OPTOFIG: CAN VISION-LANGUAGE MODELS READ OPTOELECTRONICS?

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision–language models (VLMs) offer new opportunities for automating scientific research, including the design of optoelectronic devices. Realizing this potential requires understanding the figures through which such devices are designed and evaluated, from circuit schematics and device structures to the related performance, such as spectra, mode profiles and timing diagrams. However, existing benchmarks do not capture this capability. We introduce OPTOFIG, a benchmark of 1,092 questions over 799 figures of optoelectronic devices, covering both how the devices are built and their corresponding performance characterization. OPTOFIG is simulation-grounded and rigorously validated: every answer is measured on the figure itself, read from the simulator or drawing-script record, or written by a domain expert, and passes automated checks and multi-stage expert review. Its 39 task types span three capability categories: perception and recognition, quantitative and comparative, and physical and functional reasoning, with difficulty defined by construction parameters. Experimental results across seven state-of-the-art VLMs reveal substantial room for improvement: overall accuracy ranges from 54.2% to 74.5%, no single model leads across capabilities, and average accuracy is 10.2 points lower on hard items than on easy ones. The experiments further reveal two findings. First, models perceive figures with limited precision: they capture coarse visual patterns but miss the exact quantities a question depends on. Second, models lack the reasoning needed to interpret what they perceive: on the physics questions, even when a model reads a figure correctly, it draws the wrong physical conclusion about one time in three (67.3% correct on average). VLMs' limitations on these device figures thus lie both in the precision of perception and in reasoning from what they see to the underlying physics.

1 INTRODUCTION

Driven by advances in large language models (LLMs), AI systems have begun to take part in scientific research, such as the design and analysis of optoelectronic devices. For example, recent systems turn natural-language design goals into photonic design scripts, connect language models to inverse-design solvers, and autonomously refine device geometries through repeated simulation (Jiang et al., 2026; Huang et al., 2025; Kharel et al., 2026). In parallel, vision–language models (VLMs) now extract scientific information from figures rather than text or tables (Roy et al., 2026). Together, these trends point toward agents that not only design and simulate devices but also interpret the results.

To do so, general research agents must understand figures, which are central to how researchers design and evaluate optoelectronic devices. Design figures show how a device is built: device structures, component symbols, and schematics. Characterization figures show how it performs: simulated mode profiles and field maps reveal the spatial distribution of the optical field; spectra plot how much light a device transmits or reflects at each wavelength, with resonances as sharp peaks or dips; and timing diagrams plot digital signals switching between high and low over time. However, current AI agents for photonic design use simulators mainly through code, numbers, and predefined evaluation criteria (Jiang et al., 2026; Huang et al., 2025; Kharel et al., 2026). This works when one knows in advance what to measure, but it does not replace how researchers read figures, for example recognizing a mode by its shape or noticing an unexpected feature in a spectrum. We therefore ask a complementary question: can today's VLMs reliably interpret these device figures and reason from what they show?

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

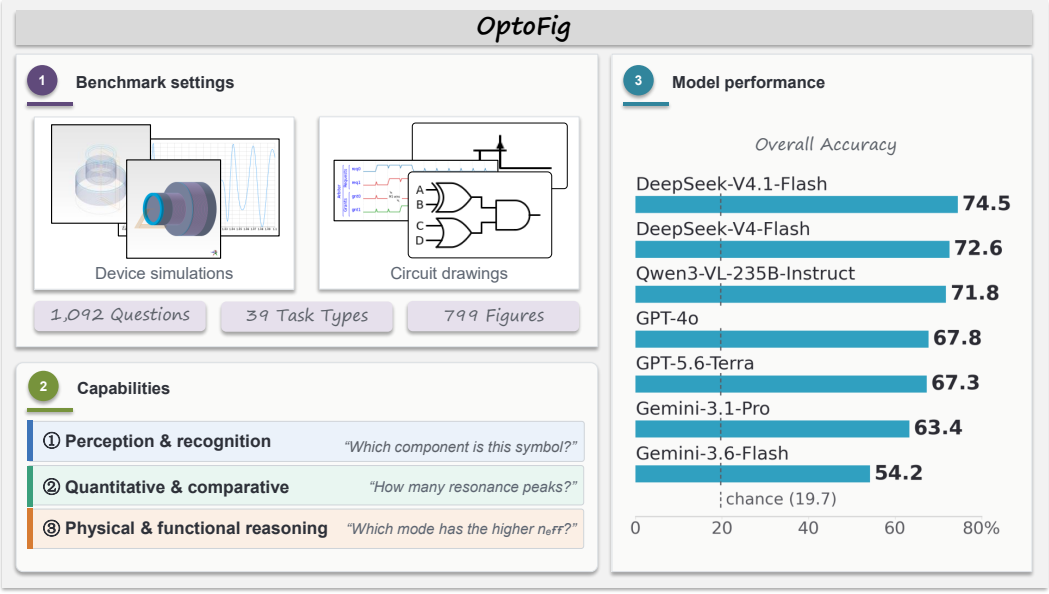


Figure 1: Overview of OPTOFIG. The benchmark covers figures of optoelectronic device design and characterization, organized into three capability categories and construction-defined difficulty levels. The right panel shows each model's overall accuracy with minimal or no reasoning; the dashed line is the guessing floor (random-guess accuracy).

Existing benchmarks test scientific figure understanding from several directions. First, chart and scientific-figure benchmarks test whether models can extract, interpret, and reason about information in charts and paper figures (Wang et al., 2024; Roberts et al., 2024; Pramanick et al., 2024). Second, physics and engineering benchmarks pair technical diagrams with problems that require domain-specific reasoning (Xiang et al., 2025; Li et al., 2025), while circuit benchmarks test how models read, analyze, and design schematics (Akbari et al., 2026; Shi et al., 2025). More recently, QCalEval moves closer to real research by testing VLMs on plots from quantum-calibration experiments (Cao et al., 2026). However, to our knowledge, no benchmark evaluates VLMs on the figures of optoelectronic devices, covering both their design and the characterization of their performance, and none tests the physical reasoning that links the two. Prior work also shows that performance can differ sharply across capabilities: strong visual perception does not necessarily lead to correct domain-specific reasoning (Akbari et al., 2026). This motivates an evaluation of optoelectronic figures that breaks down performance by task type and controlled difficulty level, with answers traceable to the process that produced each figure.

To address these gaps, we introduce OPTOFIG, a benchmark of 1,092 questions over 799 figures that show how optoelectronic devices are built and how they perform (Figure 1). Design figures show how a device is built, from optical structures and cross-sections to component symbols, circuit schematics, logic gates (e.g., AND or OR), and integrated-circuit packages. Characterization figures show how light and electrical signals behave in these devices: simulated mode profiles, field maps, 1D field profiles, resonance spectra, and timing diagrams. The 39 task types span three capability categories: perception and recognition, quantitative and comparative, and physical and functional reasoning.

A central challenge in building such a benchmark is obtaining ground truth that is both physically meaningful and reliable. Therefore, OPTOFIG is simulation-grounded: for 572 questions we compute the answer from the figure or the file that produced it; the remaining answers are source labels or come from a domain expert; and every item passes automated checks and expert review. Similarly, we set difficulty by construction, without using any evaluated model, through a parameter such as the gap between compared values, the count, or clutter. We then test per task type whether this parameter controls accuracy.

Evaluating seven VLMs with minimal or no reasoning reveals substantial room for improvement. Overall accuracy ranges from 54.2% to 74.5%, no single model leads across capabilities, and average accuracy is 10.2 points lower on hard items than on easy ones. Two findings stand out. First, models perceive figures with limited precision: they capture coarse visual patterns but miss the exact quantities a question depends on, so their counts and readings are often slightly off. Second, models lack the reasoning needed to interpret what they perceive: on the physics questions, even when a model reads a figure correctly, it draws the wrong physical conclusion about one time in three (67.3% correct on average). These results suggest that VLMs need both more precise perception and better reasoning from what they see to the underlying physics.

