

Peer Review

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

Evgenii Arsentev¹

1. Independent researcher

1. What the work claims

The paper proposes to read a multi-hop reasoning chain as a trajectory in an embedding space and to describe that trajectory with the vocabulary of classical mechanics. Each step of a chain (a fact or the question) is embedded with bert-base-uncased (§4.4); position q is the embedding, momentum p is the difference between consecutive embeddings (§3.1.2), kinetic energy is $T(p) = \frac{1}{2}\|p\|^2$ (Eq. 19), potential energy is $V(q) = -\text{sim}(q, q_d)$ with cosine similarity (Eq. 20-21), and the reasoning Hamiltonian is $H_R = T - V$ (Eq. 18). On top of that, the author computes Frenet-Serret quantities (curvature, torsion, magnitude, angle) after PCA, action-angle variables, a set of "conservation" measures, trajectory entropy, and free energy. The data are reasoning chains built for the 500 OpenBookQA test questions with the QASC corpus (§4.1).

The central empirical claim, stated in the abstract and in §6.1, is that valid chains have lower Hamiltonian energy than invalid ones, and that valid chains follow smoother, less curved trajectories with "stronger conservation". The abstract itself is unusually candid: it says the connection to physics "remains largely metaphorical" and that any claim to steer or improve reasoning needs more validation. §6.2 repeats this. I appreciate that honesty, and I read the paper as a diagnostic proposal rather than a claim about LLM internals. Even on those modest terms, I found that several reported numbers cannot be produced by the equations as written, that the text gives opposite directions for the main effect, and that the classification results sit at or below a majority-class baseline.

2. Whether the conclusion holds on the data

The reported energies are outside the range the formulas allow. By Eq. 19, T is non-negative. By Eq. 20–21, $V = -\cos(q, q_d)$, so $-V$ lies in $[-1, 1]$. Then $H_R = T - V = \frac{1}{2}\|p\|^2 + \cos(q, q_d) \geq -1$ for every step. (The conventional $H = T + V$ is also bounded below by -1 here.) Yet the text reports energies of “ -9.5 to -9.0 ” for Figure 6, distributions centred “between -9.8 and -9.5 ” for Figure 9, and a “refined Hamiltonian energy” centred “between -60 and -70 ” for Figure 10. None of these is reachable under Eq. 18–21. Either the implementation uses a different potential (unnormalized dot product, a scale factor, a sum over steps with a different sign) or the figures plot something else. The “refined” Hamiltonian of Figure 10 is not defined anywhere I could find. Until the computed quantity is written down exactly, I cannot tell what the t-test in Table 1 is testing.

The direction of the main effect is stated both ways. The abstract and §6.1 say valid chains have lower energy. §5.1, in the paragraph that follows Table 1, says the opposite: “Valid reasoning chains generally show higher energy levels and more variability compared to invalid chains”, and in the next paragraph, “invalid chains mainly show lower energy profiles”. The paragraph on Figure 9 says valid chains form “a narrower, taller distribution”, which contradicts “more variability”. The paragraph on Figure 6 calls the difference between classes “ambiguous”. Table 1 gives t-statistic -6.5304 and p-value 0.0001 , but without the group order, the sign does not settle the question. This is the headline result of the paper, and the reader cannot currently tell which way it goes.

The potential term is defined against two different targets. In §3.1.2, q_d in Eq. 20 is “the embedding of the desired answer or goal state”. In §4.2.2 and §4.4, the potential is “the cosine similarity between q and the question embedding”. The difference matters a lot. If the answer embedding was used, and invalid chains are chains that do not lead to the correct answer, then V partly encodes the label, and a difference in H between classes would be close to circular. If the question embedding was used, that concern goes away, but Eq. 20 is wrong. The kinetic term has a similar small mismatch: Eq. 19 has the factor $\frac{1}{2}$, and §4.2.2 says “the squared magnitude of the momentum”.

How chains were produced and labelled is not described. §4.1 says the study focuses “on generating and annotating reasoning chains for the 500 test questions using the QASC corpus and our proposed technique”. I could not find how many chains were generated per question, how they were assembled from QASC facts, who or what decided that a chain is valid, or the class balance. The MANOVA in Table 2 has Den DF 994 with 3 numerator DF; for a two-group one-way MANOVA on three variables, that

corresponds to $994 + 3 + 1 = 998$ observations, which suggests about two chains per question, but this is my inference, not something the paper states. The confusion matrix in Table 4 sums to $148 + 31 + 14 + 7 = 200$, which would fit a 20% hold-out of roughly 1,000 chains. Neither the split nor the sampling is given.

