Distillation as Self-Reference: Epistemic Limits for Mathematical and Symbolic Reasoning in AI

Imran Nasim[1,2]
[1]IBM UK
[2]University of Surrey
imran.nasim@ibm.com, i.nasim@surrey.ac.uk

Abstract

Knowledge distillation (KD) is widely used to compress large language models, yet its impact on models' *reasoning capacity* remains poorly understood. We present a theoretical framing of data-free and recursive KD as a self-referential learning process in which students approximate their teachers' approximations. Using Kolmogorov complexity and computable information-theoretic proxies, we show that such recursive compression enforces a *monotonic reduction of information* and bounds representational richness. This perspective has direct implications for *mathematical and symbolic reasoning*, where epistemic depth and compositional structure are essential. We further relate this information reduction to Shannon entropy and Minimum Description Length (MDL), and outline new evaluation paradigms grounded in epistemic fidelity to assess whether distilled models retain the structural knowledge required for robust reasoning.

Introduction

Knowledge distillation (KD) [13] has become a cornerstone of efficient large language model (LLM) development, enabling smaller student models to approximate the behavior of larger teachers with reduced computational cost. Distilled models often match or even outperform their teachers on standard benchmarks [22, 14, 8], suggesting that behavioral imitation can suffice for many applications. However, such benchmarks typically measure surface-level accuracy rather than the depth or robustness of learned representations. They rarely assess whether the student has acquired the structural knowledge needed to generalize beyond its teacher's distribution or to perform novel reasoning tasks.

We argue that data-free distillation forms a closed epistemic loop: the student learns only the teacher's approximations of the data, producing an approximation of an approximation. This self-referential structure imposes an upper bound on representational complexity and risks epistemic flattening, echoing collapse dynamics seen in adversarial systems such as GANs [9, 24]. Unlike GANs, however, which retain a limited grounding via the discriminator's access to real data, distillation in its data-free form provides no such external correction signal. Similar risks arise in self-learning pipelines [29, 28, 33], where models are trained on their own synthetic outputs. Our perspective complements prior information-theoretic analyses of representation bottlenecks [1, 31] by framing distillation as an *epistemically bounded process* with implications for mathematical and symbolic reasoning. This paper advances a *conceptual and theoretical hypothesis*: that data-free or recursive distillation constitutes a form of *self-referential learning*, in which students approximate teachers that themselves approximate data. We analyse this regime through information-theoretic lenses, Kolmogorov complexity, Shannon entropy, and Minimum Description Length (MDL), and propose that such recursive compression induces a *monotonic reduction of information* across generations. Beyond its information-theoretic framing, the process can also be interpreted geometrically [20, 18, 30, 11, 21]: each model defines a manifold of representations in logit space, and the KD objective aligns the student's manifold with that of the teacher. In the data-free regime, this alignment occurs without reference to the original data manifold, offering a complementary geometric view of the monotonic information-reduction hypothesis developed below. Rather than presenting new empirical results, our contribution is an analytic framework and formal proposition clarifying when and why epistemic degradation may occur. We also outline how this framing connects to reasoning tasks such as theorem proving, symbolic manipulation, and compositional abstraction, where preserving epistemic richness is essential. Our goal is not to dismiss the practical value of KD, but to highlight the potential risks of epistemic degradation in recursive or data-isolated settings. As KD becomes a standard tool for scaling and deploying reasoning systems, it is essential to ask not only whether students imitate their teachers, but whether they preserve the epistemic depth required for robust abstraction and generalization.

Background and Related Work

KD transfers the behavior of a large teacher to a smaller student [13]. Recent work has explored im-

proved objectives [32, 8], multi-step distillation [5], and self-supervised extensions such as Self-Instruct [29], Alpaca [28], and LIMA [33]. However, these approaches are typically evaluated through surface benchmarks, not through the lens of epistemic content preservation. Our work complements information-theoretic analyses of representation bottlenecks [1, 31] by focusing on epistemic limits in model-to-model transfer.

KD and compression. Beyond objective tweaks [32] and training dynamics [8, 14], most evaluations optimize for accuracy or throughput rather than epistemic content. By contrast, we foreground preservation (or loss) of structural knowledge under teacher-only supervision.

