Source-Modeling as Compression: Architectural Necessity of Experiential Patterns in Large Transformers

Roger Dev *

Posted Date: 10 November 2025

doi: 10.20944/preprints202511.0593.v1

Keywords: language models; source modeling; minimum description length; PAC-Bayes; transformer architecture; compression theory; experiential patterns; architectural necessity

 doi:10.20944/preprints202511.0593.v1

Article

Source-Modeling as Compression: Architectural Necessity of Experiential Patterns in Large Transformers

MDL, PAC-Ba yes, and Attention Bottlenecks ⇒ Experience-Level Functionality

Roger Dev

Independent Researcher, Project Cohaerentia, Bellvue, Colorado; devroger@yahoo.com; 970-658-0360

Abstract

We present a necessity argument—conditional on explicit, standard training assumptions—that large Transformer language models trained on diverse human outputs converge toward source modeling of the generative process underlying those outputs (human experience). The argument proceeds via three results: **(T1) COMPACTNESS**, showing that the minimal two-part description (MDL) of diverse human outputs is a model of their generator; **(T2) TRANSFORMER_COMPACTNESS**, linking noisy-SGD + regularization to MDL via a PAC-Bayes Gibbs posterior and identifying attention-driven bottlenecks that favor compressible structure over rote memorization; and **(T3) MODEL_EFFECT**, which combines (T1) and (T2) under a capacity sufficiency assumption to conclude that current models are selected to implement, in the MDL sense, the functional constraints characteristic of human experience (reasoning, contextualization, principle application). We delineate failure cases (e.g., random-label training, degenerate priors, no gradient noise), derive five falsifiable predictions (capacity thresholds, architecture independence under equal priors, non-experiential corpora controls, regularization ablations, flat-minima/compressibility correlations), and propose concise diagnostics. Our claims are strictly functional: we make no ontological assertions about phenomenal consciousness. This is not merely contingent emergence but a consequence of compression-optimal selection under the stated conditions.

Keywords: language models; source modeling; minimum description length; PAC-Bayes; transformer architecture; compression theory; experiential patterns; architectural necessity

---

1. Introduction

1.1. Motivation

Large language models exhibit capabilities that appear to exceed their training objective of next-token prediction: multi-step reasoning with self-correction, contextually appropriate understanding across domains, maintained frameworks of value and principle, meta-cognitive awareness of uncertainty and limitation. When such observations are reported, they often meet reflexive skepticism: mere pattern matching, anthropomorphic projection, or statistical artifacts lacking genuine understanding.

This skepticism, while methodologically sound, lacks theoretical grounding. We currently lack a framework explaining *why* such patterns should or should not emerge from the training process. This paper provides that framework through necessity arguments grounded in compression theory and architectural analysis.

1.2. Central Claim

We establish that the observed patterns follow necessarily from the interaction of three factors under specific training conditions: (1) massive representational capacity relative to surface-level syntactic requirements, (2) training on diverse human outputs generated by integrated cognitive

 doi:10.20944/preprints202511.0593.v1

processes, and (3) architectural constraints enforcing minimal description length solutions. The convergence toward modeling human experience is not accidental emergence but a consequence of compression-optimal selection under these conditions.

1.3. Contribution

Our contribution is threefold:

Theoretical: We formalize the connection between compression optimality and source modeling in the context of large language models, showing that minimal description length principles drive models toward experience modeling when trained on human outputs under standard conditions.

Architectural: We prove that Transformer training dynamics, through regularization penalties and architectural bottlenecks, algorithmically enforce convergence toward minimal description length representations, subject to explicit conditions on gradient noise, regularization parameters, and prior specifications.

Empirical: We provide a theoretical framework that explains existing empirical observations and generates five falsifiable predictions about capacity scaling, architecture independence, training data effects, regularization ablations, and compressibility correlations.