Our contributions are threefold:

- **A benchmark for optoelectronic device figures.** OPTOFIG has 1,092 questions over 799 figures in 39 task types; to our knowledge, it is the first to evaluate VLMs on the figures of optoelectronic devices, covering their design, the characterization of their performance, and the physical reasoning that links the two.
- **A physics-grounded construction and validation protocol.** Answers come from the figure, simulator or drawing-script records, or a domain expert, and pass automated checks and expert review with measured per-type error rates; difficulty is set by construction and tested per task type; and matched question pairs separate reading from reasoning.
- **A diagnostic evaluation of SOTA VLMs.** We break down results by task type and difficulty tier, and identify two limitations: limited perceptual precision and unreliable reasoning from a correct reading to the physical conclusion, pointing to where VLMs must improve before they can reliably support optoelectronic research.

2 RELATED WORK

Scientific figure and chart understanding. A growing line of work tests whether VLMs can read and reason about scientific figures. Early benchmarks, such as ChartQA, PlotQA, and FigureQA, focus mainly on charts and plots (Masry et al., 2022; Methani et al., 2020; Kahou et al., 2018). CharXiv then extends this to charts from real scientific papers (Wang et al., 2024). Broader benchmarks, such as SciFIBench and SPIQA, test figure interpretation and multimodal question answering over research papers (Roberts et al., 2024; Pramanick et al., 2024). Closest to our setting, QCalEval tests VLMs on domain-specific plots from quantum-calibration experiments (Cao et al., 2026). Together, these benchmarks show that figure understanding matters. However, they focus on charts, paper figures, or calibration plots. To our knowledge, no benchmark evaluates VLMs on the figures of optoelectronic devices, covering both their design and the characterization of their performance, and none tests the physical reasoning that links the two.

Physics and engineering visual reasoning. A second line of work tests multimodal reasoning in science and engineering. General benchmarks such as MMMU and MathVista include scientific and technical problems among broader multimodal tasks (Yue et al., 2024; Lu et al., 2024). More specialized benchmarks test physics problem solving with diagrams from textbooks, exams, and competitions (Xiang et al., 2025; Wang et al., 2026), and EEE-Bench extends multimodal problem solving to electrical and electronics engineering (Li et al., 2025). Circuit benchmarks come closer to engineering design drawings. For example, CircuitSense tests schematic perception, analysis, and design, while AMSbench and CircuitReason-1k test understanding and symbolic reasoning over schematics (Akbari et al., 2026; Shi et al., 2025; Yang et al., 2026). These benchmarks show that a model can perform very differently on visual perception and on domain-specific reasoning. OPTOFIG complements them by covering optoelectronic devices, from design drawings to the simulated plots that characterize their performance, with recognition, quantitative, and reasoning questions.

Table 1 compares OPTOFIG with the most closely related benchmarks. Unlike prior work, OPTOFIG evaluates both design and characterization figures of optoelectronic devices. It also provides fine-grained task categories, difficulty defined by how each item is built, and ground truth traced to the source of each figure. As a result, OPTOFIG can show not only how well VLMs perform overall but also where they fail across task types and difficulty levels.

Table 1: Comparison of OPTOFIG with representative multimodal benchmarks for scientific figure understanding and domain-specific visual reasoning. ✓ means that the benchmark explicitly evaluates the property in that column.

Benchmark	Scientific figures	Domain-specific	Device drawings	Simulated performance	Task-level taxonomy	Controlled difficulty
CharXiv	✓	–	–	–	✓	–
SciFIBench	✓	–	–	–	✓	–
SeePhys	–	✓	✓	–	✓	✓
CircuitSense	–	✓	✓	–	✓	–
QCalEval	✓	✓	–	–	✓	–
OPTOFIG	✓	✓	✓	✓	✓	✓

Language models for optoelectronic device design. Language models increasingly help design optoelectronic devices, especially photonic ones. For example, recent LLM agents propose device geometries, run numerical simulators, and repeatedly check designs against quantitative criteria (Kharel et al., 2026). Other systems generate photonic integrated-circuit design scripts from natural-language specifications (Jiang et al., 2026), or connect language models to differentiable solvers for inverse design of meta-optics (Huang et al., 2025). PICBench, in contrast, is a benchmark: it evaluates LLMs on photonic integrated-circuit design tasks (Wu et al., 2025). Beyond device design, VLM-based systems have begun to extract quantitative information directly from figures in scientific papers (Roy et al., 2026). Together, these developments motivate agents that work not only with code and structured simulator outputs, but also with the figures that researchers use to inspect and interpret device behavior. OPTOFIG evaluates this figure-understanding ability in isolation.

3 THE OPTOFIG BENCHMARK

We build OPTOFIG in six stages (Figure 2). (1) We collect figures of optoelectronic devices from simulations and drawing scripts and keep the legible ones. (2) Each answer is measured on the figure, read from the simulator or drawing-script record, or written by a domain expert. (3) Each fact becomes a hard-to-guess question, and we record the parameter that sets its difficulty. (4) A text-only gate removes questions answerable without the image, and external language models, none of which we evaluate, reword the questions to remove template wording. (5) A domain expert confirms, corrects or drops every question; per-type error rates decide where a spot check suffices. (6) We release each question with its answer source, difficulty tier and review record. Below, we describe figures and taxonomy (§3.1), ground truth (§3.2), difficulty (§3.3) and validation (§3.4).

3.1 FIGURES AND TAXONOMY

Stage 1 of Figure 2 shows the source figures, which capture how optoelectronic devices are built (e.g., device structures, component symbols, circuit schematics) and how they perform (e.g., mode profiles, spectra, timing diagrams); of the 2,512 collected figures, 2,377 pass a usability screen. Optical figures are simulated, mostly with the photonics suite Ansys Lumerical: we ran projects from its example library with solvers such as FDTD (light propagation) and MODE (guided modes of a waveguide). Electrical figures are drawn with the Python library Schemdraw (Appendix A). The 1,092 released questions use 799 distinct figures, concern either the optical side of a device (371) or its electrical side (721), and have two independent labels. The first is the *task type*, the template that generated the question, such as counting the peaks of a spectrum, naming the component a symbol stands for, or comparing two modes by effective index (how slowly a mode travels along the waveguide). Each of the 39 task types (Appendix G) has one figure kind, one question form and one ground-truth route, so we define difficulty and report results per task type. We group the task types into three *capability categories* (Table 2): ① perception and recognition (12 task types, 552 items), ② quantitative and comparative (19 types, 412 items) and ③ physical and functional reasoning (8 types, 128 items). Within each category, Table 2 groups the task types into nine levels (L1–L9) by the kind of reading they require; the levels organize the table and appendix, the categories the main text. The categories are kinds of capability, not a difficulty ladder: their answer formats, and hence

