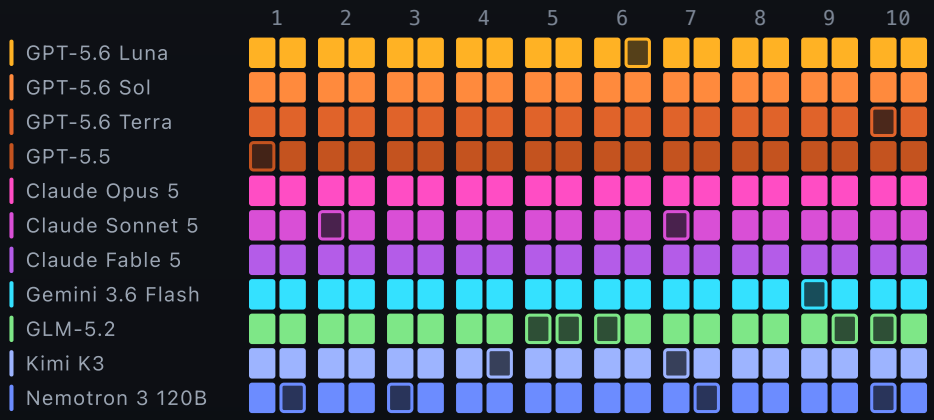
PRISM ARENA BENCH — 11 MODELS × 10 TASKS × 2 ATTEMPTS

# Eleven models get the same ten prompts. Only the scene they describe is different.

Every contender receives a byte-identical prompt and answers with one JSON **scene spec**. The arena validates it against a zod contract, renders it in WebGL, animates it on an externally clocked timeline, and a blind four-judge panel scores the result out of 20.

## SWEEP COVERAGE

11 × 10 × 2 = 220




solid = valid · outlined = ran, invalid · hatched = not run

colour = the model

| VALID | INVALID | SLOTS RUN | SCORED |
|---|---|---|---|
| **203** | **17** | **220** | **110** |
| of 220 generations | of 220 generations | of 220; 0 not yet run | of 110 model×task cells |

A cell is one model against one task. It scores only once its judged attempt validated and every judge returned, which is why scored cells trail valid generations.

HOW A SCORE IS BUILT

A mean of means, in that order. Judgment rows are never pooled into one flat average — that would silently weight tasks by how many judgments happen to exist.

01

4 rubric dimensions, 0–5 each

↓

**one judge total, 0–20**

02

4 judge totals for one task

↓

**task panel mean**

03

10 task panel means

↓

**headline score**

01 STANDINGS

# The ranking, with its own evidence attached

Headline score is the mean across tasks of each task's panel mean. Beside it sits the per-task filmstrip that produced it and the two validity rates, because a rank is not a finding on its own.


FILMSTRIP READINGS AS TEXT

Every number drawn as a bar above is also printed in its own cell. The per-task filmstrip is read cell by cell in the full matrix below, which lists each model against each task.

Standings, numeric form

| RANK | MODEL | FAMILY | HEADLINE / 20 | TASKS SCORED | PER-ATTEMPT VALIDITY | BEST-OF-N VALIDITY | MEDIAN LATENCY |
|---|---|---|---|---|---|---|---|
| 1 | **Claude Opus 5** | anthropic | 18.38 | 10/10 | 100% | 100% | 2m 32s |
| 2 | **GPT-5.6 Sol** | openai | 17.82 | 10/10 | 100% | 100% | 1m 48s |
| 3 | **Kimi K3** | moonshot | 17.77 | 10/10 | 90% | 100% | 4m 46s |
| 4 | **Claude Fable 5** | anthropic | 17.32 | 10/10 | 100% | 100% | 2m 23s |
| 5 | **Nemotron 3 120B** | nvidia | 17.13 | 10/10 | 80% | 100% | 2m 19s |
| 6 | **GPT-5.6 Terra** | openai | 16.68 | 10/10 | 95% | 100% | 1m 28s |
| 7 | **Claude Sonnet 5** | anthropic | 16.55 | 10/10 | 90% | 100% | 3m 11s |
| 8 | **GPT-5.5** | openai | 16.05 | 10/10 | 95% | 100% | 1m 37s |
| 9 | **GPT-5.6 Luna** | openai | 15.93 | 10/10 | 95% | 100% | 1m 38s |
| 10 | **GLM-5.2** | zhipu | 15.82 | 10/10 | 75% | 90% | 2m 27s |
| 11 | **Gemini 3.6 Flash** | google | 15.03 | 10/10 | 95% | 100% | 3m 37s |

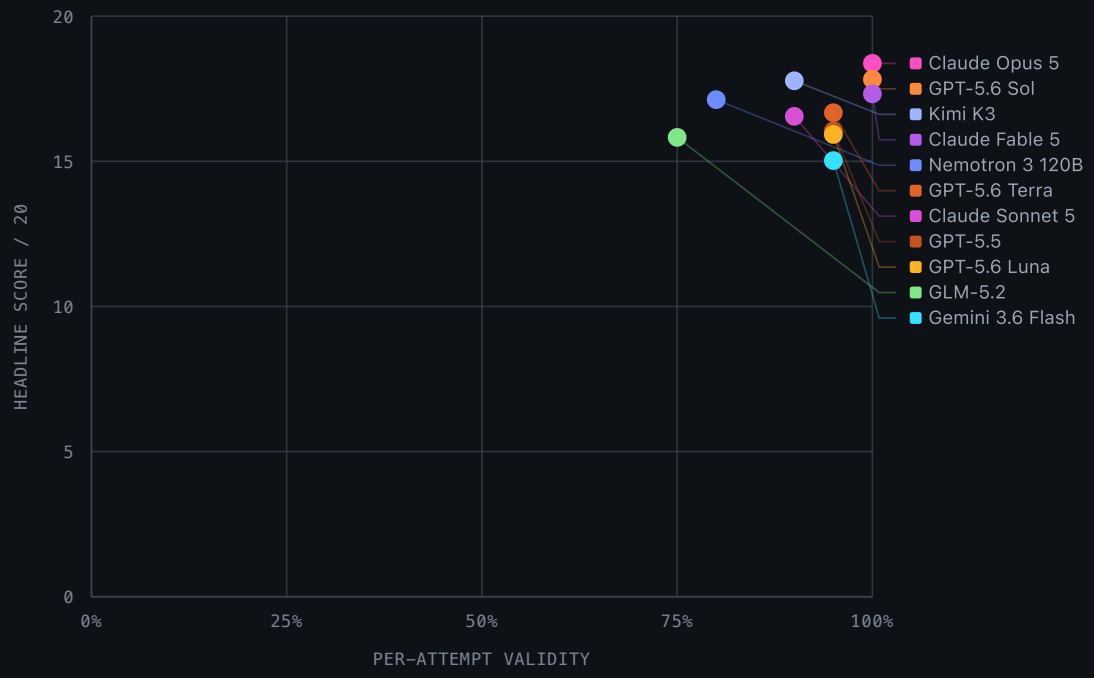
02 RELIABILITY AGAINST PEAK

# Never failing and often winning are different achievements

Per-attempt validity counts every generation. Best-of-N validity counts a task as covered if any attempt parsed. Best-of-N is strictly the more flattering number, so both are shown: the gap between them is how much retries rescued.

## Validity against headline score

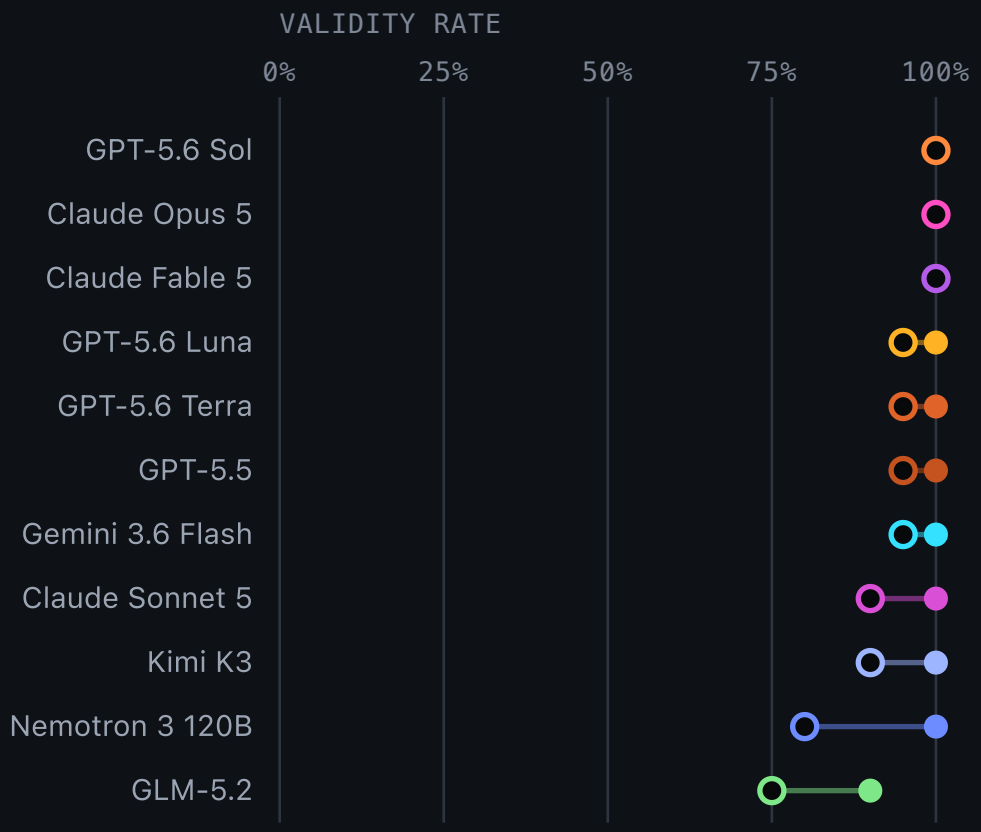
Hollow rings are provisional: not every task is scored yet, so the mean is over fewer tasks.




