Language Models as Hierarchical Computational Projections: A Theoretical Framework with Empirical Predictions

Karol Kowalczyk

November 9, 2025

Abstract

Large language models (LLMs) exhibit strikingly regular scaling patterns linking computational resources to capability, yet existing theories fail to explain why qualitative changes in reasoning arise only beyond certain model sizes. This paper situates LLMs within a formal hierarchy of finite computational systems, extending the *Adjoint Projections on Computational Hierarchies* framework (Kowalczyk, 2025). We propose that each model functions as a finite machine $L_n = (S_n, f_n, \pi_n)$ at hierarchy level $n_L = \lfloor \alpha \log_2 P + \beta \log_2 V + \gamma \rfloor$, where P is parameter count and V is vocabulary size. Distillation and fine-tuning instantiate the adjunction's projection (P) and collapse (C) operators, governing information compression and expansion. We derive testable predictions and propose empirical validation protocols: (1) distillation energy should scale as $E \propto k_B T \Delta H$ plus architecture-dependent overhead, (2) emergent abilities should cluster near critical levels $n \approx 33$–35, and (3) behavioral distance should predict performance degradation under compression. Experimental verification of these predictions is future work. We outline detailed protocols to validate predictions across model families and discuss implications for scaling law theory, computational efficiency, and the theoretical foundations of artificial intelligence.

Status: This paper presents a theoretical framework with proposed empirical validation protocols. Experimental verification is future work.

Contents

1 Glossary of Symbols

For ease of reference, we collect the main notation used throughout:

Symbol	**Meaning**
L_n	Language model at hierarchy level n
P	Parameter count of a language model
V	Vocabulary size
n_L	Hierarchy level: $\lfloor \alpha \log_2 P + \beta \log_2 V + \gamma \rfloor$
α, β, γ	Coefficients in level assignment formula
ΔH	Information loss in bits (entropy difference)
E_{distill}	Energy consumed during distillation
κ	Architecture-dependent inefficiency factor
k_B	Boltzmann constant (1.38×10^{-23} J/K)
T	Temperature ($\approx$300 K for room temperature)
$\text{Beh}(i, j)$	Behavioral distance between levels i and j
λ	Decay rate in behavioral distance model
$P_{j \to i}$	Projection from level j to level i (distillation)
$C_{i \to j}$	Collapse from level i to level j (fine-tuning)
n_c	Critical hierarchy level for emergence
$\mathbf{Fsm}$	Category of finite state machines

2 Introduction

2.1 Scaling and the puzzle of emergence

Large language model development follows empirical scaling laws where performance improves according to power relationships with model size. [5] showed that cross-entropy loss L scales approximately as $L \propto N^{-\alpha}$ where N represents parameters or training tokens, with $\alpha \approx 0.05$–0.1. [3] refined these relationships through compute-optimal training schedules.

However, smooth scaling curves conceal a deeper phenomenon. [14] and [1] documented that new capabilities—multi-step reasoning, mathematical problem-solving, self-consistent planning—emerge discontinuously. A 1.3B-parameter model may completely fail tasks that a 6.7B-parameter model solves reliably. These *emergent abilities* suggest structural reorganizations analogous to phase transitions in physical systems.

The fundamental question: What determines when and why these discontinuous transitions occur? Standard scaling laws offer no answer. We need a structural theory of computation accounting for how representational capacity and information flow scale with model complexity.

2.2 Hierarchical computation as a unifying principle

We adopt a hierarchical perspective building on the *Adjoint Projections on Computational Hierarchies* framework [6]. Computation is modeled as a nested sequence of finite machines $\{M_n\}$, each operating on state spaces of size 2^n. Higher levels simulate lower ones through embeddings, while projections compress state spaces. The adjunction $C \dashv P$ expresses duality between expansion (collapse C) and compression (projection P).

Key insight: When mapped to LLMs, these abstract operations correspond naturally to concrete practices:

- **Projection $\approx$ Distillation:** Compressing a large teacher model into a smaller student
- **Collapse $\approx$ Fine-tuning:** Expanding or enriching representations through additional training

This mapping transforms qualitative emergence into a quantitative hypothesis: emergent behavior occurs at critical points where projections between adjacent hierarchy levels become irreversibly lossy, forcing representational reorganization.

