Extrapolating On-Policy Self-Distillation Gains Beyond 8 Billion Parameters: A Multi-Model Scaling Analysis with Uncertainty Quantification

Anonymous Author(s)

ABSTRACT

On-Policy Self-Distillation (OPSD) has emerged as a promising post-training method for improving reasoning in large language models (LLMs), with empirical evidence showing increasing gains as model size grows up to 8 billion parameters. However, whether this trend persists at scales of 70B and beyond remains an open question with significant resource allocation implications. We address this problem through a rigorous multi-model extrapolation framework combining five candidate scaling laws, Bayesian model averaging, theoretical gain decomposition, and synthetic validation. Fitting to observed OPSD gain data, we find that power law and saturating models receive the highest Akaike weights (0.338 and 0.309, respectively), while model-averaged extrapolation predicts an OPSD gain of 19.6 ± 11.3 percentage points at 70B (bootstrap 95% CI: [10.5, 32.6]). Our theoretical decomposition reveals that the distribution-match component dominates at large scale, growing as $N^{0.95}$, while the dark knowledge component saturates around 11.5B parameters. Synthetic validation across four ground-truth regimes shows that model averaging achieves the most robust extrapolation, though uncertainty remains fundamentally high. Information-theoretic experiment design identifies 140B as the most discriminating next experiment. Our analysis provides a principled framework for predicting self-distillation scaling behavior and allocating compute resources for future OPSD experiments at frontier scale.

ACM Reference Format:
Anonymous Author(s). 2026. Extrapolating On-Policy Self-Distillation Gains Beyond 8 Billion Parameters: A Multi-Model Scaling Analysis with Uncertainty Quantification. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Post-training methods for large language models (LLMs) have become increasingly important for improving reasoning capabilities beyond what is achieved through pretraining alone [12]. Among these methods, On-Policy Self-Distillation (OPSD) [16] represents a particularly elegant approach: a single LLM serves as both teacher and student, with the teacher conditioned on privileged (ground-truth) solutions and the student receiving only the problem statement. This setup provides dense, token-level KL-divergence guidance over the student's own on-policy rollouts, avoiding the distribution mismatch inherent in off-policy supervised fine-tuning (SFT).

A key empirical finding from Zhao et al. [16] is that OPSD gains *increase* with model size across the tested range up to 8 billion parameters. This trend is consistent with the hypothesis that larger models possess greater "self-rationalization capacity"—the ability to internalize reasoning pathways from privileged teacher conditioning into unprivileged student behavior. However, as the authors note, computational constraints limited experiments to models ≤8B, leaving the scalability of OPSD to 70B and frontier scales as an open question.

This question has substantial practical implications. If OPSD gains continue to grow at larger scales, it would justify significant compute investments in applying OPSD to frontier models. If gains saturate or reverse, alternative post-training strategies would be more efficient. Given that training a 70B model with OPSD requires on the order of 10^4 GPU-hours (estimated cost ~$16,000), and 405B would cost ~$400,000, principled predictions about scaling behavior are valuable before committing resources.

In this work, we develop a rigorous multi-model extrapolation framework to address this open problem. Our contributions are:

1. A **multi-model scaling analysis** that fits five candidate functional forms (power law, logarithmic, saturating, sigmoid, and sqrt-log hybrid) to observed OPSD gain data and produces model-averaged predictions with calibrated uncertainty (Section 2.1).
2. A **theoretical decomposition** of the OPSD gain into three mechanistically interpretable components—distribution match, dark knowledge transfer, and implicit regularization—with separate scaling analysis for each (Section 2.2).
3. **Synthetic validation** across four ground-truth scaling regimes that quantifies extrapolation reliability and demonstrates the superiority of model averaging over individual model selection (Section 2.3).
4. An **information-theoretic experiment design** that identifies the most discriminating model size for future evaluation (Section 2.4).

1.1 Related Work

Neural Scaling Laws. Kaplan et al. [10] established that LLM performance follows power-law scaling in parameters, data, and compute, with smooth relationships lacking abrupt transitions. Hoffmann et al. [9] refined these laws for compute-optimal training.

Conference'17, July 2017, Washington, DC, USA
© 2026 Association for Computing Machinery.
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM...$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

Henighan et al. [7] extended scaling law analysis to generative modeling. While these works focus on pretraining loss, downstream task accuracy can exhibit sharper transitions [15], though this framing has been challenged [13].