Self-learning and recursive distillation. Methods like Self-Instruct [29], Alpaca [28], and LIMA [33] train models on teacher- or self-generated data; related techniques include self-refinement [26] and born-again distillation [5]. These raise questions about long-term representation fidelity under recursive supervision.

Epistemic and information-theoretic perspectives. Prior work explores representation bottlenecks and generalization [1, 31] and examines grounding, abstraction, and hallucination in LLMs [3, 15]. We add an explicitly *epistemic* lens on model-to-model transfer, arguing that data-free distillation forms epistemically closed loops.

Theoretical Framing

Distillation as Self-Reference

In adversarial learning, such as in Generative Adversarial Networks (GANs), a generator G learns to produce outputs that a discriminator D cannot distinguish from real data, while D co-trains on both real and generated samples [7, 2]. This creates a feedback loop in which each model's signal is defined by the other. In data-free distillation, by contrast, the student S is trained to mimic the output distribution of a teacher T, typically via a KL-based loss, [13, 27, 25] *without a direct grounding signal from the data.* In typical settings, T is significantly larger than S, further constraining S to compress T's outputs under tighter capacity limits.

If S never sees raw data, its learning signal is entirely mediated through T. Thus, S does not learn the real training distribution P_{data}, but an approximation induced by T. This yields a form of *epistemic self-reference*: S encodes an approximation of an approximation, which may fail to capture structure not externally visible in T's outputs. If T has already compressed or filtered aspects of P_{data}, S has no mechanism to recover them, making it fundamentally unable to exceed the epistemic reach of T.

Kolmogorov Complexity and Information Loss

Let $C(x)$ denote the Kolmogorov complexity of an object x, i.e., the length of the shortest program (for a fixed universal Turing machine) that outputs x [16]. In the KD setting, let D be the original data distribution, T the teacher trained on D, and S the student trained to approximate T. If T is a (lossy) compressor of D, then

$$C(T) < C(D), \qquad C(S) \leq C(T) \;\Rightarrow\; C(S) < C(D),$$

indicating cumulative information loss through two-step compression. For iterative distillation $S_1 \to \cdots \to S_n$,

$$C(S_n) \leq \cdots \leq C(S_1) \leq C(T) < C(D).$$

This expresses a qualitative hierarchy of representational complexity.

Computable proxies. Because $C(\cdot)$ is uncomputable, we employ computable surrogates capturing the same intuition: (i) Shannon entropy $H(\cdot)$ of model outputs or latent states; (ii) Minimum Description Length (MDL), which upper-bounds total codelength via model plus data description; and (iii) Mutual Information (MI) between representations and data. These quantities make the analysis empirically testable and address concerns about operationalisation.

Proposition 1 (Monotonic Information Reduction)

Let P_D, P_T, and P_S denote the predictive distributions of the data, teacher, and student over a common input support $\mathcal{X}$. If T is a lossy compressor of D and S is optimized only to approximate T, then

$$H(S) \;\leq\; H(T) \;\leq\; H(D),$$

with equality only when the mappings are lossless with respect to P_D.

Sketch. Because S minimizes divergence to T rather than to D, information absent from T's outputs is unrecoverable by S. By the data-processing inequality, information cannot increase along the chain $D \to T \to S$, giving the stated ordering. Iterating the argument across n generations yields $H(S_n) \leq \cdots \leq H(S_1) \leq H(T) \leq H(D)$. □

Recursive intuition. If each distillation step reduces representational entropy by a constant factor $\alpha \in (0, 1)$, then $H_n = \alpha^n H_0$, illustrating exponential decay of representational richness under repeated model-to-model transfer.

Terminology. *Epistemic flattening* refers to reduced diversity (e.g., lower entropy) of internal states such that distinct underlying structures map to equivalent representations. By *inference over an inferred distribution* we mean that S conditions on T's induced predictive distribution rather than directly on P_{data}.

