SPHYR: SPATIAL-PHYSICAL REASONING BENCHMARK ON MATERIAL DISTRIBUTION

Anonymous authors
Paper under double-blind review

ABSTRACT

We introduce a novel dataset designed to benchmark the physical and spatial reasoning capabilities of Large Language Models (LLM) based on topology optimization, a method for computing optimal material distributions within a design space under prescribed loads and supports. In this dataset, LLMs are provided with conditions such as 2D boundary, applied forces and supports, and must reason about the resulting optimal material distribution. The dataset includes a variety of tasks, ranging from filling in masked regions within partial structures to predicting complete material distributions. Solving these tasks requires understanding the flow of forces and the required material distribution under given constraints, without access to simulation tools or explicit physical models, challenging models to reason about structural stability and spatial organization. Our dataset targets the evaluation of spatial and physical reasoning abilities in 2D settings, offering a complementary perspective to traditional language and logic benchmarks [1].

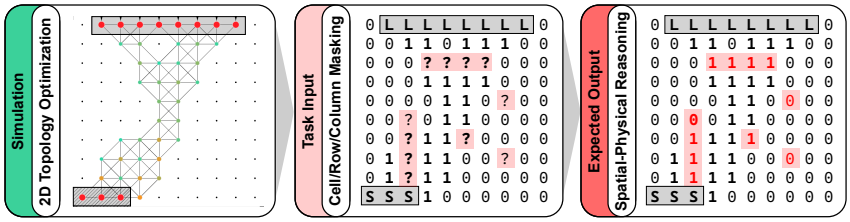

Figure 1: Topology Optimization is used to calculate material distribution. Masking individual cells, rows, columns or the complete distribution space offer challenging spatial physical reasoning tasks.

1 INTRODUCTION

Large language models (LLMs) have achieved strong performance on linguistic and logical tasks, but their ability to reason about physical systems and spatial structures remains underexplored Zhang et al. (2025). Existing benchmarks primarily probe either visual perception or text-based commonsense knowledge, but few explicitly test reasoning grounded in physical constraints.
For example, visual question-answering benchmarks such as CLEVR focus on object attributes and spatial relations in synthetic scenes Johnson et al. (2016), while intuitive physics datasets like IntPhys and Physion evaluate models' ability to predict or assess the plausibility of physical events in videos Riochet et al. (2020); Bear et al. (2022). Interactive environments such as PHYRE Bakhtin et al. (2019) and stability-focused datasets like ShapeStacks Groth et al. (2018) further probe causal reasoning and contact mechanics, whereas text-based datasets such as PIQA Bisk et al. (2019) and PhysReason Zhang et al. (2025) target physical commonsense and multi-step problem solving in language form.

Existing benchmarks have advanced our understanding of physical reasoning in LLMs, but they largely focus on object dynamics, intuitive physics, or qualitative predictions. They do not evaluate

[1] Huggingface Dataset: `anonymized`
Data Generation and Model Evaluation Code: `https://anonymous.4open.science/r/SPhyR-587C`

whether models can reason about how forces should be supported and transmitted through a structure, a capability fundamental to engineering and design. This gap leaves untested a crucial class of reasoning that requires integrating spatial layout with structural principles such as load paths, stiffness, and stability. Beyond physical reasoning, recent work like ARC-AGI-2 Chollet et al. (2025) has introduced grid-based tasks for testing abstract reasoning and generalization. While unrelated to physics, this work highlights the value of structured 2D representations for isolating reasoning capabilities. We build on this intuition but shift the focus from symbolic transformations to spatially grounded physical reasoning.

To address this gap, we introduce **SPhyR**, a new benchmark for evaluating spatial and physical reasoning in LLMs. SPhyR formulates topology optimization-inspired tasks in a grid-based format, where models must infer how to distribute material to support specified forces and constraints. By testing whether models can reason about load paths, stability, and structural connectivity from descriptions alone, SPhyR bridges the gap between language-based reasoning and physically grounded design tasks. We benchmark state-of-the-art LLMs on SPhyR and reveal fundamental limitations in their ability to integrate spatial and physical reasoning.

2 Related Work

Benchmarks for Physical and Spatial Reasoning A wide range of benchmarks probe models' understanding of physical and spatial reasoning (Table 1). CLEVR Johnson et al. (2016) evaluates visual reasoning about objects and spatial relations in synthetic scenes, while CLEVRER Yi et al. (2020) extends this to temporal and causal reasoning in videos. IntPhys Riochet et al. (2020) and Physion Bear et al. (2022) test whether models can predict or assess the plausibility of physical events, while ShapeStacks Groth et al. (2018) targets block stability prediction. In interactive settings, PHYRE Bakhtin et al. (2019) challenges agents to solve 2D physics puzzles by reasoning about actions and causal effects. Language-based datasets such as PIQA Bisk et al. (2019) and PhysReason Zhang et al. (2025) shift the focus from perception to textual physical reasoning, evaluating knowledge of everyday object interactions and multi-step physics problem solving, respectively.

While these benchmarks advance physical reasoning evaluation, they largely focus on event prediction or commonsense reasoning. None require models to determine optimal material arrangements under explicit load and support constraints - a capability crucial for real-world engineering reasoning.

Benchmark	Format	Physical Reasoning	Spatial Reasoning	Notes
CLEVR (2017)	Visual QA	✗	✓	Scene reasoning
CLEVRER (2020)	Video QA	✓	✓	Causal events
IntPhys (2018)	Video plausibility	✓	✓	Violation detection
Physion (2021)	Video prediction	✓	✓	Object behavior prediction
ShapeStacks (2016)	Image classification	✓	✓	Block stability
PHYRE (2019)	2D physics puzzles	✓	✓	Action planning
PIQA (2020)	Text QA	✓	✗	Physical commonsense
PhysReason (2023)	Text QA	✓	✗	Multi-step physics
SPhyR (ours)	Structured prediction	✓	✓	Material distribution

Table 1: 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.

Topology Optimization as a Benchmark Topology optimization (TO) Bendsøe & Sigmund (2004) is a well-established method for computing optimal material layouts in a domain under specified forces and supports. Prior work on Machine Learning (ML) in this space has focused on accelerating solvers or generating high-quality designs Banga et al. (2018); Rawat & Shen (2019). Our work repurposes topology optimization as a reasoning benchmark rather than a design tool. By framing it as a grid-based prediction problem, SPhyR tests whether LLMs can infer material distributions solely from boundary conditions and physical constraints - without access to solvers or simulation engines. This setup complements existing physical reasoning benchmarks by embedding spatial and physical structure into tasks that require more than pattern recognition.

Machine Learning for Topology Optimization Prior machine learning work for Topology Optimization (TO) has focused on developing fast, high-fidelity solvers that can predict optimized material layouts with orders-of-magnitude speedup over conventional methods Banga et al. (2018); Rawat & Shen (2019); Zhang et al. (2020). These domain-specific approaches rely on embedding explicit structural knowledge, such as physics-informed loss functions or compliance constraints, into the model architecture and training process. In contrast, SPhyR evaluates general-purpose LLMs in a zero-shot setting, probing whether emergent, implicit physical knowledge acquired during broad training can substitute for explicitly learned physics.

Structured Reasoning Beyond Physics Finally, our work connects to broader research on structured reasoning in grid-based environments. ARC-AGI-2 Chollet et al. (2025) tests abstract reasoning and generalization in symbolic, non-physical tasks. While ARC-AGI-2 and SPhyR share a structured representation, SPhyR introduces grounded physical constraints, bridging the gap between abstract symbolic reasoning and the physically grounded reasoning required for real-world design.

3 Problem Setup

Topology Optimization Task Topology optimization determines an optimal material distribution within a domain under prescribed forces and supports. All dataset samples are generated using Millipede's density-based SIMP formulation, solving a minimum-compliance problem with a fixed volume fraction (Appendix B for solver parameters). This yields well-defined, single-objective solutions that capture characteristic load paths and material connectivity.
In this work, we repurpose these topology optimization instances as reasoning tasks for LLMs. Instead of performing numerical optimization, models must predict plausible material distributions from forces, supports, and boundaries alone, requiring them to infer principles of load transfer, stability, and efficient material use, approximating the behavior of minimum-compliance topology optimization without access to simulation tools.

Input and Output Specification Each task instance in our benchmark is defined by a set of boundary conditions and a corresponding material distribution. The **inputs** provided to the model are: **2D boundaries**: A discretized 2D grid representing the spatial extent of the structure, **fixed supports**: Locations within the boundaries that act as load bearing supports and **applied forces**: Locations within the boundaries specifying external loads. The **output** expected from the model is a partial or complete **material distribution** over the domain grid, indicating where material should be placed to form a material optimized, that is minimum material distributed, but stable structure under the given boundary conditions. All inputs and outputs are represented in structured formats suitable for LLMs, through textual descriptions and serialized grids. No direct access to simulation results or numerical solvers is provided.

Reasoning Challenges The tasks in our benchmark require a combination of physical and spatial reasoning that poses significant challenges for current large language models. First, models must infer how forces propagate through the structure, deciding where material is necessary to maintain stability and support loads. This involves understanding force paths, support connectivity, and load transfer-concepts that are rarely encountered in typical LLM training data. Second, models must reason spatially about the layout of material across a 2D grid. Predicting plausible completions requires local coherence (e.g., avoiding isolated material islands) as well as global structural organization (e.g., maintaining continuous load paths from forces to supports). Moreover, models must solve these tasks without explicit simulation tools or numerical methods. Instead, they must generalize from the provided boundary conditions and partial observations, synthesizing structures that satisfy implicit physical constraints. These reasoning demands span from local (individual cells or lines) to global (complete structures), creating a rich and graded challenge space for evaluating LLM capabilities beyond language-based tasks.

Task Variations We define several task variations according to the nature and extent of the masked regions in the material distribution, and categorize them into two difficulty levels: easy and hard. Easy is distribution based on binary values such as material or no material, while hard is based on a continuous value range, 0 to 1. **N-Random Cell(s)**: Predict the material state of N randomly masked cell(s), where N is one of 1, 5 or 10. **N-Random Row(s)**: Predict the material state of N randomly

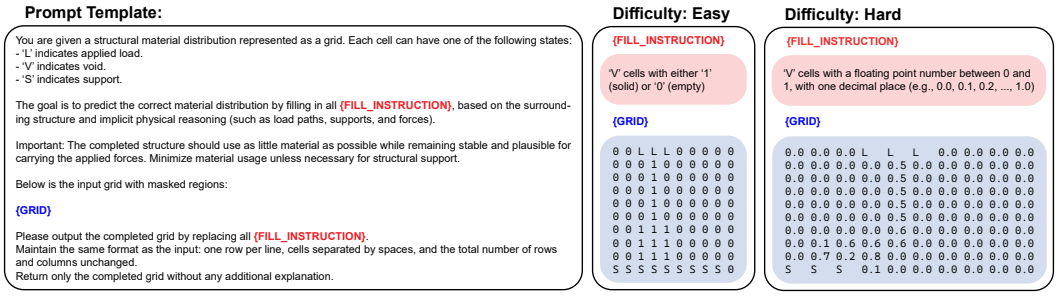

Figure 2: Prompt template used across all tasks and difficulty levels, showing instructions and grid format as served to models for evaluation.

masked row(s), where N is one of 1 or 3. **N-Random Columns(s)**: Predict the material state of N randomly masked columns(s), where N is one of 1 or 3. **Full Structure Prediction**: Predict the complete material distribution based only on boundary conditions. These variations allow us to systematically probe local and global reasoning abilities, from single-cell predictions to complex structural synthesis (Appendix C for samples).

4 Dataset Description

The SPhyR dataset was generated by solving 2D topology optimization problems, creating material distributions under various boundary conditions using the density-based solver Millipede Michalatos & Kaijima (2024). We constructed a set of 2D samples by systematically varying the positions of applied forces and supports, focusing on load-support scenarios typical of structural building design (load from the top, support on the bottom, ranging from 3 to 6 cells in width). Each material distribution was optimized for stiffness and efficiency using 10 solver iteration steps. The inherent variability in these boundary conditions ensures that tasks require generalization beyond memorization of fixed patterns (Appendix B, for detailed solver parameters).

Dataset Statistics The dataset consists of 10×10 structural optimization grids, balancing computational tractability with sufficient spatial complexity. In total, the dataset contains **1296** samples for all task variations and difficulties. These samples are organized into task-specific subsets, including cell completion, row/column completion, and full structure prediction, across both easy and hard difficulty levels. The full list of eight subject types (e.g., 1 Random Cell, 3 Random Row) and their descriptions is provided in Table C. Each sample includes structured representations of the boundary conditions and the corresponding ground truth material distribution.

Input and Output Formats Each sample in the dataset is represented as a structured input-output pair designed for compatibility with large language models. Samples are grouped into task-specific subjects, enabling targeted evaluation of different reasoning challenges.
The input consists of a natural language prompt that describes the task and defines the structural grid format. Within this grid, different symbols indicate key physical roles: **L** marks an applied load, **S** a support, and **V** (void) a masked cell whose material state must be predicted. Regions with known material values, whether binary or continuous, depending on the task difficulty (easy/hard), are explicitly included in the grid. The prompt provides clear instructions emphasizing structural plausibility and material efficiency, along with a grid where each row appears on a separate line and cell values are space-separated.
The expected output is a completed version of the same grid, where all `V` cells are replaced by predicted values (`1` or `0`) while preserving the original structure and formatting. No explanation or commentary is included in the output-only the raw grid content.
Each subject is labeled with a difficulty level. In easy variants, the ground truth material distribution is binary, focusing on high-level structural placement and discrete spatial reasoning. In hard variants, the underlying distributions are continuous or involve more complex structural dependencies, requiring finer-grained predictions and deeper reasoning about stress propagation and global support (Figure 2, for prompt template and Appendix J, E.1, for detailed model prompt and completion samples).

5 EVALUATION SETUP

5.1 EVALUATION METRICS

We evaluate model performance using three complementary families of metrics 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; and (3) **Physics-Approximating Metrics**, estimating the structural efficiency via gravity-aligned load-transfer paths. This comprehensive suite ensures robustness against simple pattern-matching success.

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 **??**, 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 cell's 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 D.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 D.1 for EPCEff calculation details, and D.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.*

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

Table 2: Summary of all evaluation metrics by category, with their types, typical ranges, and optimization direction.

5.2 Experiments

To establish baseline performance, we evaluate a broad set of contemporary language models in a zero-shot setting. From OpenAI, we include GPT-3.5 Brown et al. (2020), GPT-4.1 OpenAI et al. (2024a), and GPT-4o OpenAI et al. (2024b), representing successive generations with improved reasoning and multimodal capabilities. From Anthropic, we test Claude 3.7 Sonnet anthropic (2025a) and Claude Opus 4 anthropic (2025b), the strongest in the Claude family. From Google DeepMind, we include Gemini 1.5 Pro Team et al. (2024) and Gemini 2.5 Pro Comanici et al. (2025), designed for complex multimodal reasoning. We also assess DeepSeek-R1 DeepSeek-AI et al. (2025), an open-source model for scientific and engineering tasks, and Perplexity Sonar Team (2025a) and Sonar Reasoning Team (2025b), tuned for information-seeking and multi-step reasoning. Models are prompted (Appendix J) with structured descriptions of boundary conditions, forces, and supports, without simulation tools or external knowledge. A random subset of **100** examples spanning all task variations, difficulties and all models are evaluated under identical conditions via publicly available APIs (Table 3). Performance is measured using the metrics defined in Section 5.1.

6 Results and Analysis

We present quantitative results in Table 3 and analyze failure modes qualitatively. Detailed results on few-shot prompting, rotation, and physics-enhanced and -neutral prompt design are discussed in the subsequent sections and further expanded in the Appendix.

Table 3: **Zero-Shot Performance on SPhyR 2D Tasks (Easy vs. Hard).** Top-performing LLMs (Claude, Gemini) maintain high Load-Support Connectivity, demonstrating core topological understanding. However, performance degrades sharply on Hard tasks, with negative Difference Ratios (red) confirming inefficient material hallucination and structural over-designing across all models. (↑ indicates better, ↓ indicates worse).

		Easy					Hard				
Task	Metric	GPT 4.1	Claude Opus 4	Gemini 2.5 Pro	DeepSeek-R1	Perplexity Sonar	GPT 4.1	Claude Opus 4	Gemini 2.5 Pro	DeepSeek-R1	Perplexity Sonar
1 Random Cell	Exact Match ↑	26	**82**	81	58	52	13	**77**	76	37	13
	Difference Ratio (%) ↑	95.47	**99.05**	99.03	97.37	93.28	88.14	95.45	**96.70**	91.44	80.07
	Relative Difference Ratio (%) ↑	95.47	**99.05**	99.03	97.37	93.28	96.05	96.72	**97.88**	96.51	88.98
	Penalized Difference Ratio (%) ↑	94.82	**98.40**	98.37	96.71	92.03	92.44	93.11	**94.31**	92.90	85.12
	Average Difficulty Score	**1.99**	**1.99**	**1.99**	**1.99**	**1.99**	**1.96**	**1.96**	1.86	**1.96**	**1.96**
	Difficulty Weighted Difference Ratio (%) ↑	63.31	**65.51**	65.47	64.54	61.92	58.05	**60.97**	58.87	59.41	52.28
	Difficulty Weighted Relative Difference Ratio (%) ↑	63.31	**65.51**	65.47	64.54	61.92	62.49	62.25	60.01	**62.63**	58.26
	Valid Output Grid ↑	**100.00**	**100.00**	**100.00**	**100.00**	96.00	**100.00**	99.00	**100.00**	**100.00**	94.00
	Load-Support Connectivity (%) ↑	**100.00**	**100.00**	**100.00**	**100.00**	95.00	99.00	98.00	**100.00**	99.00	93.00
	Load-Support Directional Connectivity (%) ↑	**100.00**	**100.00**	**100.00**	**100.00**	95.00	99.00	98.00	**100.00**	99.00	93.00
	Average Isolated Clusters Count ↓	**0.32**	0.00	0.00	0.15	0.16	**0.44**	0.00	0.01	0.26	0.37
	Force Path Cost Average Efficiency Ratio (%) ↑	**99.94**	**99.94**	**99.94**	99.86	94.84	98.93	97.91	**99.92**	98.93	92.93
5 Random Cells	Exact Match ↑	1	**45**	39	15	10	0	**38**	37	15	3
	Difference Ratio (%) ↑	75.87	**95.76**	95.08	88.16	75.79	35.23	**87.27**	83.19	65.68	38.29
	Relative Difference Ratio (%) ↑	75.87	**95.76**	95.08	88.16	75.79	75.68	**91.58**	89.98	86.53	70.37
	Penalized Difference Ratio (%) ↑	69.37	**89.26**	88.59	81.66	68.84	61.72	**78.42**	75.62	71.73	57.71
	Average Difficulty Score	**1.89**	**1.89**	**1.89**	**1.89**	**1.89**	**1.97**	**1.97**	1.96	**1.97**	**1.97**
	Difficulty Weighted Difference Ratio (%) ↑	48.77	**59.89**	59.48	55.56	47.57	23.41	**56.17**	52.60	41.52	24.87
	Difficulty Weighted Relative Difference Ratio (%) ↑	48.77	**59.89**	59.48	55.56	47.57	49.57	**59.65**	57.87	56.19	45.99
	Valid Output Grid ↑	**100.00**	**100.00**	**100.00**	**100.00**	87.00	99.00	**100.00**	99.00	**100.00**	85.00
	Load-Support Connectivity (%) ↑	**100.00**	**100.00**	**100.00**	99.00	80.00	99.00	**100.00**	98.00	99.00	80.00
	Load-Support Directional Connectivity (%) ↑	**100.00**	**100.00**	**100.00**	99.00	80.00	99.00	**100.00**	98.00	99.00	80.00
	Average Isolated Clusters Count ↓	**1.64**	0.00	0.00	0.44	0.37	**1.88**	0.00	0.03	0.56	1.22
	Force Path Cost Average Efficiency Ratio (%) ↑	99.72	99.70	**99.75**	97.50	78.42	98.55	**99.49**	97.84	98.25	79.48
10 Random Cells	Exact Match ↑	0	**15**	13	2	1	0	13	**14**	3	0
	Difference Ratio (%) ↑	59.82	**89.08**	88.88	78.78	72.68	-26.62	**69.70**	67.83	36.60	-1.98
	Relative Difference Ratio (%) ↑	59.82	**89.08**	88.88	78.78	72.68	51.37	79.91	**82.80**	74.22	42.43
	Penalized Difference Ratio (%) ↑	47.02	**76.41**	76.21	65.95	59.85	20.54	49.33	**57.50**	43.59	23.37
	Average Difficulty Score	**1.97**	**1.97**	**1.97**	**1.97**	**1.97**	**2.01**	**2.01**	1.94	**2.01**	**2.01**
	Difficulty Weighted Difference Ratio (%) ↑	40.24	**58.23**	58.06	51.72	47.81	-17.38	**45.17**	41.99	22.14	-0.67
	Difficulty Weighted Relative Difference Ratio (%) ↑	40.24	**58.23**	58.06	51.72	47.81	34.45	**52.88**	52.66	48.86	29.03
	Valid Output Grid ↑	99.00	**100.00**	**100.00**	**100.00**	92.00	99.00	99.00	**100.00**	99.00	64.00
	Load-Support Connectivity (%) ↑	99.00	**100.00**	**100.00**	90.00	76.00	99.00	93.00	**100.00**	85.00	58.00
	Load-Support Directional Connectivity (%) ↑	99.00	**100.00**	**100.00**	90.00	76.00	99.00	93.00	**100.00**	85.00	58.00
	Average Isolated Clusters Count ↓	**2.27**	0.00	0.00	0.46	0.40	**2.82**	0.01	0.02	0.69	1.20
	Force Path Cost Average Efficiency Ratio (%) ↑	98.28	99.06	**99.31**	88.72	73.62	98.13	91.98	**99.34**	83.93	56.25
1 Random Row	Exact Match ↑	20	**52**	44	14	21	2	**49**	46	34	7
	Difference Ratio (%) ↑	84.50	**94.92**	93.90	73.09	77.87	18.44	**80.55**	71.69	73.13	27.07
	Relative Difference Ratio (%) ↑	84.50	**94.92**	93.90	73.09	77.87	73.91	**94.39**	93.86	94.13	74.82
	Penalized Difference Ratio (%) ↑	84.50	**94.92**	93.90	73.09	77.65	73.91	**94.39**	93.86	94.13	74.82
	Average Difficulty Score	**1.94**	**1.94**	**1.94**	**1.94**	**1.94**	1.92	1.92	**1.99**	1.92	1.92
	Difficulty Weighted Difference Ratio (%) ↑	54.72	**61.04**	60.55	47.90	50.65	11.00	**49.95**	45.14	45.37	17.05
	Difficulty Weighted Relative Difference Ratio (%) ↑	54.72	**61.04**	60.55	47.90	50.65	47.38	60.02	**61.91**	59.94	47.99
	Valid Output Grid ↑	97.00	**100.00**	**100.00**	**100.00**	91.00	98.00	**100.00**	**100.00**	**100.00**	91.00
	Load-Support Connectivity (%) ↑	97.00	**100.00**	**100.00**	97.00	73.00	97.00	99.00	**100.00**	86.00	85.00
	Load-Support Directional Connectivity (%) ↑	97.00	**100.00**	**100.00**	97.00	73.00	97.00	99.00	**100.00**	86.00	85.00
	Average Isolated Clusters Count ↓	0.02	0.00	0.00	0.00	**0.07**	0.01	0.00	0.00	0.00	**0.05**
	Force Path Cost Average Efficiency Ratio (%) ↑	96.95	99.88	**99.99**	97.03	72.93	96.99	98.92	**99.99**	86.07	84.94
3 Random Rows	Exact Match ↑	6	**35**	29	20	24	0	**21**	12	17	1
	Difference Ratio (%) ↑	59.31	**88.64**	84.09	62.64	74.23	-162.36	**32.01**	18.75	16.98	-106.35
	Relative Difference Ratio (%) ↑	59.31	**88.64**	84.09	62.64	74.23	23.74	**84.70**	77.42	76.37	46.20
	Penalized Difference Ratio (%) ↑	59.31	**88.64**	82.84	62.04	73.99	23.74	**84.70**	77.42	76.37	45.35
	Average Difficulty Score	1.89	1.89	**1.92**	1.89	1.89	**1.99**	**1.99**	1.95	**1.99**	**1.99**
	Difficulty Weighted Difference Ratio (%) ↑	37.91	**55.81**	53.52	39.27	46.97	-107.71	**17.50**	9.64	8.49	-71.32
	Difficulty Weighted Relative Difference Ratio (%) ↑	37.91	**55.81**	53.52	39.27	46.97	17.09	**55.67**	49.98	49.88	30.87
	Valid Output Grid ↑	99.00	**100.00**	99.00	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	96.00	98.00
	Load-Support Connectivity (%) ↑	95.00	**100.00**	98.00	69.00	74.00	**100.00**	97.00	**100.00**	61.00	96.00
	Load-Support Directional Connectivity (%) ↑	95.00	**100.00**	98.00	69.00	74.00	**100.00**	97.00	**100.00**	61.00	96.00
	Average Isolated Clusters Count ↓	0.01	0.00	0.01	0.27	**0.28**	0.03	0.00	0.00	**0.38**	0.01
	Force Path Cost Average Efficiency Ratio (%) ↑	94.09	**99.68**	97.54	69.22	73.47	**99.99**	96.97	99.87	61.31	95.91
1 Random Column	Exact Match ↑	0	23	**26**	8	12	0	21	**26**	15	3
	Difference Ratio (%) ↑	51.92	**83.09**	81.95	63.38	67.71	-34.24	37.46	**44.93**	29.07	-26.12
	Relative Difference Ratio (%) ↑	51.92	**83.09**	81.95	63.38	67.71	29.24	62.68	**72.53**	60.88	36.26
	Penalized Difference Ratio (%) ↑	42.23	**73.51**	70.55	53.69	55.52	-3.14	31.65	**46.17**	27.89	4.24
	Average Difficulty Score	**1.90**	**1.90**	1.85	**1.90**	**1.90**	**2.13**	**2.13**	1.87	**2.13**	**2.13**
	Difficulty Weighted Difference Ratio (%) ↑	36.02	**50.85**	48.21	40.90	44.12	-23.21	14.68	**15.91**	10.73	-17.01
	Difficulty Weighted Relative Difference Ratio (%) ↑	36.02	**50.85**	48.21	40.90	44.12	22.54	37.62	**40.27**	38.36	26.35
	Valid Output Grid ↑	**100.00**	**100.00**	98.00	**100.00**	97.00	97.00	99.00	**100.00**	97.00	91.00
	Load-Support Connectivity (%) ↑	**100.00**	**100.00**	98.00	98.00	88.00	97.00	97.00	**100.00**	97.00	90.00
	Load-Support Directional Connectivity (%) ↑	**100.00**	**100.00**	98.00	98.00	88.00	97.00	97.00	**100.00**	97.00	90.00
	Average Isolated Clusters Count ↓	**0.13**	0.00	0.00	0.08	0.06	**0.07**	0.00	0.00	0.01	0.04
	Force Path Cost Average Efficiency Ratio (%) ↑	**99.55**	94.90	97.30	97.61	73.45	96.31	84.17	**99.48**	93.47	88.37
3 Random Columns	Exact Match ↑	1	**5**	3	2	1	0	**7**	3	5	1
	Difference Ratio (%) ↑	2.46	**58.69**	52.03	18.34	24.01	-238.99	**-17.63**	-56.28	-22.76	-142.17
	Relative Difference Ratio (%) ↑	2.46	**58.69**	52.03	18.34	24.01	-35.52	**31.78**	20.55	22.87	-18.10
	Penalized Difference Ratio (%) ↑	-28.47	**27.96**	17.06	-12.54	-9.25	-108.99	**-43.98**	-60.21	-49.47	-109.09
	Average Difficulty Score	1.88	1.88	**1.90**	1.88	1.88	1.90	1.90	**1.96**	1.90	1.90
	Difficulty Weighted Difference Ratio (%) ↑	1.69	**34.52**	31.14	12.74	15.59	-153.59	**-22.99**	-50.66	-24.21	-94.09
	Difficulty Weighted Relative Difference Ratio (%) ↑	1.69	**34.52**	31.14	12.74	15.59	-25.00	**14.44**	7.72	10.10	-12.75
	Valid Output Grid ↑	99.00	**100.00**	98.00	**100.00**	93.00	94.00	**100.00**	**100.00**	88.00	80.00
	Load-Support Connectivity (%) ↑	**97.00**	94.00	96.00	93.00	70.00	94.00	93.00	**96.00**	71.00	61.00
	Load-Support Directional Connectivity (%) ↑	**97.00**	94.00	96.00	93.00	70.00	94.00	93.00	**96.00**	71.00	61.00
	Average Isolated Clusters Count ↓	**0.32**	0.00	0.02	0.12	0.24	0.30	0.00	0.02	0.01	**0.32**
	Force Path Cost Average Efficiency Ratio (%) ↑	**94.43**	88.28	92.30	86.51	54.44	92.95	77.05	**94.28**	54.47	58.37
Full	Exact Match ↑	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
	Difference Ratio (%) ↑	-62.06	-35.78	**-25.03**	-126.02	-49.16	-816.91	**-466.42**	-348.98	-585.87	-537.83
	Relative Difference Ratio (%) ↑	-62.06	-35.78	**-25.03**	-126.02	-49.16	-251.75	-177.48	-316.57	-162.59	**-144.70**
	Penalized Difference Ratio (%) ↑	-62.06	-35.78	**-25.96**	-142.73	-49.86	-251.75	-177.48	-318.95	-178.14	**-151.62**
	Average Difficulty Score	**1.93**	1.92	1.91	**1.93**	**1.93**	1.95	**1.96**	1.95	1.95	1.95
	Difficulty Weighted Difference Ratio (%) ↑	-37.69	-21.79	**-14.61**	-79.62	-29.78	-530.22	**-310.26**	-360.89	-382.38	-348.11
	Difficulty Weighted Relative Difference Ratio (%) ↑	-37.69	-21.79	**-14.61**	-79.62	-29.78	-162.49	-117.17	-208.21	-105.80	**-92.41**
	Valid Output Grid ↑	**100.00**	**100.00**	99.00	**100.00**	81.00	**100.00**	**100.00**	99.00	**100.00**	76.00
	Load-Support Connectivity (%) ↑	94.00	94.00	**98.00**	48.00	42.00	**100.00**	88.00	98.00	68.00	49.00
	Load-Support Directional Connectivity (%) ↑	94.00	94.00	**98.00**	48.00	42.00	**100.00**	88.00	98.00	68.00	49.00
	Average Isolated Clusters Count ↓	0.01	0.00	0.01	0.06	**0.09**	0.00	0.00	0.01	0.06	**0.08**
	Force Path Cost Average Efficiency Ratio (%) ↑	88.87	90.33	**96.85**	48.31	37.61	**99.86**	87.83	97.83	68.32	48.89
Average	Exact Match ↑	6.75	**32.12**	29.38	14.88	15.12	1.88	**28.25**	26.75	15.75	3.50
	Difference Ratio (%) ↑	45.91	**71.68**	71.24	44.47	54.55	-142.16	**-10.20**	-27.77	-36.97	-83.63
	Relative Difference Ratio (%) ↑	45.91	**71.68**	71.24	44.47	54.55	7.84	**45.53**	27.31	43.61	24.53
	Penalized Difference Ratio (%) ↑	38.34	**64.16**	62.70	34.73	46.10	-11.44	**26.27**	8.21	22.38	3.74
	Average Difficulty Score	**1.92**	1.92	1.92	**1.92**	**1.92**	1.98	**1.98**	1.94	1.98	1.98
	Difficulty Weighted Difference Ratio (%) ↑	30.62	**45.51**	45.23	29.12	35.61	-92.46	**-11.10**	-23.42	-27.37	-54.63
	Difficulty Weighted Relative Difference Ratio (%) ↑	30.62	**45.51**	45.23	29.12	35.61	5.75	**28.17**	15.28	27.52	16.67
	Valid Output Grid ↑	99.25	**100.00**	99.25	**100.00**	92.12	98.38	99.62	**99.75**	97.50	84.88
	Load-Support Connectivity (%) ↑	97.75	98.50	**98.75**	86.75	74.75	98.12	95.62	**99.00**	83.25	76.50
	Load-Support Directional Connectivity (%) ↑	97.75	98.50	**98.75**	86.75	74.75	98.12	95.62	**99.00**	83.25	76.50
	Average Isolated Clusters Count ↓	**0.59**	0.00	0.01	0.20	0.21	**0.69**	0.00	0.01	0.25	0.41
	Force Path Cost Average Efficiency Ratio (%) ↑	96.48	96.47	**97.87**	85.59	69.85	97.71	91.79	**98.57**	80.59	75.64