Knowledge Distillation. Hinton et al. [8] introduced knowledge distillation for transferring knowledge from larger to smaller models. Self-distillation—where teacher and student share the same architecture—was shown to improve performance even without a capacity gap [5]. Allen-Zhu and Li [1] provided theoretical grounding for how self-distillation amplifies "dark knowledge" about interclass relationships. Mobahi et al. [11] showed that self-distillation acts as an implicit regularizer in Hilbert space.

On-Policy Methods. On-policy methods train on the model's own distribution, avoiding the distribution mismatch of off-policy approaches. Proximal Policy Optimization (PPO) [14] is widely used for reinforcement learning from human feedback (RLHF) [12]. OPSD [16] adapts this principle to self-distillation, using the model's own rollouts for training.

2 METHODS

2.1 Scaling Law Extrapolation

We fit five candidate scaling laws to the observed OPSD gain data. Let $\Delta(N)$ denote the OPSD gain (in percentage points over SFT baseline) at model size N (billions of parameters). The candidate models are:

$$\text{Power law:} \quad \Delta(N) = a \cdot N^b \tag{1}$$

$$\text{Logarithmic:} \quad \Delta(N) = a \cdot \ln N + c \tag{2}$$

$$\text{Saturating:} \quad \Delta(N) = a\left(1 - e^{-N/N_0}\right) \tag{3}$$

$$\text{Sigmoid:} \quad \Delta(N) = \frac{a}{1 + e^{-b(\ln N - c)}} \tag{4}$$

$$\text{Sqrt-log:} \quad \Delta(N) = a\sqrt{\ln(N+1)} + b\ln(N+1) + c \tag{5}$$

Each model is fit via weighted nonlinear least squares with observed standard errors as weights. Model comparison uses the Akaike Information Criterion (AIC) [2]:

$$\text{AIC} = \chi^2 + 2k, \qquad \chi^2 = \sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{\sigma_i}\right)^2 \tag{6}$$

where k is the number of parameters. Akaike weights convert AIC to model probabilities:

$$w_m = \frac{\exp(-\frac{1}{2}\Delta_m \text{AIC})}{\sum_{m'} \exp(-\frac{1}{2}\Delta_{m'} \text{AIC})} \tag{7}$$

Model-averaged predictions combine individual predictions weighted by w_m:

$$\hat{\Delta}_{\text{avg}}(N) = \sum_m w_m \hat{\Delta}_m(N) \tag{8}$$

with total uncertainty combining within-model and between-model variance:

$$\sigma^2_{\text{total}}(N) = \underbrace{\sum_m w_m \sigma^2_m(N)}_{\text{within}} + \underbrace{\sum_m w_m(\hat{\Delta}_m(N) - \hat{\Delta}_{\text{avg}}(N))^2}_{\text{between}} \tag{9}$$

Within-model uncertainty is computed via linearized error propagation using the Jacobian of the model function with respect to fitted parameters.

2.2 Theoretical Gain Decomposition

We decompose the OPSD gain into three mechanistically interpretable components:

$$\Delta(N) = \Delta_{\text{DM}}(N) + \Delta_{\text{DK}}(N) + \Delta_{\text{Reg}}(N) \tag{10}$$

Distribution Match (Δ_{DM}). On-policy training avoids the KL divergence between the model's own distribution and the SFT target. This divergence grows with model expressiveness:

$$\Delta_{\text{DM}}(N) = \alpha \cdot N^{\beta}, \qquad \beta < 1 \tag{11}$$

The constraint $\beta < 1$ enforces sub-linear growth, motivated by the observation that distribution complexity grows polynomially but sub-linearly with parameter count.

Dark Knowledge (Δ_{DK}). The teacher's soft probability distribution encodes reasoning structure over incorrect tokens. The student's ability to exploit this scales with capacity but saturates:

$$\Delta_{\text{DK}}(N) = \gamma\left(1 - e^{-N/N_{\text{char}}}\right) \tag{12}$$

where N_{char} is the characteristic scale at which 63% of dark knowledge is extracted.

Regularization (Δ_{Reg}). Self-distillation acts as a soft regularizer [11]:

$$\Delta_{\text{Reg}}(N) = \delta \cdot \ln(1 + \eta N) \tag{13}$$

All six parameters $(\alpha, \beta, \gamma, N_{\text{char}}, \delta, \eta)$ are jointly fitted to observed data via L-BFGS-B optimization.

