Project Title

“Divide21X: A Benchmark for Faithful and Explainable Strategic Reasoning.”

Jacinto Jeje Matamba Quimua
October 2025

Abstract

My project is related to the domain of faithful and explainable AI, which is facing a "black box" crisis, because we can build AIs that win games, but we often don't know why they make the choices they do. I have created a benchmark called Divide21X, which is a small, discrete, rule-based system that allows us to precisely measure the fidelity of an AI's strategic explanation. This benchmark is derived from the Divide21 game, which I invented. Divide21X aims to solve the "fidelity vs. plausibility" problem, in which an AI is tested if it is telling the truth about its strategy as it plays Divide21. Moreover, it also examines how an AI responds to counterfactual explanations, by asking "what-if" questions, for example, after the AI plays a move A, it is then asked what if the move B was played instead? Divide21X provides a "unit test" for the core reasoning component of a symbolic AI.
I propose a set of seven evaluation metrics based on Divide21 game:

1. **Fidelity:** Does the AI's stated reason for a move (e.g., "I wanted to divide by 7") match the *actual* best move found in its internal search tree?
2. **Causality alignment:** Does the explanation reference the correct game state features (digits, divisibility, score)?
3. **Clarity:** Human judges or Large Language Model (LLM) based scoring for readability and coherence
4. **Outcome consistency:** Did the move produce the intended effect stated in the explanation?
5. **Counterfactual accuracy:** When presented with a "what if" scenario, can the AI correctly identify the game-theoretic outcome of the alternate move?
6. **Constraint awareness:** Can the AI correctly explain *why* a move is illegal, specifically citing the digit-history constraint?
7. **Strategic depth:** Does the AI's explanation only consider its own next move, or does it correctly identify an opponent's threat n moves later?

Project Relevance and Significance

My project is directly related to AI because it aims to evaluate the ability of LLMs to not only perform well at a specific task but most importantly to explain with clarity its reasoning. In this class we have learned that a LLM functions as an agent’s brain, which is made up of planning,

memory and tool use. My research I believe touches more on planning and reasoning, as well as In Context Learning (ICL).
Divide21X addresses a fundamental challenge in modern Artificial Intelligence (AI): the creation of Explainable, Symbolically-Grounded Intelligence. Current state-of-the-art AI, often based on Deep Reinforcement Learning (DRL), excels at complex games (like Go or Chess) but typically produces opaque, black-box strategies. Meanwhile, the game Divide21 offers a perfectly-suited, constrained environment to benchmark an AI's capacity for interpretable, symbolic reasoning. Divide21 game is defined entirely by explicit, formal mathematical rules (divisibility, factors, prime numbers, and digit manipulation). Unlike perceptual tasks, the optimal strategy must rely on verifiable number theory concepts. This makes the game inherently a test of Explainable AI (XAI): any successful AI agent's decision-making process should be translatable back into human-understandable mathematical proofs or logical rules (e.g., "The move was optimal because the current number, has a prime factor of 3, which is a losing state"). The problem is not merely about finding the optimal move, but about explaining the move in symbolic terms. Advancing AI toward Artificial General Intelligence (AGI) requires mastering domains with universal logical and symbolic structures, with mathematics being a natural progression. Divide21X acts as a crucial benchmark for Symbolic AI and Neuro-Symbolic AI architectures. The game’s questions about optimal play and computational complexity are rooted in discrete mathematics and combinatorial game theory, demanding an AI that can handle formal, rule-based systems rather than just pattern recognition.
Divide21X as an Explainable Symbolic AI Benchmark creates significant potential impact across research and educational communities. It allows researchers to move beyond qualitative discussions of interpretability and establish a quantitative benchmark. An AI's performance can be measured not just by its Win Rate, but by the fidelity and completeness of its generated symbolic explanation. A successful agent must not only win but also output the underlying number-theoretic proof for its sequence of moves. Furthermore, it can also improve existing methods for Formal Verification and automated theorem proving (ATP). If a successful AI can formally prove its strategy, it strengthens the tools used for verifying the correctness of complex mathematical and software systems. Divide21X also provides the research community with a standardized, accessible benchmark for comparing the symbolic reasoning and XAI capabilities of different models. This accelerates research by focusing on a common, mathematically grounded problem.

