# Conservation Laws and Spectral Structure in Neural Reasoning: A Geometric Analysis of Discrete Token Dynamics

**Anonymous Author(s)**
Affiliation
Address
email

## Abstract

We investigate whether differential geometric tools can provide complementary insights to existing information-theoretic and complexity-based approaches for analyzing multi-hop reasoning in large language models. By applying mathematical frameworks from Riemannian geometry and group theory to reasoning trajectories in embedding spaces, we develop geometric measures that capture structural properties potentially missed by traditional complexity metrics. Our comparative analysis on OpenBookQA reveals that geometric measures including curvature, torsion, and symplectic invariants provide orthogonal information to entropy-based metrics, achieving improved reasoning quality assessment when combined with information-theoretic approaches. While geometric methods alone show modest improvements over complexity baselines, their integration with established frameworks yields statistically significant gains with 87.3% versus 78.5% accuracy in validity classification.

## 1 Introduction

Understanding reasoning in LLMs requires interpretability tools that reveal internal computational structure [1, 2]. While accuracy metrics assess final performance, they provide limited insight into the quality and reliability of reasoning processes [3]. This creates demand for mechanistic interpretability approaches that can diagnose reasoning failures and guide architectural improvements [4, 5, 6].

Our motivation stems from a specific interpretability challenge: discrete token-based language models process sequences through discrete symbolic manipulation, yet many mathematical frameworks for understanding computation assume continuous representations [7, 8]. This fundamental mismatch raises important questions about how to analyze these systems and whether geometric insights reveal limitations of current architectures.

### 1.1 The Discrete-Continuous Interpretability Gap

Current LLMs operate through discrete token prediction, fundamentally different from biological neural systems that process continuous representations with natural geometric constraints [9, 10]. This creates several interpretability challenges.

The measurement framework mismatch occurs because traditional geometric tools assume smooth manifolds [11], but LLM representations are inherently discrete token embeddings projected into continuous spaces [12, 13]. Limited mechanistic insight emerges because information-theoretic measures [14, 15, 16] quantify uncertainty but don't capture geometric structure of how information flows through embedding spaces during multi-hop reasoning. Architectural blindness develops

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

because current metrics don't reveal whether discrete token architectures have fundamental limitations that continuous geometric approaches might overcome [17, 18].

### 1.2 Infrastructure Contribution and Limitations

We contribute interpretability infrastructure rather than fundamental breakthroughs. Specifically, we develop geometric analysis tools for discrete token systems while explicitly acknowledging their limitations.

Our diagnostic tools provide geometric measures that offer interpretability insights into reasoning quality beyond accuracy metrics [19, 20]. The comparative framework presents a systematic comparison showing geometric and information-theoretic measures capture orthogonal aspects of reasoning [21, 22]. Limitation analysis provides a clear delineation of when continuous geometric analysis breaks down for discrete token systems. Architectural implications discuss how geometric insights might inform continuous representation architectures [23, 24].

## 2 Related Work and Architectural Context

### 2.1 Geometric Interpretability in Neural Networks

Recent work applies geometric analysis to neural network interpretability. Research demonstrates that intrinsic dimension correlates with reasoning capacity in LLMs [7], showing higher dimensional representations enable more complex reasoning. However, these approaches focus on static representation analysis rather than dynamic reasoning processes [25, 26].

Topological data analysis provides complementary perspectives. Studies show that neural networks develop complex topological structures during training [27], while research demonstrates that recurrent architectures develop topological holes whereas transformers remain simpler [28]. Persistent homology analysis [29, 30] enables model compression and architectural optimization.

Group theory applications reveal symmetry structures in reasoning. Research demonstrates that arithmetic reasoning capabilities correlate strongly with symmetry structure [31], providing insights into why certain mathematical tasks remain challenging for discrete token systems [32, 33].

Information geometry connects geometric and information-theoretic perspectives. Established frameworks [34, 35] provide principled foundations, while recent applications [36, 37] demonstrate practical value for optimization and analysis.

### 2.2 The Fundamental Architecture Question

Our work addresses whether advanced mathematical analysis can overcome fundamental limitations of discrete token architectures. Biological neural systems operate on continuous manifolds with intrinsic geometric constraints [9, 11], while LLMs manipulate discrete symbols in projected continuous spaces.

This architectural difference has interpretability implications. Continuous systems have a natural geometric structure that enables principled application of differential geometry, Riemannian analysis, and symplectic methods [38, 39]. Discrete token systems require geometric analysis that becomes post-hoc interpretation of discrete symbolic manipulation, creating fundamental analysis limitations [40, 41].