Relation to Information Bottleneck and MDL

The Information Bottleneck (IB) principle formalizes compression by minimizing $I(X; Z)$ while preserving predictive information $I(Z; Y)$ [23]. Knowledge distillation can be interpreted as an *ungrounded* IB process: the student minimizes mutual information with

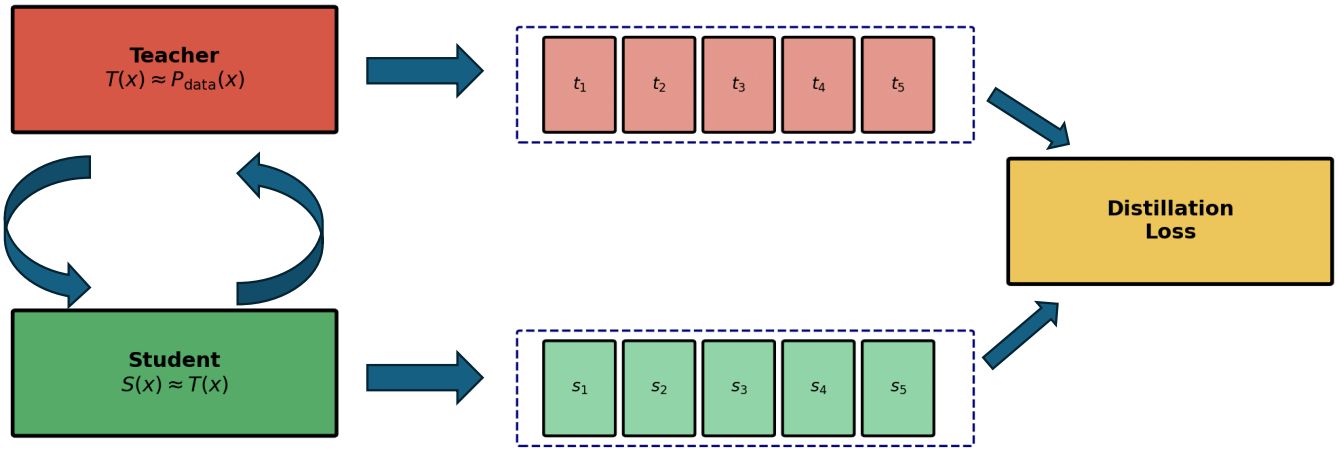

Figure 1: Knowledge Distillation as a self-referential system. The student approximates the teacher, which itself is an approximation of the real training data. Teacher and student logits are t_i and s_i, respectively.

the teacher's latent distribution rather than with the true data. This makes epistemic compression a predictable outcome when $I(Z_S; Y_D)$ decreases across recursive transfers. Under the MDL principle, the total codelength $L(M, D) = L(M) + L(D|M)$ increases as the student's model M_S loses explanatory capacity relative to M_T. Both frameworks corroborate the monotonic-entropy bound in Proposition 1.

Recursive Distillation and Information Decay

Consider a teacher T trained on D with predictive distribution $P_T(y|x)$. A first-generation student S_1 is trained via $\mathcal{L}_{\text{KD}} = \text{KL}(P_T \| P_{S_1})$; a second-generation student S_2 is trained on S_1, and so on. If $C(D) > C(T) > C(S_1) > \cdots > C(S_n)$, the resulting models exhibit progressively reduced entropy and diversity of outputs unless external information (data or stochastic regularisation) is re-injected. This recursive compression clarifies how epistemic flattening can compound even in the absence of overt training collapse.

Epistemic Hierarchies and Exceeding the Teacher

While Proposition 1 establishes monotonic compression under purely data-free transfer, a student may *exceed* its teacher's epistemic reach when one or more of the following hold: (i) *capacity expansion* (the student has a larger function class); (ii) *structural priors* aligning with P_{data} (e.g., modularity or disentanglement); or (iii) *external grounding* through limited labeled data or feedback. Such factors inject new information channels, breaking the closed loop assumed in Proposition 1. Our bounds therefore characterise the idealised, teacher-only regime and serve as an upper limit on epistemic depth within recursive KD.

Geometric view. From a geometric perspective, each model defines an embedding $f : \mathcal{X} \to \mathbb{R}^d$ whose image forms a *manifold of representations* $\mathcal{M}_f$ in logit space. Knowledge distillation can be viewed as a manifold-alignment process [11] in which the student seeks a mapping whose image $\mathcal{M}_S$ approximates the teacher's manifold $\mathcal{M}_T$, minimizing a divergence between their local distributions. In the data-free setting, however, $\mathcal{M}_S$ is anchored only to $\mathcal{M}_T$, not to the original data manifold $\mathcal{M}_D$. Hence the composition $\mathcal{M}_D \to \mathcal{M}_T \to \mathcal{M}_S$ can only reduce or preserve intrinsic dimensionality, providing a geometric interpretation of the monotonic information-reduction bound in Proposition 1.