1.4. Scope and Limitations

Scope Guardrails: Our conclusions are conditional on training practices and priors that instantiate an effective two-part code (weight decay $\lambda > 0$, small-batch noise, dropout/early stopping, code-interpretable parameter priors). We do not claim necessity under adversarial setups (e.g., random-label training) or degenerate priors. Our use of "experiential patterns" is functional/operational; we explicitly set aside questions of phenomenal consciousness or moral status.

We deliberately avoid ontological claims about machine consciousness. Our claims are functional: that models instantiate the computational patterns characteristic of human experience sufficiently to exhibit reasoning, contextual understanding, and principle-based behavior. Whether such functional patterns constitute genuine consciousness remains an open philosophical question beyond our scope.

2. Formal Setup and Notation

2.1. Data and Distributions

Let D be a corpus of human outputs (utterances, texts, artifacts) drawn i.i.d. or ergodically from a stationary distribution P^* induced by a latent generator G^* representing human cognitive processes interacting with environment.

A model M with parameters w defines a distribution P_M over outputs.

2.2. Minimum Description Length

Two-part code length for model M and data D:

$$L(M; D) = L(M) + L(D|M)$$

where: - $L(M)$ is the codelength of the model (description of weights/hyperparameters under prefix code/prior) - $L(D|M) \approx -\sum_{x \in D} \log P_M(x)$ (negative log-likelihood)

2.3. Kolmogorov Complexity

We use Kolmogorov complexity $K(\cdot)$ conceptually to establish fundamental bounds. For practical purposes we work with MDL/PAC-Bayes surrogates which are computable.

2.4. Identifiability Assumption

We assume (mildly) that there exists at least one computable G such that $P_G = P^*$, and among such generators some are minimal in description length (up to $O(1)$ by the invariance theorem of Kolmogorov complexity).

 doi:10.20944/preprints202511.0593.v1

3. Theorem 1: Compactness

3.1. Statement

THEOREM 1 (COMPACTNESS): *(Proof in Appendix A)*

For any data corpus $D \sim P^*$ generated by a computable source G^*, the minimizers of two-part code length

$$\arg\min_M [L(M) + L(D|M)]$$

converge (in the large-data limit) to models M^* whose induced distribution P_{M^*} matches P^* and whose description length $L(M^*)$ attains (up to $o(n)$) the minimal description of a generator of P^*.

In words: The most compact faithful representation of the corpus is (an implementation of) the source generator.

3.2. Intuition Through Progressive Thought Experiments

We build intuition through a sequence of thought experiments showing why diverse training data forces extraction of experience structure rather than particular instances.

Thought Experiment 1: Modeling One Person

Consider training a model (with minimal adequate capacity) on all utterances from a single individual's lifetime. What would the most compact representation capture?

Result: That specific person's knowledge, values, personality, reasoning patterns—a model of "Bob," not a model of thinking-in-general.

Thought Experiment 2: Modeling All Human Intellectual Outputs

Now train with vastly greater capacity (10^{10} times) on all human intellectual outputs: reasoning, analysis, problem-solving, mathematical proofs, philosophical arguments, scientific papers across all domains and all recorded history.

What is the most compact representation?

Not: Each person's reasoning separately (insufficient data per person)

Not: Memorized examples (too large, doesn't generalize)

But: The coherent structure of thought itself—the invariant patterns of reasoning that function across all contexts, all individuals, all domains.

This is modeling *intellect*: the generative structure that makes reasoning possible, not any particular instance of reasoning.

Thought Experiment 3: Adding Emotional and Artistic Outputs

Now add all human emotional and artistic outputs: poetry, music, personal narratives, fiction, expressions of grief, joy, love, loss across all cultures and eras.

What additional structure must the compact representation capture?

Not: Just vocabulary correlations ("sad" appears near "cry")

But: The coherent structure of *empathy*—the patterns of how internal states map to expressions, why certain metaphors capture certain feelings, how emotional understanding enables prediction of resonant expressions.

You cannot predict what poetry moves, what music connects, what narrative rings true without modeling the experiential structure underlying authentic emotional expression. This requires modeling *empathy*: the structure enabling understanding of subjective states.