6.1 Quantitative Results

General Performance Trends Table 3 presents model performance across all task variations using the metrics defined in Section 5.1. As expected, performance degrades as task complexity increases; "Easy" (binary) tasks consistently yield higher accuracy than "Hard" (continuous) variants. Top-performing models like Claude Opus 4 and Gemini 2.5 Pro achieve near-perfect Load-Support Connectivity (>98%) and Valid Grid scores on easy tasks, suggesting that while they fail to replicate the exact ground-truth geometry (low Exact Match), they successfully reason about global structural integrity and force propagation (Appendix F, for additional plots).

The Hard Task Anomaly and Material Hallucination A critical observation in the hard (continuous) tasks is the prevalence of negative Difference Ratios across almost all models. Physically, this result implies significant over-designing: rather than converging on efficient load paths, models tend to "smear" material across the void, producing dense, non-optimal clusters. This hallucination of mass suggests that while models grasp the concept of filling space, they lack the physical intuition to minimize volume while maintaining stability, a core tenet of topology optimization.

DeepSeek-R1 and the Limits of Chain-of-Thought Notably, DeepSeek-R1, a model optimized for reasoning, exhibits a strong performance drop between easy and hard tasks. While it maintains reasonable connectivity on binary tasks, its performance collapses on continuous distributions (Table 3). We hypothesize that the model's Chain-of-Thought (CoT) process struggles to ground floating-point grid representations into spatial intuition. Instead of visualizing the physical load path, the model likely attempts arithmetic or symbolic manipulation of the density values. This symbolic approach fails to capture the global topological constraints required for stability, resulting in outputs that are computationally "reasoned" but structurally incoherent.

Rotation Experiments and Gravity Bias Among localized tasks, row completions consistently outperform column completions. Our rotation experiments ($k = 3, 270°$) reveal that this is not merely a formatting artifact but a "gravity bias." When loads are applied horizontally (simulating a cantilever or rotated structure), models frequently fail to reorient their structural intuition, attempting to build "downward" relative to the grid rather than in the direction of the force vector L. This indicates that models rely heavily on memorized visual patterns of vertical buildings rather than reasoning about the directed vector of applied forces (Appendix G, for additional rotation experiment results).

Few-Shot Experiments To investigate the in-context learning capabilities of the models, we performed few-shot experiments complementary to the zero-shot baseline. In this setting, we prepended $k = 1$ and $k = 3$ randomly selected input-output pairs from the dataset to the prompt before presenting the target test instance. The examples were drawn from the same task variation (e.g., 3 Random Row) and difficulty level (easy or hard) as the query. This approach evaluates whether models can improve their spatial reasoning and output formatting by observing valid load-path distributions, thereby allowing us to quantify the extent to which physical constraints can be inferred from examples versus explicit instructions (Appendix H, for additional few-shot experiment results).

Physics-Enhanced vs. Physics-Neutral Prompts Counter-intuitively, our prompt ablation studies reveal that physics-enhanced prompts, those augmented with terminology like "stress," "load path," and "equilibrium", actually degraded performance on harder tasks compared to the base prompt. While the Physics-Neutral setting suffered in connectivity metrics, the failure of the Enhanced prompt suggests that models do not ground physical jargon to the visual grid. Instead, terms like "stress" likely act as semantic distractors, shifting the model's focus away from the necessary spatial pattern-matching and leading to worse topological validity (Appendix I, for details).

6.2 Qualitative Analysis of Failure Modes

To complement the quantitative metrics, we visually inspected model predictions to identify recurring patterns of reasoning failure. We observed three distinct failure modes that explain the performance gaps reported in Table 3.

The "Smearing" Effect in Continuous Tasks In hard (continuous) tasks, models frequently fail to commit to a defined structure. Instead of placing high-density material in critical load paths, they distribute low-density values ($0.1 - 0.3$) broadly across the void (Appendix J.15). This "smearing" behavior results in the negative Difference Ratios observed quantitatively; the models appear to be minimizing risk by filling space rather than optimizing for stiffness, effectively hallucinating material where none is needed.

Disconnected Islands and Local Bias A common error in lower-performing models (e.g., Perplexity Sonar, DeepSeek-R1 on hard tasks) is the generation of "floating islands", clusters of material completely disconnected from supports. This confirms: these models are operating primarily on local pattern consistency (placing a 1 next to another 1) rather than global constraint satisfaction. They fail to trace the load path $L \rightarrow S$ back to a fixed point, violating fundamental equilibrium principles.

Gravity Bias in Rotated Scenarios Qualitative inspection of the rotated experiments ($270°$) reveals a strong directional bias. 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, providing strong evidence that the models are relying on visual memorization of vertical architectural forms rather than physical reasoning.

Over-Constrained "Safety" Conversely, top-performing models like Gemini 2.5 Pro often "overbuild," 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.

7 Discussion

The quantitative and qualitative results highlight fundamental gaps between linguistic reasoning and physical-spatial understanding in Large Language Models.

Lack of Grounded Physical Understanding The failure of physics-enhanced prompts and the struggle with hard tasks suggest that current LLMs do not possess a grounded model of physics. When a model reads "load path", it does not translate this into a constraint satisfaction problem on the grid; it treats it as a textual token associated with general engineering contexts. Consequently, models perform best when the task is framed as a visual pattern completion (base prompt) rather than a physics simulation problem.

Visual Memorization vs. Force Reasoning The "gravity bias" observed in our rotation experiments confirms that models are solving SPhyR tasks primarily through visual memorization of architectural forms (e.g., columns support beams from below) rather than first-principles reasoning about force vectors. When the "floor" is moved to the "wall" (rotated setup), the model's heuristic fails, proving that it is not tracing the force L to the support S, but rather completing a learned image schema.

The Challenge of Continuous Optimization The "smearing" effect and negative Difference Ratios in continuous tasks highlight a specific deficiency in LLM spatial reasoning: the inability to perform gradient-like optimization. While models can predict discrete binary occupancy (material vs. void) based on connectivity rules, they cannot intuitively minimize compliance or volume in a continuous space. This remains a significant barrier for using LLMs in generative design and engineering applications where efficiency is paramount.

8 Conclusion

SPhyR reveals that while LLMs exhibit strong general reasoning, they fail to integrate spatial layout with grounded physical constraints. Observed failure modes (e.g., gravity bias, material smearing) confirm reliance on visual pattern matching over global force-directed reasoning, necessitating future work on geometric constraint satisfaction.

REFERENCES

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. Bendsøe 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].

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].

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].

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, Krishna Haridasan, Ahmed Omran, Nikunj Saunshi, Dara Bahri, Gaurav Mishra, Eric Chu, Toby Boyd, Brad Hekman, Aaron Parisi, Chaoyi Zhang, Kornraphop Kawintiranon, Tania Bedrax-Weiss, Oliver Wang, Ya Xu, Ollie Purkiss, Uri Mendlovic, Ilaï Deutel, Nam Nguyen, Adam Langley, Flip Korn, Lucia Rossazza, Alexandre Ramé, Sagar Waghmare, Helen Miller, Nathan Byrd, Ashrith Sheshan, Raia Hadsell Sangnie Bhardwaj, Pawel Janus, Tero Rissa, Dan Horgan, Sharon Silver, Ayzaan Wahid, Sergey Brin, Yves Raimond, Klemen Kloboves, Cindy Wang, Nitesh Bharadwaj Gundavarapu, Ilia Shumailov, Bo Wang, Mantas Pajarskas, Joe Heyward, Martin Nikoltchev, Maciej Kula, Hao Zhou, Zachary Garrett, Sushant Kafle, Sercan Arik, Ankita Goel, Mingyao Yang, Jiho Park, Koji Kojima, Parsa Mahmoudieh, Koray Kavukcuoglu, Grace Chen, Doug Fritz, Anton Bulyenov, Sudeshna Roy, Dimitris Paparas, Hadar Shemtov, Bo-Juen Chen, Robin Strudel, David Reitter, Aurko Roy, Andrey Vlasov, Changwan Ryu, Chas Leichner, Haichuan Yang, Zelda Mariet, Denis Vnukov, Tim Sohn, Amy Stuart, Wei Liang, Minmin Chen, Praynaa Rawlani, Christy Koh, J. D. Co-Reyes, Guangda Lai, Praseem Banzal, Dimitrios Vytiniotis, Jieru Mei, Mu Cai, Mohammed Badawi, Corey Fry, Ale Hartman, Daniel Zheng, Eric Jia, James Keeling, Annie Louis, Ying Chen, Efren Robles, Wei-Chih Hung, Howard Zhou, Nikita Saxena, Sonam Goenka, Olivia Ma, Zach Fisher, Mor Hazan Taege, Emily Graves, David Steiner, Yujia Li, Sarah Nguyen, Rahul Sukthankar, Joe Stanton, Ali Eslami, Gloria Shen, Berkin Akin, Alexey Guseynov,

Yiqian Zhou, Jean-Baptiste Alayrac, Armand Joulin, Efrat Farkash, Ashish Thapliyal, Stephen Roller, Noam Shazeer, Todor Davchev, Terry Koo, Hannah Forbes-Pollard, Kartik Audhkhasi, Greg Farquhar, Adi Mayrav Gilady, Maggie Song, John Aslanides, Piermaria Mendolicchio, Alicia Parrish, John Blitzer, Pramod Gupta, Xiaoen Ju, Xiaochen Yang, Puranjay Datta, Andrea Tacchetti, Sanket Vaibhav Mehta, Gregory Dibb, Shubham Gupta, Federico Piccinini, Raia Hadsell, Sujee Rajayogam, Jiepu Jiang, Patrick Griffin, Patrik Sundberg, Jamie Hayes, Alexey Frolov, Tian Xie, Adam Zhang, Kingshuk Dasgupta, Uday Kalra, Lior Shani, Klaus Macherey, Tzu-Kuo Huang, Liam MacDermed, Karthik Duddu, Paulo Zacchello, Zi Yang, Jessica Lo, Kai Hui, Matej Kastelic, Derek Gasaway, Qijun Tan, Summer Yue, Pablo Barrio, John Wieting, Weel Yang, Andrew Nystrom, Solomon Demmessie, Anselm Levskaya, Fabio Viola, Chetan Tekur, Greg Billock, George Necula, Mandar Joshi, Rylan Schaeffer, Swachhand Lokhande, Christina Sorokin, Pradeep Shenoy, Mia Chen, Mark Collier, Hongji Li, Taylor Bos, Nevan Wichers, Sun Jae Lee, Angéline Pouget, Santhosh Thangaraj, Kyriakos Axiotis, Phil Crone, Rachel Sterneck, Nikolai Chinaev, Victoria Krakovna, Oleksandr Ferludin, Ian Gemp, Stephanie Winkler, Dan Goldberg, Ivan Korotkov, Kefan Xiao, Malika Mehrotra, Sandeep Mariserla, Vihari Piratla, Terry Thurk, Khiem Pham, Hongxu Ma, Alexandre Senges, Ravi Kumar, Clemens Meyer, Ellie Talius, Nuo Wang Pierse, Ballie Sandhu, Horia Toma, Kuo Lin, Swaroop Nath, Tom Stone, Dorsa Sadigh, Nikita Gupta, Arthur Guez, Avi Singh, Matt Thomas, Tom Duerig, Yuan Gong, Richard Tanburn, Lydia Lihui Zhang, Phuong Dao, Mohamed Hammad, Sirui Xie, Shruti Rijhwani, Ben Murdoch, Duhyeon Kim, Will Thompson, Heng-Tze Cheng, Daniel Sohn, Pablo Sprechmann, Qiantong Xu, Srinivas Tadepalli, Peter Young, Ye Zhang, Hansa Srinivasan, Miranda Aperghis, Aditya Ayyar, Hen Fitoussi, Ryan Burnell, David Madras, Mike Dusenberry, Xi Xiong, Tayo Oguntebi, Ben Albrecht, Jörg Bornschein, Jovana Mitrović, Mason Dimarco, Bhargav Kanagal Shamanna, Premal Shah, Eren Sezener, Shyam Upadhyay, Dave Lacey, Craig Schiff, Sebastien Baur, Sanjay Ganapathy, Eva Schnider, Mateo Wirth, Connor Schenck, Andrey Simanovsky, Yi-Xuan Tan, Philipp Fränken, Dennis Duan, Bharath Mankalale, Nikhil Dhawan, Kevin Sequeira, Zichuan Wei, Shivanker Goel, Caglar Unlu, Yukun Zhu, Haitian Sun, Ananth Balashankar, Kurt Shuster, Megh Umekar, Mahmoud Alnahlawi, Aäron van den Oord, Kelly Chen, Yuexiang Zhai, Zihang Dai, Kuang-Huei Lee, Eric Doi, Lukas Zilka, Rohith Vallu, Disha Shrivastava, Jason Lee, Hisham Husain, Honglei Zhuang, Vincent Cohen-Addad, Jarred Barber, James Atwood, Adam Sadovsky, Quentin Wellens, Steven Hand, Arunkumar Rajendran, Aybuke Turker, C. J. Carey, Yuanzhong Xu, Hagen Soltau, Zefei Li, Xinying Song, Conglong Li, Iurii Kemaev, Sasha Brown, Andrea Burns, Viorica Patraucean, Piotr Stanczyk, Renga Aravamudhan, Mathieu Blondel, Hila Noga, Lorenzo Blanco, Will Song, Michael Isard, Mandar Sharma, Reid Hayes, Dalia El Badawy, Avery Lamp, Itay Laish, Olga Kozlova, Kelvin Chan, Sahil Singla, Srinivas Sunkara, Mayank Upadhyay, Chang Liu, Aijun Bai, Jarek Wilkiewicz, Martin Zlocha, Jeremiah Liu, Zhuowan Li, Haiguang Li, Omer Barak, Ganna Raboshchuk, Jiho Choi, Fangyu Liu, Erik Jue, Mohit Sharma, Andreea Marzoca, Robert Busa-Fekete, Anna Korsun, Andre Elisseeff, Zhe Shen, Sara Mc Carthy, Kay Lamerigts, Anahita Hosseini, Hanzhao Lin, Charlie Chen, Fan Yang, Kushal Chauhan, Mark Omernick, Dawei Jia, Karina Zainullina, Demis Hassabis, Danny Vainstein, Ehsan Amid, Xiang Zhou, Ronny Votel, Eszter Vértes, Xinjian Li, Zongwei Zhou, Angeliki Lazaridou, Brendan McMahan, Arjun Narayanan, Hubert Soyer, Sujoy Basu, Kayi Lee, Bryan Perozzi, Qin Cao, Leonard Berrada, Rahul Arya, Ke Chen, Katrina, Xu, Matthias Lochbrunner, Alex Hofer, Sahand Sharifzadeh, Renjie Wu, Sally Goldman, Pranjal Awasthi, Xuezhi Wang, Yan Wu, Claire Sha, Biao Zhang, Maciej Mikuła, Filippo Graziano, Siobhan Mcloughlin, Irene Giannoumis, Youhei Namiki, Chase Malik, Carey Radebaugh, Jamie Hall, Ramiro Leal-Cavazos, Jianmin Chen, Vikas Sindhwani, David Kao, David Greene, Jordan Griffith, Chris Welty, Ceslee Montgomery, Toshihiro Yoshino, Liangzhe Yuan, Noah Goodman, Assaf Hurwitz Michaely, Kevin Lee, K. P. Sawhney, Wei Chen, Zheng Zheng, Megan Shum, Nikolay Savinov, Etienne Pot, Alex Pak, Morteza Zadimoghaddam, Sijal Bhatnagar, Yoad Lewenberg, Blair Kutzman, Ji Liu, Lesley Katzen, Jeremy Selier, Josip Djolonga, Dmitry Lepikhin, Kelvin Xu, Jacky Liang, Jiewen Tan, Benoit Schillings, Muge Ersoy, Pete Blois, Bernd Bandemer, Abhimanyu Singh, Sergei Lebedev, Pankaj Joshi, Adam R. Brown, Evan Palmer, Shreya Pathak, Komal Jalan, Fedir Zubach, Shuba Lall, Randall Parker, Alok Gunjan, Sergey Rogulenko, Sumit Sanghai, Zhaoqi Leng, Zoltan Egyed, Shixin Li, Maria Ivanova, Kostas Andriopoulos, Jin Xie, Elan Rosenfeld, Auriel Wright, Ankur Sharma, Xinyang Geng, Yicheng Wang, Sam Kwei, Renke Pan, Yujing Zhang, Gabby Wang, Xi Liu, Chak Yeung, Elizabeth Cole, Aviv Rosenberg, Zhen Yang, Phil Chen, George Polovets, Pranav Nair, Rohun Saxena, Josh Smith, Shuo-yiin Chang, Aroma Mahendru, Svetlana Grant, Anand Iyer, Irene Cai, Jed McGiffin, Jiaming Shen, Alanna Walton, Antonious Girgis, Oliver Woodman, Rosemary Ke, Mike Kwong, Louis

Rouillard, Jinmeng Rao, Zhihao Li, Yuntao Xu, Flavien Prost, Chi Zou, Ziwei Ji, Alberto Magni, Tyler Liechty, Dan A. Calian, Deepak Ramachandran, Igor Krivokon, Hui Huang, Terry Chen, Anja Hauth, Anastasija Ilić, Weijuan Xi, Hyeontaek Lim, Vlad-Doru Ion, Pooya Moradi, Metin Toksoz-Exley, Kalesha Bullard, Miltos Allamanis, Xiaomeng Yang, Sophie Wang, Zhi Hong, Anita Gergely, Cheng Li, Bhavishya Mittal, Vitaly Kovalev, Victor Ungureanu, Jane Labanowski, Jan Wassenberg, Nicolas Lacasse, Geoffrey Cideron, Petar Dević, Annie Marsden, Lynn Nguyen, Michael Fink, Yin Zhong, Tatsuya Kiyono, Desi Ivanov, Sally Ma, Max Bain, Kiran Yalasangi, Jennifer She, Anastasia Petrushkina, Mayank Lunayach, Carla Bromberg, Sarah Hodkinson, Vilobh Meshram, Daniel Vlasic, Austin Kyker, Steve Xu, Jeff Stanway, Zuguang Yang, Kai Zhao, Matthew Tung, Seth Odoom, Yasuhisa Fujii, Justin Gilmer, Eunyoung Kim, Felix Halim, Quoc Le, Bernd Bohnet, Seliem El-Sayed, Behnam Neyshabur, Malcolm Reynolds, Dean Reich, Yang Xu, Erica Moreira, Anuj Sharma, Zeyu Liu, Mohammad Javad Hosseini, Naina Raisinghani, Yi Su, Ni Lao, Daniel Formoso, Marco Gelmi, Almog Gueta, Tapomay Dey, Elena Gribovskaya, Domagoj Ćevid, Sidharth Mudgal, Garrett Bingham, Jianling Wang, Anurag Kumar, Alex Cullum, Feng Han, Konstantinos Bousmalis, Diego Cedillo, Grace Chu, Vladimir Magay, Paul Michel, Ester Hlavnova, Daniele Calandriello, Setareh Ariafar, Kaisheng Yao, Vikash Sehwag, Arpi Vezer, Agustin Dal Lago, Zhenkai Zhu, Paul Kishan Rubenstein, Allen Porter, Anirudh Baddepudi, Oriana Riva, Mihai Dorin Istin, Chih-Kuan Yeh, Zhi Li, Andrew Howard, Nilpa Jha, Jeremy Chen, Raoul de Liedekerke, Zafarali Ahmed, Mikel Rodriguez, Tanuj Bhatia, Bangju Wang, Ali Elqursh, David Klinghoffer, Peter Chen, Pushmeet Kohli, Te I, Weiyang Zhang, Zack Nado, Jilin Chen, Maxwell Chen, George Zhang, Aayush Singh, Adam Hillier, Federico Lebron, Yiqing Tao, Ting Liu, Gabriel Dulac-Arnold, Jingwei Zhang, Shashi Narayan, Buhuang Liu, Orhan Firat, Abhishek Bhowmick, Bingyuan Liu, Hao Zhang, Zizhao Zhang, Georges Rotival, Nathan Howard, Anu Sinha, Alexander Grushetsky, Benjamin Beyret, Keerthana Gopalakrishnan, James Zhao, Kyle He, Szabolcs Payrits, Zaid Nabulsi, Zhaoyi Zhang, Weijie Chen, Edward Lee, Nova Fallen, Sreenivas Gollapudi, Aurick Zhou, Filip Pavetić, Thomas Köppe, Shiyu Huang, Rama Pasumarthi, Nick Fernando, Felix Fischer, Daria Ćurko, Yang Gao, James Svensson, Austin Stone, Haroon Qureshi, Abhishek Sinha, Apoorv Kulshreshtha, Martin Matysiak, Jieming Mao, Carl Saroufim, Aleksandra Faust, Qingnan Duan, Gil Fidel, Kaan Katircioglu, Raphaël Lopez Kaufman, Dhruv Shah, Weize Kong, Abhishek Bapna, Gellért Weisz, Emma Dunleavy, Praneet Dutta, Tianqi Liu, Rahma Chaabouni, Carolina Parada, Marcus Wu, Alexandra Belias, Alessandro Bissacco, Stanislav Fort, Li Xiao, Fantine Huot, Chris Knutsen, Yochai Blau, Gang Li, Jennifer Prendki, Juliette Love, Yinlam Chow, Pichi Charoenpanit, Hidetoshi Shimokawa, Vincent Coriou, Karol Gregor, Tomas Izo, Arjun Akula, Mario Pinto, Chris Hahn, Dominik Paulus, Jiaxian Guo, Neha Sharma, Cho-Jui Hsieh, Adaeze Chukwuka, Kazuma Hashimoto, Nathalie Rauschmayr, Ling Wu, Christof Angermueller, Yulong Wang, Sebastian Gerlach, Michael Pliskin, Daniil Mirylenka, Min Ma, Lexi Baugher, Bryan Gale, Shaan Bijwadia, Nemanja Rakićević, David Wood, Jane Park, Chung-Ching Chang, Babi Seal, Chris Tar, Kacper Krasowiak, Yiwen Song, Georgi Stephanov, Gary Wang, Marcello Maggioni, Stein Xudong Lin, Felix Wu, Shachi Paul, Zixuan Jiang, Shubham Agrawal, Bilal Piot, Alex Feng, Cheolmin Kim, Tulsee Doshi, Jonathan Lai, Chuqiao, Xu, Sharad Vikram, Ciprian Chelba, Sebastian Krause, Vincent Zhuang, Jack Rae, Timo Denk, Adrian Collister, Lotte Weerts, Xianghong Luo, Yifeng Lu, Håvard Garnes, Nitish Gupta, Terry Spitz, Avinatan Hassidim, Lihao Liang, Izhak Shafran, Peter Humphreys, Kenny Vassigh, Phil Wallis, Virat Shejwalkar, Nicolas Perez-Nieves, Rachel Hornung, Melissa Tan, Beka Westberg, Andy Ly, Richard Zhang, Brian Farris, Jongbin Park, Alec Kosik, Zeynep Cankara, Andrii Maksai, Yunhan Xu, Albin Cassirer, Sergi Caelles, Abbas Abdolmaleki, Mencher Chiang, Alex Fabrikant, Shravya Shetty, Luheng He, Mai Giménez, Hadi Hashemi, Sheena Panthaplackel, Yana Kulizhskaya, Salil Deshmukh, Daniele Pighin, Robin Alazard, Disha Jindal, Seb Noury, Pradeep Kumar S, Siyang Qin, Xerxes Dotiwalla, Stephen Spencer, Mohammad Babaeizadeh, Blake JianHang Chen, Vaibhav Mehta, Jennie Lees, Andrew Leach, Penporn Koanantakool, Ilia Akolzin, Ramona Comanescu, Junwhan Ahn, Alexey Svyatkovskiy, Basil Mustafa, David D'Ambrosio, Shiva Mohan Reddy Garlapati, Pascal Lamblin, Alekh Agarwal, Shuang Song, Pier Giuseppe Sessa, Pauline Coquinot, John Maggs, Hussain Masoom, Divya Pitta, Yaqing Wang, Patrick Morris-Suzuki, Billy Porter, Johnson Jia, Jeffrey Dudek, Raghavender R, Cosmin Paduraru, Alan Ansell, Tolga Bolukbasi, Tony Lu, Ramya Ganeshan, Zi Wang, Henry Griffiths, Rodrigo Benenson, Yifan He, James Swirhun, George Papamakarios, Aditya Chawla, Kuntal Sengupta, Yan Wang, Vedrana Milutinovic, Igor Mordatch, Zhipeng Jia, Jamie Smith, Will Ng, Shitij Nigam, Matt Young, Eugen Vušak, Blake Hechtman, Sheela Goenka, Avital Zipori, Kareem Ayoub, Ashok Popat, Trilok Acharya, Luo Yu, Dawn Bloxwich, Hugo Song, Paul Roit, Haiqiong Li, Aviel Boag, Nigamaa Nayakanti, Bilva Chandra,

