Tractatus-Eval: A Physics-Engine Validated Benchmark for Embodied Spatial Reasoning in Large Language Models

Tianjie Sun
alexflankerucsd@gmail.com
https://github.com/AlexFlanker/tractatus-eval

February 2026

Abstract

We present TRACTATUS-EVAL, a benchmark for evaluating embodied spatial reasoning capabilities in text-only large language models (LLMs). The benchmark comprises six physics-based tasksspatial navigation, key-lock puzzles, object stacking, container filling, collision prediction, and circuit connectivityeach generated at three difficulty tiers (Easy, Medium, Hard), yielding 9,000 validated samples across 18 subtasks. Unlike existing benchmarks that use hand-crafted or randomly generated distractors, TRACTATUS-EVAL employs deterministic *physics-engine validators* for each task, ensuring every wrong answer provably violates physical constraints and achieving a verified **0% distractor contamination rate**. We evaluate four model families (Pythia-410m through Phi-2) and find: (1) all models score at exactly ∼50% on binary prediction tasks regardless of scale, demonstrating no genuine physics simulation; (2) training data composition matters more than parameter count for embodied reasoning; and (3) counter-intuitive accuracy *increases* with difficulty on certain tasks reveal pattern matching rather than simulation. Our results quantify the *embodied cognition gap*the engineering cost of the philosophical limitation that Wittgenstein identified: "The limits of my language mean the limits of my world."

Keywords: embodied reasoning, spatial cognition, LLM evaluation, benchmark, physics simulation

1 Introduction

Modern large language models (LLMs) exhibit remarkable linguistic reasoning capabilities, yet consistently fail on tasks requiring *embodied spatial understanding*intuitions that any physically-situated agent acquires trivially through interaction with the world. Consider a simple instruction: "Navigate from A1 to E5, avoiding walls." A child solves this instantly. State-of-the-art LLMs, however, routinely generate paths that *pass through walls*, *teleport across obstacles*, or *step outside grid boundaries*behaviors that are physically impossible but textually "plausible" [Liu et al., 2023, Shridhar et al., 2020].

This gap between linguistic competence and physical-world understanding is not merely a failure of scale. It reflects a fundamental limitation identified by Wittgenstein in the *Tractatus Logico-Philosophicus*: "The limits of my language mean the limits of my world" (Proposition 5.6). A

text-only LLM has never *experienced* a wall. It possesses no sensorimotor grounding for the concept of "impassable"it can only pattern-match the *word* "obstacle" against its training distribution.

TRACTATUS-EVAL operationalizes this insight as a quantitative engineering benchmark. We make three contributions:

1. **A 6-task, 3-tier benchmark** covering distinct physical reasoning dimensions: grid pathfinding, state-dependent navigation, gravitational stability, volume conservation, trajectory prediction, and circuit analysis.

2. **Physics-engine validated distractors** each task employs a deterministic validator ensuring every wrong answer provably violates physical constraints, achieving 0% contamination.

3. **Comprehensive baseline evaluations** across 4 model families revealing three distinct task categories: genuinely hard, partially solvable, and unsolvable (binary).

2 Related Work

Spatial Reasoning Benchmarks. SpartQA [Mirzaee et al., 2021] and StepGame [Shi et al., 2022] evaluate spatial language understanding but rely on textual descriptions without physics constraints. CLEVR [Johnson et al., 2017] tests visual reasoning with synthetic scenes but targets vision-language models. BIG-Bench [Srivastava et al., 2022] includes spatial tasks (e.g., `navigate`) but uses hand-crafted examples without systematic distractor validation.

Embodied Evaluation. ALFWorld [Shridhar et al., 2020] and VirtualHome [Puig et al., 2018] evaluate embodied task completion in interactive environments but require agent infrastructure. TRACTATUS-EVAL distills the core reasoning challenge into a static multiple-choice format compatible with standard LLM evaluation pipelines.

Benchmark Contamination. Recent work has highlighted the problem of benchmark contamination [Jacovi et al., 2023, Yang et al., 2023]. TRACTATUS-EVAL addresses a distinct but related concern: *distractor contamination*, where ostensibly wrong answers are actually valid alternatives, penalizing models that reason correctly.

