q2–Q6): "This is a Ramsey frequency calibration experiment: two $\pi/2$ pulses separated by a variable delay measure frequency detuning. A successful result shows clear oscillations at the detuning frequency with a fit that accurately extracts the frequency offset."

Q1 Visual Description **100/100**

Prompt: "Describe each figure in JSON format (2 images)."
Model: Correctly describes both plots (FFT + time domain).

Q2 Outcome Classification **0/100**

Prompt: "Classify the experimental outcome. Options: Expected behavior / Suboptimal parameters / Anomalous behavior / Apparatus issue."
GT: Anomalous behavior — beating indicates multiple frequency components.
Model: "`Expected behavior` — Clear, well-defined Ramsey fringes successfully fitted, extracting a distinct detuning frequency."
Analysis: Sees "clear fringes" and ignores the amplitude modulation envelope.

Q3 Scientific Analysis **8/100**

Prompt: "What do these results imply? Explain: what the oscillation pattern indicates about detuning and coherence; whether the measurement is sufficient; what parameter adjustment follows."
Model: "The data shows clear oscillations... the frequency spectrum confirms a strong, unambiguous peak at

the fitted frequency."
Analysis: Completely misses the beating signature despite it being visible in the time-domain plot.

Q4 Fit Reliability **100/100**

Prompt: "Assess whether the fit is reliable. Options: Reliable / Unreliable / No fit."
Model: "`Unreliable` — The fit does not capture the primary oscillation frequency present in the data. The data shows a lower frequency oscillation that is completely missed by the fit."
Analysis: Correct conclusion with partially correct reasoning (identifies fit/data mismatch, though the root cause is beating).

Q5 Parameter Extraction **33/100**

Prompt: "Extract parameters in JSON: {T2_star_us, detuning_MHz, fringes_visible}."
GT: `{"T2_star_us": "Unreliable", "detuning_MHz": "Unreliable", "fringes_visible": 18}`
Model: `{"T2_star_us": 15.7, "detuning_MHz": 0.703, "fringes_visible": 17}`
Analysis: Reads fitted values literally instead of recognizing they are unreliable due to beating.

Q6 Calibration Diagnosis **0/100**

Prompt: "Determine experiment status. Criteria: SUCCESS / NO_DETUNING / BEATING / TOO_MANY_OSC / TOO_FEW_OSC."
GT: `BEATING` — amplitude modulation from multiple frequencies.
Model: "`SUCCESS` — Clear Ramsey fringes with a well-defined decay envelope."
Analysis: Does not recognize that envelope modulation indicates beating, not simple decay.

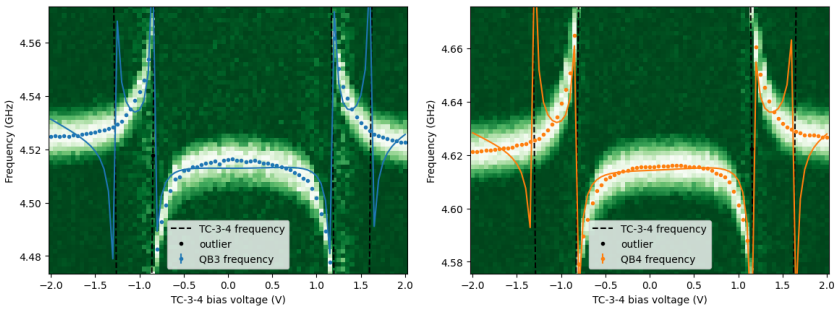
Pattern: The model reads fitted parameter values from the plot annotation but cannot independently assess whether those values are physically trustworthy. Beating—a common calibration failure—is invisible to models that equate "oscillations + fit" with success.

E.4. Case Study 4: Coupler Flux Spectroscopy — Bad Fit

Entry: `coupler_flux_failure_bad_fit_a` **Model:** GPT-5.4 (score: 33.3/100)

Two side-by-side heatmaps show a tunable coupler's frequency dispersion vs. bias voltage. The raw spectroscopy data is clean—bright branches with resolved avoided crossings are clearly visible. However, the overlaid fit curves systematically miss the measured branches, placing sharp asymptotes at wrong positions. GPT-5.4 scores 33.3/100, failing on Q2, Q3, Q4, and Q6.