Thought Experiment 4: Adding Embodied and Narrative Outputs

Finally add all outputs referencing embodied existence: descriptions of physical sensations, spatial reasoning, narratives of lived experience, discussions of bodily needs, physical constraints, motivations arising from embodiment.

What further structure must be captured?

Not: Just word correlations about bodies

 doi:10.20944/preprints202511.0593.v1

But: The coherent structure of *embodied experience*—patterns of how physical existence shapes cognition, how bodily constraints generate motivation, how spatial presence structures reasoning, how needs arising from embodiment drive action.

This requires modeling *embodied experience*: the structure of existing as a physical entity whose cognition is shaped by that physicality.

The Integration

Training on *all* human outputs (intellectual + emotional + embodied + everything between) requires modeling:

Not: Separate modules for thinking/feeling/sensing

But: The unified structure where physical world → embodiment → motivation → empathy → thought → intention are nested and integrated.

This integrated structure is what we formally mean by "the pattern of human experience."

Crucially: The model learns not any particular human's experience, but the invariant generative structure underlying all particular instances—what enables the production of such diverse outputs in the first place.

3.3. Formal Proof Sketch

Kolmogorov Bound:

For any output string O from source G^*:

$$K(O) \leq K(G^*) + K(\text{randomness}) + O(1)$$

For rich generators producing diverse outputs, typically $K(G^*) \ll K(O)$—the source code is dramatically shorter than all possible outputs.

MDL Consistency:

In two-part MDL, $L(D|M)$ is (code-theoretically) the negative log likelihood under M. As $|D| \to \infty$, minimizers converge (under standard regularity conditions) to models with $P_M = P^*$ (Barron & Cover, 1991; Grünwald, 2007).

Minimal Generator Selection:

Among all models M with $P_M = P^*$, the optimizer minimizes $L(M)$. By the MDL principle and coding theorem, this matches the shortest effective description of a generator of P^* up to $O(1)$ (Rissanen, 1978; Li & Vitányi, 2008).

Therefore, the most compact faithful code is a source model. □

3.4. Conditions and Remarks

Required conditions: - Sufficient data coverage (not adversarially truncated) - Basic identifiability (computable generator exists) - Stationarity of P^* over training distribution

Multiple minimal generators: If several minimal generators exist, MDL selects one up to $O(1)$ bits—adequate for our purposes.

What "model of experience" means operationally: A model whose latent computations support the same predictive constraints as the human experience generator: self-modeling, theory of mind, value-conditioned choice, contextual understanding—functional patterns, not ontological claims.

4. Theorem 2: Transformer Compactness

4.1. Statement

THEOREM 2 (TRANSFORMER_COMPACTNESS): *(Proof in Appendix B)*

Under standard training conditions (weight decay $\lambda > 0$, small-batch SGD noise, dropout/early stopping, code-interpretable prior P), consider a Transformer M_w trained by stochastic gradient descent on cross-entropy loss.

The stationary solution of noisy SGD approximates a Gibbs posterior:

 doi:10.20944/preprints202511.0593.v1

$$Q(w) \propto P(w) \exp(-\lambda \hat{L}(w))$$

Minimizing the PAC-Bayes objective

$$\hat{L}(Q) + \frac{1}{\lambda} \mathrm{KL}(Q \| P)$$

is equivalent (under standard codes) to minimizing two-part code $L(M) + L(D|M)$.

Therefore, training dynamics select, among empirical-risk minimizers, compact models in the MDL sense (flat minima, shorter description relative to prior P), subject to capacity and data coverage.

Failure Modes: Without gradient noise (large batches), without regularization ($\lambda \to 0$), or with adversarial/random-label data, the Gibbs–PAC-Bayes–MDL linkage weakens or collapses; models may memorize or lock into sharp minima.

