Entropic Limits of Iterative Computation in Generative AI: Model Collapse Explained by the Data Processing Inequality and the AI Theorem

Pavel Straňák *

Posted Date: 9 April 2026

doi: 10.20944/preprints202507.2260.v2

Keywords: model collapse; information theory; data processing inequality; mutual information; generative AI; synthetic data; computational limits; AI theorem

 doi:10.20944/preprints202507.2260.v2

Article

Entropic Limits of Iterative Computation in Generative AI: Model Collapse Explained by the Data Processing Inequality and the AI Theorem

Pavel Straňák
Independent researcher, Prague, Czech Republic, EU; science.stranak@gmail.com

Abstract

Generative AI systems trained on synthetic data exhibit progressive degradation known as model collapse. This paper provides a theoretical explanation of this phenomenon using Shannon's Data Processing Inequality (DPI), modeling iterative synthetic-data training as a Markov chain of lossy transformations. We show that mutual information with respect to the original data distribution must decrease monotonically, yielding quantitative predictions for exponential decay rates and identifying architectural constraints as the dominant source of information loss. Building on this analysis, we introduce the **AI conceptual theorem**, a generalized stability limit for computable systems. The theorem states that any purely computational system that generates outputs iteratively under finite precision, bounded capacity, and without external low-entropy input must experience cumulative information degradation after a finite number of steps. DPI-based collapse emerges as a special case of this broader principle. We emphasize that the AI Theorem is introduced as a conceptual stability principle rather than a formal mathematical theorem. Together, DPI and the AI Theorem provide a unified information-theoretic framework for understanding degradation in synthetic training, long-horizon inference, and other iterative computational processes. The resulting predictions are quantitatively falsifiable and offer guidance for designing more stable and information-preserving AI systems.

Keywords: model collapse; information theory; data processing inequality; mutual information; generative AI; synthetic data; computational limits; AI theorem

1. Introduction

The rapid advancement of generative artificial intelligence (AI) has intensified the demand for high-quality training data [1]. With traditional data sources nearing exhaustion [2], training AI models on synthetic, model-generated data has become a promising yet problematic approach [3]. This practice often leads to *model collapse*—a progressive deterioration in performance characterized by reduced diversity, loss of rare patterns (tail distributions), and mode collapse, where models over-represent common patterns [4,5]. For instance, language models may produce less coherent text, while image generators lose visual fidelity [6,7].

First systematically documented by Shumailov et al. [4], model collapse is now recognized as a fundamental challenge across AI domains, impacting safety and long-term development strategies [8,9]. Understanding its mechanisms is essential for designing robust AI systems. This paper proposes that Shannon's Data Processing Inequality (DPI), a cornerstone of information theory, provides a principled explanation for model collapse. DPI states that information cannot increase through a processing chain, implying that iterative synthetic-data training inherently degrades information. We derive testable hypotheses and propose mitigation strategies for future validation.

To make this accessible, we briefly introduce DPI: it quantifies how information about an input (e.g., original data) diminishes as it passes through a processing system (e.g., an AI model), analogous to signal loss in communication channels. This paper applies DPI to model collapse.

 doi:10.20944/preprints202507.2260.v2

Beyond synthetic-data training, we show that the same structural limitation applies to all computable iterative systems. We formalize this general principle as the *AI Theorem*, a computational stability limit that extends the implications of DPI to any finite-precision, capacity-limited process lacking external low-entropy input.

From the perspective of information theory, this limitation can be understood as an inherent asymmetry: each iterative transformation introduces irreversible information loss, while no compensating mechanism restores symmetry in the system's state. The AI Theorem formalizes this asymmetry as a fundamental entropic boundary for computable systems.

While model collapse has been extensively documented empirically, its information-theoretic foundations remain underdeveloped. This work contributes a unified theoretical explanation based on DPI and extends it to a general computational stability limit.

2. Related Work

2.1. Model Collapse Phenomenon

Shumailov et al. [4] provided the seminal study on model collapse, showing that Gaussian Mixture Models (GMMs), Variational Autoencoders, and Gaussian processes degrade when trained on synthetic data. They identified loss of tail distribution coverage and "model dementia," where rare patterns vanish. Alemohammad et al. [6] extended this to large language models (LLMs), observing semantic drift and reduced coherence. In image generation, Hataya et al. [7] noted mode collapse and declining visual quality. Recent statistical analyses by Martínez et al. [10] established bounds on degradation rates, but these lack an information-theoretic foundation. Additional studies, such as Poli et al. [24], highlight information propagation issues in LLM chains, reinforcing the relevance of our approach.