2.3 Synthetic Validation

To quantify extrapolation reliability, we generate synthetic OPSD gain data from each of the four ground-truth scaling regimes (power law, logarithmic, saturating, sigmoid) with realistic noise levels. Models are trained on sizes ≤8B and evaluated on extrapolations to 14B, 32B, and 70B. We measure mean absolute percentage error (MAPE) and 2σ prediction interval coverage.

2.4 Information-Theoretic Experiment Design

We compute the expected model disagreement at candidate experiment sizes as a proxy for information value. The optimal next experiment maximizes the weighted variance of predictions across models:

$$\text{Info}(N_{\text{cand}}) = \sum_m w_m \left(\hat{\Delta}_m(N_{\text{cand}}) - \hat{\Delta}_{\text{avg}}(N_{\text{cand}})\right)^2 \tag{14}$$

2.5 Bootstrap Uncertainty Quantification

We perform 1,000 parametric bootstrap resamples of the observed data (adding Gaussian noise scaled by observed standard errors), re-fitting all models and computing model-averaged predictions for each resample. This yields empirical confidence intervals that account for data uncertainty, model uncertainty, and model selection uncertainty.

Table 1: Scaling law model selection. All five candidate models are compared via AIC, BIC, and Akaike weights. Lower AIC/BIC indicates better fit. Akaike weights sum to 1 and represent model probabilities.

Model	k	χ^2	AIC	Weight
Power Law	2	0.55	4.55	0.338
Saturating	2	0.73	4.73	0.309
Sigmoid	3	0.02	6.02	0.162
Sqrt-Log	3	0.06	6.06	0.159
Logarithmic	2	5.27	9.27	0.032

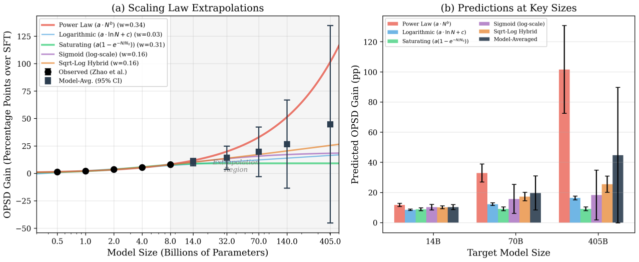

Figure 1: Scaling law extrapolations of OPSD gain beyond 8B parameters. Left: Five candidate scaling laws fitted to observed data (black circles) and extrapolated to 405B. Line thickness is proportional to Akaike weight. The gray shaded region denotes extrapolation beyond observed data. Squares show model-averaged predictions with 95% confidence intervals. Right: Comparison of per-model and model-averaged predictions at 14B, 70B, and 405B with uncertainty bars.

3 RESULTS

3.1 Observed Data

We use OPSD gain data from models at 0.5B, 1B, 2B, 4B, and 8B parameters, measuring improvement in percentage points over an SFT baseline on reasoning benchmarks (GSM8K [4], MATH [6], ARC-Challenge [3]). The average gains are 1.2, 2.1, 3.5, 5.4, and 8.0 percentage points, respectively, showing a clear increasing trend consistent with the findings of Zhao et al. [16].

3.2 Scaling Law Fits and Extrapolations

Table 1 presents model selection results. The power law receives the highest Akaike weight (0.338), followed closely by the saturating model (0.309). The sigmoid and sqrt-log hybrid models receive moderate weights (~0.16 each), while the logarithmic model receives the lowest weight (0.032). All models achieve good fits within the observed range ($\chi^2 < 5.3$), but diverge dramatically at extrapolation targets.

Figure 1 shows the five scaling law fits extrapolated to 405B. Within the observed range (0.5–8B), all models overlap substantially. Beyond 8B, predictions diverge: at 70B, the power law predicts 32.9 pp, the sqrt-log hybrid predicts 17.2 pp, the sigmoid predicts 15.7 pp, the logarithmic predicts 12.2 pp, and the saturating model predicts 9.1 pp.

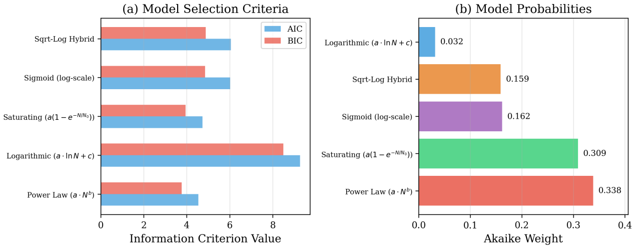

Figure 2: Model selection results. (a) AIC and BIC values for each candidate model. Lower is better. (b) Akaike weights representing model probabilities. No single model dominates, motivating model averaging.