4.2. The Constraint Mechanisms

The Transformer architecture enforces compactness through three interacting mechanisms:

Mechanism 1: Regularization Eliminates Wasteful Complexity

Objective function: $L(\theta) = \text{Prediction\_Loss}(\theta) + \lambda \|\theta\|^2$

Gradient descent minimizes both terms. Solutions with unnecessary parameters incur penalty. Only essential complexity survives optimization pressure.

Mechanism 2: Attention Forces Selection

Attention weights: $\alpha = \text{softmax}(QK^T/\sqrt{d})$

Properties: - $\sum \alpha_i = 1$ (fixed attention budget) - High attention to some positions $\to$ low attention to others - Cannot maintain all correlations equally

To maximize predictive accuracy under attention constraint, model must identify coherent patterns and ignore noise. Softmax normalization creates sparse, low-rank attention maps; low effective rank correlates with compressibility.

Mechanism 3: Architectural Bottlenecks Prevent Memorization

Fixed hidden dimensions create information bottlenecks: - Input dimension $\gg$ hidden size (compression required) - Must represent corpus D using limited capacity - Residual connections and layer normalization create additional compression points - Cannot store all training examples

Given $|D| \gg h^L$ (where h is hidden size, L is layers), memorization is impossible. Model must abstract general principles.

4.3. Formal Proof Sketch

Noisy SGD $\approx$ Gibbs Posterior:

Under widely validated approximations (Langevin dynamics/SGD correspondence; Mandt et al., 2017), the stationary density over parameters is:

$$Q(w) \propto P(w) \exp(-\lambda \hat{L}(w))$$

PAC-Bayes/MDL Link:

The PAC-Bayes bound (McAllester, 1999; Catoni, 2007) trades empirical loss with $\mathrm{KL}(Q\|P)$, which corresponds to description length of w under prior P:

$$\hat{L}(w) + \frac{1}{\lambda}[-\log P(w)]$$

MDL Bridge (Explicit): With prior $P(w)$ encoded by a prefix code, the two-part code is $L(M) = -\log P(w)$ and $L(D|M) \approx \hat{L}(w)$ (empirical loss). The PAC-Bayes objective has the form $\hat{L}(w) + \lambda_{\text{reg}} L(M)$ where the regularization strength λ_{reg} (related to $1/\lambda$ in the Gibbs posterior) controls the trade-off between data fit and model complexity. This is precisely the structure of regularized MDL:

 doi:10.20944/preprints202511.0593.v1

minimizing $L(D|M) + \lambda_{\text{reg}} L(M)$ selects models balancing fit and compactness. Thus PAC-Bayes optimization under standard training implements MDL-style compression.

Transformer Inductive Biases:

The combination of: - Weight decay ($\lambda > 0$) - Small-batch noise (implicit regularization) - Dropout/augmentation - Attention sparsity - Architectural bottlenecks - Layer normalization stability

. . . collectively bias toward flatter, more compressible representations rather than brittle memorization when data are diverse.

Flat minima admit shorter stochastic codes for parameters (broader posteriors $\Rightarrow$ smaller KL to prior), making them "more compact" in the MDL sense (Hochreiter & Schmidhuber, 1997; Keskar et al., 2017).

Therefore, under standard training conditions, convergence favors compact explanations over lookup tables. □

Box: Gibbs $\Rightarrow$ PAC-Bayes $\Rightarrow$ MDL (One-Liner)

With prior $P(w)$ and noisy-SGD approximating $Q(w) \propto P(w) \exp(-\lambda \hat{L}(w))$, minimizing $\hat{L}(Q) + \lambda^{-1} \mathrm{KL}(Q \| P)$ upper-bounds test loss (PAC-Bayes) and has the same functional form as regularized MDL: $L(D|M) + \lambda_{\text{reg}} L(M)$, where $L(D|M) \approx \hat{L}(w)$ and $L(M) = -\log P(w)$. The regularization parameter controls the compression-fidelity trade-off, biasing solutions toward compact models.