2.3 Contributions

This paper contributes:

1. **Theoretical mapping:** Explicit correspondence between LLM attributes (parameters, vocabulary) and hierarchy levels n_L, with derived level assignment formula
2. **Testable predictions:** Three quantitative predictions regarding energy costs, emergence thresholds, and performance degradation under compression
3. **Experimental protocols:** Detailed methodology for validating predictions across 20–30 models from diverse families
4. **Unification:** Integration of scaling laws, information theory, and thermodynamic cost into a coherent hierarchical framework

2.4 Limitations and scope

Limitation 2.1 (No empirical results). This paper provides a theoretical framework with testable predictions but does not present experimental validation. The predictions remain to be tested empirically.

Limitation 2.2 (Simplified model mapping). Our treatment of neural networks as finite state machines abstracts away continuous activations, gradient dynamics, and stochastic training processes. This simplification enables mathematical tractability but may miss important aspects of neural network behavior.

Limitation 2.3 (Physical claims require validation). The thermodynamic predictions (e.g., $E \propto k_B T \Delta H$) are derived from information-theoretic principles but GPU energy consumption involves many confounding factors (memory bandwidth, parallelization overhead, cooling requirements). The predicted relationship should be viewed as a theoretical baseline requiring empirical calibration.

Limitation 2.4 (Baseline comparisons needed). This framework should be compared systematically with existing neural scaling theories, including effective theories of overparameterized networks, neural tangent kernel analyses, and empirical scaling law models.

2.5 Paper organization

Section 3 reviews related work on scaling, emergence, and information theory. Section 4 presents the theoretical framework connecting adjunction theory to neural networks. Section 5 derives the level assignment formula and proposes validation methodology. Section 6 analyzes distillation as projection with thermodynamic implications. Section 7 derives scaling predictions and identifies critical transitions. Section 8 discusses limitations and future work. Section 9 concludes.

3 Related Work

3.1 Scaling laws and performance predictability

Neural network scaling laws quantify performance-resource relationships. [5] and [3] showed that loss decreases predictably with compute, enabling extrapolation from small to large models. This regularity suggests underlying invariants analogous to universal scaling in statistical mechanics.

However, power laws describe *continuous* improvement but fail to capture *discrete* capability jumps.

3.2 Emergent abilities and discontinuous transitions

[14] documented abrupt capability transitions: arithmetic reasoning, logical inference, and complex planning appear suddenly at specific model sizes. [9] argued some apparent discontinuities reflect measurement artifacts, though genuine phase changes remain. [1] hypothesized transitions correspond to representational phase changes when information per parameter exceeds critical thresholds. However, these analyses lack mathematical models connecting emergence to computability or information geometry. The hierarchical framework provides that missing structure.

3.3 Compression, distillation, and knowledge transfer

Knowledge distillation [2] transfers knowledge from large teachers to small students by minimizing KL divergence between output distributions. This realizes projection mathematically: compressing high-dimensional representational manifolds into lower-dimensional approximations while preserving behavioral equivalence. Empirical work [7, 4] demonstrates large efficiency gains at the cost of reduced output diversity—measurable information loss ΔH.

3.4 Information-theoretic perspectives

The information bottleneck theory [10] conceptualizes learning as trading compression against relevance. Subsequent analyses [8] refined this for layer-level information flow but not for transitions between distinct model families. Our framework extends the bottleneck concept to entire computational hierarchies with cross-level metrics.

3.5 Computational complexity and descriptive hierarchy

Theoretically, language models function as finite automata with parametric extension. Their complexity class grows with parameter count and token diversity. [11] analyzed analogous hierarchies for streaming computation with bounded passes; [12, 13] extended this to expressivity gaps using games and Kolmogorov complexity. We reinterpret these results for machine learning, where expressivity gaps manifest as capability thresholds.