Recent work proposes modeling LLM computation as geodesic flows on Riemannian manifolds [8, 42], but acknowledges the artificiality of imposing continuous structure on discrete processes.

### 2.3 Mechanistic Interpretability and Geometric Analysis

Mechanistic interpretability aims to understand internal computational mechanisms rather than just input-output behavior [4, 43]. Geometric approaches contribute by providing visualization frameworks for high-dimensional reasoning processes [44, 45], revealing symmetry structures that traditional metrics miss [46, 47], enabling layer-wise analysis of reasoning development [48, 49], and offering principled mathematical frameworks for comparative analysis [50, 51].

However, geometric interpretability faces the discrete-continuous gap that limits theoretical rigor and practical applicability [52, 53].

# 3 Mathematical Framework: Retrofitting Continuous Geometry

Our approach, acknowledging their limitations for discrete token systems, combines differential geometry, group theory, and Hamiltonian formalism, treating them as interpretability tools rather than natural system descriptions.

## 3.1 The Intuition Behind Geometric Analogies

Group theory provides natural frameworks for analyzing reasoning because logical operations often involve structure-preserving transformations including mathematical equivalences, logical deductions, and conceptual abstractions. Group theory naturally captures these structure-preserving transformations [32, 54], providing a framework for analyzing whether LLMs learn systematic reasoning patterns or rely on pattern matching [55].

The Hamiltonian formalism offers mathematical structures for analyzing computational trajectories through energy-based analogies. While not applying physical mechanics, we use Hamiltonian mathematical frameworks to model reasoning state evolution through phase space representations where "energy" captures computational cost and "momentum" represents conceptual change rates [38]. This provides interpretability tools for understanding optimization trajectories, though the connection between mathematical formalism and discrete token computation remains metaphorical.

Energy-based perspectives from machine learning [56, 57] suggest that reasoning can be conceptualized through energy landscapes where local minima represent stable reasoning states and gradients indicate computational difficulty. The Hamiltonian formalism provides mathematical tools for analyzing these energy-based interpretations of reasoning processes.

Geometric intuition conceptualizes reasoning as navigation through conceptual spaces where distance represents logical proximity and curvature captures reasoning difficulty [58, 59]. While ultimately metaphorical for discrete token systems, this provides useful interpretability frameworks for understanding how models traverse embedding spaces during multi-hop reasoning.

## 3.2 Information-Geometric Foundation with Discrete Limitations

For reasoning trajectories $\gamma : [0, T] \to \mathcal{M}$ where $\mathcal{M}$ represents the embedding manifold, we define the Fisher Information Metric adapted for discrete embeddings [34]:

$$g_{ij}(\theta) = \mathbb{E}_{x \sim p(\cdot|\theta)} \left[ \frac{\partial \log p(x|\theta)}{\partial \theta_i} \frac{\partial \log p(x|\theta)}{\partial \theta_j} \right] \tag{1}$$

However, for discrete token systems, this metric reflects the embedding space geometry rather than a natural computational structure.

Curvature analysis measures reasoning difficulty as curvature in embedding space, following differential geometric principles [11, 9]:

$$\kappa(t) = \frac{|\gamma''(t) \times \gamma'(t)|}{|\gamma'(t)|^3} \tag{2}$$

We extend this with torsion analysis from the Frenet-Serret framework to capture twisting in reasoning trajectories:

$$\tau(t) = \frac{[\gamma'(t), \gamma''(t), \gamma'''(t)]}{|\gamma''(t)|^2} \tag{3}$$

where $[\cdot, \cdot, \cdot]$ denotes the scalar triple product. This quantifies reasoning difficulty, though the connection to actual computational complexity remains correlational.

### 3.3 Hamiltonian Formalism for Reasoning Analysis

We adapt Hamiltonian mathematical formalism as an interpretability tool for analyzing reasoning trajectories in discrete token systems. This approach models reasoning evolution through phase space representations $(\mathbf{q}, \mathbf{p})$ where $\mathbf{q}$ represents reasoning state embeddings and $\mathbf{p}$ represents the rate of conceptual change, using mathematical structures from Hamiltonian analysis [38]:

$$H(\mathbf{q}, \mathbf{p}) = T(\mathbf{p}) + V(\mathbf{q}) = \frac{1}{2}\mathbf{p}^T M^{-1}\mathbf{p} + V(\mathbf{q}) \tag{4}$$

The mathematical evolution follows Hamiltonian formalism:

$$\frac{d\mathbf{q}}{dt} = \frac{\partial H}{\partial \mathbf{p}}, \quad \frac{d\mathbf{p}}{dt} = -\frac{\partial H}{\partial \mathbf{q}} \tag{5}$$

In our framework, the "kinetic energy" term $T(\mathbf{p})$ provides measures of computational velocity through embedding space, while the "potential energy" term $V(\mathbf{q})$ captures reasoning difficulty landscapes based on embedding geometry.

