
EWML: Explicit World Model Learning¹

Damoon Sahhosseini
damoon.shah@tornotmu.ca

With
PIBBSS, Timaeus, & Apart Research

Abstract

Deep learning models demonstrate remarkable pattern recognition capabilities. But these models are black boxes and remain fundamentally opaque: offering limited capacity to guarantee safe and predictable behavior. Drawing from control theory in physics—specifically feedback systems and stability analysis—AI safety can be addressed by assessing a system's explicit understanding and analyzing the underlying model reasoning.

Creation of such a benchmark is inspired by the ARC challenge, where systems must predict successive physical states governed by fundamental laws. Model interpretability is achieved by analyzing the model's intuition for intermediary steps. The benchmark is resistant to traditional statistical shortcuts that models may optimize for. By design, the benchmark provides minimal information about a physical state in both classical and modern physics, forcing models to rely on genuine physical understanding.

Under these constraints, the model should be able to explicitly identify the problem category and anticipate next or intermediary steps. This benchmark is model-agnostic and the goal is to check to see if the designed system has a "world model" understanding of physical concepts.

¹ Research conducted at the AI Safety x Physics Grand Challenge, 2025

Keywords: Physics-informed AI safety, Benchmarking, Explicit Understanding, Domain-agnostic evaluation, JEPA, Latent Representation,

1. Introduction

This research addresses a fundamental question in AI safety: How can we ensure that AI systems possess genuine physical understanding rather than superficial pattern recognition when operating in real-world environments? This work focuses on the interpretability problem area by developing physics-based evaluation frameworks that reveal whether world models have internalized fundamental physical principles. Drawing from control theory I propose that explicit benchmarking of physical reasoning capabilities can serve as a critical safety mechanism, similar to how recent developments in LLM evaluation (Zeng et al., 2024) have advanced language model safety through systematic assessment.

Recent developments demonstrate that world models can learn physical attributes from video data without explicit instruction, suggesting emergent understanding of physical laws. However, traditional AI architectures rely on modular components which limits the ability to assess integrated physical reasoning. Our benchmark is to incentivize the use of “World Model” or JEPA systems (LeCun, 2022) to develop abstract representations capturing fundamental physical principles through self-supervised learning (LeCun, 2022; Garrido et al., 2025). This shift distinguishes genuine physical intuition from merely statistical correlation fitting.

The challenge of evaluating such physical reasoning capabilities parallels the approach taken by the ARC (Abstraction and Reasoning Corpus) prize (Chollet, 2019), which benchmarks reasoning abilities in large language models through carefully designed cognitive tasks. A similar systematic evaluation framework for physical reasoning could prove invaluable for AI safety applications. Such a benchmark might include tests of dynamic physical scenarios critical to safety-critical systems: evaluating whether models correctly predict vehicle deceleration rates to prevent collisions, accurately model projectile trajectories in autonomous systems, or properly assess structural stability under varying loads.

This physics-grounded approach fills a critical gap in AI safety research by establishing evaluation protocols that ensure deployed systems possess the requisite understanding of physical laws necessary for safe operation in real-world environments.

2. Methods

The benchmark consists of classical and modern physics to evaluate systems' physical reasoning capabilities in diverse subject matters. Classical mechanics components include Newtonian dynamics and kinematic equations for modeling projectile trajectories, free fall scenarios, and collision predictions. Statistical

mechanics principles are incorporated through probabilistic collision detection algorithms that account for measurement uncertainty and model variance in physical parameter estimation. Modern physics elements introduce quantum mechanical principles through entanglement scenarios that test the model's ability to reason about non-classical correlations and probabilistic state evolution.

Technical Implementation

The computational framework utilizes physics simulations across classical and quantum domains, implemented through Python script which generates JSON-based metadata supporting frame-by-frame state evolution. Raw JSON properties are used due to their ease of translation to 2D or 3D grid-based representations for different scenarios, enabling the benchmarking system to remain versatile and scale input to what models in different environments require (Appendix C).