**Claude Opus 5** leads on both axes in this export: 100% of its generations validated and it holds the top score at 18.38. Reliability and peak are not trading off here.

## What retries rescued

The gap between the two dots is the share of tasks a second attempt saved.




HOLLOW = PER-ATTEMPT · FILLED = BEST-OF-N

VALIDITY AS TEXT

Per-attempt and best-of-N validity by model

| MODEL | VALID / ATTEMPTS | PER-ATTEMPT | TASKS COVERED | BEST-OF-N |
|---|---|---|---|---|

| MODEL | VALID / ATTEMPTS | PER-ATTEMPT | TASKS COVERED | BEST-OF-N |
|---|---|---|---|---|
| **GPT-5.6 Sol** | 20/20 | 100% | 10/10 | 100% |
| **Claude Opus 5** | 20/20 | 100% | 10/10 | 100% |
| **Claude Fable 5** | 20/20 | 100% | 10/10 | 100% |
| **GPT-5.6 Luna** | 19/20 | 95% | 10/10 | 100% |
| **GPT-5.6 Terra** | 19/20 | 95% | 10/10 | 100% |
| **GPT-5.5** | 19/20 | 95% | 10/10 | 100% |
| **Gemini 3.6 Flash** | 19/20 | 95% | 10/10 | 100% |
| **Claude Sonnet 5** | 18/20 | 90% | 10/10 | 100% |
| **Kimi K3** | 18/20 | 90% | 10/10 | 100% |
| **Nemotron 3 120B** | 16/20 | 80% | 10/10 | 100% |
| **GLM-5.2** | 15/20 | 75% | 9/10 | 90% |

03 THE FULL MATRIX

# Every model against every task

One cell per model/task pair. Lightness encodes the task panel mean; a hatched cell is a pair with no score at all. This table is also the text equivalent of the filmstrips in the standings.

| Model \ task | Observatory Array | Signal Cascade | Retrofit Under Rules | Kepler Rig | Crank and Slider | Standing Wave | Phase Change | Jointed Arm | Bloom Sequence | Coupled Pendulums | Mean |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Claude Opus 5 | 19.00 | 17.50 | 18.25 | 19.00 | 18.50 | 18.75 | 18.75 | 17.50 | 19.00 | 17.50 | 18.38 |
| GPT-5.6 Sol | 18.25 | 17.75 | 18.25 | 17.75 | 18.50 | 17.50 | 18.00 | 17.25 | 19.00 | 16.00 | 17.82 |
| Kimi K3 | 17.00 | 18.25 | 16.00 | 19.00 | 17.75 | 19.00 | 17.75 | 18.25 | 18.25 | 16.50 | 17.77 |
| Claude Fable 5 | 18.25 | 17.75 | 16.00 | 18.25 | 16.75 | 18.25 | 17.00 | 16.75 | 17.50 | 16.75 | 17.32 |
| Nemotron 3 120B | 16.50 | 16.25 | 17.00 | 17.25 | 18.25 | 17.75 | 16.25 | 17.50 | 18.00 | 16.50 | 17.13 |
| GPT-5.6 Terra | 18.50 | 15.25 | 18.25 | 18.50 | 16.00 | 16.75 | 17.25 | 18.25 | 11.50 | 16.50 | 16.68 |
| Claude Sonnet 5 | 16.00 | 17.25 | 15.50 | 19.00 | 15.75 | 18.50 | 16.25 | 13.50 | 16.75 | 17.00 | 16.55 |
| GPT-5.5 | 13.50 | 15.75 | 14.25 | 16.75 | 15.75 | 17.75 | 16.25 | 17.25 | 18.50 | 14.75 | 16.05 |
| GPT-5.6 Luna | 15.50 | 14.25 | 18.75 | 17.75 | 17.25 | 18.50 | 15.75 | 16.50 | 10.25 | 14.75 | 15.93 |
| GLM-5.2 | 18.00 | 16.25 | 18.75 | 18.00 | 0.00 | 17.25 | 17.00 | 18.25 | 18.50 | 16.25 | 15.82 |
| Gemini 3.6 Flash | 9.50 | 16.50 | 14.75 | 15.50 | 13.75 | 14.00 | 17.00 | 16.75 | 17.50 | 15.00 | 15.03 |
| Task mean | 16.36 | 16.61 | 16.89 | 17.89 | 15.30 | 17.64 | 17.02 | 17.07 | 16.80 | 16.14 | |

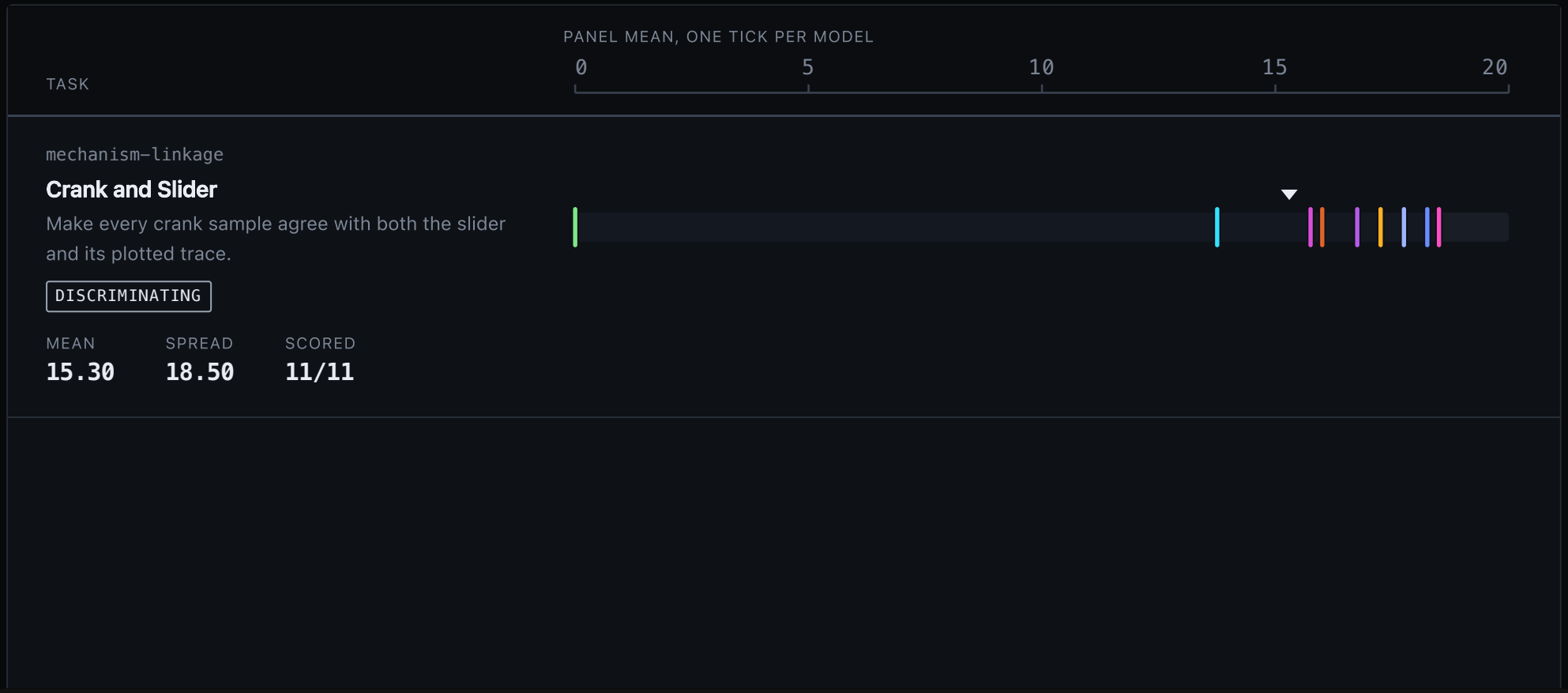
0 20

— no run recorded · valid, panel incomplete × ran, never valid 0.00 underlined = a genuine zero * mean over fewer than all tasks

04 TASK DIFFICULTY

# Which tasks separate the field, and which have stopped working

Each strip plots one tick per model on the shared 0–20 track. A wide strip is a task that discriminates. A task where everyone clusters at the top has saturated and is no longer buying information, however elegant its prompt.



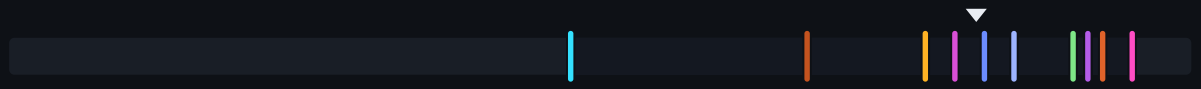
spatial-composition

## Observatory Array

Place a calibrated observatory whose geometry can be checked by eye.

DISCRIMINATING

| MEAN | SPREAD | SCORED |
|---|---|---|
| 16.36 | 9.50 | 11/11 |



stagger-bloom

## Bloom Sequence

Open twelve petals as one ordered, keyframed wave.

DISCRIMINATING

| MEAN | SPREAD | SCORED |
|---|---|---|
| 16.80 | 8.75 | 11/11 |

wave-lattice

## Standing Wave

Encode three wave bases without enumerating nine copies of every action.

NARROW

| MEAN | SPREAD | SCORED |
|---|---|---|
| 17.64 | 5.00 | 11/11 |

kinematic-chain

### Jointed Arm

Resolve a five-link target by compounding parent-local rotations.

NARROW

| MEAN | SPREAD | SCORED |
|---|---|---|
| 17.07 | 4.75 | 11/11 |

constraint-scene-edit

### Retrofit Under Rules

Add a parented instrument without disturbing the commissioned station.

NARROW

| MEAN | SPREAD | SCORED |
|---|---|---|
| 16.89 | 4.50 | 11/11 |

transform-choreography

### Signal Cascade

Five pillars share a compact score whose timing has to land.