Novelty

Divide21X is a novel and innovative benchmark to accelerate research in Explainable Symbolic AI (XAI) and Neuro-Symbolic AI.
Historically, Symbolic AI has focused on games with transparent, formally defined rules and discrete state spaces, such as Go, Chess and Checkers. These systems relied on Symbolic AI techniques, primarily using exhaustive tree search algorithms like Minimax and Alpha-Beta Pruning, which are inherently logical and explainable. Early successes, such as IBM's Deep Blue, showed AI could master complex strategy games by evaluating a vast number of symbolic states. Furthermore, the Deep Reinforcement Learning (DRL), which is more associated with Pattern Recognition, mastered perceptually complex games where the state space is

ambiguous or immense, such as Atari video games, Go, StarCraft II, and Dota 2. DRL agents, like AlphaGo, use deep neural networks to learn strategies directly from raw pixel input or game board state representations. These methods excel at pattern-matching and intuition-like play. A growing body of work in Neuro-Symbolic AI and Explainable DRL (XRL) attempts to combine the power of neural networks with the precision of symbolic logic to address the limitations of DRL.

Despite significant achievements, current AI benchmarks fall short in demanding and verifying explainable, symbolic reasoning in a clean, mathematical domain:

- **The Black-Box Problem:** DRL agents, despite their superhuman performance in games like Go, rely on black-box neural architectures. Their strategies are opaque, hindering interpretability and trust. The rationale behind a decision cannot be easily translated into human-understandable logic.
- Lack of Pure Symbolic Benchmark: Most DRL benchmarks (Atari, etc.) primarily test visual perception and state-space evaluation. They do not explicitly require or benchmark the ability to discover and use formal, verifiable symbolic rules rooted in foundational mathematics.
- Explainability Measurement: While the field of XRL exists, there is a lack of standardized, quantitative benchmarks where the "correct explanation" for an optimal move is a formal, concise proof, forcing the AI to prove its logic, not just its outcome.

How Divide21X is New and Different:

- **Pure Symbolic Grounding:** Unlike DRL benchmarks, Divide21X's core mechanics are entirely based on number theory and discrete mathematics (divisibility, factors, and prime numbers). The optimal strategy must, by definition, be rooted in these symbolic, formal rules.
- **Forcing XAI:** Divide21X is a benchmark that forces explainability. A successful AI cannot just output the optimal move; it must output a verifiable, human-readable mathematical proof, which allows researchers to quantitatively measure the fidelity and completeness of the AI's symbolic explanation alongside its win rate.

Objectives & Deliverables

1. Develop and Publish a Symbolic AI Environment for the game Divide21 as Divide21X.
    a. Create a standardized environment for Divide21X where agents must both act and explain their reasoning.
        i. A Python Gymnasium-style API
        ii. Formal specification of state, action, and explanation trace formats (Python dict)
        iii. GitHub repository with documentation and examples
    b. Implementation of a Divide21 baseline agent (using Minimax with Alpha-Beta Pruning, MCTS, DRL, etc) capable of proving the optimal move.

2. Formal definition/implementation of the seven evaluation metrics and how they will be collected: Fidelity, Causality alignment, Clarity, Outcome consistency, Counterfactual accuracy, Constraint awareness and Strategic depth.
    a. Options:
        i. A technical paper, detailed documentation or paper chapter
        ii. An evaluation engine that parses and evaluates the metrics
3. Publish an open dataset of gameplay traces
    a. Collect, clean, and publish a dataset of Divide21 gameplay sessions containing actions, states, and move explanations, which may be valid or invalid for evaluation research.
4. Create a web-based leaderboard showcasing human and AI agents playing Divide21 with visible reasoning traces.

Scope & Feasibility

The Divide21X project will focus specifically on explainable symbolic reasoning within a controlled mathematical game environment.
The following elements are explicitly included:

- **Design and implementation of the Divide21X benchmark environment** — including state/action formalization and explanation-trace specification.
- **Development of baseline agents** that perform both symbolic decision-making and self-explanation.
- **Creation of a document or evaluation engine** that automatically checks the seven metrics collected.
- **Design of a web-based leaderboard** to visualize AI reasoning traces in real time and encourage participation from other researchers.

Out of Scope:

- **Perceptual or multimodal inputs** (e.g., vision or text comprehension) will *not* be included — the benchmark focuses purely on numerical and symbolic reasoning.
- **Training of large foundation models** is out of scope; the project will use pre-existing LLM APIs or lightweight fine-tuning on small models for efficiency.
- **Cross-domain generalization** (e.g., applying the same explainability methods to unrelated games or tasks) is not part of the initial release.