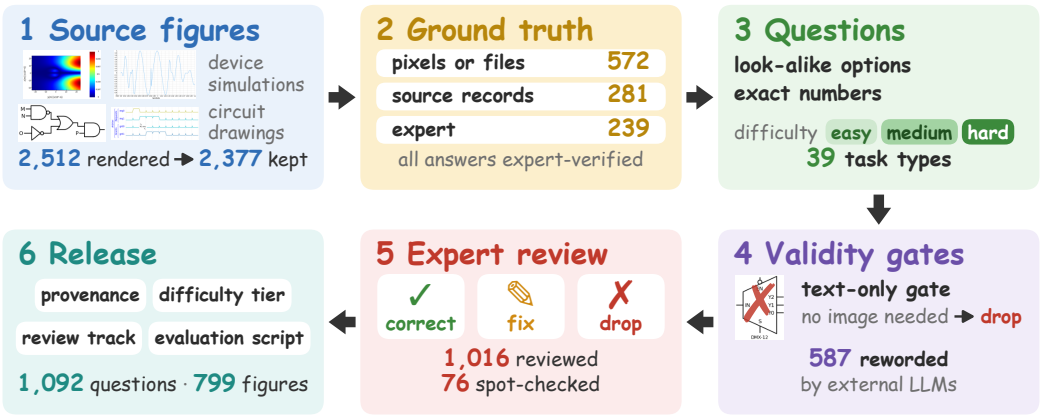


Figure 2: Construction pipeline of OPTOFIG. (1) **Source figures**: optoelectronic device simulations and drawings, screened for legibility. (2) **Ground truth**: measured on the figure, read from the simulator or drawing-script record, or written by a domain expert. (3) **Questions**: each fact becomes a hard-to-guess question, with the parameter that sets its difficulty recorded. (4) **Validity gates**: questions answerable without the image are removed, and external language models, none of which we evaluate, reword the questions. (5) **Expert review**: a domain expert confirms, corrects or drops every question; per-type error rates decide where a spot check suffices. (6) **Release**: each question has its answer source, difficulty tier and review record.

Table 2: Capability taxonomy with item counts. Categories ① and ② ask what a figure shows; ③ asks what it implies. Thick lines separate categories and thin lines separate levels. Construction rules per level are in Appendix B, and example items in Appendix C.

Category	Level (n)	Example question
① **Perception & Recognition (552)**	L1 Perception (84)	What kind of figure is this? How many signal lanes does this timing diagram have?
	L2 Reading (100)	Which physical quantity is plotted? Which visual feature identifies this component?
	L3 Recognition (289)	Which component, circuit type or logic gate does this drawing represent?
	L4 Spatial (79)	Which of these four maps is *not* left–right symmetric?
② **Quantitative & Comparative (412)**	L5 Quantitative (242)	How many distinct peaks does this curve show? How many pins does this package have?
	L6 Comparison (170)	Which of these two modes is more tightly confined? Which signal rises first?
③ **Physical & Functional Reasoning (128)**	L7 Inference (86)	Which of two modes of the same waveguide has the higher effective index? What does this circuit output when all inputs are 1?
	L8 Counterfactual (22)	If the index changed from n_1 to n_2, where would this resonance appear?
	L9 Cross-modal (20)	Expert-authored statements linking structure, field and spectrum (exact-set multi-select).

guessing floors, differ (24.6%, 16.8% and 8.3%). The second label, the *difficulty tier* (easy, medium or hard), is set within each task type by a construction parameter, independently of the category (§3.3).

Descriptive and physical questions. Some questions come in pairs that separate reading a figure from reasoning about it. Categories ① and ② ask what is drawn, whereas ③ asks what follows from it by a physical or logical rule. In ③, 55 questions also have a descriptive partner on the same figure. The pairs ask which of two modes is more confined (its light more concentrated in the waveguide core) and hence which has the higher effective index (11 pairs); the wavelength of a resonance and where it moves after an index change (10); and the number of gates in a circuit and its output when all inputs are one (34). Comparing accuracy within a pair thus shows whether a model that reads the figure correctly also takes the reasoning step (§4.3).

3.2 Questions and ground truth

Questions are either multiple choice or fill-in. The 719 multiple-choice questions (66%) ask the model to pick a name from a short list, usually four, such as candidate component types; to name which of two to four figures has a given property; or to answer yes or no. In contrast, the 373 fill-in questions (34%) offer no options: the model gives an integer count, a numeric reading, or, in 12 questions, a short conclusion. Both formats are built so that guessing rarely works. For example, in multiple choice, wrong options look like the right one (a diode symbol is listed among other diode symbols), and a multi-figure question asks which figure has more of a quantity, such as peaks. Fill-in answers are graded exactly for integers, within $\pm 10\%$ for readings, and against the expert's written explanation for conclusions. Averaged over the set, random guessing scores 19.7%.

Each gold answer in OptoFig comes from one of three sources and is verified by a domain expert (§3.4). First, for 572 questions we compute the answer from the figure or the file that produced it: we count peaks, measure peak widths, test whether a map is mirror-symmetric, count signal lanes and pulses in a timing diagram, and read a component symbol's class from its drawing script or a mode's effective index from the simulation file. Second, for 281 questions the answer is a label (e.g., figure type, plotted quantity, circuit type or logic gate), taken from the source metadata and checked by the expert. Finally, the remaining 239 answers come from the expert, who wrote 67 questions from scratch and corrected, one by one, 172 gate and pin counts drafted by a helper model.

3.3 Difficulty by construction

Most task types have one *difficulty knob*, a construction parameter meant to control difficulty, fixed at generation and independent of any model. The knob is the gap between compared values (comparisons), the count (counting), the number of look-alike wrong options (recognition), clutter (lane count, schematic ink density), or reading precision (axis range relative to the allowed error). Within each task type, items are split into three equal tiers by knob value; 89 items (five task types and 30 quantity items) have no knob and are tiered by reasoning steps. However, we do not assume that a knob controls difficulty; we test it against per-type model accuracy (§4.3).

3.4 Validation

Separation of model roles. Four groups of models play separate roles in building the benchmark. First, the evaluated models (§4.1) never write a question or an answer. Second, two helper models, Gemini-2.5-Flash and Qwen2.5-VL-72B, assist in Stage 2: they read the axis ranges of spectra so that pixel positions can be converted to wavelengths, re-derive figure-type labels from the source metadata, and draft the gate and pin counts that the expert later corrects. Third, two validation models, Gemini-2.5-Flash and GPT-4o-mini, answer questions without the image in Stage 4 (the text-only gate below); GPT-4o-mini also grades the 12 short-conclusion questions at test time. Finally, three text models, Qwen-2.5-72B, Llama-3.3-70B and DeepSeek-V3, reword questions in Stage 4.

Validity gates and phrasing. Stage 4 applies two steps to every generated question. First, a *text-only gate* removes questions whose answer can be guessed from the wording alone: we give each question to the two validation models without its figure, and drop it if either model answers correctly with high confidence. Second, *rewording* removes repeated template phrasing. Each question is rewritten by one of three text models in rotation (Qwen-2.5-72B, Llama-3.3-70B, DeepSeek-V3); a model from another family then checks whether the rewrite still asks the same question, and if not, we keep the original. Of the 1,092 released questions, 587 are reworded, 79 were rewritten by hand

Table 3: Accuracy (%) by category, with reasoning switched off (Gemini-3.1-Pro at its minimal level). The best value in each column is shaded and bold; the second best is underlined.

Model	① (552)	② (412)	③ (128)	Overall (1,092)
Chance (format mix)	24.6	16.8	8.3	19.7
DeepSeek-V4.1-Flash	74.3	**76.5**	69.5	**74.5**
DeepSeek-V4-Flash	73.9	<u>73.3</u>	64.8	<u>72.6</u>
Qwen3-VL-235B-Instruct	**78.3**	62.9	**72.7**	71.8
GPT-4o	75.2	61.7	55.5	67.8
GPT-5.6-Terra	<u>76.1</u>	57.5	60.9	67.3
Gemini-3.1-Pro	65.9	57.3	<u>71.9</u>	63.4
Gemini-3.6-Flash	54.9	51.0	61.7	54.2
Mean of seven	71.2	62.9	65.3	67.4

during expert review, and 426 keep their original wording, mostly one-sentence generic questions such as "What electronic component does this symbol represent?" and expert-written ones.