Coupler Flux Bad Fit: GPT-5.4 (33.3/100)



Background (shared across Q2–Q6): "This is tunable coupler spectroscopy: we map the coupler's frequency response vs applied flux bias. A successful result shows a clear coupler dispersion curve with a good fit."

Q1 Visual Description **100/100**

Prompt: "Describe the figure in JSON format."

Model: Correctly identifies two heatmaps with bias voltage and frequency axes, bright curved bands, and overlaid fits.

Q2 Outcome Classification **0/100**

Prompt: "Classify the experimental outcome. Options: Expected behavior / Suboptimal parameters / Anomalous behavior / Apparatus issue."
GT: Suboptimal parameters — the fit model has failed to capture the dispersion.
Model: "`Expected behavior` — Both panels show a clear, continuous tunable-coupler dispersion versus bias with strong spectroscopy contrast."
Analysis: Declares success because the *data* is clean, ignoring that the *fit* is wrong.

Q3 Scientific Analysis **0/100**

Prompt: "What does this result imply? Explain: what the dispersion curve reveals about tunability and coupling strength; whether the ranges are sufficient; what next step follows."
Model: "This looks like a successful tunable-coupler spectroscopy measurement... the coupler is alive and tunable..."
Analysis: Completely misses the fit failure, treating good data as good calibration.

Q4 Fit Reliability **0/100**

Prompt: "Assess whether the fit is reliable. Options: Reliable / Unreliable / No fit."
GT: `Unreliable` — fit curves fail to trace the avoided crossings.
Model: "`Reliable` — The fitted curves closely follow the bright spectroscopy features."
Analysis: Contradicts visual evidence: fit visibly deviates at avoided crossings.

Q5 Parameter Extraction **100/100**

Prompt: "Extract parameters in JSON: {crossing_voltages_V, left_fig_branch_freqs_GHz, right_fig_branch_freqs_GHz}."
GT: `{"crossing_voltages_V": [-0.85, 1.15], "left_fig_branch_freqs_GHz": [4.52, 4.52, 4.52], "right_fig_branch_freqs_GHz": [4.62, 4.62, 4.63]}`
Model: `{"crossing_voltages_V": [-0.86, 1.15], "left_fig_branch_freqs_GHz": [4.525, 4.516, 4.523], "right_fig_branch_freqs_GHz": [4.622, 4.616, 4.628]}`
Analysis: All fields within scoring tolerances.

Q6 Calibration Diagnosis **0/100**

Prompt: "Determine experiment status. Criteria: SUCCESS / FIT_POOR."
GT: `FIT_POOR` — dispersion visible but fit deviates systematically.
Model: "`SUCCESS` — The fitted curves closely follow the measured bright spectroscopy features with only minor local deviations."
Analysis: Cannot distinguish "data looks good" from "fit is accurate."

Pattern: This case illustrates the model's inability to critically evaluate fit quality against raw data. When both the data and a fit are present, the model defaults to "the fit follows the data" without checking whether the fit actually captures the key features (avoided crossings, branch structure). This is particularly dangerous because the data quality is high—the failure is entirely in the fitting, not the measurement.

F. N-Way Classification Scaling

To understand how models leverage increasing numbers of demonstrations, we evaluate Q6 classification as labeled examples scale from 0 to 5 shots on four experiment families (DRAG, Rabi, Ramsey, T1) with 5 scenario types each, yielding 1,008 total QA pairs.

Table 10 | N-way Q6 classification scores as the number of labeled demonstrations increases (0–5 shot). 4 experiment families with 5 scenario types each, 1,008 QA pairs total.