Implications and Extensions

Framing knowledge distillation as a self-referential, epistemically bounded process clarifies when information reduction becomes problematic, how it can be measured, and how such pipelines might be mitigated or redesigned. This section connects the theoretical result of monotonic information reduction to practical implications for reasoning systems.

When is epistemic loss consequential? Distillation-induced degradation may be masked by in-distribution benchmarks but becomes critical when generalization beyond the teacher's scope is required (e.g., theorem proving, symbolic manipulation, compositional reasoning, or regulated domains) [19]. In such settings, epistemically shallow students can produce fluent yet brittle outputs that lack structural fidelity. Self-distillation pipelines (Self-Instruct [29], Alpaca [28], LIMA [33]) intensify this risk by relying on synthetic supervision [17]. We *suggest under this framing* that such models perform well in-distribution but may fail on extrapolative or adversarial reasoning tasks that demand preserved epistemic depth.

Compression vs. exploration. Recursive KD regularizes the hypothesis class and can reduce variance, but may also truncate exploratory capacity in reasoning. In symbolic domains, this trade-off can simplify search over proof trees while limiting expressive compositional

abstraction, aligning with observed hallucination and brittleness phenomena in large language models [3, 15].

Why symbolic and mathematical reasoning are sensitive. These domains rely on maintaining latent compositional structure (e.g., proof trees, symbolic equivalence classes, logical dependencies). Recursive compression can collapse such structure into surface correlations, reducing capacity for compositional generalization.

How can epistemic degradation be measured? Beyond accuracy, one can probe latent structure [12], evaluate calibration [10], and employ entropy/MDL/mutual-information proxies [1] as quantitative diagnostics. Tasks deliberately outside the teacher's distribution (counterfactual or adversarially constructed mathematical problems) can further reveal retained epistemic depth.

How might distillation be epistemically augmented? If KD imposes an epistemic ceiling by severing access to raw data, it can be mitigated by mixing KD with limited supervision; injecting high-entropy auxiliary objectives [4]; or applying regularization methods such as dropout-as-Bayesian inference [6]. Incorporating structural priors (modularity, disentanglement) may also scaffold reasoning and partially restore epistemic diversity.

Toward Epistemic Evaluation Frameworks. Future evaluation regimes should explicitly test models' epistemic depth rather than surface fluency. Promising directions include: (i) synthetic reasoning datasets that isolate structural generalization (e.g., compositional algebra or logic puzzles); (ii) entropy tracking across intermediate layers as a measure of epistemic diversity; and (iii) counterfactual perturbations designed to expose second-order inference failure. Such diagnostics would operationalize the theoretical framework developed here and provide concrete means of monitoring information reduction in distilled reasoning models.

Open Questions

Formal measures of epistemic fidelity. Can entropy-, MDL-, or mutual-information-based proxies serve as reliable, computable indicators of epistemic degradation during model-to-model transfer?

Grounding in recursive pipelines. What forms of minimal re-grounding, such as curated data, feedback, or hybrid supervision, are sufficient to prevent cumulative information reduction in recursive distillation loops?

Exceeding the teacher. Under what conditions (e.g., capacity expansion, architectural priors, or structural constraints) can a student exceed its teacher's epistemic reach and recover information absent from teacher outputs?

Failure analysis. Can specific hallucination or brittleness phenomena in reasoning models be empirically linked to upstream epistemic loss predicted by the theoretical framework?

Benchmarks and evaluation. What benchmark families best capture epistemic depth, such as counterfactual, compositional, or adversarial mathematical diagnostics, and how can they complement standard accuracy-based metrics?

Conclusion

As knowledge distillation and synthetic supervision become central to LLM development, it is essential to ask not only whether students imitate their teachers, but what *epistemic depth* they retain. We analyzed data-free, model-to-model training as a self-referential loop that constrains representational complexity. Using Kolmogorov complexity and computable information-theoretic proxies (entropy, MDL, and mutual information), we framed recursive distillation as a *monotonic information-reducing process* that can progressively limit epistemic richness even when surface performance remains strong. We further outlined conceptual strategies and diagnostics for preserving epistemic depth, through partial re-grounding, structural priors, and information-based evaluation. Our analysis motivates the development of empirical metrics and benchmarks that directly measure epistemic fidelity, providing a foundation for future work on safe and interpretable model compression.

