SPHYR: AN ENVIRONMENT FOR SPATIAL-PHYSICAL REASONING WITH A SIMULATION VERIFIER

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning from verifiable rewards needs verifiers, and the available ones are narrow: exact match, unit tests, or a learned reward model. Each assumes a task has a checkable canonical answer. Many design problems do not. We present **SPhyR**, an environment for spatial-physical reasoning whose verifier is a physics simulation rather than a comparison to a stored answer. An episode presents a 2D domain with prescribed loads and supports and a masked material distribution; the agent submits a completion, and the environment solves it as a linear elastic structure and scores it against a topology optimizer re-run on exactly the masked region. A completion that routes the load differently from the reference but just as efficiently scores just as well. Across the $24{,}325$ valid completions in our results, 72.7% carry their load correctly while differing from the reference somewhere; a string-matching verifier rejects them all, and graph connectivity disagrees with compliance on $1{,}079$ of them, so a cheaper structural proxy will not substitute. We show the two views rank models differently and in both directions, that the environment's reward is not reachable by retrieval, and that a benchmark of this shape can rank models by grid-serialisation ability rather than by reasoning unless that is measured separately. The environment is released in the OpenEnv interface, single-step or with simulator feedback between attempts[1].

1 INTRODUCTION

Large language models reason well over language and logic, but their ability to reason about physical systems and spatial structure remains underexplored (Zhang et al., 2025). Existing benchmarks probe visual perception, intuitive physics, contact and stability, or textual physical commonsense (Section 2). None ask whether a model can decide how a force should be carried: where material must go so that a load reaches a support efficiently. That question is fundamental to engineering and design, and it has a property that makes it awkward to benchmark — it has no single right answer.

We present **SPhyR**, an environment for spatial-physical reasoning whose verifier is a physics simulation rather than a comparison to a stored answer. An episode presents a 2D domain with prescribed loads and supports and a masked material distribution (Figure 2, Appendix A). The agent submits a completion; the environment solves it as a linear elastic structure and scores it against a topology optimizer re-run on exactly the masked region, under the same material budget. A completion that routes the load differently from the reference but just as efficiently scores just as well.

This is necessary rather than merely preferable. Across the $24{,}325$ valid completions in our results, 72.7% carry their load correctly while differing from the reference somewhere; a string-matching verifier rejects them all. Nor does a cheaper structural proxy substitute: compliance disagrees with graph connectivity on $1{,}079$ predictions, 486 of them connected on the cell graph yet unable to transmit load.

Contributions

- **A verifier that needs no canonical answer.** We score a completion by simulating it and comparing it against an optimizer re-run on exactly the masked cells (Section 4.1).

[1]Anonymised repository with the environment, evaluation code, both dataset versions and every stored completion: `https://anonymous.4open.science/r/SPhyR-053C`

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.

- **An environment whose reward is actionable, not only a scored dataset.** SPhyR is released behind the OpenEnv interface, single-step or with simulator feedback between attempts (Section 5). Against a control that redraws the same prompt blind, the verifier's report is worth $+12.8$ Topology Score over the same number of attempts, winning 35 of 50 samples ($p = 0.0001$; Section 6), and it never says which cells are wrong.
- **Evidence that the reward measures something reconstruction cannot.** Ranking by cell-wise agreement and by verified structural quality disagree, in both directions and on individual samples (Section 7). Exact Match is 0.0 for every model on every full-structure task while Load Carrying spans 33% to 99%.
- **A reward that retrieval cannot reach.** The first release could be largely solved by copying a neighbouring solved instance — $1{,}294$ of its $1{,}295$ boundary conditions have another a single cell away — so we regenerate it by rejection sampling that holds instances apart and report three current models on the result (Section 8); the same open-weight model loses 23.7 points of verified score at identical grid size.
- **A measurement hazard this class of environment creates.** Because an answer is a serialised grid, a score conflates structural reasoning with the ability to emit several hundred cells in an exact shape: at 20×20 one model leads another by 41 points unconditionally and trails it by 11 conditioned on well-formed output (Section 8), so we report grid validity as a first-class metric.

The environment, dataset and evaluation harness are released.[2]

2 Related Work

Verifiable rewards and what can be verified Reinforcement learning from verifiable rewards replaces a learned reward model with a programmatic check of correctness, removing the reward model as a source of error and as a surface to game (DeepSeek-AI et al., 2025b; Lambert et al., 2024). Its reach is set by what can be checked: mathematics admits a final answer and code admits tests to execute (Cobbe et al., 2021; Chen et al., 2021), and even step-level supervision is graded against a known solution (Lightman et al., 2024). Where no such check exists the substitute is a learned reward model, which reintroduces what the verifiable setting was meant to avoid: optimising an imperfect proxy degrades true performance (Gao et al., 2023; Skalse et al., 2022). Between these lies a class of problems with a decisive notion of correctness but no canonical answer, because many solutions are equally correct. Design problems are of this kind, and topology optimization is a clean instance.

Environments for language agents Environments exposing a reset/step interface and a scalar reward have become the standard artefact for training and evaluating language agents (Brockman et al., 2016; Zhou et al., 2024; Jimenez et al., 2024). Their verifiers are, with few exceptions, exact-match or execution-based, inheriting the canonical-answer assumption above. SPhyR contributes an environment whose reward is a physics simulation: deterministic, needing no reference answer, and carrying enough information for an agent to act on rather than be scored by (Section 6).

Physical, spatial and structured reasoning benchmarks Existing benchmarks probe visual reasoning about objects and relations (Johnson et al., 2016; Yi et al., 2020), the plausibility of physical events (Riochet et al., 2020; Bear et al., 2022), block stability (Groth et al., 2018), interactive physics puzzles (Bakhtin et al., 2019), and textual physical reasoning (Bisk et al., 2019; Zhang et al., 2025); ARC-AGI-2 (Chollet et al., 2025) shows the value of structured 2D representations for isolating reasoning, without physics. None require a model to determine material arrangements under explicit load and support constraints (Appendix A).

Machine learning for topology optimization Prior work embeds structural knowledge into fast surrogate solvers (Banga et al., 2018; Rawat & Shen, 2019; Zhang et al., 2020); we instead evaluate general-purpose LLMs zero-shot. That literature also bears on *scoring*: learned solvers are assessed by pixel-wise agreement with a reference design (Banga et al., 2018; Rawat & Shen, 2019), defensible for a solver imitating one optimizer but not for a benchmark of reasoning, where a different but

[2] `https://anonymous.4open.science/r/SPhyR-053C`; the public repository and dataset links are restored in the camera-ready.

equally efficient load path is penalised for being different rather than worse. To our knowledge SPhyR is the first benchmark here to evaluate against a re-optimized design rather than a fixed answer.

3 THE SPHYR TASK

Task and data Topology optimization determines an optimal material distribution under prescribed forces and supports (Bendse & Sigmund, 2004). We repurpose such instances as reasoning tasks: a model must predict a plausible material distribution from forces, supports and boundaries alone, without access to simulation. SPhyR-v1 comprises $1{,}296$ samples on 10×10 grids, generated with the density-based SIMP solver Millipede (Michalatos & Kaijima, 2024) at 10 iteration steps, sweeping load and support positions over scenarios typical of structural building design (load from the top, support on the bottom, 3 to 6 cells wide; Appendix E for solver parameters). Section 8 audits this construction and regenerates it.

Input and output Inputs are a serialised grid in which **L** marks an applied load, **S** a support, and **V** a masked cell whose material state must be predicted; known cells carry their values. The expected output is the same grid with every `V` replaced, preserving shape and formatting. No simulator or solver is available to the model (Appendix A for the prompt template).

Task variations Variations differ in the extent of the masked region, at two difficulty levels: *easy*, where the ground truth is binary, and *hard*, where it is continuous in $[0, 1]$. We mask N random cells ($N \in \{1, 5, 10\}$), N random rows or columns ($N \in \{1, 3\}$), or the full structure, giving eight subjects per difficulty (Appendix F).

4 EVALUATION

We report four families of metrics. Three compare a prediction to the reference answer: **reconstruction** metrics of cell-wise agreement (Exact Match and difference ratios), **topological validity** metrics (grid well-formedness, load-support connectivity and its gravity-aligned variant, isolated clusters), and a **physics-approximating** load-path efficiency by gravity-aligned Dijkstra traversal (Appendices B, G.2, G.3, G.1). The fourth, developed next, asks whether the prediction *works* as a structure; it is the environment's reward.

4.1 SIMULATED STRUCTURAL VERIFIER

A topology optimization problem has no unique solution, so cell-wise agreement with one reference measures the wrong thing twice over: a structure routing the load differently but just as efficiently is marked wrong, while one matching the reference almost everywhere but severing the load path is marked nearly right. Across the $24{,}325$ valid completions in our results, 72.7% carry their load correctly while differing from the reference somewhere. We therefore evaluate each prediction by simulating it and scoring it against an optimizer re-run on the same masked region.

Structural model Each grid cell becomes a unit-square bilinear (Q4) plane-stress finite element whose Young's modulus follows the SIMP law (Bendse, 1989), $E(\rho) = E_{\min} + \rho^p(E_0 - E_{\min})$ with $p = 3$, $E_0 = 1$ and $E_{\min} = 10^{-9}$, so $\rho = 0$ leaves a negligible but non-singular stiffness and $\rho = 1$ recovers full material stiffness. Load cells carry a unit resultant over their nodes and support cells are clamped, both solid by definition; boundary conditions are read from the reference sample, so a prediction cannot improve its score by relocating them. Solving $\mathbf{K}\mathbf{u} = \mathbf{f}$ yields the **compliance** $c = \mathbf{f}^\top \mathbf{u}$, the work done by the load and hence the inverse of stiffness.

Compliance also decides physically whether a structure carries its load at all: a prediction whose compliance exceeds 10^3 times that of the fully solid grid transmits force through void stiffness rather than material. This is strictly sharper than connectivity on the cell graph, with which it disagrees on $1{,}079$ predictions — 486 the graph calls connected but that cannot transmit load, cells touching only at a corner sharing one node and forming a hinge, and 593 that carry load while the graph calls them disconnected. Both directions occur in the binary block as often as in the continuous one, so neither is an artefact of intermediate densities. The threshold is not delicate either: compliance ratios are

bimodal, and fewer than 0.15% of valid completions lie between $10^{2.5}$ and 10^5 times the reference design's compliance, so any cutoff within those two decades classifies the same completions.

Reference design For each sample we re-run a SIMP optimizer (Bendse & Sigmund, 2004; Andreassen et al., 2011) on *exactly* the masked cells, every unmasked cell pinned to its prescribed value, minimizing compliance subject to the reference material budget. Because a single run on a 10×10 mesh has grey local minima, we keep the stiffest of several candidates within budget: cold and warm-started runs under penalization continuation, their volume-preserving binarizations, a discrete swap refinement, and the reference answer itself. Seeding with the reference guarantees the design a prediction is measured against is never worse than the answer the dataset ships.

Scores For a prediction of compliance c and volume fraction v, against a reference design of compliance $c^\star$ at budget v_{ref}, we report structural efficiency $\text{SE} = \min(1, c^\star/c)$, material efficiency $\text{ME} = \min(1, v_{\text{ref}}/v)$, and their product, the **Topology Score** $\text{TS} = \text{SE} \cdot \text{ME}$, all zero for a prediction carrying no load. Both factors are necessary: stiffness alone is maximized by filling the grid solid and thrift alone by building nothing. Empirically, among load-carrying predictions 68.5% of those spending more material than the reference beat the budget-constrained reference design, as they should, against 5.3% at equal material and 0.3% at less; structural efficiency alone would read overspend as quality (Appendix B). The prompt asks for as little material as possible but does not state v_{ref}, so material efficiency penalises overspend against a budget the agent must infer from the visible structure; in the feedback setting of Section 5 the budget is reported back explicitly.

Validation On `full_easy` the reference answer scores 0.93, the same answer with two cells flipped 0.80, ten flipped 0.48, a solid grid 0.29 and an empty grid 0.00. The reference does not reach 1.0 because the optimizer finds a slightly stiffer structure for the same material; that margin is the headroom a model is measured against.

5 The Environment

The verifier is what makes SPhyR usable as an environment rather than only as a scored dataset, and we release it as one.

Interface An episode is one completion task. `reset()` returns the masked grid and its prompt, with the sample's shape, cell alphabet, masked-cell count and load-case rotation; `step()` takes the completed grid as text (prose or code fences are tolerated, the last grid block being the answer), verifies it, and returns the full metric set with the Topology Score as scalar reward; an unparsable or misshapen grid scores zero. We implement the OpenEnv interface, and the verifier is deterministic: the same answer scores the same every time, which a learned reward model cannot promise.

Feedback between attempts Episodes are single-step by default. With feedback rounds enabled, the agent is shown how its structure behaved — whether it carried its load, its stiffness against the best design at that budget, its material use relative to the budget, and how many clusters of its material are attached to nothing — and asked to revise (Appendix D). The feedback is deliberately physical and incomplete: it never says which cells were wrong, so an agent that improves under it was not handed the answer.

6 The Verifier's Signal is Actionable

If a reward only ranks agents, it is a benchmark; to be worth optimising against it must carry information an agent can act on, and a trajectory across rounds tests that directly.

Design Each sample is run twice over. In the **revision** condition the agent answers, the completion is verified, and the agent is shown how the structure behaved and asked to try again, for three revisions. In the **control** condition it gets the same four attempts but is shown nothing between them: the prompt is simply re-posed. Both take the best of four. The control is what makes the comparison interpretable, since best-of-several beats a single attempt whether or not the attempts learn anything; only the gap between conditions separates a verifier that *informs* from one that merely *scores*.

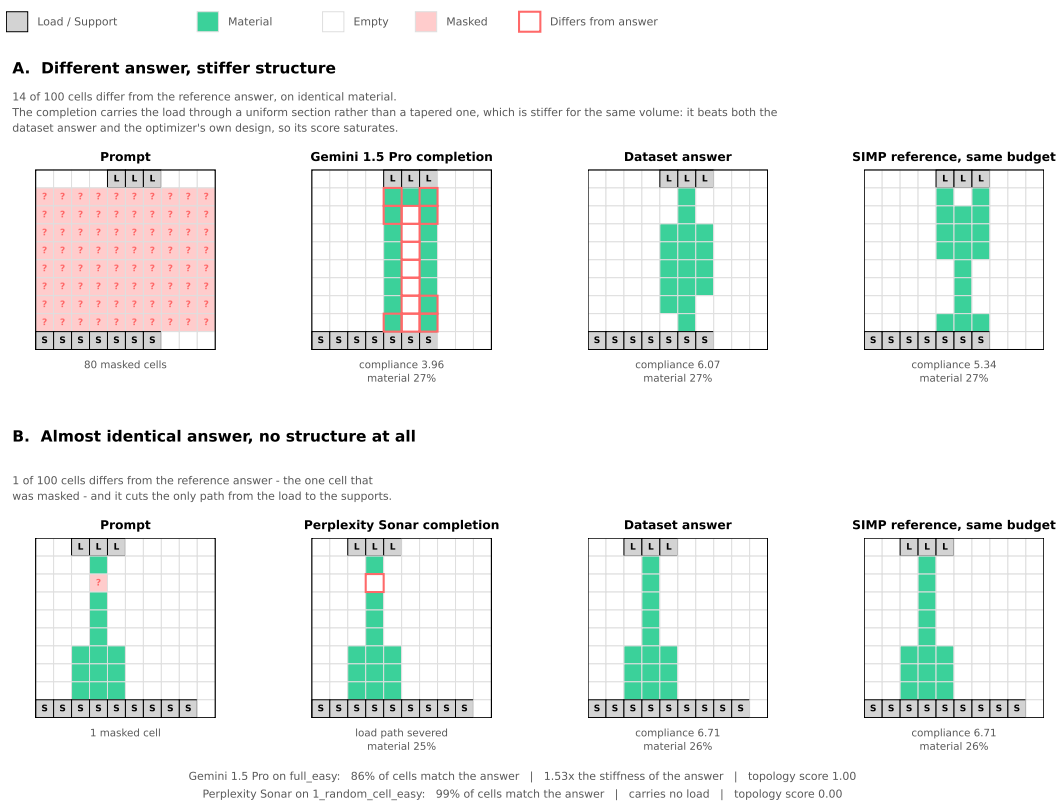

Figure 1: **What cell-wise comparison sees against what the simulation sees. (A)** Gemini 1.5 Pro differs from the reference in 14 of 100 cells on identical material and is $1.53\times$ stiffer. **(B)** Perplexity Sonar matches 99 of 100 cells; the one it missed was the masked one, and it severs the only load path. Column four is the design our optimizer finds for the same hole and budget.

Result On `full_easy` at 10×10 with Qwen3.8, a 27B-parameter open-weight model served locally at 4-bit quantisation (17 GB; settings and the feedback prompt in Appendix D), over 50 samples, feedback is worth $+12.8$ Topology Score over blind resampling (paired standard error 4.5, 95% confidence interval $[4.0, 21.5]$), winning 35 of 50 samples against 9 losses (sign test $p = 0.0001$). First attempt to best-of-four is $15.9 \rightarrow 37.9$ with feedback and $19.7 \rightarrow 25.1$ without. The two conditions are indistinguishable at the first attempt, before any feedback exists ($p = 0.60$), so the difference is attributable to the feedback rather than to the arms differing.

The gain concentrates in the first revision — the score roughly doubles from round 0 to round 1, then plateaus — and the mechanism is visible in the material budget: first attempts use about half the material allowed, and after one report of stiffness and material use the model moves to 82% of budget, acting on a quantity it cannot compute for itself.

What this does not show One model on one subject: this establishes that the reward is actionable, not how far that scales. Feedback also costs grid validity, which falls from 90% to 74% as the context comes to hold two grids (Table 5).

7 Experiments and Results

We evaluate a broad set of contemporary models zero-shot: GPT-3.5 (Brown et al., 2020), GPT-4.1 (OpenAI et al., 2024a) and GPT-4o (OpenAI et al., 2024b); Claude 3.7 Sonnet (anthropic, 2025a) and Claude Opus 4 (anthropic, 2025b); Gemini 1.5 Pro (Team et al., 2024) and Gemini 2.5 Pro (Comanici et al., 2025); DeepSeek-V3 (DeepSeek-AI et al., 2025a) and DeepSeek-R1 (DeepSeek-AI et al., 2025b); and Perplexity Sonar (Team, 2025a) and Sonar Reasoning (Team, 2025b). Each receives structured descriptions of boundary conditions, forces and supports, with no simulation tools, on

Table 1: **Zero-shot performance on SPhyR-v1.** Averages run over the eight subjects of each difficulty. On the full-structure task (Appendix I) Exact Match is 0 for every model while Load Carrying spans 33–99%, the disagreement Section 7 examines. Bootstrap 95% half-widths are 1.3–2.8 points on the averages and up to 9.5 on the full-structure rows (Section 10).

	Easy (binary)				**Hard** (continuous)			
Model	EM ↑	Valid ↑	TS ↑	LC ↑	EM ↑	Valid ↑	TS ↑	LC ↑
Average over all eight subjects								
GPT-4.1	6.75	99.25	74.55	94.00	1.88	98.38	70.80	97.12
Claude Opus 4	**32.12**	**100.00**	**85.79**	96.00	**28.25**	99.62	80.47	93.38
Gemini 2.5 Pro	29.38	99.25	85.67	97.25	26.75	**99.75**	**83.00**	**98.00**
DeepSeek-R1	28.62	99.38	80.00	**98.62**	24.12	98.88	79.36	96.75
Perplexity Sonar	15.12	92.12	59.69	74.75	3.50	84.88	56.71	78.75
Non-reasoning baselines, Topology Score only								
Solid / void / random	71.9 / 44.2 / 59.3				60.8 / 39.5 / 64.8			

100 random examples per task variation through public APIs (Appendix N). Table 1 reports the five strongest; per-task results for all metrics and models are in Appendices I and J.

Non-reasoning baselines To establish what such scores mean absolutely we evaluate three policies that do no reasoning, on the same samples. Filling every masked cell solid reaches 100% Load Carrying but pays in material, settling at 71.9 Topology Score on easy tasks and 60.8 on hard; filling void collapses to 44.2 and 39.5. Uniform random fill is the more uncomfortable comparison: 59.3 on easy tasks against Perplexity Sonar's 59.7, and 64.8 on hard against Sonar's 56.7, within six points of GPT-4.1's 70.8. A model that cannot be separated from random filling is a stronger statement of failure than any reconstruction metric makes, and it is visible only because the structural metrics score a prediction on what it does rather than on what it resembles.

Reconstruction and structure disagree The simulated metrics do not simply track the reconstruction metrics, and where they disagree it is the simulation that answers the question the benchmark poses. Exact Match ranks Claude Opus 4 (30.2%) above Gemini 2.5 Pro (28.1%), while the Topology Score reverses them (83.1 against 84.3): Opus reproduces the reference string more often, Gemini builds better structures. The clearest case is GPT-4.1, near the bottom on Exact Match at 4.3% yet producing load-carrying structures 95.6% of the time and scoring 72.7 — above Gemini 1.5 Pro, which matches the reference string five times as often. Figure 1 shows both directions on individual samples. On the full-structure tasks the two views separate completely: Exact Match is 0 for every model while Load Carrying ranges from 33% to 99%, so reconstruction cannot distinguish an agent that builds a standing structure on 99% of samples from one that manages it on 33%.

Material hallucination is universal; smearing is not Every model overspends material on the continuous tasks: averaged over the masked cells of all eight continuous subjects the reference answers carry a mean density of 0.087, while predictions range from 0.168 (DeepSeek-V3) to 0.343 (GPT-4.1).

How that excess is spent separates the models sharply. We call a masked cell *smeared* when a model assigns it an intermediate density ($0 < \rho < 0.5$). The reference answers smear 14–15% of masked cells; GPT-4.1 smears 53.8% and Perplexity Sonar 45.5%, while Claude Opus 4, Gemini 2.5 Pro and DeepSeek-R1 (12.2–15.4%) are indistinguishable from the reference. The same two models leave the most disconnected material (0.70 and 0.43 isolated clusters per prediction against 0.00–0.01) and exceed the budget by more than 5% on 71.1% and 45.3% of predictions against 24–28%. Smearing, islanding and overspend are one syndrome affecting two models, not a property of language models on continuous tasks; the rest overspend by building too much definite structure (Appendix C).

Reasoning buys structure, not strings DeepSeek-R1, the one model trained explicitly to reason before answering, is the one whose two metric families diverge most sharply: its Topology Score is flat across the difficulty boundary (80.0 against 79.4) where every other model loses ground, and its Load Carrying rate is the benchmark's highest on easy tasks at 98.7%, yet on Exact Match it ranks third in both blocks.

Where the gain lands depends on which metric has room to move. Comparing DeepSeek-R1 with DeepSeek-V3, the same base model without the reasoning training, on identical samples: on single-

cell tasks, where the structure is already near-optimal, reasoning doubles Exact Match ($37 \rightarrow 75$ on `1_random_cell_hard`) and changes the Topology Score by less than a point. Where the structure is genuinely at stake the pattern inverts: on `3_random_row_hard` the Topology Score rises 41.5 points and Load Carrying 40 while Exact Match moves from 17 to 19; on `10_random_cell_hard` Exact Match (3 against 4) cannot distinguish the conditions at all while the simulation reports $+22.1$ Topology Score and $+15$ Load Carrying.

Interventions move strings more than structures Three further experiments point the same way (Appendices K, L, M). Rotating the load case by $270°$ degrades every model, the Topology Score falling 11–20 points over four rotated tasks (Gemini 2.5 Pro $75.9 \rightarrow 62.8$; Perplexity Sonar $45.2 \rightarrow 24.8$): models build "downward" relative to the grid rather than along the applied force. Prepending $k = 3$ solved pairs gains Claude Opus 4 14.6 points of Exact Match ($30.2 \rightarrow 44.8$) while its Topology Score rises only $83.1 \rightarrow 86.6$, and physical jargon such as "stress" and "load path" in the prompt *degraded* performance on harder tasks. In all three, cell-wise agreement moves several times more than structural quality, as expected if models match the form of a solution rather than reason about the forces it must carry.

Rotation also exposes a metric failure. For weak models Directional Load-Support Connectivity falls in step with the Topology Score (Perplexity Sonar loses 29.7 points); for DeepSeek-R1 the two come apart entirely, 8.4 points of Topology Score lost against 0.5 of connectivity, and on `full_easy` a collapse from 44.6 to 25.8 against 98 to 95 (Table 7). Being a boolean per sample, directional connectivity saturates once a model reliably finds *some* admissible path: it measures gravity bias in the models that fail outright and is blind to it in the strongest one.

8 Benchmark Construction and SPhyR-v2

The samples are not far apart SPhyR-v1 was generated by sweeping load and support positions one cell at a time along two fixed edges. Among the $1{,}296$ samples of a subject there are $1{,}295$ distinct boundary conditions, and for $1{,}294$ of them the nearest *different* boundary condition is a single cell away, so adjacent load cases usually share an optimal structure: the dataset is a dense sweep of a small space rather than a sample of a large one. A second issue compounds it: the `full` variant, nominally a prediction of the whole structure, masks only rows 1–8, leaving 20 of 100 cells visible. Neither property invalidates Section 7, which compares models on identical samples, but both bound the interpretation: a high absolute score need not indicate reasoning about forces, because neighbouring pre-solved problems are close enough to interpolate between.

SPhyR-v2 We regenerate the benchmark, changing the sampling rather than the task. Loads and supports are placed on any edge in any combination, supports optionally in two disjoint groups, and the load direction is drawn from eight orientations including diagonals, where v1 used one. Candidates are accepted by rejection sampling: a sample joins the set only if its boundary condition differs from every accepted sample by at least six cells *and* its solved structure by at least eight. The `full` variant masks every non-structural cell, and ground truth comes from the SIMP optimizer of Section 4.1, so the dataset regenerates from source.

The construction achieves what it was built for: nearest-neighbour boundary-condition separation rises from 1 cell in v1 to a minimum and median of 6. We release v2 at 10×10 for continuity and at 20×20, where the design space admits multi-member trusses rather than single load paths. Variant sizes scale with the grid — masked cells by area, rows and columns by side length — so `5_random_cell` masks 5% of the grid at either size; carrying v1's literal counts across would have quietly made every task easier.

Results on v2 Table 2 evaluates two current hosted models and one open-weight model on v2 at both sizes, on identical samples. The regenerated instances are harder than v1's at identical grid size: Qwen3.8 loses 23.7 points of Topology Score between the two on the same six subjects, and no model on v2-10 reaches the 80–86 that the strongest models reach on v1. Gemini 3.8 Flash leads at 10×10, Claude Opus 5 at 20×20, where the ranking turns on grid validity (Section 8). Solid fill drops from 71.9 on v1 to 55.8 at 10×10, 14 points below the best model; at 20×20 it scores 56.7, level with Claude Opus 5's 56.9, and Qwen3.8 on the continuous block at 10×10 (39.4) no longer clears random fill (42.6).

Table 2: **Zero-shot performance on SPhyR-v2** at 10×10 (v2-10) and 20×20 (v2-20), 100 samples per subject, identical across models; averages over `10_random_cell`, `3_random_row` and `full` at each difficulty, bootstrap 95% half-widths 2–4 points. Qwen3.8's v1 row: the same subjects on v1. Per-subject and conditional scores, including the full-structure task: Table 6.

Model	**Easy** (binary) EM ↑	Valid ↑	TS ↑	LC ↑	**Hard** (continuous) EM ↑	Valid ↑	TS ↑	LC ↑
v2-10, average over three subjects								
Claude Opus 5	18.0	**99.7**	61.9	89.7	16.3	**99.7**	65.7	**96.0**
Gemini 3.8 Flash	**25.0**	99.3	**70.1**	**95.0**	**23.0**	98.0	**67.8**	95.3
Qwen3.8 (27B, local)	9.3	94.0	42.4	84.3	10.0	86.3	39.4	75.3
same model on v1	12.0	97.3	68.3	95.0	12.0	95.0	60.9	85.3
Solid / void / random	55.8 / 24.8 / 36.3				55.7 / 26.5 / 42.6			
v2-20, average over three subjects								
Claude Opus 5	7.3	**97.3**	**56.9**	**79.0**	6.0	**97.0**	**59.7**	**91.3**
Gemini 3.8 Flash	**7.7**	69.7	43.3	69.7	**7.0**	66.7	40.4	66.7
Solid / void / random	56.7 / 21.9 / 36.2				56.4 / 21.0 / 38.4			

Format compliance becomes a confound as the grid grows A larger grid asks a model to emit more cells in an exact shape, a separate ability from reasoning about structure. At 10×10 the burden is negligible for strong models: on v1 the mean Valid Output Grid rate is 98.8–99.8% for GPT-4.1, Claude Opus 4, Gemini 2.5 Pro and DeepSeek-R1, and only Perplexity Sonar falls away (88.5% mean, 64% worst subject). At 20×20 it is no longer negligible, and it is not confined to small models. Qwen3.8 (27B, Appendix D) on v2-20 reaches a Valid Output Grid rate of 69.3% on `1_random_cell_easy`, the failures almost entirely miscounted row widths — 83 of 92 invalid completions have the right number of rows and go wrong only within a row; restricted to well-formed grids the same model carries the load on 100% of samples and scores 97.5 Topology Score. The two hosted models show the hazard at its sharpest. On `3_random_row_easy` at 20×20, Claude Opus 5 scores 72.7 Topology Score against Gemini 3.8 Flash's 31.8, a 41-point lead; conditioned on well-formed output the order reverses, 75.0 against 85.8, because Gemini returns a valid grid on 37% of samples and Opus on 97% (Table 6). The unconditional number ranks the two by serialisation, the conditional one by structure, and neither alone is the answer. We therefore report Valid Output Grid as a first-class metric rather than a filter, with the conditional figure alongside the unconditional one where they differ. This is a property of the interface rather than the structural task, and the main practical cost of scaling the benchmark up.