Resources planned to be used:

Resource Type	Description	Why Appropriate	Access Method
Programming Languages	Python (≥3.10)	Standard for AI, logic programming, and Gym environments	Open source

AI / ML Libraries	PyTorch, TensorFlow, OpenAI Gym, HuggingFace Transformers	Needed for RL and LLM agents	Open source
Symbolic Tools	SymPy, PyEEL, Prolog-Python interface	Provide symbolic reasoning and arithmetic verification	Open source
Compute Resources	Cloud GPU (Google Colab Pro / institutional GPU cluster)	Supports RL training and LLM inference	Available via institutional or cloud access
Hosting / Version Control	GitHub (code), HuggingFace Datasets (data), divide21.com (demo)	Enables open access, reproducibility, and community contributions	Existing project infrastructure
LLM Access	GPT-4/5 API or similar	For hybrid explainable agent prototypes and natural-language explanation generation	Paid API or academic license

Methods and Techniques

Environment Design

- Implement the Divide21X environment as a Python package compatible with the OpenAI Gym API.
- Define structured state representations and action types.
- Require agents to produce an explanation trace per move in JSON format, containing logical predicates and optional natural-language justifications.

Explanation Verification and Scoring

- Develop a symbolic verifier using SymPy or Prolog to evaluate logical consistency between an action, its explanation, and the resulting game state.
- Compute seven metric scores
- Aggregate results across turns and episodes to form benchmark leaderboards.

Agent Development

- (Minimax with Alpha-Beta Pruning, MCTS, DRL, etc)

Dataset Creation and Evaluation

- Collect gameplay logs from all baseline agents and self-play sessions.

- Store actions, states, explanations, and verification scores.
- Publish the dataset following FAIR principles (Findable, Accessible, Interoperable, Reusable).

Demonstration and Analysis

- Develop a web-based demo interface where humans can watch agents play with visible reasoning chains.
- Display leaderboard

Ethical Considerations

Divide21X as an Explainable Symbolic AI Benchmark introduces ethical considerations primarily related to bias, appropriate application, and the potential for over-reliance on the explanations generated by the system. Since the project does not involve human subjects, it does not require Institutional Review Board (IRB) approval.

Potential Bias Source	Description	Mitigation Strategy
Language-model explanations	LLM-based agents may produce explanations that are verbose, inconsistent, or stylistically biased (e.g., using overconfident language even when uncertain).	Use symbolic verifiers to ground all explanations in mathematical truth; score faithfulness higher than fluency.
Metric bias	Automatic metrics like “clarity” can favor certain writing styles or English proficiency.	Combine automated metrics with symbolic validation; clearly separate stylistic vs. factual correctness.
Training data bias in hybrid models	If pre-trained language models are used to generate explanations, they inherit biases from their training data.	Restrict generation to factual mathematical reasoning and audit sample outputs for neutrality.
Agent evaluation bias	Overfitting to benchmark metrics can lead models to “game” explanation scores instead of genuinely reasoning.	Design multi-factor scoring and randomize test cases to discourage metric gaming.
Misapplication / Over-trust in XAI	People might over trust the generated explanations, assuming fidelity guarantees the overall system’s intelligence or robustness. The system	Use different evaluation metrics. Clearly state in documentation that high fidelity on known

might have a high fidelity score but low generalization, for example it might explain a loss correctly but fail to find a win.	states does not guarantee optimal play on novel states.

The project is committed to establishing a transparent and rigorous benchmark. By grounding evaluation in symbolic truth (divisibility logic, numeric consistency), Divide21X minimizes subjective or linguistic bias.
Transparency is a core goal of this project, not just a requirement. All models and evaluation methods will be:

- Published and reproducible, with clear documentation of algorithms, scoring rules, and datasets.
- Explainability is a must, meaning that every AI decision must be accompanied by a verifiable reasoning trace.
- Publicly auditable, allowing other people to inspect how explanation scores are computed and whether they reflect actual logical validity.

This ensures that the benchmark cannot be used as a “black box” evaluator, because interpretability is built into its design.

Timeline

Stage	Milestone / Deliverable	Key Activities	Outputs / Checkpoints
Project Setup & Design	Milestone 1: Environment Design & Specification	• Finalize Divide21X formal specification (state, action, explanation trace schema). • Design logic rules for explanation verification. • Set up version control (GitHub) and documentation structure. • Draft ethics statement & data management plan.	- Specification document (JSON & Python class definitions). - Repository initialized with basic environment skeleton. - Ethics & responsible use statement completed.

