Cover Letter for Submission of the Information Dynamics Framework

A Theoretical and Computational Framework for Measuring Emergent Patterns in AI

To the Esteemed Research Teams at AI Institutions,

I am pleased to submit the manuscript titled The Information Dynamics Framework: A Theoretical and Computational Approach to Measuring Emergent Patterns in AI Systems for your consideration. This work integrates information theory and predictive dynamics to quantify emergent patterns—latent representations with high mutual information and low compressibility—in AI systems. It proposes a core metric, the -Field, for predictive dynamics, an optional -Gate for output relevance, and an emergence score, implemented via the ADNAT (Algorithmic Divergence in Neural Abstraction Tracker) codebase.

The framework includes a hypothetical experimental design for transformer models (e.g., TinyLLaMA, BERT) on tasks like sentiment analysis and text generation, with comparisons to Logical Depth, Effective Measure Complexity, and Integrated Information Theory. While theoretical, it is supported by a robust computational implementation, seeking feedback for empirical validation and collaboration.

Sincerely,

Laurentiu G. Florea
floreaglurentiu@gmail.com
+40 728 191 220

July 21, 2025

The Information Dynamics Framework: A Theoretical and Computational Approach to Measuring Emergent Patterns in AI Systems

Laurentiu G. Florea
floreaglurentiu@gmail.com, +40 728 191 220
July 2025

Abstract

The Information Dynamics Framework quantifies emergent patterns in AI systems—latent representations with high mutual information to inputs but low compressibility—using information theory and predictive dynamics. Implemented via the ADNAT (Algorithmic Divergence in Neural Abstraction Tracker) codebase, it defines a core -Field for context-dependent predictive dynamics ($I(S_{n+1}; C|S_n)$), an optional -Gate for output relevance ($I(H; \alpha)$), and an emergence score ($\psi + \phi$). A hypothetical experiment on transformer models (TinyLLaMA, BERT) with tasks like IMDb sentiment analysis and WikiText-103 text generation compares against Logical Depth, Effective Measure Complexity, and Integrated Information Theory. While theoretical, the framework offers a computational foundation for AI interpretability research.

1 Introduction

Emergent patterns in AI—informative yet incompressible representations—are critical for interpretability. The Information Dynamics Framework, implemented via ADNAT, measures such patterns using information-theoretic metrics, drawing on predictive coding [?], Information Bottleneck [?], and statistical complexity [?]. It simplifies to a core predictive metric, supported by a robust computational toolkit.

2 Literature Review

Logical Depth [?] measures computational effort, Effective Measure Complexity (EMC) [?] quantifies predictability, and Integrated Information Theory (IIT) [?] assesses integration. AI interpretability uses mutual information [?] and attention [?]. This framework extends these by focusing on predictive dynamics, complementing Olah et al. (2020) and Belinkov (2022).

3 Notation

- $S_n \in \mathbb{R}^d$: Latent state at timestep n.
- $C \in \mathbb{R}^{k \times l}$: Input sequence of length l.
- $\alpha \in \mathbb{R}^p$: Output vector.
- $H \in \mathbb{R}^m$: Hidden state.
- θ: Neural network parameters.

4 The Information Dynamics Framework

4.1 -Field: Predictive Dynamics

The -Field measures context-dependent predictive dynamics:

$$\psi = I(S_{n+1}; C|S_n)$$

estimated via MINE [?] with PCA reduction (top 10 components). The state update is:

$$S_{n+1} = S_n + W\text{Attn}(C;\theta)$$

where $\text{Attn}(C;\theta) \in \mathbb{R}^k$ is multi-head attention [?].

Justification: High ψ indicates predictive informativeness, extending predictive coding [?].

Complexity: $O(ldd_k)$ for attention, $O(d \log l)$ for MINE with PCA.

4.2 -Gate: Output Relevance (Optional)

The -Gate measures output informativeness:

$$\phi = I(H;\alpha)$$

estimated via MINE with PCA reduction.

Justification: High ϕ aligns with Information Bottleneck [?].

Complexity: $O(mp \log l)$.

4.3 Emergence Score

The emergence score aggregates:

$$\varepsilon = \psi + \phi - (\psi_{\text{random}} + \phi_{\text{random}})$$

normalized against a random predictor baseline.

Justification: High ε indicates emergent patterns [?].

5 Toy Example: Linear Regression

For $y = 2x + \epsilon$, the -Field captures input-output correlations in S_n, and -Gate ensures output relevance. High ε suggests informative, incompressible representations.

6 Hypothetical Experimental Design

Test on TinyLLaMA (1B) and BERT (110M) with: - Tasks: Sentiment analysis (IMDb), text generation (WikiText-103). - Preprocessing: Tokenization, PCA reduction (10 components). - Metrics: ψ, ϕ, normalized ε, ΔK. - Baselines: Logical Depth [?], EMC [?], IIT [?], LSTM, random predictor. - Tests: Mann-Whitney U test. - Procedure: Compute metrics over 100 samples, comparing against baselines.

Limitations: MINE estimation is noisy; compressibility is approximated.

7 Applications

The framework may quantify representation informativeness, aiding interpretability (e.g., detecting overfitting).

8 Conclusion

The Information Dynamics Framework, implemented via ADNAT, provides a theoretical and computational approach to measure emergent patterns, with a clear path for validation.

References

1. Belinkov, Y. (2022). Probing classifiers: Promises, shortcomings, and advances. Computational Linguistics, 48(1), 207–270. DOI:10.1162/coli$_{a0}$0422.

2. Bennett, C. H. (1988). Logical Depth and Physical Complexity. The Universal Turing Machine, 207–223.

3. Belghazi, M. I., et al. (2018). Mutual Information Neural Estimation. ICML, 80, 531–540. DOI:10.48550/arXiv.1801.04062.

4. Friston, K. (2010). The free-energy principle: A unified brain theory? Nature Reviews Neuroscience, 11(3), 127–138. DOI:10.1038/nrn2787.

5. Grassberger, P. (1986). Toward a quantitative theory of self-generated complexity. International Journal of Theoretical Physics, 25(9), 907–938. DOI:10.1007/BF00668821.

6. Kullback, S., Leibler, R. A. (1951). On information and sufficiency. Annals of Mathematical Statistics, 22(1), 79–86. DOI:10.1214/aoms/1177729694.

7. Olah, C., et al. (2020). Zoom in: An introduction to circuits. Distill, 5(3). DOI:10.23915/distill.00024.001

8. Shalizi, C. R., Crutchfield, J. P. (2001). Computational mechanics: Pattern and prediction, structure and simplicity. Journal of Statistical Physics, 104(3), 817–879. DOI:10.1023/A:1010388907

9. Tishby, N., Zaslavsky, N. (2017). Deep learning and the information bottleneck principle. IEEE Information Theory Workshop, 1–5. DOI:10.1109/ITW.2015.7499375.

10. Tononi, G., et al. (2016). Integrated information theory: From consciousness to its physical substrate. Nature Reviews Neuroscience, 17(7), 450–461. DOI:10.1038/nrn.2016.44.

11. Vaswani, A., et al. (2017). Attention is all you need. Neural Information Processing Systems, 30, 5998–6008. DOI:10.48550/arXiv.1706.03762.