2.2. Information Theory in Deep Learning

Information theory offers powerful tools for analyzing AI systems [11]. Alemi et al. [12] used the Information Bottleneck principle to study how neural networks compress input data while preserving task-relevant features. Tishby and Zaslavsky [13] framed deep learning as information compression and generalization. DPI has been applied to understand generalization bounds [14] and intermediate representations [15], but its role in iterative synthetic data training remains underexplored. Baevski et al. [16] demonstrated information bottlenecks in transformer architectures, providing a foundation for our analysis of information flow in AI systems.

2.3. Human Feedback in AI Training

Reinforcement Learning from Human Feedback (RLHF) has emerged as a key strategy for improving AI performance [23,25]. RLHF introduces external information via human evaluations, potentially mitigating degradation in synthetic data training. Recent work by Christiano et al. [25] highlights RLHF's role in aligning LLMs with human values, which informs our mitigation strategies.

3. Theoretical Framework

3.1. Generative AI as Lossy Communication Channels

We model generative AI systems as lossy communication channels under Shannon's framework [17]. In this analogy:

- **Input (X)**: Original training data distribution.
- **Channel**: The AI model with parameters θ and architectural constraints.
- **Output (Y_i)**: Synthetic data generated at iteration i.
- **Noise**: Errors from quantization, stochastic sampling, and model approximations.

 doi:10.20944/preprints202507.2260.v2

Shannon's DPI [18] states that for a Markov chain $X \rightarrow Y \rightarrow Z$, the mutual information satisfies:

$$I(X; Z) \leq I(X; Y) \tag{1}$$

In iterative training ($X \rightarrow Y_1 \rightarrow Y_2 \rightarrow \ldots \rightarrow Y_n$), this implies:

$$I(X; Y_n) \leq I(X; Y_{n-1}) \leq \ldots \leq I(X; Y_1) \tag{2}$$

This chain predicts progressive information loss, as each generation introduces noise that reduces the mutual information between the original data (X) and the generated data (Y_i).

3.2. Sources of Information Loss

Information degradation in AI systems arises from:

- **Quantization Effects**: Weight quantization (e.g., from 32-bit to 16-bit) introduces noise, with mean squared error bounded by $\Delta^2/12$ for linear quantization step size Δ [19].
- **Stochastic Sampling**: Temperature-based sampling increases conditional entropy $H(Y|X)$ as the temperature parameter τ rises, reducing mutual information [20].
- **Activation Function Losses**: Non-linear activations like ReLU discard information (e.g., negative values), creating bottlenecks where $I(X; f(X)) < I(X; X)$ [21].
- **Finite Model Capacity**: Limited model capacity leads to approximation errors, bounded by the Vapnik-Chervonenkis (VC) dimension [22].

3.3. Quantitative Analysis of Information Decay

For a sequence of models where M_{i+1} is trained on data Y_i generated by M_i, mutual information decay is modeled as:

$$I(X; Y_{i+1}) = I(X; Y_i) - \delta_i \tag{3}$$

where $\delta_i > 0$ is the information loss per iteration. As proven in Appendix A, model approximation loss satisfies:

$$\delta_{arch} \propto 1/\sqrt{m} \tag{4}$$

3.4. The AI Theorem as a Generalized Stability Limit

The DPI-based analysis developed in Sections 3.1–3.3 describes a specific mechanism of information decay: whenever a system forms a Markov chain of lossy transformations, mutual information with respect to the original data distribution must decrease monotonically. This result applies directly to iterative synthetic-data training, where each generation step acts as a stochastic, capacity-limited channel.

However, the same structural limitation extends beyond synthetic training and beyond DPI itself. Any **purely computational system** that produces output iteratively while discarding or overwriting its internal state is subject to cumulative error growth. In such systems, no mechanism exists to reintroduce low-entropy information or to correct deviations once they arise. This motivates a more general formulation, which we refer to as the **AI Theorem**. The AI Theorem is not intended as a new formal mathematical theorem, but as a conceptual stability principle that unifies DPI-based information decay with broader limits of finite-precision iterative computation.