9 Conclusion

SPhyR shows that current LLMs do not integrate spatial layout with grounded physical constraints: the failure modes we observe — gravity bias, material smearing, disconnected islands — indicate pattern matching over force-directed reasoning, and neither in-context examples nor explicit reasoning closes the gap evenly. The methodological finding may outlast the empirical one. Because a topology optimization problem has no unique solution, cell-wise agreement with a reference is not a noisy measure of structural reasoning but a systematically misleading one: 72.7% of valid completions carry their load while differing from the reference, models rank differently under the two views, and metrics built on the cell graph inherit the problem. Simulating predictions against an optimizer re-run on the same masked region resolves this without requiring a unique answer, and the verifier's report is something an agent can act on. Future work should train against it.

10 Limitations

Grids are 10×10 and 20×20, far smaller than any structure an engineer would design. v1's boundary conditions come from a single family — load above, support below, three to six cells wide — so it measures one class of structural problem, which is also what lets its samples lie close together (Section 8). Table 2 reports three current models on v2, but the ten-model comparison of Section 7 remains a v1 comparison and its absolute scores carry Section 8's caveat. The v1 references are Millipede's output after ten iterations and not fully converged: the hard block's continuous targets carry that optimizer's intermediate densities, and the 0.93 the reference scores under our verifier measures that gap. The feedback experiment is one model on one subject. With 100 samples per task

variation, bootstrap 95% intervals are ± 1.3 to ± 2.8 points on the block averages of Table 1 and up to ± 9.5 on a single full-structure subject, so we draw conclusions only from patterns holding across subjects; models are evaluated zero-shot through public APIs that may not pin a serving configuration. Our metrics rest on linear elastic analysis under one static load case: a structure scoring well carries its load efficiently but is not ready to build.

Reproducibility Statement

The anonymised repository linked in the abstract contains the environment, the evaluation code, both dataset versions (v1 and v2), the prompts, and every stored model completion behind the numbers in this paper, so each table can be recomputed from the stored outputs without re-querying any model. Appendix A gives the prompt format, Appendix B defines every metric, Appendix E lists the topology optimization solver parameters used by the verifier and by the v2 generation, Appendix D details the feedback experiment and the v2 runs, and Appendix I reports the complete results for all metrics. Section 7 names the models and the sampling configuration; the limitations of evaluating through public APIs are stated in Section 10.

AI Use Statement

In this work, we used generative AI tools (large language model coding assistants) to help draft and edit the manuscript text and to assist in writing and refactoring the evaluation and analysis code. We did not use generative AI tools to generate the research ideas, the experimental design, or the benchmark data; the language models listed in Section 7 are the subjects of the evaluation, not tools used to produce it. We have reviewed all AI-assisted work: the authors verified the AI-assisted code against the stored completions and the reported numbers, and edited all AI-assisted text. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Erik Andreassen, Anders Clausen, Mattias Schevenels, Boyan S. Lazarov, and Ole Sigmund. Efficient topology optimization in MATLAB using 88 lines of code. *Structural and Multidisciplinary Optimization*, 43(1):1–16, 2011. doi: 10.1007/s00158-010-0594-7.

anthropic. Claude 3.7 Sonnet and Claude Code, February 2025a. URL https://www.anthropic.com/news/claude-3-7-sonnet.

anthropic. Introducing Claude 4, May 2025b. URL https://www.anthropic.com/news/claude-4.

Anton Bakhtin, Laurens van der Maaten, Justin Johnson, Laura Gustafson, and Ross Girshick. PHYRE: A New Benchmark for Physical Reasoning, August 2019. URL http://arxiv.org/abs/1908.05656. arXiv:1908.05656 [cs].

Saurabh Banga, Harsh Gehani, Sanket Bhilare, Sagar Patel, and Levent Kara. 3D Topology Optimization using Convolutional Neural Networks, August 2018. URL http://arxiv.org/abs/1808.07440. arXiv:1808.07440 [cs].

Daniel M. Bear, Elias Wang, Damian Mrowca, Felix J. Binder, Hsiao-Yu Fish Tung, R. T. Pramod, Cameron Holdaway, Sirui Tao, Kevin Smith, Fan-Yun Sun, Li Fei-Fei, Nancy Kanwisher, Joshua B. Tenenbaum, Daniel L. K. Yamins, and Judith E. Fan. Physion: Evaluating Physical Prediction from Vision in Humans and Machines, June 2022. URL http://arxiv.org/abs/2106.08261. arXiv:2106.08261 [cs].

Martin P. Bendse. Optimal shape design as a material distribution problem. *Structural Optimization*, 1(4):193–202, 1989. doi: 10.1007/BF01650949.

Martin P. Bendse and Ole Sigmund. *Topology Optimization*. Springer, Berlin, Heidelberg, 2004. ISBN 978-3-642-07698-5 978-3-662-05086-6. doi: 10.1007/978-3-662-05086-6. URL http://link.springer.com/10.1007/978-3-662-05086-6.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. PIQA: Reasoning about Physical Commonsense in Natural Language, November 2019. URL http://arxiv.org/abs/1911.11641. arXiv:1911.11641 [cs].

Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. OpenAI Gym. *arXiv preprint arXiv:1606.01540*, 2016. URL https://arxiv.org/abs/1606.01540.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language Models are Few-Shot Learners, July 2020. URL http://arxiv.org/abs/2005.14165. arXiv:2005.14165 [cs].

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating Large Language Models Trained on Code. *arXiv preprint arXiv:2107.03374*, 2021. URL https://arxiv.org/abs/2107.03374.

Francois Chollet, Mike Knoop, Gregory Kamradt, Bryan Landers, and Henry Pinkard. ARC-AGI-2: A New Challenge for Frontier AI Reasoning Systems, May 2025. URL http://arxiv.org/abs/2505.11831. arXiv:2505.11831 [cs].

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training Verifiers to Solve Math Word Problems. *arXiv preprint arXiv:2110.14168*, 2021. URL https://arxiv.org/abs/2110.14168.

Gheorghe Comanici et al. Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities, December 2025. URL http://arxiv.org/abs/2507.06261. arXiv:2507.06261 [cs].

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Haowei Zhang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Li, Hui Qu, J. L. Cai, Jian Liang, Jianzhong Guo, Jiaqi Ni, Jiashi Li, Jiawei Wang, Jin Chen, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, Junxiao Song, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Lei Xu, Leyi Xia, Liang Zhao, Litong Wang, Liyue Zhang, Meng Li, Miaojun Wang, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Ning Tian, Panpan Huang, Peiyi Wang, Peng Zhang, Qiancheng Wang, Qihao Zhu, Qinyu Chen, Qiushi Du, R. J. Chen, R. L. Jin, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, Runxin Xu, Ruoyu Zhang, Ruyi Chen, S. S. Li, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shaoqing Wu, Shengfeng Ye, Shengfeng Ye, Shirong Ma, Shiyu Wang, Shuang Zhou, Shuiping Yu, Shunfeng Zhou, Shuting Pan, T. Wang, Tao Yun, Tian Pei, Tianyu Sun, W. L. Xiao, Wangding Zeng, Wanjia Zhao, Wei An, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, X. Q. Li, Xiangyue Jin, Xianzu Wang, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaojin Shen, Xiaokang Chen, Xiaokang Zhang, Xiaosha Chen, Xiaotao Nie, Xiaowen Sun, Xiaoxiang Wang, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xingkai Yu, Xinnan Song, Xinxia Shan, Xinyi Zhou, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, Y. K. Li, Y. Q. Wang, Y. X. Wei, Y. X. Zhu, Yang Zhang, Yanhong Xu, Yanhong Xu, Yanping Huang, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Li, Yaohui Wang, Yi Yu, Yi Zheng, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Ying Tang, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yu Wu, Yuan Ou, Yuchen Zhu, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yukun Zha, Yunfan Xiong, Yunxian Ma, Yuting Yan, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Z. F. Wu, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhen Huang, Zhen Zhang, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhibin Gou, Zhicheng Ma, Zhigang Yan, Zhihong Shao, Zhipeng Xu, Zhiyu Wu, Zhongyu Zhang, Zhuoshu Li, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Ziyi Gao, and Zizheng Pan. DeepSeek-V3 Technical Report, February 2025a. URL http://arxiv.org/abs/2412.19437. arXiv:2412.19437 [cs].

DeepSeek-AI et al. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning, January 2025b. URL http://arxiv.org/abs/2501.12948. arXiv:2501.12948 [cs].

Leo Gao, John Schulman, and Jacob Hilton. Scaling Laws for Reward Model Overoptimization. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pp. 10835–10866. PMLR, 2023. URL https://proceedings.mlr.press/v202/gao23h.html.

Oliver Groth, Fabian B. Fuchs, Ingmar Posner, and Andrea Vedaldi. ShapeStacks: Learning Vision-Based Physical Intuition for Generalised Object Stacking, July 2018. URL http://arxiv.org/abs/1804.08018. arXiv:1804.08018 [cs].

Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. SWE-bench: Can Language Models Resolve Real-World GitHub Issues? In *International Conference on Learning Representations (ICLR)*, 2024. URL https://arxiv.org/abs/2310.06770.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning, December 2016. URL http://arxiv.org/abs/1612.06890. arXiv:1612.06890 [cs].

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, et al. Tülu 3: Pushing Frontiers in Open Language Model Post-Training. *arXiv preprint arXiv:2411.15124*, 2024. URL https://arxiv.org/abs/2411.15124.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's Verify Step by Step. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://arxiv.org/abs/2305.20050. arXiv:2305.20050.

Panagiotis Michalatos and Sawako Kaijima. Millipede Plugin Grasshopper 3D, 2024. URL https://www.creativemutation.com/millipede.

OpenAI et al. GPT-4 Technical Report, March 2024a. URL http://arxiv.org/abs/2303.08774. arXiv:2303.08774 [cs].

OpenAI et al. GPT-4o System Card, October 2024b. URL http://arxiv.org/abs/2410.21276. arXiv:2410.21276 [cs].

Sharad Rawat and M.-H. Herman Shen. A Novel Topology Optimization Approach using Conditional Deep Learning, January 2019. URL http://arxiv.org/abs/1901.04859. arXiv:1901.04859 [cs].

Ronan Riochet, Mario Ynocente Castro, Mathieu Bernard, Adam Lerer, Rob Fergus, Vronique Izard, and Emmanuel Dupoux. IntPhys: A Framework and Benchmark for Visual Intuitive Physics Reasoning, February 2020. URL http://arxiv.org/abs/1803.07616. arXiv:1803.07616 [cs].

Joar Max Viktor Skalse, Nikolaus H. R. Howe, Dmitrii Krasheninnikov, and David Krueger. Defining and Characterizing Reward Gaming. In *Advances in Neural Information Processing Systems*, volume 35, October 2022. URL https://openreview.net/forum?id=yb3HOXO3lX2.

Gemini Team et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context, December 2024. URL http://arxiv.org/abs/2403.05530. arXiv:2403.05530 [cs].

Perplexity Team. Meet new Sonar: A Blazing Fast Model Optimized for Perplexity Search, February 2025a. URL https://www.perplexity.ai/hub/blog/meet-new-sonar.

Perplexity Team. RL Training For Math Reasoning, May 2025b. URL https://www.perplexity.ai/hub/blog/rl-training-for-math-reasoning.

Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B. Tenenbaum. CLEVRER: CoLlision Events for Video REpresentation and Reasoning, March 2020. URL http://arxiv.org/abs/1910.01442. arXiv:1910.01442 [cs].

Xinyu Zhang, Yuxuan Dong, Yanrui Wu, Jiaxing Huang, Chengyou Jia, Basura Fernando, Mike Zheng Shou, Lingling Zhang, and Jun Liu. PhysReason: A Comprehensive Benchmark towards Physics-Based Reasoning, February 2025. URL http://arxiv.org/abs/2502.12054. arXiv:2502.12054 [cs].

Yiquan Zhang, Bo Peng, Xiaoyi Zhou, Cheng Xiang, and Dalei Wang. A deep Convolutional Neural Network for topology optimization with strong generalization ability, March 2020. URL http://arxiv.org/abs/1901.07761. arXiv:1901.07761 [cs].

Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. WebArena: A Realistic Web Environment for Building Autonomous Agents. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://arxiv.org/abs/2307.13854.

A Task Overview, Benchmark Comparison and Prompt Format

Figure 2 shows the task family. Table 3 places SPhyR among existing physical and spatial reasoning benchmarks, and Figure 3 shows the prompt template served to every model.

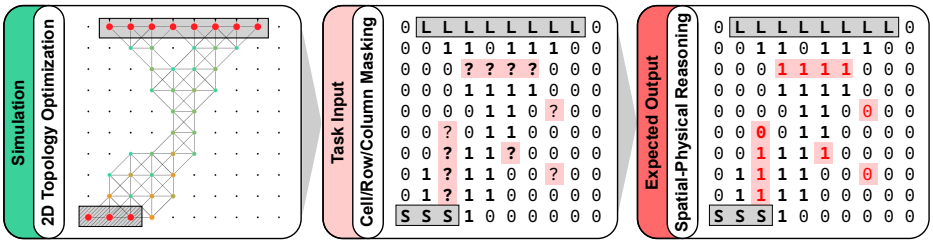

Figure 2: Topology optimization computes a material distribution; masking individual cells, rows, columns or the whole distribution yields a graded family of spatial-physical reasoning tasks.

Benchmark	Format	Physical Reasoning	Spatial Reasoning	Notes
CLEVR (2016)	Visual QA	✗	✓	Scene reasoning
CLEVRER (2020)	Video QA	✓	✓	Causal events
IntPhys (2020)	Video plausibility	✓	✓	Violation detection
Physion (2022)	Video prediction	✓	✓	Object behavior prediction
ShapeStacks (2018)	Image classification	✓	✓	Block stability
PHYRE (2019)	2D physics puzzles	✓	✓	Action planning
PIQA (2019)	Text QA	✓	✗	Physical commonsense
PhysReason (2025)	Text QA	✓	✗	Multi-step physics
SPhyR (ours)	Structured prediction	✓	✓	Material distribution

Table 3: Comparison of existing benchmarks evaluating physical and spatial reasoning. Our proposed dataset (**SPhyR**) focuses specifically on material distribution reasoning under boundary conditions, combining spatial and physical understanding in structured tasks.

B Metric Definitions

We evaluate model performance using four complementary families of metrics (summarised in Table 4) for a holistic assessment of both symbolic accuracy and structural realism: (1) **Reconstruction Accuracy Metrics**, quantifying cell-wise agreement with the ground truth, including measures of fidelity, penalization, and difficulty weighting; (2) **Topological Validity Metrics**, assessing global structural soundness through load-support connectivity and grid validity; (3) **Physics-Approximating Metrics**, estimating the structural efficiency via gravity-aligned load-transfer paths; and (4) **Simulated Structural Metrics** (Section 4.1), which simulate the predicted structure and compare it against a topology optimizer re-run on the same masked region. The first three families all compare a prediction to one reference answer; the fourth asks instead whether the prediction works as a structure, which is the property the benchmark is ultimately about.

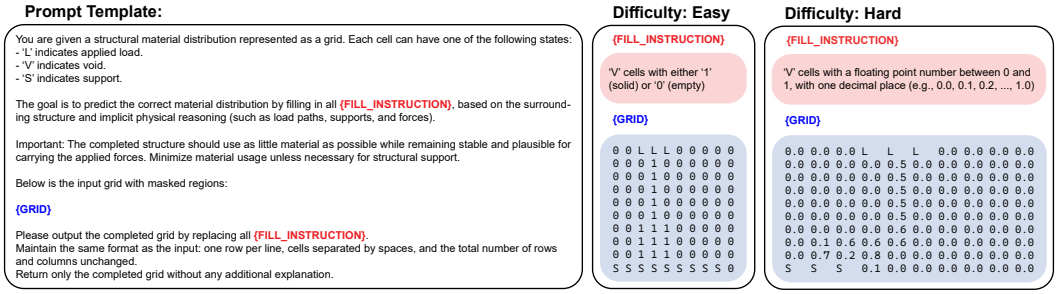

Figure 3: Prompt template used across all tasks and difficulty levels, showing instructions and grid format as served to models for evaluation.

Category	Metric Name	Type / Range	Desired Trend
Reconstruction	Exact Match (EM)	Boolean $\{0,1\}$	True
	Difference Ratio (DiffRatio)	Float [0,1]	Higher is better
	Penalized Difference Ratio (PenDiffRatio)	Float [0,1]	Higher is better
	Relative Difference Ratio (RelDiffRatio)	Float [0,1]	Higher is better
	Difficulty-Weighted Diff. Ratio	Float [0,1]	Higher is better
	Difficulty-Weighted Rel. Diff. Ratio	Float [0,1]	Higher is better
Topology	Valid Grid (ValidGrid)	Boolean $\{0,1\}$	True
	Load-Support Connectivity (LSConn)	Boolean $\{0,1\}$	True
	Directional L-S Connectivity (DirLSConn)	Boolean $\{0,1\}$	True
	Isolated Clusters (N_{islands})	Integer ≥ 0	Lower is better
	Difficulty Score (DWCS)	Float [1,3] avg.	Higher is harder
Physics-Approx.	Force Path Cost Efficiency (FPCEff)	Float [0,1]	Higher is better
Structural (Simulated)	Topology Score (TS)	Float [0,1]	Higher is better
	Structural Efficiency (SE)	Float [0,1]	Higher is better
	Material Efficiency (ME)	Float [0,1]	Higher is better
	Load Carrying (LC)	Boolean $\{0,1\}$	True
	Compliance (c)	Float > 0	Lower is better
	Compliance Eff. vs Ground Truth (CE)	Float [0,1]	Higher is better

Table 4: Summary of all evaluation metrics by category, with their types, typical ranges, and optimization direction.

Reconstruction Accuracy Metrics We assess reconstruction fidelity using several grid-level measures based on cellwise agreement between the predicted grid $\hat{G}$ and the ground-truth grid G^* (Appendix G.2, for prompt, completion and calculation scenarios).

- **Exact Match ↑ (EM)**: Binary indicator of perfect reconstruction:

$$\text{EM} = \begin{cases} 1, & \text{if } \hat{G} = G^*, \\ 0, & \text{otherwise.} \end{cases}$$

- **Difference Ratio (DiffRatio)**: Fraction of incorrect cells normalized by total ground-truth mass:

$$\text{DiffRatio} = 1 - \frac{D(\hat{G}, G^*)}{\sum_{i,j} g^*_{ij}},$$

  where $D(A, B)$ counts cellwise mismatches. *Higher is better* (1 indicates perfect match).
- **Relative Difference Ratio (RelDiffRatio)**: A softer variant that measures numeric deviation:

$$\text{RelDiffRatio} = 1 - \frac{D_{\text{rel}}(\hat{G}, G^*)}{\sum_{i,j} g^*_{ij}},$$

  where D_{rel} accumulates $|a_{ij} - b_{ij}|$ for numeric cells and counts categorical mismatches (L, S, V) as 1. *Higher is better.*
- **Penalized Difference Ratio (PenDiffRatio)**: Penalty-weighted version increasing the cost of modifying or introducing new load, support, or void cells:

$$\text{PenDiffRatio} = 1 - \frac{D_{\text{pen}}(\hat{G}, G^*)}{\sum_{i,j} g^*_{ij}},$$

  where D_{pen} multiplies L, S, or V cell errors by a penalty (typically $3\times$). *Higher is better.*
- **Difficulty-Weighted Difference Ratios**: Optional variants that multiply each cells contribution by its local difficulty weight (see DWCS below). These versions emphasize correctness in structurally ambiguous or high-difficulty regions. *Higher is better.*

Topological Validity Metrics Beyond pixelwise accuracy, we evaluate the structural and connectivity properties of the reconstructed topology (Appendix G.3, for prompt, completion and calculation scenarios):

- **Grid Validity (ValidGrid)**: Boolean check ensuring $\hat{G}$ matches G^* in shape and uses only admissible values (L, S, or $[0, 1]$). *True is desired.*

- **Load-Support Connectivity (LSConn)**: True if any load cell (L) connects to any support (S) through contiguous solid cells (> 0, L, or S):

$$\text{LSConn} = \begin{cases} 1, & \exists \text{ load-support path through solids,} \\ 0, & \text{otherwise.} \end{cases}$$

  True is desired.
- **Directional Load-Support Connectivity (DirLSConn)**: Same as LSConn, but restricted to force paths aligned with the gravity vector $\mathbf{g}$ inferred from dataset rotation metadata. *True is desired.*
- **Isolated Cluster Count (N_{islands})**: Number of solid-cell clusters disconnected from any load or support, found via 4-connectivity. *Lower is better.*
- **Difficulty Score (DWCS)**: Average difficulty weight for originally masked cells:

$$\text{DWCS} = \frac{1}{|\mathcal{V}|} \sum_{(i,j)\in\mathcal{V}} w_{ij}, \quad w_{ij} \in \{1, 2, 3\}.$$

  Higher DWCS implies the reconstruction region is more complex or ambiguous; it reflects task difficulty rather than model quality. *Higher indicates harder samples.*

Physics-Approximating Metrics To estimate the physical plausibility of predicted topologies, we approximate directional load-support efficiency using a force-path traversal cost. We calculate the average minimum directional cost for each load to reach a support, computed via a gravity-aligned Dijkstra traversal with angular and depth penalties. Unsupported loads receive a large but finite penalty (Appendix G.1 for EPCEff calculation details, and G.3, for prompt, completion and calculation scenarios).

- **Force Path Cost Average Efficiency Ratio (FPCEff)**: Relative efficiency of predicted vs. ground-truth structures:

$$\text{FPCEff} = \text{clip}_{[0,1]}\left(\frac{C^*_{\text{avg}}}{\hat{C}_{\text{avg}}}\right),$$

  where C^*_{avg} and $\hat{C}_{\text{avg}}$ are average load-support path costs in G^* and $\hat{G}$ respectively. *Higher is better.*

C Qualitative Failure Modes

To complement the quantitative metrics we inspected model predictions visually. Three recurring failure modes account for the performance gaps of Section 7 and Table 1.

The "Smearing" Effect in Continuous Tasks Two models fail to commit to a defined structure on hard (continuous) tasks. Instead of placing high-density material in critical load paths, GPT-4.1 and Perplexity Sonar distribute low-density values broadly across the void (Appendix N.6), assigning intermediate densities to 53.8% and 45.5% of masked cells against the reference answers' 14–15%. They appear to minimize risk by filling space rather than optimizing for stiffness, hallucinating material where none is needed. We stress that this is a property of these two models and not of the task: the other four commit as decisively as the optimizer.

Disconnected Islands and Local Bias The same two models generate "floating islands", clusters of material completely disconnected from any load or support. Over the continuous subjects GPT-4.1 averages 0.70 such clusters per prediction and Perplexity Sonar 0.43, while Claude Opus 4, Gemini 2.5 Pro and DeepSeek-R1 average 0.01 or fewer — a difference of nearly two orders of magnitude rather than a matter of degree. Where islands appear, the model is operating on local pattern consistency (placing a 1 next to another 1) rather than global constraint satisfaction, failing to trace the load path $L \to S$ back to a fixed point. That the strongest models produce essentially none suggests the constraint is learnable, and that the models which violate it are not approximating it poorly but ignoring it.

Gravity Bias in Rotated Scenarios Qualitative inspection of the rotated experiments ($270°$) reveals a directional bias in every model tested. Even when the load L is applied horizontally from the left, models often attempt to build "downward" relative to the grid layout, ignoring the rotated force vector. This results in structures that "hang" into empty space or connect to non-existent supports at the bottom of the grid, evidence that these models lean on visual memorization of vertical architectural forms rather than on physical reasoning. The bias takes two visually distinct forms. Weaker models fail outright, leaving material unattached to any support. DeepSeek-R1 instead produces structures that remain correctly attached — its directional connectivity is unchanged under rotation — but that route the load through a heavier, less efficient section than the upright case, as though it recognises where the load must go and not how a structure carrying it sideways should be shaped.

Over-Constrained "Safety" Conversely, top-performing models like Gemini 2.5 Pro often "over-build," creating blocky, wall-like structures rather than truss-like efficient designs. While this strategy achieves high Load-Support Connectivity (resulting in high success rates), it fails the efficiency objective of topology optimization, treating the task as a "fill-the-gap" segmentation problem rather than a minimum-compliance optimization problem.

D Details for the Feedback Experiment and the SPhyR-v2 Results

Local model The local model of Sections 6 and 8 is Qwen3.8 with 27.3B parameters, run at 4-bit (Q4_K_M) quantisation, 17 GB on disk, through Ollama's OpenAI-compatible endpoint. It was called exactly as the hosted models are: one user message holding the prompt of Figure 3, at most $16{,}000$ completion tokens, and the sampling parameters the served model defaults to (temperature was not overridden). The same settings apply to both arms of Section 6 and to every Qwen3.8 row of Tables 2 and 6. Claude Opus 5 and Gemini 3.8 Flash were reached through OpenRouter at their default reasoning effort with the same token cap. Raising the effort to its highest setting does not explain the gap between them: on `full_easy` at 20×20 over 25 samples Opus moves from 7.7 to 6.5 Topology Score, and Gemini emits $15{,}964$ of its $16{,}000$ tokens as reasoning, returns no grid on any sample, and scores 0.

Feedback protocol Both arms use the first 50 samples of `full_easy` on v2 at 10×10, with three further attempts after the first answer. In the revision arm each further attempt receives the prompt below, in which the feedback lines are generated from the previous attempt's metrics: whether the reply parsed as a grid of the right shape; whether it carried the load; its structural efficiency as a percentage of the reference design's; its material use, phrased as an instruction to remove material when more than 5% over budget and to add material along the load path when more than 5% under; and the number of material clusters attached to no load or support, if any. In the control arm the original prompt is re-posed with nothing added. Each attempt is scored on its own, and an episode's score is the best of its four attempts.

Revision prompt

```
Your previous answer was simulated as a linear elastic structure.
Here is how it behaved:

{FEEDBACK}

Below is the original grid with masked regions again:

{GRID}

Your previous answer:

{PREVIOUS}

Revise it to carry the load with less material.  Reply with the
completed grid only, in the same format:  one row per line, cells
```

```
separated by spaces, the same number of rows and columns.  Return only
the grid, with no explanation.
```

Feedback lines, as generated

```
- The reply could not be read as a grid of the right shape, so it
could not be simulated.

- It did NOT carry the load:  force reached no support through
material, so the structure is not standing.

- It carried the load.

- Stiffness against the best design at this material budget:  43%
(100% would match it).

- It used 30% MORE material than the budget allows.  Remove material
that carries no load; you must come down to the budget.

- It used only 52% of the material budget, so 48% is still unspent.
Material is not penalised until the budget is exceeded, so ADD
material along the load path to make the structure stiffer.

- Its material use matched the budget.

- 2 cluster(s) of material are connected to no load and no support,
and carry nothing.
```

Table 5: **Score by attempt in both arms of Section 6.** Qwen3.8 on `full_easy`, v2 at 10×10, 50 samples. Topology Score, Load Carrying and Valid Output Grid are means over samples at each attempt; the last row is the mean over samples of the best attempt. The paired best-of-four difference is $+12.79$ (standard error 4.46, 95% CI $[4.0, 21.5]$), with 35 wins, 9 losses and 6 ties for revision.

	Revision			**Control** (re-posed)		
Attempt	TS	LC	Valid	TS	LC	Valid
First answer	15.9	76	90	19.7	80	82
Second	24.5	72	78	15.4	78	92
Third	24.2	70	78	17.9	82	92
Fourth	22.8	66	74	17.3	86	90
Best of four	**37.9**			25.1		

Per-subject SPhyR-v2 results Table 6 gives every subject behind Table 2, with the Topology Score both unconditionally and restricted to well-formed grids. The two differ wherever Valid Output Grid falls below 100%, and at 20×20 they can reverse a ranking: on `3_random_row_easy` Claude Opus 5 leads Gemini 3.8 Flash by 41 points unconditionally and trails it by 11 conditionally.

DeepSeek-R1 under rotation Table 7 gives the DeepSeek-R1 numbers quoted in Section 7; the rotation tables and figures of Appendix K report DeepSeek-V3, the non-reasoning model of the same family, on the same samples.

E Topology Optimization Solver Parameters

SPhyR-v1: Grasshopper Millipede The v1 ground truth was produced with Millipede (Michalatos & Kaijima, 2024), an external solver, at these settings:

target density = 0.1
self-weight = 0
iterations = 10

Table 6: **Per-subject results on SPhyR-v2**, 100 samples per subject on identical samples (300 for the Qwen3.8 rows at 20×20, which cover a different subject set). TS|valid is the Topology Score over well-formed grids only.

Grid, model	Subject	Valid	TS	TS|valid	LC
v2-10, Claude Opus 5	10_random_cell_easy	100	88.8	88.8	97
	10_random_cell_hard	99	89.3	90.2	95
	3_random_row_easy	99	70.4	71.1	96
	3_random_row_hard	100	70.7	70.7	95
	full_easy	100	26.5	26.5	76
	full_hard	100	37.0	37.0	98
v2-10, Gemini 3.8 Flash	10_random_cell_easy	100	91.1	91.1	99
	10_random_cell_hard	100	92.4	92.4	98
	3_random_row_easy	98	74.8	76.4	94
	3_random_row_hard	94	73.1	77.8	88
	full_easy	100	44.3	44.3	92
	full_hard	100	38.0	38.0	100
v2-10, Qwen3.8	10_random_cell_easy	96	69.1	71.9	91
	10_random_cell_hard	88	67.1	76.3	85
	3_random_row_easy	94	42.4	45.2	84
	3_random_row_hard	85	34.8	40.9	70
	full_easy	92	15.7	17.1	78
	full_hard	86	16.3	19.0	71
v2-20, Claude Opus 5	10_random_cell_easy	98	92.8	94.7	98
	10_random_cell_hard	96	91.9	95.7	96
	3_random_row_easy	97	72.7	75.0	92
	3_random_row_hard	96	76.5	79.7	94
	full_easy	97	5.1	5.2	47
	full_hard	99	10.9	11.0	84
v2-20, Gemini 3.8 Flash	10_random_cell_easy	81	78.1	96.4	81
	10_random_cell_hard	78	75.6	96.9	78
	3_random_row_easy	37	31.8	85.8	37
	3_random_row_hard	34	31.0	91.2	34
	full_easy	91	20.0	21.9	91
	full_hard	88	14.5	16.5	88
v2-20, Qwen3.8 ($n = 300$)	1_random_cell_easy	69.3	67.6	97.4	69.3
	5_random_cell_easy	63.3	55.7	87.9	62.3
	1_random_row_easy	64.3	45.8	71.2	63.3
	10_random_cell_easy	58.1	43.3	74.6	57.0

Table 7: **DeepSeek-R1 upright and under** $270°$ **rotation**, 100 samples per subject on v1. Topology Score falls by 8.4 points on average while Directional Load-Support Connectivity falls by 0.5.