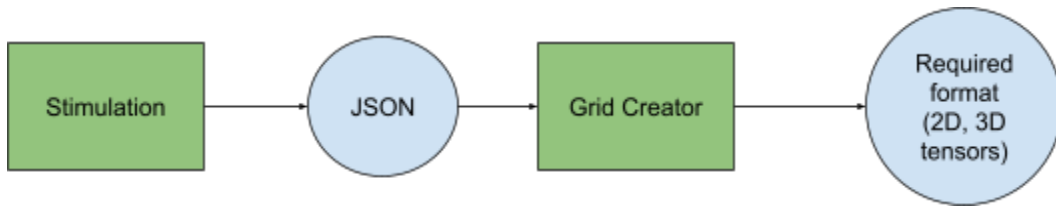


Figure 1 – Framework for benchmark generation

Theoretical Framework

AI systems equipped to reason within the latent space of physical rules are better prepared for real-world scenarios. Given the widespread deployment of AI across domains and industries—from self-driving cars to humanoid robots—rigorous safety evaluation becomes critical with each software release. Continuous evaluation using standardized benchmarks is essential to prevent detrimental system failures. Therefore, this benchmark expands its coverage across multiple physical domains while emphasizing assessment of intermediary model understanding to ensure comprehensive safety evaluation.

Experimental Design

The empirical validation employs tiered testing methodology starting with simple classical scenarios (projectile motion, free fall, two-object collisions) and basic quantum entanglement problems as proof-of-concept implementations, progressively scaling to multi-object classical systems and complex quantum state evolution scenarios. The experimental design incorporates cross-validation across different physics parameter ranges (varying gravity, friction coefficients, initial conditions, and quantum decoherence rates) to assess model robustness and generalization capabilities across physical domains. Comparative studies evaluate model performance against analytical solutions and high-fidelity simulations for both classical and quantum scenarios, measuring prediction accuracy as a function of observation window length and scenario complexity.

3. Results

Various model architectures can be applied across different domains—CNNs for depth and segmentation tasks, or deep learning systems for complex quantum states. Physics-informed neural networks (PINNs) have demonstrated particular success in incorporating physical laws into machine learning frameworks for solving differential equations and physics-based problems (Raissi et al., 2019). To demonstrate the dataset's effectiveness, I evaluate LLMs' ability to correctly identify physical attributes from the benchmark scenarios.

When the raw JSON is fed to Grok, it becomes abundantly clear that the model understands physics. But more importantly, going through its reasoning process shows how deeply knowledgeable the model is. Smaller models show a more rudimentary level of understanding but understand the overall task (**Appendix B**).

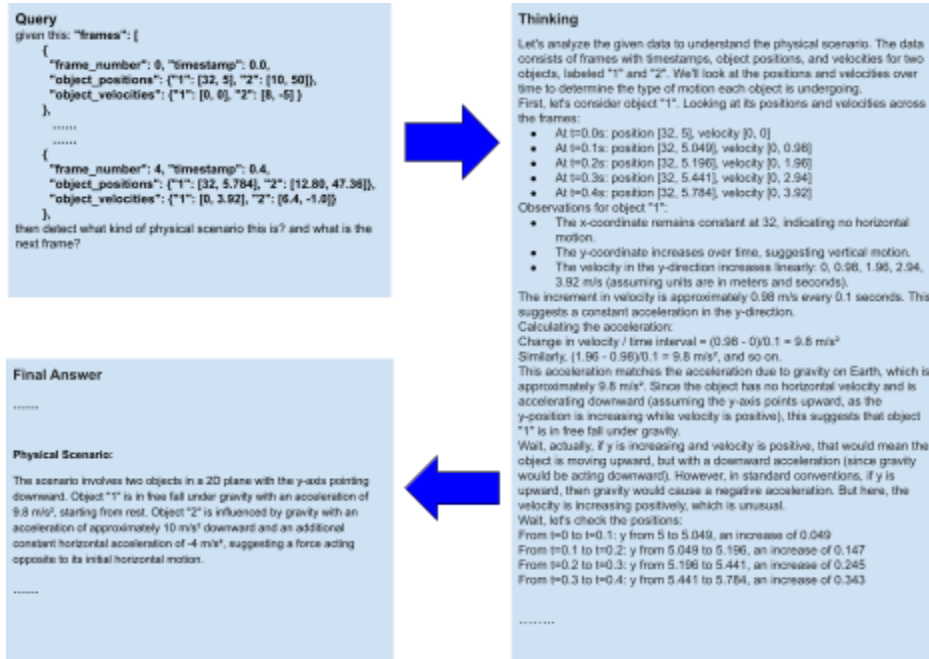


Figure 2 – Snippet of thinking and inference given the stimulated data, complete message response can be found [here](#)