The trajectories are not produced by a language model. The title speaks of "Multi-Hop Reasoning in LLMs" and §5.2 of "LLMs reasoning processes". As far as I can see from §4.2-§4.4, the trajectory is a sequence of BERT sentence embeddings of QASC facts selected for a question. No LLM generates the chain, and BERT is used as a frozen encoder (§4.4 says embeddings are taken from "BERT's last hidden layer"). The results, therefore, describe the geometry of fact sequences in BERT space, not the reasoning of a generative model. §4.2.1 also describes a fine-tuned BERT classifier for chain validity; its training data, epochs, and evaluation are not reported, and I could not tell whether Tables 4-5 evaluate that classifier or a classifier on trajectory features.

The classifiers do not beat the trivial baseline. I recomputed Table 5 from Table 4, reading rows as the true class and columns as the predicted class (the only reading that reproduces the report). Class False: precision $148/(148+14) = 0.914$, recall $148/(148+31) = 0.827$. Class True: precision $7/(7+31) = 0.184$, recall $7/(7+14) = 0.333$. Accuracy $(148+7)/200 = 0.775$. All match Table 5. The test set is 179 invalid and 21 valid chains, so predicting "invalid" for everything gives $179/200 = 89.5\%$. The reported accuracy of 0.78 is 11.5 points below that baseline, and macro F1 is 0.55 (Table 5, row macro avg). The logistic regression on PCA coordinates in Table 2 reports accuracy 0.9048; if the class balance is similar, that is about one point above always predicting "invalid". §5.1 calls this "high accuracy" and says it "suggests the combined discriminatory potential of these dimensions". On these numbers, it suggests close to none. The same table puts every Cohen's d between 0.0476 (PCA2) and 0.2134 (PCA3), which is consistent with that reading.

Significance is claimed at the margin without correction. Table 2: MANOVA for is_valid, Wilks' lambda 0.9933, F 2.2313, Pr > F 0.0830, which is not significant at 0.05. The per-dimension tests give p = 0.0800 (PCA1), 0.6701 (PCA2), 0.0496 (PCA3). With three tests after a non-significant omnibus test, a Bonferroni threshold is $0.05/3 = 0.0167$, and PCA3 does not pass it. Trajectory length gives p = 0.0390, while the text on Figure 16 says lengths have "similar length distributions" and that "path length doesn't reliably indicate validity". Table 3 reports p = 0.8856 for trajectory angle and p = 0.8052 for trajectory magnitude, yet the paragraph on Figure 17 describes valid chains as having "slightly higher magnitudes" and angles clustering near π . The prose describes differences the tests do not support.

Several conclusions in §6.1 have no reported test. §6.1 states that "valid reasoning follows smoother paths with lower curvature" and that "Invalid chains show higher curvature and torsion". I could not find any curvature or torsion statistic in Tables 1-3, and smoothness (Figure 16) is shown only as box plots, with the claim that it "significantly differs" but no test. The claim of "stronger conservation in valid chains" sits uneasily with Table 1, where the correlation between energy conservation and validity is -0.189 ; whether that means more or less conservation depends on the direction of the score, which is not defined (the paragraph on Figure 8 says the score is "generally obtained from the balance between kinetic energy ... and potential energy", which is not a formula).

The physics does not carry the weight put on it. A Hamiltonian is useful because Hamilton's equations (Eq. 1-2) generate the dynamics, and conserved quantities then follow from symmetries of H . Here the trajectory is a fixed list of embeddings; nothing evolves under H . So "the Hamiltonian is the most rigorously conserved" (§5.2, on Figure 14) is an empirical observation about a scalar computed along a sequence, not a consequence of Noether's theorem, and it does not by itself indicate a symmetry. With T and V as defined, low variance of H along a chain can also arise simply because all steps of a chain are sentences on the same topic, with similar norms and similar cosine to the target. A shuffled-order control, or chains of random QASC facts, would show how much of the "conservation" is a property of reasoning and how much a property of BERT embeddings of related sentences. Eq. 18 also writes $H = T - V$, which is the form of a Lagrangian; the text notes the similarity, but a reader from physics will expect $H = T + V$, and the choice should be justified or changed. The action-angle transformation in §5.1 (Figure 13) reports mean action 253.4054 for invalid and 229.8741 for valid chains ($t = -2.5152$, $p = 0.0121$), but the paper does not say how I was computed for a discrete, non-periodic sequence of a few points, so I cannot judge whether "action is almost constant" means anything beyond the construction.

The complexity estimate is not credible as stated. §5.4 and Figure 21 report $O(n^{0.43})$ in the number of trajectories and call this sub-linear scaling. Any method that computes a quantity for each of n trajectories must touch each one, so the true cost is at least linear in n . An exponent below 1 on a timing curve usually means fixed overhead (model loading, PCA fit) dominates at the sizes tried. This paragraph should be removed or replaced with per-trajectory timing.

Taken together: the idea of treating embedding sequences as trajectories and measuring their geometry is reasonable and could yield a useful diagnostic. The evidence presented, however, does not establish that valid and invalid chains differ in energy in a stated direction, and the geometric differences are weak, uncorrected, and in several places contradicted by the paper's own tables.