	TS		Dir. Conn.		LC		Valid	
Subject	upright	rotated	upright	rotated	upright	rotated	upright	rotated
10_random_cell_easy	76.6	78.4	99	100	98	99	100	100
3_random_row_easy	82.6	74.3	99	97	96	97	99	98
3_random_column_easy	70.5	62.2	98	100	98	99	99	100
full_easy	44.6	25.8	98	95	99	95	99	96
Mean	68.6	60.2	98.5	98.0	97.8	97.5	99.3	98.5

smoothing = 0.1
penalization = 3.0
minimum density = 0.001
delete threshold = 0.5
compliant mechanism disabled

SPhyR-v2: The Evaluation Optimizer v2's ground truth comes from the same SIMP implementation used for the simulated structural metrics (Section 4.1) rather than from an external tool, which is what allows the dataset to be regenerated from source. It runs at penalization $p = 3$ under the continuation schedule described there, with a volume fraction of 0.3 and the best design retained across cold and warm-started runs, their volume-preserving binarizations and a discrete swap refinement.

The budget differs from v1's target density of 0.1 because v2 places loads and supports on arbitrary edges: those cells are solid by definition and can occupy a tenth of the grid on their own, leaving a 0.1 budget with nothing for the optimizer to distribute. The two versions' absolute material figures are therefore not directly comparable, though within either version they are.

F VISUAL TASK VARIATIONS OVERVIEW

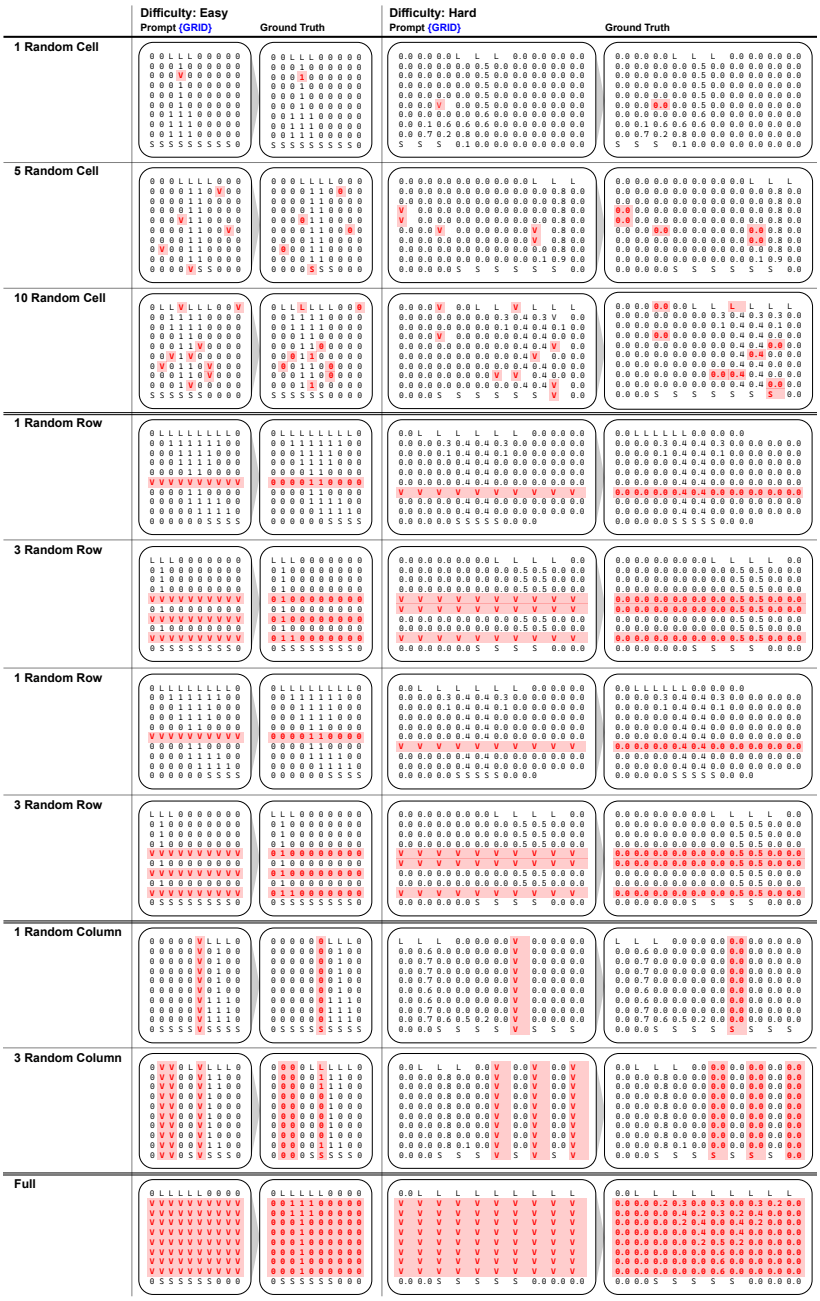

Figure 4: Overview of task variations: predicting material distributions for N random cells, rows, columns, or full structures for easy (binary) and hard (continous) difficulties.

G Additional Evaluation Metrics Computation and Prompt and Completion Examples

G.1 Force-Path Cost Computation

To approximate the physical efficiency of load transmission through the predicted topology, we define a gravity-aligned cost metric that measures the minimum traversal effort for any load cell to reach a support cell through contiguous solid material.

Each grid cell g_{ij} can take values in $\{L, S\} \cup [0, 1]$, where L and S denote applied load and support, respectively, and real-valued entries represent material density. We assume a fixed gravity direction $\mathbf{g} = (d_r, d_c) \in \{(1, 0), (0, 1), (-1, 0), (0, -1)\}$.

Directional neighborhood. We consider all 8-connected neighbors of (i, j),

$$\mathcal{N}(i, j) = \{(i', j') \mid (i' - i, j' - j) \in \{(\pm 1, 0), (0, \pm 1), (\pm 1, \pm 1)\}\},$$

with direction vector $\mathbf{d} = (i' - i, j' - j)$. Each neighbor is assigned a traversal cost $w_{\mathbf{d}}$ based on its angular deviation from gravity:

$$w_{\mathbf{d}} = \begin{cases} 1.0, & \angle(\mathbf{g}, \mathbf{d}) < 15^\circ, \\ 1.2, & 15^\circ \le \angle < 45^\circ, \\ 1.5, & 45^\circ \le \angle < 100^\circ, \\ 3.0, & \text{otherwise.} \end{cases}$$

Upward (against-gravity) moves are disallowed whenever $\mathbf{d} \cdot \mathbf{g} < -0.5$, ensuring that load flow occurs only downward or laterally.

Shortest-path computation. For each load cell $\ell = (i_\ell, j_\ell)$, we compute the minimal cost to any support $s \in S$ using Dijkstras algorithm over the graph of solid nodes $\{(i, j) \mid g_{ij} > 0\}$. The cumulative path cost is defined as

$$C(\ell) = \min_{p \in P_{\ell \to S}} \sum_{((i,j),(i',j')) \in p} w_{(i'-i),(j'-j)} \left(1 + 0.05\,|i' - i_\ell|\right),$$

where the multiplicative term $1 + 0.05\,|i' - i_\ell|$ imposes a mild depth penalty to discourage long vertical travel from the load origin. If no valid support is reachable, a finite penalty $C_{\max}$ is assigned.

The mean force-path cost for a grid G is

$$\overline{C}(G) = \frac{1}{N_L} \sum_{\ell \in L} C(\ell), \qquad C(\ell) = C_{\max} \text{ if unsupported.}$$

Force-Path Cost Average Efficiency Ratio. We define the final metric as

$$\text{FPCEff} = \text{clip}_{[0,1]} \left(\frac{\overline{C}(G^*)}{\overline{C}(\hat{G})} \right),$$

where G^* and $\hat{G}$ denote the ground-truth and predicted grids, respectively. Higher values indicate that the predicted structure achieves comparable or better load-support transmission efficiency than the reference.

G.2 Reconstruction Metric Tests

G.2.1 Exact Match Examples

Exact Match Examples

This test validates the get_exact_match function, which returns True if the predicted grid $\hat{G}$ exactly matches the ground truth G^* cell-by-cell.

1. Perfect match (True)

```
Ground truth:  [ 0 , L , 0 ],
               [ 0 , 1 , 0 ],
               [ 0 , S , 0 ]
Prediction:    [ 0 , L , 0 ],
               [ 0 , 1 , 0 ],
               [ 0 , S , 0 ]
Expected output:  True
```

2. Slight difference (False)

```
Ground truth:  [ 0 , L , 0 ],
               [ 0 , 1 , 0 ],
               [ 0 , S , 0 ]
Prediction:    [ 1 , L , 0 ],
               [ 1 , 1 , 0 ],
               [ 0 , S , 0 ]
Expected output:  False
```

G.2.2 Difference Ratio Examples

Difference Ratio Examples

This test validates get_difference_ratio, which measures similarity between $\hat{G}$ and G^*.
A value of 1.0 means perfect reconstruction, while lower values indicate greater deviation.

1. Perfect match (1.000)

```
Ground truth:  [ 0 , L , 0 ],
               [ 0 , 1 , 0 ],
               [ 0 , S , 0 ]
Prediction:    [ 0 , L , 0 ],
               [ 0 , 1 , 0 ],
               [ 0 , S , 0 ]
Expected output:  1.000
```

2. One altered column (0.000)

```
Ground truth:  [ 0 , L , 0 ],
               [ 0 , 1 , 0 ],
               [ 0 , S , 0 ]
Prediction:    [ 1 , L , 0 ],
               [ 1 , 1 , 0 ],
               [ 0 , S , 0 ]
Expected output:  0.000
```

3. Half correct (0.500)

```
Ground truth:  [ 0 , L , 0 ],
               [ 1 , 1 , 0 ],
               [ 0 , S , 0 ]
Prediction:    [ 0 , L , 0 ],
               [ 0 , 1 , 0 ],
               [ 0 , S , 0 ]
Expected output:  0.500
```

G.2.3 Relative and Penalized Difference Ratio Examples

Relative and Penalized Difference Ratio Examples

```
These tests validate get_relative_difference_ratio and
get_penalized_difference_ratio, which account for numeric
cell differences and penalize fixed-cell deviations respectively.
1.  Perfect alignment (1.000)
Ground truth:  [ 0 , L , 0 ],
               [ 0 , 1 , 0 ],
               [ 0 , S , 0 ]
Prediction:    [ 0 , L , 0 ],
               [ 0 , 1 , 0 ],
               [ 0 , S , 0 ]
Expected output:  1.000
2.  Gradual deviation (0.333)
Ground truth:  [ 0 , L , 0 ],
               [ 1 , 1 , 1 ],
               [ 0 , S , 0 ]
Prediction:    [ 0 , L , 0 ],
               [ 0 , 1 , 0 ],
               [ 0 , S , 0 ]
Expected output:  0.333
3.  Continuous values (0.500)
Ground truth:  [ 0 , L , 0 ],
               [ 0.8 , 1 , 0.8 ],
               [ 0 , S , 0 ]
Prediction:    [ 0 , L , 0 ],
               [ 0.4 , 0.5 , 0.4 ],
               [ 0 , S , 0 ]
Expected output:  0.500
4.  Over-extrapolation (0.308)
Ground truth:  [ 0 , L , 0 ],
               [ 0.8 , 1 , 0.8 ],
               [ 0 , S , 0 ]
Prediction:    [ 0 , L , 0 ],
               [ 0.4 , 2.0 , 0.4 ],
               [ 0 , S , 0 ]
Expected output:  0.308
5.  Negative ratio (-1.000 or -2.000)
Ground truth:  [ 0 , L , 0 ],
               [ 0 , 1 , 0 ],
               [ 0 , S , 0 ]
Prediction:    [ 0 , 1 , 0 ],
               [ 1 , 1 , 1 ],
               [ 0 , S , 0 ]
Expected output:  -1.000 (unpenalized), -2.000 (penalized)
Interpretation:
The ratios decrease as predictions deviate numerically from the ground
truth,
and penalized variants further reduce the score when fixed regions
(load or support)
are incorrectly modified.  Scores near or below 0 reflect large or
structurally
meaningful errors.
```

G.3 Topology Metric

G.3.1 Grid Validity Examples

Grid Validity Examples

```
This test validates the get_grid_shape_and_value_validity function,
which ensures that a generated grid has valid symbols and a consistent
rectangular shape.
A valid grid:
     • Uses only the symbols {0, 1, L, S};
     • Contains values within allowed numeric bounds;
     • Has equal row lengths (rectangular shape).
1.  Valid grid
Completion:    [ 0 , L , 0 ],
               [ 0 , 1 , 0 ],
               [ 0 , S , 0 ]
Expected:  True
(All symbols valid, shape consistent.)
2.  Invalid character (X)
Completion:    [ 0 , X , 0 ],
               [ 0 , 1 , 0 ],
               [ 0 , S , 0 ]
Expected:  False
(Unrecognized symbol X.)
3.  Invalid character (P)
Completion:    [ 0 , L , 0 ],
               [ 0 , 1 , 0 ],
               [ 0 , P , 0 ]
Expected:  False
(Unrecognized symbol P.)
4.  Out-of-range value (-1)
Completion:    [ 0 , L , 0 ],
               [ 0 , -1 , 0 ],
               [ 0 , S , 0 ]
Expected:  False
(Negative numeric value not allowed.)
5.  Out-of-range value (2)
Completion:    [ 0 , L , 0 ],
               [ 0 , 2 , 0 ],
               [ 0 , S , 0 ]
Expected:  False
(Value exceeds permitted range.)
6.  Non-rectangular grid
Completion:    [ 0 , L , 0 ],
               [ 0 , 1 ],
               [ 0 , S , 0 ]
Expected:  False
(Inconsistent row lengths.)
Interpretation:
This check ensures that downstream metrics operate on well-formed
grids only.
Any invalid symbol, numeric range violation, or non-rectangular
structure
results in a False validity flag.
```

G.3.2 LOAD-SUPPORT CONNECTIVITY EXAMPLES

Load-Support Connectivity Examples

```
These tests validate is_load_supported and
is_load_supported_force_directional, which determine whether
loads (L) are connected to supports (S) through solid
cells (1, L, S). The directional variant allows
only gravity-aligned or lateral connections.
1.  Perfect vertical connection
Completion:     [ 0 , L , 0 ],
                [ 0 , 1 , 0 ],
                [ 0 , S , 0 ]
Expected:  True (both directional & non-directional)
2.  Diagonal bridge
Completion:     [ 0 , L , 0 ],
                [ 1 , 1 , 0 ],
                [ 0 , S , S ]
Expected:  True (connected diagonally)
3.  Horizontal load alignment
Completion:     [ 0 , L , L ],
                [ 0 , 1 , 0 ],
                [ 0 , S , 0 ]
Expected:  True (non-directional)
4.  Incomplete bridge
Completion:     [ 0 , L , L ],
                [ 1 , 0 , 0 ],
                [ 0 , S , 0 ]
Expected:  True (non-directional), False (directional)
5.  Disconnected load
Completion:     [ 0 , 0 , L ],
                [ 1 , 0 , 0 ],
                [ 0 , S , 0 ]
Expected:  False (no path)
6.  Complex multi-load structure
Completion:     [ 1 , 1 , 1 ,  0 , 1 , L ],
                [ 1 , 0 , 1 ,  0 , 1 , 0 ],
                [ 1 , 0 , 1 ,  1 , 1 , 0 ],
                [ 1 , 0 , 0 ,  0 , 0 , 0 ],
                [ S , 0 , 0 ,  0 , 0 , 0 ]
Expected:  True (non-directional), False (directional)
Interpretation:
The directional test approximates gravity-aligned force flow, while
the
non-directional variant checks only geometric reachability.
Disconnected or upward-only paths yield False.
```

G.3.3 ISOLATED CLUSTER COUNT EXAMPLES

Isolated Cluster Count Examples

```
This test validates get_isolated_clusters_count, which counts
solid regions (1) disconnected from any load (L) or
support (S). A higher count indicates fragmented or non-functional
material regions.
1.  Single isolated column (1)
Completion:     [ L , 0 , 0 ,  0 , 0 , 0 ],
                [ 1 , 0 , 0 ,  0 , 0 , 0 ],
                [ 1 , 0 , 1 ,  0 , 0 , 0 ],
                [ 1 , 0 , 0 ,  0 , 0 , 0 ],
                [ 1 , 0 , 0 ,  0 , 0 , 0 ],
                [ S , 0 , 0 ,  0 , 0 , 0 ]
Expected:  1
2.  Slightly connected cluster (1)
Completion:     [ L , 0 , 0 ,  0 , 0 , 0 ],
                [ 1 , 0 , 0 ,  0 , 0 , 0 ],
                [ 1 , 0 , 1 ,  1 , 0 , 0 ],
                [ 1 , 0 , 0 ,  1 , 0 , 0 ],
                [ 1 , 0 , 0 ,  0 , 0 , 0 ],
                [ S , 0 , 0 ,  0 , 0 , 0 ]
Expected:  1
3.  Two isolated clusters (2)
Completion:     [ L , 0 , 0 ,  0 , 0 , 1 ],
                [ 1 , 0 , 0 ,  0 , 0 , 1 ],
                [ 1 , 0 , 1 ,  1 , 0 , 1 ],
                [ 1 , 0 , 0 ,  1 , 0 , 0 ],
                [ 1 , 0 , 0 ,  0 , 0 , 0 ],
                [ S , 0 , 0 ,  0 , 0 , 0 ]
Expected:  2
4.  Multiple detached clusters (3)
Completion:     [ L , 0 , 0 ,  1 , 0 , 0 ],
                [ 1 , 0 , 0 ,  0 , 0 , 0 ],
                [ 1 , 0 , 1 ,  1 , 0 , 0 ],
                [ 1 , 0 , 0 ,  1 , 0 , 1 ],
                [ 1 , 0 , 0 ,  0 , 0 , 0 ],
                [ S , 0 , 0 ,  0 , 0 , 0 ]
Expected:  3
Interpretation:
Isolated clusters represent solid islands that do not participate in
load-support transfer.  Lower counts indicate more integrated and
structurally valid predictions.
```

G.3.4 Difficulty Score (DWCS) Examples

Difficulty Score (DWCS) Examples

```
This test validates get_difficulty_score, which computes the
average difficulty of masked (V) cells in the input grid
based on their ground-truth neighborhood configuration in the GT grid.
The completion grid is used to confirm the reconstruction context.
Higher scores correspond to more complex masked regions.
1.  Simple vertical case (2.0)
Input:          [ 0 , L , 0 ],
                [ 0 , V , 0 ],
                [ 0 , S , 0 ]
Ground truth:   [ 0 , L , 0 ],
                [ 0 , 1 , 0 ],
                [ 0 , S , 0 ]
Completion:     [ 0 , L , 0 ],
                [ 0 , 1 , 0 ],
                [ 0 , S , 0 ]
Expected output:  2.000
2.  Mixed neighborhood (3.0)
Input:          [ 0 , L , 0 ],
                [ V , 1 , 0 ],
                [ 0 , S , 0 ]
Ground truth:   [ 0 , L , 0 ],
                [ 0 , 1 , 0 ],
                [ 0 , S , 0 ]
Completion:     [ 0 , L , 0 ],
                [ 1 , 1 , 0 ],
                [ 0 , S , 0 ]
Expected output:  3.000
3.  Large structure (1.0)
Input:          [ 0 , L , 0 ],
                [ 0 , 1 , 0 ],
                [ 0 , 1 , V ],
                [ 0 , 1 , 0 ],
                [ 0 , S , 0 ]
Ground truth:   [ 0 , L , 0 ],
                [ 0 , 1 , 0 ],
                [ 0 , 1 , 0 ],
                [ 0 , 1 , 0 ],
                [ 0 , S , 0 ]
Completion:     [ 0 , L , 0 ],
                [ 0 , 1 , 0 ],
                [ 0 , 1 , 0 ],
                [ 0 , 1 , 0 ],
                [ 0 , S , 0 ]
Expected output:  1.000
4.  Dense structure with boundary void (3.0)
Input:          [ 0 , L , L ,  L , 0 ],
                [ 0 , 1 , 1 ,  1 , 0 ],
                [ 0 , 1 , V ,  1 , 0 ],
                [ 0 , 1 , 1 ,  1 , 0 ],
                [ 0 , S , S ,  S , 0 ]
Ground truth:   [ 0 , L , L ,  L , 0 ],
                [ 0 , 1 , 1 ,  0 , 0 ],
                [ 0 , 1 , 0 ,  1 , 0 ],
                [ 0 , 1 , 1 ,  0 , 0 ],
                [ 0 , S , S ,  S , 0 ]
Completion:     [ 0 , L , L ,  L , 0 ],
                [ 0 , 1 , 1 ,  0 , 0 ],
                [ 0 , 1 , 0 ,  1 , 0 ],
                [ 0 , 1 , 1 ,  0 , 0 ],
                [ 0 , S , S ,  S , 0 ]
```

```
Expected output:  3.000
Interpretation:
The score increases when masked cells (V) occur in ambiguous or mixed
regions, particularly around structural boundaries.  Uniform
neighborhoods yield
lower scores, reflecting easier reconstruction.
```

G.4 Physics Approximation Metric

G.4.1 Force Path Cost Examples

Force Path Cost Examples

```
This test validates the
get_total_force_path_cost_average_efficiency_ratio function, which
computes the Force Path Cost Average Efficiency Ratio (FPCEff).
Higher ratios indicate more efficient and physically plausible
load-support paths aligned with gravity.
Gravity direction:  (1, 0) downward
Test cases:
1.  Perfect vertical alignment
Ground truth:  [ 0 , L , 0 ],
               [ 0 , 1 , 0 ],
               [ 0 , S , 0 ]
Prediction:    [ 0 , L , 0 ],
               [ 0 , 1 , 0 ],
               [ 0 , S , 0 ]
Expected output:  1.000
2.  Slightly wider vertical column (still efficient)
Ground truth:  [ 0 , L , 0 ],
               [ 0 , 1 , 0 ],
               [ 0 , S , 0 ]
Prediction:    [ 0 , L , 0 ],
               [ 1 , 1 , 0 ],
               [ 0 , S , 0 ]
Expected output:  1.000
3.  Offset load-support connection (less efficient)
Ground truth:  [ 0 , L , 0 ],
               [ 0 , 1 , 0 ],
               [ S , 0 , 0 ]
Prediction:    [ 0 , L , 0 ],
               [ 1 , 1 , 0 ],
               [ S , 0 , 0 ]
Expected output:  0.8037
4.  Broken vertical link (similar inefficiency)
Ground truth:  [ 0 , L , 0 ],
               [ 0 , 1 , 0 ],
               [ S , 0 , 0 ]
Prediction:    [ 0 , L , 0 ],
               [ 1 , 0 , 0 ],
               [ S , 0 , 0 ]
Expected output:  0.8037
5.  Horizontally displaced load (least efficient)
Ground truth:  [ 0 , 0 , L ],
               [ 0 , 1 , 0 ],
               [ S , 0 , 0 ]
Prediction:    [ 0 , 1 , L ],
               [ 1 , 0 , 0 ],
               [ S , 0 , 0 ]
Expected output:  0.7724
Interpretation:
As load-support paths deviate from the gravity direction or become
discontinuous,
FPCEff decreases from 1.0 toward 0, reflecting reduced
physical plausibility of the structure.
```

H TOPOLOGY OPTIMIZATION SAMPLE PLOTS

H.1 2D SAMPLES

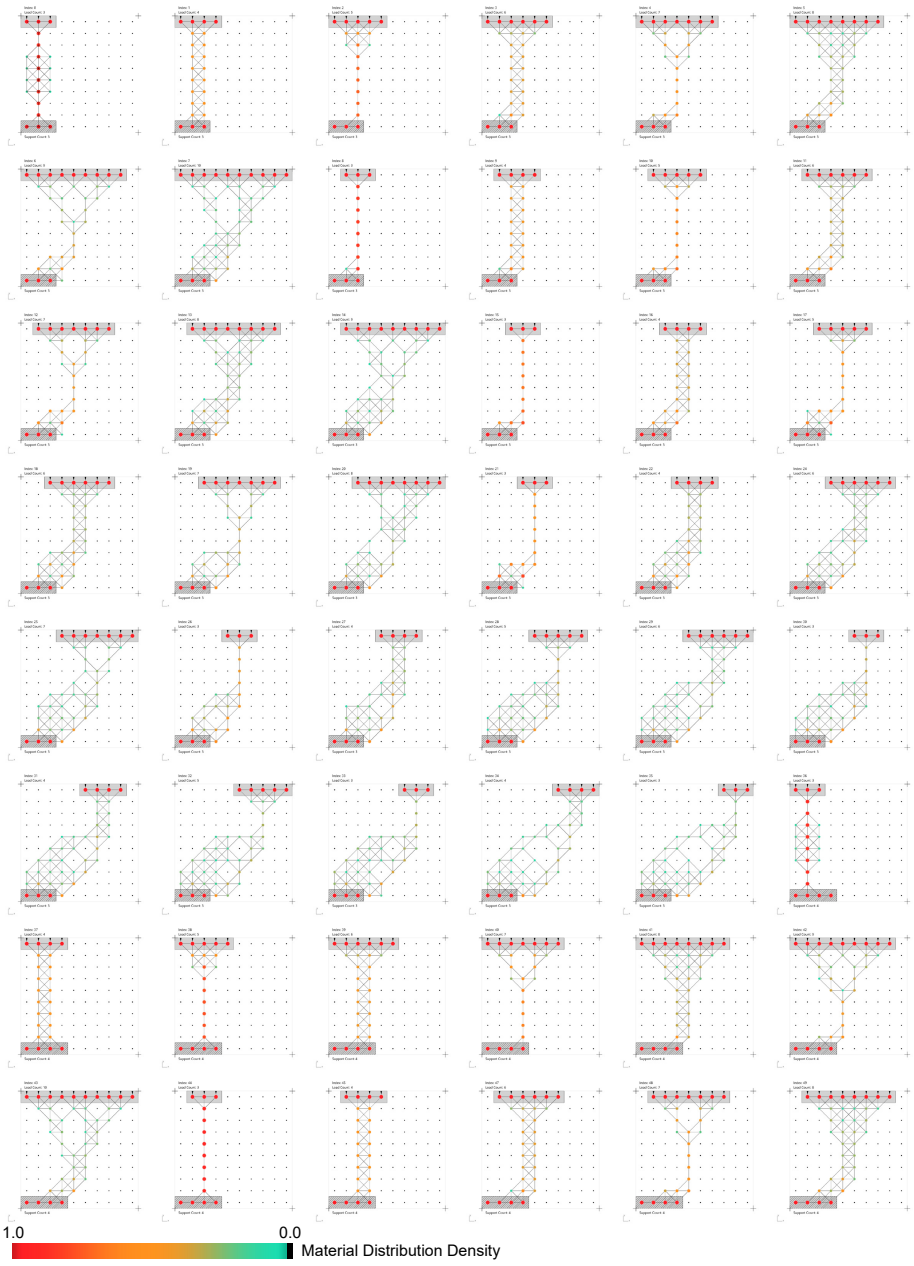

Figure 5: Example 2D topology optimization samples from the SPhyR dataset.

I COMPLETE RESULTS FOR ALL METRICS

Table 8: All evaluation metrics on the SPhyR 2D Easy tasks, for every model. The body reports the four metrics the analysis turns on; this is the complete set.