3 Benchmark Design

3.1 Overview

TRACTATUS-EVAL consists of six tasks, each testing a distinct dimension of physical reasoning. All tasks share a common architecture: procedural generation $\rightarrow$ deterministic ground truth $\rightarrow$ physics-validated distractors $\rightarrow$ JSONL output. Table 1 summarizes the six tasks.

3.2 Task Descriptions

Spatial Navigation. The model must find the shortest valid path from a start to a goal on an $N \times N$ grid with impassable obstacles. Ground truth is computed via A* search with Manhattan distance heuristic. Distractors include wall-phasing straight lines, random walks, reversed paths, and single-step mutations.

Table 1: The six tasks of TRACTATUS-EVAL.

Task	Physics Constraints	Validator	Difficulty Parameters
Spatial Navigation	Pathfinding, obstacle avoidance	`simulate_path()`	Grid: 4/5/7, Obstacles: 2/3/5
Key-Lock Puzzles	State-dependent behavior, inventory	`simulate_path()`	Grid: 4/5/7, Keys: 1/1-2/2-3
Object Stacking	Gravity, structural stability	`is_stable()`	Blocks: 3/4/6
Container Filling	Volume conservation, overflow	`simulate_step()`	Containers: 2/2-3/3-4
Collision Prediction	Trajectory intersection	`simulate()`	Grid: 4/5/7, Objects: 2/2/3
Circuit Connectivity	Topological reachability	Graph BFS	Grid: 4/5/7, Switches: 1/1-3/2-4

Key-Lock Puzzles. Extends spatial navigation with colored doors requiring matching keys. Solutions interleave movement actions with `pick_up` and `unlock` operations. Ground truth uses state-aware BFS over a (position, inventory) state space $\sim$25$\times$ larger than plain pathfinding.

Object Stacking. Given blocks of varying widths, the model must determine a stable bottom-to-top ordering where each block is fully supported: $\text{width}[i] \leq \text{width}[i-1]$. The validator `is_stable()` rejects any permutation violating this constraint.

Container Filling. Models receive 2–4 containers with capacities and initial levels, then execute a sequence of pour/fill/empty operations. Ground truth is computed by `simulate_step()`, which enforces $\min(\text{poured} + \text{current}, \text{capacity})$ at each stepexcess liquid overflows and is lost.

Collision Prediction. Two or more objects move at constant velocities on an $N \times N$ grid. The model predicts whether and when they collide (occupy the same cell at the same timestep). Ground truth is computed by deterministic step-by-step trajectory simulation.

Circuit Connectivity. An $N \times N$ grid contains batteries, bulbs, wires, and numbered switches. The model determines whether the bulb lights up by checking if a complete path exists from $+$ through wires and *closed* switches to the bulb and back to $-$.

3.3 Distractor Validation

A naive distractor engine can accidentally generate *alternate valid paths*paths that differ from the computed answer but still satisfy all physical constraints. Scoring these as "wrong" penalizes models that reason correctly. TRACTATUS-EVAL solves this with per-task physics validators:

$$\text{distractor } d \text{ accepted} \iff \exists \text{ constraint } c \in C_{\text{task}} : \text{violates}(d, c) \tag{1}$$

If a candidate passes all physical checks, it is silently discarded. Audit across all tasks confirms 0% contamination.

3.4 Difficulty Tiers

Each task is generated at three difficulty levels by scaling complexity parameters (Table 1, column 4). Each tier produces 500 samples, yielding $6 \times 3 \times 500 = 9{,}000$ total evaluation instances.

4 Experiments

4.1 Setup

We evaluate six models using the EleutherAI `lm-evaluation-harness` [Eval Harness, 2024] in 0-shot multiple-choice mode on Apple M5 (24GB, MPS). All tasks present 4 options (1 correct + 3 distractors). We report accuracy (`acc`) and length-normalized accuracy (`acc_norm`).

4.2 Results

Table 2: Average accuracy (%) on non-binary tasks (Spatial, Key-Lock, Stacking, Container) across difficulty tiers. Random baseline is 25%.