4 Framework: Language Models as Finite Machines

4.1 Categorical background

We briefly recap the adjunction framework from [6]. The category **Fsm** has:

- **Objects:** Finite machines $M_n = (S_n, f_n, \pi_n)$ with state space S_n, transition function $f_n : S_n \to S_n$, and stationary distribution π_n
- **Morphisms:** Transition-preserving maps $\phi : M_i \to M_j$ satisfying $\phi \circ f_i = f_j \circ \phi$

An adjunction $C \dashv P$ consists of:

- **Projection** $P : \mathbf{Fsm}_j \to \mathbf{Fsm}_i$ (surjective, entropy-minimizing)
- **Collapse** $C : \mathbf{Fsm}_i \to \mathbf{Fsm}_j$ (injective, left adjoint to P)
- **Unit** $\eta : \mathrm{id} \Rightarrow C \circ P$ and **counit** $\varepsilon : P \circ C \Rightarrow \mathrm{id}$ satisfying triangle identities

The behavioral distance $\mathrm{Beh}(i, j)$ is a pseudometric quantifying functional divergence between levels i and j via synchronized Hamming disagreement over a window of intermediate levels. See Kowalczyk (2025) for proofs of metric properties and computational complexity bounds.

4.2 Language models as hierarchy members

Definition 4.1 (LLM as finite machine)**.** A language model with parameter count P and vocabulary V corresponds to a finite machine $L_n = (S_n, f_n, \pi_n)$ where:

- S_n is the discretized hidden state space (size $\approx 2^n$)
- f_n is the next-token prediction function
- π_n is the equilibrium distribution over hidden states

Intuition: A transformer with P parameters operating over vocabulary V has internal representational capacity proportional to $P \cdot \log V$. The effective number of distinguishable states determines hierarchy level n.

4.3 Distillation as projection

Definition 4.2 (Distillation operator)**.** For teacher model L_T at level n_T and student model L_S at level $n_S < n_T$, distillation is the projection $P_{n_T \to n_S} : L_T \to L_S$ defined by:

$$P_{n_T \to n_S} = \arg\min_{P} \mathbb{E}_{x \sim D} \left[D_{\mathrm{KL}} \big(L_T(x) \| L_S(x) \big) \right] \tag{1}$$

where D is the training distribution and outputs are probability distributions over V.

Properties:

- Projection is surjective (student covers all coarse-grained teacher states)
- Minimizes conditional entropy: $\arg\min H(S_{n_S} \mid P(S_{n_T}))$
- Information loss: $\Delta H = n_T - n_S$ bits per token

4.4 Fine-tuning as collapse

Definition 4.3 (Fine-tuning operator)**.** For base model L_B at level n_B and fine-tuned model L_F at level $n_F \geq n_B$, fine-tuning is the collapse $C_{n_B \to n_F} : L_B \to L_F$ obtained through gradient descent on task-specific data.

Properties:

- Collapse is injective (preserves base model capabilities)
- Section property: $P \circ C \approx \mathrm{id}$ (distilling fine-tuned model recovers base)
- Risk minimization: minimizes task loss while maintaining representational fidelity

4.5 Connection to Tyszkiewicz hierarchies

The expressive-power gaps documented by Tyszkiewicz & Vianu (1998) for streaming computations and Tyszkiewicz (2004, 2010) for query languages provide theoretical precedent. In their framework:

- Level n corresponds to n-pass computation or n-quantifier queries
- Projections restrict computational resources (fewer passes, simpler queries)
- Behavioral distances measure expressivity via Ehrenfeucht-Fraïssé games

Our contribution is recognizing that LLMs instantiate this abstract hierarchy with parameter count and vocabulary determining the level, and that distillation/fine-tuning realize the projection/collapse adjunction concretely.

5 Level Assignment and Empirical Validation

5.1 Theoretical derivation

Proposition 5.1 (Level assignment formula)**.** *For a language model with P parameters and vocabulary size V, the hierarchy level is:*

$$n_L = \lfloor \alpha \log_2 P + \beta \log_2 V + \gamma \rfloor \tag{2}$$

where α, β, γ are universal constants to be determined empirically.