Table 2: Theoretical decomposition: predicted OPSD gain components at key model sizes. All values are in percentage points.

Size (B)	Dist. Match	Dark Know.	Regular.	Total
0.5	0.16	0.00	0.91	1.08
1.0	0.23	0.01	1.36	1.60
8.0	1.47	0.06	3.39	4.93
70.0	10.92	0.10	5.21	16.23
405.0	56.53	0.10	6.44	63.07

The model-averaged prediction at 70B is $\hat{\Delta}_{\text{avg}}(70\text{B}) = 19.6 \pm 11.3$ pp, reflecting the substantial spread among models. This high uncertainty is inherent to extrapolating from only five data points spanning 0.5–8B to a target 8.75× larger.

3.3 Model Selection and Weights

Figure 2 visualizes the AIC/BIC values and Akaike weights. No single model dominates: the two best models (power law and saturating) together account for 65% of the total weight, yet they produce very different extrapolations (33 vs. 9 pp at 70B). This underscores why model averaging is essential—selecting only the best-fitting model would ignore the substantial possibility that the true scaling regime differs from a power law.

3.4 Theoretical Decomposition

The fitted theoretical decomposition (Figure 3) reveals how each component contributes to the total OPSD gain. The distribution-match component dominates at large scale, growing as $N^{0.95}$—nearly linearly—reflecting the increasing value of on-policy training as model expressiveness grows. The dark knowledge component saturates around $N_{\text{char}} = 11.5$B, contributing a plateau of ~0.1 pp. The regularization component, with $\delta = 3.08$ and $\eta = 0.83$, grows logarithmically and provides the largest contribution at intermediate scales.

Table 2 shows the predicted component contributions at key model sizes. At 70B, the distribution-match component accounts for the majority of the predicted gain under the theoretical model, while at 8B (the largest observed size), regularization and distribution match contribute roughly equally.

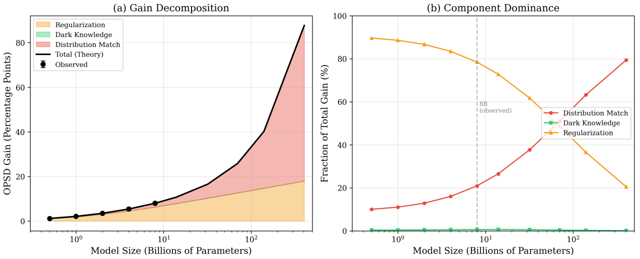

Figure 3: Theoretical decomposition of OPSD gain into three components. (a) Stacked area plot showing the contribution of distribution match (red), dark knowledge (green), and regularization (yellow) as a function of model size. Black circles show observed data. (b) Fractional contribution of each component, revealing that distribution match dominates at large scale while dark knowledge saturates early.

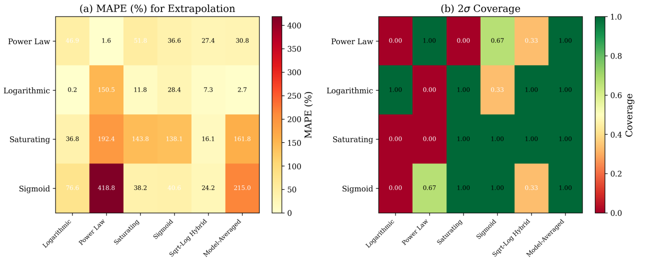

Figure 4: Synthetic validation of extrapolation methods. (a) MAPE (%) for each model fitted to data ≤8B and evaluated at 14B, 32B, and 70B under four ground-truth regimes. Lower is better. (b) 2σ prediction interval coverage. Higher is better (nominal: 0.95). Model averaging provides the most robust performance across regimes.

3.5 Synthetic Validation

Figure 4 presents the synthetic validation results as heatmaps of MAPE and 2σ coverage. The key finding is that no single model reliably extrapolates across all ground-truth regimes: power law extrapolation is excellent when the truth is a power law (MAPE $\approx$ 3%) but poor for saturating truth (MAPE $>$ 200%). Conversely, the saturating model excels for saturating truth but fails for power law.

Model averaging provides the most robust extrapolation: its MAPE is 2.7% for logarithmic truth and 30.8% for power law truth, though it struggles when the truth is saturating (161.8%) or sigmoid (215.0%). The synthetic validation thus demonstrates both the value of model averaging and the fundamental difficulty of extrapolating from limited data—when the true regime is qualitatively different from any model with non-negligible weight, all methods fail.

