Qeios

Peer Review

Review of: "Geometric Analysis of Reasoning Trajectories: A Phase Space Approach to Understanding Valid and Invalid Multi-Hop Reasoning in LLMs"

Hong-Kun Zhang[1]

1. University of Massachusetts at Amherst, United States

This paper uses Hamiltonian mechanics for the first time to analyze reasoning in LLMs. It defines "kinetic energy" (cost of state transitions) and "potential energy" (how related each step is to the question). It builds a phase-space model, giving a new view on reasoning processes.

The authors mixed classical mechanics with AI reasoning; their idea of "energy balance" (kinetic vs. potential) is clever. They also use curvature and torsion (from differential geometry) to study reasoning paths. They show that valid reasoning has lower curvature (smoother) and torsion (less twisty). However, linking math concepts (like phase space) to neural networks feels forced. This part needs a clearer explanation.

1) In the Method section, energy plots (Figures 6–9) show valid reasoning clusters in lower energy bands. But the overlap between valid and invalid makes it hard to separate them perfectly.

2) Only the OBQA dataset is used. Tests on other domains (e.g., legal or medical texts) are needed.

3) Classification performance is imbalanced (valid recall = 0.33; invalid recall = 0.83), highlighting challenges in identifying high-quality reasoning (Table 5).

4) The term "phase space" is not well-explained. Readers unfamiliar with physics may struggle.

5) While energy/geometry metrics distinguish reasoning validity, their alignment with specific logical errors (e.g., fallacies) remains unclear. Future work should link geometric features to interpretable reasoning failures.

doi.org/10.32388/RIYOC6

6) The math is interesting, but the explanations are not very clear. More examples would help. It would be more helpful if the authors compared results with existing explainability methods (e.g., LIME, SHAP).

Despite limitations in data scope and physical analogies, its methodological rigor and empirical findings make it a landmark contribution to explainable AI. Future explorations in multimodal tasks and cognitive alignment could further amplify its impact.

Declarations

Potential competing interests: No potential competing interests to declare.