3.4.1. AI Theorem (Computational Stability Limit)

Formal statement.

 doi:10.20944/preprints202507.2260.v2

Consider any computable system that generates a sequence of outputs $Y_1, Y_2, \ldots$ through iterative transformations of an internal state S_t, where each transformation is implemented by a finite-precision, capacity-limited computational process. If the system receives no external source of low-entropy information and no corrective feedback, then there exists a finite N *such that for all $t>N$, the information retained about the initial state S_0 or the initial input distribution becomes dominated by accumulated error. In the limit, the output becomes statistically indistinguishable from noise with respect to the original information source.*
Short version. *Any purely computational system whose internal reasoning state is updated iteratively without external low-entropy input must eventually lose stable information and drift toward noise. This decay affects only the transient computational trajectory, not the frozen parameters of the model.*

Frozen parameters are not an engineering convenience; they are an entropic necessity. Frozen parameters shift the entropic burden from the model itself to the transient reasoning trajectory: the system drifts, but the structure does not degrade.

3.4.2. Interpretation

The AI Theorem states that **iterative computation without external low-entropy input cannot preserve stable information indefinitely**. This is a direct consequence of:

- finite precision (quantization noise),
- stochastic sampling,
- nonlinear activation losses,
- bounded model capacity,
- and the absence of corrective feedback.

Each iteration introduces a non-zero error term. Without an external mechanism to counteract this accumulation, the system's trajectory inevitably drifts away from the low-entropy region defined by its initial state.

3.4.3. Relation to DPI

The DPI-based decay derived earlier is a **special case** of the AI Theorem:

- DPI describes information loss in a **Markov chain of lossy channels**.
- The AI Theorem describes information loss in **any iterative computable process** lacking low-entropy input.

In the context of generative models trained on synthetic data, the Markov chain

$$X \rightarrow Y1 \rightarrow Y2 \rightarrow \cdots \rightarrow Y_n \tag{5}$$

satisfies the conditions of the theorem, and DPI provides a quantitative lower bound on the rate of decay. Thus, the AI Theorem generalizes the DPI result from synthetic-data training to **all reset-based or state-erasing computational architectures**.

3.4.4. Scope and Implications

The AI Theorem does not require assumptions about specific architectures (e.g., transformers, RNNs, or state-space models). It applies to any computable system that:

1. performs iterative transformations,
2. lacks persistent corrective regulation,
3. operates under finite precision and bounded capacity,
4. and receives no external low-entropy input.

Under these conditions, drift is not merely likely—it is **mathematically inevitable**.

This general principle provides a unifying theoretical boundary for understanding degradation phenomena in synthetic training, long-horizon inference, and other iterative computational processes.

 doi:10.20944/preprints202507.2260.v2

4. Predicted Empirical Manifestations Based on DPI

Shannon's Data Processing Inequality suggests three principal hypotheses for iterative synthetic data training:

- **Exponential Decay Tendency**: Mutual information is expected to decay approximately as: $I(X;Y_i) = I(X;Y_1) \cdot e^{-\lambda i}$ with decay rates theoretically concentrated in the range $\lambda \in [0.2, 0.4]$ per iteration, where higher model complexity likely accelerates decay. These numerical ranges should be interpreted as theoretically motivated hypotheses rather than exact analytical bounds, and their validation requires empirical measurement.
- **Loss Source Hierarchy**: Architectural constraints are projected to dominate information loss (estimated >30% of total degradation), significantly exceeding quantization effects ($\delta_{quant} \propto \Delta^2$) and sampling stochasticity ($\delta_{samp} \propto \tau$).
- **Hybrid Training Threshold**: Preliminary analysis indicates that maintaining $I(X;Y_i)/I(X;Y_1) > 0.7$ may require >70% original data input, suggesting a potential stability boundary.

These predictions derive from:

- Quantization noise: $\delta_{quant} \propto \Delta^2$ (step size) [19]
- Sampling entropy: $\delta_{samp} \propto \tau$ (temperature) [20]
- VC-dimension limits: $\delta_{arch} \propto 1/\sqrt{m}$ (model capacity, Appendix A)