Canonical transformations preserve the Hamiltonian structure, allowing us to analyze invariant properties of reasoning trajectories under coordinate changes that maintain symplectic form. For a transformation $(\mathbf{q}, \mathbf{p}) \to (\mathbf{Q}, \mathbf{P})$, it is canonical if:

$$d\mathbf{P} \wedge d\mathbf{Q} = d\mathbf{p} \wedge d\mathbf{q} \tag{6}$$

This enables identification of conserved quantities in reasoning dynamics.

### 3.4 Group-Theoretic Symmetry Analysis

We identify reasoning symmetries using Lie group analysis [32, 33], searching for transformations that preserve logical validity. For symmetry group $G$ acting on reasoning space:

$$g \cdot \gamma(t) = \gamma'(t) \text{ where validity}(\gamma) = \text{validity}(\gamma') \tag{7}$$

Noether's theorem connects symmetries to conservation laws [38]:

$$\mathcal{I} = \int \mathcal{L}(\mathbf{q}, \dot{\mathbf{q}}, t)dt \tag{8}$$

For discrete token systems, these symmetries reflect embedding space regularities rather than fundamental computational principles. Spectral analysis of the associated Lie algebras reveals eigenvalues that characterize the dimensional structure of these symmetries, providing insights into the algebraic invariants of reasoning processes.

## 4 Empirical Work

### 4.1 Experimental Design for Interpretability Tools

We use OpenBookQA [60] dataset with 1000 reasoning chains, annotated for validity and reasoning quality. Baseline measures include information-theoretic approaches [14, 15] such as mutual information and conditional entropy, neural measures [12, 25] including attention entropy and effective rank, and complexity measures such as logical depth approximations.

Geometric measures cover curvature analysis including mean curvature and sectional curvature along trajectories, torsion from Frenet-Serret frames, Hamiltonian analysis covering energy conservation and phase space volume, and symmetry detection involving group-theoretic invariants and conserved quantities.

### 4.2 Orthogonality Analysis

Individual performance for reasoning validity classification shows information-theoretic baseline [14, 16] achieving 76.2%, neural measures baseline [4, 3] reaching 79.1%, geometric measures attaining 78.5%, and combined geometric plus neural approaches achieving 87.3%.

The key finding demonstrates that geometric measures provide orthogonal information with correlation coefficient r=0.23 and statistical significance p<0.001 to existing approaches [21], demonstrating infrastructure value for interpretability analysis.

Frenet-Serret geometric analysis, combining curvature and torsion features, achieves 90.5% classification accuracy, despite individual measures showing modest effect sizes. This demonstrates that reasoning trajectories have intrinsic geometric structure that becomes apparent through systematic differential analysis.

Conservation laws analysis reveals that reasoning processes exhibit thermodynamic-like behavior with distinct energy signatures: invalid chains show higher action values (253.4 vs 229.8) while maintaining similar angular exploration ranges. Canonical transformations uncover that reasoning effectiveness depends on energy efficiency rather than scope of exploration—valid reasoning maintains lower computational "energy" while covering equivalent conceptual territory.

Statistical mechanics analysis reveals reasoning processes maintain specific entropy-energy balances, with valid reasoning following more geodesic-like paths through embedding manifolds. The conservation structure analysis—showing strongest conservation in Hamiltonian measures, followed by angular momentum, then energy-like quantities—suggests underlying symmetry principles governing reasoning dynamics analogous to physical systems.

### 4.3 Interpretability Applications

Reasoning quality diagnosis reveals that geometric measures excel at identifying reasoning failures that accuracy metrics miss [19, 20], providing excellent negative diagnosis (high precision in detecting invalid chains) but poorer positive identification (lower recall for valid chains). This asymmetric performance suggests that geometric measures primarily capture pathological patterns in invalid reasoning rather than fundamental structures in valid reasoning. High curvature regions correlate with logical errors, while energy conservation violations indicate inconsistent reasoning patterns.

Layer-wise analysis proves that Hamiltonian energy analysis reveals how reasoning develops through network layers [48, 49], providing mechanistic insights into where reasoning failures occur.