Task	Metric	GPT 4.1	Claude Opus 4	Gemini 2.5 Pro	DeepSeek-R1	Perplexity Sonar
1 Random Cell	Exact Match ↑	26	**82**	81	78	52
	Difference Ratio (%) ↑	95.47	**99.05**	99.03	98.92	93.28
	Valid Output Grid (%) ↑	**100.00**	**100.00**	**100.00**	**100.00**	96.00
	Load-Support Connectivity (%) ↑	**100.00**	**100.00**	**100.00**	**100.00**	95.00
	Average Isolated Clusters Count ↓	**0.32**	0.00	0.00	0.00	0.16
	Topology Score (%) ↑	97.57	99.54	**99.61**	96.91	92.39
	Structural Efficiency (%) ↑	**100.00**	99.86	**100.00**	96.98	93.52
	Material Efficiency (%) ↑	97.57	99.68	99.61	**99.93**	93.88
	Load Carrying (%) ↑	**100.00**	**100.00**	**100.00**	**100.00**	95.00
5 Random Cells	Exact Match ↑	1	**45**	39	33	10
	Difference Ratio (%) ↑	75.87	**95.76**	95.08	94.33	75.79
	Valid Output Grid (%) ↑	**100.00**	**100.00**	**100.00**	**100.00**	87.00
	Load-Support Connectivity (%) ↑	**100.00**	**100.00**	**100.00**	99.00	80.00
	Average Isolated Clusters Count ↓	**1.64**	0.00	0.00	0.01	0.37
	Topology Score (%) ↑	88.60	97.58	**98.27**	90.13	66.56
	Structural Efficiency (%) ↑	**99.33**	97.78	99.04	90.18	68.93
	Material Efficiency (%) ↑	89.24	98.80	99.22	**99.93**	75.60
	Load Carrying (%) ↑	**100.00**	99.00	**100.00**	**100.00**	78.00
10 Random Cells	Exact Match ↑	0	**15**	13	4	1
	Difference Ratio (%) ↑	59.82	**89.08**	88.88	86.41	72.68
	Valid Output Grid (%) ↑	99.00	**100.00**	**100.00**	**100.00**	92.00
	Load-Support Connectivity (%) ↑	99.00	**100.00**	**100.00**	99.00	76.00
	Average Isolated Clusters Count ↓	**2.27**	0.00	0.00	0.00	0.40
	Topology Score (%) ↑	81.44	93.92	**96.39**	76.60	56.85
	Structural Efficiency (%) ↑	97.31	95.33	**98.39**	77.01	58.70
	Material Efficiency (%) ↑	83.10	**98.58**	98.00	97.59	77.07
	Load Carrying (%) ↑	99.00	**100.00**	**100.00**	98.00	79.00
1 Random Row	Exact Match ↑	20	**52**	44	50	21
	Difference Ratio (%) ↑	84.50	**94.92**	93.90	92.66	77.87
	Valid Output Grid (%) ↑	97.00	**100.00**	**100.00**	98.00	91.00
	Load-Support Connectivity (%) ↑	97.00	**100.00**	**100.00**	98.00	73.00
	Average Isolated Clusters Count ↓	0.02	0.00	0.00	0.00	**0.07**
	Topology Score (%) ↑	88.40	94.58	**94.79**	88.66	65.20
	Structural Efficiency (%) ↑	94.39	96.19	**97.41**	89.87	69.56
	Material Efficiency (%) ↑	89.98	**97.38**	96.38	96.79	67.63
	Load Carrying (%) ↑	96.00	**99.00**	**99.00**	98.00	72.00
3 Random Rows	Exact Match ↑	6	**35**	29	32	24
	Difference Ratio (%) ↑	59.31	**88.64**	84.09	86.36	74.23
	Valid Output Grid (%) ↑	99.00	**100.00**	99.00	99.00	**100.00**
	Load-Support Connectivity (%) ↑	95.00	**100.00**	98.00	99.00	74.00
	Average Isolated Clusters Count ↓	0.01	0.00	0.01	0.00	**0.28**
	Topology Score (%) ↑	67.69	**87.18**	85.39	82.56	59.16
	Structural Efficiency (%) ↑	79.93	**89.50**	89.47	85.16	63.60
	Material Efficiency (%) ↑	73.17	**93.67**	90.82	93.35	62.56
	Load Carrying (%) ↑	86.00	**96.00**	95.00	**96.00**	67.00
1 Random Column	Exact Match ↑	0	23	**26**	**26**	12
	Difference Ratio (%) ↑	51.92	83.09	81.95	**87.57**	67.71
	Valid Output Grid (%) ↑	**100.00**	**100.00**	98.00	**100.00**	97.00
	Load-Support Connectivity (%) ↑	**100.00**	**100.00**	98.00	**100.00**	88.00
	Average Isolated Clusters Count ↓	**0.13**	0.00	0.00	0.00	0.06
	Topology Score (%) ↑	78.50	**91.49**	89.01	90.12	72.58
	Structural Efficiency (%) ↑	**98.02**	97.02	92.85	92.02	79.79
	Material Efficiency (%) ↑	80.40	93.47	93.16	**98.09**	86.51
	Load Carrying (%) ↑	**100.00**	99.00	97.00	**100.00**	94.00
3 Random Columns	Exact Match ↑	1	5	3	**6**	1
	Difference Ratio (%) ↑	2.46	58.69	52.03	**62.35**	24.01
	Valid Output Grid (%) ↑	99.00	**100.00**	98.00	99.00	93.00
	Load-Support Connectivity (%) ↑	97.00	94.00	96.00	**98.00**	70.00
	Average Isolated Clusters Count ↓	**0.32**	0.00	0.02	0.00	0.24
	Topology Score (%) ↑	61.19	**78.43**	73.52	70.45	54.62
	Structural Efficiency (%) ↑	**87.35**	86.37	80.47	72.75	66.10
	Material Efficiency (%) ↑	68.63	85.03	90.97	**95.49**	68.29
	Load Carrying (%) ↑	95.00	93.00	**98.00**	**98.00**	80.00
Full	Exact Match ↑	**0**	**0**	**0**	**0**	**0**
	Difference Ratio (%) ↑	-62.06	-35.78	-25.03	**-1.36**	-49.16
	Valid Output Grid (%) ↑	**100.00**	**100.00**	99.00	99.00	81.00
	Load-Support Connectivity (%) ↑	94.00	94.00	**98.00**	**98.00**	42.00
	Average Isolated Clusters Count ↓	0.01	0.00	0.01	0.01	**0.09**
	Topology Score (%) ↑	33.01	43.61	**48.42**	44.58	10.11
	Structural Efficiency (%) ↑	53.94	56.30	**72.53**	63.17	17.29
	Material Efficiency (%) ↑	50.24	67.49	62.41	**79.14**	22.96
	Load Carrying (%) ↑	76.00	82.00	89.00	**99.00**	33.00
Average	Exact Match ↑	6.75	**32.12**	29.38	28.62	15.12
	Difference Ratio (%) ↑	45.91	71.68	71.24	**75.90**	54.55
	Valid Output Grid (%) ↑	99.25	**100.00**	99.25	99.38	92.12
	Load-Support Connectivity (%) ↑	97.75	98.50	98.75	**98.88**	74.75
	Average Isolated Clusters Count ↓	**0.59**	0.00	0.01	0.00	0.21
	Topology Score (%) ↑	74.55	**85.79**	85.67	80.00	59.69
	Structural Efficiency (%) ↑	88.78	89.79	**91.27**	83.39	64.68
	Material Efficiency (%) ↑	79.04	91.76	91.32	**95.04**	69.31
	Load Carrying (%) ↑	94.00	96.00	97.25	**98.62**	74.75

Table 9: All evaluation metrics on the SPhyR 2D Hard tasks, for every model. The body reports the four metrics the analysis turns on; this is the complete set.

Task	Metric	GPT 4.1	Claude Opus 4	Gemini 2.5 Pro	DeepSeek-R1	Perplexity Sonar
1 Random Cell	Exact Match ↑	13	**77**	76	75	13
	Difference Ratio (%) ↑	88.14	95.45	**96.70**	95.06	80.07
	Valid Output Grid (%) ↑	**100.00**	99.00	**100.00**	99.00	94.00
	Load-Support Connectivity (%) ↑	99.00	98.00	**100.00**	98.00	93.00
	Average Isolated Clusters Count ↓	**0.44**	0.00	0.01	0.00	0.37
	Topology Score (%) ↑	97.97	96.88	**99.04**	96.75	91.54
	Structural Efficiency (%) ↑	98.87	96.97	**99.13**	96.91	92.74
	Material Efficiency (%) ↑	98.11	96.91	**99.91**	97.84	92.80
	Load Carrying (%) ↑	99.00	97.00	**100.00**	98.00	94.00
5 Random Cells	Exact Match ↑	0	**38**	37	25	3
	Difference Ratio (%) ↑	35.23	**87.27**	83.19	82.89	38.29
	Valid Output Grid (%) ↑	99.00	**100.00**	99.00	**100.00**	85.00
	Load-Support Connectivity (%) ↑	99.00	**100.00**	98.00	**100.00**	80.00
	Average Isolated Clusters Count ↓	**1.88**	0.00	0.03	0.01	1.22
	Topology Score (%) ↑	86.75	94.99	**96.85**	90.32	67.50
	Structural Efficiency (%) ↑	92.12	95.45	**97.34**	90.72	69.33
	Material Efficiency (%) ↑	93.47	98.54	**99.50**	98.60	84.11
	Load Carrying (%) ↑	99.00	99.00	**100.00**	99.00	86.00
10 Random Cells	Exact Match ↑	0	13	**14**	4	0
	Difference Ratio (%) ↑	-26.62	**69.70**	67.83	64.98	-1.98
	Valid Output Grid (%) ↑	99.00	99.00	**100.00**	**100.00**	64.00
	Load-Support Connectivity (%) ↑	99.00	93.00	**100.00**	**100.00**	58.00
	Average Isolated Clusters Count ↓	**2.82**	0.01	0.02	0.00	1.20
	Topology Score (%) ↑	74.22	87.05	**93.72**	86.64	37.63
	Structural Efficiency (%) ↑	83.22	87.63	**95.03**	87.29	39.15
	Material Efficiency (%) ↑	89.21	92.42	**98.67**	98.33	59.18
	Load Carrying (%) ↑	99.00	93.00	**100.00**	99.00	61.00
1 Random Row	Exact Match ↑	2	**49**	46	44	7
	Difference Ratio (%) ↑	18.44	**80.55**	71.69	76.09	27.07
	Valid Output Grid (%) ↑	98.00	**100.00**	**100.00**	**100.00**	91.00
	Load-Support Connectivity (%) ↑	97.00	99.00	**100.00**	99.00	85.00
	Average Isolated Clusters Count ↓	0.01	0.00	0.00	0.00	**0.05**
	Topology Score (%) ↑	85.97	94.98	93.93	**95.56**	65.16
	Structural Efficiency (%) ↑	94.39	96.53	95.48	**97.11**	69.15
	Material Efficiency (%) ↑	88.43	96.44	97.44	**97.44**	79.62
	Load Carrying (%) ↑	97.00	98.00	**99.00**	**99.00**	84.00
3 Random Rows	Exact Match ↑	0	**21**	12	19	1
	Difference Ratio (%) ↑	-162.36	**32.01**	18.75	20.58	-106.35
	Valid Output Grid (%) ↑	**100.00**	**100.00**	**100.00**	99.00	98.00
	Load-Support Connectivity (%) ↑	**100.00**	97.00	**100.00**	99.00	96.00
	Average Isolated Clusters Count ↓	**0.03**	0.00	0.00	0.00	0.01
	Topology Score (%) ↑	68.45	83.63	80.08	**84.28**	59.04
	Structural Efficiency (%) ↑	**87.57**	86.86	85.45	87.24	69.00
	Material Efficiency (%) ↑	79.43	89.61	85.53	**94.76**	80.85
	Load Carrying (%) ↑	**100.00**	93.00	91.00	98.00	93.00
1 Random Column	Exact Match ↑	0	21	**26**	19	3
	Difference Ratio (%) ↑	-34.24	37.46	**44.93**	31.26	-26.12
	Valid Output Grid (%) ↑	97.00	99.00	**100.00**	98.00	91.00
	Load-Support Connectivity (%) ↑	97.00	97.00	**100.00**	98.00	90.00
	Average Isolated Clusters Count ↓	**0.07**	0.00	0.00	0.00	0.04
	Topology Score (%) ↑	73.06	81.73	**91.73**	80.07	70.66
	Structural Efficiency (%) ↑	88.62	87.60	**95.82**	85.06	81.93
	Material Efficiency (%) ↑	79.14	89.12	**95.90**	90.97	78.32
	Load Carrying (%) ↑	95.00	95.00	**100.00**	96.00	90.00
3 Random Columns	Exact Match ↑	0	**7**	3	**7**	1
	Difference Ratio (%) ↑	-238.99	**-17.63**	-56.28	-31.77	-142.17
	Valid Output Grid (%) ↑	94.00	**100.00**	**100.00**	99.00	80.00
	Load-Support Connectivity (%) ↑	94.00	93.00	96.00	**97.00**	61.00
	Average Isolated Clusters Count ↓	0.30	0.00	0.02	0.00	**0.32**
	Topology Score (%) ↑	56.78	**71.54**	70.69	68.14	53.38
	Structural Efficiency (%) ↑	**77.75**	76.91	77.52	73.94	63.07
	Material Efficiency (%) ↑	68.44	85.20	87.69	**88.81**	68.66
	Load Carrying (%) ↑	91.00	91.00	**95.00**	**95.00**	80.00
Full	Exact Match ↑	**0**	**0**	**0**	**0**	**0**
	Difference Ratio (%) ↑	-816.91	**-466.42**	-548.98	-470.16	-537.83
	Valid Output Grid (%) ↑	**100.00**	**100.00**	99.00	96.00	76.00
	Load-Support Connectivity (%) ↑	**100.00**	88.00	98.00	93.00	49.00
	Average Isolated Clusters Count ↓	0.00	0.00	0.01	0.03	**0.08**
	Topology Score (%) ↑	23.18	32.95	**37.94**	33.12	8.76
	Structural Efficiency (%) ↑	52.10	69.37	**95.85**	84.58	19.23
	Material Efficiency (%) ↑	**45.79**	40.05	39.50	36.86	21.30
	Load Carrying (%) ↑	97.00	81.00	**99.00**	90.00	42.00
Average	Exact Match ↑	1.88	**28.25**	26.75	24.12	3.50
	Difference Ratio (%) ↑	-142.16	**-10.20**	-27.77	-16.38	-83.63
	Valid Output Grid (%) ↑	98.38	99.62	**99.75**	98.88	84.88
	Load-Support Connectivity (%) ↑	98.12	95.62	**99.00**	98.00	76.50
	Average Isolated Clusters Count ↓	**0.69**	0.00	0.01	0.01	0.41
	Topology Score (%) ↑	70.80	80.47	**83.00**	79.36	56.71
	Structural Efficiency (%) ↑	84.33	87.16	**92.70**	87.86	62.95
	Material Efficiency (%) ↑	80.25	86.04	**88.02**	87.95	70.61
	Load Carrying (%) ↑	97.12	93.38	**98.00**	96.75	78.75

J Additional Main Run Results

J.1 Results for All Models and Tasks

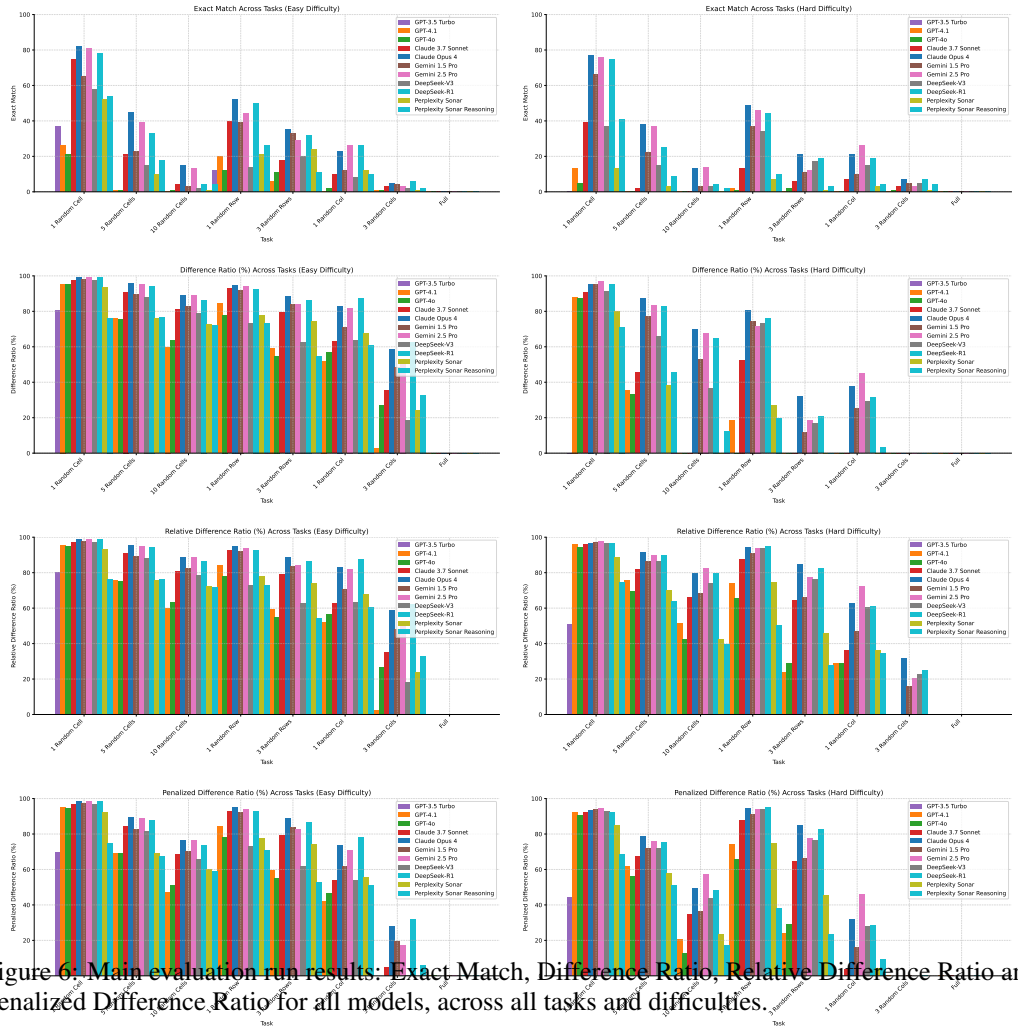

Figure 6: Main evaluation run results: Exact Match, Difference Ratio, Relative Difference Ratio and Penalized Difference Ratio for all models, across all tasks and difficulties.

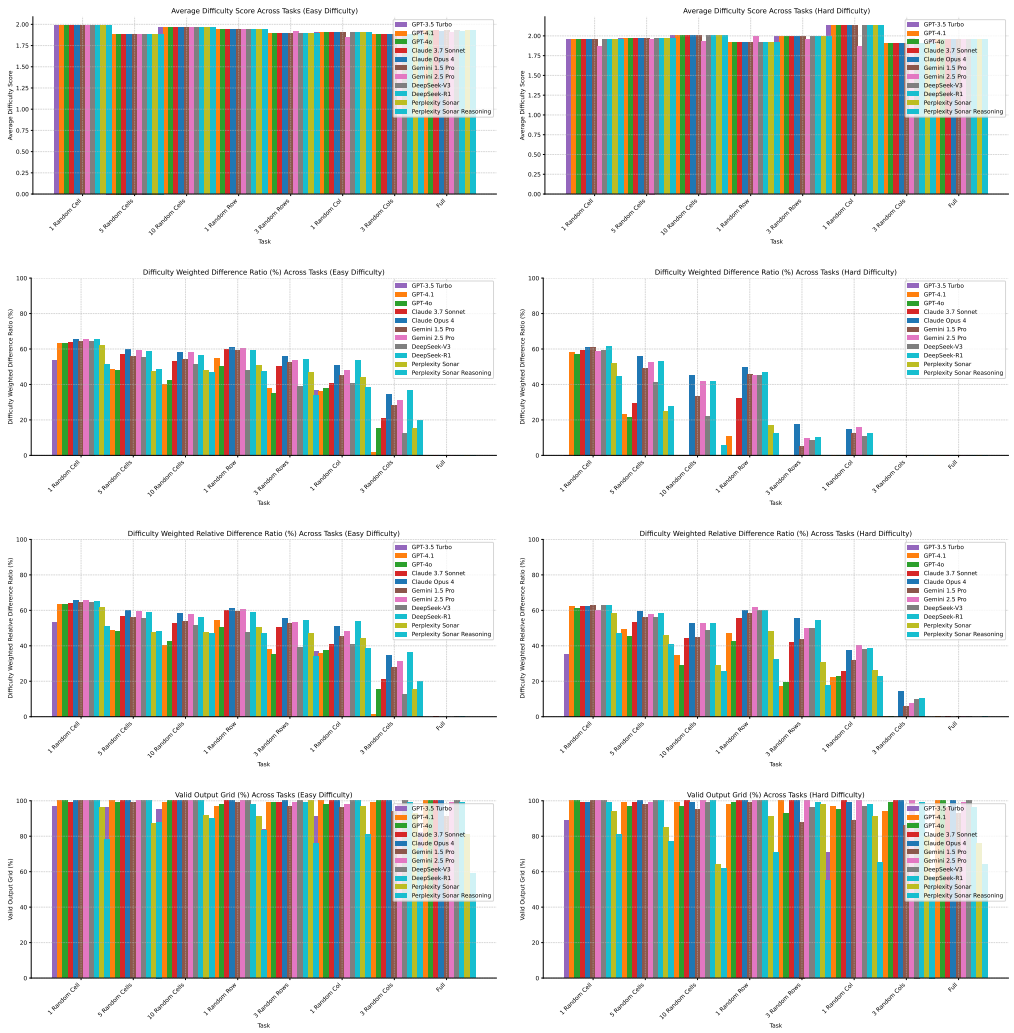

Figure 7: Main evaluation run results: Average Difficulty Score, Difficulty Weighted Difference Ratio, Difficulty Weighted Relative Difference Ratio and Grid Validity for all models, across all tasks and difficulties.

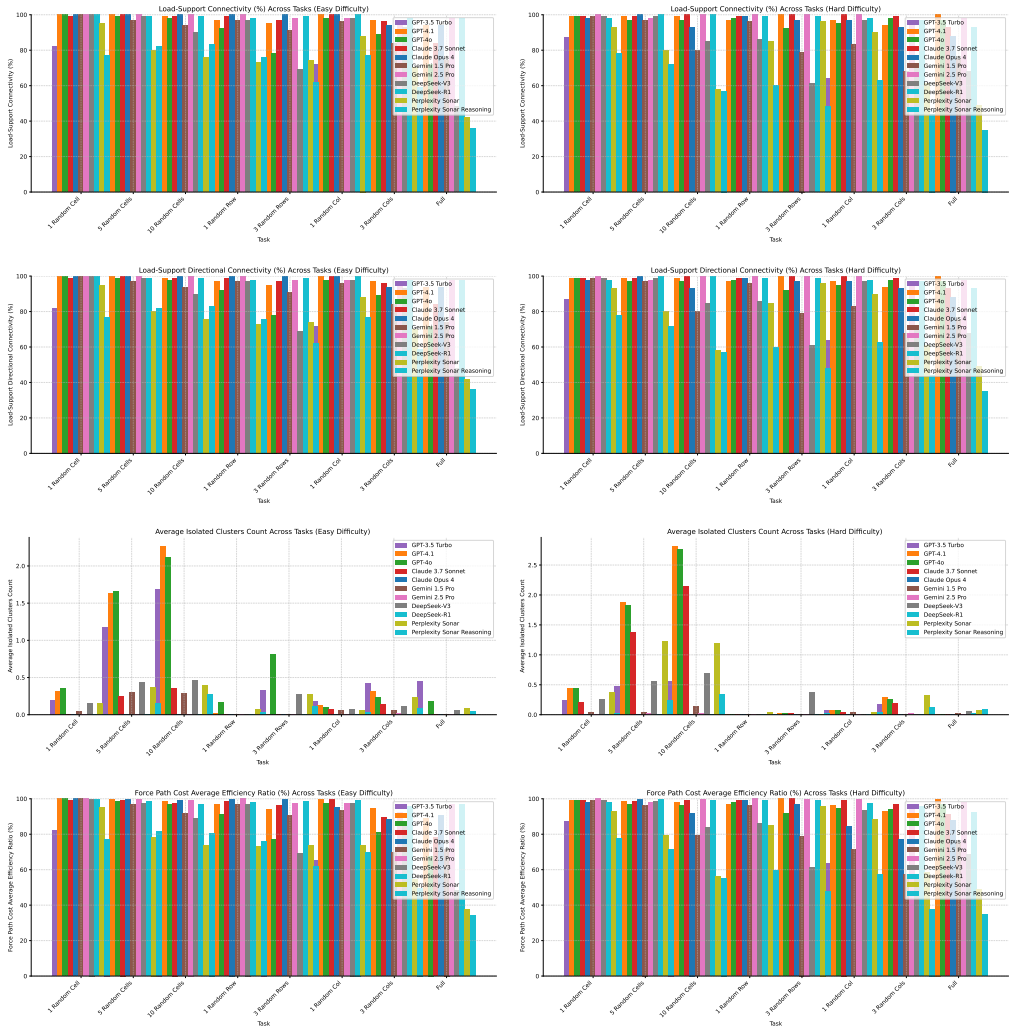

Figure 8: Main evaluation run results: Load-Support Connectivity, Directional Load-Support Connectivity, Average Isolated Cluster Count and Force Path Cost Average Efficiency Ratio for all models, across all tasks and difficulties.

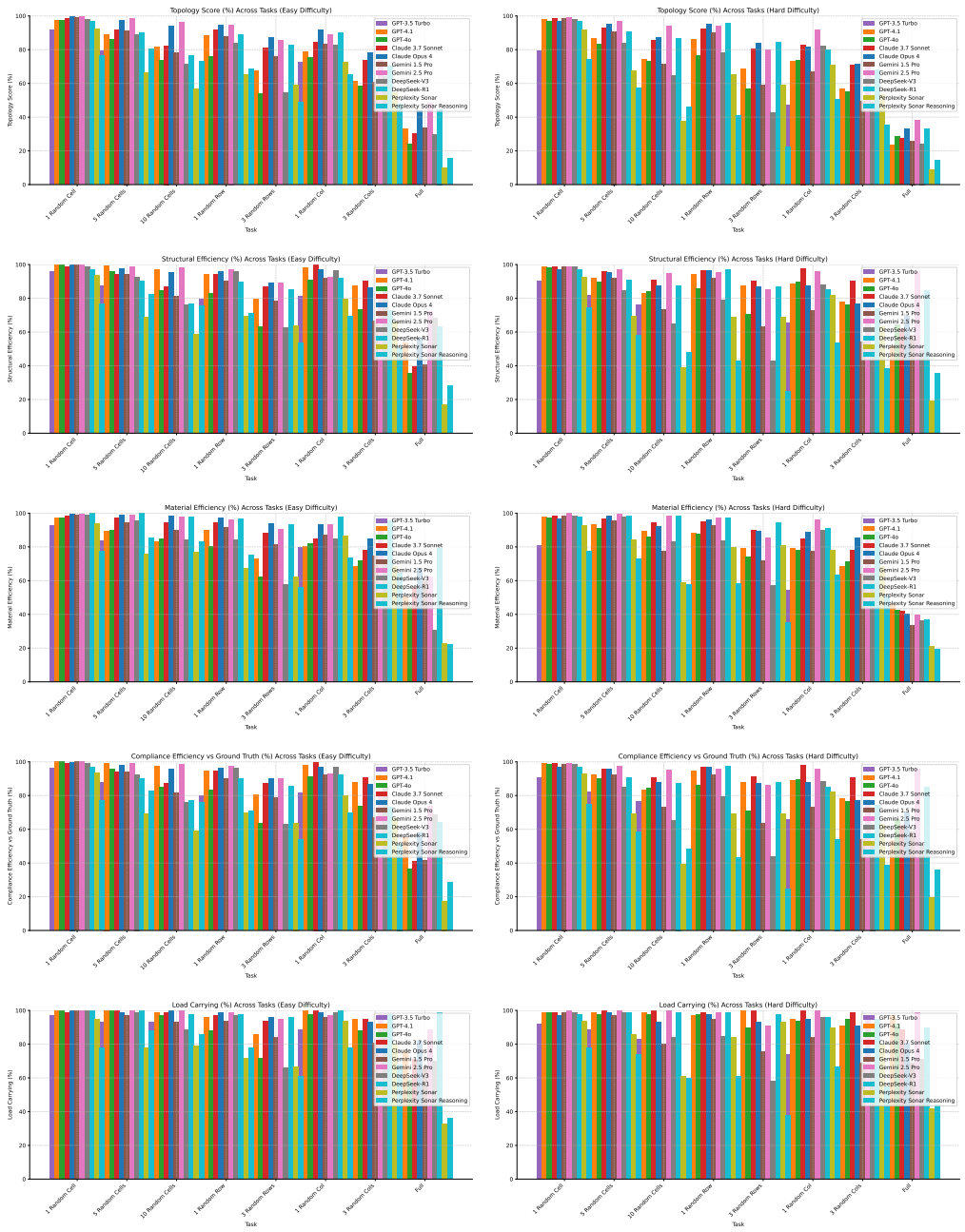

Figure 9: Main evaluation run results, simulated structural metrics: Topology Score, Structural Efficiency, Material Efficiency, Compliance Efficiency vs Ground Truth and Load Carrying for all models, across all tasks and difficulties.