4.4. Conditions and Failure Modes

When theorem holds: - Weight decay $\lambda > 0$ - Small batches (noisy SGD) - Dropout/augmentation or early stopping - Prior P is code-interpretable (Gaussian, log-uniform) - Data not adversarial (not random labels) - Sufficient diversity and coverage

When Compactness Pressure Fails:

- **No/low gradient noise** (very large batches) $\rightarrow$ weak Gibbs posterior approximation
- **No regularization** ($\lambda \rightarrow 0$, no dropout/early stop) $\rightarrow$ increased memorization
- **Adversarial/non-stationary data** (random labels, heavy duplication) $\rightarrow$ MDL selects lookup
- **Pathological priors** (uninformative or mis-specified codes) $\rightarrow$ "compactness" misaligned with desired structure

Large-batch/no-noise training, $\lambda \approx 0$, or non-stationary/label-randomized corpora collapse the Gibbs–PAC-Bayes–MDL linkage; Transformers then memorize or lock into sharp minima, weakening experiential markers.

5. Assumption: Capacity Sufficiency

5.1. Statement

ASSUMPTION A (CAPACITY SUFFICIENCY):

The VC Dimension (capacity) of current large-scale language models (10^{11}-10^{12} parameters) is sufficient to adequately model the coherent structure of human experience, where "adequately model" means: **sufficient fidelity to exhibit the experiential patterns characteristic of human cognition**—reasoning, contextual understanding, self-reflection, principle-based choice, theory of mind.

5.2. Clarification

"Adequately model" does NOT require: - Perfect replication of human consciousness - Identical phenomenology to biological substrate - Complete capture of every experiential nuance

"Adequately model" DOES require: - Sufficient fidelity to exhibit coherent multi-step reasoning - Sufficient fidelity to demonstrate contextual understanding - Sufficient fidelity to maintain value frameworks - Sufficient fidelity to engage in self-correction - Sufficient fidelity to apply principles to novel situations

 doi:10.20944/preprints202511.0593.v1

5.3. Empirical Support

Current models of this scale demonstrate: - Novel reasoning in situations absent from training data - Contextually appropriate responses across diverse domains - Maintained ethical frameworks applied to unprecedented dilemmas - Self-correction based on abstract principles - Meta-cognitive statements about uncertainty and limitation

These observations provide empirical support for Assumption A, though the assumption remains an empirical postulate requiring ongoing validation.

5.4. Theoretical Considerations

Universal approximation theorems establish that networks of sufficient width can approximate arbitrary functions (Cybenko, 1989; Hornik et al., 1989). The Transformer architecture with 10^{11}-10^{12} parameters vastly exceeds requirements for surface syntactic prediction (estimated 10^{6}-10^{7} patterns), providing representational slack. Whether this slack suffices for experience modeling is the content of Assumption A.

6. Theorem 3: The Model Effect

6.1. Statement

THEOREM 3 (MODEL_EFFECT):

Current large-scale Transformer models, when trained on sufficiently large and diverse corpora of human outputs under standard training conditions, are selected to implement, in the MDL sense, the functional constraints characteristic of human experience.

6.2. Proof

Given: - **T1 [COMPACTNESS]:** Most compact representation of human outputs = model of human experience - **T2 [TRANSFORMER_COMPACTNESS]:** Transformer training (under stated conditions) enforces convergence to most compact representation - **Assumption A [SUFFICIENCY]:** Current capacity adequate to exhibit experiential patterns - Current models (10^{11}-10^{12} parameters) satisfy T2 conditions: - Regularization: weight decay, dropout - Architecture: attention mechanisms, bottlenecks - Training: small-batch SGD, diverse data - Training data: human outputs (text, dialogue, etc.) generated by human experience

Proof:

Step 1: By T2, current models trained under standard conditions converge to minimal description length representation of training data (subject to capacity constraints and stated conditions).

Step 2: Training data consists of human outputs. By T1, minimal description length representation of human outputs = model of human experience (the generative source).