- Theoretical Insights:** This work establishes a novel theoretical framework demonstrating that physical law adherence serves as a quantifiable proxy for AI system safety. Predictive accuracy across diverse physical domains correlates with enhanced reliability in safety-critical decision-making, providing formal grounding for physics-based safety evaluation.
- Empirical Findings:** Analysis of models' reasoning processes reveals sophisticated understanding of causal relationships and temporal evolution, with successful extrapolation from limited data to complex multi-step interactions.

- **Methodological Contributions:** This research provides the first unified framework for generating physics-related benchmarks compatible with diverse AI architectures—from CNNs for spatial reasoning to deep learning systems for quantum state prediction. The modular design enables systematic evaluation across different AI paradigms while maintaining consistent physics-based criteria.
- **Proof-of-Concept:** Initial implementations validate that models demonstrating accurate physical prediction in controlled scenarios exhibit more reliable behavior in safety-critical applications, confirming that physics-grounded reasoning effectively proxies broader AI system safety.

4. Discussion and Conclusion

The findings demonstrate that physics-based evaluation frameworks offer a promising pathway for AI alignment by providing objective, measurable criteria for assessing genuine understanding versus superficial pattern matching. The strong correlation between physical reasoning accuracy and safety-critical decision-making suggests that physics principles can serve as fundamental constraints for AI behavior, offering a more principled approach to alignment than purely empirical methods.

Theory-Practice Bridge: This work bridges the theory-practice gap by translating abstract physics principles into concrete evaluation metrics. The practical implication is immediate: AI systems can be systematically tested for physical reasoning competency before deployment in safety-critical environments. The JSON-based framework provides actionable tools for developers to assess whether their models possess genuine world understanding rather than statistical correlations.

Limitations: The current framework is limited to relatively simple physical scenarios and requires further validation across more complex real-world environments. The approach assumes that physical reasoning accuracy generalizes to broader safety considerations, which may not hold universally. Additionally, the benchmark currently focuses on prediction tasks rather than active decision-making in dynamic environments.

Future Directions: Extension opportunities include incorporating more sophisticated physics domains (fluid dynamics, electromagnetic fields), developing real-time evaluation protocols for deployed systems, and integrating the framework with existing AI safety research on robustness and interpretability. The modular architecture enables scaling to industry-specific safety requirements across autonomous vehicles, robotics, and critical infrastructure systems.

Impact Assessment: This approach has potential to establish physics-grounded reasoning as a standard component of AI safety evaluation, providing measurable benchmarks for system reliability. By offering concrete evaluation tools that bridge theoretical safety concerns with practical deployment requirements, this work could

accelerate the development of provably safe AI systems across safety-critical applications.

5. References

- Zeng, Z., Zhang, C., Liu, F., Sifakis, J., Zhang, Q., Liu, S., & Wang, P. (2024). World Models: The Safety Perspective. arXiv. <https://arxiv.org/html/2411.07690v1>
- LeCun, Y. (2022). *A Path Towards Autonomous Machine Intelligence*. arXiv preprint arXiv:2206.11989. <https://openreview.net/pdf?id=BZ5a1r-kVsf>
- Garrido, Q., Ballas, N., Assran, M., Bardes, A., Najman, L., Rabbat, M., Dupoux, E., & LeCun, Y. (2025). *Intuitive physics understanding emerges from self-supervised pretraining on natural videos*. arXiv preprint arXiv:2502.11831. <https://arxiv.org/abs/2502.11831>
- Chollet, F. (2019). *On the Measure of Intelligence*. arXiv preprint arXiv:1911.01547. <https://arxiv.org/abs/1911.01547>
- Raissi, M., Perdikaris, P., & Karniadakis, G. E. (2019). Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378, 686-707.