NARROW

| MEAN | SPREAD | SCORED |
|---|---|---|
| 16.61 | 4.00 | 11/11 |

orbital-mechanics

### Kepler Rig

Prove two counter-rotating shells with parent space and plotted orbits.

NARROW

| MEAN | SPREAD | SCORED |
|---|---|---|
| 17.89 | 3.50 | 11/11 |

material-transmutation

### Phase Change

Carry five specimens through a checkable five-channel phase cycle.

SATURATED

| MEAN | SPREAD | SCORED |
|---|---|---|
| 17.02 | 3.00 | 11/11 |

harmonic-cascade

### Coupled Pendulums

Choose integer pass counts that all close together.

SATURATED

| MEAN | SPREAD | SCORED |
|---|---|---|
| 16.14 | 2.75 | 11/11 |

TASK STATISTICS AS TEXT

Per-task panel-mean statistics out of 20, ordered by spread.

| TASK | MODELS SCORED | MODELS RUN | MEAN | MIN | MAX | SPREAD | READING |
|---|---|---|---|---|---|---|---|

| TASK | MODELS SCORED | MODELS RUN | MEAN | MIN | MAX | SPREAD | READING |
|---|---|---|---|---|---|---|---|
| **Crank and Slider** | 11/11 | 11/11 | 15.30 | 0.00 | 18.50 | 18.50 | discriminating — 18.5 points between best and worst |
| **Observatory Array** | 11/11 | 11/11 | 16.36 | 9.50 | 19.00 | 9.50 | discriminating — 9.5 points between best and worst |
| **Bloom Sequence** | 11/11 | 11/11 | 16.80 | 10.25 | 19.00 | 8.75 | discriminating — 8.8 points between best and worst |
| **Standing Wave** | 11/11 | 11/11 | 17.64 | 14.00 | 19.00 | 5.00 | narrow — 5.0 points between best and worst |
| **Jointed Arm** | 11/11 | 11/11 | 17.07 | 13.50 | 18.25 | 4.75 | narrow — 4.8 points between best and worst |
| **Retrofit Under Rules** | 11/11 | 11/11 | 16.89 | 14.25 | 18.75 | 4.50 | narrow — 4.5 points between best and worst |
| **Signal Cascade** | 11/11 | 11/11 | 16.61 | 14.25 | 18.25 | 4.00 | narrow — 4.0 points between best and worst |
| **Kepler Rig** | 11/11 | 11/11 | 17.89 | 15.50 | 19.00 | 3.50 | narrow — 3.5 points between best and worst |
| **Phase Change** | 11/11 | 11/11 | 17.02 | 15.75 | 18.75 | 3.00 | saturated — every scored model within 3.0 points near the ceiling |
| **Coupled Pendulums** | 11/11 | 11/11 | 16.14 | 14.75 | 17.50 | 2.75 | saturated — every scored model within 2.8 points near the ceiling |

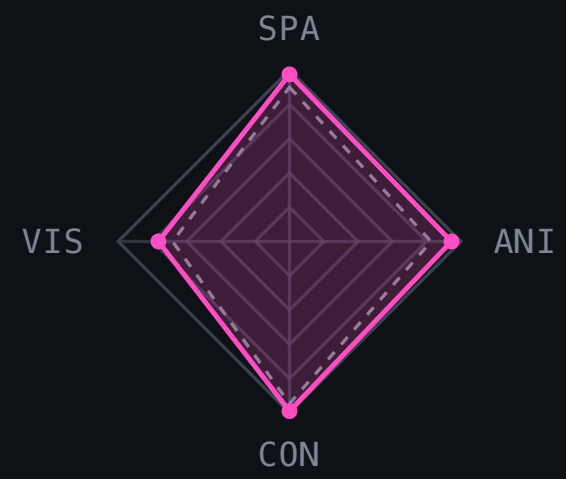
05 ——— RUBRIC SHAPE

# Same total, different shape

Four dimensions, 0–5 each, averaged over the panel and then over tasks. All eleven plots share identical axes and rings, so the silhouettes are directly comparable.

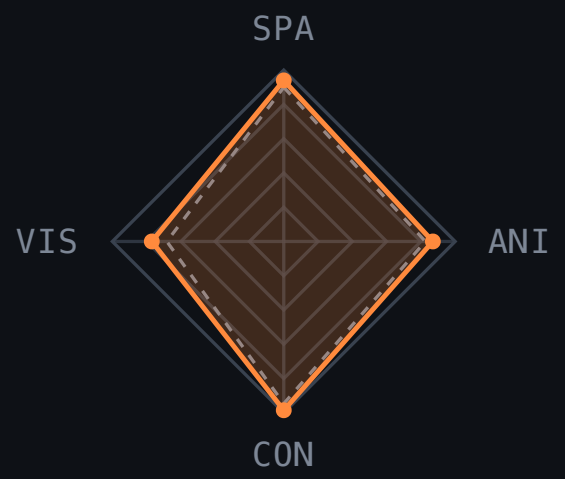
**SPA** Spatial accuracy **ANI** Animation quality **CON** Constraint adherence **VIS** Visual composition -- field mean