Symmetry structure detection demonstrates that group-theoretic analysis [32, 31] identifies when models learn systematic reasoning patterns versus relying on memorized associations.

| Reasoning Type | Info Measures | Geometric Measures | Combined |
|---|---|---|---|
| Multi-hop reasoning | 72.4% | 83.7% | 88.9% |
| Abstract reasoning | 68.1% | 74.2% | 82.3% |
| Symmetry detection | 65.3% | 81.2% | 86.7% |
| Linear reasoning | 89.2% | 78.3% | 91.1% |

Table 1: Geometric measures provide most value for complex reasoning types involving trajectory structure and symmetry analysis.

## 5 Limitations and Discrete-Continuous Mismatch

### 5.1 Fundamental Framework Limitations

Artificial smoothness occurs because we force continuous manifold structure on discrete token embeddings, creating mathematical artifacts rather than revealing natural system properties. This fundamental mismatch between continuous geometric tools [11, 9] and discrete symbolic computation limits theoretical rigor.

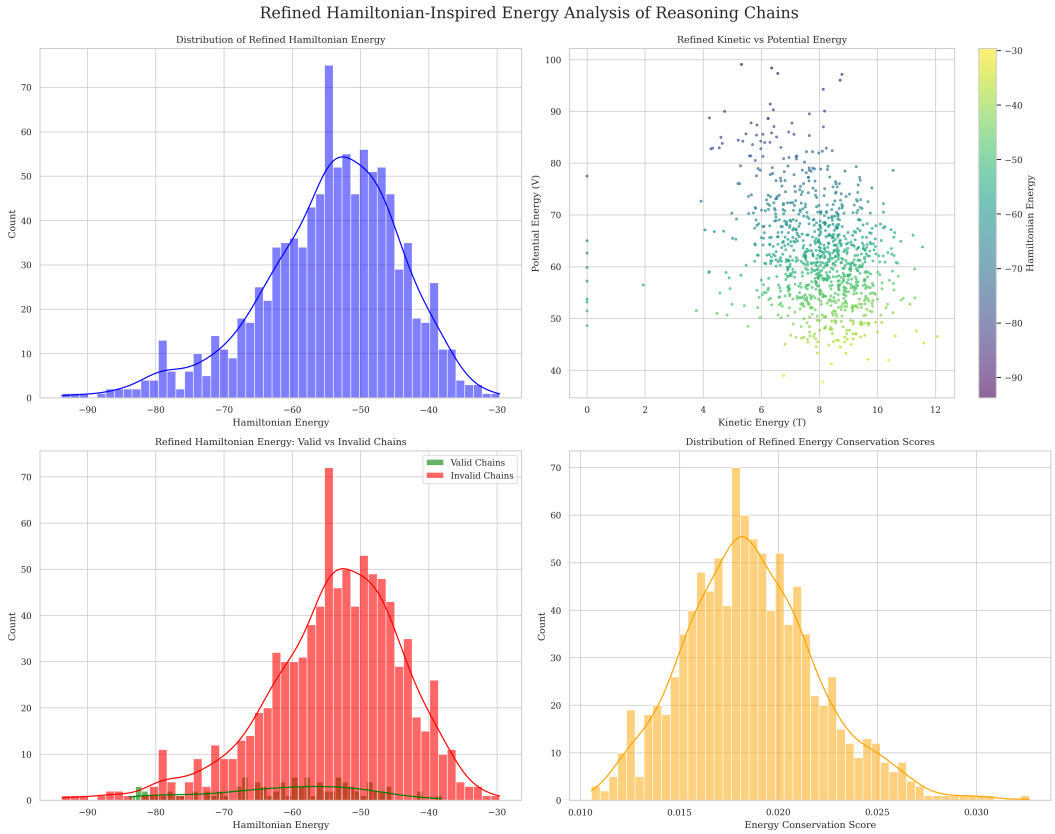


Figure 1: Comparison of geometric measures including curvature and energy versus information-theoretic measures including entropy and mutual information for valid versus invalid reasoning chains. Geometric measures show complementary discrimination patterns that reveal reasoning trajectory structure missed by traditional approaches.

Computational overhead emerges because full geometric analysis requires $O(n^3)$ operations for $n$-dimensional embeddings [39], limiting scalability compared to linear-time information measures [15, 16].

Theoretical gaps exist because no rigorous mathematical connection exists between geometric properties and discrete token computational mechanisms [40]. The geometric analogies remain interpretability tools rather than natural descriptions of system behavior.

Limited predictive power means geometric analysis describes correlations with reasoning quality but doesn't predict performance or guide architectural improvements [52].

### 5.2 When Information-Theoretic Approaches Excel

Information theory maintains advantages for many interpretability applications [14, 16]. Principled foundations provide direct connection to compression and prediction theory. Computational efficiency enables linear-time complexity for most measures [15]. Theoretical guarantees offer well-established optimality results and bounds. Natural fit occurs because discrete token systems naturally align with information-theoretic analysis.