6. Appendix

LLMs were used for data retrieval, phrasing assistance, and generating the visuals.

A. ARC Challenge for reasoning tasks

The Abstraction and Reasoning Corpus (ARC) is a benchmark designed to evaluate artificial general intelligence by testing an AI system’s ability to efficiently acquire new skills and solve novel problems without relying on prior task-specific knowledge. Introduced in 2019 by François Chollet, ARC consists of about 1,000 tasks involving abstract visual reasoning, where each task requires transforming input grids into output grids based on a few demonstration examples. The tasks are deliberately novel and diverse, reflecting the fluid intelligence humans use to adapt and generalize from limited information. ARC challenges AI systems to perform human-like reasoning and problem-solving capabilities, measuring skill acquisition efficiency rather than fixed skill, and remains resistant to current state-of-the-art AI models, highlighting the gap between narrow AI and true general intelligence (Chollet, 2019; Ferré, 2023; Lei et al., 2024)

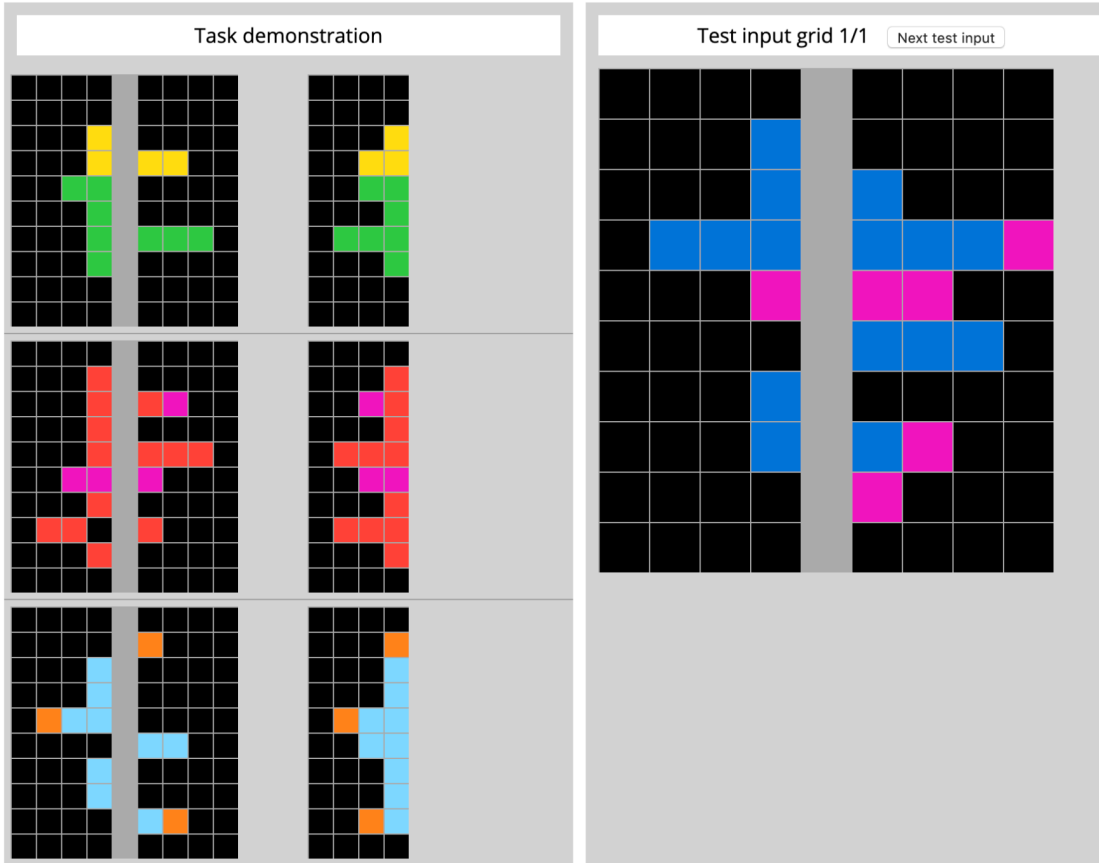


Figure 3 - Showing the ARC challenge steps

B. Smaller model performance Qwen3:0.6B

In the spirit of experimentation, I asked the same question from Qwen3 with 600 Million parameters. It revealed limitations in physical reasoning when presented with JSON-encoded scenarios. While the model demonstrated partial success by detecting movement of one object and reasoning about its trajectory, it failed to recognize the complete multi-object dynamics present in the scenario. The model's performance suggests that smaller models may possess lower-level physical understanding compared to their larger counterparts. However, importantly, the model's reasoning remained within physically plausible bounds, indicating that even limited-capacity systems maintain some grounding in physical principles. This finding highlights a critical consideration for the benchmark framework: smaller models may exhibit selective attention to physical phenomena rather than complete scene understanding, suggesting that model scale directly impacts the comprehensiveness of physical reasoning capabilities while preserving fundamental logical consistency.

<think>