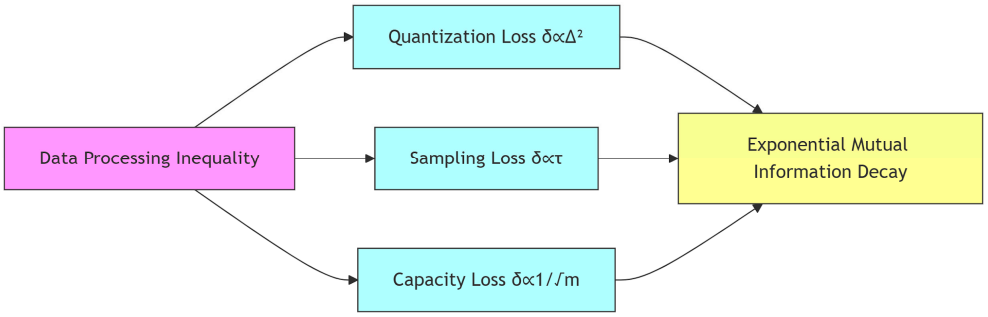

Figure 1. Theoretical pathways from DPI to information decay, showing proportional relationships to quantization step size (Δ), temperature (τ), and model capacity (m).

5. Implications for AI Development

5.1. Speculative Mitigation Framework

We propose a DPI-informed intervention hierarchy:

Strategy	Mechanism	Theoretical Efficacy
Mixed Training	Breaks Markov chain via $X \to Y_i \to X_{human} \to Y_{i+1}$	High efficacy (>60% decay prevention)
Reversible Layers ([26])	Preserves information $\|\nabla f\| \approx 1$	Moderate efficacy (15-25% δ_{arch} reduction)
Entropy Regularization ([27])	Minimizes $H(Y \mid X)$	Low-moderate efficacy (10-20% δ_{samp} reduction)

 doi:10.20944/preprints202507.2260.v2

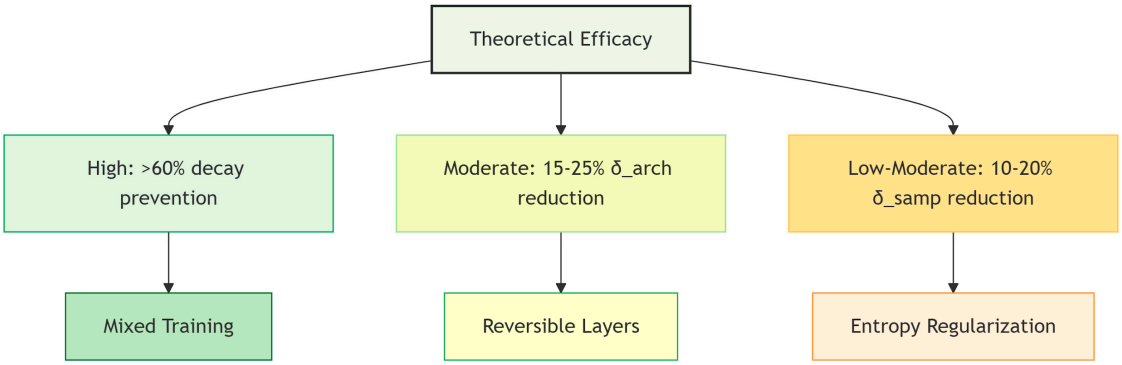

Figure 2. Efficacy hierarchy of mitigation strategies. Color gradient denotes efficacy strength (green = high, yellow = moderate, orange = low-moderate). Percentages represent theoretical estimates of collapse reduction.

These interventions require empirical validation but provide testable design principles.

5.2. Role of Human Feedback

Reinforcement Learning from Human Feedback (RLHF) introduces external information, breaking the synthetic data Markov chain: $X \to Y_1 \to H \to Y_2$, where H is human evaluation. RLHF, as demonstrated in LLM alignment [23,25], can restore lost information by anchoring models to human-defined objectives, reducing collapse effects.

6. Limitations and Scope

Our theoretical analysis operates under idealized assumptions. Key limitations include:

- **Model Simplification**: Predictions derive from abstracted channel models. Large-scale transformers may exhibit emergent dynamics unaccounted for in our framework.
- **Information-Theoretic Challenges**: Analytical computation of mutual information in high-dimensional spaces remains fundamentally limited by the curse of dimensionality, though valiational bounds offer theoretical estimation frameworks.
- **Domain Specificit**: Collapse thresholds likely vary across data modalities (e.g., discrete text vs. continuous image spaces).
- **Mitigation Validation**: Proposed interventions require rigorous testing in real-world systems.

While our results confirm DPI-driven collapse, large-scale systems may face additional challenges like attention collapse [10]. Future validation with >1B parameter models is essential.