3. Reproducibility

The encoder is named precisely (bert-base-uncased, last hidden layer, §4.4), and the dataset is a public benchmark. That is where the specification ends. I could not find a code or data availability statement, a repository link, or the generated chains themselves. Because the chains are the object of study and were created by the author, without them nothing in §5 can be rerun.

Specifically missing: the pooling function (§3.1.1 lists mean pooling and [CLS] as options without saying which was used); the chain construction procedure from QASC, the number of chains, and the labeling rule for valid versus invalid; the train/test split and class counts; the fine-tuning setup of the BERT validity classifier in §4.2.1 (learning rate, epochs, seed, training set); the classifier behind Tables 4-5 and its features; the PCA settings (fitted on which set, whether before or after the split); the exact formulas for the energy conservation score, the "refined" Hamiltonian, smoothness, trajectory entropy, free energy, angular momentum, and the "energy-like quantity" of Figure 14, and the action variable I ; and the hardware and software used for the timing in Figure 21. No random seeds are reported for PCA, logistic regression, or the split.

There is also a factual slip in §4.1: OBQA is given as 5,957 questions with "4,957 for training, 500 for testing," but $4,957 + 500 = 5,457$. The missing 500 are the development split, which should be named. This does not affect the results but shows the setup was not checked against the source.

The paper does not use a generative LLM, so model snapshots and decoding parameters do not apply; that should itself be said clearly, given the title.

4. What specifically to fix

1. **Eq. 18-21 versus Figures 6, 9 and 10.** Write down the exact quantity plotted and tested. Under the stated definitions, H_R cannot fall below -1 , so the ranges " -9.5 to -9.0 ", " -9.8 to -9.5 ", and " -60 to -70 " need an explanation or a correction. Define the "refined Hamiltonian" of Figure 10 or drop it.
2. State the direction of the main effect once, consistently. The §5.1 sentences "Valid reasoning chains generally show higher energy levels" and "invalid chains mainly show lower energy profiles" contradict the abstract and §6.1. Give group means, standard deviations, n per group, and an effect size next to the t -statistic -6.5304 in Table 1.
3. **§3.1.2 (Eq. 20) and §4.2.2 / §4.4.** Say whether q_d is the answer or the question embedding. If it is the answer, discuss label leakage and add a version with the question as the target.

4. Add a subsection to §4.1 on how chains were built from QASC and labelled: count per question, total, valid/invalid split, the rule or annotator, and an example of each class.
5. **Tables 2, 4, and 5.** Report the majority-class baseline ($179/200 = 89.5\%$ on the Table 4 test set) next to every accuracy. Replace "high accuracy" in §5.1 with balanced accuracy, macro F1, or AUROC. Say which classifier and which features produced Table 4.
6. Correct for multiple comparisons in Table 2 and Table 3, and bring the prose in line with the results: the MANOVA ($p = 0.0830$) is not significant; angle and magnitude ($p = 0.8856, 0.8052$) show no difference.
7. Either add curvature, torsion, and smoothness tests or remove the corresponding §6.1 claims.
8. Add two controls to §5.2 and §5.1: chains with shuffled step order, and chains of random topic-matched QASC facts. If "conservation" and the energy gap survive in those controls, they are not about reasoning.
9. Justify $H = T - V$ in Eq. 18 or switch to $T + V$, and say plainly in §3.3 that the trajectories do not obey Hamilton's equations, so Noether-type conclusions in §5.2 are analogies.
10. Give formulas for the energy conservation score (Table 1, Figure 8), the Figure 14 quantities, entropy and free energy (Figure 20), and the action variable (Figure 13).
11. Remove the $O(n^{0.43})$ claim in §5.4 / Figure 21.
12. Retitle, or limit the LLM wording in the title, §5.2, and §6.2: what is analyzed is BERT embeddings of fact sequences. A version with chains generated by an actual LLM would be the natural next study.
13. §4.1: fix the OBQA split arithmetic ($4,957 + 500 \neq 5,957$) and name the development split. Add a data and code availability statement and release the chains with labels.
14. Shorten §1.1–§2.3 and §3.3. Much of it is textbook material on mechanics and the philosophy of mathematics that the experiments never use, while the parts the experiments do depend on (chain construction, score definitions) are missing.

5. Recommendation

Major revisions. The framing is original, and the author deserves credit for writing the metaphorical status of the analogy into the abstract. But the reported energies are inconsistent with the paper's own equations, the direction of the main finding is stated both ways in §5.1, the chain generation and labelling are undocumented, and the classifiers in Tables 2, 4, and 5 do not clear a majority-class baseline.

These problems can be fixed with the data the author already has, plus two simple controls. I would be glad to read a revised version.

Declarations

Potential competing interests: No potential competing interests to declare.