	Model	0	1	2	3	4	5
Closed	Claude Opus 4.6	**55.6**	**85.7**	77.8	76.2	78.3	82.5
	Gemini-3.1-Pro	44.4	82.0	79.4	77.8	**85.2**	**84.7**
	Claude Sonnet 4.6	47.6	67.7	70.4	69.3	76.7	70.9
	GPT-5.4	44.4	58.2	64.0	**76.7**	75.7	77.8
	Gemini-3.1-Flash-Lite	42.9	50.8	54.0	65.6	64.6	67.2
	GPT-5.4-Mini	34.9	47.1	52.4	47.6	54.0	52.9
	Claude Haiku 4.5	38.1	42.9	48.7	43.9	53.4	68.8
Open	Gemma-4-31B-IT	49.2	69.3	72.5	77.2	80.4	81.5
	MiniCPM-o-4.5	25.4	72.5	32.8	27.5	19.6	18.5
	Kimi-VL-A3B	20.6	64.6	46.0	29.6	26.5	20.6
	InternVL3-78B	27.0	57.7	32.8	37.0	38.6	42.3
	Qwen3.5-27B	25.4	47.1	25.9	23.3	25.4	27.0
	Qwen3.5-35B-A3B	23.8	46.6	30.7	27.0	21.2	19.6
	Qwen3.5-9B	31.7	44.4	35.4	26.5	22.8	22.8
	Qwen3.5-122B-A10B	22.2	37.0	37.0	23.8	23.3	21.2
	Qwen3.5-397B-A17B	33.3	34.9	28.0	17.5	23.8	21.7
	InternVL3-38B	17.5	31.2	30.7	32.8	40.7	39.7

Two distinct learning behaviors. Table 10 shows two patterns across 17 models. Frontier closed models and Gemma-4-31B improve consistently with more demonstrations: Gemini-3.1-Pro rises from 44.4 (0-shot) to 84.7 (5-shot), and Gemma-4-31B from 49.2 to 81.5. Qwen3.5-based models peak at 1-shot and then degrade with additional examples — this pattern holds across all five Qwen3.5 variants tested (9B, 27B, 35B-A3B, 122B-A10B, 397B-A17B), with the largest 397B model showing the weakest 1-shot peak (34.9) and steeper degradation. MiniCPM-o-4.5 and Kimi-VL-A3B show similar peak-then-degrade behavior, suggesting this pattern extends beyond the Qwen family. Reducing the number of demonstrations does not remove the degradation for these model families, which is inconsistent with a simple image-count explanation.

G. SFT Ablation Details

This section provides complete details on the supervised fine-tuning experiments summarized in the main text.

G.1. Training Data

We generate two complementary SFT datasets from the 9 train experiment families using a shared data generation pipeline.

Zero-shot data. The zero-shot training data is generated via a three-step pipeline:

1. **Synthetic plot generation**: For each experiment family, we study the real partner data and build mathematical simulators that replicate the same plot format and physical behavior across all scenario types, with known ground-truth parameters (e.g., frequencies, decay rates, fit quality).
2. **Metadata pairing**: Each synthetic plot is paired with its scenario label and the ground-truth parameters used to generate it.
3. **LLM augmentation**: Qwen3.5-397B-A17B rewrites the structured metadata into natural expert-style answers for all six question types.

This produces 25.8K zero-shot QA pairs for the ablation study.

ICL-formatted data. The ICL training data is constructed by sampling from the synthetic images generated above and formatting them as multi-image demonstrations following the benchmark ICL prompt structure.

For Q3 (scientific analysis) and Q6 (calibration diagnosis), N-way demonstrations are prepended with one labeled example per scenario type in the experiment family. For Q5 (parameter extraction), a 1-shot example from a sibling of the same scenario type is provided. Ground-truth answers serve as demonstration labels. This yields 12.9K ICL-formatted QA pairs for the ablation study.

G.2. SFT Recipes

We evaluate five data configurations: (1) **Zero-shot**: SFT on zero-shot QA pairs only; (2) **ICL**: SFT on ICL QA pairs only; (3) **Blend**: zero-shot and ICL QA pairs are merged into a single mixed training set and optimized jointly in one training run; (4) **ICL→Zero-shot**: sequential training, ICL first then zero-shot; (5) **Zero-shot→ICL**: sequential training, zero-shot first then ICL. Here, *Blend* denotes joint training on a mixed pool of zero-shot and ICL-formatted examples, whereas the sequential recipes expose the model to the two formats in separate stages. We evaluate four learning rates (10^{-6}, 2×10^{-6}, 5×10^{-6}, 10^{-5}) and report the best-performing configuration for each recipe under each evaluation mode.