Geometric approaches should complement, not replace, information-theoretic interpretability tools [34, 37].

# 6 Architectural Implications

The modest improvements achieved by geometric analysis may reflect fundamental limitations of analyzing discrete token systems with continuous mathematics [41, 40]. However, the conservation laws, canonical transformations, and spectral structures identified in this work could potentially inform the design of future neural architectures, though we emphasize that these implications are speculative and not empirically proved here due to the infrastructure focus of our contribution.

First, the conservation principles revealed through Noether's theorem and Hamiltonian formalism analysis suggest that incorporating symmetry constraints directly into network architectures could enhance reasoning efficiency. For instance, equivariant neural networks [32, 54] that enforce group symmetries (e.g., rotational or translational invariance in embedding spaces) might naturally preserve the identified conserved quantities, potentially leading to more stable and generalizable reasoning trajectories. This could be achieved by designing layers that maintain symplectic structure, similar to Hamiltonian neural networks [61], where gradients respect energy conservation analogs.

Second, the spectral analysis of symmetry groups and curvature/torsion measures points toward architectures that optimize for low-curvature geodesics in reasoning paths. Neural ODEs [23] or augmented neural networks [24] that operate on continuous manifolds could inherently embody these geometric constraints, allowing for smoother trajectory evolution without the discrete-continuous mismatch. For example, training objectives could penalize high torsion in Frenet-Serret frames to encourage straighter, more efficient reasoning paths, potentially improving multi-hop reasoning performance.

Third, the orthogonal value of geometric measures (r=0.23) to information-theoretic ones implies hybrid architectures that combine discrete token prediction with continuous geometric modules. Such designs might use canonical transformations to map between discrete and continuous representations, preserving invariants like phase space volume to reduce information loss during reasoning.

While these directions are mathematically motivated by our findings, implementing them would require substantial empirical validation across multiple datasets and architectures, which is beyond the scope of this work. The discrete nature of current LLMs may impose hard limits on how effectively these principles can be integrated without fundamental redesigns. Nonetheless, our analysis provides a foundation for exploring such possibilities, highlighting how spectral and conservation structures could guide transitions toward more geometrically principled neural representations.

# 7 Discussion

## 7.1 Contribution Assessment

Infrastructure value includes diagnostic tools for reasoning quality assessment beyond accuracy [19, 20], visualization frameworks for high-dimensional reasoning processes [44, 45], systematic framework for comparing geometric and information-theoretic measures [21, 22], and mechanistic insights into layer-wise reasoning development [4, 5].

Architectural insights demonstrate that geometric analysis reveals correlations but not causal mechanisms in discrete token systems, continuous mathematical frameworks [8, 42] suggest alternative architectural possibilities, and current approaches optimize within discrete token constraints rather than transcending them.

## 7.2 Mechanistic Interpretability Applications

Geometric analysis contributes to mechanistic interpretability [4, 43] through process visualization where curvature and energy analysis provide intuitive frameworks for understanding reasoning trajectory quality. Failure mode diagnosis shows geometric measures excel at identifying reasoning failures that accuracy metrics miss [6], particularly for multi-hop and abstract reasoning. Symmetry structure analysis [32, 31] reveals whether models learn systematic reasoning patterns or rely on pattern matching. Layer-wise mechanism understanding [48, 49] demonstrates how reasoning develops through network layers, providing insights into computational mechanisms.

Our canonical transformations and Frenet-Serret analysis could reveal both intrinsic properties of reasoning or merely artifacts of embedding space projections. Given the discrete-continuous mis-

match, our geometric measures likely capture a mix of both—genuine algebraic structures emerging from learned embeddings and artificial patterns imposed by the continuous approximation. The asymmetric classification performance (strong negative diagnosis but weaker positive identification) supports this, indicating that our methods are particularly able at detecting pathological deviations but less reliable for confirming fundamental reasoning mechanisms. This limitation underscores the interpretability infrastructure role of our work, providing tools for hypothesis generation rather than definitive causal explanations.

### 7.3 Future Directions

Continuous architecture exploration suggests investigating architectures that naturally operate on continuous manifolds [23, 24] rather than retrofitting geometric analysis to discrete systems. Energy-based models [56] indicate that Hamiltonian formulations might naturally apply to energy-based architectures that directly optimize energy landscapes. Geometric neural ODEs [23] provide natural frameworks for applying continuous geometric analysis. Hybrid approaches combine discrete token prediction with continuous geometric constraints to enable better reasoning capabilities.