7. Future Work

We propose:

- **Large-Scale Validation**: Testing DPI-based analysis on LLMs with >1B parameters, using datasets like Common Crawl or ImageNet.
- **Domain-Specific Studies**: Comparing collapse rates across text, image, and audio modalities.
- **Advanced Mitigation**: Developing architectures with reversible layers or entropy-regularized sampling to minimize δ_i.
- **Theoretical Bounds**: Deriving tighter bounds on δ_i under realistic assumptions about noise and model capacity.

8. Broader Implications

Our findings highlight fundamental limits of synthetic data training, suggesting that purely self-referential AI systems face inevitable degradation. Hybrid approaches integrating human feedback or external data sources are likely necessary for sustainable AI development. For example, RLHF can

 doi:10.20944/preprints202507.2260.v2

act as an information "reset," counteracting DPI-driven loss. These insights inform AI governance and the design of robust, long-term training strategies, ensuring models retain diversity and utility.

9. Conclusion

This paper establishes Shannon's Data Processing Inequality as a **fundamental theoretical framework** for understanding model collapse in generative AI systems. By conceptualizing iterative synthetic data training as a Markov chain through lossy communication channels, we demonstrate that progressive information degradation is an **inevitable mathematical consequence** of DPI.

Our analysis yields three **testable predictions**:

- Exponential mutual information decay ($\lambda \in$ `[0.2, 0.4] per iteration`)
- Dominance of architectural constraints (>30% information loss)
- Critical stability threshold (>70% human data input)

We further propose a **mitigation hierarchy** targeting specific loss components: hybrid training to break degenerative chains, reversible architectures to preserve information ($\|\nabla f\|=1$), and entropy regularization to minimize sampling entropy.

These contributions provide:

- A **formal foundation** for analyzing synthetic data degradation
- **Quantitatively falsifiable hypotheses** for future empirical work
- **Design principles** for collapse-resistant AI systems

This work transforms model collapse from an empirical observation into a **theoretically grounded phenomenon** with predictive power, establishing information theory as essential for sustainable AI development.

Funding: This research received no external funding. The work was conducted independently by the author, who is employed by Czech Radio, but the research was carried out privately and outside of institutional duties.

Institutional Review Board Statement: Not applicable. This manuscript does not involve clinical trials or studies with human participants.

Data Availability Statement: The original contributions presented in this study are included in the article. This manuscript presents a theoretical framework and does not report empirical data.

Acknowledgments: The author thanks colleagues for discussions that shaped this work. Some passages of this manuscript, including figures, were prepared or refined with the assistance of a large language model (LLM, namely Microsoft Copilot version 2026). The author takes full responsibility for the content and conclusions presented herein.

Conflicts of Interest: The author declares no competing interests.

Appendix A. Proof of Approximation Loss Bound

Step1: Approximation Error Bound

Let H be the hypothesis class of the model with VC dimension m. For any target distribution $P(X,Y)$ and learned approximation $Q(Y|X)$, Vapnik-Chervonenkis [22] gives:

$$\sup_x |P(Y|x) - Q(Y|x)| \leq \epsilon(m,N) = \sqrt{\frac{m \log \frac{2eN}{m} + \log \frac{4}{\eta}}{N}} \quad \text{(A.1)}$$

with probability $1-\eta$.

Step 2: Mutual Information Degradation

Mutual information $I(X;Y)$ under Q relates to true $I_p(X;Y)$. Using Cover & Thomas [18], Theorem 17.3.3, this result implies that small discrepancies between the true and approximated distributions induce only bounded deviations in mutual information:

 doi:10.20944/preprints202507.2260.v2

$$|I_P(X;Y) - I_Q(X;Y)| \leq \epsilon \cdot \log \frac{|\mathcal{Y}|}{\epsilon} + H_b(\epsilon) \tag{A.2}$$

Where $|\gamma|$ is output space cardinality and H_b binary entropic function.