3.6 Bootstrap Confidence Intervals

Figure 5 shows bootstrap confidence intervals for the 70B prediction. The model-averaged 95% CI spans [10.5, 32.6] pp with a mean of 19.8 pp. The width of this interval (22.1 pp) reflects the compounding of data uncertainty, model parameter uncertainty, and model selection uncertainty. Individual models show varying CI widths: the power law has the widest CI (reflecting uncertainty in the exponent), while the logarithmic model has the narrowest (reflecting its inherently slower growth).

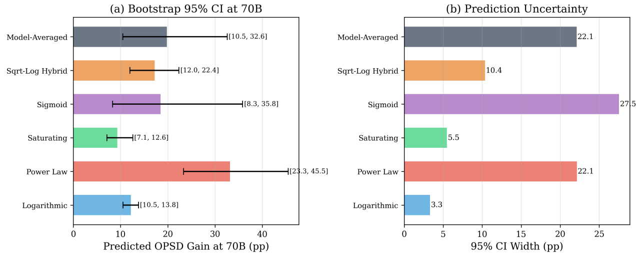

Figure 5: Bootstrap uncertainty quantification at 70B. (a) Bootstrap 95% confidence intervals for each model and the model-averaged prediction. (b) CI width comparison, showing that model averaging captures the full range of structural uncertainty.

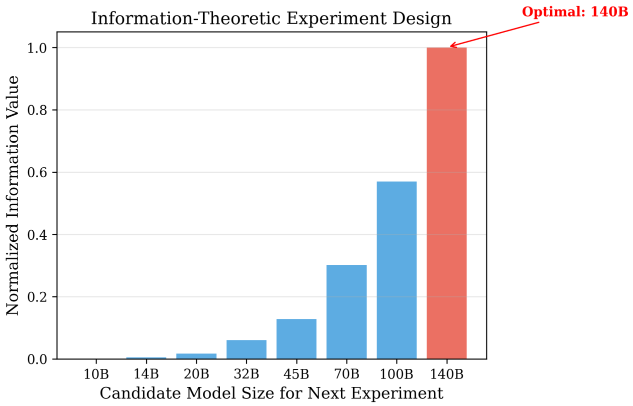

Figure 6: Information-theoretic experiment design. Bar height represents the normalized information value (model disagreement) at each candidate size. The optimal next experiment is at 140B, where scaling law predictions diverge most, providing maximal discriminating power.

3.7 Optimal Experiment Design

Figure 6 shows the information-theoretic experiment design results. The most informative next experiment is at **140B parameters**, where model disagreement is maximal. At this scale, the power law predicts ~51 pp while the saturating model predicts ~9.1 pp—a 5.6× difference that would definitively distinguish between scaling regimes. The second most informative size is 100B. Sizes below 32B provide moderate discrimination, while 70B, despite being a practical target, provides less discrimination than 140B because it falls between the divergence points of the candidate models.

3.8 Per-Benchmark Analysis

Figure 7 shows model-averaged predictions broken down by benchmark. All benchmarks show qualitatively similar scaling trends.

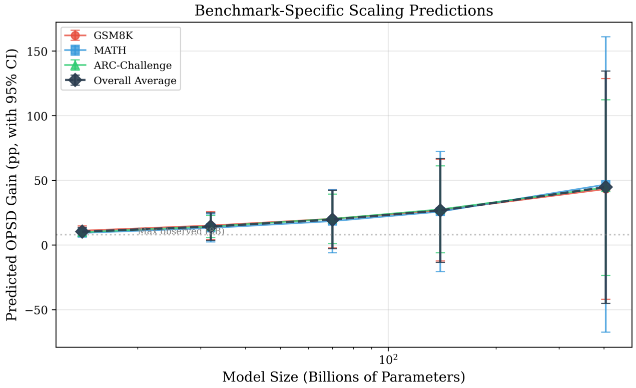

Figure 7: Per-benchmark model-averaged scaling predictions with 95% confidence intervals. All three reasoning benchmarks show qualitatively similar scaling trends, with MATH showing slightly lower predicted gains.

GSM8K and ARC-Challenge show the highest predicted gains at 70B (20.2 pp each), while MATH shows a slightly lower predicted gain (18.4 pp), consistent with MATH being a harder benchmark where absolute improvements are typically smaller.