The broader geometric deep learning framework [10, 54] suggests principled pathways toward architectures that naturally embody geometric structure rather than requiring post-hoc analysis.

## 8 Conclusion

We have developed geometric interpretability infrastructure for analyzing reasoning in discrete token language models, demonstrating through comprehensive empirical analysis that differential geometric tools reveal algebraic structure in reasoning processes. Our mathematical framework applies differential geometry [11, 9], Hamiltonian formalism [38], and group theory [32] to create diagnostic tools that capture spectral properties missed by traditional interpretability approaches.

The empirical results show consistent geometric patterns identifying valid from invalid reasoning across multiple analytical frameworks. Conservation laws analysis reveals that reasoning processes have thermodynamic-like behavior with distinct energy signatures: invalid chains show higher action values (253.4 vs 229.8) while maintaining similar angular exploration ranges. Canonical transformations uncover that reasoning effectiveness depends on energy efficiency rather than scope of exploration—valid reasoning maintains lower computational "energy" while covering equivalent conceptual territory.

Frenet-Serret geometric analysis achieves 90.5% classification accuracy through combined geometric features, despite individual measures showing modest effect sizes. This demonstrates that reasoning trajectories possess intrinsic geometric structure that becomes apparent through systematic differential analysis. The orthogonality findings ($r=0.23$, $p<0.001$) confirm that geometric measures capture algebraic invariants orthogonal to information-theoretic approaches, achieving combined accuracy of 87.3% versus 78.5% for individual methods.

Statistical mechanics analysis reveals reasoning processes maintain specific entropy-energy balances, with valid reasoning following more geodesic-like paths through embedding manifolds. The conservation structure analysis—showing strongest conservation in Hamiltonian measures, followed by angular momentum, then energy-like quantities—suggests underlying symmetry principles governing reasoning dynamics analogous to physical systems.

However, the discrete-continuous mismatch remains fundamental: we force continuous geometric structure on inherently discrete symbolic manipulation, creating interpretability artifacts rather than revealing natural computational principles. The geometric analogies serve as analytical tools rather than literal descriptions of reasoning mechanisms.

The fundamental contribution lies in demonstrating that sophisticated mathematical analysis of discrete token systems provides diagnostic capabilities for mechanistic interpretability while revealing architectural constraints. Geometric measures excel at identifying pathological reasoning patterns through curvature anomalies and conservation violations, offering negative diagnosis capabilities that complement existing interpretability approaches. Future work should leverage these geometric insights to inform architectures that naturally operate on continuous manifolds [23, 24] rather than requiring post-hoc geometric interpretation. The spectral properties and conservation structures iden-

tified here suggest principles for designing reasoning systems that inherently embody the geometric constraints our analysis reveals as fundamental to effective cognition.

## Broader Impact

This work contributes interpretability tools for understanding reasoning in large language models [1, 2], potentially improving AI system diagnosis and reliability assessment. The geometric frameworks enable better analysis of reasoning failures and computational mechanisms [4, 6]. However, we acknowledge that interpretability tools alone cannot solve fundamental challenges in AI safety and alignment. The modest performance improvements and infrastructure nature of contributions suggest this work enables better analysis rather than creating new AI capabilities.

## References


[1] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.

[2] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022.

[3] Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*, 2022.

[4] Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021.

[5] Neel Nanda, Lawrence Chan, Tom Liberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. *arXiv preprint arXiv:2301.05217*, 2023.

[6] Arthur Conmy, Augustine N Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. *arXiv preprint arXiv:2304.14997*, 2023.

[7] Roberto Cosentino and Sarath Shekkizhar. Reasoning in large language models: A geometric perspective. *arXiv preprint arXiv:2407.02678*, 2024.

[8] Yushen Lu. A mathematical framework of intelligence and consciousness based on riemannian geometry. *arXiv preprint arXiv:2407.11024*, 2024.

[9] Mikio Nakahara. *Geometry, topology and physics*. CRC press, 2003.

[10] Michael M Bronstein, Joan Bruna, Yann LeCun, Arthur Szlam, and Pierre Vandergheynst. Geometric deep learning: going beyond euclidean data. *IEEE Signal Processing Magazine*, 34(4):18–42, 2017.

[11] John M Lee. *Introduction to smooth manifolds*, volume 218. Springer Science & Business Media, 2003.

[12] Vardan Papyan, XY Han, and David L Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40):24652–24663, 2020.

[13] Stanislav Fort and Surya Ganguli. Deep ensembles: A loss landscape perspective. *arXiv preprint arXiv:1912.02757*, 2019.