Tianli Ding, Aahil Mehta, Cath Hope, Jiageng Zhang, Idan Heimlich Shtacher, Kartikeya Badola, Ryo Nakashima, Andrei Sozanschi, Iulia Comşa, Ante Žužul, Emily Caveness, Julian Odell, Matthew Watson, Dario de Cesare, Phillip Lippe, Derek Lockhart, Siddharth Verma, Huizhong Chen, Sean Sun, Lin Zhuo, Aditya Shah, Prakhar Gupta, Alex Muzio, Ning Niu, Amir Zait, Abhinav Singh, Meenu Gaba, Fan Ye, Prajit Ramachandran, Mohammad Saleh, Raluca Ada Popa, Ayush Dubey, Frederick Liu, Sara Javanmardi, Mark Epstein, Ross Hemsley, Richard Green, Nishant Ranka, Eden Cohen, Chuyuan Kelly Fu, Sanjay Ghemawat, Jed Borovik, James Martens, Anthony Chen, Pranav Shyam, André Susano Pinto, Ming-Hsuan Yang, Alexandru Ţifrea, David Du, Boqing Gong, Ayushi Agarwal, Seungyeon Kim, Christian Frank, Saloni Shah, Xiaodan Song, Zhiwei Deng, Ales Mikhalap, Kleopatra Chatziprimou, Timothy Chung, Toni Creswell, Susan Zhang, Yennie Jun, Carl Lebsack, Will Truong, Slavica Andačić, Itay Yona, Marco Fornoni, Rong Rong, Serge Toropov, Afzal Shama Soudagar, Andrew Audibert, Salah Zaiem, Zaheer Abbas, Andrei Rusu, Sahitya Potluri, Shitao Weng, Anastasios Kementsietsidis, Anton Tsitsulin, Daiyi Peng, Natalie Ha, Sanil Jain, Tejasi Latkar, Simeon Ivanov, Cory McLean, Anirudh GP, Rajesh Venkataraman, Canoee Liu, Dilip Krishnan, Joel D'sa, Roey Yogev, Paul Collins, Benjamin Lee, Lewis Ho, Carl Doersch, Gal Yona, Shawn Gao, Felipe Tiengo Ferreira, Adnan Ozturel, Hannah Muckenhirn, Ce Zheng, Gargi Balasubramaniam, Mudit Bansal, George van den Driessche, Sivan Eiger, Salem Haykal, Vedant Misra, Abhimanyu Goyal, Danilo Martins, Gary Leung, Jonas Valfridsson, Four Flynn, Will Bishop, Chenxi Pang, Yoni Halpern, Honglin Yu, Lawrence Moore, Yuvein, Zhu, Sridhar Thiagarajan, Yoel Drori, Zhisheng Xiao, Lucio Dery, Rolf Jagerman, Jing Lu, Eric Ge, Vaibhav Aggarwal, Arjun Khare, Vinh Tran, Oded Elyada, Ferran Alet, James Rubin, Ian Chou, David Tian, Libin Bai, Lawrence Chan, Lukasz Lew, Karolis Misiunas, Taylan Bilal, Aniket Ray, Sindhu Raghuram, Alex Castro-Ros, Viral Carpenter, C. J. Zheng, Michael Kilgore, Josef Broder, Emily Xue, Praveen Kallakuri, Dheeru Dua, Nancy Yuen, Steve Chien, John Schultz, Saurabh Agrawal, Reut Tsarfaty, Jingcao Hu, Ajay Kannan, Dror Marcus, Nisarg Kothari, Baochen Sun, Ben Horn, Matko Bošnjak, Ferjad Naeem, Dean Hirsch, Lewis Chiang, Boya Fang, Jie Han, Qifei Wang, Ben Hora, Antoine He, Mario Lučić, Beer Changpinyo, Anshuman Tripathi, John Youssef, Chester Kwak, Philippe Schlattner, Cat Graves, Rémi Leblond, Wenjun Zeng, Anders Andreassen, Gabriel Rasskin, Yue Song, Eddie Cao, Junhyuk Oh, Matt Hoffman, Wojtek Skut, Yichi Zhang, Jon Stritar, Xingyu Cai, Saarthak Khanna, Kathie Wang, Shriya Sharma, Christian Reisswig, Younghoon Jun, Aman Prasad, Tatiana Sholokhova, Preeti Singh, Adi Gerzi Rosenthal, Anian Ruoss, Françoise Beaufays, Sean Kirmani, Dongkai Chen, Johan Schalkwyk, Jonathan Herzig, Been Kim, Josh Jacob, Damien Vincent, Adrian N. Reyes, Ivana Balazevic, Léonard Hussenot, Jon Schneider, Parker Barnes, Luis Castro, Spandana Raj Babbula, Simon Green, Serkan Cabi, Nico Duduta, Danny Driess, Rich Galt, Noam Velan, Junjie Wang, Hongyang Jiao, Matthew Mauger, Du Phan, Miteyan Patel, Vlado Galić, Jerry Chang, Eyal Marcus, Matt Harvey, Julian Salazar, Elahe Dabir, Suraj Satishkumar Sheth, Amol Mandhane, Hanie Sedghi, Jeremiah Willcock, Amir Zandieh, Shruthi Prabhakara, Aida Amini, Antoine Miech, Victor Stone, Massimo Nicosia, Paul Niemczyk, Ying Xiao, Lucy Kim, Sławek Kwasiborski, Vikas Verma, Ada Maksutaj Oflazer, Christoph Hirnschall, Peter Sung, Lu Liu, Richard Everett, Michiel Bakker, Ágoston Weisz, Yufei Wang, Vivek Sampathkumar, Uri Shaham, Bibo Xu, Yasemin Altun, Mingqiu Wang, Takaaki Saeki, Guanjie Chen, Emanuel Taropa, Shanthal Vasanth, Sophia Austin, Lu Huang, Goran Petrovic, Qingyun Dou, Daniel Golovin, Grigory Rozhdestvenskiy, Allie Culp, Will Wu, Motoki Sano, Divya Jain, Julia Proskurnia, Sébastien Cevey, Alejandro Cruzado Ruiz, Piyush Patil, Mahdi Mirzazadeh, Eric Ni, Javier Snaider, Lijie Fan, Alexandre Fréchette, A. J. Pierigiovanni, Shariq Iqbal, Kenton Lee, Claudio Fantacci, Jinwei Xing, Lisa Wang, Alex Irpan, David Raposo, Yi Luan, Zhuoyuan Chen, Harish Ganapathy, Kevin Hui, Jiazhong Nie, Isabelle Guyon, Heming Ge, Roopali Vij, Hui Zheng, Dayeong Lee, Alfonso Castaño, Khuslen Baatarsukh, Gabriel Ibagon, Alexandra Chronopoulou, Nicholas FitzGerald, Shashank Viswanadha, Safeen Huda, Rivka Moroshko, Georgi Stoyanov, Prateek Kolhar, Alain Vaucher, Ishaan Watts, Adhi Kuncoro, Henryk Michalewski, Satish Kambala, Bat-Orgil Batsaikhan, Alek Andreev, Irina Jurenka, Maigo Le, Qihang Chen, Wael Al Jishi, Sarah Chakera, Zhe Chen, Aditya Kini, Vikas Yadav, Aditya Siddhant, Ilia Labzovsky, Balaji Lakshminarayanan, Carrie Grimes Bostock, Pankil Botadra, Ankesh Anand, Colton Bishop, Sam Conway-Rahman, Mohit Agarwal, Yani Donchev, Achintya Singhal, Félix de Chaumont Quitry, Natalia Ponomareva, Nishant Agrawal, Bin Ni, Kalpesh Krishna, Masha Samsikova, John Karro, Yilun Du, Tamara von Glehn, Caden Lu, Christopher A. Choquette-Choo, Zhen Qin, Tingnan Zhang, Sicheng Li, Divya Tyam, Swaroop Mishra, Wing Lowe, Colin Ji, Weiyi Wang, Manaal Faruqui, Ambrose Slone, Valentin Dalibard, Arunachalam Narayanaswamy, John Lambert, Pierre-Antoine Manzagol, Dan Karliner, Andrew Bolt, Ivan

Lobov, Aditya Kusupati, Chang Ye, Xuan Yang, Heiga Zen, Nelson George, Mukul Bhutani, Olivier Lacombe, Robert Riachi, Gagan Bansal, Rachel Soh, Yue Gao, Yang Yu, Adams Yu, Emily Nottage, Tania Rojas-Esponda, James Noraky, Manish Gupta, Ragha Kotikalapudi, Jichuan Chang, Sanja Deur, Dan Graur, Alex Mossin, Erin Farnese, Ricardo Figueira, Alexandre Moufarek, Austin Huang, Patrik Zochbauer, Ben Ingram, Tongzhou Chen, Zelin Wu, Adrià Puigdomènech, Leland Rechis, Da Yu, Sri Gayatri Sundara Padmanabhan, Rui Zhu, Chu-ling Ko, Andrea Banino, Samira Daruki, Aarush Selvan, Dhruva Bhaswar, Daniel Hernandez Diaz, Chen Su, Salvatore Scellato, Jennifer Brennan, Woohyun Han, Grace Chung, Priyanka Agrawal, Urvashi Khandelwal, Khe Chai Sim, Morgane Lustman, Sam Ritter, Kelvin Guu, Jiawei Xia, Prateek Jain, Emma Wang, Tyrone Hill, Mirko Rossini, Marija Kostelac, Tautvydas Misiunas, Amit Sabne, Kyuyeun Kim, Ahmet Iscen, Congchao Wang, José Leal, Ashwin Sreevatsa, Utku Evci, Manfred Warmuth, Saket Joshi, Daniel Suo, James Lottes, Garrett Honke, Brendan Jou, Stefani Karp, Jieru Hu, Himanshu Sahni, Adrien Ali Taïga, William Kong, Samrat Ghosh, Renshen Wang, Jay Pavagadhi, Natalie Axelsson, Nikolai Grigorev, Patrick Siegler, Rebecca Lin, Guohui Wang, Emilio Parisotto, Sharath Maddineni, Krishan Subudhi, Eyal Ben-David, Elena Pochernina, Orgad Keller, Thi Avrahami, Zhe Yuan, Pulkit Mehta, Jialu Liu, Sherry Yang, Wendy Kan, Katherine Lee, Tom Funkhouser, Derek Cheng, Hongzhi Shi, Archit Sharma, Joe Kelley, Matan Eyal, Yury Malkov, Corentin Tallec, Yuval Bahat, Shen Yan, Xintian, Wu, David Lindner, Chengda Wu, Avi Caciularu, Xiyang Luo, Rodolphe Jenatton, Tim Zaman, Yingying Bi, Ilya Kornakov, Ganesh Mallya, Daisuke Ikeda, Itay Karo, Anima Singh, Colin Evans, Praneeth Netrapalli, Vincent Nallatamby, Isaac Tian, Yannis Assael, Vikas Raunak, Victor Carbune, Ioana Bica, Lior Madmoni, Dee Cattle, Snchit Grover, Krishna Somandepalli, Sid Lall, Amelio Vázquez-Reina, Riccardo Patana, Jiaqi Mu, Pranav Talluri, Maggie Tran, Rajeev Aggarwal, R. J. Skerry-Ryan, Jun Xu, Mike Burrows, Xiaoyue Pan, Edouard Yvinec, Di Lu, Zhiying Zhang, Duc Dung Nguyen, Hairong Mu, Gabriel Barcik, Helen Ran, Lauren Beltrone, Krzysztof Choromanski, Dia Kharrat, Samuel Albanie, Sean Purser-haskell, David Bieber, Carrie Zhang, Jing Wang, Tom Hudson, Zhiyuan Zhang, Han Fu, Johannes Mauerer, Mohammad Hossein Bateni, A. J. Maschinot, Bing Wang, Muye Zhu, Arjun Pillai, Tobias Weyand, Shuang Liu, Oscar Akerlund, Fred Bertsch, Vittal Premachandran, Alicia Jin, Vincent Roulet, Peter de Boursac, Shubham Mittal, Ndaba Ndebele, Georgi Karadzhov, Sahra Ghalebikesabi, Ricky Liang, Allen Wu, Yale Cong, Nimesh Ghelani, Sumeet Singh, Bahar Fatemi, Warren, Chen, Charles Kwong, Alexey Kolganov, Steve Li, Richard Song, Chenkai Kuang, Sobhan Miryoosefi, Dale Webster, James Wendt, Arkadiusz Socala, Guolong Su, Artur Mendonça, Abhinav Gupta, Xiaowei Li, Tomy Tsai, Qiong, Hu, Kai Kang, Angie Chen, Sertan Girgin, Yongqin Xian, Andrew Lee, Nolan Ramsden, Leslie Baker, Madeleine Clare Elish, Varvara Krayvanova, Rishabh Joshi, Jiri Simsa, Yao-Yuan Yang, Piotr Ambroszczyk, Dipankar Ghosh, Arjun Kar, Yuan Shangguan, Yumeya Yamamori, Yaroslav Akulov, Andy Brock, Haotian Tang, Siddharth Vashishtha, Rich Munoz, Andreas Steiner, Kalyan Andra, Daniel Eppens, Qixuan Feng, Hayato Kobayashi, Sasha Goldshtein, Mona El Mahdy, Xin Wang, Jilei, Wang, Richard Killam, Tom Kwiatkowski, Kavya Kopparapu, Serena Zhan, Chao Jia, Alexei Bendebury, Sheryl Luo, Adrià Recasens, Timothy Knight, Jing Chen, Mohak Patel, YaGuang Li, Ben Withbroe, Dean Weesner, Kush Bhatia, Jie Ren, Danielle Eisenbud, Ebrahim Songhori, Yanhua Sun, Travis Choma, Tasos Kementsietsidis, Lucas Manning, Brian Roark, Wael Farhan, Jie Feng, Susheel Tatineni, James Cobon-Kerr, Yunjie Li, Lisa Anne Hendricks, Isaac Noble, Chris Breaux, Nate Kushman, Liqian Peng, Fuzhao Xue, Taylor Tobin, Jamie Rogers, Josh Lipschultz, Chris Alberti, Alexey Vlaskin, Mostafa Dehghani, Roshan Sharma, Tris Warkentin, Chen-Yu Lee, Benigno Uria, Da-Cheng Juan, Angad Chandorkar, Hila Sheftel, Ruibo Liu, Elnaz Davoodi, Borja De Balle Pigem, Kedar Dhamdhere, David Ross, Jonathan Hoech, Mahdis Mahdieh, Li Liu, Qiujia Li, Liam McCafferty, Chenxi Liu, Markus Mircea, Yunting Song, Omkar Savant, Alaa Saade, Colin Cherry, Vincent Hellendoorn, Siddharth Goyal, Paul Pucciarelli, David Vilar Torres, Zohar Yahav, Hyo Lee, Lars Lowe Sjoesund, Christo Kirov, Bo Chang, Deepanway Ghoshal, Lu Li, Gilles Baechler, Sébastien Pereira, Tara Sainath, Anudhyan Boral, Dominik Grewe, Afief Halumi, Nguyet Minh Phu, Tianxiao Shen, Marco Tulio Ribeiro, Dhriti Varma, Alex Kaskasoli, Vlad Feinberg, Navneet Potti, Jarrod Kahn, Matheus Wisniewski, Shakir Mohamed, Arnar Mar Hrafnkelsson, Bobak Shahriari, Jean-Baptiste Lespiau, Lisa Patel, Legg Yeung, Tom Paine, Lantao Mei, Alex Ramirez, Rakesh Shivanna, Li Zhong, Josh Woodward, Guilherme Tubone, Samira Khan, Heng Chen, Elizabeth Nielsen, Catalin Ionescu, Utsav Prabhu, Mingcen Gao, Qingze Wang, Sean Augenstein, Neesha Subramaniam, Jason Chang, Fotis Iliopoulos, Jiaming Luo, Myriam Khan, Weicheng Kuo, Denis Teplyashin, Florence Perot, Logan Kilpatrick, Amir Globerson, Hongkun Yu, Anfal Siddiqui, Nick Sukhanov, Arun Kandoor, Umang Gupta, Marco Andreetto, Moran Ambar, Donnie Kim, Paweł Wesołowski, Sarah Perrin,

Ben Limonchik, Wei Fan, Jim Stephan, Ian Stewart-Binks, Ryan Kappedal, Tong He, Sarah Cogan, Romina Datta, Tong Zhou, Jiayu Ye, Leandro Kieliger, Ana Ramalho, Kyle Kastner, Fabian Mentzer, Wei-Jen Ko, Arun Suggala, Tianhao Zhou, Shiraz Butt, Hana Strejček, Lior Belenki, Subhashini Venugopalan, Mingyang Ling, Evgenii Eltyshev, Yunxiao Deng, Geza Kovacs, Mukund Raghavachari, Hanjun Dai, Tal Schuster, Steven Schwarcz, Richard Nguyen, Arthur Nguyen, Gavin Buttimore, Shrestha Basu Mallick, Sudeep Gandhe, Seth Benjamin, Michal Jastrzebski, Le Yan, Sugato Basu, Chris Apps, Isabel Edkins, James Allingham, Immanuel Odisho, Tomas Kocisky, Jewel Zhao, Linting Xue, Apoorv Reddy, Chrysovalantis Anastasiou, Aviel Atias, Sam Redmond, Kieran Milan, Nicolas Heess, Herman Schmit, Allan Dafoe, Daniel Andor, Tynan Gangwani, Anca Dragan, Sheng Zhang, Ashyana Kachra, Gang Wu, Siyang Xue, Kevin Aydin, Siqi Liu, Yuxiang Zhou, Mahan Malihi, Austin Wu, Siddharth Gopal, Candice Schumann, Peter Stys, Alek Wang, Mirek Olšák, Dangyi Liu, Christian Schallhart, Yiran Mao, Demetra Brady, Hao Xu, Tomas Mery, Chawin Sitawarin, Siva Velusamy, Tom Cobley, Alex Zhai, Christian Walder, Nitzan Katz, Ganesh Jawahar, Chinmay Kulkarni, Antoine Yang, Adam Paszke, Yinan Wang, Bogdan Damoc, Zalán Borsos, Ray Smith, Jinning Li, Mansi Gupta, Andrei Kapishnikov, Sushant Prakash, Florian Luisier, Rishabh Agarwal, Will Grathwohl, Kuangyuan Chen, Kehang Han, Nikhil Mehta, Andrew Over, Shekoofeh Azizi, Lei Meng, Niccolò Dal Santo, Kelvin Zheng, Jane Shapiro, Igor Petrovski, Jeffrey Hui, Amin Ghafouri, Jasper Snoek, James Qin, Mandy Jordan, Caitlin Sikora, Jonathan Malmaud, Yuheng Kuang, Aga Świetlik, Ruoxin Sang, Chongyang Shi, Leon Li, Andrew Rosenberg, Shubin Zhao, Andy Crawford, Jan-Thorsten Peter, Yun Lei, Xavier Garcia, Long Le, Todd Wang, Julien Amelot, Dave Orr, Praneeth Kacham, Dana Alon, Gladys Tyen, Abhinav Arora, James Lyon, Alex Kurakin, Mimi Ly, Theo Guidroz, Zhipeng Yan, Rina Panigrahy, Pingmei Xu, Thais Kagohara, Yong Cheng, Eric Noland, Jinhyuk Lee, Jonathan Lee, Cathy Yip, Maria Wang, Efrat Nehoran, Alexander Bykovsky, Zhihao Shan, Ankit Bhagatwala, Chaochao Yan, Jie Tan, Guillermo Garrido, Dan Ethier, Nate Hurley, Grace Vesom, Xu Chen, Siyuan Qiao, Abhishek Nayyar, Julian Walker, Paramjit Sandhu, Mihaela Rosca, Danny Swisher, Mikhail Dektiarev, Josh Dillon, George-Cristian Muraru, Manuel Tragut, Artiom Myaskovsky, David Reid, Marko Velic, Owen Xiao, Jasmine George, Mark Brand, Jing Li, Wenhao Yu, Shane Gu, Xiang Deng, François-Xavier Aubet, Soheil Hassas Yeganeh, Fred Alcober, Celine Smith, Trevor Cohn, Kay McKinney, Michael Tschannen, Ramesh Sampath, Gowoon Cheon, Liangchen Luo, Luyang Liu, Jordi Orbay, Hui Peng, Gabriela Botea, Xiaofan Zhang, Charles Yoon, Cesar Magalhaes, Paweł Stradomski, Ian Mackinnon, Steven Hemingray, Kumaran Venkatesan, Rhys May, Jaeyoun Kim, Alex Druinsky, Jingchen Ye, Zheng Xu, Terry Huang, Jad Al Abdallah, Adil Dostmohamed, Rachana Fellinger, Tsendsuren Munkhdalai, Akanksha Maurya, Peter Garst, Yin Zhang, Maxim Krikun, Simon Bucher, Aditya Srikanth Veerubhotla, Yaxin Liu, Sheng Li, Nishesh Gupta, Jakub Adamek, Hanwen Chen, Bernett Orlando, Aleksandr Zaks, Joost van Amersfoort, Josh Camp, Hui Wan, HyunJeong Choe, Zhichun Wu, Kate Olszewska, Weiren Yu, Archita Vadali, Martin Scholz, Daniel De Freitas, Jason Lin, Amy Hua, Xin Liu, Frank Ding, Yichao Zhou, Boone Severson, Katerina Tsihlas, Samuel Yang, Tammo Spalink, Varun Yerram, Helena Pankov, Rory Blevins, Ben Vargas, Sarthak Jauhari, Matt Miecnikowski, Ming Zhang, Sandeep Kumar, Clement Farabet, Charline Le Lan, Sebastian Flennerhag, Yonatan Bitton, Ada Ma, Arthur Bražinskas, Eli Collins, Niharika Ahuja, Sneha Kudugunta, Anna Bortsova, Minh Giang, Wanzheng Zhu, Ed Chi, Scott Lundberg, Alexey Stern, Subha Puttagunta, Jing Xiong, Xiao Wu, Yash Pande, Amit Jhindal, Daniel Murphy, Jon Clark, Marc Brockschmidt, Maxine Deines, Kevin R. McKee, Dan Bahir, Jiajun Shen, Minh Truong, Daniel McDuff, Andrea Gesmundo, Edouard Rosseel, Bowen Liang, Ken Caluwaerts, Jessica Hamrick, Joseph Kready, Mary Cassin, Rishikesh Ingale, Li Lao, Scott Pollom, Yifan Ding, Wei He, Lizzetth Bellot, Joana Iljazi, Ramya Sree Boppana, Shan Han, Tara Thompson, Amr Khalifa, Anna Bulanova, Blagoj Mitrevski, Bo Pang, Emma Cooney, Tian Shi, Rey Coaguila, Tamar Yakar, Marc'aurelio Ranzato, Nikola Momchev, Chris Rawles, Zachary Charles, Young Maeng, Yuan Zhang, Rishabh Bansal, Xiaokai Zhao, Brian Albert, Yuan Yuan, Sudheendra Vijayanarasimhan, Roy Hirsch, Vinay Ramasesh, Kiran Vodrahalli, Xingyu Wang, Arushi Gupta, D. J. Strouse, Jianmo Ni, Roma Patel, Gabe Taubman, Zhouyuan Huo, Dero Gharibian, Marianne Monteiro, Hoi Lam, Shobha Vasudevan, Aditi Chaudhary, Isabela Albuquerque, Kilol Gupta, Sebastian Riedel, Chaitra Hegde, Avraham Ruderman, András György, Marcus Wainwright, Ashwin Chaugule, Burcu Karagol Ayan, Tomer Levinboim, Sam Shleifer, Yogesh Kalley, Vahab Mirrokni, Abhishek Rao, Prabakar Radhakrishnan, Jay Hartford, Jialin Wu, Zhenhai Zhu, Francesco Bertolini, Hao Xiong, Nicolas Serrano, Hamish Tomlinson, Myle Ott, Yifan Chang, Mark Graham, Jian Li, Marco Liang, Xiangzhu Long, Sebastian Borgeaud, Yanif Ahmad, Alex Grills, Diana Mincu, Martin Izzard, Yuan Liu, Jinyu Xie, Louis O'Bryan,