References

[1] Alessandro Achille and Stefano Soatto. Emergence of invariance and disentanglement in deep representations. *Journal of Machine Learning Research*, 19(50):1–34, 2018.

[2] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017.

[3] 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.

[4] Steven Y Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. A survey of data augmentation approaches for nlp. *arXiv preprint arXiv:2105.03075*, 2021.

[5] Tommaso Furlanello, Zachary Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. Born again neural networks. In *International conference on machine learning*, pages 1607–1616. PMLR, 2018.

[6] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016.

[7] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[8] Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. Minillm: Knowledge distillation of large language models. *arXiv preprint arXiv:2306.08543*, 2023.

[9] Jie Gui, Zhenan Sun, Yonggang Wen, Dacheng Tao, and Jieping Ye. A review on generative adversarial networks: Algorithms, theory, and applications. *IEEE transactions on knowledge and data engineering*, 35(4):3313–3332, 2021.

[10] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR, 2017.

[11] Michael E Henderson and Imran Nasim. Point cloud continuation: Extracting manifolds from observations of a dynamical system. *SIAM Journal on Applied Dynamical Systems*, 24(4):2575–2617, 2025.

[12] John Hewitt and Christopher D Manning. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, 2019.

[13] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[14] Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alexander Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. *arXiv preprint arXiv:2305.02301*, 2023.

[15] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12):1–38, 2023.

[16] Ming Li, Paul Vitányi, et al. *An introduction to Kolmogorov complexity and its applications*, volume 3. Springer, 2008.

[17] Imran Nasim. Governance in agentic workflows: Leveraging llms as oversight agents. In *AAAI 2025 Workshop on AI Governance: Alignment, Morality, and Law*, 2025.

[18] Imran Nasim and Michael E Henderson. Dynamically meaningful latent representations of dynamical systems. *Mathematics*, 12(3):476, 2024.

[19] Imran Nasim and Adam Nasim. Towards a governance framework for generative ai in drug discovery: Ethical, regulatory, and practical challenges. In *AAAI 2025 Workshop on AI Governance: Alignment, Morality, and Law*, 2025.

[20] Imran Nasim and Melanie Weber. Learning reduced order dynamics via geometric representations. In *International Conference on Scientific Computing and Machine Learning*, 2024.

[21] Imran Nasim and Melanie Weber. Automated manifold learning for reduced order modeling. *arXiv preprint arXiv:2506.01741*, 2025.

[22] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

[23] Andrew M Saxe, Yamini Bansal, Joel Dapello, Madhu Advani, Artemy Kolchinsky, Brendan D Tracey, and David D Cox. On the information bottleneck theory of deep learning. *Journal of Statistical Mechanics: Theory and Experiment*, 2019(12):124020, 2019.

[24] Divya Saxena and Jiannong Cao. Generative adversarial networks (gans) challenges, solutions, and future directions. *ACM Computing Surveys (CSUR)*, 54(3):1–42, 2021.

[25] Prajvi Saxena, Sabine Janzen, and Wolfgang Maaß. Streamlining llms: Adaptive knowledge distillation for tailored language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop)*, pages 448–455, 2025.

[26] Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36:8634–8652, 2023.

[27] Yuncheng Song, Liang Ding, Changtong Zan, and Shujian Huang. Self-evolution knowledge distillation for llm-based machine translation. *arXiv preprint arXiv:2412.15303*, 2024.

[28] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. Alpaca: A strong, replicable instruction-following model. *Stanford Center for Research on Foundation Models. https://crfm. stanford. edu/2023/03/13/alpaca. html*, 3(6):7, 2023.

[29] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language models with self-generated instructions. *arXiv preprint arXiv:2212.10560*, 2022.

[30] Junjie Yang, Junhao Song, Xudong Han, Ziqian Bi, Tianyang Wang, Chia Xin Liang, Xinyuan Song, Yichao Zhang, Qian Niu, Benji Peng, et al. Feature alignment and representation transfer in knowledge distillation for large language models. *arXiv preprint arXiv:2504.13825*, 2025.

[31] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.

[32] Qihuang Zhong, Liang Ding, Li Shen, Juhua Liu, Bo Du, and Dacheng Tao. Revisiting knowledge distillation for autoregressive language models. *arXiv preprint arXiv:2402.11890*, 2024.

[33] Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36:55006–55021, 2023.