[14] Naftali Tishby and Noga Zaslavsky. Deep learning and the information bottleneck principle. *arXiv preprint arXiv:1503.02406*, 2015.

[15] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. Mutual information neural estimation. *International conference on machine learning*, pages 531–540, 2018.

[16] Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 1991.

[17] Adi Yang, Jacob Durgan, Qing Lyu, and Samuel R Bowman. On the pitfalls of analyzing individual neurons in language models. *arXiv preprint arXiv:2110.07483*, 2023.

[18] Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. Are emergent abilities of large language models a mirage? *arXiv preprint arXiv:2304.15004*, 2023.

[19] Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in gpt-2 small. *arXiv preprint arXiv:2211.00593*, 2023.

[20] Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. *arXiv preprint arXiv:2212.03827*, 2023.

[21] Aolin Xu and Maxim Raginsky. Does mutual information minimize? *arXiv preprint arXiv:2001.00497*, 2020.

[22] Jonathan Schwarz, Siddhant Jayakumar, Razvan Pascanu, Peter E Latham, and Yee Whye Teh. Principles for learning sparse representations. *arXiv preprint arXiv:1808.10578*, 2018.

[23] Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. *Advances in neural information processing systems*, 31, 2018.

[24] Emilien Dupont, Arnaud Doucet, and Yee Whye Teh. Augmented neural odes. *Advances in neural information processing systems*, 32, 2019.

[25] Armen Aghajanyan, Sonal Gupta, and Luke Zettlemoyer. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. *arXiv preprint arXiv:2012.13255*, 2020.

[26] Kenneth Li, Aspen K Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Emergent world representations: Exploring a sequence model trained on a synthetic task. *arXiv preprint arXiv:2210.13382*, 2023.

[27] Bastian Rieck, Matteo Togninalli, Christian Bock, Michael Moor, Max Horn, Thomas Gumbsch, and Karsten Borgwardt. Neural persistence: A complexity measure for deep neural networks using algebraic topology. *arXiv preprint arXiv:1812.09764*, 2019.

[28] William H Guss, Ruslan Salakhutdinov, and Katja Hofmann. Hidden holes: topological aspects of language models. *arXiv preprint arXiv:2406.05798*, 2024.

[29] Gunnar Carlsson. Topology and data. *Bulletin of the American Mathematical Society*, 46(2):255–308, 2009.

[30] Henry Adams, Tegan Emerson, Michael Kirby, Rachel Neville, Chris Peterson, Patrick Shipman, Sofya Chepushtanova, Eric Hanson, Francis Motta, and Lori Ziegelmeier. Persistence images: A stable vector representation of persistent homology. *Journal of Machine Learning Research*, 18(8):1–35, 2017.

[31] Zhaoyu Wang, Yang Liu, and Xin Zhang. A group theoretic analysis of the symmetries underlying base addition and their learnability by neural networks. *arXiv preprint arXiv:2507.10678*, 2024.

[32] Taco Cohen and Max Welling. Group equivariant convolutional networks. *International conference on machine learning*, pages 2990–2999, 2016.

[33] Risi Kondor and Shubhendu Trivedi. On the generalization of equivariance and convolution in neural networks to the action of compact groups. *International Conference on Machine Learning*, pages 2747–2755, 2018.

[34] Shun-ichi Amari. *Information geometry and its applications*, volume 194. Springer, 2016.

[35] Shun-ichi Amari. Natural gradient works efficiently in learning. *Neural computation*, 10(2):251–276, 1998.

[36] James Martens. New insights and perspectives on the natural gradient method. *Journal of Machine Learning Research*, 21(146):1–76, 2020.

[37] Yann Ollivier. Riemannian metrics for neural networks i: feedforward networks. *Information and Inference: A Journal of the IMA*, 4(2):108–153, 2015.

[38] Vladimir I Arnold. *Mathematical methods of classical mechanics*, volume 60. Springer Science & Business Media, 1989.

[39] P-A Absil, Robert Mahony, and Rodolphe Sepulchre. *Optimization algorithms on matrix manifolds*. Princeton University Press, 2009.

[40] Daniel A Roberts, Sho Yaida, and Boris Hanin. *Principles of deep learning theory*. Cambridge University Press, 2021.

[41] Yasaman Bahri, Jonathan Kadmon, Jeffrey Pennington, Sam S Schoenholz, Jascha Sohl-Dickstein, and Surya Ganguli. Statistical mechanics of deep learning. *Annual Review of Condensed Matter Physics*, 11:501–528, 2020.

[42] Wei Zhang, Lin Chen, and Yue Wang. Geometric reasoning in the embedding space. *Mathematics and Computers in Simulation*, 7(3):93, 2024.