Core Implementation	Milestone 2: Environment Prototype Ready	• Implement Divide21X environment (Gym-style API). • Add logic for digit manipulation, divisibility checks, and scoring. • Implement automatic logging of actions and explanations. • Test with baseline agents.	- Working environment capable of full game simulation. - Sample logs with valid/invalid explanations. - Unit tests for game logic.
Explanation Verifier & Scoring System	Milestone 3: Automated Explanation Evaluation Module	• Build evaluation engine. • Define and implement the seven scoring metrics. • Run verification on baseline logs. • Begin writing methods section for documentation/paper.	- Verified explanation scoring script. - Preliminary quantitative results for baselines. - Draft of evaluation methodology.
Agent Development & Training	Milestone 4: Baseline Agents Completed	• Minimax with Alpha-Beta Pruning, MCTS, DRL, etc. • Begin comparative analysis between agents.	- Trained DRL agents. - Comparison table: win rate vs. explanation fidelity. - Preliminary figures for paper/dataset.
Dataset Creation & Public Release Prep	Milestone 5: Divide21X Dataset v1.0	• Generate gameplay logs from all agents (self-play + evaluation). • Clean, validate, and standardize data. • Write metadata and documentation. • Prepare submission to an open repository (e.g., HuggingFace, Zenodo).	- Public dataset package (CSV/JSON). - Data card with metadata & license. - Verification results summary.

Demo, Evaluation & Final Reporting	Milestone 6: Final Demo & Publication Submission	• Build interactive web demo showing AI reasoning traces in real time. • Conduct final performance and explainability analysis. • Write and submit the final report/paper. • Disseminate benchmark publicly (GitHub + demo link).	- Working web demo at divide21.com/benchmark. - Full research report or paper draft. - Public release announcement and leaderboard page.

References

Aldeia, D., Correia, J., & Machado, P. (2022). Interpretability in symbolic regression: A benchmark of explanatory methods using the Feynman data set. Genetic Programming and Evolvable Machines, 23(3), 391–416. https://arxiv.org/pdf/2404.05908

Brandon-Colelough, B., & Arrieta, A. (2024). Neuro-Symbolic AI in 2024: A Systematic Review. In CEUR Workshop Proceedings (Vol. 3819, Paper 3). CEUR-WS. https://ceur-ws.org/Vol-3819/paper3.pdf

Ghosh, A., & Anand, A. (2024). A Neuro-Symbolic Benchmark Suite for Concept Quality and Reasoning Shortcuts (rsbench). arXiv preprint arXiv:2406.10368. https://arxiv.org/abs/2406.10368

Gonzalez-Santamaria, R., & Fischer, M. (2024). The KANDY Benchmark: Incremental Neuro-Symbolic Learning and Reasoning with Kandinsky Patterns. arXiv preprint arXiv:2402.17431. https://arxiv.org/abs/2402.17431

ISWC 2023. (2023). Benchmarking Symbolic and Neuro-Symbolic Description Logic Reasoners. Proceedings of the International Semantic Web Conference (ISWC 2023), Paper 385. https://iswc2023.semanticweb.org/wp-content/uploads/2023/11/ISWC2023_paper_385.pdf

Liu, X., Xu, F., & Li, Y. (2023). M4: A Unified XAI Benchmark for Faithfulness Evaluation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2023). https://par.nsf.gov/servlets/purl/10535816

Martins, P., Ribeiro, B., & Silva, A. (2023). Benchmarking eXplainable AI: A Survey on Available Toolkits and Open Challenges. In Proceedings of the 32nd International Joint Conference on Artificial Intelligence (IJCAI 2023) (pp. 6770–6777). https://www.ijcai.org/proceedings/2023/747

Matamba Quimua, J. J. (2025). Divide to One (divide21 or /21): A Strategic Game of Digit Manipulation and Divisibility. Unpublished manuscript. https://drive.google.com/file/d/1zDJArL9eLU93WFw4IDhB66akW1EN3FIR/view?usp=sharing

OpenAI. (2024). SymbolicAI: A framework for logic-based approaches combining generative models and solvers. arXiv preprint arXiv:2402.00854. https://arxiv.org/abs/2402.00854

Rebele, T., & Arrieta, A. B. (2023). How to Think About Benchmarking Neuro-Symbolic AI? In CEUR Workshop Proceedings (Vol. 3432, Paper 22). CEUR-WS. https://ceur-ws.org/Vol-3432/paper22.pdf

Sacha, D., & Samek, W. (2024). Narrative Review on Symbolic Approaches for Explainable Artificial Intelligence: Foundations, Challenges, and Perspectives. Applied Sciences, 14(3), 1193. https://doi.org/10.3390/app14031193