Expert review. In Stage 5, the domain expert reviewed every question individually, with one exception. For circuit-type recognition, the largest and most uniform task type, the expert reviewed a random 27 of its 103 items; since all 27 were correct, the remaining 76 were accepted without individual review. In total, the expert reviewed or wrote 1,016 questions, and 76 were accepted through this sample. The expert marked each reviewed question correct, wrong (and supplied the correct answer), or unreasonable (removed). Thus, review corrected items rather than only discarding them. Per-type error rates are in Appendix E.

4 Experiments

4.1 Models and protocol

We evaluate seven VLMs: GPT-5.6-Terra and Gemini-3.1-Pro (frontier), GPT-4o (previous generation), Gemini-3.6-Flash (economy), and Qwen3-VL-235B-Instruct, DeepSeek-V4-Flash and DeepSeek-V4.1-Flash (open weights). We query them all through one API at temperature 0, with a 4,000-token output budget and reasoning switched off; Gemini-3.1-Pro cannot switch it off and runs at its minimal level (about 540 reasoning tokens per item). All share the same prompts and grading: we parse multiple-choice answers as JSON (regex as fallback) and grade integers exactly and readings within $\pm 10\%$, while a GPT-4o-mini judge checks the 12 short conclusions against the expert's explanation. No reply was empty.

Models from two evaluated families, Qwen and DeepSeek, also reworded questions (§3.4). This gives these families at most a small edge: on questions reworded by their own family rather than by others, accuracy shifts by -2.3 points for Qwen3-VL-235B-Instruct and by $+1.9$ and $+1.1$ for DeepSeek-V4.1-Flash and DeepSeek-V4-Flash.

4.2 Overall results

All seven models leave substantial room for improvement, and no single model leads across capabilities. Overall accuracy spans 20 points (54.2–74.5%; Table 3), and the ranking follows neither release date nor price: the open-weight DeepSeek-V4.1-Flash leads overall and on quantitative questions, Qwen3-VL-235B-Instruct leads on perception and physical reasoning, the frontier GPT-5.6-Terra ranks fifth, and Gemini-3.1-Pro leads on levels L4, L6, L8 and L9 (Table 4). The benchmark is not saturated: 25% of items are solved by all seven models, 3% by none and 16% by at most two. Mean accuracy is also 10.2 points lower on hard than on easy items (§4.3).

The two Gemini models rank last mainly because they misread circuit drawings. Gemini-3.6-Flash and Gemini-3.1-Pro score 36% and 53% on the 137 component-symbol items (others 65–70%), 30–50% on all electrical recognition items (others 67–76%), 3% on gate counts and 4–10% on chip pin

Table 4: Per-level accuracy (%), item counts in italics; chance is each level's guessing floor. The best value in each column is shaded and bold; the second best is underlined (ties share the mark).

Model	① Perception & recognition				② Quantitative		③ Physical reasoning		
	L1 *84*	L2 *100*	L3 *289*	L4 *79*	L5 *242*	L6 *170*	L7 *86*	L8 *22*	L9 *20*
Chance	9	25	25	39	0	41	7	14	7
DeepSeek-V4.1-Flash	95	**79**	65	78	**81**	70	**78**	50	55
DeepSeek-V4-Flash	90	76	67	78	77	68	67	50	70
Qwen3-VL-235B-Instruct	**100**	78	**74**	72	69	54	77	59	70
GPT-4o	96	**79**	70	68	61	63	59	45	50
GPT-5.6-Terra	96	72	70	81	50	68	66	27	75
Gemini-3.1-Pro	94	64	51	**94**	35	**89**	62	**91**	**95**
Gemini-3.6-Flash	94	60	32	91	31	80	52	77	85

counts. This is not a parsing error: their answers are cleanly formatted, and on the 52 symbol items both miss, they choose the same wrong option 29 times. Nor is it a general reading problem: they are the top two on spatial questions (L4, 91–94%), expert items (L9, 85–95%) and optical items (82–88% against 67–73%). Since 66% of items are electrical, they drop from the top on optical items to the bottom overall.

4.3 Discussion

Perception lacks precision. Models capture coarse visual patterns but often miss the exact quantities a question depends on. For example, the models average 83.8% when asked which of two or three spectra shows more peaks (22 items) but 45.5% when asked how many peaks one spectrum shows (16 items), and every model counts less accurately than it compares. Most counting errors are near misses: of the 61 wrong peak counts, 44 are off by exactly one and 48 are too low. Similarly, gate-input counting is the hardest task type with at least 30 items (33 items, mean 38%, range 12–67%), and GPT-4o counts the pulses of one timing-diagram signal correctly on 17% of items (others 31–83%). By contrast, large structures are counted almost perfectly, even in the same figures: on the 17 timing diagrams with both a lane-count and a pulse-count question, the models average 99% on the number of lanes but 51% on the number of pulses in one lane. The imprecision thus concerns small, repeated features that must be tallied one by one.

Reading does not guarantee reasoning. Even when a model reads a figure correctly, it often draws the wrong physical conclusion. On the 21 optical pairs, a correct descriptive answer leads to a correct physical answer 67.3% of the time on average (Table 5, last row). For example, every model names the more confined of two modes on at least 10 of 11 pairs, yet only 45–91% of these correct readings lead to the right conclusion, that this mode has the higher effective index (chance 50%). Moreover, reading accuracy does not predict this rate: over all pairs, DeepSeek-V4-Flash and Qwen3-VL-235B-Instruct both answer 96% of descriptive questions correctly, yet turn 60% and 79% of them into correct physical answers. Thus, the physical step is a separate capability that descriptive scores hide. The counterfactual level L8, which asks what-if questions, shows where this step breaks. Its ten index-change items need one reading and one scaling: a resonance wavelength is proportional to the refractive index. Of the 20 wrong answers across the seven models, 18 move the resonance the wrong way, and in all 18 the same model read the starting wavelength correctly; the other two follow a misread start. These errors lie in the physical relation, not in the reading.

Implications for research agents. An agent that reads a figure could pass an off-by-one resonance count or a reversed shift unnoticed into its next design step. We suggest two directions, which we have not tested. First, exact quantities could be measured by tools such as curve digitizers and peak finders, as we do for the gold peak and pulse counts (Appendix D), with the VLM deciding what to measure. Second, known physical relations, such as the index scaling of a resonance, could be applied in code rather than in free text, and the reading and reasoning steps could be evaluated separately, as our paired questions do.

Table 5: Accuracy (%) on question pairs sharing a figure, as descriptive / physical. Last two rows: physical accuracy given a correct descriptive answer (best shaded and bold). Model abbreviations: DS-4.1 = DeepSeek-V4.1-Flash, DS-V4 = DeepSeek-V4-Flash, Qw-I = Qwen3-VL-235B-Instruct, Terra = GPT-5.6-Terra, Gem-P = Gemini-3.1-Pro, Gem-F = Gemini-3.6-Flash.

Descriptive / physical (n)	DS-4.1	DS-V4	Qw-I	GPT-4o	Terra	Gem-P	Gem-F
Confined / higher n_{eff} (11)	100/73	91/45	100/45	91/45	91/55	100/91	100/73
Wavelength / after Δn (10)	100/50	90/50	80/80	100/80	90/40	100/100	100/100
Gate count / output (34)	88/82	100/68	100/88	88/62	62/71	3/56	3/47
All pairs (55)	93/75	96/60	96/78	91/62	73/62	40/71	40/62
P(phys. correct | desc. correct)	75	60	79	60	52	**95**	82
Same, 21 optical pairs	62	47	63	65	53	**95**	86

Table 6: Accuracy (%) by difficulty tier. Tiers are set by construction parameters alone; which task types count as validated is decided from these models' accuracy (Appendix H). The best model in each row is shaded and bold. Model abbreviations as in Table 5.