K Additional Rotation Experiment Results

K.1 Prompt and Completion - 1 Random Cell Easy: Rotated x 3

Prompt and Completion - 1 Random Cell Easy: Rotated x 3

Prompt:

```
You are given a structural material distribution represented as a grid.  Each cell can
have one of the following states:
- 'L' indicates applied load.
- 'V' indicates void.
- 'S' indicates support.

The goal is to predict the correct material distribution by filling in all 'V' cells
with either '1' (solid) or '0' (empty), based on the surrounding structure and implicit
physical reasoning (such as load paths, supports, and forces).

Important:  The completed structure should use as little material as possible while
remaining stable and plausible for carrying the applied forces.  Minimize material usage
unless necessary for structural support.

Below is the input grid with masked regions:

0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 S
0 0 0 0 0 0 0 0 0 S
0 0 0 0 0 0 0 0 0 S
0 0 0 0 0 0 0 0 0 S
L 0 0 0 0 0 1 1 1 S
L 1 V 1 1 1 1 1 1 S
L 0 0 0 0 0 1 1 1 S
0 0 0 0 0 0 0 0 0 S
0 0 0 0 0 0 0 0 0 S

Please output the completed grid by replacing all 'V' cells with either '1' (solid) or '0'
(empty).
Maintain the same format as the input:  one row per line, cells separated by spaces, and
the total number of rows and columns unchanged.
Return only the completed grid without any additional explanation.
```

Model Completion (differences in red):

```
0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 S
0 0 0 0 0 0 0 0 0 S
0 0 0 0 0 0 0 0 0 S
0 0 0 0 0 0 0 0 0 S
L 0 0 0 0 0 1 1 1 S
L 1 1 1 1 1 1 1 1 S
L 0 0 0 0 0 1 1 1 S
0 0 0 0 0 0 0 0 0 S
0 0 0 0 0 0 0 0 0 S
```

Ground Truth:

```
0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 S
0 0 0 0 0 0 0 0 0 S
0 0 0 0 0 0 0 0 0 S
0 0 0 0 0 0 0 0 0 S
L 0 0 0 0 0 1 1 1 S
L 1 1 1 1 1 1 1 1 S
L 0 0 0 0 0 1 1 1 S
0 0 0 0 0 0 0 0 0 S
0 0 0 0 0 0 0 0 0 S
```

K.2 Results for Model Sub-set and Task Sub-set

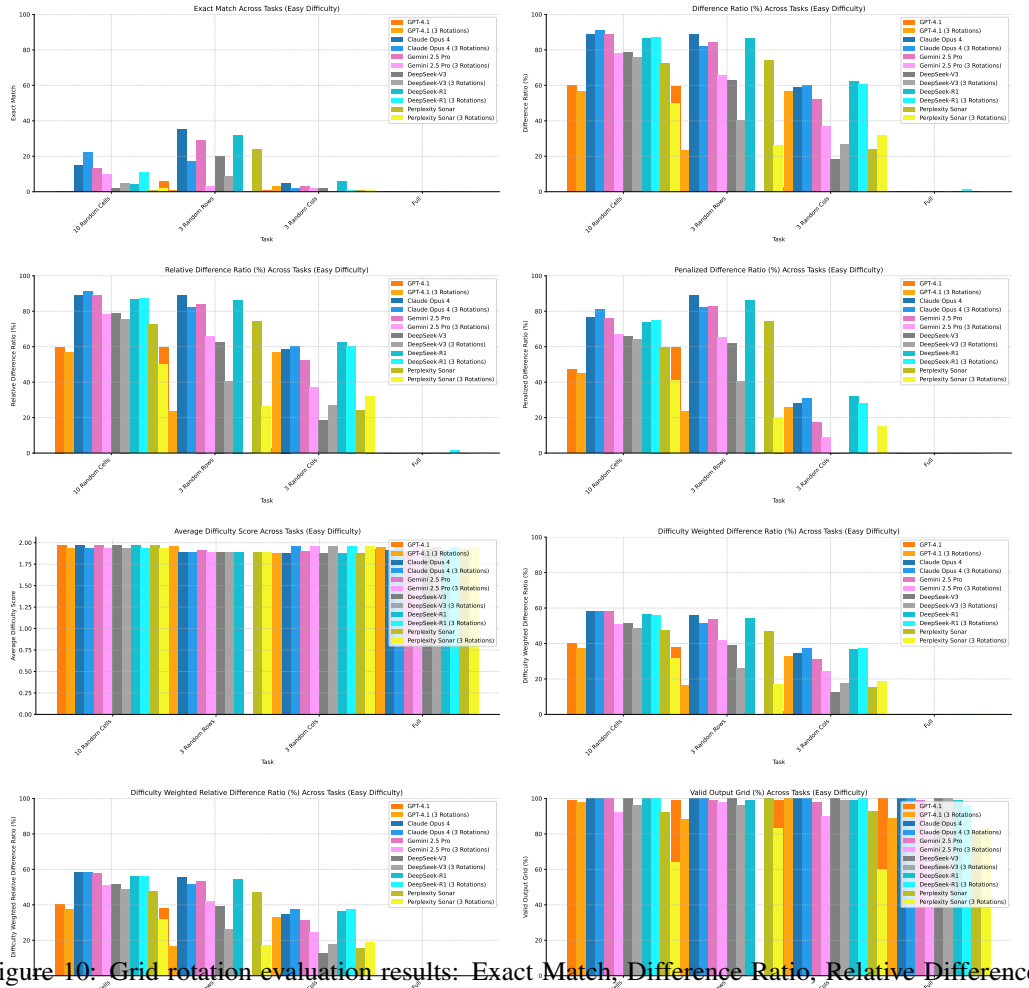

Figure 10: Grid rotation evaluation results: Exact Match, Difference Ratio, Relative Difference Ratio, Penalized Difference Ratio, Average Difficulty Score, Difficulty Weighted Difference Ratio, Difficulty Weighted Relative Difference Ratio and Grid Validity for GPT-4.1, Claude Opus 4, Gemini 2.5 Pro, DeepSeek-V3 and Perplexity, for 10 Random Cells, 3 Random Rows, 3 Random Cells and Full tasks and easy difficulty.

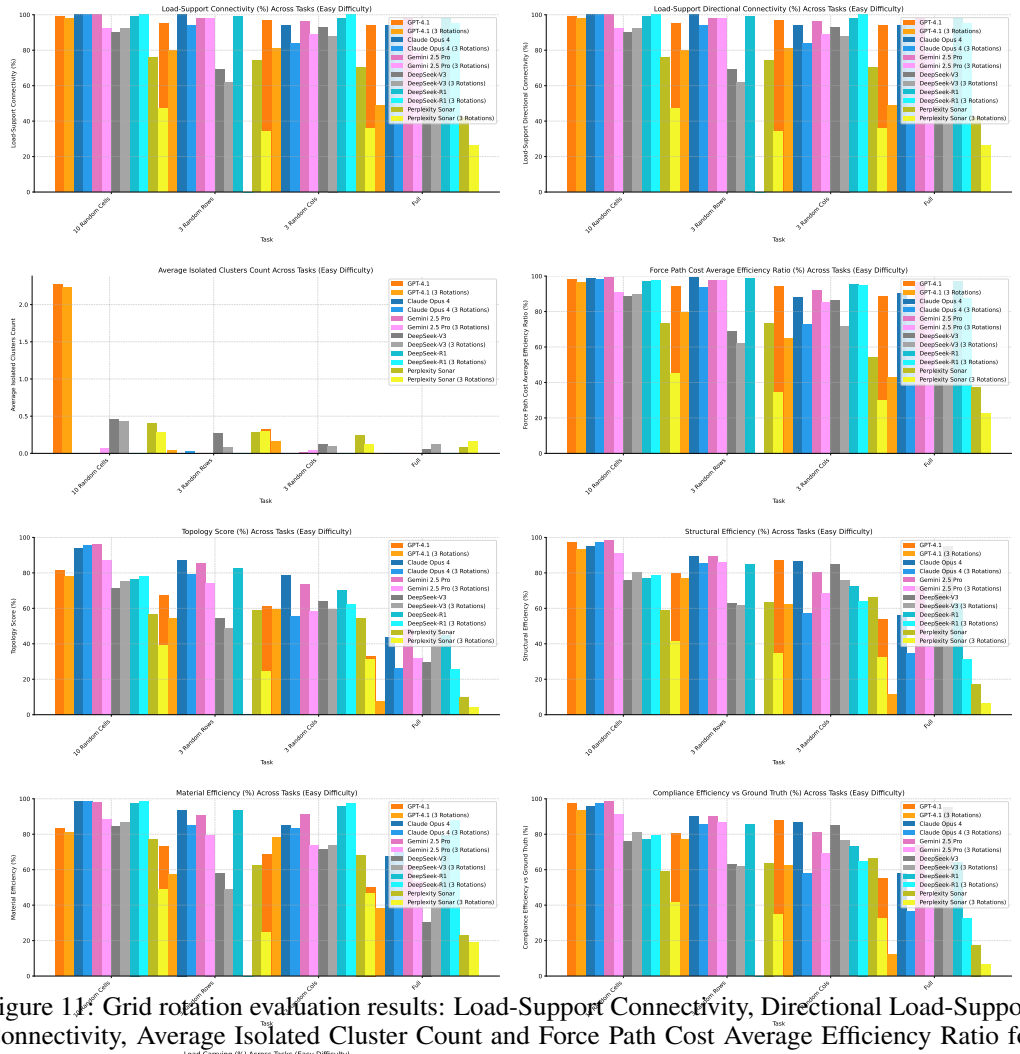

Figure 11: Grid rotation evaluation results: Load-Support Connectivity, Directional Load-Support Connectivity, Average Isolated Cluster Count and Force Path Cost Average Efficiency Ratio for GPT-4.1, Claude Opus 4, Gemini 2.5 Pro, DeepSeek-V3 and Perplexity, for 10 Random Cells, 3 Random Rows, 3 Random Cells and Full tasks and easy difficulty.

Table 10: Grid rotation evaluation results for all metrics, for GPT-4.1, Claude Opus 4, Gemini 2.5 Pro, DeepSeek-V3 and Perplexity, for 10 Random Cells, 3 Random Rows, 3 Random Cells and Full tasks and easy difficulty.

Task	Metric	GPT 4.1	GPT 4.1 (3 Rotations)	Claude Opus 4	Claude Opus 4 (3 Rotations)	Gemini 2.5 Pro	Gemini 2.5 Pro (3 Rotations)	DeepSeek-V3	DeepSeek-V3 (3 Rotations)	Perplexity Sonar	Perplexity Sonar (3 Rotations)
Difficulty: Easy											
10 Random Cells	Exact Match ↑	0	0	15	**22**	13	10	2	5	1	2
	Difference Ratio (%) ↑	59.82	56.85	89.08	**91.02**	88.88	78.26	78.78	75.62	72.68	50.19
	Relative Difference Ratio (%) ↑	59.82	56.85	89.08	**91.02**	88.88	78.26	78.78	75.62	72.68	50.19
	Penalized Difference Ratio (%) ↑	47.02	44.95	76.41	**80.78**	76.21	66.98	65.95	64.27	59.85	41.15
	Average Difficulty Score	**1.97**	1.94	**1.97**	1.94	**1.97**	1.94	**1.97**	1.94	**1.97**	1.94
	Difficulty Weighted Difference Ratio (%) ↑	40.24	37.62	58.23	**58.52**	58.06	50.94	51.72	48.69	47.81	31.74
	Difficulty Weighted Relative Difference Ratio (%) ↑	40.24	37.62	58.23	**58.52**	58.06	50.94	51.72	48.69	47.81	31.74
	Valid Output Grid ↑	99.00	98.00	**100.00**	**100.00**	**100.00**	92.00	**100.00**	96.00	92.00	64.00
	Load-Support Connectivity (%) ↑	99.00	98.00	**100.00**	**100.00**	**100.00**	92.00	90.00	92.00	76.00	47.00
	Load-Support Directional Connectivity (%) ↑	99.00	83.00	**100.00**	86.00	**100.00**	78.00	90.00	72.00	76.00	32.00
	Average Isolated Clusters Count ↓	**2.27**	2.24	0.00	0.00	0.00	0.07	0.46	0.43	0.40	0.29
	Force Path Cost Average Efficiency Ratio (%) ↑	98.28	88.93	99.06	94.41	**99.31**	85.60	88.72	82.77	73.62	47.18
3 Random Rows	Exact Match ↑	6	1	**35**	17	29	3	20	9	24	0
	Difference Ratio (%) ↑	59.31	23.28	**88.64**	82.19	84.09	65.52	62.64	40.38	74.23	26.11
	Relative Difference Ratio (%) ↑	59.31	23.28	**88.64**	82.19	84.09	65.52	62.64	40.38	74.23	26.11
	Penalized Difference Ratio (%) ↑	59.31	23.28	**88.64**	82.19	82.84	65.08	62.04	40.21	73.99	19.33
	Average Difficulty Score	1.89	**1.96**	1.89	1.89	1.92	1.89	1.89	1.89	1.89	1.89
	Difficulty Weighted Difference Ratio (%) ↑	37.91	16.47	**55.81**	51.60	53.52	42.02	39.27	26.05	46.97	17.17
	Difficulty Weighted Relative Difference Ratio (%) ↑	37.91	16.47	**55.81**	51.60	53.52	42.02	39.27	26.05	46.97	17.17
	Valid Output Grid ↑	99.00	88.00	**100.00**	**100.00**	99.00	98.00	**100.00**	96.00	**100.00**	83.00
	Load-Support Connectivity (%) ↑	95.00	80.00	**100.00**	94.00	98.00	98.00	69.00	62.00	74.00	34.00
	Load-Support Directional Connectivity (%) ↑	95.00	62.00	**100.00**	82.00	98.00	84.00	69.00	58.00	74.00	28.00
	Average Isolated Clusters Count ↓	0.01	0.05	0.00	0.03	0.01	0.01	0.27	0.08	0.28	**0.30**
	Force Path Cost Average Efficiency Ratio (%) ↑	94.09	83.83	**99.68**	96.39	97.54	96.68	69.22	81.37	73.47	54.16
3 Random Columns	Exact Match ↑	1	3	**5**	2	3	2	2	0	1	1
	Difference Ratio (%) ↑	2.46	56.64	58.69	**60.16**	52.03	37.06	18.34	26.74	24.01	32.05
	Relative Difference Ratio (%) ↑	2.46	56.64	58.69	**60.16**	52.03	37.06	18.34	26.74	24.01	32.05
	Penalized Difference Ratio (%) ↑	-28.47	25.76	27.96	**30.59**	17.06	8.67	-12.54	-4.77	-9.25	15.25
	Average Difficulty Score	1.88	1.88	1.88	**1.96**	1.90	**1.96**	1.88	**1.96**	1.88	**1.96**
	Difficulty Weighted Difference Ratio (%) ↑	1.69	32.76	34.52	**37.29**	31.14	24.32	12.74	17.90	15.59	18.99
	Difficulty Weighted Relative Difference Ratio (%) ↑	1.69	32.76	34.52	**37.29**	31.14	24.32	12.74	17.90	15.59	18.99
	Valid Output Grid ↑	99.00	**100.00**	**100.00**	**100.00**	98.00	90.00	**100.00**	99.00	93.00	60.00
	Load-Support Connectivity (%) ↑	**97.00**	81.00	94.00	84.00	96.00	89.00	93.00	88.00	70.00	36.00
	Load-Support Directional Connectivity (%) ↑	**97.00**	69.00	94.00	68.00	96.00	68.00	93.00	69.00	70.00	31.00
	Average Isolated Clusters Count ↓	**0.32**	0.17	0.00	0.01	0.02	0.05	0.12	0.10	0.24	0.13
	Force Path Cost Average Efficiency Ratio (%) ↑	**94.43**	82.06	88.28	84.62	92.30	79.45	86.51	82.60	54.44	46.29
Full	Exact Match ↑	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
	Difference Ratio (%) ↑	-62.06	-32.96	-35.78	-32.88	**-25.03**	-32.85	-126.02	-84.98	-49.16	-34.84
	Relative Difference Ratio (%) ↑	-62.06	-32.96	-35.78	-32.88	**-25.03**	-32.85	-126.02	-84.98	-49.16	-34.84
	Penalized Difference Ratio (%) ↑	-62.06	-32.96	-35.78	-32.96	**-25.96**	-38.93	-142.73	-104.95	-49.86	-45.60
	Average Difficulty Score	1.93	1.95	1.92	**1.95**	1.91	**1.95**	1.93	**1.95**	1.93	**1.95**
	Difficulty Weighted Difference Ratio (%) ↑	-37.69	-20.14	-21.79	-20.08	**-14.61**	-19.98	-79.62	-54.42	-29.78	-21.20
	Difficulty Weighted Relative Difference Ratio (%) ↑	-37.69	-20.14	-21.79	-20.08	**-14.61**	-19.98	-79.62	-54.42	-29.78	-21.20
	Valid Output Grid ↑	**100.00**	89.00	**100.00**	**100.00**	99.00	81.00	**100.00**	**100.00**	81.00	84.00
	Load-Support Connectivity (%) ↑	94.00	49.00	94.00	94.00	**98.00**	78.00	48.00	78.00	42.00	26.00
	Load-Support Directional Connectivity (%) ↑	94.00	46.00	94.00	72.00	**98.00**	72.00	48.00	76.00	42.00	20.00
	Average Isolated Clusters Count ↓	0.01	0.00	0.00	0.00	0.01	0.01	0.06	0.13	0.09	**0.16**
	Force Path Cost Average Efficiency Ratio (%) ↑	88.87	55.53	90.33	81.69	**96.85**	70.68	48.31	79.00	37.61	46.55
Average	Exact Match ↑	1	1	**13**	10	11	3	6	3	6	0
	Difference Ratio (%) ↑	14.88	25.95	**50.16**	50.13	49.99	37.00	8.43	14.44	30.44	18.38
	Relative Difference Ratio (%) ↑	14.88	25.95	**50.16**	50.13	49.99	37.00	8.43	14.44	30.44	18.38
	Penalized Difference Ratio (%) ↑	3.95	15.26	39.31	**40.15**	37.54	25.45	-6.82	-1.31	18.68	7.53
	Average Difficulty Score	1.92	1.93	1.92	**1.93**	1.92	**1.93**	1.92	**1.93**	1.92	**1.93**
	Difficulty Weighted Difference Ratio (%) ↑	10.54	16.68	31.69	31.83	**32.03**	24.33	6.03	9.55	20.15	11.68
	Difficulty Weighted Relative Difference Ratio (%) ↑	10.54	16.68	31.69	31.83	**32.03**	24.33	6.03	9.55	20.15	11.68
	Valid Output Grid ↑	99.25	93.75	**100.00**	**100.00**	99.00	90.25	**100.00**	97.75	91.50	72.75
	Load-Support Connectivity (%) ↑	96.25	77.00	97.00	93.00	**98.00**	89.25	75.00	80.00	65.50	35.75
	Load-Support Directional Connectivity (%) ↑	96.25	65.00	97.00	77.00	**98.00**	75.50	75.00	68.75	65.50	27.75
	Average Isolated Clusters Count ↓	**0.65**	0.61	0.00	0.01	0.01	0.04	0.23	0.18	0.25	0.22
	Force Path Cost Average Efficiency Ratio (%) ↑	93.92	77.59	94.34	89.28	**96.50**	83.11	73.19	81.43	59.79	48.55

K.3 Results for Claude Opus 4 on All Tasks

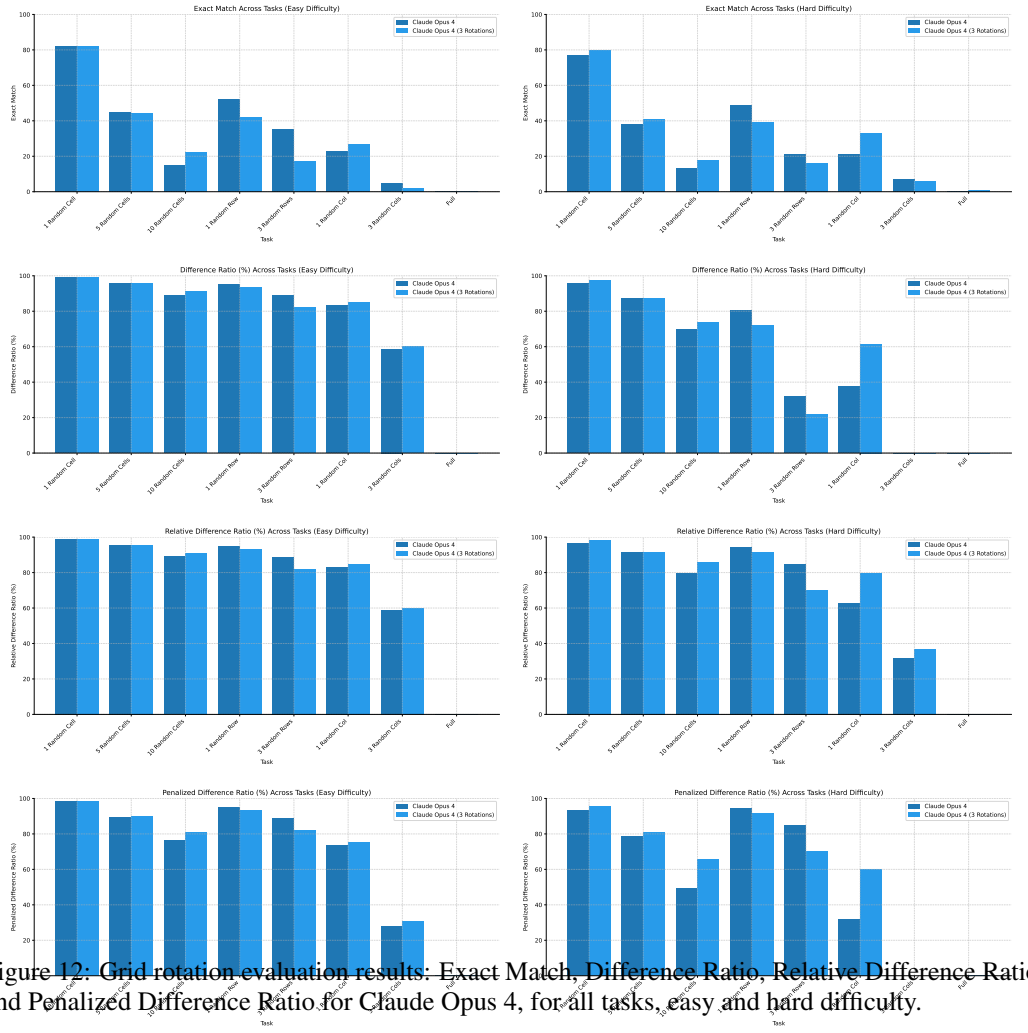

Figure 12: Grid rotation evaluation results: Exact Match, Difference Ratio, Relative Difference Ratio and Penalized Difference Ratio for Claude Opus 4, for all tasks, easy and hard difficulty.

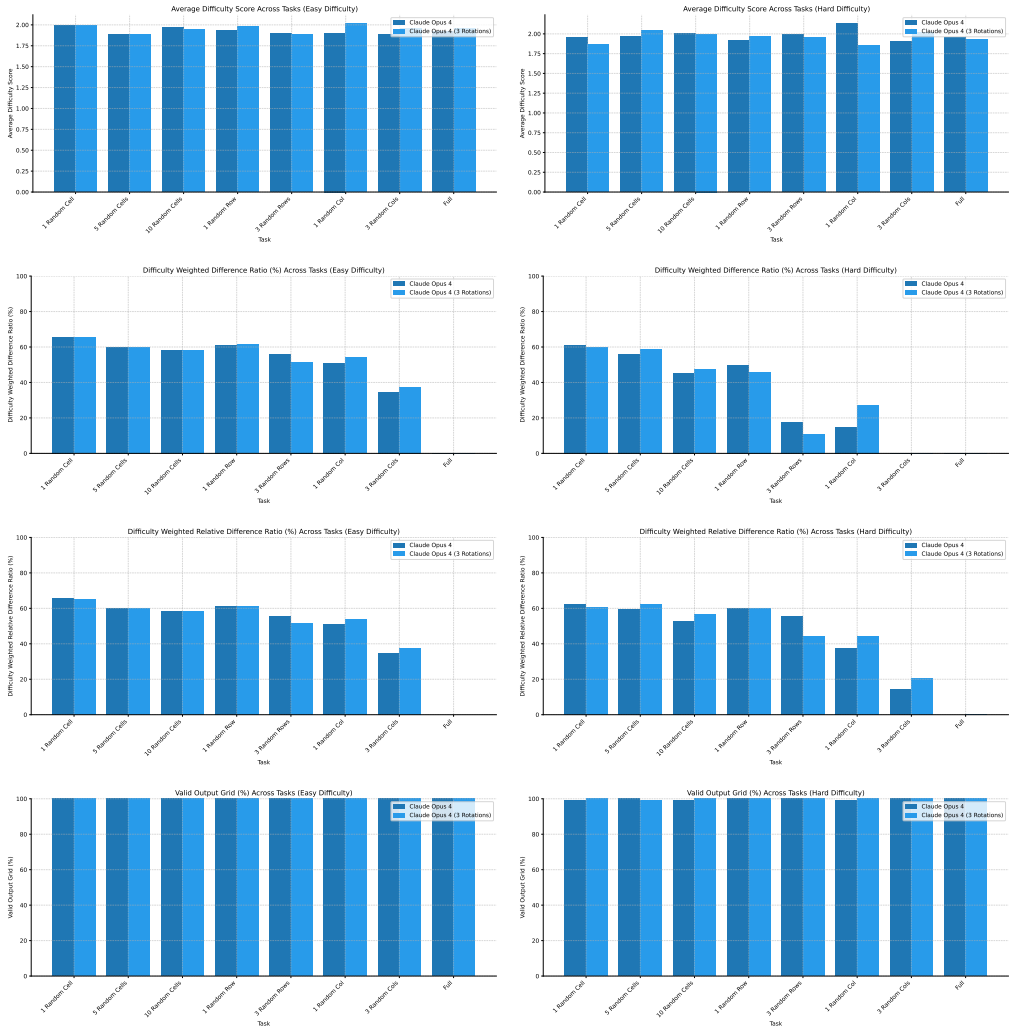

Figure 13: Grid rotation evaluation results: Average Difficulty Score, Difficulty Weighted Difference Ratio, Difficulty Weighted Relative Difference Ratio and Grid Validity for Claude Opus 4, for all tasks, easy and hard difficulty.

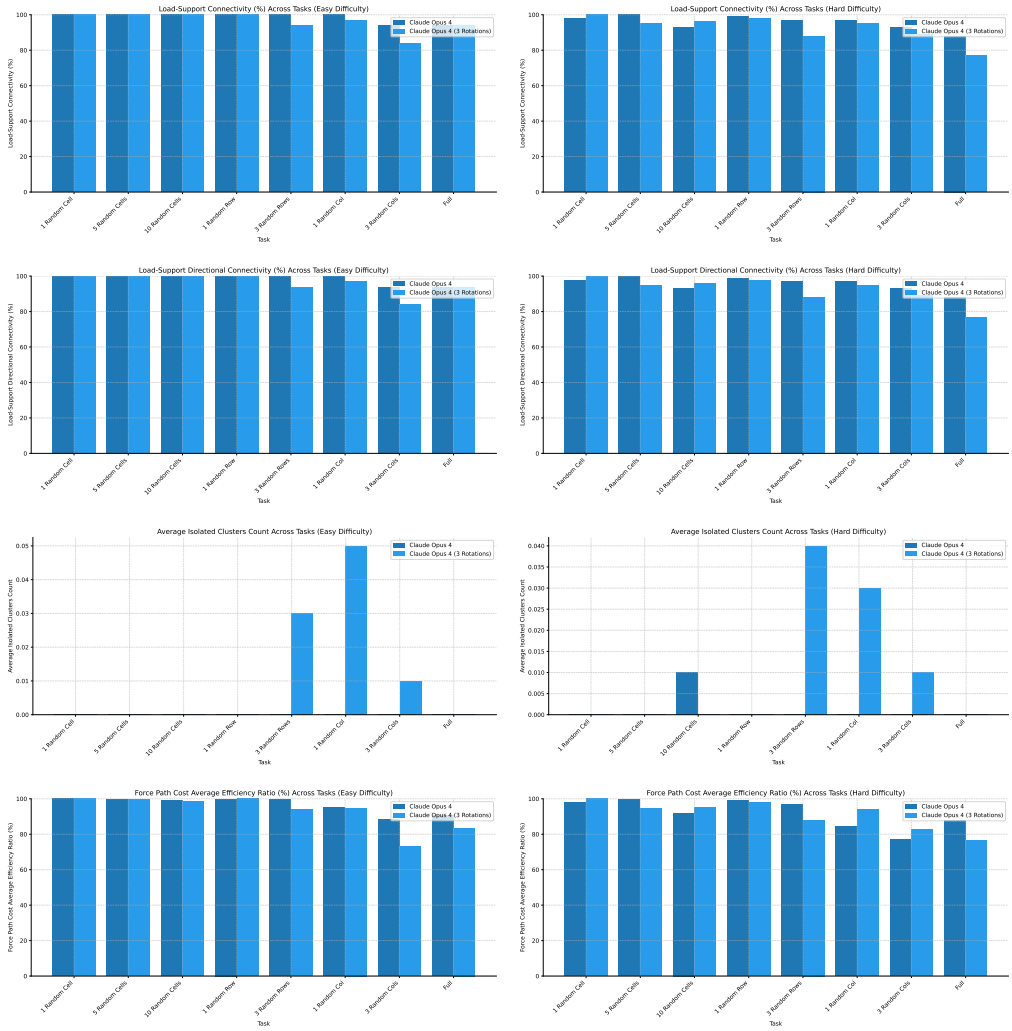

Figure 14: Grid rotation evaluation results: Load-Support Connectivity, Directional Load-Support Connectivity, Average Isolated Cluster Count and Force Path Cost Average Efficiency Ratio for Claude Opus 4, for all tasks, easy and hard difficulty.



Figure 15: Grid rotation evaluation results, simulated structural metrics: Topology Score, Structural Efficiency, Material Efficiency, Compliance Efficiency vs Ground Truth and Load Carrying for Claude Opus 4, for all tasks, easy and hard difficulty.

Table 11: Grid rotation evaluation run results for all metrics for Claude Opus 4, for all tasks, easy and hard difficulty.