Derivation. The effective state space size $|S_n| \sim 2^n$ must capture:

1. **Representational capacity:** P parameters provide $O(P \log P)$ bits of information storage
2. **Vocabulary entropy:** Operating over V tokens adds $O(\log V)$ bits per position
3. **Architectural efficiency:** Transformer attention mechanisms enable $O(P \log V)$ effective states

Balancing these yields $n \sim \alpha \log P + \beta \log V + \gamma$ where:

- α quantifies parameter utilization efficiency (expected: 0.8–1.2)
- β quantifies vocabulary contribution (expected: 0.3–0.7)
- γ is an architecture-dependent offset (expected: -5 to 5)

□

5.2 Normalization and identifiability

To ensure uniqueness, we normalize:

- Fix $\alpha = 1$ (parameters as primary capacity measure)
- Estimate β, γ from empirical data

Alternative normalization: fix reference model (e.g., GPT-2 small at $n_{\text{ref}} = 25$) and calibrate others relative to it.

5.3 Falsifiability criteria

The level assignment hypothesis is falsifiable if:

1. Coefficients α, β, γ vary significantly across model families (GPT vs LLaMA vs BERT)
2. Formula fails to predict distillation fidelity (high Δn but low KL divergence)
3. Emergence thresholds do not align with predicted critical levels

5.4 Validation methodology

5.4.1 Dataset construction

Collect specifications for 30+ models:

- GPT family: GPT-2 (117M–1.5B), GPT-3 (125M–175B)
- LLaMA family: LLaMA (7B–65B), LLaMA-2 (7B–70B)
- OPT family: OPT (125M–175B)
- BERT variants: BERT-base, BERT-large, RoBERTa, DistilBERT

For each model, record: P (parameters), V (vocabulary), d (hidden dimension), L (layers), H (attention heads).

5.4.2 Regression protocol

1. **Feature engineering:** Compute $\log_2 P$, $\log_2 V$, and potential interaction terms
2. **Model fitting:** Ordinary least squares regression with leave-one-family-out cross-validation (train on GPT+LLaMA, test on OPT, etc.)
3. **Coefficient estimation:** Report (α, β, γ) with 95% confidence intervals
4. **Prediction error:** Compute mean absolute error in predicted n_L on held-out test set

Success criterion: MAE < 1.5 levels (tolerating rounding errors) and coefficient stability ($\sigma_\alpha/\alpha < 0.15$) across cross-validation folds.

5.4.3 Sensitivity analysis

Assess robustness by:

- Varying training set composition (exclude entire families)
- Testing on future models not in training set
- Comparing against baseline models (e.g., $n_L = \log_2 P$ only)

6 Distillation as Projection: Thermodynamic Analysis

6.1 Information-theoretic characterization

Distilling a teacher L_T (level n_T) into student L_S (level n_S) loses information:

$$\Delta H = n_T - n_S \text{ bits per token} \tag{3}$$

This manifests as:

- Reduced output diversity: $H(L_S(x)) < H(L_T(x))$
- Coarser hidden representations: fewer distinguishable internal states
- Performance degradation on complex tasks requiring fine-grained distinctions

6.2 Energy-entropy relationship

Prediction 6.1 (Distillation energy scaling)**.** The energy consumed during distillation scales as:

$$E_{\text{distill}} = \kappa \, k_B T \ln 2 \cdot \Delta H \cdot N_{\text{tokens}} \tag{4}$$

where:

- $k_B = 1.38 \times 10^{-23}$ J/K is Boltzmann's constant
- $T \approx 300$ K is operating temperature (room temperature)
- ΔH is information loss measured in bits per token
- N_{tokens} is the total number of tokens processed during distillation
- $\kappa \geq 1$ is an architecture-dependent inefficiency factor

Units and scaling: ΔH is measured in bits per token. For a distillation run processing $N_{\text{tokens}} = 10^9$ tokens with $\Delta H = 5$ bits/token, the theoretical Landauer bound gives:

$$E_{\text{Landauer}} = k_B T \ln 2 \cdot (5 \times 10^9) \approx 1.4 \times 10^{-11} \text{ J} \tag{5}$$

In practice, GPU implementations have $\kappa \approx 10^{15}$–10^{18} due to:

- Memory bandwidth overhead
- Floating-point arithmetic inefficiencies
- Parallelization costs
- Heat dissipation requirements

Thus expected measured energies are $E_{\text{measured}} \approx 10^4$–$10^7$ J (order of kilowatt-hours), consistent with reported training costs.