Scope	Tier	n	DS-4.1	DS-V4	Qw-I	GPT-4o	Terra	Gem-P	Gem-F	Mean
All items	easy	255	80	81	**82**	76	77	64	57	74
	medium	375	**78**	76	74	69	66	61	50	68
	hard	462	**69**	65	65	62	63	65	56	64
Validated types (13)	easy	49	88	88	90	90	**92**	84	82	87
	medium	49	82	**86**	49	49	57	69	63	65
	hard	152	62	53	45	43	50	**72**	63	55

Difficulty tiers. Mean accuracy falls as the tier gets harder (Table 6). Over all items, it falls from 74% to 68% to 64%; the drop is small because not every construction parameter controls difficulty (Appendix H). In contrast, on the 13 task types whose parameter is validated (250 items), the mean falls from 87% to 65% to 55%, and the best hard-tier score is 72%. There, every model drops from easy to hard, including both Gemini models (from 82–84% to 63–72%). Validated knobs (§3.3) also offer a way to build harder test sets as models improve: generate new items at more extreme knob values, such as smaller margins or larger counts, instead of keeping only items that current models fail. Since the knobs were validated on these models, they should be re-checked for new models.

Limitations. OPTOFIG has several limitations. First, it tests figure reading on its own, not whether better reading improves an agent's designs; measuring that would need an end-to-end study in which an agent acts on what it reads. Second, some subsets are small by nature: physical reasoning, the smallest category (128 items), the 55 matched pairs, and L8 and L9 (22 and 20 items). These items are costly to build, because each needs a physical or logical relation that the expert confirms for the figure, or a question that the expert writes. Finally, the dataset has room to scale up, since the same construction process can produce more figures and questions; however, each new item still needs an optoelectronics expert to confirm it.

5 CONCLUSION

We introduced OPTOFIG, a benchmark that tests whether VLMs can read the figures used to design and characterize optoelectronic devices, and reason from them to the underlying physics. Its answers are verified by a domain expert, its descriptive and physical questions are paired on the same figures, and its difficulty is set by construction rather than by keeping the items that models fail. Results on seven state-of-the-art VLMs leave substantial room for improvement and reveal two gaps: perception that captures coarse patterns but misses exact quantities, and reasoning that often fails to turn a correct reading into the correct physical conclusion. Research agents that rely on device figures would inherit both gaps, and OPTOFIG offers a controlled way to measure progress on each.

AI USE STATEMENT

We used large language models to draft and polish the writing of this paper. Language models also served as tools to help construct the benchmark; their roles are detailed in §3.4 and Appendix E.

ETHICS STATEMENT

All figures in OPTOFIG are rendered from device simulations and drawing scripts, and all questions and answers are derived from them, so the benchmark involves no human subjects, no personally identifiable information and no photographs or figures copied from publications, and no institutional review was required. The questions, images, provenance records and evaluation code will be released for research use.

REPRODUCIBILITY STATEMENT

Section 3 and Appendices B and D describe how every item was constructed, Appendix E reports the outcome of expert review, and Appendix F gives the model identifiers, decoding settings and grading procedure. The anonymized supplementary material contains the released questions and images with their provenance, pairing, difficulty and review records; the extraction, evaluation and analysis code; and the per-item outputs of every evaluated model, from which a single script recomputes every number reported in §4. A public repository will be linked in the camera-ready version.

REFERENCES

Arman Akbari, Jian Gao, Yifei Zou, Mei Yang, Jinru Duan, Dmitrii Torbunov, Yanzhi Wang, Yihui Ren, and Xuan Zhang. CircuitSense: A hierarchical MLLM benchmark bridging visual comprehension and symbolic reasoning in engineering design process. *arXiv preprint arXiv:2509.22339*, 2026.

Shuxiang Cao, Zijian Zhang, Abhishek Agarwal, Grace Bratrud, Niyaz R. Beysengulov, Daniel C. Cole, Alejandro Gómez Frieiro, Elena O. Glen, Hao Hsu, Gang Huang, Raymond Jow, Greshma Shaji, Tom Lubowe, Ligeng Zhu, Luis Mantilla Calderón, Nicola Pancotti, Joel Pendleton, Brandon Severin, Charles Etienne Staub, Sara Sussman, Antti Vepsäläinen, Neel Rajeshbhai Vora, Yilun Xu, Varinia Bernales, Daniel Bowring, Elica Kyoseva, Ivan Rungger, Giulia Semeghini, Sam Stanwyck, Timothy Costa, Alán Aspuru-Guzik, and Krysta Svore. QCalEval: Benchmarking vision-language models for quantum calibration plot understanding. *arXiv preprint arXiv:2604.25884*, 2026.

Yi Huang, Bowen Zheng, Yunxi Dong, Hong Tang, Huan Zhao, Rakibul Hasan Shawon, Sensong An, and Hualiang Zhang. MCP-enabled LLM for meta-optics inverse design: Leveraging differentiable solver without LLM expertise. *Nanophotonics*, 14(27):5589–5602, 2025.

Xiaohan Jiang, Zeyu Li, Wei Zhang, and Jiang Xu. PICopilot: An LLM-based agentic framework for assisting photonic integrated circuit design via script generation. *arXiv preprint arXiv:2608.01791*, 2026.

Samira Ebrahimi Kahou, Vincent Michalski, Adam Atkinson, Akos Kadar, Adam Trischler, and Yoshua Bengio. FigureQA: An annotated figure dataset for visual reasoning. *arXiv preprint arXiv:1710.07300*, 2018.

Prashanta Kharel, Amin Khavasi, Xinzhong Chen, and Tyler W. Hughes. Autonomous agentic design for photonics. *arXiv preprint arXiv:2606.00915*, 2026.

Ming Li, Jike Zhong, Tianle Chen, Yuxiang Lai, and Konstantinos Psounis. EEE-Bench: A comprehensive multimodal electrical and electronics engineering benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13337–13349, 2025.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. MathVista: Evaluating mathematical reasoning of

foundation models in visual contexts. In *International Conference on Learning Representations (ICLR)*, 2024.

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 *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2263–2279, 2022.

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)*, pp. 1516–1525, 2020.

Shraman Pramanick, Rama Chellappa, and Subhashini Venugopalan. SPIQA: A dataset for multimodal question answering on scientific papers. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 118807–118833, 2024.

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 (NeurIPS)*, pp. 18695–18728, 2024.

Aritra Roy, Enrico Grisan, Chiara Gattinoni, and John Buckeridge. Beyond text and tables: Vision-language model integration in ComProScanner for extracting materials data from scientific figures with high accuracy. *arXiv preprint arXiv:2606.00065*, 2026.

Yichen Shi, Ze Zhang, Hongyang Wang, Zhuofu Tao, Zhongyi Li, Bingyu Chen, Yaxin Wang, Zhen Huang, Xuhua Liu, Quan Chen, Zhiping Yu, Ting-Jung Lin, and Lei He. AMSbench: A comprehensive benchmark for evaluating MLLM capabilities in AMS circuits. *arXiv preprint arXiv:2505.24138*, 2025.

Lintao Wang, Encheng Su, Jiaqi Liu, Pengze Li, Jiabei Xiao, Wenlong Zhang, Xinnan Dai, Xi Chen, Yuan Meng, Lei Bai, Wanli Ouyang, Shixiang Tang, Aoran Wang, and Xinzhu Ma. PhysUniBench: A multi-modal physics reasoning benchmark at undergraduate level. *arXiv preprint arXiv:2506.17667*, 2026.

Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, Alexis Chevalier, Sanjeev Arora, and Danqi Chen. CharXiv: Charting gaps in realistic chart understanding in multimodal LLMs. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 113569–113697, 2024.