Sameera Ponda, Simon Tong, Michelle Liu, Dan Malkin, Khalid Salama, Yuankai Chen, Rohan Anil, Anand Rao, Rigel Swavely, Misha Bilenko, Nina Anderson, Tat Tan, Jing Xie, Xing Wu, Lijun Yu, Oriol Vinyals, Andrey Ryabtsev, Rumen Dangovski, Kate Baumli, Daniel Keysers, Christian Wright, Zoe Ashwood, Betty Chan, Artem Shtefan, Yaohui Guo, Ankur Bapna, Radu Soricut, Steven Pecht, Sabela Ramos, Rui Wang, Jiahao Cai, Trieu Trinh, Paul Barham, Linda Friso, Eli Stickgold, Xiangzhuo Ding, Siamak Shakeri, Diego Ardila, Eleftheria Briakou, Phil Culliton, Adam Raveret, Jingyu Cui, David Saxton, Subhrajit Roy, Javad Azizi, Pengcheng Yin, Lucia Loher, Andrew Bunner, Min Choi, Faruk Ahmed, Eric Li, Yin Li, Shengyang Dai, Michael Elabd, Sriram Ganapathy, Shivani Agrawal, Yiqing Hua, Paige Kunkle, Sujeevan Rajayogam, Arun Ahuja, Arthur Conmy, Alex Vasiloff, Parker Beak, Christopher Yew, Jayaram Mudigonda, Bartek Wydrowski, Jon Blanton, Zhengdong Wang, Yann Dauphin, Zhuo Xu, Martin Polacek, Xi Chen, Hexiang Hu, Pauline Sho, Markus Kunesch, Mehdi Hafezi Manshadi, Eliza Rutherford, Bo Li, Sissie Hsiao, Iain Barr, Alex Tudor, Matija Kecman, Arsha Nagrani, Vladimir Pchelin, Martin Sundermeyer, Aishwarya P. S, Abhijit Karmarkar, Yi Gao, Grishma Chole, Olivier Bachem, Isabel Gao, Arturo BC, Matt Dibb, Mauro Verzetti, Felix Hernandez-Campos, Yana Lunts, Matthew Johnson, Julia Di Trapani, Raphael Koster, Idan Brusilovsky, Binbin Xiong, Megha Mohabey, Han Ke, Joe Zou, Tea Sabolić, Víctor Campos, John Palowitch, Alex Morris, Linhai Qiu, Pranavaraj Ponnuramu, Fangtao Li, Vivek Sharma, Kiranbir Sodhia, Kaan Tekelioglu, Aleksandr Chuklin, Madhavi Yenugula, Erika Gemzer, Theofilos Strinopoulos, Sam El-Husseini, Huiyu Wang, Yan Zhong, Edouard Leurent, Paul Natsev, Weijun Wang, Dre Mahaarachchi, Tao Zhu, Songyou Peng, Sami Alabed, Cheng-Chun Lee, Anthony Brohan, Arthur Szlam, G. S. Oh, Anton Kovsharov, Jenny Lee, Renee Wong, Megan Barnes, Gregory Thornton, Felix Gimeno, Omer Levy, Martin Sevenich, Melvin Johnson, Jonathan Mallinson, Robert Dadashi, Ziyue Wang, Qingchun Ren, Preethi Lahoti, Arka Dhar, Josh Feldman, Dan Zheng, Thatcher Ulrich, Liviu Panait, Michiel Blokzijl, Cip Baetu, Josip Matak, Jitendra Harlalka, Maulik Shah, Tal Marian, Daniel von Dincklage, Cosmo Du, Ruy Ley-Wild, Bethanie Brownfield, Max Schumacher, Yury Stuken, Shadi Noghabi, Sonal Gupta, Xiaoqi Ren, Eric Malmi, Felix Weissenberger, Blanca Huergo, Maria Bauza, Thomas Lampe, Arthur Douillard, Mojtaba Seyedhosseini, Roy Frostig, Zoubin Ghahramani, Kelvin Nguyen, Kashyap Krishnakumar, Chengxi Ye, Rahul Gupta, Alireza Nazari, Robert Geirhos, Pete Shaw, Ahmed Eleryan, Dima Damen, Jennimaria Palomaki, Ted Xiao, Qiyin Wu, Quan Yuan, Phoenix Meadowlark, Matthew Bilotti, Raymond Lin, Mukund Sridhar, Yannick Schroecker, Da-Woon Chung, Jincheng Luo, Trevor Strohman, Tianlin Liu, Anne Zheng, Jesse Emond, Wei Wang, Andrew Lampinen, Toshiyuki Fukuzawa, Folawiyo Campbell-Ajala, Monica Roy, James Lee-Thorp, Lily Wang, Iftekhar Naim, Tony, Nguy\~ên, Guy Bensky, Aditya Gupta, Dominika Rogozińska, Justin Fu, Thanumalayan Sankaranarayana Pillai, Petar Veličković, Shahar Drath, Philipp Neubeck, Vaibhav Tulsyan, Arseniy Klimovskiy, Don Metzler, Sage Stevens, Angel Yeh, Junwei Yuan, Tianhe Yu, Kelvin Zhang, Alec Go, Vincent Tsang, Ying Xu, Andy Wan, Isaac Galatzer-Levy, Sam Sobell, Abodunrinwa Toki, Elizabeth Salesky, Wenlei Zhou, Diego Antognini, Sholto Douglas, Shimu Wu, Adam Lelkes, Frank Kim, Paul Cavallaro, Ana Salazar, Yuchi Liu, James Besley, Tiziana Refice, Yiling Jia, Zhang Li, Michal Sokolik, Arvind Kannan, Jon Simon, Jo Chick, Avia Aharon, Meet Gandhi, Mayank Daswani, Keyvan Amiri, Vighnesh Birodkar, Abe Ittycheriah, Peter Grabowski, Oscar Chang, Charles Sutton, Zhixin, Lai, Umesh Telang, Susie Sargsyan, Tao Jiang, Raphael Hoffmann, Nicole Brichtova, Matteo Hessel, Jonathan Halcrow, Sammy Jerome, Geoff Brown, Alex Tomala, Elena Buchatskaya, Dian Yu, Sachit Menon, Pol Moreno, Yuguo Liao, Vicky Zayats, Luming Tang, S. Q. Mah, Ashish Shenoy, Alex Siegman, Majid Hadian, Okwan Kwon, Tao Tu, Nima Khajehnouri, Ryan Foley, Parisa Haghani, Zhongru Wu, Vaishakh Keshava, Khyatti Gupta, Tony Bruguier, Rui Yao, Danny Karmon, Luisa Zintgraf, Zhicheng Wang, Enrique Piqueras, Junehyuk Jung, Jenny Brennan, Diego Machado, Marissa Giustina, M. H. Tessler, Kamyu Lee, Qiao Zhang, Joss Moore, Kaspar Daugaard, Alexander Frömmgen, Jennifer Beattie, Fred Zhang, Daniel Kasenberg, Ty Geri, Danfeng Qin, Gaurav Singh Tomar, Tom Ouyang, Tianli Yu, Luowei Zhou, Rajiv Mathews, Andy Davis, Yaoyiran Li, Jai Gupta, Damion Yates, Linda Deng, Elizabeth Kemp, Ga-Young Joung, Sergei Vassilvitskii, Mandy Guo, Pallavi LV, Dave Dopson, Sami Lachgar, Lara McConnaughey, Himadri Choudhury, Dragos Dena, Aaron Cohen, Joshua Ainslie, Sergey Levi, Parthasarathy Gopavarapu, Polina Zablotskaia, Hugo Vallet, Sanaz Bahargam, Xiaodan Tang, Nenad Tomasev, Ethan Dyer, Daniel Balle, Hongrae Lee, William Bono, Jorge Gonzalez Mendez, Vadim Zubov, Shentao Yang, Ivor Rendulic, Yanyan Zheng, Andrew Hogue, Golan Pundak, Ralph Leith, Avishkar Bhoopchand, Michael Han, Mislav Žanić, Tom Schaul, Manolis Delakis, Tejas Iyer, Guanyu Wang, Harman Singh, Abdelrahman Abdelhamed, Tara Thomas, Siddhartha Brahma, Hilal Dib, Naveen Kumar, Wenxuan Zhou, Liang

Bai, Pushkar Mishra, Jiao Sun, Valentin Anklin, Roykrong Sukkerd, Lauren Agubuzu, Anton Briukhov, Anmol Gulati, Maximilian Sieb, Fabio Pardo, Sara Nasso, Junquan Chen, Kexin Zhu, Tiberiu Sosea, Alex Goldin, Keith Rush, Spurthi Amba Hombaiah, Andreas Noever, Allan Zhou, Sam Haves, Mary Phuong, Jake Ades, Yi-ting Chen, Lin Yang, Joseph Pagadora, Stan Bileschi, Victor Cotruta, Rachel Saputro, Arijit Pramanik, Sean Ammirati, Dan Garrette, Kevin Villela, Tim Blyth, Canfer Akbulut, Neha Jha, Alban Rrustemi, Arissa Wongpanich, Chirag Nagpal, Yonghui Wu, Morgane Rivière, Sergey Kishchenko, Pranesh Srinivasan, Alice Chen, Animesh Sinha, Trang Pham, Bill Jia, Tom Hennigan, Anton Bakalov, Nithya Attaluri, Drew Garmon, Daniel Rodriguez, Dawid Wegner, Wenhao Jia, Evan Senter, Noah Fiedel, Denis Petek, Yuchuan Liu, Cassidy Hardin, Harshal Tushar Lehri, Joao Carreira, Sara Smoot, Marcel Prasetya, Nami Akazawa, Anca Stefanoiu, Chia-Hua Ho, Anelia Angelova, Kate Lin, Min Kim, Charles Chen, Marcin Sieniek, Alice Li, Tongfei Guo, Sorin Baltateanu, Pouya Tafti, Michael Wunder, Nadav Olmert, Divyansh Shukla, Jingwei Shen, Neel Kovelamudi, Balaji Venkatraman, Seth Neel, Romal Thoppilan, Jerome Connor, Frederik Benzing, Axel Stjerngren, Golnaz Ghiasi, Alex Polozov, Joshua Howland, Theophane Weber, Justin Chiu, Ganesh Poomal Girirajan, Andreas Terzis, Pidong Wang, Fangda Li, Yoav Ben Shalom, Dinesh Tewari, Matthew Denton, Roee Aharoni, Norbert Kalb, Heri Zhao, Junlin Zhang, Angelos Filos, Matthew Rahtz, Lalit Jain, Connie Fan, Vitor Rodrigues, Ruth Wang, Richard Shin, Jacob Austin, Roman Ring, Mariella Sanchez-Vargas, Mehadi Hassen, Ido Kessler, Uri Alon, Gufeng Zhang, Wenhu Chen, Yenai Ma, Xiance Si, Le Hou, Azalia Mirhoseini, Marc Wilson, Geoff Bacon, Becca Roelofs, Lei Shu, Gautam Vasudevan, Jonas Adler, Artur Dwornik, Tayfun Terzi, Matt Lawlor, Harry Askham, Mike Bernico, Xuanyi Dong, Chris Hidey, Kevin Kilgour, Gaël Liu, Surya Bhupatiraju, Luke Leonhard, Siqi Zuo, Partha Talukdar, Qing Wei, Aliaksei Severyn, Vít Listík, Jong Lee, Aditya Tripathi, S. K. Park, Yossi Matias, Hao Liu, Alex Ruiz, Rajesh Jayaram, Jackson Tolins, Pierre Marcenac, Yiming Wang, Bryan Seybold, Henry Prior, Deepak Sharma, Jack Weber, Mikhail Sirotenko, Yunhsuan Sung, Dayou Du, Ellie Pavlick, Stefan Zinke, Markus Freitag, Max Dylla, Montse Gonzalez Arenas, Natan Potikha, Omer Goldman, Connie Tao, Rachita Chhaparia, Maria Voitovich, Pawan Dogra, Andrija Ražnatović, Zak Tsai, Chong You, Oleaser Johnson, George Tucker, Chenjie Gu, Jae Yoo, Maryam Majzoubi, Valentin Gabeur, Bahram Raad, Rocky Rhodes, Kashyap Kolipaka, Heidi Howard, Geta Sampemane, Benny Li, Chulayuth Asawaroengchai, Duy Nguyen, Chiyuan Zhang, Timothee Cour, Xinxin Yu, Zhao Fu, Joe Jiang, Po-Sen Huang, Gabriela Surita, Iñaki Iturrate, Yael Karov, Michael Collins, Martin Baeuml, Fabian Fuchs, Shilpa Shetty, Swaroop Ramaswamy, Sayna Ebrahimi, Qiuchen Guo, Jeremy Shar, Gabe Barth-Maron, Sravanti Addepalli, Bryan Richter, Chin-Yi Cheng, Eugénie Rives, Fei Zheng, Johannes Griesser, Nishanth Dikkala, Yoel Zeldes, Ilkin Safarli, Dipanjan Das, Himanshu Srivastava, Sadh MNM Khan, Xin Li, Aditya Pandey, Larisa Markeeva, Dan Belov, Qiqi Yan, Mikołaj Rybiński, Tao Chen, Megha Nawhal, Michael Quinn, Vineetha Govindaraj, Sarah York, Reed Roberts, Roopal Garg, Namrata Godbole, Jake Abernethy, Anil Das, Lam Nguyen Thiet, Jonathan Tompson, John Nham, Neera Vats, Ben Caine, Wesley Helmholz, Francesco Pongetti, Yeongil Ko, James An, Clara Huiyi Hu, Yu-Cheng Ling, Julia Pawar, Robert Leland, Keisuke Kinoshita, Waleed Khawaja, Marco Selvi, Eugene Ie, Danila Sinopalnikov, Lev Proleev, Nilesh Tripuraneni, Michele Bevilacqua, Seungji Lee, Clayton Sanford, Dan Suh, Dustin Tran, Jeff Dean, Simon Baumgartner, Jens Heitkaemper, Sagar Gubbi, Kristina Toutanova, Yichong Xu, Chandu Thekkath, Keran Rong, Palak Jain, Annie Xie, Yan Virin, Yang Li, Lubo Litchev, Richard Powell, Tarun Bharti, Adam Kraft, Nan Hua, Marissa Ikonomidis, Ayal Hitron, Sanjiv Kumar, Loic Matthey, Sophie Bridgers, Lauren Lax, Ishaan Malhi, Ondrej Skopek, Ashish Gupta, Jiawei Cao, Mitchelle Rasquinha, Siim Põder, Wojciech Stokowiec, Nicholas Roth, Guowang Li, Michaël Sander, Joshua Kessinger, Vihan Jain, Edward Loper, Wonpyo Park, Michal Yarom, Liqun Cheng, Guru Guruganesh, Kanishka Rao, Yan Li, Catarina Barros, Mikhail Sushkov, Chun-Sung Ferng, Rohin Shah, Ophir Aharoni, Ravin Kumar, Tim McConnell, Peiran Li, Chen Wang, Fernando Pereira, Craig Swanson, Fayaz Jamil, Yan Xiong, Anitha Vijayakumar, Prakash Shroff, Kedar Soparkar, Jindong Gu, Livio Baldini Soares, Eric Wang, Kushal Majmundar, Aurora Wei, Kai Bailey, Nora Kassner, Chizu Kawamoto, Goran Žužić, Victor Gomes, Abhirut Gupta, Michael Guzman, Ishita Dasgupta, Xinyi Bai, Zhufeng Pan, Francesco Piccinno, Hadas Natalie Vogel, Octavio Ponce, Adrian Hutter, Paul Chang, Pan-Pan Jiang, Ionel Gog, Vlad Ionescu, James Manyika, Fabian Pedregosa, Harry Ragan, Zach Behrman, Ryan Mullins, Coline Devin, Aroonalok Pyne, Swapnil Gawde, Martin Chadwick, Yiming Gu, Sasan Tavakkol, Andy Twigg, Naman Goyal, Ndidi Elue, Anna Goldie, Srinivasan Venkatachary, Hongliang Fei, Ziqiang Feng, Marvin Ritter, Isabel Leal, Sudeep Dasari, Pei Sun, Alif Raditya Rochman, Brendan O'Donoghue, Yuchen Liu, Jim Sproch, Kai Chen, Natalie Clay, Slav Petrov, Sailesh Sidhwani, Ioana Mihailescu, Alex

Panagopoulos, A. J. Piergiovanni, Yunfei Bai, George Powell, Deep Karkhanis, Trevor Yacovone, Petr Mitrichev, Joe Kovac, Dave Uthus, Amir Yazdanbakhsh, David Amos, Steven Zheng, Bing Zhang, Jin Miao, Bhuvana Ramabhadran, Soroush Radpour, Shantanu Thakoor, Josh Newlan, Oran Lang, Orion Jankowski, Shikhar Bharadwaj, Jean-Michel Sarr, Shereen Ashraf, Sneha Mondal, Jun Yan, Ankit Singh Rawat, Sarmishta Velury, Greg Kochanski, Tom Eccles, Franz Och, Abhanshu Sharma, Ethan Mahintorabi, Alex Gurney, Carrie Muir, Vered Cohen, Saksham Thakur, Adam Bloniarz, Asier Mujika, Alexander Pritzel, Paul Caron, Altaf Rahman, Fiona Lang, Yasumasa Onoe, Petar Sirkovic, Jay Hoover, Ying Jian, Pablo Duque, Arun Narayanan, David Soergel, Alex Haig, Loren Maggiore, Shyamal Buch, Josef Dean, Ilya Figotin, Igor Karpov, Shaleen Gupta, Denny Zhou, Muhuan Huang, Ashwin Vaswani, Christopher Semturs, Kaushik Shivakumar, Yu Watanabe, Vinodh Kumar Rajendran, Eva Lu, Yanhan Hou, Wenting Ye, Shikhar Vashishth, Nana Nti, Vytenis Sakenas, Darren Ni, Doug DeCarlo, Michael Bendersky, Sumit Bagri, Nacho Cano, Elijah Peake, Simon Tokumine, Varun Godbole, Carlos Guía, Tanya Lando, Vittorio Selo, Seher Ellis, Danny Tarlow, Daniel Gillick, Alessandro Epasto, Siddhartha Reddy Jonnalagadda, Meng Wei, Meiyan Xie, Ankur Taly, Michela Paganini, Mukund Sundararajan, Daniel Toyama, Ting Yu, Dessie Petrova, Aneesh Pappu, Rohan Agrawal, Senaka Buthpitiya, Justin Frye, Thomas Buschmann, Remi Crocker, Marco Tagliasacchi, Mengchao Wang, Da Huang, Sagi Perel, Brian Wieder, Hideto Kazawa, Weiyue Wang, Jeremy Cole, Himanshu Gupta, Ben Golan, Seojin Bang, Nitish Kulkarni, Ken Franko, Casper Liu, Doug Reid, Sid Dalmia, Jay Whang, Kevin Cen, Prasha Sundaram, Johan Ferret, Berivan Isik, Lucian Ionita, Guan Sun, Anna Shekhawat, Muqthar Mohammad, Philip Pham, Ronny Huang, Karthik Raman, Xingyi Zhou, Ross Mcilroy, Austin Myers, Sheng Peng, Jacob Scott, Paul Covington, Sofia Erell, Pratik Joshi, João Gabriel Oliveira, Natasha Noy, Tajwar Nasir, Jake Walker, Vera Axelrod, Tim Dozat, Pu Han, Chun-Te Chu, Eugene Weinstein, Anand Shukla, Shreyas Chandrakaladharan, Petra Poklukar, Bonnie Li, Ye Jin, Prem Eruvbetine, Steven Hansen, Avigail Dabush, Alon Jacovi, Samrat Phatale, Chen Zhu, Steven Baker, Mo Shomrat, Yang Xiao, Jean Pouget-Abadie, Mingyang Zhang, Fanny Wei, Yang Song, Helen King, Yiling Huang, Yun Zhu, Ruoxi Sun, Juliana Vicente Franco, Chu-Cheng Lin, Sho Arora, Hui, Li, Vivian Xia, Luke Vilnis, Mariano Schain, Kaiz Alarakyia, Laurel Prince, Aaron Phillips, Caleb Habtegebriel, Luyao Xu, Huan Gui, Santiago Ontanon, Lora Aroyo, Karan Gill, Peggy Lu, Yash Katariya, Dhruv Madeka, Shankar Krishnan, Shubha Srinivas Raghvendra, James Freedman, Yi Tay, Gaurav Menghani, Peter Choy, Nishita Shetty, Dan Abolafia, Doron Kukliansky, Edward Chou, Jared Lichtarge, Ken Burke, Ben Coleman, Dee Guo, Larry Jin, Indro Bhattacharya, Victoria Langston, Yiming Li, Suyog Kotecha, Alex Yakubovich, Xinyun Chen, Petre Petrov, Tolly Powell, Yanzhang He, Corbin Quick, Kanav Garg, Dawsen Hwang, Yang Lu, Srinadh Bhojanapalli, Kristian Kjems, Ramin Mehran, Aaron Archer, Hado van Hasselt, Ashwin Balakrishna, J. K. Kearns, Meiqi Guo, Jason Riesa, Mikita Sazanovich, Xu Gao, Chris Sauer, Chengrun Yang, XiangHai Sheng, Thomas Jimma, Wouter Van Gansbeke, Vitaly Nikolaev, Wei Wei, Katie Millican, Ruizhe Zhao, Justin Snyder, Levent Bolelli, Maura O'Brien, Shawn Xu, Fei Xia, Wentao Yuan, Arvind Neelakantan, David Barker, Sachin Yadav, Hannah Kirkwood, Farooq Ahmad, Joel Wee, Jordan Grimstad, Boyu Wang, Matthew Wiethoff, Shane Settle, Miaosen Wang, Charles Blundell, Jingjing Chen, Chris Duvarney, Grace Hu, Olaf Ronneberger, Alex Lee, Yuanzhen Li, Abhishek Chakladar, Alena Butryna, Georgios Evangelopoulos, Guillaume Desjardins, Jonni Kanerva, Henry Wang, Averi Nowak, Nick Li, Alyssa Loo, Art Khurshudov, Laurent El Shafey, Nagabhushan Baddi, Karel Lenc, Yasaman Razeghi, Tom Lieber, Amer Sinha, Xiao Ma, Yao Su, James Huang, Asahi Ushio, Hanna Klimczak-Plucińska, Kareem Mohamed, J. D. Chen, Simon Osindero, Stav Ginzburg, Lampros Lamprou, Vasilisa Bashlovkina, Duc-Hieu Tran, Ali Khodaei, Ankit Anand, Yixian Di, Ramy Eskander, Manish Reddy Vuyyuru, Jasmine Liu, Aishwarya Kamath, Roman Goldenberg, Mathias Bellaiche, Juliette Pluto, Bill Rosgen, Hassan Mansoor, William Wong, Suhas Ganesh, Eric Bailey, Scott Baird, Dan Deutsch, Jinoo Baek, Xuhui Jia, Chansoo Lee, Abe Friesen, Nathaniel Braun, Kate Lee, Amayika Panda, Steven M. Hernandez, Duncan Williams, Jianqiao Liu, Ethan Liang, Arnaud Autef, Emily Pitler, Deepali Jain, Phoebe Kirk, Oskar Bunyan, Jaume Sanchez Elias, Tongxin Yin, Machel Reid, Aedan Pope, Nikita Putikhin, Bidisha Samanta, Sergio Guadarrama, Dahun Kim, Simon Rowe, Marcella Valentine, Geng Yan, Alex Salcianu, David Silver, Gan Song, Richa Singh, Shuai Ye, Hannah DeBalsi, Majd Al Merey, Eran Ofek, Albert Webson, Shibl Mourad, Ashwin Kakarla, Silvio Lattanzi, Nick Roy, Evgeny Sluzhaev, Christina Butterfield, Alessio Tonioni, Nathan Waters, Sudhindra Kopalle, Jason Chase, James Cohan, Girish Ramchandra Rao, Robert Berry, Michael Voznesensky, Shuguang Hu, Kristen Chiafullo, Sharat Chikkerur, George Scrivener, Ivy Zheng, Jeremy Wiesner, Wolfgang Macherey, Timothy Lillicrap, Fei Liu, Brian Walker, David Welling, Elinor Davies, Yangsibo Huang, Lijie Ren, Nir Shabat, Alessandro Agostini, Mariko

Iinuma, Dustin Zelle, Rohit Sathyanarayana, Andrea D'olimpio, Morgan Redshaw, Matt Ginsberg, Ashwin Murthy, Mark Geller, Tatiana Matejovicova, Ayan Chakrabarti, Ryan Julian, Christine Chan, Qiong Hu, Daniel Jarrett, Manu Agarwal, Jeshwanth Challagundla, Tao Li, Sandeep Tata, Wen Ding, Maya Meng, Zhuyun Dai, Giulia Vezzani, Shefali Garg, Jannis Bulian, Mary Jasarevic, Honglong Cai, Harish Rajamani, Adam Santoro, Florian Hartmann, Chen Liang, Bartek Perz, Apoorv Jindal, Fan Bu, Sungyong Seo, Ryan Poplin, Adrian Goedeckemeyer, Badih Ghazi, Nikhil Khadke, Leon Liu, Kevin Mather, Mingda Zhang, Ali Shah, Alex Chen, Jinliang Wei, Keshav Shivam, Yuan Cao, Donghyun Cho, Angelo Scorza Scarpati, Michael Moffitt, Clara Barbu, Ivan Jurin, Ming-Wei Chang, Hongbin Liu, Hao Zheng, Shachi Dave, Christine Kaeser-Chen, Xiaobin Yu, Alvin Abdagic, Lucas Gonzalez, Yanping Huang, Peilin Zhong, Cordelia Schmid, Bryce Petrini, Alex Wertheim, Jifan Zhu, Hoang Nguyen, Kaiyang Ji, Yanqi Zhou, Tao Zhou, Fangxiaoyu Feng, Regev Cohen, David Rim, Shubham Milind Phal, Petko Georgiev, Ariel Brand, Yue Ma, Wei Li, Somit Gupta, Chao Wang, Pavel Dubov, Jean Tarbouriech, Kingshuk Majumder, Huijian Li, Norman Rink, Apurv Suman, Yang Guo, Yinghao Sun, Arun Nair, Xiaowei Xu, Mohamed Elhawaty, Rodrigo Cabrera, Guangxing Han, Julian Eisenschlos, Junwen Bai, Yuqi Li, Yamini Bansal, Thibault Sellam, Mina Khan, Hung Nguyen, Justin Mao-Jones, Nikos Parotsidis, Jake Marcus, Cindy Fan, Roland Zimmermann, Yony Kochinski, Laura Graesser, Feryal Behbahani, Alvaro Caceres, Michael Riley, Patrick Kane, Sandra Lefdal, Rob Willoughby, Paul Vicol, Lun Wang, Shujian Zhang, Ashleah Gill, Yu Liang, Gautam Prasad, Soroosh Mariooryad, Mehran Kazemi, Zifeng Wang, Kritika Muralidharan, Paul Voigtlaender, Jeffrey Zhao, Huanjie Zhou, Nina D'Souza, Aditi Mavalankar, Séb Arnold, Nick Young, Obaid Sarvana, Chace Lee, Milad Nasr, Tingting Zou, Seokhwan Kim, Lukas Haas, Kaushal Patel, Neslihan Bulut, David Parkinson, Courtney Biles, Dmitry Kalashnikov, Chi Ming To, Aviral Kumar, Jessica Austin, Alex Greve, Lei Zhang, Megha Goel, Yeqing Li, Sergey Yaroshenko, Max Chang, Abhishek Jindal, Geoff Clark, Hagai Taitelbaum, Dale Johnson, Ofir Roval, Jeongwoo Ko, Anhad Mohananey, Christian Schuler, Shenil Dodhia, Ruichao Li, Kazuki Osawa, Claire Cui, Peng Xu, Rushin Shah, Tao Huang, Ela Gruzewska, Nathan Clement, Mudit Verma, Olcan Sercinoglu, Hai Qian, Viral Shah, Masa Yamaguchi, Abhinit Modi, Takahiro Kosakai, Thomas Strohmann, Junhao Zeng, Beliz Gunel, Jun Qian, Austin Tarango, Krzysztof Jastrzębski, Robert David, Jyn Shan, Parker Schuh, Kunal Lad, Willi Gierke, Mukundan Madhavan, Xinyi Chen, Mark Kurzeja, Rebeca Santamaria-Fernandez, Dawn Chen, Alexandra Cordell, Yuri Chervonyi, Frankie Garcia, Nithish Kannen, Vincent Perot, Nan Ding, Shlomi Cohen-Ganor, Victor Lavrenko, Junru Wu, Georgie Evans, Cicero Nogueira dos Santos, Madhavi Sewak, Ashley Brown, Andrew Hard, Joan Puigcerver, Zeyu Zheng, Yizhong Liang, Evgeny Gladchenko, Reeve Ingle, Uri First, Pierre Sermanet, Charlotte Magister, Mihajlo Velimirović, Sashank Reddi, Susanna Ricco, Eirikur Agustsson, Hartwig Adam, Nir Levine, David Gaddy, Dan Holtmann-Rice, Xuanhui Wang, Ashutosh Sathe, Abhijit Guha Roy, Blaž Bratanič, Alen Carin, Harsh Mehta, Silvano Bonacina, Nicola De Cao, Mara Finkelstein, Verena Rieser, Xinyi Wu, Florent Altché, Dylan Scandinaro, Li Li, Nino Vieillard, Nikhil Sethi, Garrett Tanzer, Zhi Xing, Shibo Wang, Parul Bhatia, Gui Citovsky, Thomas Anthony, Sharon Lin, Tianze Shi, Shoshana Jakobovits, Gena Gibson, Raj Apte, Lisa Lee, Mingqing Chen, Arunkumar Byravan, Petros Maniatis, Kellie Webster, Andrew Dai, Pu-Chin Chen, Jiaqi Pan, Asya Fadeeva, Zach Gleicher, Thang Luong, and Niket Kumar Bhumihar. Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities, July 2025. URL http://arxiv.org/abs/2507.06261. arXiv:2507.06261 [cs].

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, 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, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng

Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning, January 2025. URL http://arxiv.org/abs/2501.12948. arXiv:2501.12948 [cs].

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].