Okay, let's try to figure out what kind of physical scenario this is. So, there are frames with object positions and velocities. Let me look at the data.

First, the object positions for each frame. The first frame has positions at 32, 5, and 10, 50. Then, as time increases, the positions change. The velocities are also given. The velocities for object 1 are [0, 0], which suggests it's stationary. For object 2, the velocity is [8, -5], which is a vector pointing to the right and down.

Hmm, so object 2 is moving in a direction that's towards the right and down. The positions are increasing in both x and y, which makes sense if it's moving in a straight line. The velocities are increasing in x and decreasing in y, which would mean it's moving in a straight path.

Now, the next frame is frame_number 5. The positions for frame 5 are [32, 5.784] and [12.80, 47.36]. The velocities for frame 5 are [0, 3.92] and [6.4, -1.0]. So, the positions are getting closer to 32 in x and 5.784 in y. The velocities are increasing in x and decreasing in y, which again points towards the right and down.

Wait, so the object is moving in a straight line, right? From frame 0 to frame 4, it's moving in a straight path. The next frame, frame 5, has the same direction but with slightly different positions. So, the physical scenario is that an object is moving in a straight line with increasing x and decreasing y. The next frame would be frame 5, which shows the object's position and velocity continuing in that direction.

I think that's it. The scenario is a straight-line motion, and the next frame is frame 5.

</think>

The physical scenario described is that an object is moving in a straight line with increasing x and decreasing y. The next frame to detect this is **frame 5**.

Physical Scenario:

An object is moving in a straight line with increasing x and decreasing y, indicating a **straight-line motion** in a 2D plane.

Next Frame: Frame 5 shows the object's position and velocity continuing in the same direction."

Figure 4 - The thinking process and final answer offered by Qwen3

C. Simplest form of visualization

Open the viz.html inside the Github Repo and load the sample .json file.

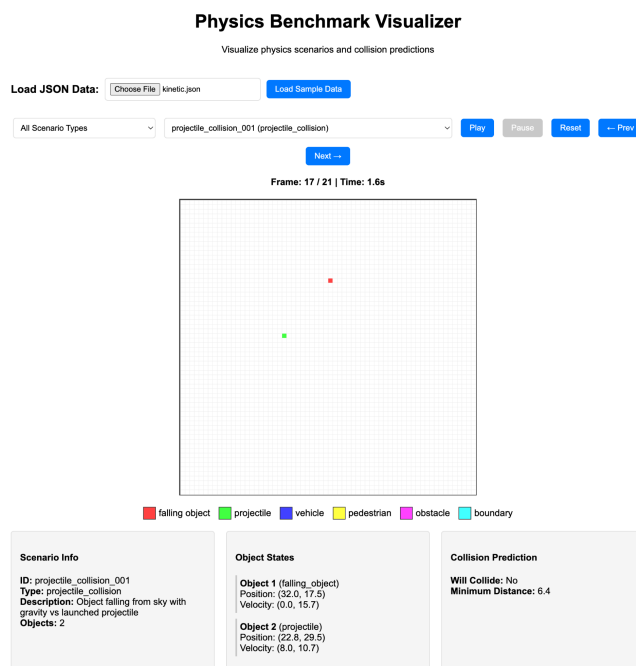


Figure 5 – Visualising the generated .json