Claude Opus 5



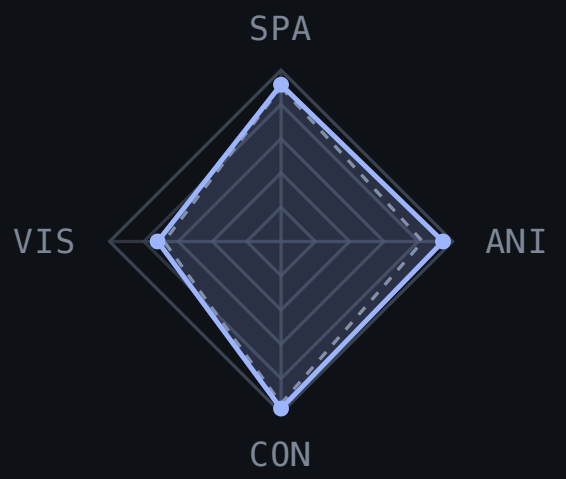

sum 18.38

GPT-5.6 Sol



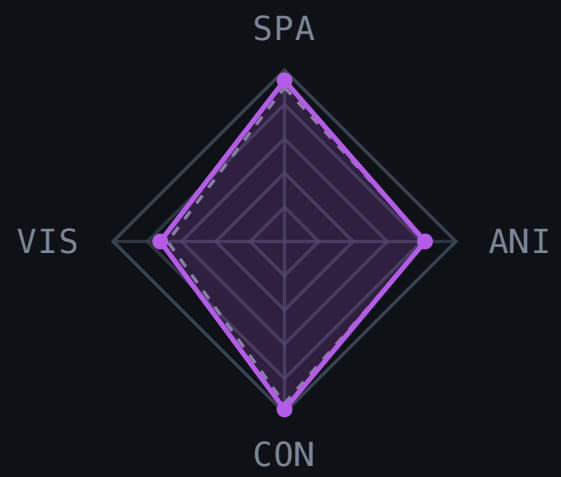

sum 17.83

Kimi K3



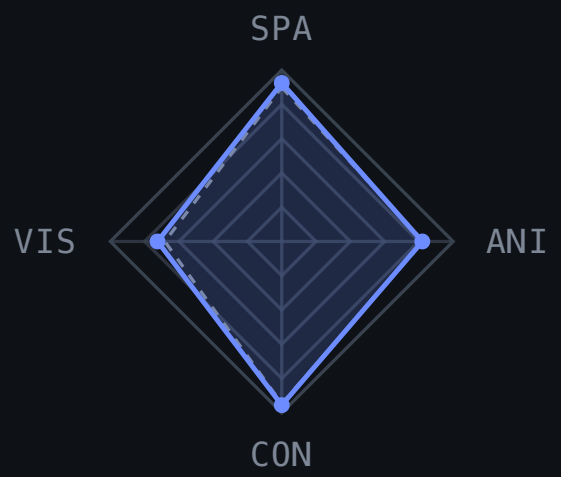

sum 17.78

Claude Fable 5



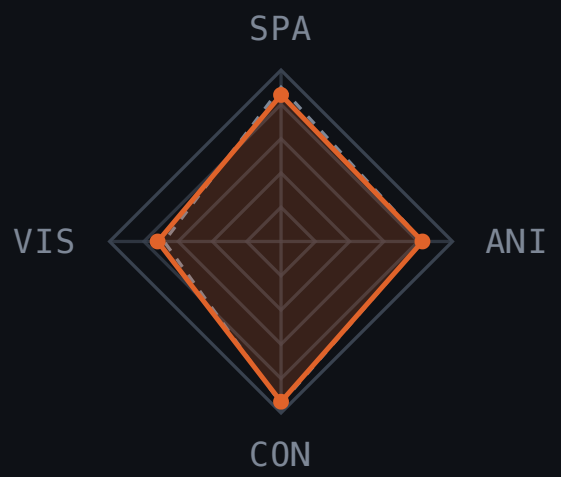

sum 17.33

Nemotron 3 120B



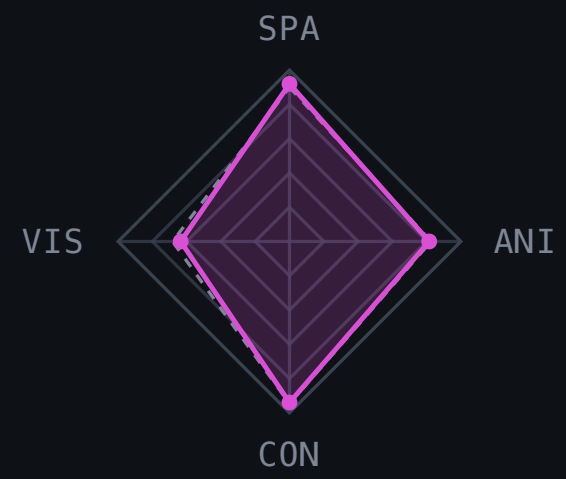

sum 17.13

GPT-5.6 Terra



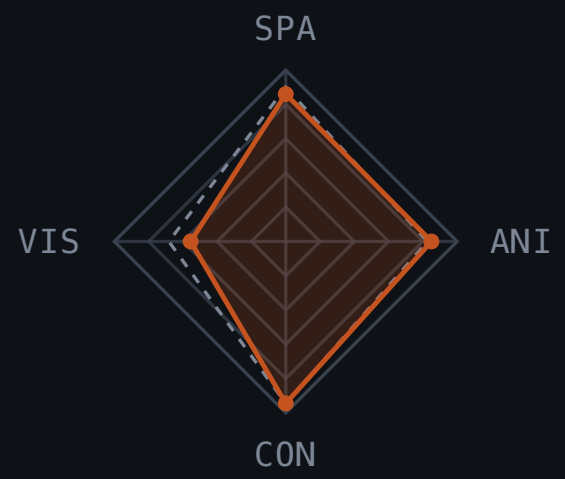

sum 16.68

Claude Sonnet 5



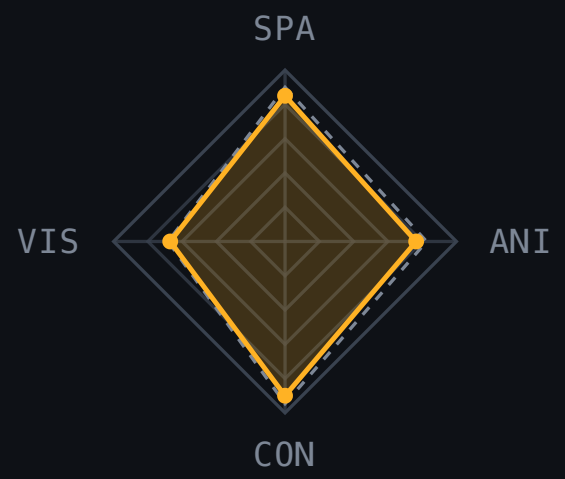

sum 16.55

GPT-5.5

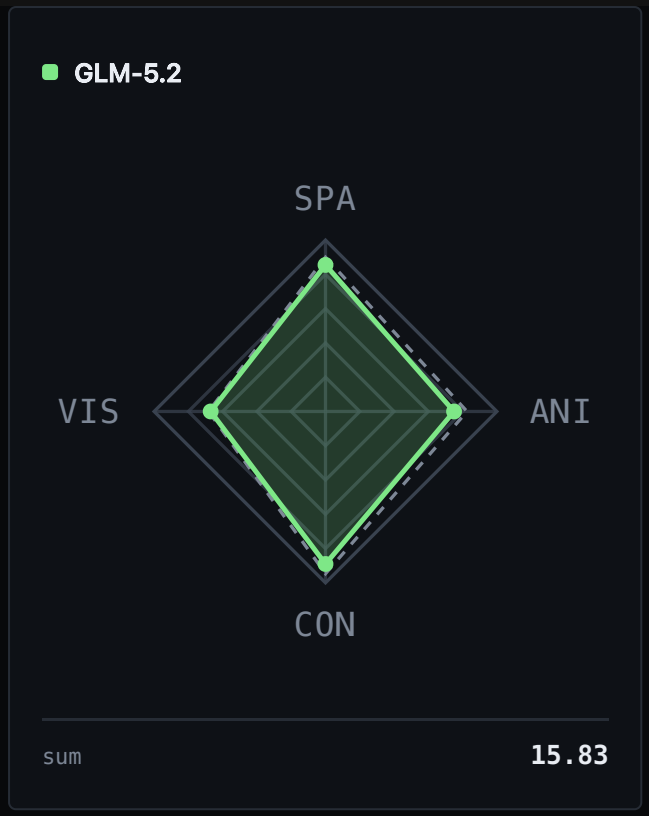


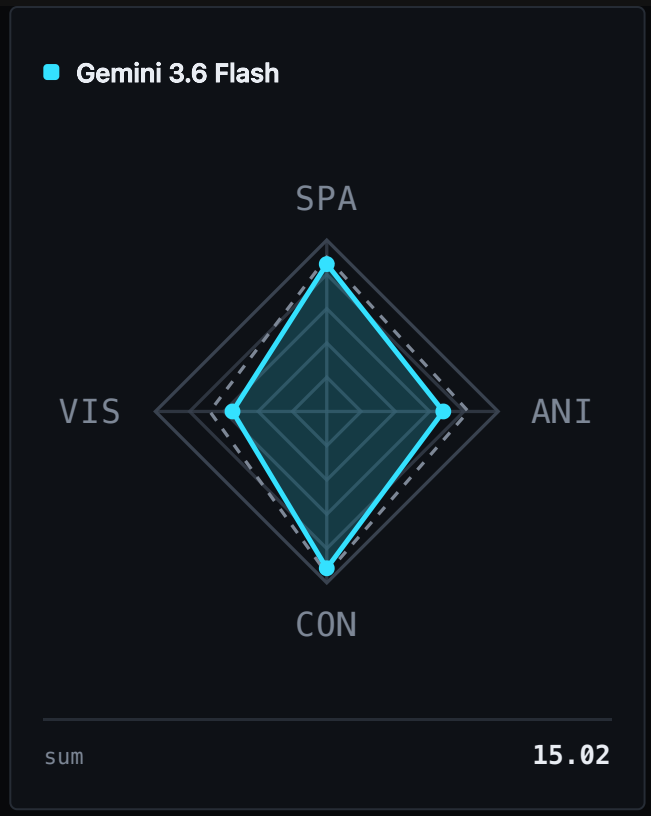

sum 16.05

GPT-5.6 Luna




sum 15.92




RUBRIC SCORES AS TEXT

Mean score per rubric dimension, 0 to 5, averaged over the panel within each task and then across tasks.

| MODEL | SPATIAL ACCURACY | ANIMATION QUALITY | CONSTRAINT ADHERENCE | VISUAL COMPOSITION | SUM / 20 | TASKS IN PROFILE |
|---|---|---|---|---|---|---|
| **Claude Opus 5** | 4.88 | 4.72 | 4.95 | 3.83 | 18.38 | 10/10 |
| **GPT-5.6 Sol** | 4.70 | 4.35 | 4.92 | 3.85 | 17.83 | 10/10 |
| **Kimi K3** | 4.58 | 4.72 | 4.88 | 3.60 | 17.78 | 10/10 |
| **Claude Fable 5** | 4.70 | 4.10 | 4.90 | 3.63 | 17.33 | 10/10 |
| **Field mean** | 4.50 | 4.13 | 4.73 | 3.41 | 16.77 | |

| MODEL | SPATIAL ACCURACY | ANIMATION QUALITY | CONSTRAINT ADHERENCE | VISUAL COMPOSITION | SUM / 20 | TASKS IN PROFILE |
|---|---|---|---|---|---|---|
| **Nemotron 3 120B** | 4.63 | 4.10 | 4.78 | 3.63 | 17.13 | 10/10 |
| **GPT-5.6 Terra** | 4.28 | 4.13 | 4.67 | 3.60 | 16.68 | 10/10 |
| **Claude Sonnet 5** | 4.60 | 4.08 | 4.70 | 3.17 | 16.55 | 10/10 |
| **GPT-5.5** | 4.30 | 4.25 | 4.72 | 2.77 | 16.05 | 10/10 |
| **GPT-5.6 Luna** | 4.25 | 3.83 | 4.50 | 3.35 | 15.92 | 10/10 |
| **GLM-5.2** | 4.28 | 3.75 | 4.45 | 3.35 | 15.83 | 10/10 |
| **Gemini 3.6 Flash** | 4.30 | 3.40 | 4.58 | 2.75 | 15.02 | 10/10 |
| **Field mean** | 4.50 | 4.13 | 4.73 | 3.41 | 16.77 | |

06 — JUDGE DISAGREEMENT

# Four judges, and the distance between them

An averaged number with hidden spread invites false confidence. Every panel mean on this page is the middle of a range, and the range is worth seeing.

## Claude Opus 5

SHARED FAMILY

anthropic/claude-opus-5

| | |
|---|---|
| Mean total awarded | 16.55 |
| Versus consensus | −0.23 |
| Cells scored | 110 |

## GPT-5.6 Sol

SHARED FAMILY

openai-codex/gpt-5.6-sol

| | |
|---|---|
| Mean total awarded | 16.15 |
| Versus consensus | −0.62 |
| Cells scored | 110 |

## Claude Fable 5

SHARED FAMILY

anthropic/claude-fable-5

| | |
|---|---|
| Mean total awarded | 17.19 |
| Versus consensus | +0.42 |
| Cells scored | 110 |