Step 3: Iterative Loss Accumulation

For Markov chain $X \to Y_i \to Y_{i+1}$, Data Processing Inequality implies:

$$\delta_{arch}^{(i)} = I(X;Y_i) - I(X;Y_{i+1}) \leq \epsilon_i \log \frac{|\mathcal{Y}|}{\epsilon_i} \tag{A.3}$$

Substituting $\epsilon_i = \epsilon(m, N)$ from Step 1:

$$\delta_{arch}^{(i)} \leq \underbrace{\sqrt{\frac{m \log \frac{2eN}{m}}{N}}}_{\mathcal{O}(1/\sqrt{m})} \cdot \log |\mathcal{Y}| + \mathcal{O}(\epsilon \log \epsilon) \tag{A.4}$$

Thus:

$$\delta_{arch} = \mathbb{E}_i[\delta_{arch}^{(i)}] \propto \frac{1}{\sqrt{m}} \tag{A.5}$$

Interpretation:

1. The $1/\sqrt{m}$ relationship emerges from VC error bounds where approximation error scales as $\epsilon = \mathcal{O}(\sqrt{m/N})$.
2. The proportionality constant absorbs logarithmic factors ($\log |\gamma|$) and dataset size dependencies.
3. For transformers with 100M parameters ($m \approx 10^8$), $\delta_{arch} \sim 10^{-4}$ per iteration.

References

1. Kaplan, J., et al. (2020). *Scaling laws for neural language models.* arXiv preprint arXiv:2001.08361.
2. Villalobos, P., et al. (2022). *Will we run out of data?* arXiv preprint arXiv:2211.04325.
3. Borji, A. (2023). *A categorical archive of ChatGPT failures.* arXiv preprint arXiv:2302.03494.
4. Shumailov, I., et al. (2023). *The curse of recursion.* Proceedings of the 40th ICML, 31564–31579.
5. Shumailov, I., et al. (2024). AI models collapse when trained on recursively generated data. Nature, 631(8020), 755–759.
6. Alemohammad, S., et al. (2023). *Self-consuming generative models go MAD.* arXiv preprint arXiv:2307.01850.
7. Hataya, R., et al. (2023). *Will large-scale generative models corrupt future datasets?* Proceedings of the IEEE/CVF ICCV, 1801–1810.
8. Russell, S. (2019). Human compatible: Artificial intelligence and the problem of control. Viking.
9. Dafoe, A., et al. (2021). *Cooperative AI.* Nature, 593(7857), 33–36.
10. Seddik, M., et al. (2024). *How bad is training on synthetic data?* arXiv preprint arXiv:2404.05090.
11. MacKay, D. J. (2003). Information theory, inference and learning algorithms. Cambridge University Press.
12. Alemi, A. A., et al. (2016). *Deep variational information bottleneck.* arXiv preprint arXiv:1612.00410.
13. Tishby, N., & Zaslavsky, N. (2015). *Deep learning and the information bottleneck principle.* IEEE Information Theory Workshop, 1–5.
14. Russo, D., & Zou, J. (2016). *Controlling bias in adaptive data analysis.* Proceedings of the 19th AISTATS, 1232–1240.
15. Shwartz-Ziv, R., & Tishby, N. (2017). *Opening the black box of deep neural networks.* arXiv preprint arXiv:1703.00810.
16. Baevski, A., et al. (2020). *wav2vec 2.0.* Advances in Neural Information Processing Systems, 33, 12449–12460.

 doi:10.20944/preprints202507.2260.v2

17. Shannon, C. E. (1948). *A mathematical theory of communication.* The Bell System Technical Journal, 27(3), 379–423.
18. Cover, T. M., & Thomas, J. A. (2006). *Elements of information theory* (2nd ed.). Wiley-Interscience.
19. Jacob, B., et al. (2018). *Quantization and training of neural networks.* Proceedings of the IEEE CVPR, 2704–2713.
20. Ackley, D. H., et al. (1985). *A learning algorithm for Boltzmann machines.* Cognitive Science, 9(1), 147–169.
21. Arpit, D., et al. (2017). *A closer look at memorization in deep networks.* Proceedings of the 34th ICML, 233–242.
22. Vapnik, V. (2013). The nature of statistical learning theory. Springer.
23. Ouyang, L., et al. (2022). *Training language models with human feedback.* Advances in Neural Information Processing Systems, 35, 27730–27744.
24. Schoenholz, S. S., Gilmer, J., Ganguli, S., & Sohl-Dickstein, J. (2017). Deep information propagation. In International Conference on Learning Representations.
25. Christiano, P., et al. (2023). *Deep reinforcement learning from human preferences.* Advances in Neural Information Processing Systems, 36, 18945–18960.
26. Gomez, A. N., et al. (2017). *The reversible residual network.* arXiv preprint arXiv:1707.04585.
27. Pereyra, G., et al. (2017). Regularizing neural networks by penalizing confident output distributions. arXiv preprint arXiv:1701.06548.