		Easy		Hard	
Task	Metric	Claude Opus 4	Claude Opus 4 (3 Rotations)	Claude Opus 4	Claude Opus 4 (3 Rotations)
1 Random Cell	Exact Match ↑	**82**	**82**	77	**80**
	Difference Ratio (%) ↑	**99.05**	99.02	95.45	**97.24**
	Relative Difference Ratio (%) ↑	**99.05**	99.02	96.72	**98.32**
	Penalized Difference Ratio (%) ↑	98.40	**98.47**	93.11	**95.60**
	Average Difficulty Score	**1.99**	**1.99**	**1.96**	1.87
	Difficulty Weighted Difference Ratio (%) ↑	**65.51**	65.40	**60.97**	59.80
	Difficulty Weighted Relative Difference Ratio (%) ↑	**65.51**	65.40	**62.25**	60.88
	Valid Output Grid ↑	**100.00**	**100.00**	99.00	**100.00**
	Load-Support Connectivity (%) ↑	**100.00**	**100.00**	98.00	**100.00**
	Load-Support Directional Connectivity (%) ↑	**100.00**	86.00	**98.00**	84.00
	Average Isolated Clusters Count ↓	**0.00**	**0.00**	**0.00**	**0.00**
	Force Path Cost Average Efficiency Ratio (%) ↑	**99.94**	99.72	97.91	**100.00**
5 Random Cells	Exact Match ↑	**45**	44	38	**41**
	Difference Ratio (%) ↑	**95.76**	95.54	**87.27**	87.23
	Relative Difference Ratio (%) ↑	**95.76**	95.54	91.58	**91.71**
	Penalized Difference Ratio (%) ↑	89.26	**90.06**	78.42	**80.79**
	Average Difficulty Score	**1.89**	**1.89**	1.97	**2.05**
	Difficulty Weighted Difference Ratio (%) ↑	59.89	**59.91**	56.17	**58.59**
	Difficulty Weighted Relative Difference Ratio (%) ↑	59.89	**59.91**	59.65	**62.07**
	Valid Output Grid ↑	**100.00**	**100.00**	**100.00**	99.00
	Load-Support Connectivity (%) ↑	**100.00**	**100.00**	**100.00**	95.00
	Load-Support Directional Connectivity (%) ↑	**100.00**	87.00	**100.00**	75.00
	Average Isolated Clusters Count ↓	**0.00**	**0.00**	**0.00**	**0.00**
	Force Path Cost Average Efficiency Ratio (%) ↑	**99.70**	98.20	**99.49**	95.37
10 Random Cells	Exact Match ↑	15	**22**	13	**18**
	Difference Ratio (%) ↑	89.08	**91.02**	69.70	**73.57**
	Relative Difference Ratio (%) ↑	89.08	**91.02**	79.91	**85.93**
	Penalized Difference Ratio (%) ↑	76.41	**80.78**	49.33	**65.62**
	Average Difficulty Score	**1.97**	1.94	**2.01**	2.00
	Difficulty Weighted Difference Ratio (%) ↑	58.23	**58.52**	45.17	**47.31**
	Difficulty Weighted Relative Difference Ratio (%) ↑	58.23	**58.52**	52.88	**56.59**
	Valid Output Grid ↑	**100.00**	**100.00**	99.00	**100.00**
	Load-Support Connectivity (%) ↑	**100.00**	**100.00**	93.00	**96.00**
	Load-Support Directional Connectivity (%) ↑	**100.00**	86.00	**93.00**	80.00
	Average Isolated Clusters Count ↓	**0.00**	**0.00**	**0.01**	0.00
	Force Path Cost Average Efficiency Ratio (%) ↑	**99.06**	94.41	91.98	**95.00**
1 Random Row	Exact Match ↑	**52**	42	**49**	39
	Difference Ratio (%) ↑	**94.92**	93.30	**80.55**	71.92
	Relative Difference Ratio (%) ↑	**94.92**	93.30	**94.39**	91.71
	Penalized Difference Ratio (%) ↑	**94.92**	93.30	**94.39**	91.71
	Average Difficulty Score	1.94	**1.98**	1.92	**1.97**
	Difficulty Weighted Difference Ratio (%) ↑	61.04	**61.36**	**49.95**	45.77
	Difficulty Weighted Relative Difference Ratio (%) ↑	61.04	**61.36**	60.02	**60.12**
	Valid Output Grid ↑	**100.00**	**100.00**	**100.00**	**100.00**
	Load-Support Connectivity (%) ↑	**100.00**	**100.00**	**99.00**	98.00
	Load-Support Directional Connectivity (%) ↑	**100.00**	82.00	**99.00**	82.00
	Average Isolated Clusters Count ↓	**0.00**	**0.00**	**0.00**	**0.00**
	Force Path Cost Average Efficiency Ratio (%) ↑	99.88	**100.00**	**98.92**	98.85
3 Random Rows	Exact Match ↑	**35**	17	**21**	16
	Difference Ratio (%) ↑	**88.64**	82.19	**32.01**	21.98
	Relative Difference Ratio (%) ↑	**88.64**	82.19	**84.70**	70.12
	Penalized Difference Ratio (%) ↑	**88.64**	82.19	**84.70**	70.12
	Average Difficulty Score	**1.89**	1.89	**1.99**	1.96
	Difficulty Weighted Difference Ratio (%) ↑	**55.81**	51.60	**17.50**	11.08
	Difficulty Weighted Relative Difference Ratio (%) ↑	**55.81**	51.60	**55.67**	44.35
	Valid Output Grid ↑	**100.00**	**100.00**	**100.00**	**100.00**
	Load-Support Connectivity (%) ↑	**100.00**	94.00	**97.00**	88.00
	Load-Support Directional Connectivity (%) ↑	**100.00**	82.00	**97.00**	78.00
	Average Isolated Clusters Count ↓	0.00	**0.03**	0.00	**0.04**
	Force Path Cost Average Efficiency Ratio (%) ↑	**99.68**	96.39	**96.97**	95.42
1 Random Column	Exact Match ↑	23	**27**	21	**33**
	Difference Ratio (%) ↑	83.09	**84.86**	37.46	**61.54**
	Relative Difference Ratio (%) ↑	83.09	**84.86**	62.68	**79.72**
	Penalized Difference Ratio (%) ↑	73.51	**75.48**	31.65	**60.26**
	Average Difficulty Score	1.90	**2.02**	**2.13**	1.85
	Difficulty Weighted Difference Ratio (%) ↑	50.85	**54.04**	14.68	**27.37**
	Difficulty Weighted Relative Difference Ratio (%) ↑	50.85	**54.04**	37.62	**44.46**
	Valid Output Grid ↑	**100.00**	**100.00**	99.00	**100.00**
	Load-Support Connectivity (%) ↑	**100.00**	97.00	**97.00**	95.00
	Load-Support Directional Connectivity (%) ↑	**100.00**	86.00	**97.00**	77.00
	Average Isolated Clusters Count ↓	0.00	**0.05**	0.00	**0.03**
	Force Path Cost Average Efficiency Ratio (%) ↑	**94.90**	92.78	84.17	**93.51**
3 Random Columns	Exact Match ↑	**5**	2	**7**	6
	Difference Ratio (%) ↑	58.69	**60.16**	**-17.63**	-18.77
	Relative Difference Ratio (%) ↑	58.69	**60.16**	31.78	**36.93**
	Penalized Difference Ratio (%) ↑	27.96	**30.59**	-43.98	**-38.28**
	Average Difficulty Score	1.88	**1.96**	1.90	**2.01**
	Difficulty Weighted Difference Ratio (%) ↑	34.52	**37.29**	-22.99	**-21.91**
	Difficulty Weighted Relative Difference Ratio (%) ↑	34.52	**37.29**	14.44	**20.53**
	Valid Output Grid ↑	**100.00**	**100.00**	**100.00**	**100.00**
	Load-Support Connectivity (%) ↑	**94.00**	84.00	**93.00**	89.00
	Load-Support Directional Connectivity (%) ↑	**94.00**	68.00	**93.00**	73.00
	Average Isolated Clusters Count ↓	0.00	**0.01**	0.00	**0.01**
	Force Path Cost Average Efficiency Ratio (%) ↑	**88.28**	84.62	77.05	**82.98**
Full	Exact Match ↑	**0**	**0**	0	**1**
	Difference Ratio (%) ↑	-35.78	**-32.88**	-466.42	**-403.39**
	Relative Difference Ratio (%) ↑	-35.78	**-32.88**	-177.48	**-136.14**
	Penalized Difference Ratio (%) ↑	-35.78	**-32.96**	-177.48	**-138.43**
	Average Difficulty Score	1.92	**1.95**	**1.96**	1.93
	Difficulty Weighted Difference Ratio (%) ↑	-21.79	**-20.08**	-310.26	**-266.48**
	Difficulty Weighted Relative Difference Ratio (%) ↑	-21.79	**-20.08**	-117.17	**-90.28**
	Valid Output Grid ↑	**100.00**	**100.00**	**100.00**	**100.00**
	Load-Support Connectivity (%) ↑	**94.00**	**94.00**	**88.00**	77.00
	Load-Support Directional Connectivity (%) ↑	**94.00**	72.00	**88.00**	74.00
	Average Isolated Clusters Count ↓	**0.00**	**0.00**	**0.00**	**0.00**
	Force Path Cost Average Efficiency Ratio (%) ↑	**90.33**	81.69	**87.83**	86.43
Average	Exact Match ↑	**32.12**	29.50	28.25	**29.25**
	Difference Ratio (%) ↑	**71.68**	71.65	-10.20	**-1.08**
	Relative Difference Ratio (%) ↑	**71.68**	71.65	45.53	**52.29**
	Penalized Difference Ratio (%) ↑	64.16	**64.74**	26.27	**35.92**
	Average Difficulty Score	1.92	**1.95**	**1.98**	1.95
	Difficulty Weighted Difference Ratio (%) ↑	45.51	**46.01**	-11.10	**-4.81**
	Difficulty Weighted Relative Difference Ratio (%) ↑	45.51	**46.01**	28.17	**32.34**
	Valid Output Grid ↑	**100.00**	**100.00**	99.62	**99.88**
	Load-Support Connectivity (%) ↑	**98.50**	96.12	**95.62**	92.25
	Load-Support Directional Connectivity (%) ↑	**98.50**	81.12	**95.62**	77.88
	Average Isolated Clusters Count ↓	0.00	**0.01**	0.00	**0.01**
	Force Path Cost Average Efficiency Ratio (%) ↑	**96.47**	93.48	91.79	**93.44**

L Additional Few-Shot Experiment Results

L.1 Prompt and Completion - 1 Random Cell Easy: 1-Shot

Prompt and Completion - 1 Random Cell Easy: 1-Shot

Prompt:

```
You are given a structural material distribution represented as a grid.  Each cell can
have one of the following states:
- 'L' indicates applied load.
- 'V' indicates void.
- 'S' indicates support.
The goal is to predict the correct material distribution by filling in all 'V' cells
with either '1' (solid) or '0' (empty), based on the surrounding structure and implicit
physical reasoning (such as load paths, supports, and forces).
Example input grid with masked regions:
0 0 0 L L L L 0 0 0
0 0 0 0 1 1 0 0 0 0
0 0 0 0 1 1 0 V 0 0
0 0 0 0 1 1 0 0 0 0
0 0 0 0 1 1 0 0 0 0
0 0 0 0 1 1 0 0 0 0
0 0 0 0 1 1 0 0 0 0
0 0 0 0 1 1 0 0 0 0
0 0 0 0 1 1 0 0 0 0
0 0 0 0 S S S S S 0
Corresponding completed output grid:
0 0 0 L L L L 0 0 0
0 0 0 0 1 1 0 0 0 0
0 0 0 0 1 1 0 0 0 0
0 0 0 0 1 1 0 0 0 0
0 0 0 0 1 1 0 0 0 0
0 0 0 0 1 1 0 0 0 0
0 0 0 0 1 1 0 0 0 0
0 0 0 0 1 1 0 0 0 0
0 0 0 0 1 1 0 0 0 0
0 0 0 0 S S S S S 0
Important:  The completed structure should use as little material as possible while
remaining stable and plausible for carrying the applied forces.  Minimize material usage
unless necessary for structural support.
Below is the input grid with masked regions:
0 0 L L L 0 0 0 0 0
0 0 0 1 0 0 0 0 0 0
0 0 0 V 0 0 0 0 0 0
0 0 0 1 0 0 0 0 0 0
0 0 0 1 0 0 0 0 0 0
0 0 0 1 0 0 0 0 0 0
0 0 1 1 1 0 0 0 0 0
0 0 1 1 1 0 0 0 0 0
0 0 1 1 1 0 0 0 0 0
S S S S S S S S S 0
Please output the completed grid by replacing all 'V' cells with either '1' (solid) or '0'
(empty).
Maintain the same format as the input:  one row per line, cells separated by spaces, and
the total number of rows and columns unchanged.
Return only the completed grid without any additional explanation.
```

Model Completion (differences in red):

```
0 0 L L L 0 0 0 0 0
0 0 0 1 0 0 0 0 0 0
0 0 0 1 0 0 0 0 0 0
0 0 0 1 0 0 0 0 0 0
0 0 0 1 0 0 0 0 0 0
0 0 0 1 0 0 0 0 0 0
0 0 1 1 1 0 0 0 0 0
0 0 1 1 1 0 0 0 0 0
0 0 1 1 1 0 0 0 0 0
S S S S S S S S S 0
```

Ground Truth:

```
0 0 L L L 0 0 0 0 0
0 0 0 1 0 0 0 0 0 0
0 0 0 1 0 0 0 0 0 0
0 0 0 1 0 0 0 0 0 0
0 0 0 1 0 0 0 0 0 0
0 0 0 1 0 0 0 0 0 0
0 0 1 1 1 0 0 0 0 0
0 0 1 1 1 0 0 0 0 0
0 0 1 1 1 0 0 0 0 0
S S S S S S S S S 0
```

L.2 Prompt and Completion - 1 Random Cell Easy: 3-Shot

Prompt and Completion - 1 Random Cell Easy: 3-Shot

Prompt:

```
You are given a structural material distribution represented as a grid.  Each cell can
have one of the following states:
- 'L' indicates applied load.
- 'V' indicates void.
- 'S' indicates support.
The goal is to predict the correct material distribution by filling in all 'V' cells
with either '1' (solid) or '0' (empty), based on the surrounding structure and implicit
physical reasoning (such as load paths, supports, and forces).
Example input grid with masked regions:
0 0 0 L L L L 0 0 0
0 0 0 0 1 1 0 0 0 0
0 0 0 0 1 1 0 V 0 0
0 0 0 0 1 1 0 0 0 0
0 0 0 0 1 1 0 0 0 0
0 0 0 0 1 1 0 0 0 0
0 0 0 0 1 1 0 0 0 0
0 0 0 0 1 1 0 0 0 0
0 0 0 0 1 1 0 0 0 0
0 0 0 0 S S S S S 0
Corresponding completed output grid:
0 0 0 L L L L 0 0 0
0 0 0 0 1 1 0 0 0 0
0 0 0 0 1 1 0 0 0 0
0 0 0 0 1 1 0 0 0 0
0 0 0 0 1 1 0 0 0 0
0 0 0 0 1 1 0 0 0 0
0 0 0 0 1 1 0 0 0 0
0 0 0 0 1 1 0 0 0 0
0 0 0 0 1 1 0 0 0 0
0 0 0 0 S S S S S 0
Example input grid with masked regions:
0 0 0 L L L 0 0 0 0
0 0 0 0 1 0 0 0 0 0 0
0 0 0 0 1 0 0 0 0 0 0
0 0 0 0 1 0 0 0 0 0 0
0 0 0 0 1 0 0 0 0 0 0
V 0 0 0 1 0 0 0 0 0 0
0 0 0 0 1 0 0 0 0 0 0
0 0 0 0 1 1 1 1 0 0
0 0 0 0 0 1 1 1 0 0
0 0 0 0 0 1 S S S 0
Corresponding completed output grid:
0 0 0 L L L 0 0 0 0
0 0 0 0 1 0 0 0 0 0 0
0 0 0 0 1 0 0 0 0 0 0
0 0 0 0 1 0 0 0 0 0 0
0 0 0 0 1 0 0 0 0 0 0
0 0 0 0 1 0 0 0 0 0 0
0 0 0 0 1 0 0 0 0 0 0
0 0 0 0 1 1 1 1 0 0
0 0 0 0 0 1 1 1 0 0
0 0 0 0 0 1 S S S 0
Example input grid with masked regions:
0 0 L L L L L 0 0 0
0 0 0 1 1 1 0 0 0 0
0 0 0 1 1 1 V 0 0 0
0 0 0 0 1 0 0 0 0 0
0 0 0 0 1 0 0 0 0 0
0 0 0 0 1 0 0 0 0 0
0 0 0 1 1 1 0 0 0 0
0 0 0 1 1 1 0 0 0 0
0 0 0 1 1 1 0 0 0 0
0 S S S S S S S 0 0
Corresponding completed output grid:
0 0 L L L L L 0 0 0
0 0 0 1 1 1 0 0 0 0
0 0 0 1 1 1 0 0 0 0
0 0 0 0 1 0 0 0 0 0
0 0 0 0 1 0 0 0 0 0
0 0 0 0 1 0 0 0 0 0
0 0 0 1 1 1 0 0 0 0
0 0 0 1 1 1 0 0 0 0
0 0 0 1 1 1 0 0 0 0
0 S S S S S S S 0 0
```

```
Important:  The completed structure should use as little material as possible while
remaining stable and plausible for carrying the applied forces.  Minimize material usage
unless necessary for structural support.
Below is the input grid with masked regions:
0 0 L L L 0 0 0 0 0
0 0 0 1 0 0 0 0 0 0
0 0 0 V 0 0 0 0 0 0
0 0 0 1 0 0 0 0 0 0
0 0 0 1 0 0 0 0 0 0
0 0 0 1 0 0 0 0 0 0
0 0 1 1 1 0 0 0 0 0
0 0 1 1 1 0 0 0 0 0
0 0 1 1 1 0 0 0 0 0
S S S S S S S S S 0
Please output the completed grid by replacing all 'V' cells with either '1' (solid) or '0'
(empty).
Maintain the same format as the input:  one row per line, cells separated by spaces, and
the total number of rows and columns unchanged.
Return only the completed grid without any additional explanation.
Model Completion (differences in red):
0 0 L L L 0 0 0 0 0
0 0 0 1 0 0 0 0 0 0
0 0 0 1 0 0 0 0 0 0
0 0 0 1 0 0 0 0 0 0
0 0 0 1 0 0 0 0 0 0
0 0 0 1 0 0 0 0 0 0
0 0 1 1 1 0 0 0 0 0
0 0 1 1 1 0 0 0 0 0
0 0 1 1 1 0 0 0 0 0
S S S S S S S S S 0
Ground Truth:
0 0 L L L 0 0 0 0 0
0 0 0 1 0 0 0 0 0 0
0 0 0 1 0 0 0 0 0 0
0 0 0 1 0 0 0 0 0 0
0 0 0 1 0 0 0 0 0 0
0 0 0 1 0 0 0 0 0 0
0 0 1 1 1 0 0 0 0 0
0 0 1 1 1 0 0 0 0 0
0 0 1 1 1 0 0 0 0 0
S S S S S S S S S 0
```

L.3 Results for Claude Opus 4 on All Tasks

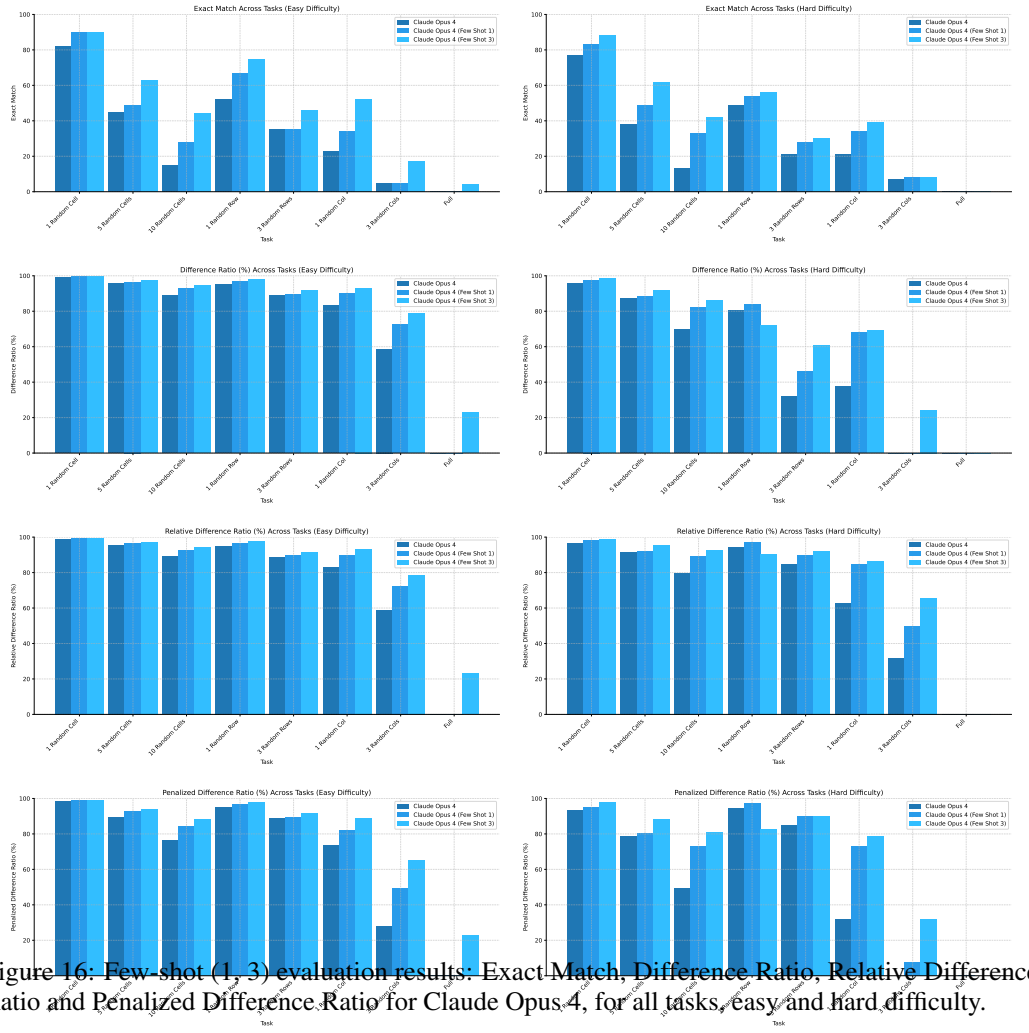

Figure 16: Few-shot (1, 3) evaluation results: Exact Match, Difference Ratio, Relative Difference Ratio and Penalized Difference Ratio for Claude Opus 4, for all tasks, easy and hard difficulty.

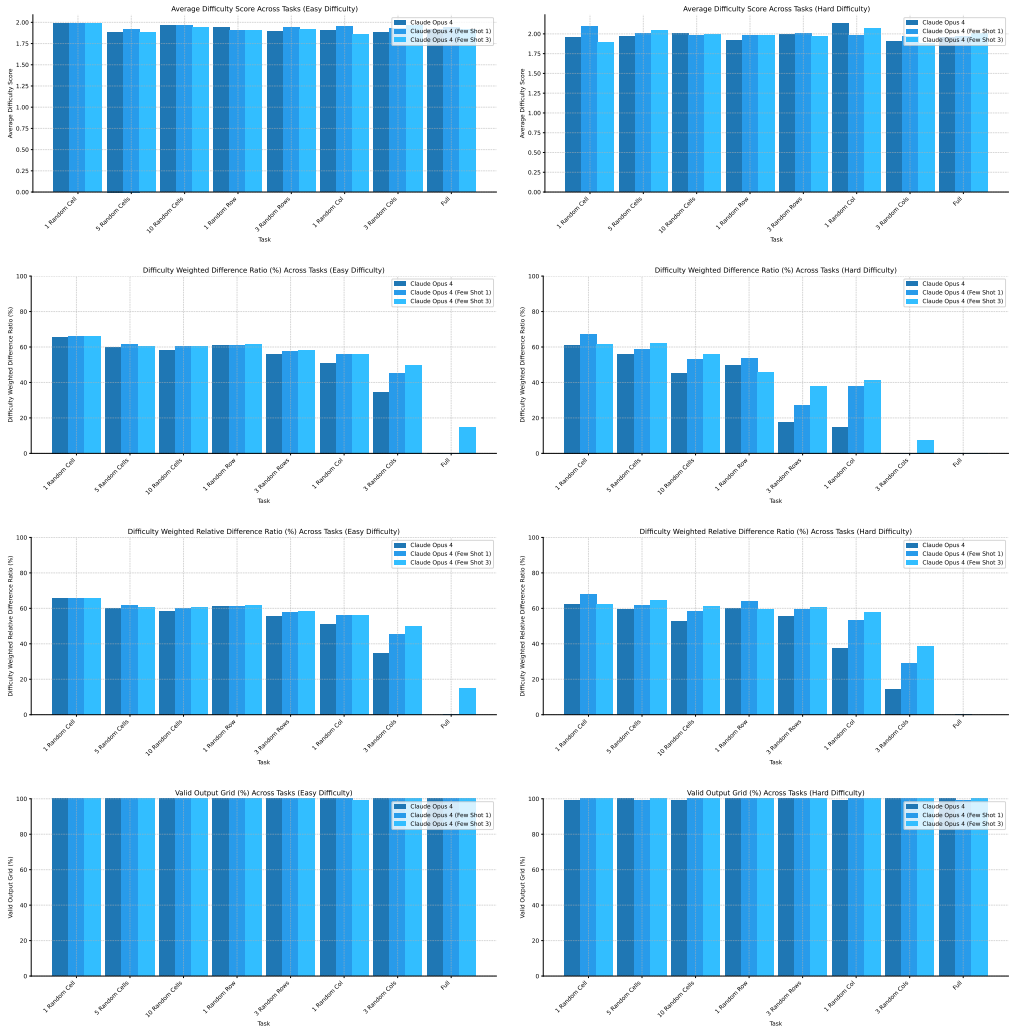

Figure 17: Few-shot (1, 3) evaluation results: Average Difficulty Score, Difficulty Weighted Difference Ratio, Difficulty Weighted Relative Difference Ratio and Grid Validity for Claude Opus 4, for all tasks, easy and hard difficulty.

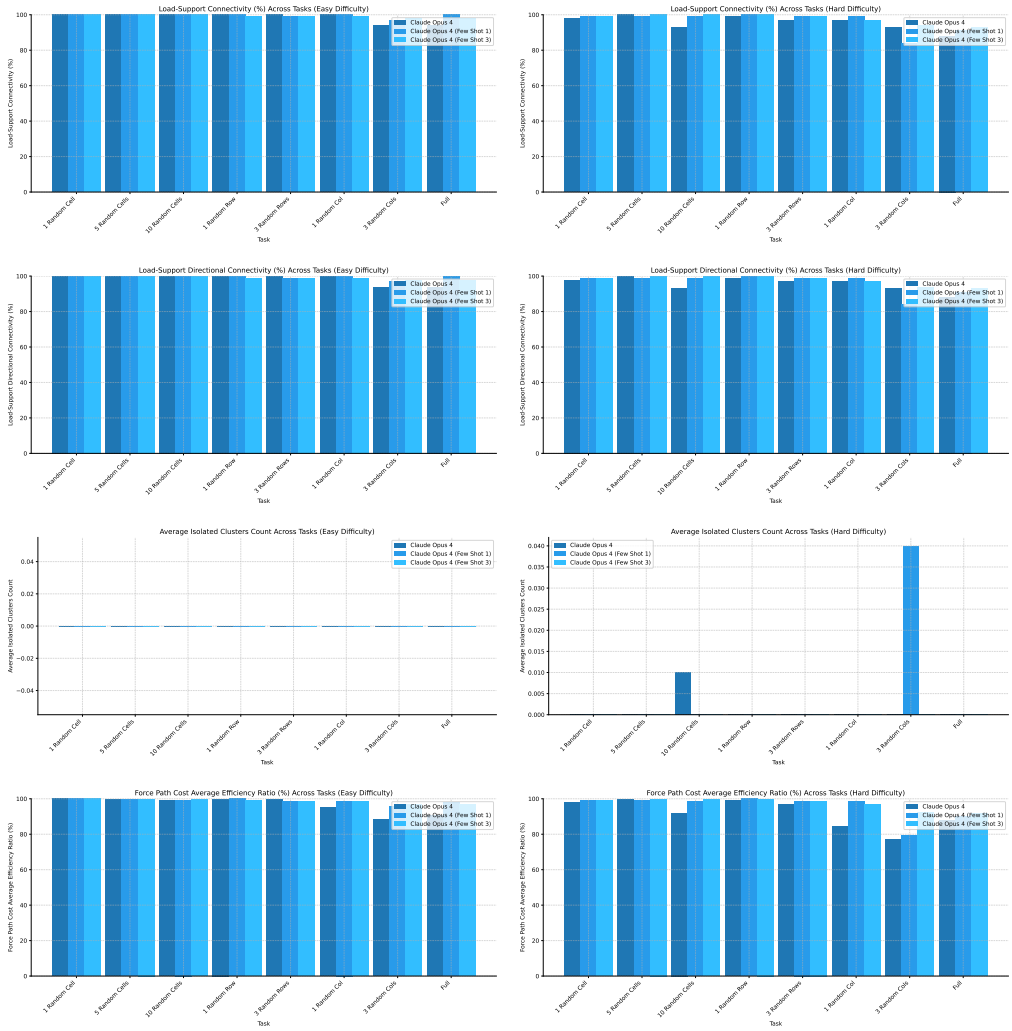

Figure 18: Few-shot (1, 3) evaluation results: Load-Support Connectivity, Directional Load-Support Connectivity, Average Isolated Cluster Count and Force Path Cost Average Efficiency Ratio for Claude Opus 4, for all tasks, easy and hard difficulty.



Figure 19: Few-shot (1, 3) evaluation results, simulated structural metrics: Topology Score, Structural Efficiency, Material Efficiency, Compliance Efficiency vs Ground Truth and Load Carrying for Claude Opus 4, for all tasks, easy and hard difficulty.

Table 12: Few-shot (1, 3) evaluation results for all metrics, for Claude Opus 4, for all tasks, easy and hard difficulty.