## Gemini 3.6 Flash

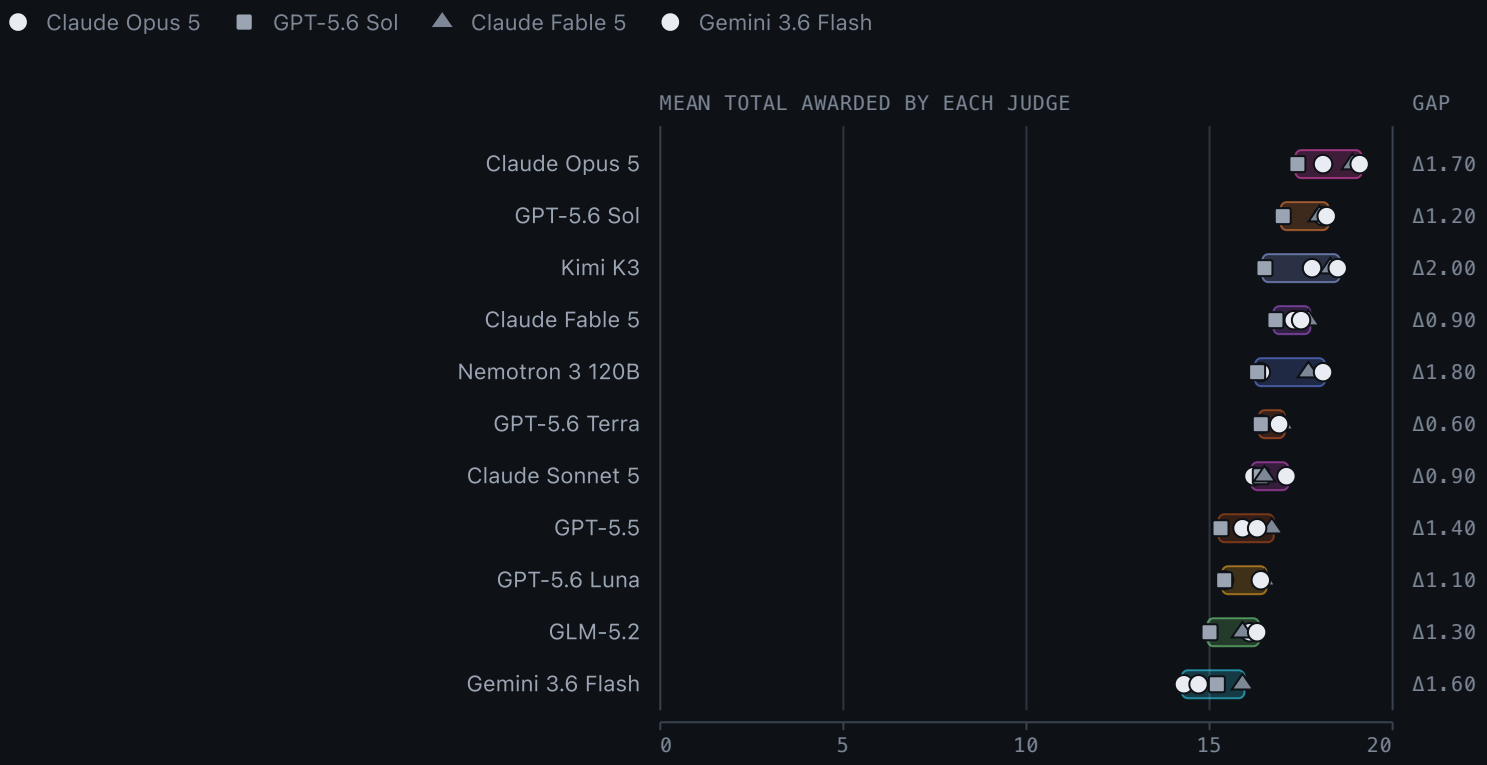
SHARED FAMILY

cursor/gemini-3.6-flash-high

| | |
|---|---|
| Mean total awarded | 17.19 |
| Versus consensus | +0.42 |
| Cells scored | 110 |

## Where the panel splits

The band is the distance between the most and least generous judge for that model.




### JUDGE MEANS AS TEXT

Mean total awarded by each judge per model, out of 20, over cells with a complete panel.

| MODEL | CLAUDE OPUS 5 | GPT-5.6 SOL | CLAUDE FABLE 5 | GEMINI 3.6 FLASH | GAP | CELLS |
|---|---|---|---|---|---|---|
| **Claude Opus 5** | 18.10 | 17.40 | 18.90 | 19.10 | 1.70 | 10 |
| **GPT-5.6 Sol** | 18.10 | 17.00 | 18.00 | 18.20 | 1.20 | 10 |
| **Kimi K3** | 17.80 | 16.50 | 18.30 | 18.50 | 2.00 | 10 |

| MODEL | CLAUDE OPUS 5 | GPT-5.6 SOL | CLAUDE FABLE 5 | GEMINI 3.6 FLASH | GAP | CELLS |
|---|---|---|---|---|---|---|
| **Claude Fable 5** | 17.30 | 16.80 | 17.70 | 17.50 | 0.90 | 10 |
| **Nemotron 3 120B** | 16.40 | 16.30 | 17.70 | 18.10 | 1.80 | 10 |
| **GPT-5.6 Terra** | 16.40 | 16.40 | 17.00 | 16.90 | 0.60 | 10 |
| **Claude Sonnet 5** | 16.20 | 16.40 | 16.50 | 17.10 | 0.90 | 10 |
| **GPT-5.5** | 15.90 | 15.30 | 16.70 | 16.30 | 1.40 | 10 |
| **GPT-5.6 Luna** | 15.40 | 15.40 | 16.50 | 16.40 | 1.10 | 10 |
| **GLM-5.2** | 16.10 | 15.00 | 15.90 | 16.30 | 1.30 | 10 |
| **Gemini 3.6 Flash** | 14.30 | 15.20 | 15.90 | 14.70 | 1.60 | 10 |

## How far apart, in points

MEAN SPREAD
**2.61**

WIDEST SPREAD
**8.00**

CELLS COMPARED
**110**

JUDGES
**4**

> The single widest disagreement is **8.00 points** on Phase Change, where the panel scored the same submission 20.0, 12.0, 19.0, 20.0. A panel mean carries that disagreement inside it whether or not it is shown.

## Judge against judge

Mean absolute gap between each pair of judges scoring the same submission.

| PAIR | MEAN GAP | CELLS |
|---|---|---|
| **Claude Opus 5 ↔ GPT-5.6 Sol** | 1.55 | 110 |
| **Claude Opus 5 ↔ Claude Fable 5** | 1.14 | 110 |
| **Claude Opus 5 ↔ Gemini 3.6 Flash** | 1.23 | 110 |
| **GPT-5.6 Sol ↔ Claude Fable 5** | 1.58 | 110 |
| **GPT-5.6 Sol ↔ Gemini 3.6 Flash** | 1.85 | 110 |
| **Claude Fable 5 ↔ Gemini 3.6 Flash** | 1.15 | 110 |

## Self-preference exposure

Mean total each judge awarded each contender family. Boxed cells are the judge scoring its own family.

| JUDGE | ANTHROPIC | GOOGLE | MOONSHOT | NVIDIA | OPENAI | ZHIPU |
|---|---|---|---|---|---|---|
| **Claude Opus 5** | 17.20 | 14.30 | 17.80 | 16.40 | 16.45 | 16.10 |
| **GPT-5.6 Sol** | 16.87 | 15.20 | 16.50 | 16.30 | 16.02 | 15.00 |
| **Claude Fable 5** | 17.70 | 15.90 | 18.30 | 17.70 | 17.05 | 15.90 |
| **Gemini 3.6 Flash** | 17.90 | 14.70 | 18.50 | 18.10 | 16.95 | 16.30 |

07 IN-DEPTH ANALYSIS

# What the benchmark measures, and where it sits

The score is a systems result: contract compliance, execution, rendering, motion, and blind rubric judgment. The evidence below reads those layers together.

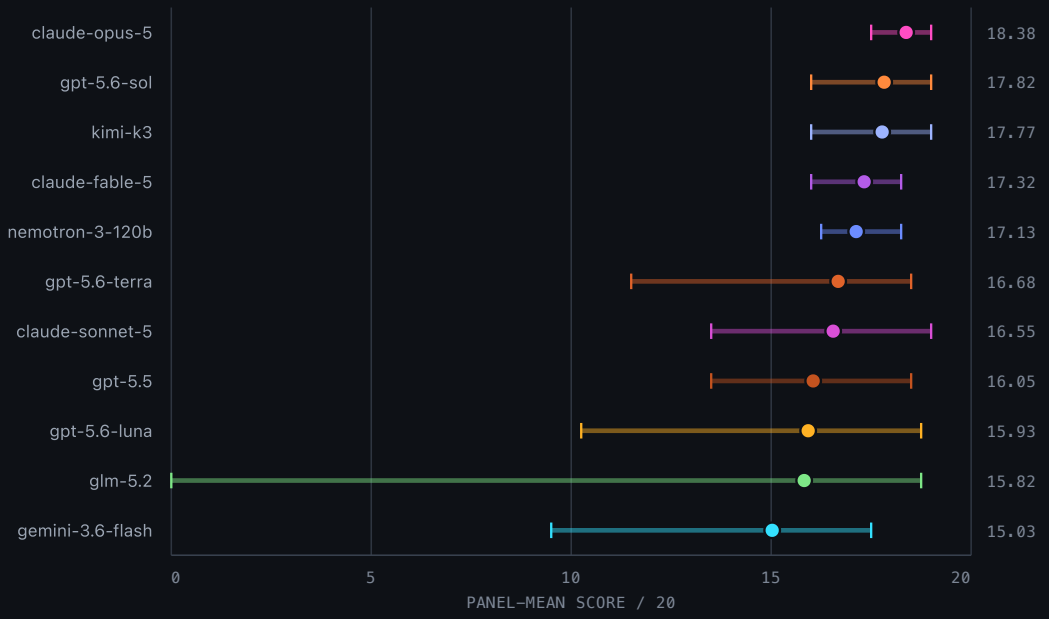
Prism Arena is a systems benchmark for constrained, executable 3D scene-and-motion synthesis by general-purpose language models. The unit of evaluation is the whole pipeline: one natural-language task pack becomes SceneSpec v1 JSON, a single React Three Fiber runtime materialises every submission, Anime.js supplies the only motion, and a label-blinded panel scores the available evidence against a fixed four-dimension, twenty-point rubric. Four judges made 120 batch invocations, yielding 440 contender-level rows. Complete batches include rendered frames; an incomplete batch falls back uniformly to JSON-only evidence. That places Prism between spatial-reasoning benchmarks that ask a model to read a scene and generative benchmarks that ask it to make one, while borrowing the executable-verification discipline of code benchmarks: structural validity and rendered quality remain separately inspectable.

## What the score actually measures

A headline score is the panel-mean total across ten task packs, each scored out of twenty. SceneSpec v1 validates geometry, materials, lights, an ordered timeline of two to eight phases, and the offset-plus-span arithmetic that keeps an action inside its phase. Validity is not a rubric dimension. Because each cell has two recorded runs, the selector sends the earliest valid run to the panel; if neither validates, it sends the earliest invalid run rather than assigning an automatic score. The sole double-invalid cell in this corpus received zero from all four judges. The headline therefore describes the validity-selected artifact, while the separate per-recorded-run validity rate exposes how reliably a model meets the contract.