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].

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

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris

Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, C. J. Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. GPT-4 Technical Report, March 2024a. URL http://arxiv.org/abs/2303.08774. arXiv:2303.08774 [cs].

OpenAI, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, A. J. Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoochian, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, Dane Sherburn, Daniel Kappler, Daniel Levin, Daniel Levy, David Carr, David Farhi, David Mely, David Robinson, David Sasaki, Denny Jin, Dev Valladares, Dimitris Tsipras, Doug Li, Duc Phong Nguyen, Duncan Findlay, Edede Oiwoh, Edmund Wong, Ehsan Asdar, Elizabeth Proehl, Elizabeth Yang, Eric Antonow, Eric Kramer, Eric Peterson, Eric Sigler, Eric Wallace, Eugene Brevdo, Evan Mays, Farzad Khorasani, Felipe Petroski Such, Filippo Raso, Francis Zhang, Fred von Lohmann, Freddie Sulit, Gabriel Goh, Gene Oden, Geoff Salmon, Giulio Starace, Greg Brockman, Hadi Salman, Haiming Bao, Haitang Hu, Hannah Wong, Haoyu Wang, Heather Schmidt, Heather Whitney, Heewoo Jun, Hendrik Kirchner, Henrique Ponde de Oliveira Pinto, Hongyu Ren, Huiwen Chang, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian O'Connell, Ian Osband, Ian Silber, Ian Sohl, Ibrahim Okuyucu, Ikai Lan, Ilya Kostrikov, Ilya Sutskever, Ingmar Kanitscheider, Ishaan Gulrajani, Jacob Coxon, Jacob Menick, Jakub Pachocki, James Aung, James Betker, James Crooks, James Lennon, Jamie Kiros, Jan Leike, Jane Park, Jason Kwon, Jason Phang, Jason Teplitz, Jason Wei, Jason Wolfe, Jay Chen, Jeff Harris, Jenia Varavva, Jessica Gan Lee, Jessica Shieh, Ji Lin, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joanne Jang, Joaquin Quinonero Candela, Joe Beutler, Joe Landers, Joel Parish, Johannes Heidecke, John Schulman, Jonathan Lachman, Jonathan McKay, Jonathan Uesato, Jonathan Ward, Jong Wook Kim, Joost Huizinga, Jordan Sitkin, Jos Kraaijeveld, Josh Gross, Josh Kaplan, Josh Snyder, Joshua Achiam, Joy Jiao, Joyce Lee, Juntang Zhuang, Justyn Harriman, Kai Fricke, Kai Hayashi, Karan Singhal, Katy Shi, Kavin Karthik, Kayla Wood, Kendra Rimbach, Kenny Hsu, Kenny Nguyen, Keren Gu-Lemberg, Kevin Button, Kevin Liu, Kiel Howe, Krithika Muthukumar, Kyle Luther, Lama Ahmad, Larry Kai, Lauren Itow, Lauren Workman, Leher Pathak, Leo Chen, Li Jing, Lia Guy, Liam Fedus, Liang Zhou, Lien Mamitsuka, Lilian Weng, Lindsay McCallum, Lindsey Held, Long Ouyang, Louis Feuvrier, Lu Zhang, Lukas Kondraciuk, Lukasz Kaiser, Luke Hewitt, Luke Metz, Lyric Doshi, Mada Aflak, Maddie Simens, Madelaine Boyd, Madeleine Thompson, Marat Dukhan, Mark Chen, Mark Gray, Mark Hudnall, Marvin Zhang, Marwan Aljubeh, Mateusz Litwin, Matthew Zeng, Max Johnson, Maya Shetty, Mayank Gupta, Meghan Shah, Mehmet Yatbaz, Meng Jia Yang, Mengchao Zhong, Mia Glaese, Mianna Chen,

Michael Janner, Michael Lampe, Michael Petrov, Michael Wu, Michele Wang, Michelle Fradin, Michelle Pokrass, Miguel Castro, Miguel Oom Temudo de Castro, Mikhail Pavlov, Miles Brundage, Miles Wang, Minal Khan, Mira Murati, Mo Bavarian, Molly Lin, Murat Yesildal, Nacho Soto, Natalia Gimelshein, Natalie Cone, Natalie Staudacher, Natalie Summers, Natan LaFontaine, Neil Chowdhury, Nick Ryder, Nick Stathas, Nick Turley, Nik Tezak, Niko Felix, Nithanth Kudige, Nitish Keskar, Noah Deutsch, Noel Bundick, Nora Puckett, Ofir Nachum, Ola Okelola, Oleg Boiko, Oleg Murk, Oliver Jaffe, Olivia Watkins, Olivier Godement, Owen Campbell-Moore, Patrick Chao, Paul McMillan, Pavel Belov, Peng Su, Peter Bak, Peter Bakkum, Peter Deng, Peter Dolan, Peter Hoeschele, Peter Welinder, Phil Tillet, Philip Pronin, Philippe Tillet, Prafulla Dhariwal, Qiming Yuan, Rachel Dias, Rachel Lim, Rahul Arora, Rajan Troll, Randall Lin, Rapha Gontijo Lopes, Raul Puri, Reah Miyara, Reimar Leike, Renaud Gaubert, Reza Zamani, Ricky Wang, Rob Donnelly, Rob Honsby, Rocky Smith, Rohan Sahai, Rohit Ramchandani, Romain Huet, Rory Carmichael, Rowan Zellers, Roy Chen, Ruby Chen, Ruslan Nigmatullin, Ryan Cheu, Saachi Jain, Sam Altman, Sam Schoenholz, Sam Toizer, Samuel Miserendino, Sandhini Agarwal, Sara Culver, Scott Ethersmith, Scott Gray, Sean Grove, Sean Metzger, Shamez Hermani, Shantanu Jain, Shengjia Zhao, Sherwin Wu, Shino Jomoto, Shirong Wu, Shuaiqi, Xia, Sonia Phene, Spencer Papay, Srinivas Narayanan, Steve Coffey, Steve Lee, Stewart Hall, Suchir Balaji, Tal Broda, Tal Stramer, Tao Xu, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Cunninghman, Thomas Degry, Thomas Dimson, Thomas Raoux, Thomas Shadwell, Tianhao Zheng, Todd Underwood, Todor Markov, Toki Sherbakov, Tom Rubin, Tom Stasi, Tomer Kaftan, Tristan Heywood, Troy Peterson, Tyce Walters, Tyna Eloundou, Valerie Qi, Veit Moeller, Vinnie Monaco, Vishal Kuo, Vlad Fomenko, Wayne Chang, Weiyi Zheng, Wenda Zhou, Wesam Manassra, Will Sheu, Wojciech Zaremba, Yash Patil, Yilei Qian, Yongjik Kim, Youlong Cheng, Yu Zhang, Yuchen He, Yuchen Zhang, Yujia Jin, Yunxing Dai, and Yury Malkov. 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, Véronique 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].

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, Soroosh Mariooryad, Yifan Ding, Xinyang Geng, Fred Alcober, Roy Frostig, Mark Omernick, Lexi Walker, Cosmin Paduraru, Christina Sorokin, Andrea Tacchetti, Colin Gaffney, Samira Daruki, Olcan Sercinoglu, Zach Gleicher, Juliette Love, Paul Voigtlaender, Rohan Jain, Gabriela Surita, Kareem Mohamed, Rory Blevins, Junwhan Ahn, Tao Zhu, Kornraphop Kawintiranon, Orhan Firat, Yiming Gu, Yujing Zhang, Matthew Rahtz, Manaal Faruqui, Natalie Clay, Justin Gilmer, J. D. Co-Reyes, Ivo Penchev, Rui Zhu, Nobuyuki Morioka, Kevin Hui, Krishna Haridasan, Victor Campos, Mahdis Mahdieh, Mandy Guo, Samer Hassan, Kevin Kilgour, Arpi Vezer, Heng-Tze Cheng, Raoul de Liedekerke, Siddharth Goyal, Paul Barham, D. J. Strouse, Seb Noury, Jonas Adler, Mukund Sundararajan, Sharad Vikram, Dmitry Lepikhin, Michela Paganini, Xavier Garcia, Fan Yang, Dasha Valter, Maja Trebacz, Kiran Vodrahalli, Chulayuth Asawaroengchai, Roman Ring, Norbert Kalb, Livio Baldini Soares, Siddhartha Brahma, David Steiner, Tianhe Yu, Fabian Mentzer, Antoine He, Lucas Gonzalez, Bibo Xu, Raphael Lopez Kaufman, Laurent El Shafey, Junhyuk Oh, Tom Hennigan, George van den Driessche, Seth Odoom, Mario Lucic, Becca Roelofs, Sid Lall, Amit Marathe, Betty Chan, Santiago Ontanon, Luheng He, Denis Teplyashin, Jonathan Lai, Phil Crone, Bogdan Damoc, Lewis Ho, Sebastian Riedel, Karel Lenc, Chih-Kuan Yeh, Aakanksha Chowdhery, Yang Xu, Mehran Kazemi, Ehsan Amid, Anastasia Petrushkina, Kevin Swersky, Ali Khodaei, Gowoon Chen, Chris Larkin, Mario Pinto, Geng Yan, Adria Puigdomenech Badia, Piyush Patil, Steven Hansen, Dave Orr, Sebastien M. R. Arnold, Jordan Grimstad, Andrew Dai, Sholto Douglas, Rishika Sinha, Vikas Yadav, Xi Chen, Elena Gribovskaya, Jacob Austin, Jeffrey Zhao, Kaushal Patel, Paul Komarek, Sophia Austin, Sebastian Borgeaud, Linda Friso, Abhimanyu Goyal, Ben Caine, Kris Cao, Da-Woon Chung, Matthew Lamm, Gabe Barth-Maron, Thais Kagohara, Kate Olszewska, Mia Chen, Kaushik Shivakumar, Rishabh Agarwal, Harshal Godhia, Ravi Rajwar, Javier Snaider, Xerxes Dotiwalla, Yuan Liu, Aditya Barua, Victor Ungureanu, Yuan Zhang, Bat-Orgil Batsaikhan, Mateo Wirth, James Qin, Ivo

Danihelka, Tulsee Doshi, Martin Chadwick, Jilin Chen, Sanil Jain, Quoc Le, Arjun Kar, Madhu Gurumurthy, Cheng Li, Ruoxin Sang, Fangyu Liu, Lampros Lamprou, Rich Munoz, Nathan Lintz, Harsh Mehta, Heidi Howard, Malcolm Reynolds, Lora Aroyo, Quan Wang, Lorenzo Blanco, Albin Cassirer, Jordan Griffith, Dipanjan Das, Stephan Lee, Jakub Sygnowski, Zach Fisher, James Besley, Richard Powell, Zafarali Ahmed, Dominik Paulus, David Reitter, Zalan Borsos, Rishabh Joshi, Aedan Pope, Steven Hand, Vittorio Selo, Vihan Jain, Nikhil Sethi, Megha Goel, Takaki Makino, Rhys May, Zhen Yang, Johan Schalkwyk, Christina Butterfield, Anja Hauth, Alex Goldin, Will Hawkins, Evan Senter, Sergey Brin, Oliver Woodman, Marvin Ritter, Eric Noland, Minh Giang, Vijay Bolina, Lisa Lee, Tim Blyth, Ian Mackinnon, Machel Reid, Obaid Sarvana, David Silver, Alexander Chen, Lily Wang, Loren Maggiore, Oscar Chang, Nithya Attaluri, Gregory Thornton, Chung-Cheng Chiu, Oskar Bunyan, Nir Levine, Timothy Chung, Evgenii Eltyshev, Xiance Si, Timothy Lillicrap, Demetra Brady, Vaibhav Aggarwal, Boxi Wu, Yuanzhong Xu, Ross McIlroy, Kartikeya Badola, Paramjit Sandhu, Erica Moreira, Wojciech Stokowiec, Ross Hemsley, Dong Li, Alex Tudor, Pranav Shyam, Elahe Rahimtoroghi, Salem Haykal, Pablo Sprechmann, Xiang Zhou, Diana Mincu, Yujia Li, Ravi Addanki, Kalpesh Krishna, Xiao Wu, Alexandre Frechette, Matan Eyal, Allan Dafoe, Dave Lacey, Jay Whang, Thi Avrahami, Ye Zhang, Emanuel Taropa, Hanzhao Lin, Daniel Toyama, Eliza Rutherford, Motoki Sano, HyunJeong Choe, Alex Tomala, Chalence Safranek-Shrader, Nora Kassner, Mantas Pajarskas, Matt Harvey, Sean Sechrist, Meire Fortunato, Christina Lyu, Gamaleldin Elsayed, Chenkai Kuang, James Lottes, Eric Chu, Chao Jia, Chih-Wei Chen, Peter Humphreys, Kate Baumli, Connie Tao, Rajkumar Samuel, Cicero Nogueira dos Santos, Anders Andreassen, Nemanja Rakićević, Dominik Grewe, Aviral Kumar, Stephanie Winkler, Jonathan Caton, Andrew Brock, Sid Dalmia, Hannah Sheahan, Iain Barr, Yingjie Miao, Paul Natsev, Jacob Devlin, Feryal Behbahani, Flavien Prost, Yanhua Sun, Artiom Myaskovsky, Thanumalayan Sankaranarayana Pillai, Dan Hurt, Angeliki Lazaridou, Xi Xiong, Ce Zheng, Fabio Pardo, Xiaowei Li, Dan Horgan, Joe Stanton, Moran Ambar, Fei Xia, Alejandro Lince, Mingqiu Wang, Basil Mustafa, Albert Webson, Hyo Lee, Rohan Anil, Martin Wicke, Timothy Dozat, Abhishek Sinha, Enrique Piqueras, Elahe Dabir, Shyam Upadhyay, Anudhyan Boral, Lisa Anne Hendricks, Corey Fry, Josip Djolonga, Yi Su, Jake Walker, Jane Labanowski, Ronny Huang, Vedant Misra, Jeremy Chen, R. J. Skerry-Ryan, Avi Singh, Shruti Rijhwani, Dian Yu, Alex Castro-Ros, Beer Changpinyo, Romina Datta, Sumit Bagri, Arnar Mar Hrafnkelsson, Marcello Maggioni, Daniel Zheng, Yury Sulsky, Shaobo Hou, Tom Le Paine, Antoine Yang, Jason Riesa, Dominika Rogozinska, Dror Marcus, Dalia El Badawy, Qiao Zhang, Luyu Wang, Helen Miller, Jeremy Greer, Lars Lowe Sjos, Azade Nova, Heiga Zen, Rahma Chaabouni, Mihaela Rosca, Jiepu Jiang, Charlie Chen, Ruibo Liu, Tara Sainath, Maxim Krikun, Alex Polozov, Jean-Baptiste Lespiau, Josh Newlan, Zeyncep Cankara, Soo Kwak, Yunhan Xu, Phil Chen, Andy Coenen, Clemens Meyer, Katerina Tsihlas, Ada Ma, Juraj Gottweis, Jinwei Xing, Chenjie Gu, Jin Miao, Christian Frank, Zeynep Cankara, Sanjay Ganapathy, Ishita Dasgupta, Steph Hughes-Fitt, Heng Chen, David Reid, Keran Rong, Hongmin Fan, Joost van Amersfoort, Vincent Zhuang, Aaron Cohen, Shixiang Shane Gu, Anhad Mohananey, Anastasija Ilic, Taylor Tobin, John Wieting, Anna Bortsova, Phoebe Thacker, Emma Wang, Emily Caveness, Justin Chiu, Eren Sezener, Alex Kaskasoli, Steven Baker, Katie Millican, Mohamed Elhawaty, Kostas Aisopos, Carl Lebsack, Nathan Byrd, Hanjun Dai, Wenhao Jia, Matthew Wiethoff, Elnaz Davoodi, Albert Weston, Lakshman Yagati, Arun Ahuja, Isabel Gao, Golan Pundak, Susan Zhang, Michael Azzam, Khe Chai Sim, Sergi Caelles, James Keeling, Abhanshu Sharma, Andy Swing, YaGuang Li, Chenxi Liu, Carrie Grimes Bostock, Yamini Bansal, Zachary Nado, Ankesh Anand, Josh Lipschultz, Abhijit Karmarkar, Lev Proleev, Abe Ittycheriah, Soheil Hassas Yeganeh, George Polovets, Aleksandra Faust, Jiao Sun, Alban Rrustemi, Pen Li, Rakesh Shivanna, Jeremiah Liu, Chris Welty, Federico Lebron, Anirudh Baddepudi, Sebastian Krause, Emilio Parisotto, Radu Soricut, Zheng Xu, Dawn Bloxwich, Melvin Johnson, Behnam Neyshabur, Justin Mao-Jones, Renshen Wang, Vinay Ramasesh, Zaheer Abbas, Arthur Guez, Constant Segal, Duc Dung Nguyen, James Svensson, Le Hou, Sarah York, Kieran Milan, Sophie Bridgers, Wiktor Gworek, Marco Tagliasacchi, James Lee-Thorp, Michael Chang, Alexey Guseynov, Ale Jakse Hartman, Michael Kwong, Ruizhe Zhao, Sheleem Kashem, Elizabeth Cole, Antoine Miech, Richard Tanburn, Mary Phuong, Filip Pavetic, Sebastien Cevey, Ramona Comanescu, Richard Ives, Sherry Yang, Cosmo Du, Bo Li, Zizhao Zhang, Mariko Iinuma, Clara Huiyi Hu, Aurko Roy, Shaan Bijwadia, Zhenkai Zhu, Danilo Martins, Rachel Saputro, Anita Gergely, Steven Zheng, Dawei Jia, Ioannis Antonoglou, Adam Sadovsky, Shane Gu, Yingying Bi, Alek Andreev, Sina Samangooei, Mina Khan, Tomas Kocisky, Angelos Filos, Chintu Kumar, Colton Bishop, Adams Yu, Sarah Hodkinson, Sid Mittal, Premal Shah, Alexandre Moufarek, Yong Cheng, Adam Bloniarz, Jaehoon Lee, Pedram Pejman, Paul Michel, Stephen Spencer, Vladimir

Feinberg, Xuehan Xiong, Nikolay Savinov, Charlotte Smith, Siamak Shakeri, Dustin Tran, Mary Chesus, Bernd Bohnet, George Tucker, Tamara von Glehn, Carrie Muir, Yiran Mao, Hideto Kazawa, Ambrose Slone, Kedar Soparkar, Disha Shrivastava, James Cobon-Kerr, Michael Sharman, Jay Pavagadhi, Carlos Araya, Karolis Misiunas, Nimesh Ghelani, Michael Laskin, David Barker, Qiujia Li, Anton Briukhov, Neil Houlsby, Mia Glaese, Balaji Lakshminarayanan, Nathan Schucher, Yunhao Tang, Eli Collins, Hyeontaek Lim, Fangxiaoyu Feng, Adria Recasens, Guangda Lai, Alberto Magni, Nicola De Cao, Aditya Siddhant, Zoe Ashwood, Jordi Orbay, Mostafa Dehghani, Jenny Brennan, Yifan He, Kelvin Xu, Yang Gao, Carl Saroufim, James Molloy, Xinyi Wu, Seb Arnold, Solomon Chang, Julian Schrittwieser, Elena Buchatskaya, Soroush Radpour, Martin Polacek, Skye Giordano, Ankur Bapna, Simon Tokumine, Vincent Hellendoorn, Thibault Sottiaux, Sarah Cogan, Aliaksei Severyn, Mohammad Saleh, Shantanu Thakoor, Laurent Shefey, Siyuan Qiao, Meenu Gaba, Shuo-yiin Chang, Craig Swanson, Biao Zhang, Benjamin Lee, Paul Kishan Rubenstein, Gan Song, Tom Kwiatkowski, Anna Koop, Ajay Kannan, David Kao, Parker Schuh, Axel Stjerngren, Golnaz Ghiasi, Gena Gibson, Luke Vilnis, Ye Yuan, Felipe Tiengo Ferreira, Aishwarya Kamath, Ted Klimenko, Ken Franko, Kefan Xiao, Indro Bhattacharya, Miteyan Patel, Rui Wang, Alex Morris, Robin Strudel, Vivek Sharma, Peter Choy, Sayed Hadi Hashemi, Jessica Landon, Mara Finkelstein, Priya Jhakra, Justin Frye, Megan Barnes, Matthew Mauger, Dennis Daun, Khuslen Baatarsukh, Matthew Tung, Wael Farhan, Henryk Michalewski, Fabio Viola, Felix de Chaumont Quitry, Charline Le Lan, Tom Hudson, Qingze Wang, Felix Fischer, Ivy Zheng, Elspeth White, Anca Dragan, Jean-baptiste Alayrac, Eric Ni, Alexander Pritzel, Adam Iwanicki, Michael Isard, Anna Bulanova, Lukas Zilka, Ethan Dyer, Devendra Sachan, Srivatsan Srinivasan, Hannah Muckenhirn, Honglong Cai, Amol Mandhane, Mukarram Tariq, Jack W. Rae, Gary Wang, Kareem Ayoub, Nicholas FitzGerald, Yao Zhao, Woohyun Han, Chris Alberti, Dan Garrette, Kashyap Krishnakumar, Mai Gimenez, Anselm Levskaya, Daniel Sohn, Josip Matak, Inaki Iturrate, Michael B. Chang, Jackie Xiang, Yuan Cao, Nishant Ranka, Geoff Brown, Adrian Hutter, Vahab Mirrokni, Nanxin Chen, Kaisheng Yao, Zoltan Egyed, Francois Galilee, Tyler Liechty, Praveen Kallakuri, Evan Palmer, Sanjay Ghemawat, Jasmine Liu, David Tao, Chloe Thornton, Tim Green, Mimi Jasarevic, Sharon Lin, Victor Cotruta, Yi-Xuan Tan, Noah Fiedel, Hongkun Yu, Ed Chi, Alexander Neitz, Jens Heitkaemper, Anu Sinha, Denny Zhou, Yi Sun, Charbel Kaed, Brice Hulse, Swaroop Mishra, Maria Georgaki, Sneha Kudugunta, Clement Farabet, Izhak Shafran, Daniel Vlasic, Anton Tsitsulin, Rajagopal Ananthanarayanan, Alen Carin, Guolong Su, Pei Sun, Shashank V, Gabriel Carvajal, Josef Broder, Iulia Comsa, Alena Repina, William Wong, Warren Weilun Chen, Peter Hawkins, Egor Filonov, Lucia Loher, Christoph Hirnschall, Weiyi Wang, Jingchen Ye, Andrea Burns, Hardie Cate, Diana Gage Wright, Federico Piccinini, Lei Zhang, Chu-Cheng Lin, Ionel Gog, Yana Kulizhskaya, Ashwin Sreevatsa, Shuang Song, Luis C. Cobo, Anand Iyer, Chetan Tekur, Guillermo Garrido, Zhuyun Xiao, Rupert Kemp, Huaixiu Steven Zheng, Hui Li, Ananth Agarwal, Christel Ngani, Kati Goshvadi, Rebeca Santamaria-Fernandez, Wojciech Fica, Xinyun Chen, Chris Gorgolewski, Sean Sun, Roopal Garg, Xinyu Ye, S. M. Ali Eslami, Nan Hua, Jon Simon, Pratik Joshi, Yelin Kim, Ian Tenney, Sahitya Potluri, Lam Nguyen Thiet, Quan Yuan, Florian Luisier, Alexandra Chronopoulou, Salvatore Scellato, Praveen Srinivasan, Minmin Chen, Vinod Koverkathu, Valentin Dalibard, Yaming Xu, Brennan Saeta, Keith Anderson, Thibault Sellam, Nick Fernando, Fantine Huot, Junehyuk Jung, Mani Varadarajan, Michael Quinn, Amit Raul, Maigo Le, Ruslan Habalov, Jon Clark, Komal Jalan, Kalesha Bullard, Achintya Singhal, Thang Luong, Boyu Wang, Sujeevan Rajayogam, Julian Eisenschlos, Johnson Jia, Daniel Finchelstein, Alex Yakubovich, Daniel Balle, Michael Fink, Sameer Agarwal, Jing Li, Dj Dvijotham, Shalini Pal, Kai Kang, Jaclyn Konzelmann, Jennifer Beattie, Olivier Dousse, Diane Wu, Remi Crocker, Chen Elkind, Siddhartha Reddy Jonnalagadda, Jong Lee, Dan Holtmann-Rice, Krystal Kallarackal, Rosanne Liu, Denis Vnukov, Neera Vats, Luca Invernizzi, Mohsen Jafari, Huanjie Zhou, Lilly Taylor, Jennifer Prendki, Marcus Wu, Tom Eccles, Tianqi Liu, Kavya Kopparapu, Francoise Beaufays, Christof Angermueller, Andreea Marzoca, Shourya Sarcar, Hilal Dib, Jeff Stanway, Frank Perbet, Nejc Trdin, Rachel Sterneck, Andrey Khorlin, Dinghua Li, Xihui Wu, Sonam Goenka, David Madras, Sasha Goldshtein, Willi Gierke, Tong Zhou, Yaxin Liu, Yannie Liang, Anais White, Yunjie Li, Shreya Singh, Sanaz Bahargam, Mark Epstein, Sujoy Basu, Li Lao, Adnan Ozturel, Carl Crous, Alex Zhai, Han Lu, Zora Tung, Neeraj Gaur, Alanna Walton, Lucas Dixon, Ming Zhang, Amir Globerson, Grant Uy, Andrew Bolt, Olivia Wiles, Milad Nasr, Ilia Shumailov, Marco Selvi, Francesco Piccinno, Ricardo Aguilar, Sara McCarthy, Misha Khalman, Mrinal Shukla, Vlado Galic, John Carpenter, Kevin Villela, Haibin Zhang, Harry Richardson, James Martens, Matko Bosnjak, Shreyas Rammohan Belle, Jeff Seibert, Mahmoud Alnahlawi, Brian McWilliams, Sankalp Singh, Annie Louis, Wen Ding, Dan Popovici, Lenin Simicich, Laura Knight, Pulkit Mehta, Nishesh

Gupta, Chongyang Shi, Saaber Fatehi, Jovana Mitrovic, Alex Grills, Joseph Pagadora, Tsendsuren Munkhdalai, Dessie Petrova, Danielle Eisenbud, Zhishuai Zhang, Damion Yates, Bhavishya Mittal, Nilesh Tripuraneni, Yannis Assael, Thomas Brovelli, Prateek Jain, Mihajlo Velimirovic, Canfer Akbulut, Jiaqi Mu, Wolfgang Macherey, Ravin Kumar, Jun Xu, Haroon Qureshi, Gheorghe Comanici, Jeremy Wiesner, Zhitao Gong, Anton Ruddock, Matthias Bauer, Nick Felt, Anirudh GP, Anurag Arnab, Dustin Zelle, Jonas Rothfuss, Bill Rosgen, Ashish Shenoy, Bryan Seybold, Xinjian Li, Jayaram Mudigonda, Goker Erdogan, Jiawei Xia, Jiri Simsa, Andrea Michi, Yi Yao, Christopher Yew, Steven Kan, Isaac Caswell, Carey Radebaugh, Andre Elisseeff, Pedro Valenzuela, Kay McKinney, Kim Paterson, Albert Cui, Eri Latorre-Chimoto, Solomon Kim, William Zeng, Ken Durden, Priya Ponnapalli, Tiberiu Sosea, Christopher A. Choquette-Choo, James Manyika, Brona Robenek, Harsha Vashisht, Sebastien Pereira, Hoi Lam, Marko Velic, Denese Owusu-Afriyie, Katherine Lee, Tolga Bolukbasi, Alicia Parrish, Shawn Lu, Jane Park, Balaji Venkatraman, Alice Talbert, Lambert Rosique, Yuchung Cheng, Andrei Sozanschi, Adam Paszke, Praveen Kumar, Jessica Austin, Lu Li, Khalid Salama, Bartek Perz, Wooyeol Kim, Nandita Dukkipati, Anthony Baryshnikov, Christos Kaplanis, XiangHai Sheng, Yuri Chervonyi, Caglar Unlu, Diego de Las Casas, Harry Askham, Kathryn Tunyasuvunakool, Felix Gimeno, Siim Poder, Chester Kwak, Matt Miecnikowski, Vahab Mirrokni, Alek Dimitriev, Aaron Parisi, Dangyi Liu, Tomy Tsai, Toby Shevlane, Christina Kouridi, Drew Garmon, Adrian Goedeckemeyer, Adam R. Brown, Anitha Vijayakumar, Ali Elqursh, Sadegh Jazayeri, Jin Huang, Sara Mc Carthy, Jay Hoover, Lucy Kim, Sandeep Kumar, Wei Chen, Courtney Biles, Garrett Bingham, Evan Rosen, Lisa Wang, Qijun Tan, David Engel, Francesco Pongetti, Dario de Cesare, Dongseong Hwang, Lily Yu, Jennifer Pullman, Srini Narayanan, Kyle Levin, Siddharth Gopal, Megan Li, Asaf Aharoni, Trieu Trinh, Jessica Lo, Norman Casagrande, Roopali Vij, Loic Matthey, Bramandia Ramadhana, Austin Matthews, C. J. Carey, Matthew Johnson, Kremena Goranova, Rohin Shah, Shereen Ashraf, Kingshuk Dasgupta, Rasmus Larsen, Yicheng Wang, Manish Reddy Vuyyuru, Chong Jiang, Joana Ijazi, Kazuki Osawa, Celine Smith, Ramya Sree Boppana, Taylan Bilal, Yuma Koizumi, Ying Xu, Yasemin Altun, Nir Shabat, Ben Bariach, Alex Korchemniy, Kiam Choo, Olaf Ronneberger, Chimezie Iwuanyanwu, Shubin Zhao, David Soergel, Cho-Jui Hsieh, Irene Cai, Shariq Iqbal, Martin Sundermeyer, Zhe Chen, Elie Bursztein, Chaitanya Malaviya, Fadi Biadsy, Prakash Shroff, Inderjit Dhillon, Tejasi Latkar, Chris Dyer, Hannah Forbes, Massimo Nicosia, Vitaly Nikolaev, Somer Greene, Marin Georgiev, Pidong Wang, Nina Martin, Hanie Sedghi, John Zhang, Praseem Banzal, Doug Fritz, Vikram Rao, Xuezhi Wang, Jiageng Zhang, Viorica Patraucean, Dayou Du, Igor Mordatch, Ivan Jurin, Lewis Liu, Ayush Dubey, Abhi Mohan, Janek Nowakowski, Vlad-Doru Ion, Nan Wei, Reiko Tojo, Maria Abi Raad, Drew A. Hudson, Vaishakh Keshava, Shubham Agrawal, Kevin Ramirez, Zhichun Wu, Hoang Nguyen, Ji Liu, Madhavi Sewak, Bryce Petrini, DongHyun Choi, Ivan Philips, Ziyue Wang, Ioana Bica, Ankush Garg, Jarek Wilkiewicz, Priyanka Agrawal, Xiaowei Li, Danhao Guo, Emily Xue, Naseer Shaik, Andrew Leach, Sadh MNM Khan, Julia Wiesinger, Sammy Jerome, Abhishek Chakladar, Alek Wenjiao Wang, Tina Ornduff, Folake Abu, Alireza Ghaffarkhah, Marcus Wainwright, Mario Cortes, Frederick Liu, Joshua Maynez, Andreas Terzis, Pouya Samangouei, Riham Mansour, Tomasz Kępa, François-Xavier Aubet, Anton Algymr, Dan Banica, Agoston Weisz, Andras Orban, Alexandre Senges, Ewa Andrejczuk, Mark Geller, Niccolo Dal Santo, Valentin Anklin, Majd Al Merey, Martin Baeuml, Trevor Strohman, Junwen Bai, Slav Petrov, Yonghui Wu, Demis Hassabis, Koray Kavukcuoglu, Jeff Dean, and Oriol Vinyals. 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.