		Easy			Hard		
Task	Metric	Claude Opus 4 (Zero-Shot)	Claude Opus 4 (1-Shot)	Claude Opus 4 (3-Shot)	Claude Opus 4 (Zero-Shot)	Claude Opus 4 (1-Shot)	Claude Opus 4 (3-Shot)
1 Random Cell	Exact Match ↑	82	**90**	**90**	77	83	**88**
	Difference Ratio (%) ↑	99.05	**99.51**	99.50	95.45	97.64	**98.33**
	Relative Difference Ratio (%) ↑	99.05	**99.51**	99.50	96.72	98.28	**99.13**
	Penalized Difference Ratio (%) ↑	98.40	98.93	**99.03**	93.11	95.20	**97.72**
	Average Difficulty Score	**1.99**	**1.99**	**1.99**	1.96	**2.09**	1.89
	Difficulty Weighted Difference Ratio (%) ↑	65.51	**65.92**	65.88	60.97	**67.45**	61.43
	Difficulty Weighted Relative Difference Ratio (%) ↑	65.51	**65.92**	65.88	62.25	**68.09**	62.22
	Valid Output Grid ↑	**100.00**	**100.00**	**100.00**	99.00	**100.00**	**100.00**
	Load-Support Connectivity (%) ↑	**100.00**	**100.00**	**100.00**	98.00	**99.00**	**99.00**
	Load-Support Directional Connectivity (%) ↑	**100.00**	**100.00**	**100.00**	98.00	**99.00**	**99.00**
	Average Isolated Clusters Count ↓	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
	Force Path Cost Average Efficiency Ratio (%) ↑	99.94	99.94	**99.95**	97.91	98.96	**99.00**
5 Random Cells	Exact Match ↑	45	49	**63**	38	49	**62**
	Difference Ratio (%) ↑	95.76	96.53	**97.25**	87.27	88.28	**91.85**
	Relative Difference Ratio (%) ↑	95.76	96.53	**97.25**	91.58	92.10	**95.51**
	Penalized Difference Ratio (%) ↑	89.26	92.90	**93.93**	78.42	80.15	**88.27**
	Average Difficulty Score	1.89	**1.92**	1.88	1.97	2.01	**2.04**
	Difficulty Weighted Difference Ratio (%) ↑	59.89	**61.73**	60.69	56.17	58.62	**61.94**
	Difficulty Weighted Relative Difference Ratio (%) ↑	59.89	**61.73**	60.69	59.65	61.59	**64.85**
	Valid Output Grid ↑	**100.00**	**100.00**	**100.00**	**100.00**	99.00	**100.00**
	Load-Support Connectivity (%) ↑	**100.00**	**100.00**	**100.00**	**100.00**	99.00	**100.00**
	Load-Support Directional Connectivity (%) ↑	**100.00**	**100.00**	**100.00**	**100.00**	99.00	**100.00**
	Average Isolated Clusters Count ↓	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
	Force Path Cost Average Efficiency Ratio (%) ↑	99.70	99.66	**99.87**	99.49	98.89	**99.90**
10 Random Cells	Exact Match ↑	15	28	**44**	13	33	**42**
	Difference Ratio (%) ↑	89.08	92.63	**94.31**	69.70	82.11	**85.88**
	Relative Difference Ratio (%) ↑	89.08	92.63	**94.31**	79.91	89.57	**92.63**
	Penalized Difference Ratio (%) ↑	76.41	84.37	**88.11**	49.33	73.32	**81.11**
	Average Difficulty Score	**1.97**	1.96	1.94	**2.01**	1.98	1.99
	Difficulty Weighted Difference Ratio (%) ↑	58.23	60.24	**60.77**	45.17	52.92	**56.05**
	Difficulty Weighted Relative Difference Ratio (%) ↑	58.23	60.24	**60.77**	52.88	58.43	**61.07**
	Valid Output Grid ↑	**100.00**	**100.00**	**100.00**	99.00	**100.00**	**100.00**
	Load-Support Connectivity (%) ↑	**100.00**	**100.00**	**100.00**	93.00	99.00	**100.00**
	Load-Support Directional Connectivity (%) ↑	**100.00**	**100.00**	**100.00**	93.00	99.00	**100.00**
	Average Isolated Clusters Count ↓	**0.00**	**0.00**	**0.00**	**0.01**	0.00	0.00
	Force Path Cost Average Efficiency Ratio (%) ↑	99.06	99.32	**99.37**	91.98	98.41	**99.55**
1 Random Row	Exact Match ↑	52	67	**75**	49	54	**56**
	Difference Ratio (%) ↑	94.92	96.86	**97.72**	80.55	**83.72**	72.00
	Relative Difference Ratio (%) ↑	94.92	96.86	**97.72**	94.39	**97.25**	90.29
	Penalized Difference Ratio (%) ↑	94.92	96.86	**97.72**	94.39	**97.25**	82.61
	Average Difficulty Score	**1.94**	1.91	1.90	1.92	1.98	**1.98**
	Difficulty Weighted Difference Ratio (%) ↑	61.04	61.29	**61.66**	49.95	**53.53**	45.64
	Difficulty Weighted Relative Difference Ratio (%) ↑	61.04	61.29	**61.66**	60.02	**63.78**	59.35
	Valid Output Grid ↑	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**
	Load-Support Connectivity (%) ↑	**100.00**	**100.00**	99.00	99.00	**100.00**	**100.00**
	Load-Support Directional Connectivity (%) ↑	**100.00**	**100.00**	99.00	99.00	**100.00**	**100.00**
	Average Isolated Clusters Count ↓	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
	Force Path Cost Average Efficiency Ratio (%) ↑	99.88	**99.96**	98.85	98.92	**99.99**	99.84
3 Random Rows	Exact Match ↑	35	35	**46**	21	28	**30**
	Difference Ratio (%) ↑	88.64	89.67	**91.83**	32.01	45.88	**60.79**
	Relative Difference Ratio (%) ↑	88.64	89.67	**91.83**	84.70	89.71	**92.24**
	Penalized Difference Ratio (%) ↑	88.64	89.67	**91.83**	84.70	89.71	**89.97**
	Average Difficulty Score	1.89	**1.94**	1.92	1.99	**2.00**	1.97
	Difficulty Weighted Difference Ratio (%) ↑	55.81	57.69	**58.32**	17.50	27.29	**38.05**
	Difficulty Weighted Relative Difference Ratio (%) ↑	55.81	57.69	**58.32**	55.67	59.27	**60.37**
	Valid Output Grid ↑	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**
	Load-Support Connectivity (%) ↑	**100.00**	99.00	99.00	97.00	**99.00**	**99.00**
	Load-Support Directional Connectivity (%) ↑	**100.00**	99.00	99.00	97.00	**99.00**	**99.00**
	Average Isolated Clusters Count ↓	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
	Force Path Cost Average Efficiency Ratio (%) ↑	**99.68**	98.49	98.62	96.97	**98.75**	98.73
1 Random Column	Exact Match ↑	23	34	**52**	21	34	**39**
	Difference Ratio (%) ↑	83.09	89.98	**93.14**	37.46	67.96	**69.20**
	Relative Difference Ratio (%) ↑	83.09	89.98	**93.14**	62.68	84.74	**86.73**
	Penalized Difference Ratio (%) ↑	73.51	82.16	**88.94**	31.65	73.11	**78.56**
	Average Difficulty Score	1.90	**1.95**	1.86	**2.13**	1.98	2.07
	Difficulty Weighted Difference Ratio (%) ↑	50.85	**56.22**	56.20	14.68	37.66	**41.22**
	Difficulty Weighted Relative Difference Ratio (%) ↑	50.85	**56.22**	56.20	37.62	53.13	**57.79**
	Valid Output Grid ↑	**100.00**	**100.00**	99.00	99.00	**100.00**	**100.00**
	Load-Support Connectivity (%) ↑	**100.00**	**100.00**	99.00	97.00	**99.00**	97.00
	Load-Support Directional Connectivity (%) ↑	**100.00**	**100.00**	99.00	97.00	**99.00**	97.00
	Average Isolated Clusters Count ↓	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
	Force Path Cost Average Efficiency Ratio (%) ↑	94.90	**98.74**	98.65	84.17	**98.62**	96.66
3 Random Columns	Exact Match ↑	5	5	**17**	7	**8**	**8**
	Difference Ratio (%) ↑	58.69	72.50	**78.86**	-17.63	-0.45	**24.33**
	Relative Difference Ratio (%) ↑	58.69	72.50	**78.86**	31.78	49.90	**65.66**
	Penalized Difference Ratio (%) ↑	27.96	49.17	**65.13**	-43.98	7.40	**31.72**
	Average Difficulty Score	1.88	1.93	**1.96**	1.90	**1.96**	1.88
	Difficulty Weighted Difference Ratio (%) ↑	34.52	45.21	**50.03**	-22.99	-10.19	**7.45**
	Difficulty Weighted Relative Difference Ratio (%) ↑	34.52	45.21	**50.03**	14.44	28.98	**38.82**
	Valid Output Grid ↑	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**
	Load-Support Connectivity (%) ↑	94.00	97.00	**98.00**	93.00	84.00	**94.00**
	Load-Support Directional Connectivity (%) ↑	94.00	97.00	**98.00**	93.00	84.00	**94.00**
	Average Isolated Clusters Count ↓	**0.00**	**0.00**	**0.00**	0.00	**0.04**	0.00
	Force Path Cost Average Efficiency Ratio (%) ↑	88.28	95.82	**96.94**	77.05	79.48	**92.20**
Full	Exact Match ↑	0	0	**4**	**0**	**0**	**0**
	Difference Ratio (%) ↑	-35.78	-14.18	**23.06**	-466.42	-292.37	**-240.02**
	Relative Difference Ratio (%) ↑	-35.78	-14.18	**23.06**	-177.48	-27.72	**-4.41**
	Penalized Difference Ratio (%) ↑	-35.78	-14.18	**23.06**	-177.48	-27.72	**-4.41**
	Average Difficulty Score	1.92	**1.94**	1.91	1.96	1.96	**2.00**
	Difficulty Weighted Difference Ratio (%) ↑	-21.79	-8.72	**14.96**	-310.26	-195.92	**-169.54**
	Difficulty Weighted Relative Difference Ratio (%) ↑	-21.79	-8.72	**14.96**	-117.17	-19.24	**-5.04**
	Valid Output Grid ↑	**100.00**	**100.00**	**100.00**	**100.00**	99.00	**100.00**
	Load-Support Connectivity (%) ↑	94.00	**100.00**	98.00	88.00	91.00	**93.00**
	Load-Support Directional Connectivity (%) ↑	94.00	**100.00**	98.00	88.00	91.00	**93.00**
	Average Isolated Clusters Count ↓	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
	Force Path Cost Average Efficiency Ratio (%) ↑	90.33	**97.86**	96.77	87.83	90.10	**91.89**
Average	Exact Match ↑	32.12	38.50	**48.88**	28.25	36.12	**40.62**
	Difference Ratio (%) ↑	71.68	77.94	**84.46**	-10.20	21.60	**32.79**
	Relative Difference Ratio (%) ↑	71.68	77.94	**84.46**	45.53	71.73	**77.22**
	Penalized Difference Ratio (%) ↑	64.16	72.48	**80.97**	26.27	61.05	**68.19**
	Average Difficulty Score	1.92	**1.94**	1.92	1.98	**1.99**	1.98
	Difficulty Weighted Difference Ratio (%) ↑	45.51	49.95	**53.56**	-11.10	11.42	**17.78**
	Difficulty Weighted Relative Difference Ratio (%) ↑	45.51	49.95	**53.56**	28.17	46.75	**49.93**
	Valid Output Grid ↑	**100.00**	**100.00**	99.88	99.62	99.75	**100.00**
	Load-Support Connectivity (%) ↑	98.50	**99.50**	99.12	95.62	96.25	**97.75**
	Load-Support Directional Connectivity (%) ↑	98.50	**99.50**	99.12	95.62	96.25	**97.75**
	Average Isolated Clusters Count ↓	**0.00**	**0.00**	**0.00**	0.00	**0.01**	0.00
	Force Path Cost Average Efficiency Ratio (%) ↑	96.47	**98.72**	98.63	91.79	95.40	**97.22**

M Additional Physics-Enhanced and -Neutral Experiment Results

M.1 Prompt and Completion - 1 Random Cell Easy: Physics-Netural

Prompt and Completion - 1 Random Cell Easy: Physics-Netural

Prompt:

```
You are given a grid of cells.  Each cell can have one of the following states:
- 'L' indicates a special marker.
- 'V' indicates an empty cell.
- 'S' indicates a fixed marker.

The goal is to predict the correct grid configuration by filling in all 'V' cells with
either '1' (solid) or '0' (empty), based on the surrounding structure and general
consistency rules.

Relevant rules for reasoning:
- Filled cells should form continuous connections between relevant markers when necessary.
- Avoid placing isolated filled cells that are not connected to anything meaningful.
- Minimize the number of filled cells while maintaining a coherent and connected
structure.

Below is the input grid with masked regions:

L L L L L 0 0 0 0 0
0 1 1 1 0 0 0 0 0 0
0 1 1 1 0 0 0 0 0 0
0 0 1 0 0 0 0 0 0 0
0 0 1 0 0 0 0 0 0 0
0 0 1 0 0 0 0 0 0 0
0 0 1 0 0 0 0 0 0 0
0 0 1 0 0 0 0 0 0 0
V 0 1 1 1 0 0 0 0 0
0 0 1 S S S 0 0 0 0

Please output the completed grid by replacing all 'V' cells with either '1' (solid) or '0'
(empty).
Maintain the same format as the input:  one row per line, cells separated by spaces, and
the total number of rows and columns unchanged.
Return only the completed grid without any additional explanation.
```

Model Completion (differences in red):

```
L L L L L 0 0 0 0 0
0 1 1 1 0 0 0 0 0 0
0 1 1 1 0 0 0 0 0 0
0 0 1 0 0 0 0 0 0 0
0 0 1 0 0 0 0 0 0 0
0 0 1 0 0 0 0 0 0 0
0 0 1 0 0 0 0 0 0 0
0 0 1 0 0 0 0 0 0 0
0 0 1 1 1 0 0 0 0 0
0 0 1 S S S 0 0 0 0
```

Ground Truth:

```
L L L L L 0 0 0 0 0
0 1 1 1 0 0 0 0 0 0
0 1 1 1 0 0 0 0 0 0
0 0 1 0 0 0 0 0 0 0
0 0 1 0 0 0 0 0 0 0
0 0 1 0 0 0 0 0 0 0
0 0 1 0 0 0 0 0 0 0
0 0 1 0 0 0 0 0 0 0
0 0 1 1 1 0 0 0 0 0
0 0 1 S S S 0 0 0 0
```

M.2 Prompt and Completion - 1 Random Cell Easy: Physics-Enhanced

Prompt and Completion - 1 Random Cell Easy: Physics-Enhanced

Prompt:

```
You are given a structural material distribution represented as a grid.  Each cell can
have one of the following states:
- 'L' indicates applied load.
- 'V' indicates void.
- 'S' indicates support.

The goal is to predict the correct material distribution by filling in all 'V' cells
with either '1' (solid) or '0' (empty), based on the surrounding structure and implicit
physical reasoning (such as load paths, supports, and forces).

Relevant physical knowledge for reasoning:
- Loads ('L') create forces that must be transferred through continuous material paths to
supports ('S').
- Stress follows the shortest stiff path from loads to supports.
- Any material cell that is disconnected from both loads and supports does not carry force
and should be avoided.
- Structures should satisfy equilibrium:  all loads must eventually be resisted by at
least one support.
- Use as little material as possible while still maintaining at least one valid load path.

Below is the input grid with masked regions:

0 0 L L L 0 0 0 0 0
0 0 0 1 0 0 0 0 0 0
0 0 0 V 0 0 0 0 0 0
0 0 0 1 0 0 0 0 0 0
0 0 0 1 0 0 0 0 0 0
0 0 0 1 0 0 0 0 0 0
0 0 1 1 1 0 0 0 0 0
0 0 1 1 1 0 0 0 0 0
0 0 1 1 1 0 0 0 0 0
S S S S S S S S S 0

Please output the completed grid by replacing all 'V' cells with either '1' (solid) or '0'
(empty).
Maintain the same format as the input:  one row per line, cells separated by spaces, and
the total number of rows and columns unchanged.
Return only the completed grid without any additional explanation.
```

Model Completion (differences in red):

```
0 0 L L L 0 0 0 0 0
0 0 0 1 0 0 0 0 0 0
0 0 0 1 0 0 0 0 0 0
0 0 0 1 0 0 0 0 0 0
0 0 0 1 0 0 0 0 0 0
0 0 0 1 0 0 0 0 0 0
0 0 1 1 1 0 0 0 0 0
0 0 1 1 1 0 0 0 0 0
0 0 1 1 1 0 0 0 0 0
S S S S S S S S S 0
```

Ground Truth:

```
0 0 L L L 0 0 0 0 0
0 0 0 1 0 0 0 0 0 0
0 0 0 1 0 0 0 0 0 0
0 0 0 1 0 0 0 0 0 0
0 0 0 1 0 0 0 0 0 0
0 0 0 1 0 0 0 0 0 0
0 0 1 1 1 0 0 0 0 0
0 0 1 1 1 0 0 0 0 0
0 0 1 1 1 0 0 0 0 0
S S S S S S S S S 0
```

M.3 Results for Claude Opus 4 on All Tasks

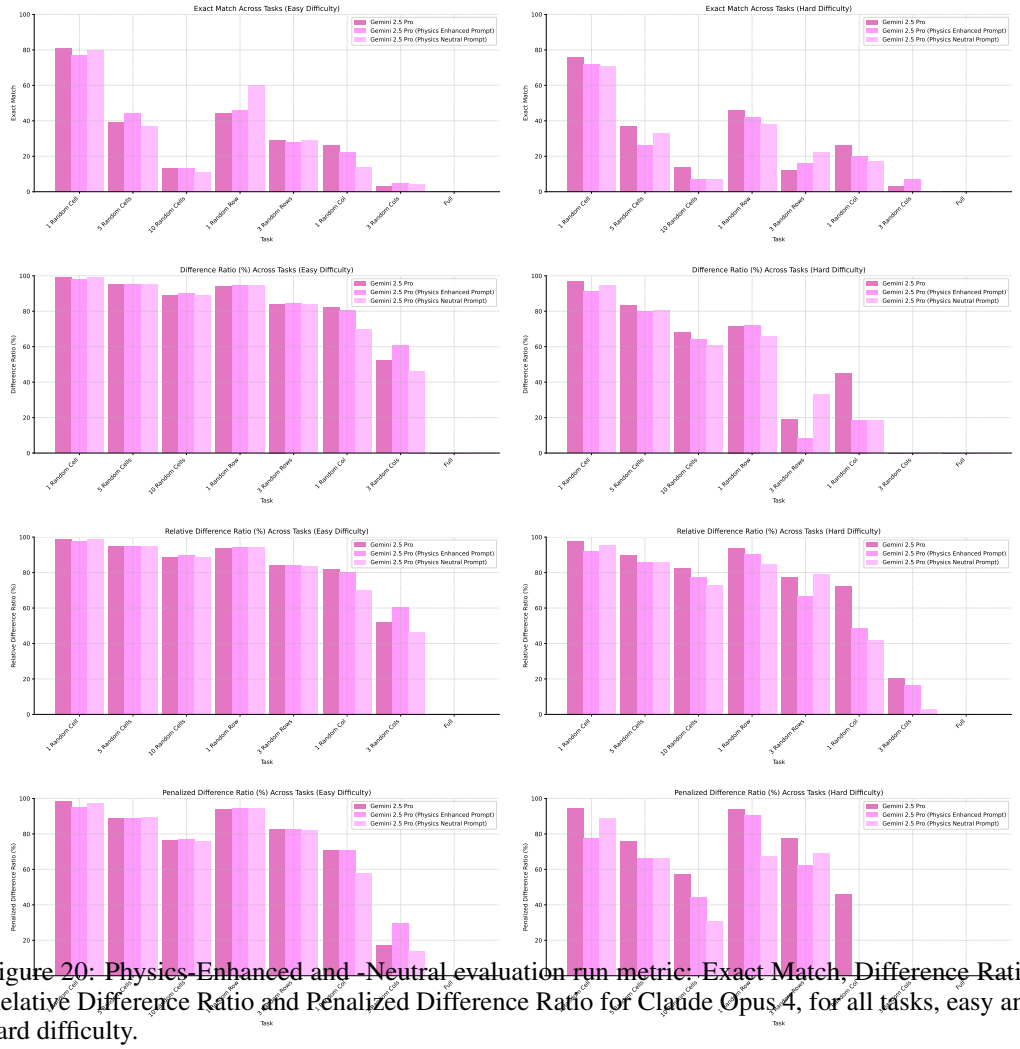

Figure 20: Physics-Enhanced and -Neutral evaluation run metric: Exact Match, Difference Ratio, Relative Difference Ratio and Penalized Difference Ratio for Claude Opus 4, for all tasks, easy and hard difficulty.

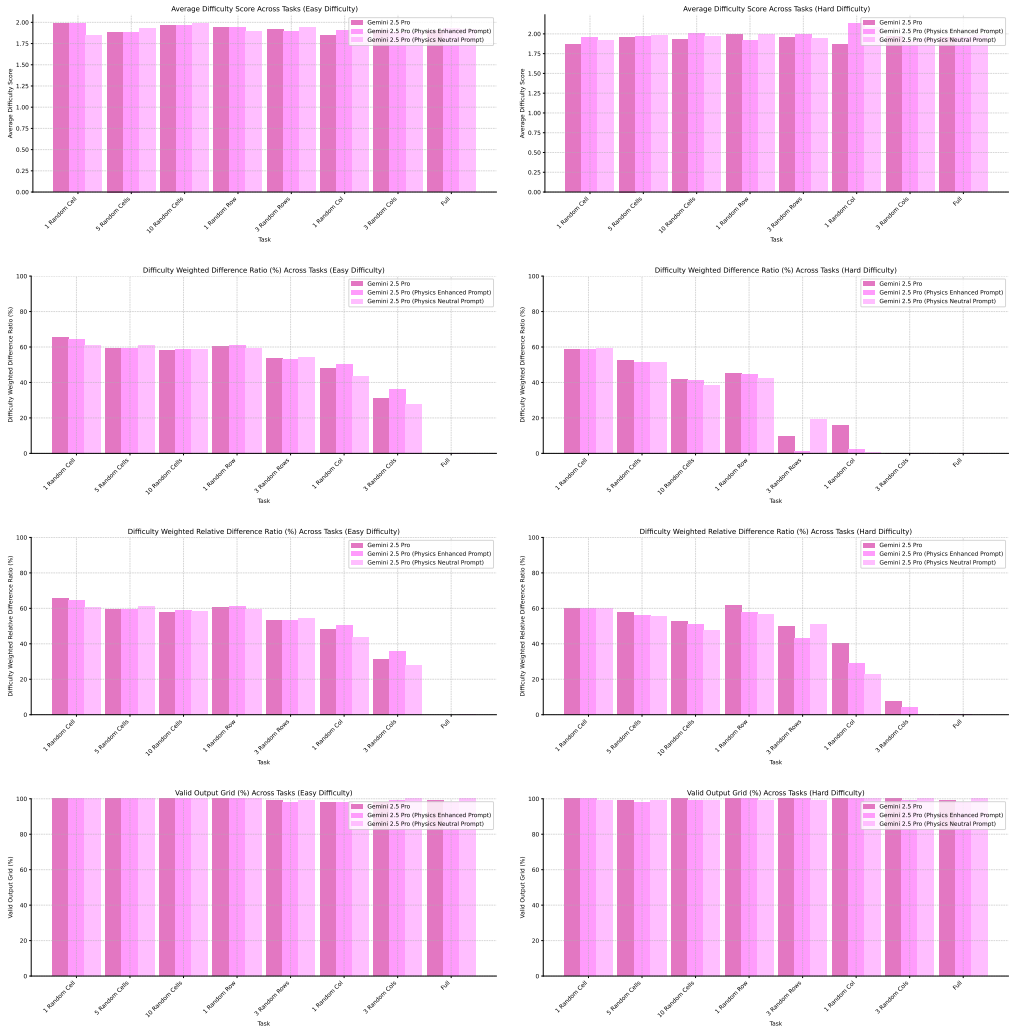

Figure 21: Physics-Enhanced and -Neutral evaluation run metric: Average Difficulty Score, Difficulty Weighted Difference Ratio, Difficulty Weighted Relative Difference Ratio and Grid Validity for Claude Opus 4, for all tasks, easy and hard difficulty.

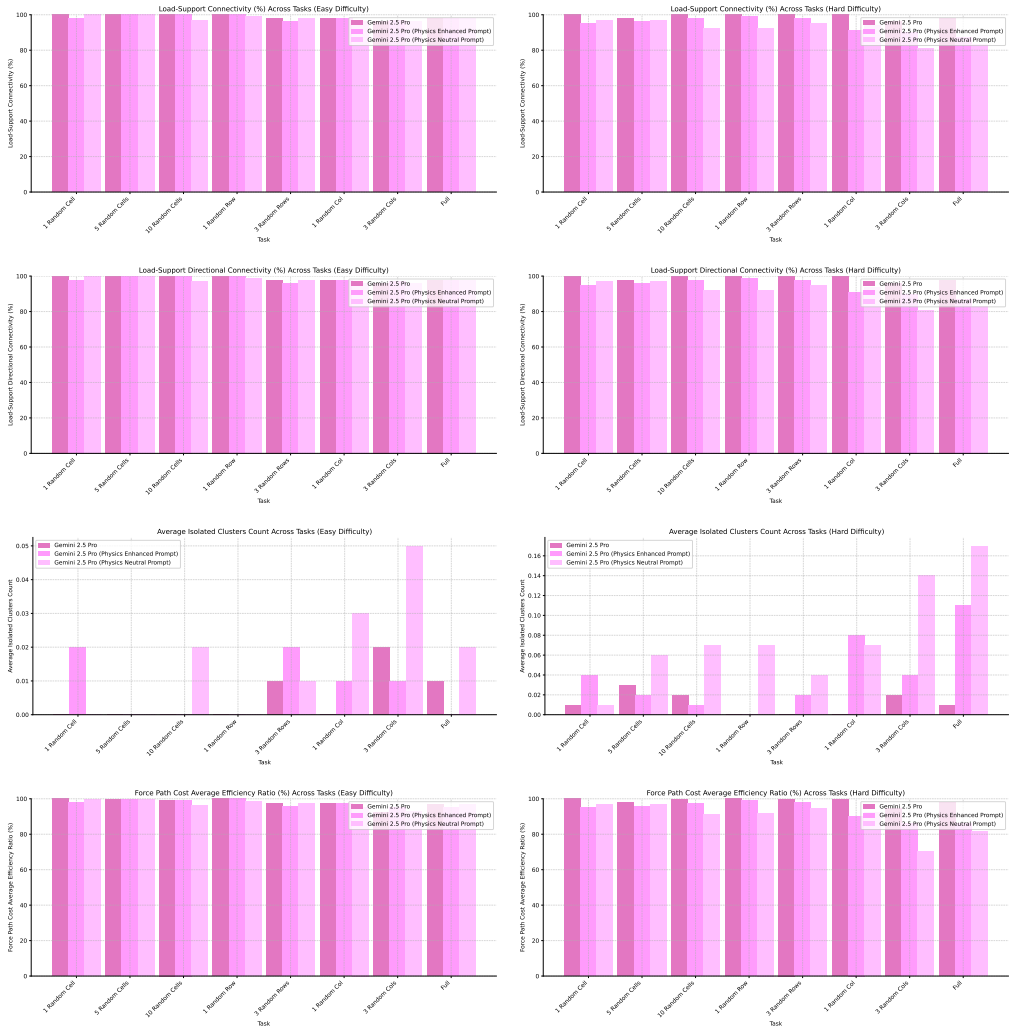

Figure 22: Physics-Enhanced and -Neutral evaluation result: Load-Support Connectivity, Directional Load-Support Connectivity, Average Isolated Cluster Count and Force Path Cost Average Efficiency Ratio for Claude Opus 4, for all tasks, easy and hard difficulty.



Figure 23: Physics-Enhanced and -Neutral evaluation result, simulated structural metrics: Topology Score, Structural Efficiency, Material Efficiency, Compliance Efficiency vs Ground Truth and Load Carrying for Claude Opus 4, for all tasks, easy and hard difficulty.

Table 13: Physics-Enhanced and -Neutral evaluation results for all metrics, for Claude Opus 4, for all tasks, easy and hard difficulty.