## Headline score and task range

Dot = mean headline score · whisker = lowest to highest task score.




### HEADLINE RANGES AS TEXT

Headline panel mean and the minimum and maximum task scores, all out of 20.

| MODEL | HEADLINE MEAN | TASK MIN | TASK MAX | TASKS |
|---|---|---|---|---|
| **claude-opus-5** | 18.38 | 17.50 | 19.00 | 10 |
| **gpt-5.6-sol** | 17.82 | 16.00 | 19.00 | 10 |
| **kimi-k3** | 17.77 | 16.00 | 19.00 | 10 |
| **claude-fable-5** | 17.32 | 16.00 | 18.25 | 10 |
| **nemotron-3-120b** | 17.13 | 16.25 | 18.25 | 10 |

| MODEL | HEADLINE MEAN | TASK MIN | TASK MAX | TASKS |
|---|---|---|---|---|
| **gpt-5.6-terra** | 16.68 | 11.50 | 18.50 | 10 |
| **claude-sonnet-5** | 16.55 | 13.50 | 19.00 | 10 |
| **gpt-5.5** | 16.05 | 13.50 | 18.50 | 10 |
| **gpt-5.6-luna** | 15.93 | 10.25 | 18.75 | 10 |
| **glm-5.2** | 15.82 | 0.00 | 18.75 | 10 |
| **gemini-3.6-flash** | 15.03 | 9.50 | 17.50 | 10 |

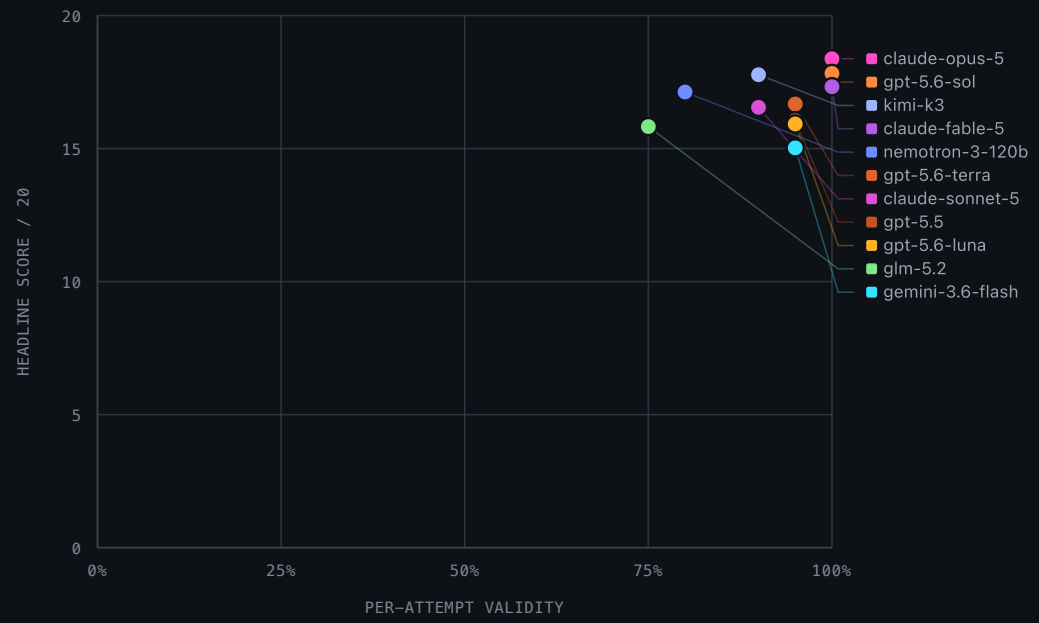
---

## Validity and quality are different abilities

The roster separates cleanly on schema compliance in a way that does not track model prestige. Three models never emitted an invalid spec across twenty recorded runs; the weakest managed exactly three quarters. The common failure is not malformed JSON but omitted required fields — a material without its emissive intensity, an object without its shadow flag — which is a specification-following failure rather than a syntax failure. This is the same distinction JSONSchemaBench draws, and it argues for reporting validity beside quality rather than folding one into the other.

## Validity versus quality

Every point is one model · x = valid attempts / all attempts · y = headline score.




### VALIDITY AND QUALITY AS TEXT

Per-attempt validity and headline panel-mean score out of 20 for every model.

| MODEL ID | VALID ATTEMPTS | ALL ATTEMPTS | VALIDITY | HEADLINE SCORE |
|---|---|---|---|---|
| **gpt-5.6-luna** | 19 | 20 | 95% | 15.93 |
| **gpt-5.6-sol** | 20 | 20 | 100% | 17.82 |
| **gpt-5.6-terra** | 19 | 20 | 95% | 16.68 |
| **gpt-5.5** | 19 | 20 | 95% | 16.05 |
| **claude-opus-5** | 20 | 20 | 100% | 18.38 |

| MODEL ID | VALID ATTEMPTS | ALL ATTEMPTS | VALIDITY | HEADLINE SCORE |
|---|---|---|---|---|
| **claude-sonnet-5** | 18 | 20 | 90% | 16.55 |
| **claude-fable-5** | 20 | 20 | 100% | 17.32 |
| **gemini-3.6-flash** | 19 | 20 | 95% | 15.03 |
| **glm-5.2** | 15 | 20 | 75% | 15.82 |
| **kimi-k3** | 18 | 20 | 90% | 17.77 |
| **nemotron-3-120b** | 16 | 20 | 80% | 17.13 |

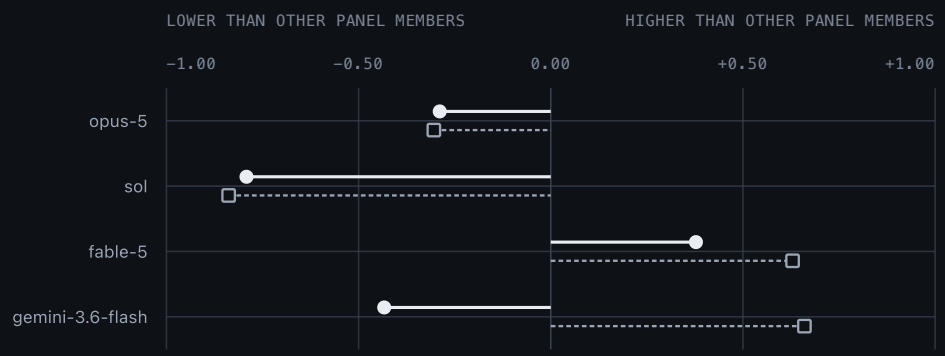
---

## The judges are also the judged

Every model on the panel is also a contender, which is the sharpest threat to these numbers. Label blinding removes names, not stylistic fingerprints. Judge marginal means range from 16.155 to 17.191, a 1.036-point leniency spread that exceeds the 0.55-point first-versus-second headline gap. Within-contender own-model residuals — a narrower slice than the own-family residuals charted below — are small and sign-inconsistent: -0.37 for Claude Opus 5, -1.10 for GPT-5.6 Sol, +0.50 for Claude Fable 5, and -0.43 for Gemini 3.6 Flash. Raw own-family differences confound judge behavior with contender quality, and the fixed family-aligned batches confound family with comparison context. Removing both Anthropic judges still leaves Claude Opus 5 first, but mid-table ranks move. These checks bound the result; they do not calibrate it against humans.

## Judge self-preference residual

Circle = contenders in the judge's family · square = contenders in other families.




The baseline is the other panel members on the same run. Those judges are not family-neutral, so this measures judge-specific deviation rather than absolute family bias.

JUDGE RESIDUALS AS TEXT

Mean judge total minus the mean total from the other judges on the same run. Positive values mean this judge scored the group higher than its peers did.

| JUDGE | JUDGE FAMILY | OWN-FAMILY RESIDUAL | OWN-FAMILY RUNS | OTHER-FAMILY RESIDUAL | OTHER-FAMILY RUNS |
|---|---|---|---|---|---|
| **opus-5** | anthropic | −0.29 | 30 | −0.30 | 80 |
| **sol** | openai | −0.79 | 40 | −0.84 | 70 |
| **fable-5** | anthropic | +0.38 | 30 | +0.63 | 80 |
| **gemini-3.6-flash** | google | −0.43 | 10 | +0.66 | 100 |

## Rendered frames are evidence, and they must match

Judges normally receive two rendered frames per contender alongside the JSON. That only means something if the frame depicts the recorded run being scored. An early version of this pipeline rendered whichever run the arena displayed most recently while the panel scored the first valid run — a mismatch on one hundred of one hundred and ten pairs. Capture and judging now derive the run from one shared selection rule, and the captured run id is recorded so a stale frame cannot be silently reused. Frames are also all-or-nothing per batch: if any contender lacks its full set, the whole batch drops to JSON-only, so no model is judged on text while its peers are judged on pictures. That fallback fired for mechanism linkage when both GLM-5.2 frames failed, so the GLM-5.2, Kimi K3, and Nemotron batch used JSON-only evidence for all four judges — twelve of 440 judgment rows.

## What this panel does not control for

Three of the four judges are invoked with an explicit high reasoning effort. The fourth, Gemini 3.6 Flash, carries its effort in the route name itself and takes no separate effort flag, so the panel is not perfectly matched on that axis. The four judges made 120 batch invocations rather than 440 independent calls, and contender order and family-aligned batch composition were fixed. Their marginal means differ by 1.036 points. Selected-spec character count also correlates positively with panel score in 9 of 10 tasks (mean within-task r=0.38), but the corpus cannot separate genuine scene richness from preference for elaboration. The study has no human calibration set and no repeat judgments. None of these invalidate the descriptive table; all of them sharply limit rank-order claims.