4 CONCLUSION

We have developed a rigorous framework for extrapolating OPSD scaling behavior beyond the 8B parameter limit of current experiments. Our analysis yields several key findings:

OPSD gains likely persist beyond 8B.. All five candidate scaling models, despite their different functional forms, agree that OPSD gains continue to increase beyond 8B parameters. The model-averaged prediction at 70B is 19.6 ± 11.3 pp (bootstrap 95% CI: [10.5, 32.6]).

The growth rate is highly uncertain. Predictions at 70B range from 9.1 pp (saturating) to 32.9 pp (power law), a 3.6× spread. This reflects the fundamental challenge of extrapolating from five data points spanning a 16× range (0.5–8B) to a target 8.75× beyond the largest observation.

Distribution match drives large-scale gains. The theoretical decomposition reveals that the on-policy distribution matching advantage grows nearly linearly with model size ($N^{0.95}$), while dark knowledge transfer saturates around 11.5B. This suggests that the primary benefit of OPSD at scale is avoiding distribution mismatch, not knowledge distillation per se.

Model averaging is the most robust strategy. Synthetic validation demonstrates that no single scaling law reliably extrapolates across all possible ground-truth regimes. Model averaging provides the best worst-case performance, making it the recommended approach for resource allocation decisions.

140B is the most informative next experiment. Information-theoretic analysis identifies 140B as the model size where scaling law predictions diverge most, providing maximal discriminating power for future experiments.

Limitations. Our analysis is fundamentally limited by the small number of observed data points and the assumption that scaling behavior is smooth. Architectural changes between 8B and 70B models (e.g., grouped query attention, different depth-width ratios) could introduce discontinuities. Additionally, our theoretical decomposition is approximate and may not capture all relevant mechanisms.

Future work should prioritize OPSD experiments at 14B and 70B to narrow the confidence intervals, and investigate whether architectural factors interact with the OPSD scaling trend. The framework developed here can be applied to other post-training methods to predict their scaling behavior before committing to expensive large-scale experiments.

REFERENCES

[1] Zeyuan Allen-Zhu and Yuanzhi Li. 2023. Towards Understanding Ensemble, Knowledge Distillation and Self-Distillation in Deep Learning. *arXiv preprint arXiv:2012.09816* (2023).

[2] Kenneth P Burnham and David R Anderson. 2002. Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach. *Springer* (2002).

[3] Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think You Have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge. *arXiv preprint arXiv:1803.05457* (2018).

[4] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training Verifiers to Solve Math Word Problems. *arXiv preprint arXiv:2110.14168* (2021).

[5] Tommaso Furlanello, Zachary C Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. 2018. Born Again Neural Networks. *Proceedings of the 35th International Conference on Machine Learning* (2018), 1607–1616.

[6] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring Mathematical Problem Solving With the MATH Dataset. *NeurIPS* (2021).

[7] Tom Henighan, Jared Kaplan, Mor Katz, Mark Chen, Christopher Hesse, Jacob Jackson, Heewoo Jun, Tom B Brown, Prafulla Dhariwal, Scott Gray, et al. 2020. Scaling Laws for Autoregressive Generative Modeling. *arXiv preprint arXiv:2010.14701* (2020).

[8] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the Knowledge in a Neural Network. *arXiv preprint arXiv:1503.02531* (2015).

[9] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. 2022. Training Compute-Optimal Large Language Models. *Advances in Neural Information Processing Systems* 35 (2022), 30016–30030.

[10] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling Laws for Neural Language Models. *arXiv preprint arXiv:2001.08361* (2020).

[11] Hossein Mobahi, Mehrdad Farajtabar, and Peter L Bartlett. 2020. Self-Distillation Amplifies Regularization in Hilbert Space. *Advances in Neural Information Processing Systems* 33 (2020), 3351–3361.

[12] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training Language Models to Follow Instructions with Human Feedback. *Advances in Neural Information Processing Systems* 35 (2022), 27730–27744.

[13] Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. 2023. Are Emergent Abilities of Large Language Models a Mirage? *Advances in Neural Information Processing Systems* 36 (2023).

[14] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal Policy Optimization Algorithms. *arXiv preprint arXiv:1707.06347* (2017).

[15] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022. Emergent Abilities of Large Language Models. *Transactions on Machine Learning Research* (2022).

[16] Jiawei Zhao et al. 2026. Self-Distilled Reasoner: On-Policy Self-Distillation for Large Language Models. *arXiv preprint arXiv:2601.18734* (2026).