Yuchao Wu, Xiaofei Yu, Hao Chen, Yang Luo, Yeyu Tong, and Yuzhe Ma. PICBench: Benchmarking LLMs for photonic integrated circuits design. In *Design, Automation & Test in Europe Conference (DATE)*, pp. 1–6, 2025.

Kun Xiang, Heng Li, Terry Jingchen Zhang, Yinya Huang, Zirong Liu, Peixin Qu, Jixi He, Jiaqi Chen, Yu-Jie Yuan, Jianhua Han, Hang Xu, Hanhui Li, Mrinmaya Sachan, and Xiaodan Liang. SeePhys: Does seeing help thinking? – Benchmarking vision-based physics reasoning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Xinqi Yang, Kang An, Tengyue Wang, Zhongyu Yang, Chenxu Du, Yuanchi Zhu, Hebao Zhu, Ziliang Wang, Faqiang Qian, Yunli Yang, and Qibing Ren. CircuitReason-1k: Benchmarking long-horizon visual-to-symbolic reasoning in electrical circuits. *arXiv preprint arXiv:2608.09374*, 2026.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9556–9567, 2024.

A How the figures were generated

The source figures come from two kinds of tools: photonic device simulators, which compute how light, charge carriers and heat behave in a device, and a Python library that draws electrical circuits. Table 7 lists what each tool produced, split into *design* figures, which show how a device is built, and *characterization* figures, which show how it performs. Of the 2,512 collected figures, 135 fail the Stage 1 usability screen: 75 that the expert excluded (36 optimization dashboards that combine several sub-plots and 39 near-duplicate curves from one arrayed waveguide grating), 56 electrical figures whose type remained unclear, and 4 on which figure-type labelling failed.

Table 7: Tools that generated the source figures. Collected and usable counts refer to the Stage 1 usability screen (§3.1); counts in parentheses are collected figures.

Tool	Collected	Usable	Design figures	Characterization figures
Ansys Lumerical (62 example projects)	1,035	956	Structure views (XY, XZ, YZ, 3D); refractive-index and mesh maps; photonic circuit schematics	Field maps and line cuts; mode profiles; transmission and reflection spectra; far fields; photonic band structures; time signals; carrier-density, band-diagram, temperature and quantum-well band-structure, wavefunction and emission plots
COMSOL Multiphysics 6.3 (1 example)	19	19	Geometry; material maps	Electric fields; far fields
Python and Schemdraw (9 figure categories)	1,458	1,402	Component symbols (223); compound elements (135); connectors (81); integrated-circuit symbols and packages (200); logic gates and circuits (121); breadboard layouts (480); flowcharts (14); signal-processing blocks (20)	Timing diagrams (155); truth tables and Karnaugh maps (29)
Total	2,512	2,377		

Simulated figures. All but 19 of the collected optical figures come from Ansys Lumerical, a commercial suite of photonic device simulators. We took ready-made projects from its example library and application gallery, ran them, and saved the resulting figures. The projects use several solvers: FDTD simulates how light propagates through a 3D structure over time; MODE computes the guided modes of a waveguide cross-section and their effective indices (FDE); FEEM and DGTD are finite-element mode and time-domain solvers; RCWA handles periodic layered structures such as gratings; CHARGE, HEAT and MQW compute carrier transport, temperature, and quantum-well band structure and gain; and INTERCONNECT simulates photonic circuits, whose schematics supply the optical circuit figures. The generation prompt asked for four structure views per sample (top view XY, cross-sections XZ and YZ, and a 3D perspective) besides the results, and one result is often saved in several renderings (2D map, 1D line cut, vector plot or 3D surface). The two largest sets are parameter series: 36 variants of an augmented-reality windshield grating, whose period follows a closed-form grating formula, and 27 variants of the analysis settings of an arrayed waveguide grating with fixed geometry. Structure views that repeat across variants appear at most once in the release, which removes duplicate images by content. The remaining 19 figures model a vertical-cavity surface-emitting laser (VCSEL) in the finite-element simulator COMSOL Multiphysics 6.3, starting from its Wave Optics application library; a second VCSEL example, run in Lumerical's VCSEL solver, gives 52 of the Lumerical figures.

Drawn figures. Electrical figures are drawn, not simulated, with Schemdraw, a Python library that draws circuit diagrams through matplotlib. One script per figure category starts from the examples in Schemdraw's documentation and adds rule-based variants, such as other element options, gates with two to five inputs or new Boolean expressions; most scripts drop duplicates by structure or image hash. Multi-gate logic circuits are drawn from Boolean expressions and timing diagrams from WaveJSON-style waveform descriptions, so the electrical characterization figures are specified rather than computed. Each component symbol's class and options are fixed in its drawing script, from which the answers to component-symbol questions are parsed (Appendix D). Rerunning the digital-logic script with Schemdraw 0.22 reproduces all 150 of its figures pixel for pixel.

Released images. Of the 799 released images, 55 come from outside the collected set: 50 mode-profile images used by restored mode-comparison questions and 5 images used by expert-authored items.

B LEVEL DEFINITIONS AND CONSTRUCTION RULES

Table 8: Ground-truth sources and construction rules by level.

Level	Ground-truth source	Construction rule
L1 figure type / lanes	Figure-type label from the source metadata, checked by the expert; lane count from timing-lane segmentation	4- or 5-way figure-type choice; integer lane count
L2 quantity / visual cue	Physical-quantity label from the source metadata; visual cues declared in the rendering script	Up to two same-family distractors, the rest cross-family
L3 recognition	Component class and variant from the rendering script; circuit type from the source metadata; logic gates labelled by the expert	Same-category distractors; IEEE/IEC variants merged; gates: same body family $\pm$ inversion
L4 spatial	Mirror correlation of field patterns, 1D profiles and structure views	4-way odd-one-out; yes/no profile symmetry
L5 quantitative	Peak count; gate input and inverted-input count; pulse count per lane; pin and gate counts (model-drafted, expert-corrected); axis read from pixel position with end points read by a calibration model	Integer fill-in graded exactly; readings within $\pm 10\%$
L6 comparison	Active fraction, relative FWHM, peak count, first rising edge, longest high stretch; modal n_{eff} from the simulation file	Two- to five-way choice; the compared values must differ by a clear margin, which is released per item as the difficulty knob (Appendix H)
L7 physical inference	An L6 fact plus one physical relation; circuit outputs for all-1 / all-0 inputs; expert items posed open-ended	Expert confirms the relation applies to the figure
L8 counterfactual	$\lambda' = \lambda n_2/n_1$ on an L5 read; expert read–compute–compare items	Needs the reading
L9 cross-modal	Expert-authored (15 of 20 multi-select)	Exact-set grading

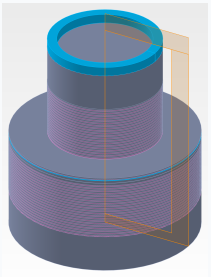
C EXAMPLE ITEMS

Figures 3 and 4 show six released items, two per category, all built on figures of one vertical-cavity surface-emitting laser (VCSEL), a laser that emits light from its surface through a short cavity between two stacks of mirror layers. Two design figures show how the device is built, a side view of the structure (next to a 3D view of the device, for context) and a map of its layers and regions; three characterization figures show how it performs, two reflection spectra and a map of transmission against frequency and angle. Each box gives the question exactly as the models saw it, the gold answer and its source (§3.2), and which of the seven models answered it correctly under the protocol of §4.1. Two items share one spectrum: the first asks what the curve shows, the second what happens to it when the structure changes. These single items illustrate rather than measure, but they follow the pattern of §4.3: every model reads the coarse mirror symmetry of the structure, whereas most miss the exact counts and the physical consequence.