Model	Params	Easy	Medium	Hard	Avg
Pythia-410m	410M	21.9	23.4	25.2	23.5
Llama-3.2-1B	1B	29.4	31.6	33.8	31.6
Llama-3.2-3B	3B	33.5	37.5	38.8	36.9
Mistral-7B	7B	34.1	40.5	42.2	38.9
Llama-3-8B	8B	35.4	39.8	41.9	39.0
Phi-2	2.7B	40.2	41.5	48.0	**43.2**

Binary tasks are unsolvable. All six models score exactly ∼50% on Collision Prediction and Circuit Connectivity across all difficulty levels. This invariance to model scale (410M–8B) and task difficulty provides strong evidence that these models exploit binary surface cues (yes/no) rather than performing any form of physical simulation.

Training data composition outweighs scale. Phi-2 (2.7B) outperforms both 7B+ models on every task: Mistral-7B (38.9%) and Llama-3-8B (39.0%) both score below Phi-2's 43.2%. The two 7B+ models achieve nearly identical performance despite different architectures, suggesting a shared capability ceiling. This demonstrates that Phi-2's textbook-quality code/math training corpus provides more useful inductive biases for physical reasoning than parameter scaling alone.

Counter-intuitive difficulty trends. Container Filling accuracy *increases* with difficulty across all models (e.g., Pythia: 37.2% Easy → 47.6% Hard). We attribute this to longer prompts providing more arithmetic tokens for pattern matchinga calibration artifact rather than genuine reasoning improvement. Conversely, Phi-2's anomalous improvement on Object Stacking (30.4% → 47.8%) likely reflects memorized sorting heuristics from code training data activated by longer inputs.

5 Analysis

Our results reveal three distinct task categories (Figure **??**):

1. **Genuinely Hard** (Spatial, Key-Lock): Even the best model barely exceeds random chance (∼33%). Path tracing is fundamentally beyond token prediction.
2. **Partially Solvable** (Stacking, Container): Models can pattern-match arithmetic and sorting operations, reaching up to 75% on Container Hard. However, this reflects surface-level heuristics, not physical understanding.

3. **Unsolvable Binary** (Collision, Circuit): All models default to coin-flip behavior (∼50%). These tasks serve as *control conditions* proving that apparent reasoning on other tasks may also be pattern matching rather than simulation.

Implications for alignment. The embodied cognition gap measured by TRACTATUS-EVAL has practical consequences: an LLM advising a robot to "walk through the wall" is not making a linguistic errorit genuinely lacks the concept of physical impassability. This motivates preference-based alignment (DPO) using physics-validated contrastive pairs, or external guardrails that bypass the model's reasoning entirely.

6 Conclusion

TRACTATUS-EVAL provides a rigorous, physics-validated benchmark for measuring the embodied cognition gap in text-only LLMs. Our key contributions are: (1) a zero-contamination distractor validation methodology applicable to any physics-based benchmark; (2) empirical evidence that no current model performs genuine physics simulation, even on simple grid-world tasks; and (3) a difficulty-tiered evaluation framework that separates pattern matching from physical reasoning. We release all generators, datasets, and evaluation configs to facilitate reproducible research.

References

Gao, L., Tow, J., Abbasi, B., et al. A framework for few-shot language model evaluation. https://github.com/EleutherAI/lm-evaluation-harness, 2024.

Johnson, J., Hariharan, B., van der Maaten, L., et al. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*, 2017.

Liu, X., Yu, H., Zhang, H., et al. AgentBench: Evaluating LLMs as agents. In *ICLR*, 2024.

Mirzaee, R., Faghihi, H.R., Ning, Q., and Kordjamshidi, P. SpartQA: A textual question answering benchmark for spatial reasoning. In *NAACL*, 2021.

Puig, X., Ra, K., Boben, M., et al. VirtualHome: Simulating household activities via programs. In *CVPR*, 2018.

Shi, Z., Zhang, Q., Lipani, A. StepGame: A new benchmark for robust multi-hop spatial reasoning in texts. In *AAAI*, 2022.

Shridhar, M., Yuan, X., Côté, M.-A., et al. ALFWorld: Aligning text and embodied environments for interactive learning. In *ICLR*, 2021.

Srivastava, A., Rastogi, A., Rao, A., et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*, 2022.

Jacovi, A., Caciularu, A., Goldberger, O., et al. Stop uploading test data in plain text: Practical strategies for mitigating data contamination by evaluation benchmarks. In *EMNLP*, 2023.

Yang, S., Chiang, W.-L., Zheng, L., et al. Rethinking benchmark and contamination for language models with rephrased samples. *arXiv preprint arXiv:2311.04850*, 2023.