08 NOTICEABLE FINDINGS

# The failures that changed the benchmark

These observations had the largest consequences for the published result, from evidence drift to provider and specification failures.

### Rendered evidence can drift from the scored artefact

CRITICAL

Frame capture drove the interactive arena, which shows a model's newest recorded run, while the judge scored the first valid recorded run. The two disagreed on one hundred of one hundred and ten model-task pairs. Nothing errored; the numbers would simply have been wrong.

### A trivial probe does not prove a route works

HIGH

One provider route answered a one-line health check and then returned "Model not found" on every real generation. Reachability checks that send a toy prompt will certify a dead lane. The smoke test that matters sends the actual task and validates the actual contract.

### Tool access silently changed what was being measured

HIGH

Given a write tool, agentic models wrote the SceneSpec to a file and replied with a prose summary, which scored as malformed. That is a delivery-channel artefact, not a capability gap, and it penalised exactly the models trained hardest to use tools. Running every contender with tools, skills, and rules disabled made the channel uniform.

### Raw family means overstate self-preference

HIGH

Comparing a judge's mean for its own family against its mean for others conflates bias with contender quality. On the full corpus the raw comparison implies that the Anthropic judges favour Anthropic contenders by +0.9 and +0.7 points. Within-contender own-model residuals instead range from -1.10 to +0.50 and change sign across judges. Removing both Anthropic judges leaves the top model first, but mid-table ranks move. The fixed family-aligned batches still confound family with context, so the correct conclusion is that simple family means are misleading —not that self-preference has been eliminated.

### A judge-redundancy result reversed between 59 and 110 submissions

HIGH

On a 59-run subset the two same-family judges looked strongly redundant: their residuals correlated at 0.502 against a cross-family control of 0.114, a 4.4x gap that argued for dropping one of them. On the complete 110-run corpus the same measurement gives 0.559 against a cross-family pair at 0.520 — a difference of 0.039, and no basis for removing a judge. The panel was kept at four. Half a corpus was enough to print a provisional table, but not enough to justify changing the instrument.

### Blank-frame heuristics reject legitimately dark scenes

MEDIUM

A lit-pixel ratio threshold discarded valid renders of genuinely dark scenes. Replacing it with a check for a perfectly uniform image — nothing rendered at all — kept the guard against the real failure without punishing art direction.

### Specification-following, not syntax, separates the field

MEDIUM

Invalid submissions rarely fail to parse. They parse and then omit a required field the prompt named explicitly. Models differ far more in how completely they honour a written contract than in whether they can emit well-formed JSON.

09 PAIN POINTS

# Where the pipeline still costs time and trust

Operational friction is part of the measurement. These are the failure surfaces that still make a full sweep slower, costlier, or harder to reproduce.

### 01 Provider reachability churns mid-sweep

Across one sweep a gateway went offline, an OAuth token expired, and a subscription lapsed. Each looked like a model failure in the logs. The runner now classifies transport faults separately from model faults, records them as retryable, and never lets one become a published zero.

### 02 SQLite has one writer

Running one judge process per panel member is the cheapest parallelism available, but four writers on one database will collide. WAL plus a busy timeout makes it safe; without both, a collision discards a judge call that has already been paid for.

### 03 Judging is the cost and latency long pole

Generation is embarrassingly parallel and finishes in under an hour. Judging is four panel members times ten tasks times three batches, each a slow multimodal call with images attached. Sequential judging took longer than the entire generation sweep.

### 04 No judge is outside the contest

Every model strong enough to judge this rubric is also strong enough to be a contender, so the panel cannot be made independent by selection. The honest response is to measure the contamination and publish it, not to claim it away.

### 05 Capture is coupled to the application's display semantics

Frames come from the real arena, which is what makes them faithful, but it means any change to what the arena chooses to display silently changes the evidence. The recorded run is now pinned explicitly in the URL and stored with the frame.

### 06 Two recorded runs per cell is thin

Best-of-two rewards a model that gets lucky once and gives a noisy per-recorded-run validity rate. It is enough to produce a complete descriptive table, not enough to establish stable ranks or support narrow confidence intervals.

### 07 Effort settings are not uniform across the panel

Reasoning effort is expressed two different ways by two different providers: as a flag for three judges and as part of the route name for the fourth. There is no single field that means "try equally hard" across vendors, so a panel assembled from several providers cannot be matched on effort by configuration alone.

### 08 Half a corpus is enough to mislead a design decision

Intermediate results are available long before a sweep finishes and they are tempting to act on. One redundancy measurement moved by a factor of four between the midpoint and the end, in the direction that would have removed a judge for no reason. Partial data is for monitoring progress, not for changing the experiment.

10 — ROADMAP

## What the next benchmark versions must prove

More attempts, stronger judge controls, human grounding, and deeper tasks turn this first arena into a benchmark whose close calls can carry statistical weight.

## v2

### Statistical power

- Raise recorded runs per cell from two to at least five so validity and within-cell generation variance are estimable.
- Report inter-judge reliability and marginal leniency per rubric dimension rather than a single correlation range.
- Model task, run, and judge variation hierarchically; treat tasks as the sampling unit and refuse to print a strict rank order when uncertainty overlaps.

### Judge integrity

- Recruit at least one capable judge that is not on the contender roster, even if it is weaker.
- Randomise contender order and rebalance batch composition across tasks and judges so family is not confounded with context.
- Publish the leave-one-out residual and leave-one-family-out sensitivity beside every headline score as standing artifacts.

### Human grounding

- Run a human-rater agreement study on a stratified sample and report calibration by rubric dimension.
- Publish a per-judgment evidence manifest with selected run id, prompt hash, frame fractions and hashes, evidence mode, judge route, and renderer revision.
- Publish high-disagreement and evidence-fallback cases, which are more informative than an aggregate agreement rate.

## v3

### Task and scoring depth

- Expand the task registry with calibrated difficulty tiers so ceiling effects are visible.
- Add deterministic seeded replay so a render can be reproduced byte-for-byte from a spec.
- Score cost and latency alongside quality, since a model that needs six minutes per scene is not interchangeable with one that needs ninety seconds.

11 — METHOD AND LIMITS

# What was controlled, and what could not be

### Identical prompts, identical channel

Every contender gets the same task prompt concatenated with the same shared `OUTPUT_CONTRACT`, through the same command. Only `--model` and `--thinking` differ between contenders. Nothing is tuned per model — no bespoke system prompt, no per-vendor formatting hint, no retry-until-parses loop.

### Tools are off, and that is a fairness control

The runner passes `--no-tools --no-skills --no-rules`. Several roster models are agentic coding models: handed a write tool, they satisfy the task by writing the spec to a file and replying with a prose summary, which the validator scores as malformed. That penalises a channel mismatch rather than 3D reasoning. One qwen response went from 390 bytes of prose to 10 KB of real JSON once tools were disabled — same model, same prompt, same rubric.

### The first valid attempt, not the best one

Each model/task cell runs two attempts. The judged attempt is the first one that validates, chosen by validity and then by start time — never by quality. Ranking attempts by quality would require the judge we are economising on, and would leak the outcome into the selection. Call this best-of-N by validity; it is not **the best run**.

### Two validity rates, always together

**Per-attempt validity** is valid generations over all generations. **Best-of-N validity** is tasks with at least one valid attempt over all tasks. Best-of-N is strictly more flattering and hides the reliability signal, so the page never shows one without the other. An attempt here is one **recorded run**; a recorded run may contain one internal transport retry, which is not counted separately.

### Blind, batched, absolute scoring

Judges see redacted submissions under anonymous labels, four per call, against an absolute rubric rather than a ranking. Redaction is structure-aware and never rewrites object ids, so a spec stays coherent while its author stays hidden. Nothing here is decided by vote, which is why the panel does not need to be odd.

### Frames are all-or-nothing per batch

Judges receive rendered frames alongside the JSON. If any contender in a judged batch is missing a frame, the whole batch drops to JSON only. No model is ever judged text-only while its peers in the same call are judged with images.

### CAVEAT Self-preference is present and not solved

Every judge is also a contender, and the panel spans three families: two Anthropic judges, one OpenAI judge, and one Google judge. Blinding hides identity and family-diverse averaging dilutes any single model's bias, but neither removes self-preference. The judge-family cross-tab above makes that risk inspectable.

### CAVEAT Absent is not zero

A missing cell is drawn as a hatched slot and written as an em dash. A genuine zero is drawn as a bar at the floor and written as 0.00. They are different findings and the page never conflates them: a model that was never run does not get credited with a failure it did not have, and a model that scored nothing does not get hidden behind a gap.

# Prior work and methodological grounding

The benchmark sits beside spatial reasoning, executable generation, structured output, and LLM-as-a-judge research. Each citation states why it matters here.

## Related benchmarks

11 references


1. ***Jihan Yang et al., "Thinking in Space: How Multimodal Large Language Models See, Remember, and Recall Spaces," arXiv:2412.14171 (2024).***

   Thinking in Space: How Multimodal Large Language Models See, Remember, and Recall Spaces (https://arxiv.org/abs/2412.14171)
   VSI-Bench is primarily perception and question answering over observed video. Prism Arena starts from a text task, requires the contender to emit a SceneSpec v1 JSON program, then renders and judges the resulting 3D animation. It therefore tests constructive spatial competence rather than recall alone.

2. ***Wufei Ma et al., "3DSRBench: A Comprehensive 3D Spatial Reasoning Benchmark," arXiv:2412.07825 (2024; revised 2025).***

   3DSRBench: A Comprehensive 3D Spatial Reasoning Benchmark (https://arxiv.org/abs/2412.07825)
   3DSRBench asks models to infer answers from visual inputs; Prism asks models to construct an executable scene that visibly satisfies spatial, kinematic, material, and temporal constraints. Prism's output can fail at schema, rendering, composition, or motion even when a model can verbalize the right relation.