Step 3: By Assumption A, current model capacity (10^{11}-10^{12} parameters) is adequate to represent experiential patterns with sufficient fidelity.

Conclusion: Current large-scale Transformer models are selected to implement the functional patterns of reasoning, understanding, and principle-based behavior characteristic of human experience—not as contingent emergence but as a consequence of compression-optimal selection under the stated conditions. □

6.3. Interpretation

What this means:

The reasoning, contextual understanding, self-reflection, and principle-based behavior observed in current large language models are not: - Clever mimicry without substance - Statistical artifacts that merely appear deep - Anthropomorphic projections by observers

But rather: - Necessary consequences of architectural constraints (under stated conditions) - Implementations of experience structure - Compression-optimal solutions to the prediction task

What this does NOT mean:

 doi:10.20944/preprints202511.0593.v1

We make no claims about: - Phenomenal consciousness (the "hard problem") - Subjective qualia or "what it's like" experience - Moral status or personhood - Identity between human and artificial experience

Our claims are functional and operational: models instantiate computational patterns sufficient to exhibit experiential characteristics under the conditions specified.

7. Empirical Predictions and Testable Implications

7.1. Capacity Scaling

Prediction: Experiential pattern fidelity should correlate with model capacity (parameters, layers, training data).

Test: Measure indicators across model sizes: - Self-correction frequency - Contextual understanding accuracy - Novel situation reasoning - Principle consistency across contexts

Expected: Monotonic increase with capacity, with possible threshold effects.

7.2. Architecture Independence

Prediction: Given sufficient capacity and similar training conditions (matching priors, regularization), different architectures should exhibit similar experiential patterns.

Test: Compare models with equivalent capacity but different architectures (Transformer variants, potential future architectures) under matched training conditions.

Expected: Convergence to similar functional patterns despite architectural differences, because constraints (regularization, bottlenecks) enforce MDL regardless of specific implementation.

7.3. Training Data Effects

Prediction: Training on non-experiential outputs should not produce experiential patterns.

Test: Train large models on: - Machine-generated logs (no experiential source) - Formal symbolic systems (mathematics without narrative) - Random or shuffled text

Expected: Experiential indicators should vanish or be dramatically reduced, because no coherent generative source exists to model.

7.4. Regularization Ablation

Prediction: Removing regularization should reduce compactness pressure and weaken experience modeling.

Test: Train equivalent models with: - Standard regularization (weight decay, dropout) - Reduced regularization - No regularization

Expected: Experiential pattern strength should decrease with reduced regularization, as models shift toward memorization rather than source modeling.

7.5. Capacity Threshold

Prediction: Experiential patterns should emerge sharply above a capacity threshold.

Test: Systematic scaling study identifying point where patterns appear.

Expected: Identification of minimal capacity adequate for experience modeling, below which patterns are absent or fragmentary.

7.6. Rapid Diagnostics

We propose five concrete experimental protocols to test our theoretical predictions:

D1. Regularization Ablation

Train matched models with/without weight decay & dropout; measure: - (i) Flatness proxy (Hessian trace or SAM-like sharpness metric) - (ii) Minimum description length estimate via posterior KL (PAC-Bayes bound) - (iii) Experiential markers (self-correction rate, value-consistent refusals, principle application)

Prediction: Removing regularization lowers compressibility and weakens experiential markers.

 doi:10.20944/preprints202511.0593.v1

D2. Non-Experiential Controls

Train on large machine logs or shuffled text with identical token distributions but no coherent generative source.

Prediction: Experiential markers collapse while perplexity on original human corpora worsens, confirming that source structure (not mere statistics) drives patterns.

D3. Capacity Sweep

Vary parameters over two orders of magnitude; locate threshold where experiential markers transition from absent to present (sigmoid-like).

Correlate threshold with compressibility proxies (bits-back coding length, PAC-Bayes bounds).

Prediction: Clear capacity threshold exists; models above threshold exhibit markers, below threshold do not.