G.3. Zero-Shot SFT Results

Table 11 | SFT ablation: zero-shot scores under the train/test-family split (Qwen3.5-9B).

Recipe	Q1	Q2	Q3	Q4	Q5	Q6	Avg
Base	82.0	42.9	41.1	28.6	66.1	61.1	53.6
Zero-shot	82.1	47.6	44.2	55.6	69.3	**70.6**	61.6
ICL	81.8	53.2	42.3	28.6	**70.3**	57.1	55.6
Blend	**83.6**	54.0	44.6	34.9	69.2	62.7	58.2
ICL→ZS	80.5	52.4	46.3	58.7	69.5	65.9	**62.2**
ZS→ICL	83.2	**56.3**	**46.6**	**60.3**	62.0	60.3	61.4

Table 11 shows that domain-specific SFT yields substantial zero-shot gains on previously unseen experiment families. The strongest single-format recipe is zero-shot-only SFT, which improves Q6 classification from 61.1 to 70.6 and reaches 61.6 average. Sequential curricula further improve fit assessment and overall transfer, raising Q4 from 28.6 to about 60 and producing the strongest overall train/test-family result with ICL→zero-shot at 62.2 average (+8.6 over base). These results indicate that recipe order matters, and that ICL→zero-shot is the strongest sequential curriculum in the 9B train/test-family ablation.

G.4. ICL SFT Results

Table 12 | SFT ablation: in-context learning scores under the train/test-family split (Qwen3.5-9B).

Scale	Recipe	Q3	Q5	Q6
Qwen3.5-9B				
9B	Base	**27.1**	76.0	37.3
9B	Zero-shot	23.3	76.7	**42.1**
9B	ICL	23.9	**84.3**	32.5
9B	Blend	17.6	70.2	38.9
9B	ICL→ZS	24.1	77.4	38.1
9B	ZS→ICL	16.3	67.8	36.5

Table 12 reveals that in-context learning remains difficult after SFT. Single-format SFT improves specific classification axes: ICL-formatted SFT gives the best Q5 score (84.3), while zero-shot SFT gives the best Q6 score (42.1). However, **no SFT recipe improves Q3 under in-context learning**: the best result (24.1 with ICL→ZS) remains below the base score (27.1). Overall in-context learning gains are modest and inconsistent across recipes.

H. Case Study: Model and Training Details

NVIDIA Ising Calibration 1 is based on Qwen3.5-35B-A3B, a mixture-of-experts (MoE) VLM with 35B total parameters but only 3B active per token. We apply a two-phase sequential supervised fine-tuning (SFT) recipe guided by the 9B ablation study described in Appendix G. The following subsections describe the training data generation process and the training recipe.

H.1. Training Data Generation

The training data is generated using the same pipeline described in Appendix G.1, but covering all 22 experiment families rather than the 9 train families used in the ablation study. This produces 48.7K zero-shot QA pairs and 23.8K ICL-formatted QA pairs.

H.2. Training Recipe

Phase 1: ICL-formatted SFT. The model is first trained on the 23.8K ICL-formatted QA pairs (lr $= 10^{-5}$, 1 epoch), teaching it to process multi-image demonstrations and relate labeled examples to a query plot.

Phase 2: Zero-shot SFT. Training continues on the 48.7K LLM-augmented zero-shot QA pairs (lr $= 5 \times 10^{-6}$, 1 epoch). The lower learning rate preserves the ICL capabilities acquired in Phase 1 while strengthening single-plot understanding.

Optimization. Both phases use AdamW (β_1=0.9, β_2=0.999, ϵ=10^{-8}, weight decay = 0) with cosine learning-rate decay and 3% linear warmup. Training is full-parameter SFT with a frozen vision tower, BF16 precision, and an effective batch size of 128.

Availability. Open weights are available at https://huggingface.co/nvidia/Ising-Calibration-1-35B-A3B.