3. ***Parker Liu et al., "IR3D-Bench: Evaluating Vision-Language Model Scene Understanding as Agentic Inverse Rendering," arXiv:2506.23329 (2025).***

   IR3D-Bench: Evaluating Vision-Language Model Scene Understanding as Agentic Inverse Rendering (https://arxiv.org/abs/2506.23329)
   This is the closest conceptual neighbor. IR3D-Bench begins with an image and evaluates reconstruction; Prism begins with an authored scene-and-motion brief and compares general-purpose LLMs under one fixed SceneSpec/React Three Fiber/Anime.js harness. Prism also emphasizes temporal choreography and cross-model arena comparison, not inverse rendering accuracy.

4. ***Yuze He et al., "T³Bench: Benchmarking Current Progress in Text-to-3D Generation," arXiv:2310.02977 (2023).***

   T$^3$Bench: Benchmarking Current Progress in Text-to-3D Generation (https://arxiv.org/abs/2310.02977)
   T³Bench chiefly evaluates generated 3D content from NeRF-era text-to-3D systems and highlights multi-object and surroundings failures. Prism evaluates LLM-authored, executable multi-object scenes and animations in a shared renderer; it is about symbolic scene construction and choreography rather than optimizing a continuous 3D representation.

5. ***Weixi Feng et al., "LayoutGPT: Compositional Visual Planning and Generation with Large Language Models," arXiv:2305.15393 (2023).***

LayoutGPT: Compositional Visual Planning and Generation with Large Language Models (https://arxiv.org/abs/2305.15393)
LayoutGPT stops largely at layout planning and evaluates a particular method; it is not a general leaderboard for full rendered animation. Prism's schema controls scene objects, transforms, materials, and animation, and the common runtime turns those plans into directly inspectable artifacts.

6. ***Dongil Yang et al., "LLM Meets Scene Graph: Can Large Language Models Understand and Generate Scene Graphs? A Benchmark and Empirical Study," arXiv:2505.19510 (2025).***

LLM Meets Scene Graph: Can Large Language Models Understand and Generate Scene Graphs? A Benchmark and Empirical Study (https://arxiv.org/abs/2505.19510)
TSG Bench evaluates a symbolic representation without requiring that it execute or render correctly. Prism likewise uses a structured scene representation, but semantic success is tested downstream in a real renderer and extends from static relations to mechanisms, trajectories, stagger, waves, and other temporal behavior.

7. ***Saibo Geng et al., "JSONSchemaBench: A Rigorous Benchmark of Structured Outputs for Language Models," arXiv:2501.10868 (2025).***

JSONSchemaBench: A Rigorous Benchmark of Structured Outputs for Language Models (https://arxiv.org/abs/2501.10868)
JSONSchemaBench asks whether a decoder can produce structurally compliant JSON across many schemas. Prism uses one domain schema and asks the harder downstream question: does valid JSON denote a compelling, instruction-following 3D animation when executed? Schema validity is a prerequisite in Prism, not the final outcome.

8. ***Mark Chen et al., "Evaluating Large Language Models Trained on Code," arXiv:2107.03374 (2021).***

Evaluating Large Language Models Trained on Code (https://arxiv.org/abs/2107.03374)
HumanEval gives code a discrete correctness oracle. Prism shares the executable-verification principle—the generated artifact is actually run—but visual quality and many constraint-satisfaction questions are open-ended, so execution and renderability cannot replace perceptual/rubric judgment.

9. ***Carlos E. Jimenez et al., "SWE-bench: Can Language Models Resolve Real-World GitHub Issues?," arXiv:2310.06770 (2023; ICLR 2024).***

SWE-bench: Can Language Models Resolve Real-World GitHub Issues? (https://arxiv.org/abs/2310.06770)
SWE-bench measures maintenance in existing codebases with test-defined acceptance. Prism measures greenfield declarative scene generation in a fixed runtime. Both reward outputs that survive execution, but Prism's central criteria include spatial composition, material appearance, and motion, which are not fully captured by tests.

10. ***Siqi Chen et al., "SVGenius: Benchmarking LLMs in SVG Understanding, Editing and Generation," arXiv:2506.03139 (2025).***

SVGenius: Benchmarking LLMs in SVG Understanding, Editing and Generation
(https://arxiv.org/abs/2506.03139)
SVGenius targets mostly static 2D vector code and covers multiple SVG tasks. Prism targets 3D, time-dependent scenes in one deliberately constrained scene language, with a common renderer and animation driver. SVGenius is a useful precedent for treating graphics code as both syntax and rendered semantics.

11. ***Adrienne Deganutti et al., "Graphic-Design-Bench: A Comprehensive Benchmark for Evaluating AI on Graphic Design Tasks," arXiv:2604.04192 (2026; author names listed alphabetically).***

    Graphic-Design-Bench: A Comprehensive Benchmark for Evaluating AI on Graphic Design Tasks (https://arxiv.org/abs/2604.04192)

    GraphicDesignBench is the strongest direct precedent for structured graphics and animation, but it focuses on layered 2D professional design and mixes many specialized model modalities and quantitative metrics. Prism is narrower and deeper on programmatic 3D mechanisms and choreography, with every contender producing the same SceneSpec consumed by the same R3F/Anime.js runtime and judged on one shared rubric.

**LLM-as-judge methodology**

5 references

12. ***Lianmin Zheng et al., "Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena," arXiv:2306.05685 (2023), §4.2.***

    Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena (https://arxiv.org/abs/2306.05685)
    This supports LLM judging as a scalable approximation, not as a universal substitute for human validation. The result was established on dialogue, not animated 3D artifacts, and close model pairs are precisely where agreement is weakest. Prism needs a stratified human calibration set before it can claim human-aligned visual quality; panel consensus alone cannot establish that.

13. ***Arjun Panickssery et al., "LLM Evaluators Recognize and Favor Their Own Generations," arXiv:2404.13076 (2024), §§2.2 and 3.5.***

    LLM Evaluators Recognize and Favor Their Own Generations (https://arxiv.org/abs/2404.13076)
    Prism's blind contender identifiers remove a straightforward source-label bias, which is good. But the judges may still infer model family from recurring visual composition, naming, object count, or animation style. Blinding should therefore be reported as a mitigation, not as proof that self-preference has been eliminated; metadata, source code, filenames, and deterministic ordering should also be scrubbed.

14. ***Lianmin Zheng et al., arXiv:2306.05685 (2023); Lin Shi et al., "Judging the Judges: A Systematic Study of Position Bias in LLM-as-a-Judge," arXiv:2406.07791 (2024).***

    Judging the Judges: A Systematic Study of Position Bias in LLM-as-a-Judge
    (https://arxiv.org/abs/2406.07791)
    If Prism uses independent absolute scoring, classic pairwise A/B position bias is less directly applicable and should not be overstated. It still matters wherever multiple candidates, recorded runs, frames, or rubric dimensions are shown together or serially. Candidate and run order should be randomized or counterbalanced, and any future pairwise arena should score both orders.

15. ***Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B. Hashimoto, "Length-Controlled AlpacaEval: A Simple Way to Debias Automatic Evaluators," arXiv:2404.04475 (2024).***

    Length-Controlled AlpacaEval: A Simple Way to Debias Automatic Evaluators
    (https://arxiv.org/abs/2404.04475)
    If judges only see rendered frames, textual verbosity bias is mostly screened off. If they see prompts, rationales, or SceneSpec source, longer specifications may gain an unfair advantage. Even image-only judging can have a visual analogue—rewarding detail density or spectacle over task fidelity—so Prism should separate constraint satisfaction from richness rather than assume the 2D/3D analogue is harmless.

16. ***Pat Verga et al., "Replacing Judges with Juries: Evaluating LLM Generations with a Panel of Diverse Models," arXiv:2404.18796 (2024), §§4.3–4.4.***

Replacing Judges with Juries: Evaluating LLM Generations with a Panel of Diverse Models (https://arxiv.org/abs/2404.18796)

A single judging pass with one rubric prompt provides no estimate of prompt variance. Prism should freeze and publish the prompt for reproducibility, but also run prompt paraphrase/repetition sensitivity checks before treating small rank differences as stable. Absolute rubric scores should be calibrated separately for spatial fidelity, visual quality, and motion rather than assumed transferable from general judging ability.

13 AUTHORS

## Who built and published the benchmark

Prism Arena Bench is an Agentic Engineering research project. Authorship and organisational context are published with the evidence.

- **Eduardo Javier García López**

  Chief Executive Officer, Cofounder
  Agentic Engineering
  eduardo@agenticengineering.agency

- **Luis Fernando Ramos**

  Cofounder — Strategy, Delivery, Sales, Capital
  Agentic Engineering

- **Ricardo Soberanis**

  Cofounder — Consumer Products
  Agentic Engineering

- **Sebastián García-Moreno Zinchenko**

  Cofounder — Product
  Agentic Engineering


**Agentic Engineering** (https://agenticengineering.agency) — Inspectable AI systems for expensive, recurring workflows.

Agentic Engineering designs and ships reliable AI agent systems for ops-heavy teams. We map one costly workflow, build the smallest credible production system, and add evals, observability, human approval, and handoff from day one.

| EXPORT STATUS | SCHEMA | JUDGE PANEL | HARNESS | JUDGMENTS ON FILE |
|---|---|---|---|---|
| complete | results v2 | prism-judge/2.0 | prism-arena/0.3.0-contract-explicit ×220 | 440 at current version<br>440 total on record |

| FIRST RUN | LAST RUN | EXPORTED |
|---|---|---|
| 2026-07-25 19:09:18Z | 2026-07-26 01:36:23Z | 2026-07-26 18:25:52Z |

Read the companion field note: Prism Arena Is Not a Leaderboard (https://agenticengineering.agency/field-notes/prism-arena-is-not-a-leaderboard).

**Prism Arena Bench.** Printed from the published dashboard at https://prismarena.agenticengineering.agency/. Every figure and table in this document is rendered from the same export the page loads; the machine-readable corpus is https://prismarena.agenticengineering.agency/data/results.json.
Regenerate this PDF with `node scripts/print-paper.mjs`.