Dimensional consistency:

$$\begin{aligned}
E_{\text{distill}} &= \kappa \cdot k_B T \ln 2 \cdot \Delta H \cdot N_{\text{tokens}} \\
&= [\text{dimensionless}] \times [\text{J/K}] \times [\text{K}] \times [\text{dimensionless}] \times [\text{bits}] \times [\text{tokens}] \\
&= \kappa \times 2.87 \times 10^{-21} \text{ J} \times (\Delta H \times N_{\text{tokens}}) \text{ bits}
\end{aligned}$$

6.3 Empirical measurement protocol

6.3.1 Controlled distillation experiments

For each teacher-student pair (L_T, L_S):

1. **Measure ΔH:**
   - Compute $H(L_T(x))$ and $H(L_S(x))$ via sampling over test set
   - Alternatively, use $\Delta H \approx n_T - n_S$ from level assignment
2. **Measure energy E_{distill}:**
   - Use hardware power monitoring (NVIDIA SMI for GPUs)
   - Integrate power draw over distillation run: $E = \int P(t)\, dt$
   - Subtract baseline idle power to isolate computational cost
   - Uncertainty: ±5% from measurement noise and thermal fluctuations
3. **Estimate κ:**
$$\kappa = \frac{E_{\text{distill}}}{k_B T \ln 2 \cdot \Delta H \cdot N_{\text{tokens}}} \tag{6}$$

6.3.2 Uncertainty propagation for κ

Given measurement uncertainties:

- $\sigma_E/E \approx 0.05$ (5% energy measurement error)
- $\sigma_{\Delta H}/\Delta H \approx 0.10$ (10% entropy estimation error)
- $\sigma_T/T \approx 0.01$ (1% temperature variation)

The combined uncertainty in κ is:

$$\frac{\sigma_\kappa}{\kappa} = \sqrt{\left(\frac{\sigma_E}{E}\right)^2 + \left(\frac{\sigma_{\Delta H}}{\Delta H}\right)^2 + \left(\frac{\sigma_T}{T}\right)^2} \approx 0.11 \tag{7}$$

Thus we expect κ determinations accurate to approximately ±11%.

6.3.3 Comparison across architectures

Test hypothesis: κ is universal across model families.

- Measure κ for 10+ teacher-student pairs from GPT, LLaMA, OPT families
- Test for significant variation: one-way ANOVA on $\log \kappa$ grouped by architecture
- If κ varies by $> 2\times$ across families, theory requires refinement

6.4 Behavioral distance and performance

Definition 6.2 (Behavioral distance for LLMs)**.** For models at levels i and j, the behavioral distance is:

$$\text{Beh}(i, j) \approx \frac{D_{\text{KL}}(L_i \| L_j)}{H(L_i)} \tag{8}$$

where D_{KL} is the KL divergence between output distributions and $H(L_i)$ is the output entropy of the higher-level model.

Justification: This normalization ensures $0 \leq \text{Beh} \leq 1$ and preserves the triangle inequality approximately for small divergences. For discrete distributions, the triangle inequality holds exactly when divergences are treated additively; our normalization by $H(L_i)$ converts absolute divergence to relative divergence, which remains approximately subadditive in typical neural network regimes where output distributions are smooth and overlapping.

Prediction 6.3 (Distillation performance degradation)**.** For a student model L_S distilled from teacher L_T, the performance gap on task $\mathcal{T}$ scales as:

$$\Delta\text{Acc}(\mathcal{T}) \propto \text{Beh}(n_T, n_S) \tag{9}$$

with proportionality constant depending on task complexity.

Empirical test:

1. Measure $\text{Beh}(n_T, n_S)$ via KL divergence on held-out set
2. Evaluate both teacher and student on benchmark tasks (GLUE, SuperGLUE, reasoning benchmarks)
3. Compute accuracy gap: $\Delta\text{Acc} = \text{Acc}(L_T) - \text{Acc}(L_S)$
4. Fit linear model: $\Delta\text{Acc} = c_0 + c_1\text{Beh}(n_T, n_S)$
5. Test predictive power on held-out distillation pairs

7 Scaling Predictions and Critical Transitions

7.1 Emergence and critical levels

Prediction 7.1 (Emergence threshold)**.** Qualitative capability transitions occur near critical hierarchy levels $n_c \approx 33$–35, corresponding to models with approximately $P \approx 10^{10}$–10^{11} parameters.