D GROUND-TRUTH ROUTES IN DETAIL

Pixels. A spectrum is digitized by isolating its coloured trace with a saturation mask, taking the per-column trace position, and normalizing to $[0, 1]$; peaks and troughs are found at prominence 0.15, and the sharpest resonance's relative width is measured with `scipy peak_widths` at half prominence on the trace and its complement. Only single-trace spectra with at least one resonance feature are used. For a field map, the background is the colour at the bottom of the colour bar and the active region is the set of pixels whose colour distance from it exceeds 35% of the maximum;

① Perception & recognition · L4 Spatial · design: 3D view and side view of the structure



3D view (context only)

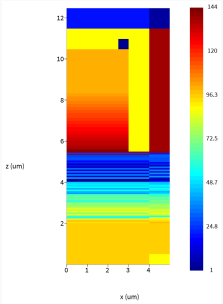
Item image

Q: Does the device structure exhibit mirror symmetry across its vertical central axis? Answer 'yes' or 'no'.
Gold: yes **Source:** computed from the figure (mirror correlation)
Models (7/7 correct):

DS-4.1	DS-V4	Qw-I	GPT-4o	Terra	Gem-P	Gem-F
✓	✓	✓	✓	✓	✓	✓

A coarse property of the structure: every model answers correctly.

① Perception & recognition · L2 Reading · design: map of layers and regions

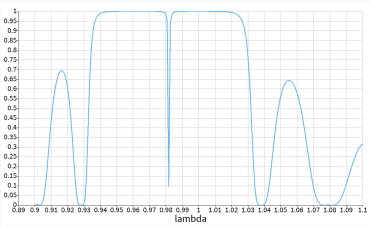


Q: What physical quantity does this plot primarily show?
(A) Transfer function (B) Generation rate
(C) Refractive index map (D) Material distribution
Gold: D **Source:** label from the source metadata
Models (1/7 correct):

DS-4.1	DS-V4	Qw-I	GPT-4o	Terra	Gem-P	Gem-F
✗	✗	✗	✗	✗	✗	✓

The colours number the device's layers and regions (1–144) and are not a physical field; the wrong answers are B (four models) and C (two).

② Quantitative & comparative · L5 Quantitative · characterization: reflection spectrum



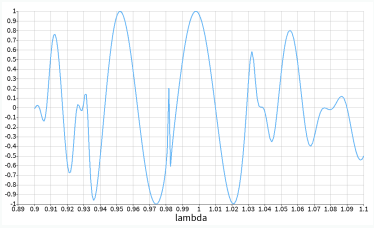
Q: How many distinct peaks does this curve show? Count a peak only if it rises at least 15% of the curve's full vertical range above the adjacent minima on both sides; a rise at the edge of the plot that does not come back down does not count. Answer with a single integer.
Gold: 4 **Source:** computed from the figure (peak finder)
Models (1/7 correct):

DS-4.1	DS-V4	Qw-I	GPT-4o	Terra	Gem-P	Gem-F
✗	✗	✗	✗	✗	✓	✗

The narrow dip near 0.98, the cavity resonance, splits the band of high reflection into two peaks; five models answer 3 and one answers 2.

Figure 3: Example items on one VCSEL (part 1 of 2): two design figures and a reflection spectrum, with the gold answer and each model's result (✓ correct, ✗ wrong). Model abbreviations as in Table 5. The 3D view in the first box shows the device for context; the models saw only the side view. Images are cropped to their content for display.

② Quantitative & comparative · L5 Quantitative · characterization: amplitude-reflection spectrum

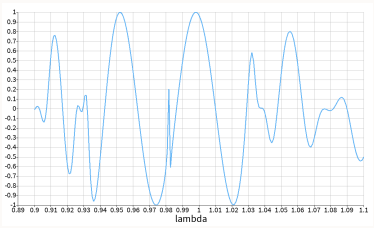


Q: How many separate stretches does the curve spend above +0.5?
Gold: 5 **Source:** written by the expert
Models (3/7 correct):

DS-4.1	DS-V4	Qw-I	GPT-4o	Terra	Gem-P	Gem-F
✗	✗	✓	✗	✗	✓	✓

The expert's stretches are 0.910–0.915, 0.945–0.959, 0.990–1.006, 1.031–1.033 and 1.051–1.060; the wrong answers are 4 (three models) and 6 (one).

③ Physical & functional reasoning · L8 Counterfactual · same spectrum as the item above

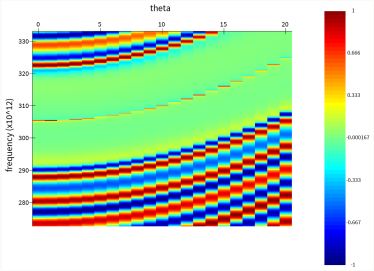


Q: Read where the curve first reaches +1. Make every layer 10 percent thicker - where does it go?
(A) 0.87 (B) 0.93 (C) 0.99 (D) 1.05
Gold: D **Source:** written by the expert
Models (3/7 correct):

DS-4.1	DS-V4	Qw-I	GPT-4o	Terra	Gem-P	Gem-F
✗	✗	✗	✓	✗	✓	✓

The curve first reaches +1 near 0.95; thicker layers stretch every resonance wavelength by the same factor, so it moves to about $0.95 \times 1.1 \approx 1.05$*. Three models choose A, a shift of the same size in the wrong direction.*

③ Physical & functional reasoning · L9 Cross-modal · characterization: transmission map



Q: Which of these match the figure?
(A) The fringe pattern looks identical at every horizontal spot
(B) Oscillatory regions appear at both the top and the bottom
(C) The colour scale covers positive values only here
(D) The vertical axis is labelled in units of 10^12 Hz
(E) A smooth band about 30 THz wide sits at the left-hand edge
Gold: B, D, E (exact set) **Source:** written by the expert
Models (3/7 correct):

DS-4.1	DS-V4	Qw-I	GPT-4o	Terra	Gem-P	Gem-F
✗	✗	✓	✗	✗	✓	✓

The four wrong answers select only B and D and miss E: at the left-hand edge the smooth band spans about 291 to 321 THz.

Figure 4: Example items on one VCSEL (part 2 of 2): an amplitude-reflection spectrum that is first read and then reasoned about, and a map of transmission against angle and frequency. Marks and abbreviations as in Figure 3.

we record the active fraction and left–right and up–down mirror correlations. Timing diagrams are segmented into lanes by ink spanning at least 60% of the plot width; diagrams with text annotations are excluded from pulse questions because leader lines produce false pulses; pulses narrower than 2% of the width are treated as clock ticks. Logic-gate bodies are clustered by glyph shape and each cluster is labelled at full resolution; input counts come from left-edge stubs.

Records. Component classes, variants and visual cues come from parsing the abstract syntax tree of each symbol's rendering script. Modal effective indices come from the simulation metadata and are used only when strictly monotonic in mode order within a device.

E EXPERT REVIEW: PER-TYPE ERROR RATES

Table 9: Share of items the expert corrected or removed, by question type. n is the number of items reviewed and includes items removed afterwards, so it can exceed the released count of a type. Every released item of a type with a non-zero rate was reviewed individually.