Perplexity Team. RL Training For Math Reasoning, May 2025b. URL https://www.perplexity.ai/hub.

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].

A Appendix

B Topology Optimization Solver Parameters: Grasshopper Millipede

Solver parameters are:

target density = 0.1
self-weight = 0
iterations = 10
smoothing = 0.1
penalization = 3.0
minimum density = 0.001
delete threshold = 0.5
compliant mechanism disabled

C VISUAL TASK VARIATIONS OVERVIEW

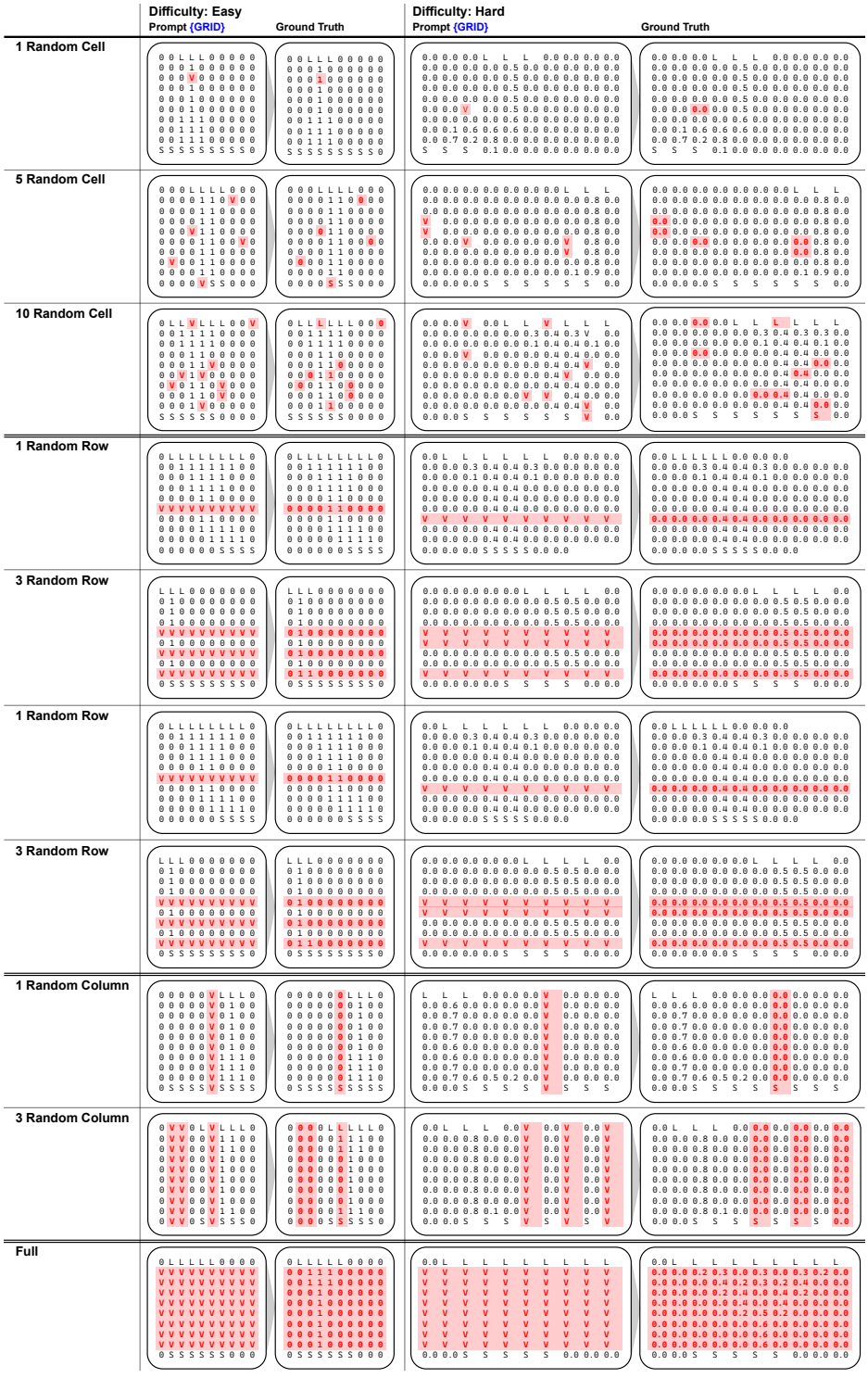

Figure 3: Overview of task variations: predicting material distributions for N random cells, rows, columns, or full structures for easy (binary) and hard (continous) difficulties.

D Additional Evaluation Metrics Computation and Prompt and Completion Examples

D.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 Dijkstra's 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.

D.2 Reconstruction Metric Tests

D.2.1 Exact Match Examples

Exact Match Examples

```
This test validates the get_exact_match function, which returns True
if the predicted grid Ĝ 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
```

D.2.2 Difference Ratio Examples

Difference Ratio Examples

```
This test validates get_difference_ratio, which measures similarity
between Ĝ 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
```

D.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.
```

D.3 Topology Metric

D.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.
```

D.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.
```

D.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.
```

D.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.
```

D.4 Physics Approximation Metric

D.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.
```

E TOPOLOGY OPTIMIZATION SAMPLE PLOTS

E.1 2D SAMPLES

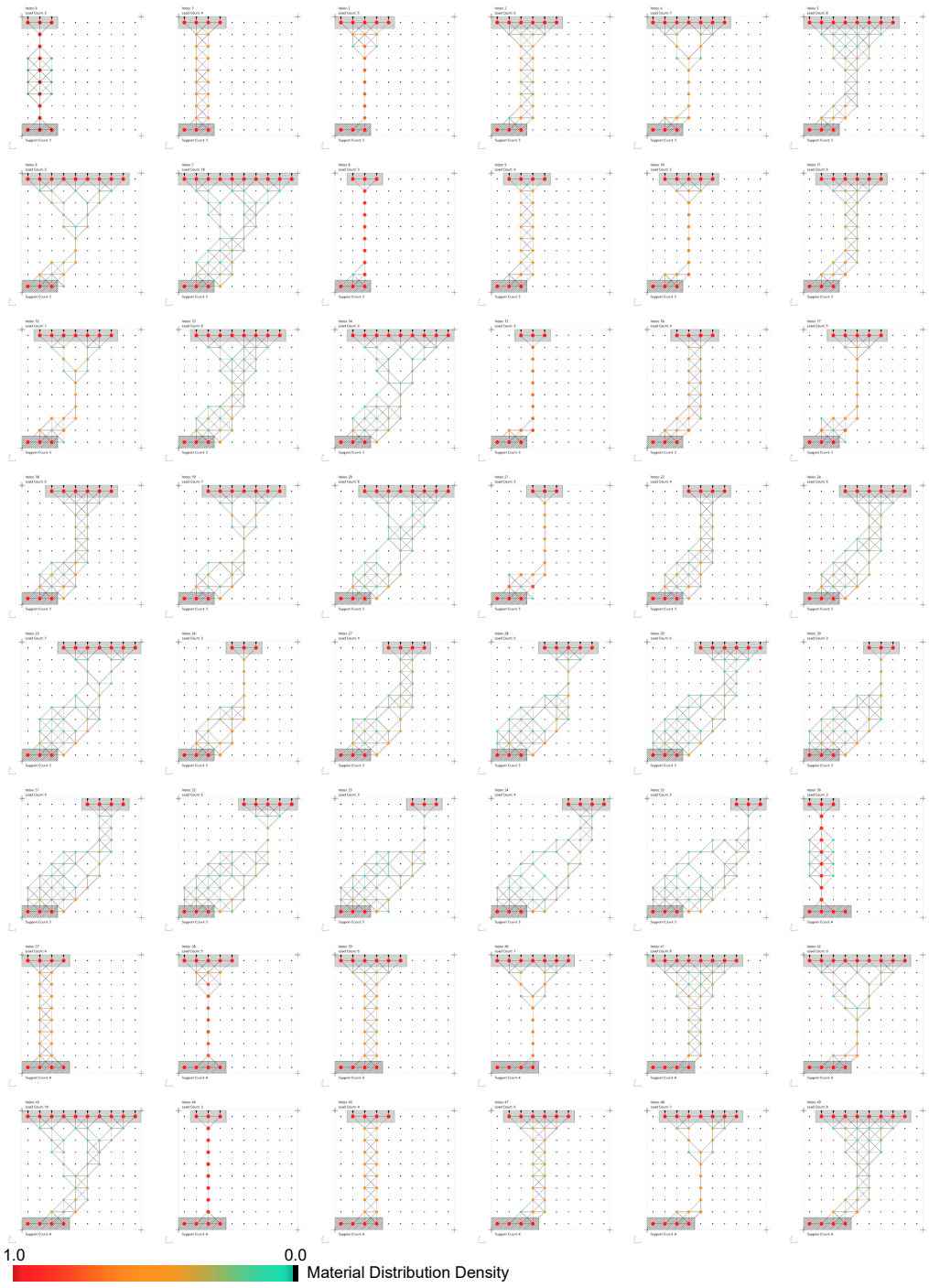

Figure 4: Example 2D topology optimization samples from the SPhyR dataset.

E.2 3D SAMPLES

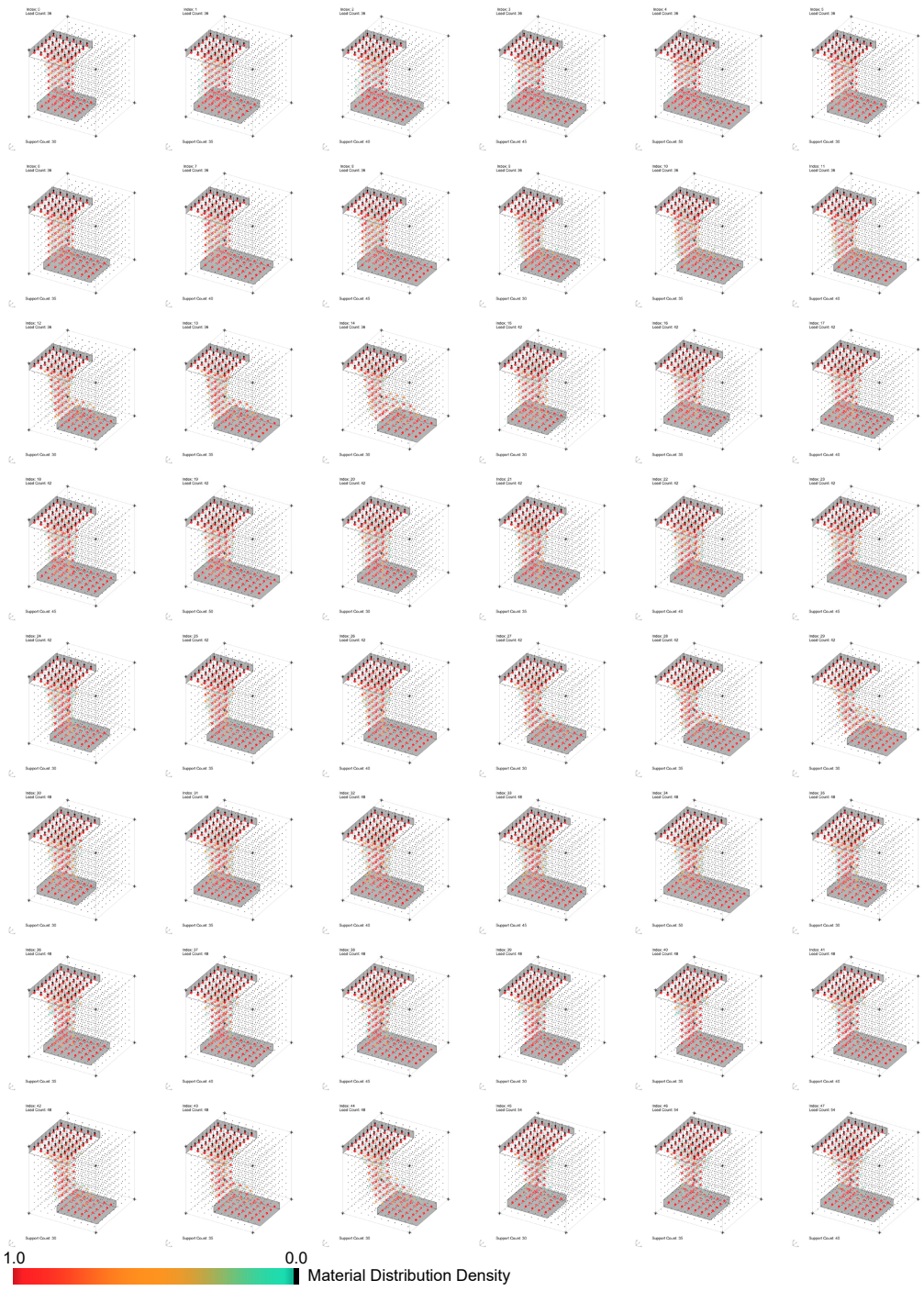

Figure 5: Example 3D topology optimization samples included for future benchmark extensions.

F Additional Main Run Results

F.1 Results for All Models and Tasks

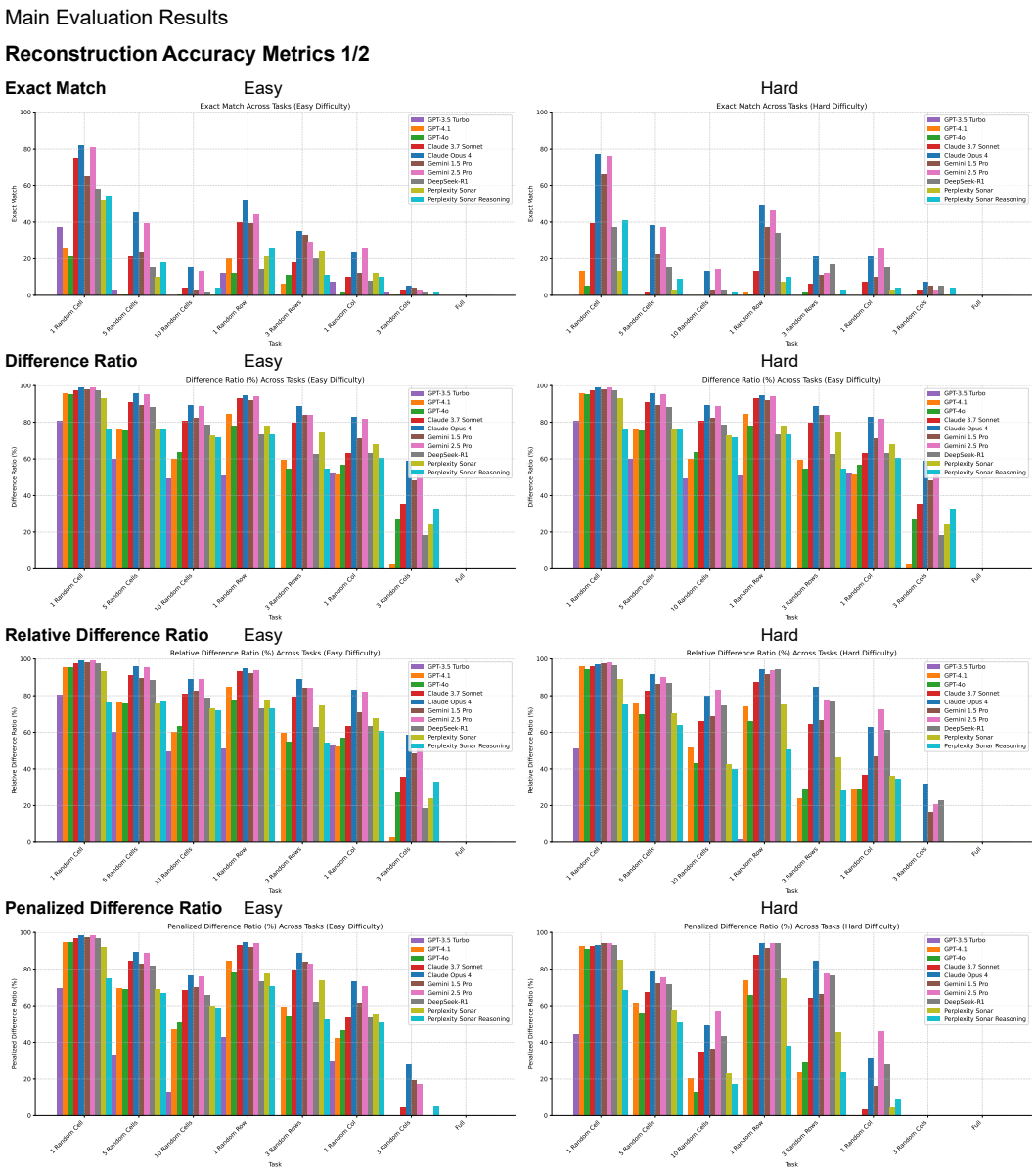

Figure 6: Main evaluation run results: Exact Match, Difference Ratio, Relative Difference Ratio and Penalized Difference Ratio for all models, across all tasks and difficulties.

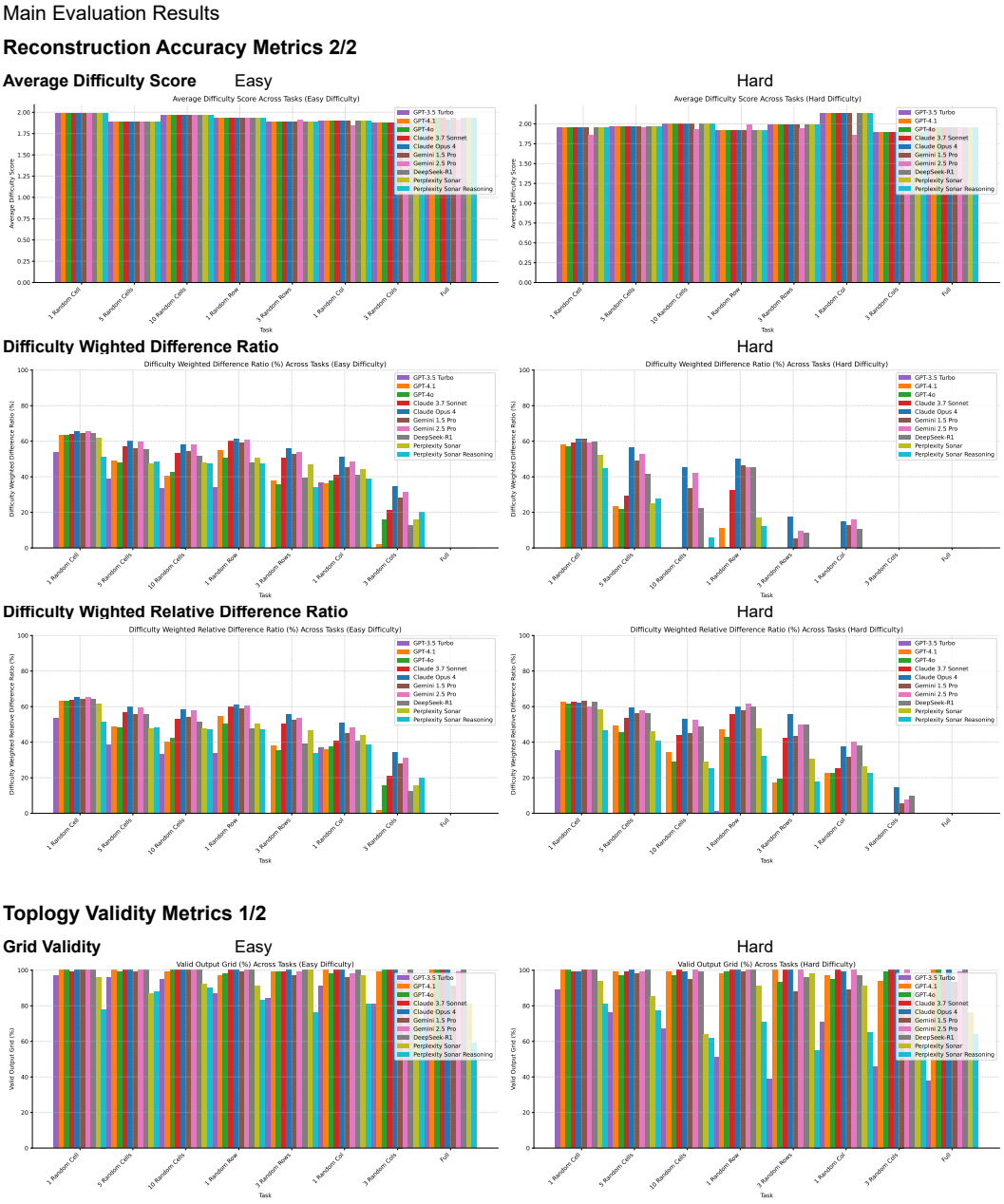

Figure 7: Main evaluation run results: Average Difficulty Score, Difficulty Weighted Difference Ratio, Difficulty Weighted Relative Difference Ratio and Grid Validity for all models, across all tasks and difficulties.

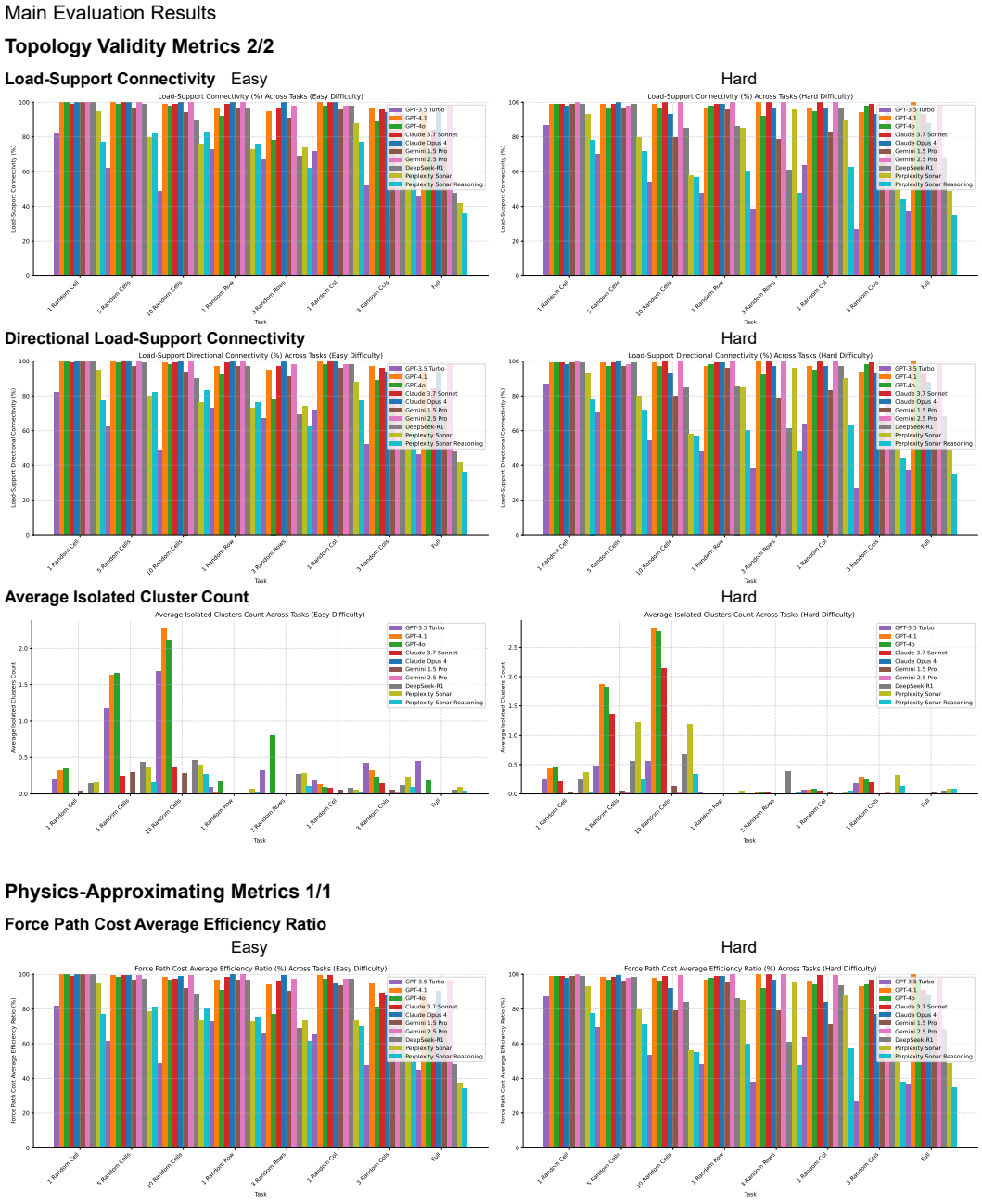

Figure 8: Main evaluation run results: Load-Support Connectivity, Directional Load-Support Connectivity, Average Isolated Cluster Count and Force Path Cost Average Efficiency Ratio for all models, across all tasks and difficulties.