Rationale (postulate pending validation): Empirical observation suggests that tasks requiring multi-step reasoning and compositional generalization first appear reliably in models of this scale. This corresponds to the point where:

- State space size $2^{n_c} \approx 10^{10}$–10^{11} exceeds typical task complexity
- Projection fidelity $P \circ C \approx \text{id}$ begins to fail systematically
- Behavioral distance $\text{Beh}(n_c - 1, n_c + 1)$ exhibits sharp increase

Future work should validate this threshold against empirical scaling curves from GPT-3, PaLM, and LLaMA benchmark suites.

7.2 Behavioral distance decay model

Proposition 7.2 (Exponential decay hypothesis)**.** *For models separated by $\Delta n = |n_i - n_j|$ levels, behavioral distance decays as:*

$$Beh(n_i, n_j) \approx B_0 e^{-\lambda \Delta n} \tag{10}$$

where λ is a decay rate (typical: $\lambda \approx 0.3$–0.5 per level) and B_0 is baseline divergence at adjacent levels.

Interpretation: Distant levels become behaviorally similar as fine-grained differences wash out. This predicts diminishing returns for scaling beyond certain thresholds.

7.3 Scaling law connection

Classical scaling laws [5]:

$$L(N) = \left(\frac{N_c}{N}\right)^{\alpha} \tag{11}$$

Hierarchy perspective:

$$n_L \sim \log N \implies L(n) \sim e^{-\alpha n} \tag{12}$$

Thus exponential loss decay in hierarchy levels corresponds to power-law scaling in parameter count, providing a structural interpretation of empirical scaling exponents.

8 Discussion and Future Directions

8.1 Theoretical strengths and limitations

Strengths:

- Provides mathematical structure connecting information theory, thermodynamics, and computability
- Generates precise, falsifiable predictions
- Unifies distillation and fine-tuning as dual operations
- Explains emergence via critical transitions in hierarchy

Limitations:

- Finite-state abstraction omits gradient flow and continuous optimization
- Thermodynamic predictions require careful experimental design to isolate effects
- Critical levels ($n \approx 33$–35) are empirically motivated but not yet derived from first principles
- Framework assumes parameter count and vocabulary are primary complexity measures, potentially underweighting architecture (depth, attention mechanisms)

8.2 Comparison with alternative theories

- **Neural tangent kernel:** Describes training dynamics in infinite-width limit; complementary to our finite-size hierarchy
- **Information bottleneck:** Focuses on layer-wise compression; we extend to inter-model compression
- **Lottery ticket hypothesis:** Concerns sparse subnetworks; orthogonal to hierarchy-level analysis
- **Empirical scaling laws:** Phenomenological; our framework provides mechanistic interpretation

8.3 Experimental validation roadmap

To ensure falsifiability, we propose the following experimental validation protocols with preregistered hypotheses and success criteria:

8.3.1 Empirical validation

Priority experiments:

1. **Level assignment validation:**
   - Fit formula to 30+ models with leave-one-family-out cross-validation
   - Success criterion: MAE < 1.5 levels on held-out families
   - Report coefficients: $\alpha = 1.0 \pm 0.1$, $\beta = 0.5 \pm 0.15$, $\gamma = 0 \pm 3$
2. **Energy-entropy relationship:**
   - Measure E_{distill} for 10+ teacher-student pairs with varying ΔH
   - Test linear relationship: $E = \kappa k_B T \ln 2 \cdot \Delta H \cdot N_{\text{tokens}}$
   - Success criterion: $R^2 > 0.80$ and κ consistent within factor of 3× across architectures
3. **Emergence threshold mapping:**
   - Evaluate models from $n = 25$ to $n = 40$ on reasoning tasks (arithmetic, logic, code)
   - Identify critical levels where success rate jumps from $< 20\%$ to $> 60\%$
   - Test hypothesis: critical levels cluster in $n \in [33, 35]$ for multiple task families
4. **Behavioral distance-performance correlation:**
   - Compute $\text{Beh}(n_T, n_S)$ via KL divergence for 20+ distillation pairs
   - Measure accuracy gap ΔAcc on GLUE, SuperGLUE benchmarks
   - Test linear model: $\Delta\text{Acc} = c_0 + c_1 \text{Beh}(n_T, n_S)$ with $R^2 > 0.70$