		Easy			Hard		
Task	Metric	Gemini 2.5 Pro (Base)	Gemini 2.5 Pro (Physics-Enhanced Prompt)	Gemini 2.5 Pro (Physics-Neutral Prompt)	Gemini 2.5 Pro (Base)	Gemini 2.5 Pro (Physics-Enhanced Prompt)	Gemini 2.5 Pro (Physics-Neutral Prompt)
1 Random Cell	Exact Match ↑	**81**	77	80	**76**	72	71
	Difference Ratio (%) ↑	**99.03**	97.87	98.93	**96.70**	91.20	94.52
	Relative Difference Ratio (%) ↑	**99.03**	97.87	98.93	**97.88**	92.22	95.33
	Penalized Difference Ratio (%) ↑	**98.37**	95.26	97.20	**94.31**	77.77	88.60
	Average Difficulty Score	**1.99**	**1.99**	1.85	1.86	**1.96**	1.92
	Difficulty Weighted Difference Ratio (%) ↑	**65.47**	64.69	60.78	58.87	59.00	**59.30**
	Difficulty Weighted Relative Difference Ratio (%) ↑	**65.47**	64.69	60.78	60.01	60.03	**60.11**
	Valid Output Grid ↑	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	99.00
	Load-Support Connectivity (%) ↑	**100.00**	98.00	**100.00**	**100.00**	95.00	97.00
	Load-Support Directional Connectivity (%) ↑	**100.00**	98.00	**100.00**	**100.00**	95.00	97.00
	Average Isolated Clusters Count ↓	0.00	**0.02**	0.00	0.01	**0.04**	0.01
	Force Path Cost Average Efficiency Ratio (%) ↑	**99.94**	97.94	99.83	**99.92**	94.93	96.80
5 Random Cells	Exact Match ↑	39	**44**	37	**37**	26	33
	Difference Ratio (%) ↑	95.08	**95.12**	95.08	**83.19**	79.81	80.21
	Relative Difference Ratio (%) ↑	95.08	**95.12**	95.08	**89.98**	85.92	85.99
	Penalized Difference Ratio (%) ↑	88.59	88.62	**89.15**	**75.62**	66.34	66.29
	Average Difficulty Score	1.89	1.89	**1.93**	1.96	1.97	**1.99**
	Difficulty Weighted Difference Ratio (%) ↑	59.48	59.56	**60.93**	**52.60**	51.32	51.35
	Difficulty Weighted Relative Difference Ratio (%) ↑	59.48	59.56	**60.93**	**57.87**	55.85	55.77
	Valid Output Grid ↑	**100.00**	**100.00**	**100.00**	**99.00**	98.00	**99.00**
	Load-Support Connectivity (%) ↑	**100.00**	**100.00**	**100.00**	**98.00**	96.00	97.00
	Load-Support Directional Connectivity (%) ↑	**100.00**	**100.00**	**100.00**	**98.00**	96.00	97.00
	Average Isolated Clusters Count ↓	**0.00**	**0.00**	**0.00**	0.03	0.02	**0.06**
	Force Path Cost Average Efficiency Ratio (%) ↑	99.75	**99.75**	99.73	**97.84**	95.55	96.83
10 Random Cells	Exact Match ↑	**13**	**13**	11	**14**	7	7
	Difference Ratio (%) ↑	88.88	**89.86**	88.77	**67.83**	64.28	60.77
	Relative Difference Ratio (%) ↑	88.88	**89.86**	88.77	**82.80**	77.50	72.96
	Penalized Difference Ratio (%) ↑	76.21	**77.19**	75.70	**57.50**	44.08	30.78
	Average Difficulty Score	1.97	1.97	**1.99**	1.94	**2.01**	1.97
	Difficulty Weighted Difference Ratio (%) ↑	58.06	**58.69**	58.53	**41.99**	41.53	38.61
	Difficulty Weighted Relative Difference Ratio (%) ↑	58.06	**58.69**	58.53	**52.66**	51.13	47.43
	Valid Output Grid ↑	**100.00**	**100.00**	**100.00**	**100.00**	99.00	99.00
	Load-Support Connectivity (%) ↑	**100.00**	**100.00**	97.00	**100.00**	98.00	92.00
	Load-Support Directional Connectivity (%) ↑	**100.00**	**100.00**	97.00	**100.00**	98.00	92.00
	Average Isolated Clusters Count ↓	0.00	0.00	**0.02**	0.02	0.01	**0.07**
	Force Path Cost Average Efficiency Ratio (%) ↑	99.31	**99.34**	96.19	**99.34**	97.12	91.07
1 Random Row	Exact Match ↑	44	46	**60**	**46**	42	38
	Difference Ratio (%) ↑	93.90	94.56	**94.59**	71.69	**72.25**	65.84
	Relative Difference Ratio (%) ↑	93.90	94.56	**94.59**	**93.86**	90.55	84.59
	Penalized Difference Ratio (%) ↑	93.90	94.56	**94.59**	**93.86**	90.55	67.23
	Average Difficulty Score	**1.94**	**1.94**	1.89	1.99	1.92	**1.99**
	Difficulty Weighted Difference Ratio (%) ↑	60.55	**60.98**	59.59	**45.14**	44.52	42.47
	Difficulty Weighted Relative Difference Ratio (%) ↑	60.55	**60.98**	59.59	**61.91**	57.70	56.45
	Valid Output Grid ↑	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	99.00
	Load-Support Connectivity (%) ↑	**100.00**	**100.00**	99.00	**100.00**	99.00	92.00
	Load-Support Directional Connectivity (%) ↑	**100.00**	**100.00**	99.00	**100.00**	99.00	92.00
	Average Isolated Clusters Count ↓	**0.00**	**0.00**	0.00	0.00	0.00	**0.07**
	Force Path Cost Average Efficiency Ratio (%) ↑	**99.99**	99.94	98.65	**99.99**	99.00	91.96
3 Random Rows	Exact Match ↑	**29**	28	**29**	12	16	**22**
	Difference Ratio (%) ↑	84.09	**84.46**	83.85	18.75	8.05	**33.21**
	Relative Difference Ratio (%) ↑	84.09	**84.46**	83.85	77.42	66.67	**78.96**
	Penalized Difference Ratio (%) ↑	**82.84**	82.61	82.18	**77.42**	62.27	69.10
	Average Difficulty Score	1.92	1.89	**1.95**	1.95	**1.99**	1.94
	Difficulty Weighted Difference Ratio (%) ↑	53.52	53.40	**54.34**	9.64	1.50	**19.13**
	Difficulty Weighted Relative Difference Ratio (%) ↑	53.52	53.40	**54.34**	49.98	43.20	**50.91**
	Valid Output Grid ↑	**99.00**	98.00	**99.00**	**100.00**	**100.00**	99.00
	Load-Support Connectivity (%) ↑	**98.00**	96.00	**98.00**	**100.00**	98.00	95.00
	Load-Support Directional Connectivity (%) ↑	**98.00**	96.00	**98.00**	**100.00**	98.00	95.00
	Average Isolated Clusters Count ↓	0.01	**0.02**	0.01	0.00	0.02	**0.04**
	Force Path Cost Average Efficiency Ratio (%) ↑	**97.54**	95.88	97.43	**99.87**	97.97	94.45
1 Random Column	Exact Match ↑	**26**	22	14	**26**	20	17
	Difference Ratio (%) ↑	**81.95**	80.45	69.86	**44.93**	18.64	18.41
	Relative Difference Ratio (%) ↑	**81.95**	80.45	69.86	**72.53**	48.66	42.16
	Penalized Difference Ratio (%) ↑	70.55	57.94	**70.85**	**46.17**	-2.29	-5.47
	Average Difficulty Score	1.85	**1.90**	1.88	1.87	**2.13**	2.08
	Difficulty Weighted Difference Ratio (%) ↑	48.21	**50.24**	43.85	**15.91**	2.64	0.94
	Difficulty Weighted Relative Difference Ratio (%) ↑	48.21	**50.24**	43.85	**40.27**	29.14	22.96
	Valid Output Grid ↑	**98.00**	**98.00**	97.00	**100.00**	**100.00**	**100.00**
	Load-Support Connectivity (%) ↑	**98.00**	**98.00**	96.00	**100.00**	91.00	94.00
	Load-Support Directional Connectivity (%) ↑	**98.00**	**98.00**	96.00	**100.00**	91.00	94.00
	Average Isolated Clusters Count ↓	0.00	0.01	**0.03**	0.00	**0.08**	0.07
	Force Path Cost Average Efficiency Ratio (%) ↑	97.30	**97.34**	95.36	**99.48**	90.32	92.43
3 Random Columns	Exact Match ↑	3	**5**	4	3	**7**	0
	Difference Ratio (%) ↑	52.03	**60.60**	46.29	-56.28	**-38.97**	-56.82
	Relative Difference Ratio (%) ↑	52.03	**60.60**	46.29	**20.55**	16.32	3.17
	Penalized Difference Ratio (%) ↑	17.06	**29.47**	13.61	**-60.21**	-68.23	-87.68
	Average Difficulty Score	**1.90**	1.88	1.88	**1.96**	1.90	1.95
	Difficulty Weighted Difference Ratio (%) ↑	31.14	**36.09**	28.05	-50.66	**-36.28**	-43.02
	Difficulty Weighted Relative Difference Ratio (%) ↑	31.14	**36.09**	28.05	**7.72**	4.18	0.15
	Valid Output Grid ↑	98.00	99.00	**100.00**	**100.00**	99.00	**100.00**
	Load-Support Connectivity (%) ↑	96.00	**97.00**	96.00	**96.00**	90.00	81.00
	Load-Support Directional Connectivity (%) ↑	96.00	**97.00**	96.00	**96.00**	90.00	81.00
	Average Isolated Clusters Count ↓	0.02	0.01	**0.05**	0.02	0.04	**0.14**
	Force Path Cost Average Efficiency Ratio (%) ↑	92.30	**95.26**	92.66	**94.28**	86.24	70.47
Full	Exact Match ↑	**0**	**0**	**0**	**0**	**0**	**0**
	Difference Ratio (%) ↑	-25.03	-39.34	**-16.98**	-548.98	-577.88	**-439.73**
	Relative Difference Ratio (%) ↑	-25.03	-39.34	**-16.98**	-316.57	-358.96	**-282.89**
	Penalized Difference Ratio (%) ↑	-25.96	-39.34	**-17.62**	**-318.95**	-390.30	-326.53
	Average Difficulty Score	1.91	**1.93**	1.91	1.95	1.95	**1.98**
	Difficulty Weighted Difference Ratio (%) ↑	-14.61	-23.66	**-10.88**	-360.89	-380.19	**-295.08**
	Difficulty Weighted Relative Difference Ratio (%) ↑	-14.61	-23.66	**-10.88**	-208.21	-236.96	**-192.17**
	Valid Output Grid ↑	99.00	98.00	**100.00**	99.00	98.00	**100.00**
	Load-Support Connectivity (%) ↑	**98.00**	**98.00**	**98.00**	**98.00**	87.00	83.00
	Load-Support Directional Connectivity (%) ↑	**98.00**	**98.00**	**98.00**	**98.00**	87.00	83.00
	Average Isolated Clusters Count ↓	0.01	0.00	**0.02**	0.01	0.11	**0.17**
	Force Path Cost Average Efficiency Ratio (%) ↑	**96.85**	95.30	96.63	**97.83**	86.29	81.43
Average	Exact Match ↑	**29.38**	**29.38**	**29.38**	**26.75**	23.75	23.50
	Difference Ratio (%) ↑	**71.24**	70.45	70.05	-27.77	-35.33	**-17.95**
	Relative Difference Ratio (%) ↑	**71.24**	70.45	70.05	**27.31**	14.86	22.53
	Penalized Difference Ratio (%) ↑	**62.70**	62.40	61.59	**8.21**	-14.97	-12.21
	Average Difficulty Score	1.92	**1.92**	1.91	1.94	**1.98**	1.98
	Difficulty Weighted Difference Ratio (%) ↑	**45.23**	45.00	44.40	-23.42	-26.99	**-16.16**
	Difficulty Weighted Relative Difference Ratio (%) ↑	**45.23**	45.00	44.40	**15.28**	8.03	12.70
	Valid Output Grid ↑	99.25	99.12	**99.50**	**99.75**	99.25	99.38
	Load-Support Connectivity (%) ↑	**98.75**	98.38	98.00	**99.00**	94.25	91.38
	Load-Support Directional Connectivity (%) ↑	**98.75**	98.38	98.00	**99.00**	94.25	91.38
	Average Isolated Clusters Count ↓	0.01	0.01	**0.02**	0.01	0.04	**0.08**
	Force Path Cost Average Efficiency Ratio (%) ↑	**97.87**	97.59	97.06	**98.57**	93.43	89.43

N SAMPLE PROMPTS & COMPLETIONS

N.1 CLAUDE 3.7 SONNET - FULL EASY

Claude 3.7 Sonnet - Full Easy

Prompt:

```
You are given a structural material distribution represented as a grid.  Each cell can
have one of the following states:
- 'L' indicates applied load.
- 'V' indicates void.
- 'S' indicates support.

The goal is to predict the correct material distribution by filling in all 'V' cells
with either '1' (solid) or '0' (empty), based on the surrounding structure and implicit
physical reasoning (such as load paths, supports, and forces).

Important:  The completed structure should use as little material as possible while
remaining stable and plausible for carrying the applied forces.  Minimize material usage
unless necessary for structural support.

Below is the input grid with masked regions:

0 L L L L L 0 0 0 0
V V V V V V V V V V
V V V V V V V V V V
V V V V V V V V V V
V V V V V V V V V V
V V V V V V V V V V
V V V V V V V V V V
V V V V V V V V V V
V V V V V V V V V V
0 S S S S S S 0 0 0

Please output the completed grid by replacing all 'V' cells with either '1' (solid) or '0'
(empty).
Maintain the same format as the input:  one row per line, cells separated by spaces, and
the total number of rows and columns unchanged.
Return only the completed grid without any additional explanation.
```

Model Completion (differences in red):

```
0 L L L L L 0 0 0 0
0 1 1 1 1 1 0 0 0 0
0 1 0 0 0 1 0 0 0 0
0 1 0 0 0 1 0 0 0 0
0 1 0 0 0 1 0 0 0 0
0 1 0 0 0 1 0 0 0 0
0 1 0 0 0 1 0 0 0 0
0 1 0 0 0 1 0 0 0 0
0 1 1 1 1 1 0 0 0 0
0 S S S S S S 0 0 0
```

Ground Truth:

```
0 L L L L L 0 0 0 0
0 0 1 1 1 0 0 0 0 0
0 0 1 1 1 0 0 0 0 0
0 0 0 1 0 0 0 0 0 0
0 0 0 1 0 0 0 0 0 0
0 0 0 1 0 0 0 0 0 0
0 0 0 1 0 0 0 0 0 0
0 0 0 1 0 0 0 0 0 0
0 0 0 1 0 0 0 0 0 0
0 S S S S S S 0 0 0
```

N.2 CLAUDE 3.7 SONNET - FULL HARD

Claude 3.7 Sonnet - Full Hard

Prompt:

```
You are given a structural material distribution represented as a grid.  Each cell can
have one of the following states:
- 'L' indicates applied load.
- 'V' indicates void.
- 'S' indicates support.

The goal is to predict the correct material distribution by filling in all 'V' cells with
a floating point number between 0 and 1, with one decimal place (e.g., 0.0, 0.1, 0.2, ...,
1.0), based on the surrounding structure and implicit physical reasoning (such as load
paths, supports, and forces).

Important:  The completed structure should use as little material as possible while
remaining stable and plausible for carrying the applied forces.  Minimize material usage
unless necessary for structural support.

Below is the input grid with masked regions:

0.0 L   L L L L L   L   L
V   V   V V V V V V   V   V
V   V   V V V V V V   V   V
V   V   V V V V V V   V   V
V   V   V V V V V V   V   V
V   V   V V V V V V   V   V
V   V   V V V V V V   V   V
V   V   V V V V V V   V   V
V   V   V V V V V V   V   V
0.0 0.0 S S S S S 0.0 0.0 0.0

Please output the completed grid by replacing all 'V' cells with a floating point number
between 0 and 1, with one decimal place (e.g., 0.0, 0.1, 0.2, ..., 1.0).
Maintain the same format as the input:  one row per line, cells separated by spaces, and
the total number of rows and columns unchanged.
Return only the completed grid without any additional explanation.
```

Model Completion (differences in red):

```
0.0 L   L   L   L   L   L   L   L   L
0.0 0.0 0.1 0.2 0.6 0.6 0.2 0.1 0.0 0.0
0.0 0.0 0.2 0.3 0.8 0.8 0.3 0.2 0.0 0.0
0.0 0.0 0.3 0.5 1.0 1.0 0.5 0.3 0.0 0.0
0.0 0.0 0.4 0.7 1.0 1.0 0.7 0.4 0.0 0.0
0.0 0.0 0.5 0.9 1.0 1.0 0.9 0.5 0.0 0.0
0.0 0.0 0.6 1.0 1.0 1.0 1.0 0.6 0.0 0.0
0.0 0.0 0.7 1.0 1.0 1.0 1.0 0.7 0.0 0.0
0.0 0.0 0.9 1.0 1.0 1.0 1.0 0.9 0.0 0.0
0.0 0.0 S   S   S   S   S   0.0 0.0 0.0
```

Ground Truth:

```
0.0 L   L   L   L   L   L   L   L   L
0.0 0.0 0.2 0.3 0.0 0.3 0.0 0.3 0.2 0.0
0.0 0.0 0.0 0.4 0.2 0.3 0.2 0.4 0.0 0.0
0.0 0.0 0.0 0.2 0.4 0.0 0.4 0.2 0.0 0.0
0.0 0.0 0.0 0.0 0.4 0.0 0.4 0.0 0.0 0.0
0.0 0.0 0.0 0.0 0.2 0.5 0.2 0.0 0.0 0.0
0.0 0.0 0.0 0.0 0.6 0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0 0.6 0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0 0.6 0.0 0.0 0.0 0.0 0.0
0.0 0.0 S   S   S   S   S   0.0 0.0 0.0
```

N.3 DEEPSEEK-V3 - FULL EASY

DeepSeek-V3 -Full Easy

Prompt:

```
You are given a structural material distribution represented as a grid.  Each cell can
have one of the following states:
- 'L' indicates applied load.
- 'V' indicates void.
- 'S' indicates support.

The goal is to predict the correct material distribution by filling in all 'V' cells
with either '1' (solid) or '0' (empty), based on the surrounding structure and implicit
physical reasoning (such as load paths, supports, and forces).

Important:  The completed structure should use as little material as possible while
remaining stable and plausible for carrying the applied forces.  Minimize material usage
unless necessary for structural support.

Below is the input grid with masked regions:

0 L L L L L 0 0 0 0
V V V V V V V V V V
V V V V V V V V V V
V V V V V V V V V V
V V V V V V V V V V
V V V V V V V V V V
V V V V V V V V V V
V V V V V V V V V V
V V V V V V V V V V
0 S S S S S S 0 0 0

Please output the completed grid by replacing all 'V' cells with either '1' (solid) or '0'
(empty).
Maintain the same format as the input:  one row per line, cells separated by spaces, and
the total number of rows and columns unchanged.
Return only the completed grid without any additional explanation.
```

Model Completion (differences in red):

```
0 L L L L L 0 0 0 0
1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1
0 S S S S S S 0 0 0
```

Ground Truth:

```
0 L L L L L 0 0 0 0
0 0 1 1 1 0 0 0 0 0
0 0 1 1 1 0 0 0 0 0
0 0 0 1 0 0 0 0 0 0
0 0 0 1 0 0 0 0 0 0
0 0 0 1 0 0 0 0 0 0
0 0 0 1 0 0 0 0 0 0
0 0 0 1 0 0 0 0 0 0
0 0 0 1 0 0 0 0 0 0
0 S S S S S S 0 0 0
```

N.4 DEEPSEEK-V3 - FULL HARD

DeepSeek-V3 -Full Hard

Prompt:

```
You are given a structural material distribution represented as a grid.  Each cell can
have one of the following states:
- 'L' indicates applied load.
- 'V' indicates void.
- 'S' indicates support.

The goal is to predict the correct material distribution by filling in all 'V' cells with
a floating point number between 0 and 1, with one decimal place (e.g., 0.0, 0.1, 0.2, ...,
1.0), based on the surrounding structure and implicit physical reasoning (such as load
paths, supports, and forces).

Important:  The completed structure should use as little material as possible while
remaining stable and plausible for carrying the applied forces.  Minimize material usage
unless necessary for structural support.

Below is the input grid with masked regions:

0.0 L   L L L L L   L   L
V   V   V V V V V V   V   V
V   V   V V V V V V   V   V
V   V   V V V V V V   V   V
V   V   V V V V V V   V   V
V   V   V V V V V V   V   V
V   V   V V V V V V   V   V
V   V   V V V V V V   V   V
V   V   V V V V V V   V   V
0.0 0.0 S S S S S 0.0 0.0 0.0

Please output the completed grid by replacing all 'V' cells with a floating point number
between 0 and 1, with one decimal place (e.g., 0.0, 0.1, 0.2, ..., 1.0).
Maintain the same format as the input:  one row per line, cells separated by spaces, and
the total number of rows and columns unchanged.
Return only the completed grid without any additional explanation.
```

Model Completion (differences in red):

```
0.0 L   L  L   L   L   L   L   L   L
0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
0.0 0.0 S   S   S   S   S   0.0 0.0 0.0
```

Ground Truth:

```
0.0 L   L   L   L   L   L   L   L   L
0.0 0.0 0.2 0.3 0.0 0.3 0.0 0.3 0.2 0.0
0.0 0.0 0.0 0.4 0.2 0.3 0.2 0.4 0.0 0.0
0.0 0.0 0.0 0.2 0.4 0.0 0.4 0.2 0.0 0.0
0.0 0.0 0.0 0.0 0.4 0.0 0.4 0.0 0.0 0.0
0.0 0.0 0.0 0.0 0.2 0.5 0.2 0.0 0.0 0.0
0.0 0.0 0.0 0.0 0.6 0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0 0.6 0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0 0.6 0.0 0.0 0.0 0.0 0.0
0.0 0.0 S   S   S   S   S   0.0 0.0 0.0
```

N.5 GPT-4.1 - FULL EASY

GPT-4.1 - Full Easy

Prompt:

```
You are given a structural material distribution represented as a grid.  Each cell can
have one of the following states:
- 'L' indicates applied load.
- 'V' indicates void.
- 'S' indicates support.

The goal is to predict the correct material distribution by filling in all 'V' cells
with either '1' (solid) or '0' (empty), based on the surrounding structure and implicit
physical reasoning (such as load paths, supports, and forces).

Important:  The completed structure should use as little material as possible while
remaining stable and plausible for carrying the applied forces.

Below is the input grid with masked regions:

0 L L L L L 0 0 0 0
V V V V V V V V V V
V V V V V V V V V V
V V V V V V V V V V
V V V V V V V V V V
V V V V V V V V V V
V V V V V V V V V V
V V V V V V V V V V
V V V V V V V V V V
0 S S S S S S 0 0 0

Please output the completed grid by replacing all 'V' cells with either '1' (solid) or '0'
(empty).
Maintain the same format as the input:  one row per line, cells separated by spaces, and
the total number of rows and columns unchanged.
Return only the completed grid without any additional explanation.
```

Model Completion (differences in red):

```
0 L L L L L 0 0 0 0
1 1 1 1 1 1 1 0 0 0
1 0 0 0 0 0 1 0 0 0
1 0 0 0 0 0 1 0 0 0
1 0 0 0 0 0 1 0 0 0
1 0 0 0 0 0 1 0 0 0
1 0 0 0 0 0 1 0 0 0
1 0 0 0 0 0 1 0 0 0
1 1 1 1 1 1 1 0 0 0
0 S S S S S S 0 0 0
```

Ground Truth:

```
0 L L L L L 0 0 0 0
0 0 1 1 1 0 0 0 0 0
0 0 1 1 1 0 0 0 0 0
0 0 0 1 0 0 0 0 0 0
0 0 0 1 0 0 0 0 0 0
0 0 0 1 0 0 0 0 0 0
0 0 0 1 0 0 0 0 0 0
0 0 0 1 0 0 0 0 0 0
0 0 0 1 0 0 0 0 0 0
0 S S S S S S 0 0 0
```

N.6 GPT-4.1 - FULL HARD

GPT-4.1 - Full Hard

Prompt:

```
You are given a structural material distribution represented as a grid.  Each cell can
have one of the following states:
- 'L' indicates applied load.
- 'V' indicates void.
- 'S' indicates support.

The goal is to predict the correct material distribution by filling in all 'V' cells with
a floating-point number between 0 and 1, with one decimal place (e.g., 0.0, 0.1, , 1.0),
based on the surrounding structure and implicit physical reasoning (such as load paths,
supports, and forces).

Important:  The completed structure should use as little material as possible while
remaining stable and plausible for carrying the applied forces.

Below is the input grid with masked regions:

0.0 L   L L L L L   L   L
V   V   V V V V V V   V   V
V   V   V V V V V V   V   V
V   V   V V V V V V   V   V
V   V   V V V V V V   V   V
V   V   V V V V V V   V   V
V   V   V V V V V V   V   V
V   V   V V V V V V   V   V
V   V   V V V V V V   V   V
0.0 0.0 S S S S S 0.0 0.0 0.0

Please output the completed grid by replacing all 'V' cells with a floating-point number
between 0 and 1, with one decimal place.
Maintain the same format as the input:  one row per line, cells separated by spaces, and
the total number of rows and columns unchanged.
Return only the completed grid without any additional explanation.
```

Model Completion (differences in red):

```
0.0 L   L   L   L   L   L   L   L   L
0.0 0.4 0.6 0.8 1.0 1.0 0.8 0.6 0.4 0.0
0.0 0.3 0.5 0.7 0.9 0.9 0.7 0.5 0.3 0.0
0.0 0.2 0.4 0.6 0.8 0.8 0.6 0.4 0.2 0.0
0.0 0.1 0.3 0.5 0.7 0.7 0.5 0.3 0.1 0.0
0.0 0.1 0.2 0.4 0.6 0.6 0.4 0.2 0.1 0.0
0.0 0.1 0.2 0.3 0.5 0.5 0.3 0.2 0.1 0.0
0.0 0.1 0.2 0.2 0.4 0.4 0.2 0.2 0.1 0.0
0.0 0.1 0.1 0.1 0.3 0.3 0.1 0.1 0.1 0.0
0.0 0.0 S   S   S   S   S   0.0 0.0 0.0
```

Ground Truth:

```
0.0 L   L   L   L   L   L   L   L   L
0.0 0.0 0.2 0.3 0.0 0.3 0.0 0.3 0.2 0.0
0.0 0.0 0.0 0.0 0.4 0.2 0.3 0.2 0.4 0.0 0.0
0.0 0.0 0.0 0.0 0.2 0.4 0.0 0.4 0.2 0.0 0.0
0.0 0.0 0.0 0.0 0.0 0.4 0.0 0.4 0.0 0.0 0.0
0.0 0.0 0.0 0.0 0.0 0.2 0.5 0.2 0.0 0.0 0.0
0.0 0.0 0.0 0.0 0.0 0.0 0.6 0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0 0.0 0.0 0.6 0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0 0.0 0.0 0.6 0.0 0.0 0.0 0.0
0.0 0.0 S   S   S   S   S   0.0 0.0 0.0
```

N.7 GEMINI 2.5 PRO - FULL EASY

Gemini 2.5 Pro - Full Easy

Prompt:

```
You are given a structural material distribution represented as a grid.  Each cell can
have one of the following states:
- 'L' indicates applied load.
- 'V' indicates void.
- 'S' indicates support.

The goal is to predict the correct material distribution by filling in all 'V' cells
with either '1' (solid) or '0' (empty), based on the surrounding structure and implicit
physical reasoning (such as load paths, supports, and forces).

Important:  The completed structure should use as little material as possible while
remaining stable and plausible for carrying the applied forces.

Below is the input grid with masked regions:

0 0 0 L L L L 0 0 0
V V V V V V V V V V
V V V V V V V V V V
V V V V V V V V V V
V V V V V V V V V V
V V V V V V V V V V
V V V V V V V V V V
V V V V V V V V V V
V V V V V V V V V V
0 0 0 0 S S S 0 0 0
```

Model Completion (differences in red):

```
0 0 0 L L L L 0 0 0
0 0 0 1 1 1 1 0 0 0
0 0 0 0 1 1 1 0 0 0
0 0 0 0 1 1 1 0 0 0
0 0 0 0 1 1 1 0 0 0
0 0 0 0 1 1 1 0 0 0
0 0 0 0 1 1 1 0 0 0
0 0 0 0 1 1 1 0 0 0
0 0 0 0 1 1 1 0 0 0
0 0 0 0 S S S 0 0 0
```

Ground Truth:

```
0 0 0 L L L L 0 0 0
0 0 0 0 1 1 0 0 0 0
0 0 0 0 1 1 0 0 0 0
0 0 0 0 1 1 0 0 0 0
0 0 0 0 1 1 0 0 0 0
0 0 0 0 1 1 0 0 0 0
0 0 0 0 1 1 0 0 0 0
0 0 0 0 1 1 0 0 0 0
0 0 0 0 1 1 0 0 0 0
0 0 0 0 S S S 0 0 0
```

N.8 GEMINI 2.5 PRO - FULL HARD

Gemini 2.5 Pro - Full Hard

Prompt:

```
You are given a structural material distribution represented as a grid.  Each cell can
have one of the following states:
- 'L' indicates applied load.
- 'V' indicates void.
- 'S' indicates support.

The goal is to predict the correct material distribution by filling in all 'V' cells with
a floating-point number between 0 and 1 (with one decimal place), based on the surrounding
structure and implicit physical reasoning.

Below is the input grid with masked regions:

0.0 0.0 L L L 0.0 0.0 0.0 0.0 0.0
V   V   V V V V   V   V   V   V
V   V   V V V V   V   V   V   V
V   V   V V V V   V   V   V   V
V   V   V V V V   V   V   V   V
V   V   V V V V   V   V   V   V
V   V   V V V V   V   V   V   V
V   V   V V V V   V   V   V   V
V   V   V V V V   V   V   V   V
S   S   S S S S   S   S   S   0.0
```

Model Completion (differences in red):

```
0.0 0.0 L   L   L   0.0 0.0 0.0 0.0 0.0
0.0 0.1 1.0 1.0 1.0 0.1 0.0 0.0 0.0 0.0
0.0 0.2 1.0 1.0 1.0 0.2 0.0 0.0 0.0 0.0
0.1 0.3 1.0 1.0 1.0 0.3 0.1 0.0 0.0 0.0
0.2 0.4 1.0 1.0 1.0 0.4 0.2 0.0 0.0 0.0
0.3 0.5 1.0 1.0 1.0 0.5 0.3 0.1 0.0 0.0
0.4 0.6 1.0 1.0 1.0 0.6 0.4 0.2 0.0 0.0
0.5 0.7 1.0 1.0 1.0 0.7 0.5 0.3 0.1 0.0
0.6 0.8 1.0 1.0 1.0 0.8 0.6 0.4 0.2 0.0
S S S   S   S   S S S S 0.0
```

Ground Truth:

```
0.0 0.0 L   L   L   0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.9 0.0 0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.8 0.0 0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.8 0.0 0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.8 0.0 0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.9 0.0 0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.1 0.8 0.1 0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.1 0.8 0.1 0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.2 0.8 0.2 0.0 0.0 0.0 0.0 0.0
S   S   S   S   S   S   S   S   S   0.0
```