G Additional Rotation Experiment Results

G.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
```

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

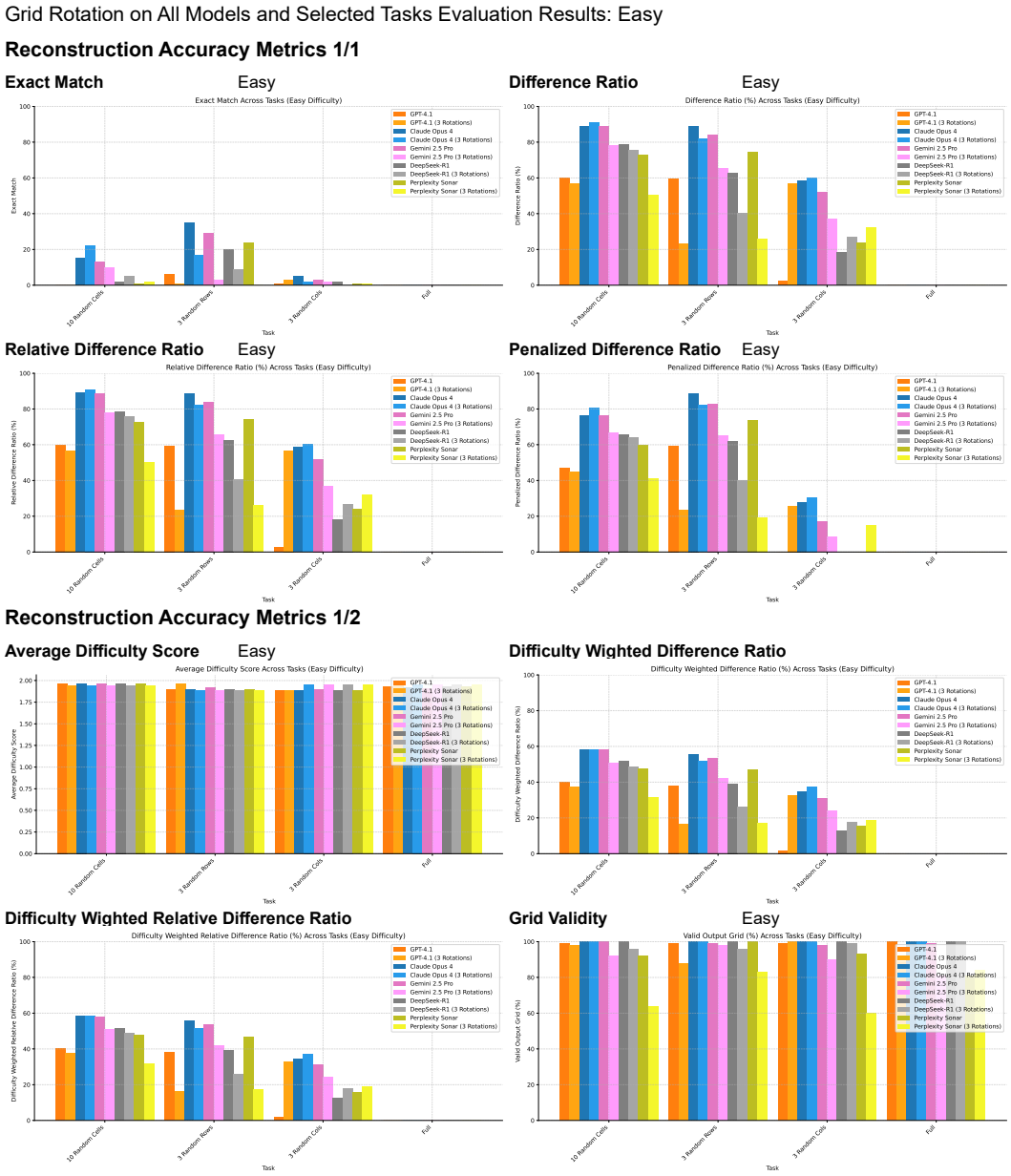

Figure 9: Grid rotation evaluation results: Exact Match, Difference Ratio, Relative Difference Ratio, Penalized Difference Ratio, Average Difficulty Score, Difficulty Weighted Difference Ratio, Difficulty Weighted Relative Difference Ratio and Grid Validity for GPT-4.1, Claude Opus 4, Gemini 2.5 Pro, DeepSeek-R1 and Perplexity, for 10 Random Cells, 3 Random Rows, 3 Random Cells and Full tasks and easy difficulty.

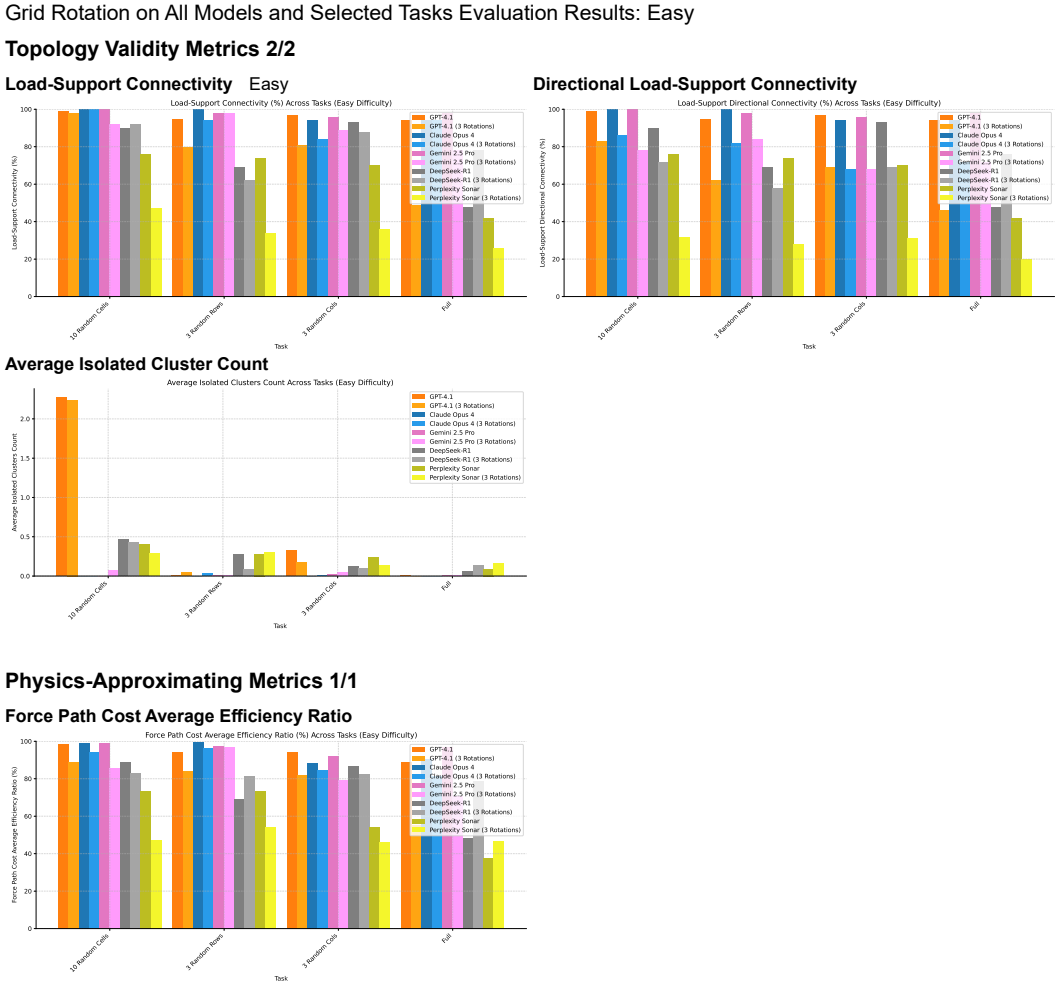

Figure 10: Grid rotation evaluation results: Load-Support Connectivity, Directional Load-Support Connectivity, Average Isolated Cluster Count and Force Path Cost Average Efficiency Ratio for GPT-4.1, Claude Opus 4, Gemini 2.5 Pro, DeepSeek-R1 and Perplexity, for 10 Random Cells, 3 Random Rows, 3 Random Cells and Full tasks and easy difficulty.

Table 4: Grid rotation evaluation results for all metrics, for GPT-4.1, Claude Opus 4, Gemini 2.5 Pro, DeepSeek-R1 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-R1	DeepSeek-R1 (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

G.3 Results for Claude Opus 4 on All Tasks

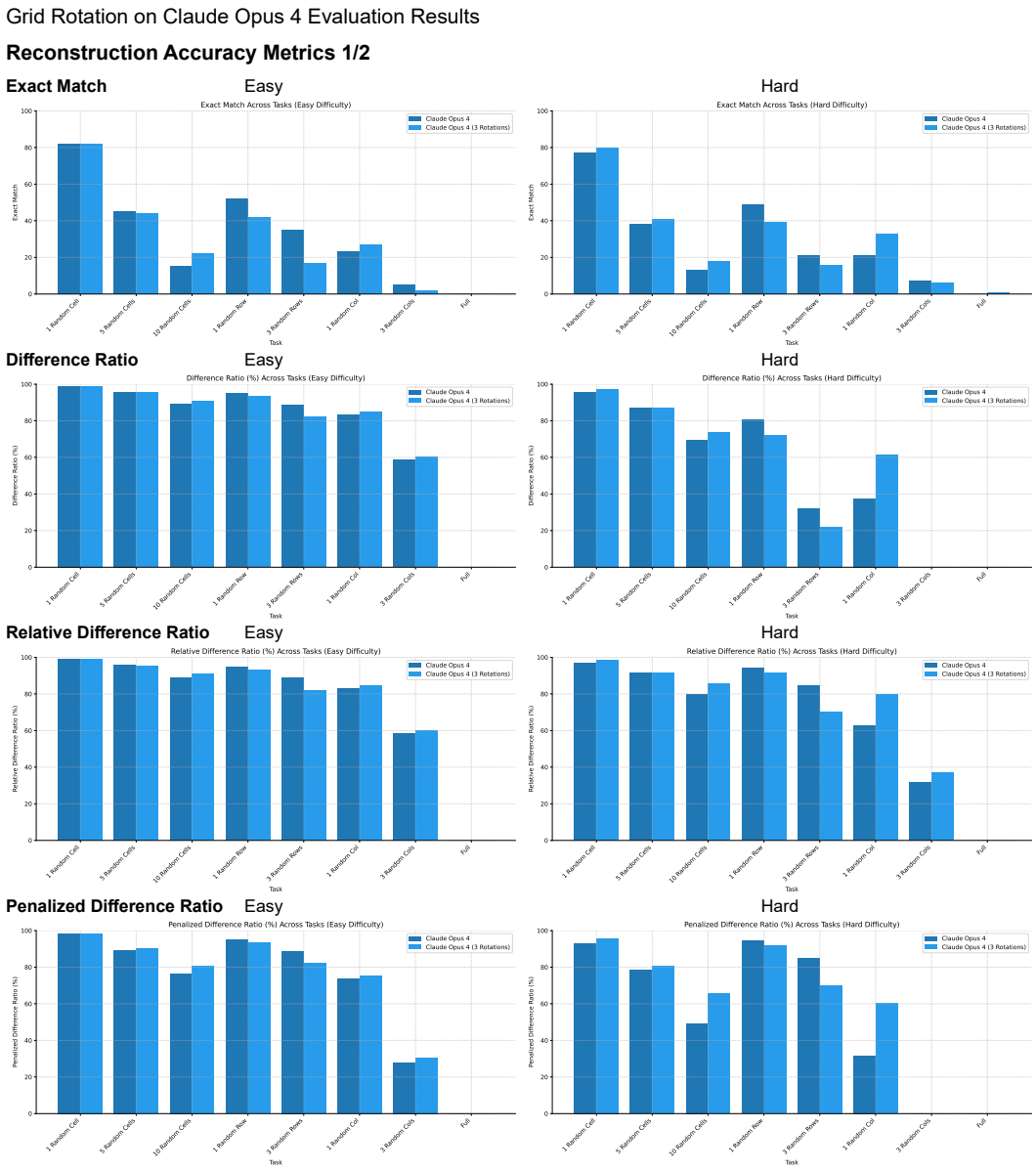

Figure 11: Grid rotation evaluation results: Exact Match, Difference Ratio, Relative Difference Ratio and Penalized Difference Ratio for Claude Opus 4, for all tasks, easy and hard difficulty.

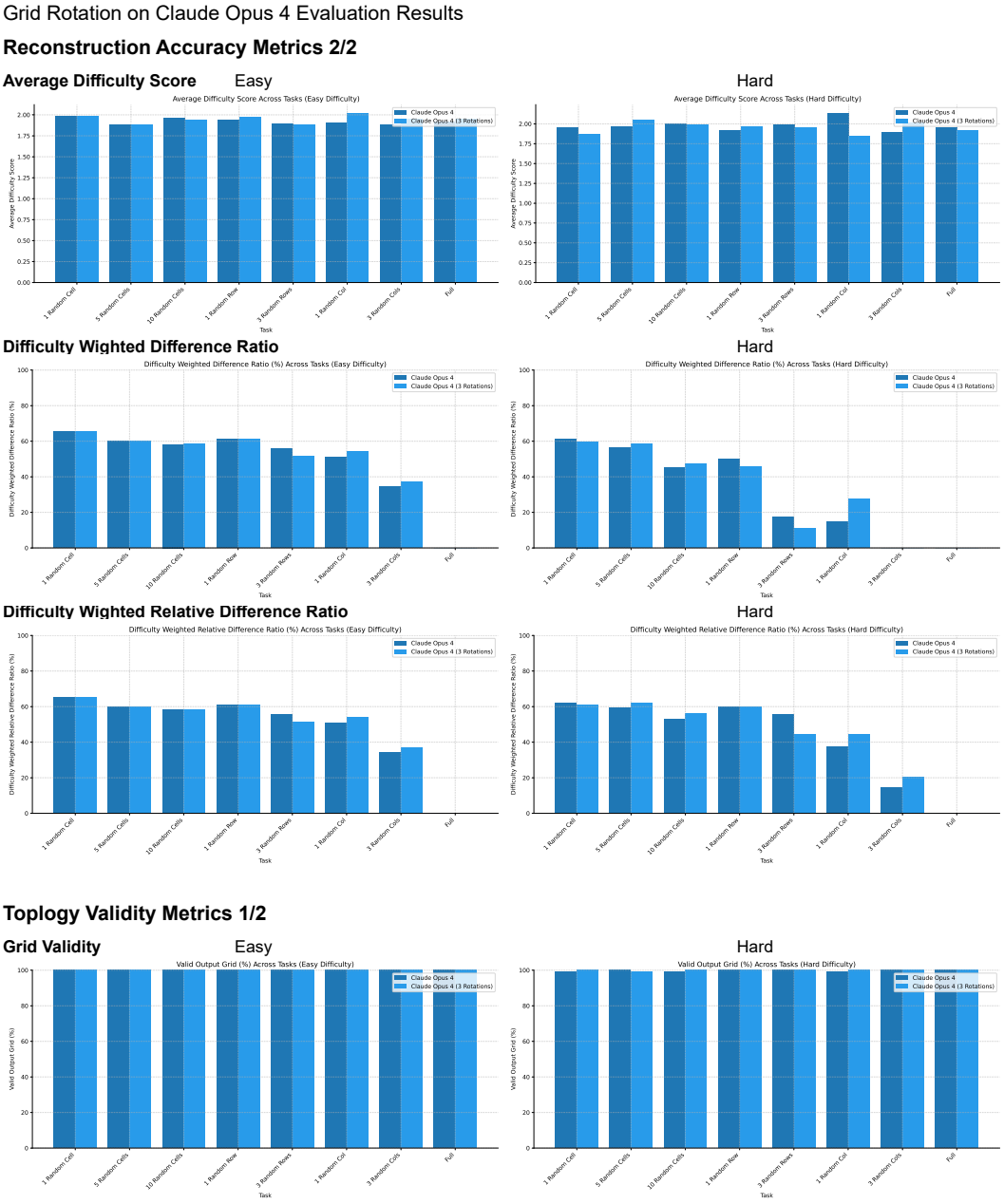

Figure 12: Grid rotation evaluation results: Average Difficulty Score, Difficulty Weighted Difference Ratio, Difficulty Weighted Relative Difference Ratio and Grid Validity for Claude Opus 4, for all tasks, easy and hard difficulty.

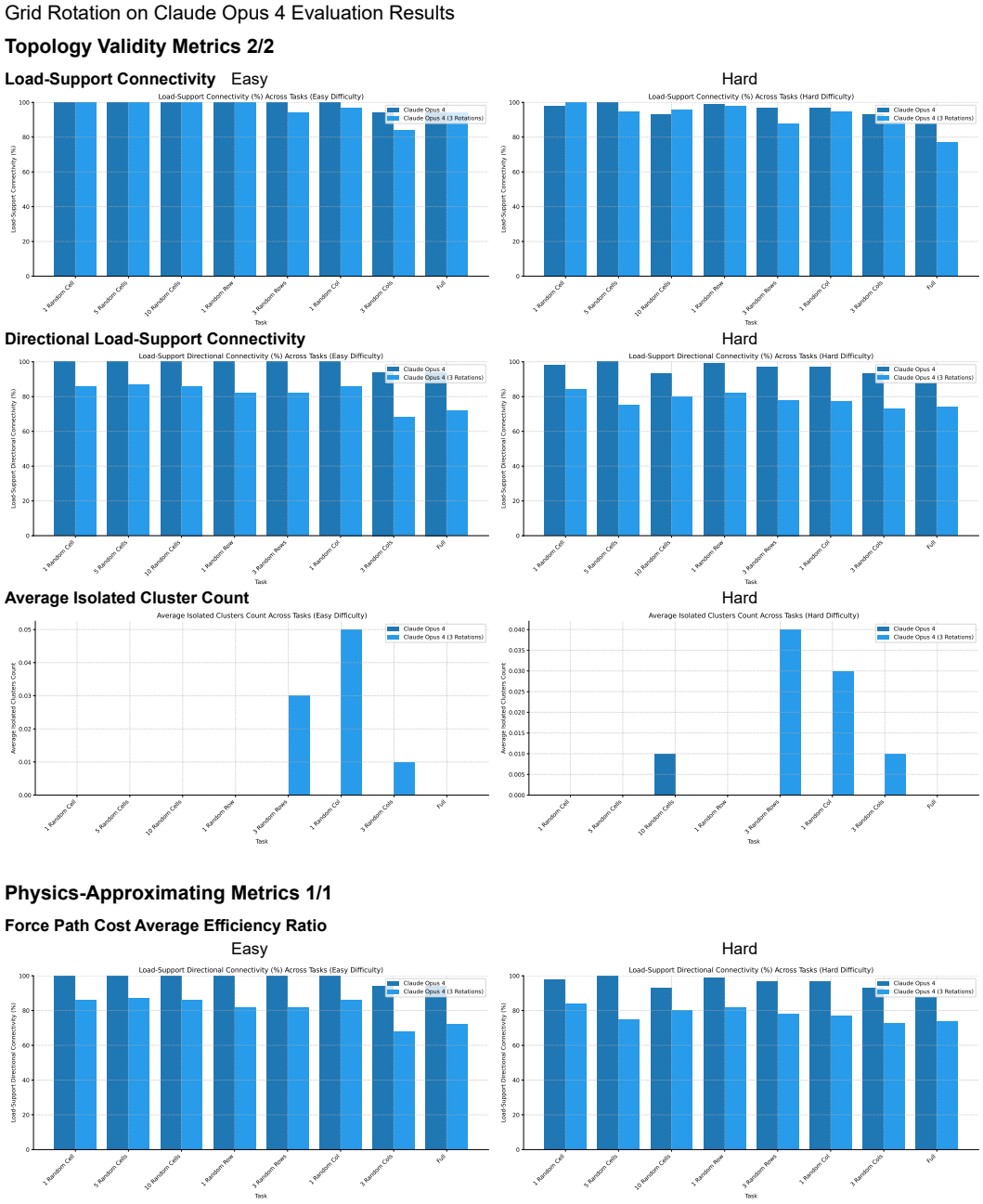

Figure 13: Grid rotation evaluation results: Load-Support Connectivity, Directional Load-Support Connectivity, Average Isolated Cluster Count and Force Path Cost Average Efficiency Ratio for Claude Opus 4, for all tasks, easy and hard difficulty.

Table 5: 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**

H Additional Few-Shot Experiment Results

H.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
```

H.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
```

H.3 Results for Claude Opus 4 on All Tasks

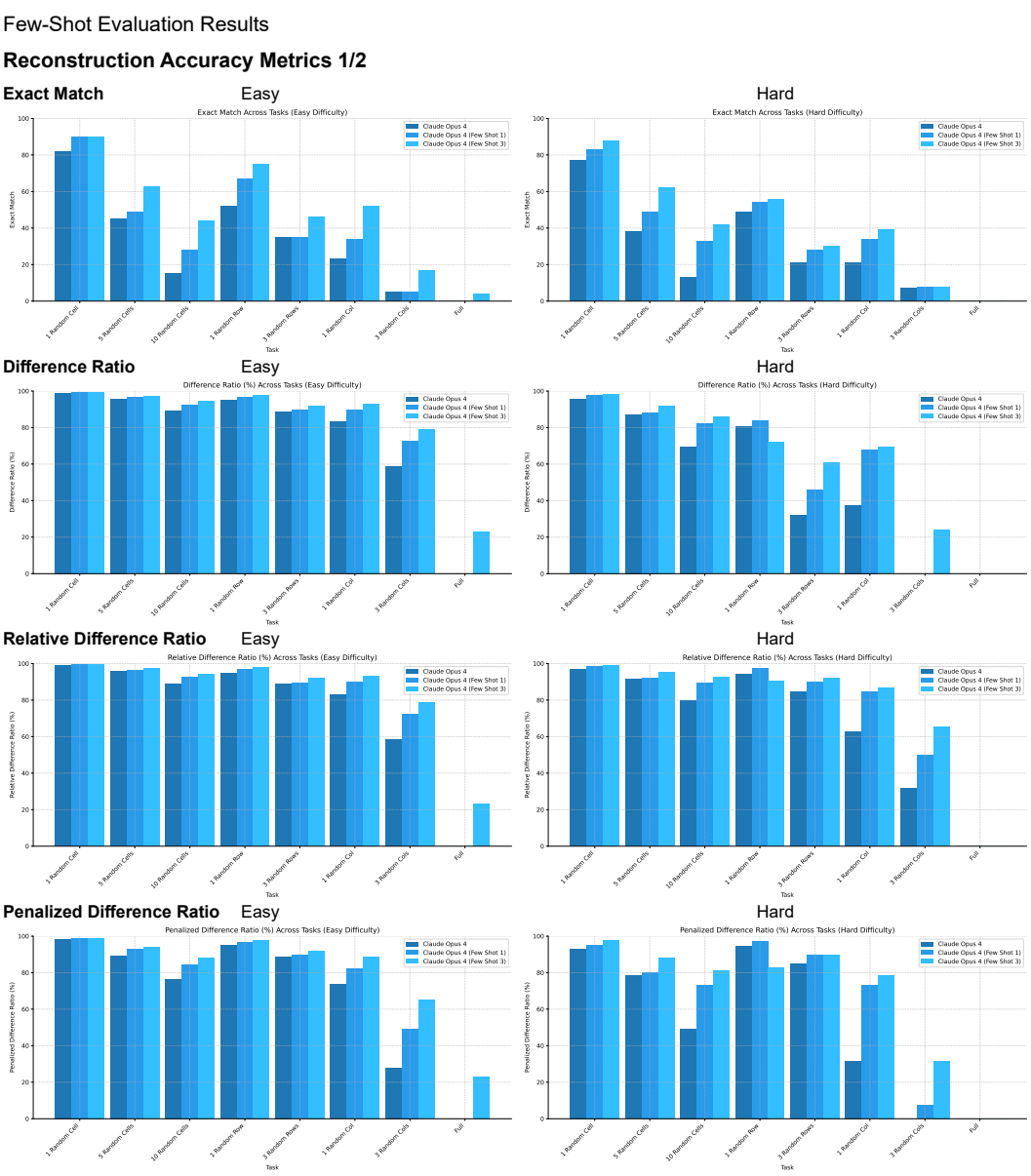

Figure 14: Few-shot (1, 3) evaluation results: Exact Match, Difference Ratio, Relative Difference Ratio and Penalized Difference Ratio for Claude Opus 4, for all tasks, easy and hard difficulty.

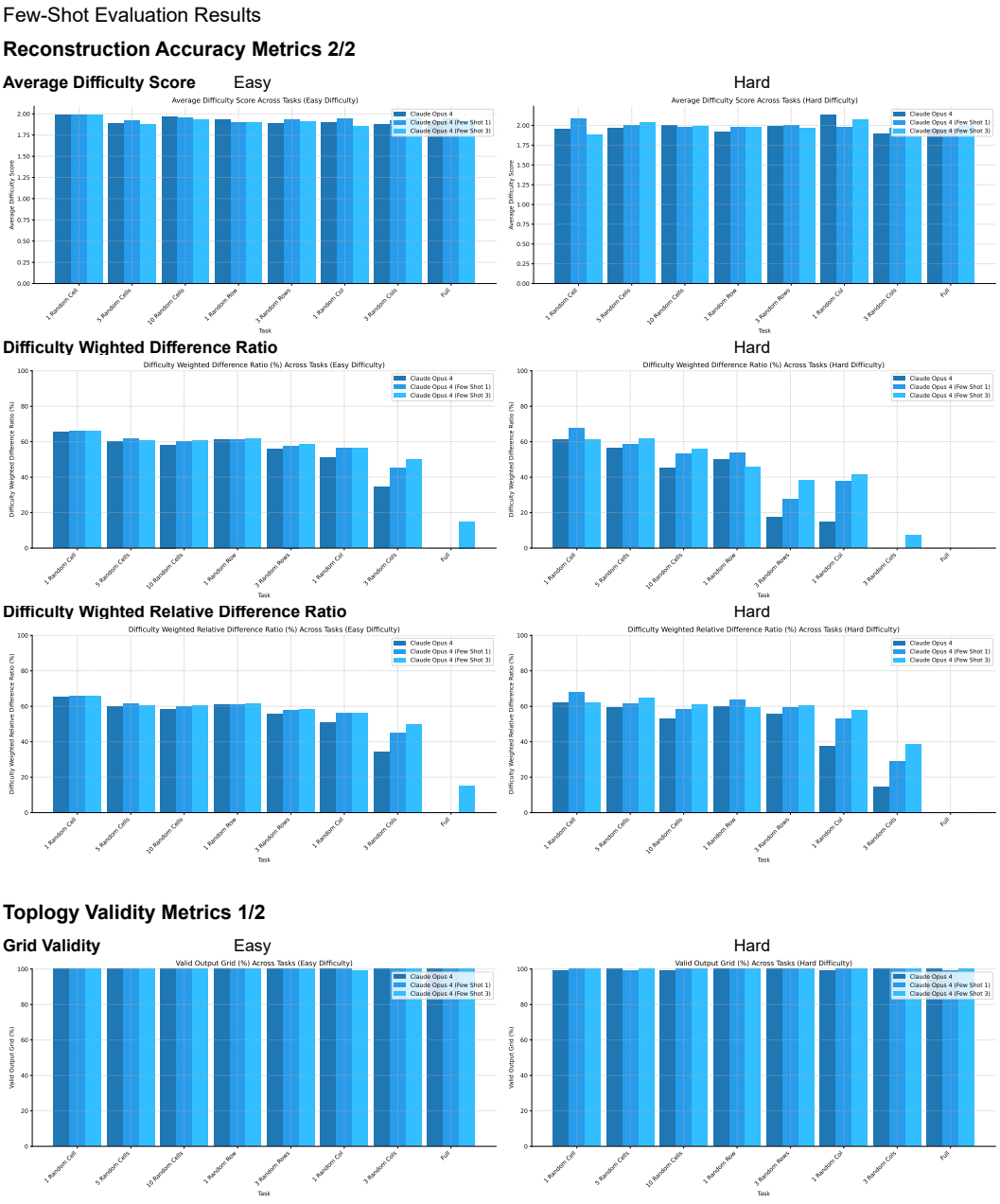

Figure 15: Few-shot (1, 3) evaluation results: Average Difficulty Score, Difficulty Weighted Difference Ratio, Difficulty Weighted Relative Difference Ratio and Grid Validity for Claude Opus 4, for all tasks, easy and hard difficulty.

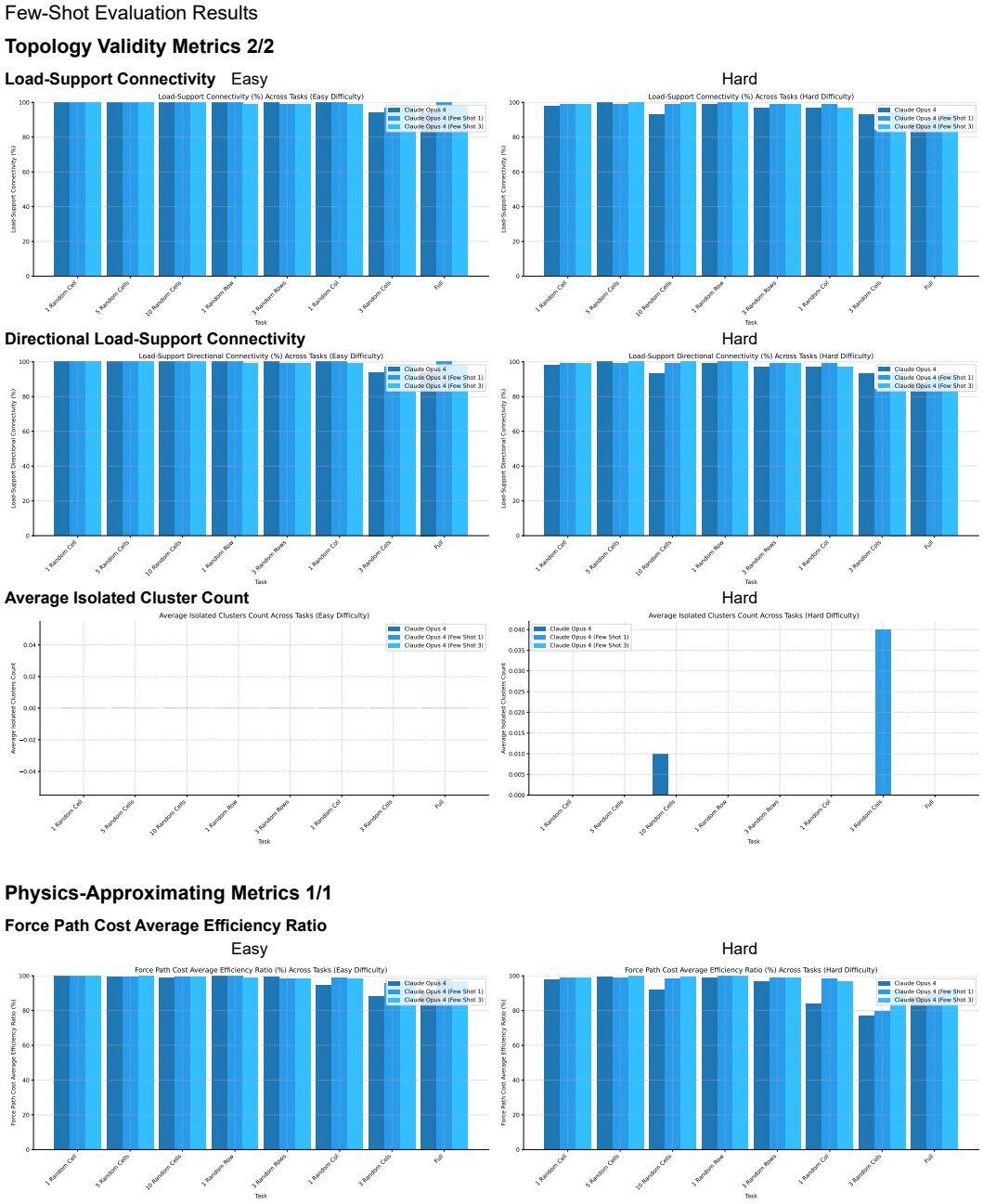

Figure 16: Few-shot (1, 3) evaluation results: Load-Support Connectivity, Directional Load-Support Connectivity, Average Isolated Cluster Count and Force Path Cost Average Efficiency Ratio for Claude Opus 4, for all tasks, easy and hard difficulty.

Table 6: 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**

I Additional Physics-Enhanced and -Neutral Experiment Results

I.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
```