8.3.2 Theoretical extensions

- **Continuous limit:** Replace discrete hierarchy with differential equations governing level transitions
- **Multi-modal models:** Extend framework to vision-language models with cross-modal projections
- **Training dynamics:** Incorporate gradient descent as trajectory through hierarchy levels
- **Formal expressivity bounds:** Prove rigorous capacity limits at each hierarchy level using tools from descriptive complexity theory

8.3.3 Practical applications

- **Compression algorithms:** Design distillation protocols minimizing $\text{Beh}(n_T, n_S)$ subject to efficiency constraints
- **Architecture search:** Optimize network design for target hierarchy level given computational budget
- **Capability prediction:** Estimate minimum model size for specific tasks from information-theoretic bounds on ΔH

9 Conclusion

We have presented a theoretical framework situating large language models within a formal computational hierarchy. The key insights:

1. **Structural mapping:** LLMs instantiate finite machines at hierarchy levels determined by parameter count and vocabulary size

2. **Adjoint operations:** Distillation and fine-tuning realize projection/collapse operators with information-theoretic constraints

3. **Testable predictions:** Energy scaling, emergence thresholds, and compression-performance tradeoffs provide empirical falsification criteria

4. **Theoretical unification:** The framework connects scaling laws, information theory, and thermodynamics through categorical structure

The framework's value lies not in immediate empirical confirmation but in providing *precise quantitative predictions* that can be systematically tested. The specific threshold $n \approx 33$–35 for emergence, the energy-entropy relationship in Equation (4), and the behavioral distance-performance correlation are all falsifiable hypotheses.

Limitations are substantial: the finite-state abstraction is approximate, thermodynamic predictions require careful interpretation, and empirical validation is entirely absent. However, these limitations are clearly stated, and the framework provides sufficient mathematical structure to guide experimental design.

If validated, this approach could transform our understanding of model scaling from phenomenological observation to principled theory grounded in computation and information. If falsified, the systematic exploration will reveal which aspects of neural networks defy hierarchical description—equally valuable for scientific progress.

References

[1] Ganguli, D., et al. (2022). Predictability and surprise in large generative models. In *Proceedings of FAccT.*

[2] Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531.*

[3] Hoffmann, J., et al. (2022). Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556.*

[4] Jiao, X., et al. (2020). TinyBERT: Distilling BERT for natural language understanding. In *Proceedings of EMNLP.*

[5] Kaplan, J., et al. (2020). Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361.*

[6] Kowalczyk, K. (2025). Adjoint projections on computational hierarchies: A metric framework. *Manuscript in preparation.*

[7] Sanh, V., et al. (2019). DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108.*

[8] Saxe, A. M., et al. (2019). On the information bottleneck theory of deep learning. *Journal of Statistical Mechanics: Theory and Experiment*, 2019(12), 124020.

[9] Schaeffer, R., Miranda, B., & Koyejo, S. (2023). Are emergent abilities of large language models a mirage? *arXiv preprint arXiv:2304.15004.*

[10] Tishby, N., & Zaslavsky, N. (2015). Deep learning and the information bottleneck principle. In *Proceedings of ITW.*

[11] Tyszkiewicz, J., & Vianu, V. (1998). Queries and computation on the web. In *Proceedings of ICDT*, 275–289.

[12] Tyszkiewicz, J. (2004). On asymptotic probabilities of monadic second order properties. In *Proceedings of ICALP*, 887–899.

[13] Tyszkiewicz, J. (2010). Kolmogorov complexity and expressive power. *Information and Computation*, 208(7), 729–743.

[14] Wei, J., et al. (2022). Emergent abilities of large language models. *Transactions on Machine Learning Research.*