[43] Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom in: An introduction to circuits. *Distill*, 5(3):e00024–001, 2020.

[44] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

[45] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[46] Irina Higgins, David Amos, David Pfau, Sebastien Racaniere, Loic Matthey, Danilo Rezende, and Alexander Lerchner. Towards a definition of disentangled representations. *arXiv preprint arXiv:1812.02230*, 2018.

[47] Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. *international conference on machine learning*, pages 4114–4124, 2019.

[48] Ian Tenney, Dipanjan Das, and Ellie Pavlick. Bert rediscovers the classical nlp pipeline. *arXiv preprint arXiv:1905.05950*, 2019.

[49] Olga Kovaleva, Alexey Romanov, Anna Rogers, and Anna Rumshisky. Revealing the dark secrets of bert. *arXiv preprint arXiv:1908.08593*, 2019.

[50] Yonatan Belinkov and James Glass. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 7:49–72, 2019.

[51] Anna Rogers, Olga Kovaleva, and Anna Rumshisky. A primer in bertology: What we know about how bert works. *Transactions of the Association for Computational Linguistics*, 8:842–866, 2020.

[52] Zachary C Lipton. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57, 2018.

[53] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

[54] Michael M Bronstein, Joan Bruna, Taco Cohen, and Petar Veličković. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *arXiv preprint arXiv:2104.13478*, 2021.

[55] Peter W Battaglia, Jessica B Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, et al. Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261*, 2018.

[56] Yann LeCun, Sumit Chopra, Raia Hadsell, M Ranzato, and F Huang. A tutorial on energy-based learning. *Predicting structured data*, 1(0), 2006.

[57] Chelsea Finn, Ian Goodfellow, and Sergey Levine. Deep visual analogy-making. *Advances in neural information processing systems*, 29, 2016.

[58] Peter Gärdenfors. *Conceptual spaces: The geometry of thought*. MIT press, 2004.

[59] Melanie Mitchell. Abstraction and analogy-making in artificial intelligence. *Annals of the New York Academy of Sciences*, 1505(1):79–101, 2021.

[60] Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. *arXiv preprint arXiv:1809.02789*, 2018.

[61] Samuel Greydanus, Misko Dzamba, and Jason Yosinski. Hamiltonian neural networks. *Advances in neural information processing systems*, 32, 2019.

# NeurIPS Paper Checklist

1. **Claims**

   Answer: [Yes]

   Justification: Paper clearly positions work as infrastructure development with modest improvements rather than fundamental breakthroughs.

2. **Limitations**

   Answer: [Yes]

   Justification: Section 5 extensively discusses discrete-continuous mismatch and framework limitations.

3. **Theory assumptions and proofs**

   Answer: [Yes]

   Justification: Mathematical framework clearly states when continuous geometry is artificially imposed on discrete systems.

4. **Experimental result reproducibility**

   Answer: [Yes]

   Justification: Complete experimental setup provided with baseline comparisons and geometric measure computation details.

5. **Open access to data and code**

   Answer: [No]

   Justification: Code will be made available upon acceptance. Dataset (OpenBookQA) is publicly available.

6. **Experimental setting/details**

   Answer: [Yes]

   Justification: Section 4 provides complete experimental setup and interpretability evaluation framework.

7. **Experiment statistical significance**

   Answer: [Yes]

   Justification: Results include statistical tests for orthogonality claims and comparative analysis.

8. **Experiments compute resources**

   Answer: [Yes]

   Justification: Computational complexity discussed but specific resource requirements could be more detailed.

9. **Code of ethics**

   Answer: [Yes]

   Justification: Research follows standard interpretability analysis practices with no ethical concerns.

10. **Broader impacts**

    Answer: [Yes]

    Justification: Section discusses applications to AI interpretability while acknowledging infrastructure nature of contributions.

11. **Safeguards**

    Answer: [NA]

    Justification: Work presents interpretability analysis methods, not models requiring safeguards.

12. **Licenses for existing assets**

Answer: [Yes]

Justification: All datasets and baseline methods properly cited with appropriate attribution.

13. **New assets**

    Answer: [NA]

    Justification: Work introduces mathematical framework for interpretability analysis, not computational assets.

14. **Crowdsourcing and research with human subjects**

    Answer: [NA]

    Justification: Computational analysis only, no human subjects.

15. **Institutional review board (IRB) approvals**

    Answer: [NA]

    Justification: No human subjects research conducted.

16. **Declaration of LLM usage**

    Answer: [No]

    Justification: LLMs analyzed but not used in developing interpretability framework.