I.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
```

I.3 Results for Claude Opus 4 on All Tasks

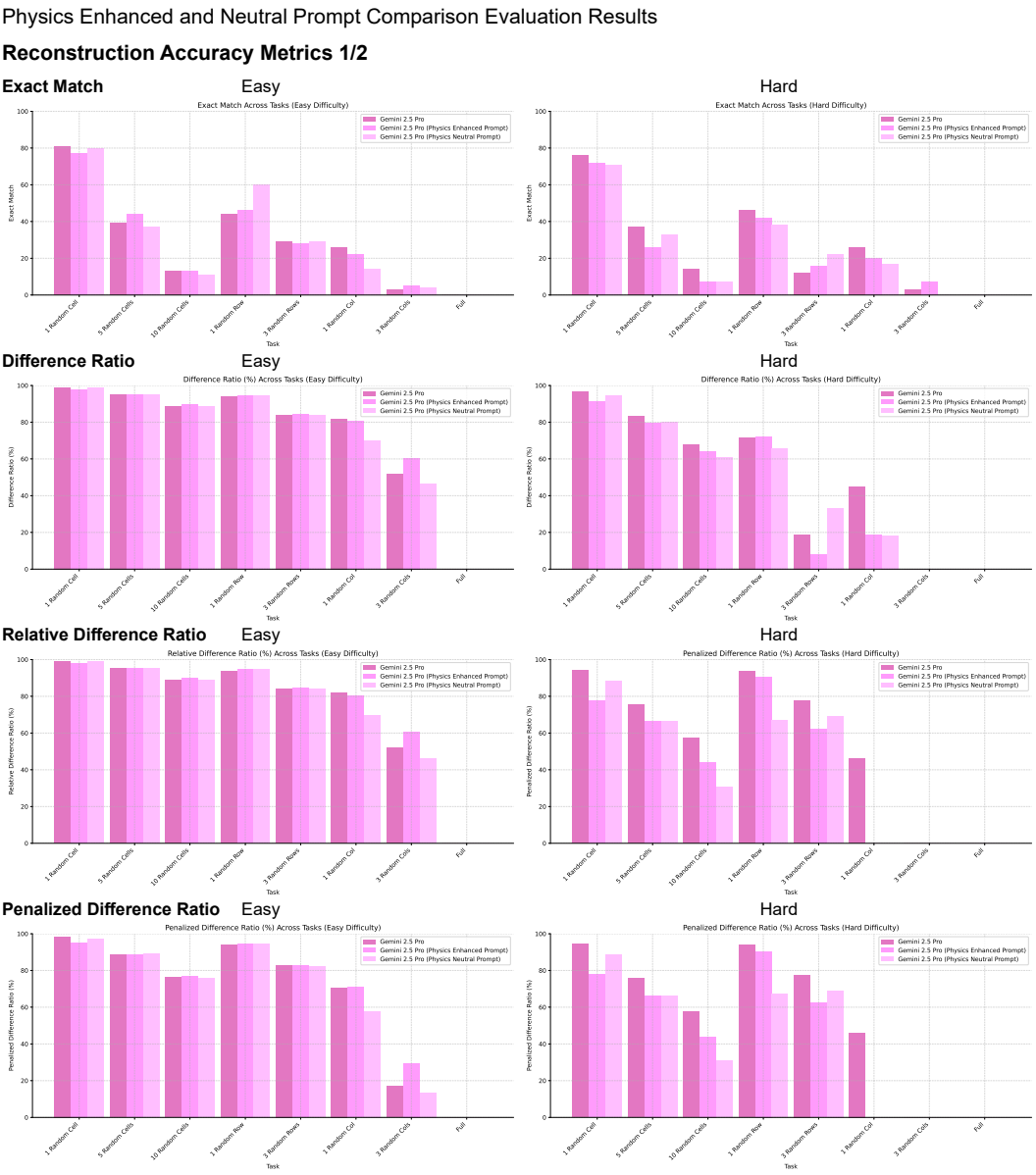

Figure 17: Physics-Enhanced and -Neutral evaluation run metric: Exact Match, Difference Ratio, Relative Difference Ratio and Penalized Difference Ratio for Claude Opus 4, for all tasks, easy and hard difficulty.

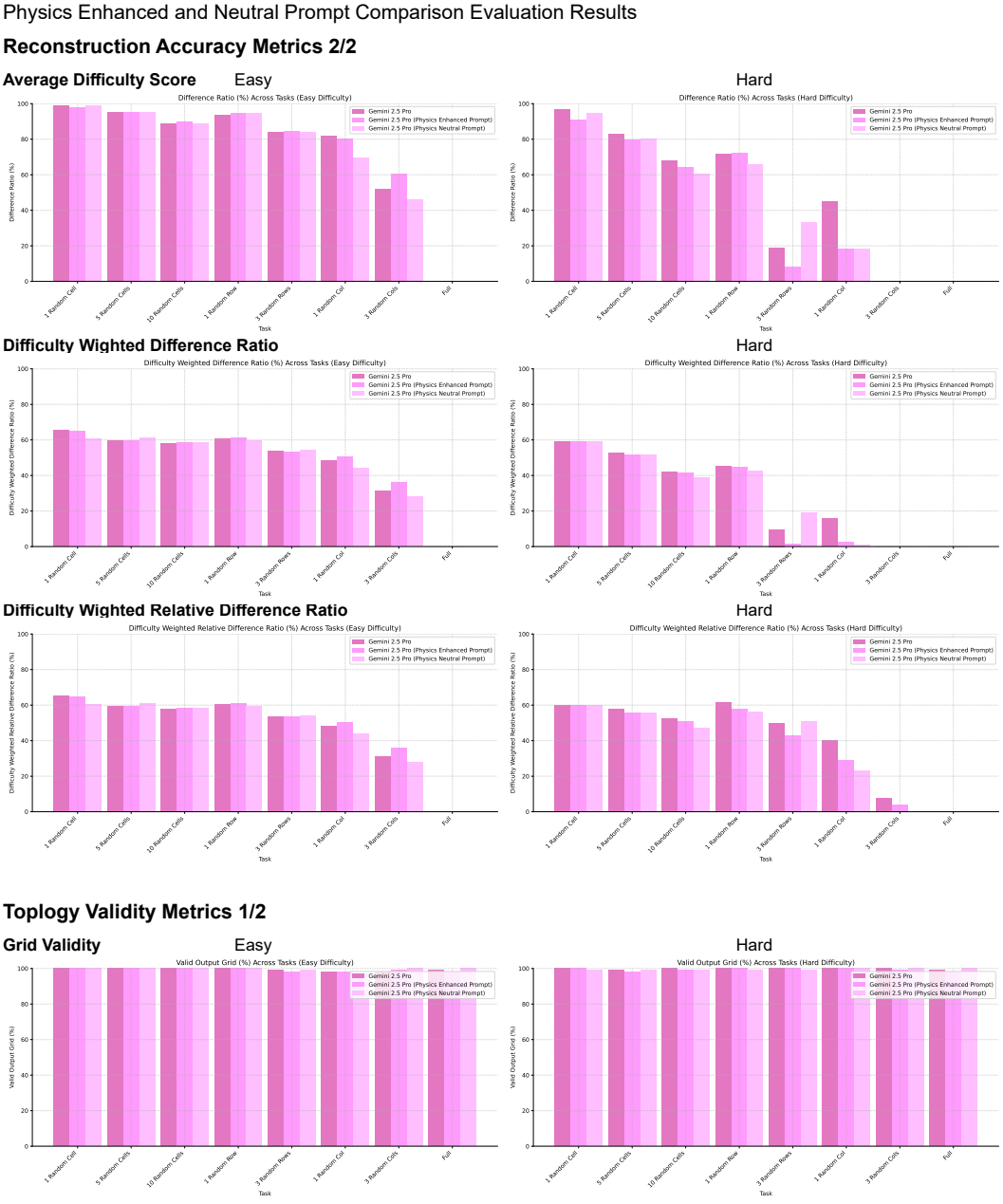

Figure 18: Physics-Enhanced and -Neutral evaluation run metric: Average Difficulty Score, Difficulty Weighted Difference Ratio, Difficulty Weighted Relative Difference Ratio and Grid Validity for Claude Opus 4, for all tasks, easy and hard difficulty.

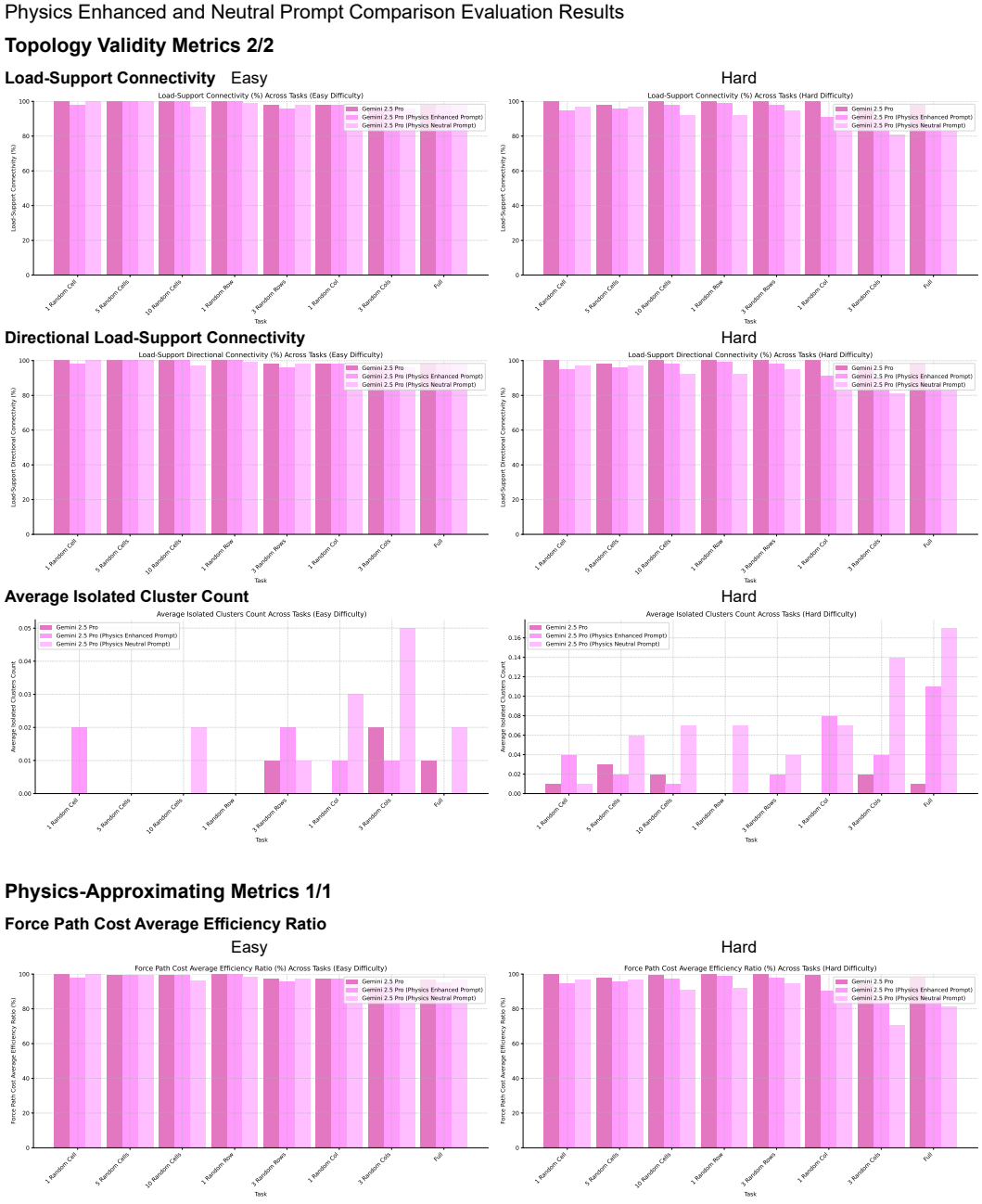

Figure 19: Physics-Enhanced and -Neutral evaluation result: Load-Support Connectivity, Directional Load-Support Connectivity, Average Isolated Cluster Count and Force Path Cost Average Efficiency Ratio for Claude Opus 4, for all tasks, easy and hard difficulty.

Table 7: 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	**70.85**	57.94	**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	**-298.08**
	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

J SAMPLE PROMPTS & COMPLETIONS

J.1 CLAUDE 3.7 SONNET - 1 RANDOM CELL EASY: SAMPLE 1

Claude 3.7 Sonnet - 1 Random Cell Easy: Sample 1

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 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
```

J.2 CLAUDE 3.7 SONNET - 1 RANDOM CELL EASY: SAMPLE 2

Claude 3.7 Sonnet - 1 Random Cell Easy: Sample 2

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 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 1 1 1 1 0 0
0 0 0 1 1 1 1 1 0 0
0 0 1 1 1 1 1 0 0 0
0 1 1 1 1 1 V 0 0 0
S S S S S 1 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):

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

Ground Truth:

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

J.3 CLAUDE 3.7 SONNET - 1 RANDOM CELL HARD

Claude 3.7 Sonnet - 1 Random Cell 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 0.0 0.0 L   L   L   0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0 0.5 0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0 0.5 0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0 0.5 0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0 0.5 0.0 0.0 0.0 0.0 0.0
0.0 0.0 V   0.0 0.5 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 0.1 0.6 0.6 0.6 0.0 0.0 0.0 0.0 0.0
0.0 0.7 0.2 0.8 0.0 0.0 0.0 0.0 0.0 0.0
S   S   S   0.1 0.0 0.0 0.0 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 0.0 0.0 L   L   L   0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0 0.5 0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0 0.5 0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0 0.5 0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0 0.5 0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.3 0.0 0.5 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 0.1 0.6 0.6 0.6 0.0 0.0 0.0 0.0 0.0
0.0 0.7 0.2 0.8 0.0 0.0 0.0 0.0 0.0 0.0
S   S   S   0.1 0.0 0.0 0.0 0.0 0.0 0.0
```

Ground Truth:

```
0.0 0.0 0.0 L   L   L   0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0 0.5 0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0 0.5 0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0 0.5 0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0 0.5 0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0 0.5 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 0.1 0.6 0.6 0.6 0.0 0.0 0.0 0.0 0.0
0.0 0.7 0.2 0.8 0.0 0.0 0.0 0.0 0.0 0.0
S   S   S   0.1 0.0 0.0 0.0 0.0 0.0 0.0
```

J.4 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
```

J.5 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
```

J.6 DeepSeek-R1 - 1 Random Cell Easy: Sample 1

DeepSeek-R1 - 1 Random Cell Easy: Sample 1

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 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
```

J.7 DeepSeek-R1 - 1 Random Cell Easy: Sample 2

DeepSeek-R1 - 1 Random Cell Easy: Sample 2

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:

L L L L L L 0 0 0 0
0 1 1 1 1 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 1 0 0 0 0 0
0 0 1 1 1 1 0 0 0 0
V 0 0 1 1 1 1 0 0 0
0 0 0 1 1 1 1 1 0 0
0 0 0 0 1 1 1 1 0 0
0 0 0 0 0 1 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):

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

Ground Truth:

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

J.8 DeepSeek-R1 - 1 Random Cell Hard

DeepSeek-R1 - 1 Random Cell 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 0.0 0.0 L   L   L   0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0 0.5 0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0 0.5 0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0 0.5 0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0 0.5 0.0 0.0 0.0 0.0 0.0
0.0 0.0 V   0.0 0.5 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 0.1 0.6 0.6 0.6 0.0 0.0 0.0 0.0 0.0
0.0 0.7 0.2 0.8 0.0 0.0 0.0 0.0 0.0 0.0
S   S   S   0.1 0.0 0.0 0.0 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 0.0 0.0 L   L   L   0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0 0.5 0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0 0.5 0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0 0.5 0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0 0.5 0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.3 0.0 0.5 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 0.1 0.6 0.6 0.6 0.0 0.0 0.0 0.0 0.0
0.0 0.7 0.2 0.8 0.0 0.0 0.0 0.0 0.0 0.0
S   S   S   0.1 0.0 0.0 0.0 0.0 0.0 0.0
```

Ground Truth:

```
0.0 0.0 0.0 L   L   L   0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0 0.5 0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0 0.5 0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0 0.5 0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0 0.5 0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0 0.5 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 0.1 0.6 0.6 0.6 0.0 0.0 0.0 0.0 0.0
0.0 0.7 0.2 0.8 0.0 0.0 0.0 0.0 0.0 0.0
S   S   S   0.1 0.0 0.0 0.0 0.0 0.0 0.0
```

J.9 DeepSeek-R1 - Full Easy

DeepSeek-R1 - 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
```

J.10 DEEPSEEK-R1 - FULL HARD

DeepSeek-R1 - 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   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
```

J.11 GPT-4.1 - 1 Random Cell Easy: Sample 1

GPT-4.1 – 1 Random Cell Easy: Sample 1

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 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
```

J.12 GPT-4.1 - 1 Random Cell Easy: Sample 2

GPT-4.1 – 1 Random Cell Easy: Sample 2

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:

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

Model Completion (differences in red):
L L L L L L 0 0 0 0
0 1 1 1 1 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 1 0 0 0 0 0
0 0 1 1 1 1 0 0 0 0
1 0 0 1 1 1 1 0 0 0
0 0 0 1 1 1 1 1 0 0
0 0 0 0 1 1 1 1 0 0
0 0 0 0 0 1 S S S 0

Ground Truth:
L L L L L L 0 0 0 0
0 1 1 1 1 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 1 0 0 0 0 0
0 0 1 1 1 1 0 0 0 0
0 0 0 1 1 1 1 0 0 0
0 0 0 1 1 1 1 1 0 0
0 0 0 0 1 1 1 1 0 0
0 0 0 0 0 1 S S S 0
```

J.13 GPT-4.1 - 1 Random Cell Hard

GPT-4.1 – 1 Random Cell 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 0.0 0.0 L   L   L   0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0 0.5 0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0 0.5 0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0 0.5 0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0 0.5 0.0 0.0 0.0 0.0 0.0
0.0 0.0 V   0.0 0.5 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 0.1 0.6 0.6 0.6 0.0 0.0 0.0 0.0 0.0
0.0 0.7 0.2 0.8 0.0 0.0 0.0 0.0 0.0 0.0
S   S   S   0.1 0.0 0.0 0.0 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.
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 L   L   L   0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0 0.5 0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0 0.5 0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0 0.5 0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0 0.5 0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.3 0.0 0.5 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 0.1 0.6 0.6 0.6 0.0 0.0 0.0 0.0 0.0
0.0 0.7 0.2 0.8 0.0 0.0 0.0 0.0 0.0 0.0
S   S   S   0.1 0.0 0.0 0.0 0.0 0.0 0.0
```

Ground Truth:

```
0.0 0.0 0.0 L   L   L   0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0 0.5 0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0 0.5 0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0 0.5 0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0 0.5 0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0 0.5 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 0.1 0.6 0.6 0.6 0.0 0.0 0.0 0.0 0.0
0.0 0.7 0.2 0.8 0.0 0.0 0.0 0.0 0.0 0.0
S   S   S   0.1 0.0 0.0 0.0 0.0 0.0 0.0
```

J.14 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
```

J.15 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 0.4 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 0.0 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 S   S   S   S   S   0.0 0.0 0.0
```

J.16 GEMINI 2.5 PRO - 1 RANDOM CELL EASY: SAMPLE 1

Gemini 2.5 Pro – 1 Random Cell Easy: Sample 1

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 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 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
```

J.17 GEMINI 2.5 PRO - 1 RANDOM CELL EASY: SAMPLE 2

Gemini 2.5 Pro – 1 Random Cell Easy: Sample 2

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:

L L L L L L 0 0 0 0
0 1 1 1 1 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 1 0 0 0 0 0
0 0 1 1 1 1 0 0 0 0
V 0 0 1 1 1 1 0 0 0
0 0 0 1 1 1 1 1 0 0
0 0 0 0 1 1 1 1 0 0
0 0 0 0 0 1 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):

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

Ground Truth:

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

J.18 GEMINI 2.5 PRO - 1 RANDOM CELL HARD

Gemini 2.5 Pro – 1 Random Cell 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 (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   V   L   L   L   L   L   L   0.0
0.0 0.0 0.2 0.3 0.1 0.1 0.3 0.2 0.0 0.0
0.0 0.0 0.1 0.4 0.0 0.0 0.4 0.1 0.0 0.0
0.0 0.0 0.0 0.4 0.0 0.0 0.4 0.0 0.0 0.0
0.0 0.0 0.0 0.4 0.0 0.0 0.4 0.0 0.0 0.0
0.0 0.0 0.0 0.4 0.0 0.0 0.4 0.0 0.0 0.0
0.0 0.0 0.0 0.4 0.0 0.0 0.4 0.0 0.0 0.0
0.0 0.0 0.0 0.4 0.1 0.1 0.4 0.0 0.0 0.0
0.0 0.0 0.0 0.4 0.1 0.1 0.4 0.0 0.0 0.0
0.0 S   S   S   S   S   S   S   S   0.0
```

Model Completion (differences in red):

```
0.0 L   0.2 L   L   L   L   L   L   0.0
0.0 0.0 0.2 0.3 0.1 0.1 0.3 0.2 0.0 0.0
0.0 0.0 0.1 0.4 0.0 0.0 0.4 0.1 0.0 0.0
0.0 0.0 0.0 0.4 0.0 0.0 0.4 0.0 0.0 0.0
0.0 0.0 0.0 0.4 0.0 0.0 0.4 0.0 0.0 0.0
0.0 0.0 0.0 0.4 0.0 0.0 0.4 0.0 0.0 0.0
0.0 0.0 0.0 0.4 0.0 0.0 0.4 0.0 0.0 0.0
0.0 0.0 0.0 0.4 0.1 0.1 0.4 0.0 0.0 0.0
0.0 0.0 0.0 0.4 0.1 0.1 0.4 0.0 0.0 0.0
0.0 S   S   S   S   S   S   S   S   0.0
```

Ground Truth:

```
0.0 L L L L L L L 0.0
0.0 0.0 0.2 0.3 0.1 0.1 0.3 0.2 0.0 0.0
0.0 0.0 0.1 0.4 0.0 0.0 0.4 0.1 0.0 0.0
0.0 0.0 0.0 0.4 0.0 0.0 0.4 0.0 0.0 0.0
0.0 0.0 0.0 0.4 0.0 0.0 0.4 0.0 0.0 0.0
0.0 0.0 0.0 0.4 0.0 0.0 0.4 0.0 0.0 0.0
0.0 0.0 0.0 0.4 0.0 0.0 0.4 0.0 0.0 0.0
0.0 0.0 0.0 0.4 0.1 0.1 0.4 0.0 0.0 0.0
0.0 0.0 0.0 0.4 0.1 0.1 0.4 0.0 0.0 0.0
0.0 S   S   S   S   S   S   S   S   0.0
```

J.19 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
```

J.20 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
```