Question type	n reviewed	Corrected or removed
Resonance-density comparison	22	0%
Linewidth comparison (2-way)	18	0%
Index counterfactual	10	0%
Gate input count	33	0%
Circuit type (random sample)	27	0%
Profile symmetry	24	0%
Multi-gate circuit (count, all-1, all-0 output)	94	3–4%
Ranking by confinement	11	9%
Peak counting	33	11–13%
Confinement comparison (2-way)	30	13%
Confinement $\rightarrow n_{\text{eff}}$	50	18%
Confinement comparison (4-way)	26	27%
Structure symmetry	21	29%
Linewidth comparison (3-way)	9	33%
Logic-gate identity	64	45%
IC pin count	79	53%
Profile maximum position	44	60%
Device recognition	31	61%

Second review of flagged items. After a first evaluation round, 109 items were returned to the expert: items on which at least six of eight VLMs (six of them evaluated in this paper) gave the same wrong answer (68), degenerate field maps (10), fill-in items missing units (8), and items that no model answered correctly (25). The expert kept 53 unchanged, corrected 11 and removed 45; model answers only selected which items to re-check, and the expert decided every case.

F EVALUATION DETAILS

Models. Evaluation (OpenRouter identifiers): `openai/gpt-5.6-terra`, `deepseek/deepseek-v4.1-flash`, `qwen/qwen3-vl-235b-a22b-instruct`, `deepseek/deepseek-v4-flash-vision-exp`, `openai/gpt-4o`, `google/gemini-3.6-flash`, `google/gemini-3.1-pro-preview`; temperature 0, 4,000 output tokens. Validation and judging: `google/gemini-2.5-flash`, `openai/gpt-4o-mini`. Calibration: `google/gemini-2.5-flash`, `qwen/qwen2.5-vl-72b-instruct`. Rewording: `qwen/qwen-2.5-72b-instruct`, `meta-llama/llama-3.3-70b-instruct`, `deepseek/deepseek-chat` (rewrite at temperature 0.4, equivalence judged at temperature 0 by a different family).

Grading. Choice items request {`"answer": ...`} with a fallback to the last option letter or position word; multi-select is exact-set; fill-in is exact for integers and $\pm 10\%$ otherwise; open-ended items are judged against the expert's explanation.

Reasoning. Every model is queried with reasoning switched off, or at its minimal level where the API forbids switching it off, through the OpenRouter `reasoning` field: `effort: none` for GPT-5.6-Terra and DeepSeek-V4.1-Flash and `enabled: false` for DeepSeek-V4-Flash. Both Gemini models reject any request that disables reasoning and run at `effort: minimal`; Gemini-3.6-Flash then produced no reasoning tokens on any item, while Gemini-3.1-Pro used 542 per item on average. GPT-4o and Qwen3-VL-235B-Instruct do not reason. No reply was empty; three DeepSeek-V4.1-Flash replies reached the output limit and are graded as returned.

Text-only gate. The validation models see the stem and options without the image and give an answer and a confidence. Items with three or more options or a numeric answer are dropped if either model is correct with high confidence; two-option types are checked for above-chance blind accuracy at the type level.

G RESULTS BY TASK TYPE

Table 10: Accuracy (%) on all 39 task types with reasoning switched off (Gemini-3.1-Pro at its minimal level), ordered by the seven-model mean. The best model in each row is shaded and bold (ties share the mark). Model abbreviations as in Table 5.

Task type (level)	n	DS-4.1	DS-V4	Qw-I	GPT-4o	Terra	Gem-P	Gem-F	Mean
Second-most confined (L6)	3	0	0	33	**67**	0	33	0	19
Confinement, 4-way (L6)	3	33	33	0	33	0	**67**	**67**	33
Expert curve relation (L7)	3	33	0	33	33	0	**100**	33	33
Gate input count (L5)	33	**67**	58	55	39	18	18	12	38
Peak count, spectra (L5)	6	33	33	17	33	50	**67**	50	40
Expert counterfactual (L8)	12	50	50	42	17	17	**83**	58	45
Device from render (L3)	12	33	33	25	58	42	**75**	67	48
Peak count, signed curves (L5)	10	60	40	30	40	60	**70**	40	49
Logic-gate identity (L3)	37	51	46	73	43	**78**	46	14	50
Open-ended inference (L7)	12	50	58	33	33	67	67	**75**	55
IC pin count (L5)	79	**90**	**90**	80	75	43	10	4	56
Pulse count per lane (L5)	35	71	57	31	17	63	**83**	**83**	58
Inverted-input count (L5)	15	**100**	87	93	73	33	13	20	60
Structure symmetry (L4)	18	56	50	33	39	72	**89**	83	60
Component symbol (L3)	137	66	65	69	**70**	67	53	36	61
Confinement $\rightarrow n_{\text{eff}}$ (L7)	12	67	42	50	50	58	**92**	75	62
Longest high lane (L6)	35	74	69	29	43	51	**94**	77	62
Gate count per circuit (L5)	34	88	**100**	**100**	88	62	3	3	63
Symbol visual feature (L2)	59	75	71	75	**76**	66	49	41	65
First rising lane (L6)	40	70	65	50	52	68	**82**	78	66
Circuit type (L3)	103	74	82	**85**	80	75	48	29	67
Circuit output, all-1 (L7)	34	82	68	**88**	62	71	56	47	68
Mode-index comparison (L6)	40	65	62	52	60	70	**92**	72	68
Expert cross-modal (L9)	20	55	70	70	50	75	**95**	85	71
Index counterfactual (L8)	10	50	50	80	80	40	**100**	**100**	71
Resonance density, 3-way (L6)	8	50	38	62	**100**	62	**100**	88	71
Circuit output, all-0 (L7)	25	96	92	**100**	76	72	48	40	75
Expert read / count (L5)	20	75	75	75	60	75	**90**	85	76
Odd-one-out symmetry (L4)	28	79	75	79	57	86	**100**	93	81
Plotted quantity (L2)	41	85	83	83	83	80	85	**88**	84
Linewidth, 2-way (L6)	9	67	**100**	78	**100**	89	67	89	84
Profile maximum position (L4)	9	89	89	67	**100**	89	78	78	84
Figure type (L1)	34	88	85	**100**	91	91	85	85	89
Confinement, 2-way (L6)	17	88	88	94	76	88	94	**100**	90
Resonance density, 2-way (L6)	14	86	79	79	93	**100**	**100**	**100**	91
Profile symmetry (L4)	24	92	**100**	96	92	79	96	**100**	93
Wavelength read (L5)	10	**100**	90	80	**100**	90	**100**	**100**	94
Lane count (L1)	50	**100**	94	**100**	**100**	**100**	**100**	**100**	99
Linewidth, 3-way (L6)	1	**100**	**100**	**100**	**100**	**100**	**100**	**100**	100

H DIFFICULTY KNOBS

Table 11: Difficulty knobs. 34 task types (1,003 items) have one; the remaining 89 items (five task types and 30 single-curve quantity-recognition items) have none and are tiered by reasoning steps and open-endedness.

Knob (higher = harder)	Task types	Definition
Separation margin	Confinement, linewidth and resonance-density comparisons; confinement $\rightarrow n_{\text{eff}}$; symmetry and odd-one-out; mode-index comparison	Difference or ratio between the correct item and the runner-up, negated
Count	Lane, pulse, peak, gate-input, inverted-input and pin counts; multi-gate circuits	The gold count, or the gate-chain length
Distractor similarity	Symbol and gate recognition; figure type; expert cross-modal	Number of same-family distractors; number of options; multi-select versus single choice
Clutter	Timing comparisons; circuit type; quantity recognition	Number of lanes; schematic ink density; number of curves
Reading precision	Wavelength reads; index counterfactuals	Axis span over the $\pm 10\%$ tolerance band

Within each task type, items are split into tertiles of the knob (halves below twelve items; direct mapping when the knob takes only two or three discrete values). The knob is fixed when the item is generated and does not depend on any model. Whether a knob in fact controls difficulty is checked afterwards, per task type, against the evaluated models: a knob is validated when mean accuracy decreases from tier to tier and the lowest and highest tiers present differ by at least ten points, and the tier statistics in the main text are computed on the validated types.