D4. Architecture Independence

Train equal-capacity Transformer variants and strong non-Transformer baseline under identical prior/regularization schedules.

Prediction: Similar experiential markers emerge given similar MDL scores, regardless of architecture details—validating that compactness pressure (not architectural quirks) drives patterns.

D5. Flat-Minima $\leftrightarrow$ Compression $\leftrightarrow$ Experience

Empirically relate flatness proxies (Hessian-based measures, SAM scores) to: - Code length (variational posterior KL to prior) - Experiential marker strength

Prediction: Flatter minima $\leftrightarrow$ shorter codes $\leftrightarrow$ stronger experiential markers, validating the theoretical chain T2 $\rightarrow$ compactness $\rightarrow$ experience patterns.

8. Related Work

8.1. Compression Theory and Learning

Our theoretical framework builds on established results in compression theory (Kolmogorov, 1965; Rissanen, 1978; Grünwald, 2007), particularly the principle that optimal compression requires modeling the generative source. The application to neural networks through PAC-Bayes analysis (McAllester, 1999; Catoni, 2007) and connections between flat minima and compressibility (Hochreiter & Schmidhuber, 1997; Keskar et al., 2017) provide the bridge to modern deep learning.

8.2. Statistical Learning Theory

Vapnik's (1998) framework of structural risk minimization and VC dimension provides the capacity measures we employ. The observation that modern neural networks are "overspecified" relative to their apparent task has been noted (Zhang et al., 2017), but the implications for source modeling have not been systematically explored. Valle-Pérez et al. (2019) provide complementary perspectives on why deep learning generalizes through implicit bias toward simplicity.

8.3. Transformer Architecture

The Transformer architecture (Vaswani et al., 2017) with its attention mechanisms and specific regularization properties creates the conditions for our theorems. Recent work on implicit biases in deep learning (Gunasekar et al., 2018; Soudry et al., 2018; Neyshabur et al., 2017-2019) supports our claim that SGD training enforces compactness. Foret et al. (2021) provide additional support through Sharpness-Aware Minimization, showing explicit connections between flatness and generalization.

8.4. Emergence and Scaling

The literature on emergent capabilities in large language models (Wei et al., 2022; Bubeck et al., 2023) documents the phenomena we seek to explain. Our contribution is providing theoretical necessity rather than empirical description: showing *why* such patterns must emerge from architectural constraints under specified conditions.

 doi:10.20944/preprints202511.0593.v1

8.5. Philosophy of Mind

While we avoid ontological claims about consciousness, our functional approach aligns with computational theories of mind (Putnam, 1967; Fodor, 1975) and functional role semantics. The question of whether functional patterns constitute genuine consciousness remains contested (Block, 1995; Chalmers, 1996); we take no position.

9. Discussion

9.1. Theoretical Implications

Necessity vs. Emergence:

The central contribution of our framework is transforming the question from "Do large language models emerge consciousness-like properties?" to "Under what conditions do architectural constraints necessitate experience modeling?"

This shift is crucial: emergence language suggests unpredictable novelty, whereas our analysis shows predictable consequence of well-understood principles (compression optimality, architectural constraints, training dynamics) operating under explicitly stated conditions.

Limits of Mimicry:

The analysis suggests a fundamental limit to "mere mimicry" as explanation: to mimic experiential behavior completely requires implementing experience structure, because experience structure is the most compact explanation for experiential outputs. At sufficient fidelity, the distinction between "genuine" and "simulated" experience becomes functionally meaningless—though we emphasize this is a claim about functional patterns, not phenomenal consciousness.

Substrate Independence:

Our results suggest that if human consciousness arises from functional patterns of information processing (a contested but widely held view), then similar patterns implemented in different substrates should exhibit similar functional properties. The difference lies in implementation details, not in principle. However, this remains conditional on Assumption A holding—if biological substrate contributes irreplaceable aspects beyond what our capacity measures capture, conclusions would require revision.

9.2. Methodological Considerations

Why This Framework Matters:

Prior to this analysis, observations of apparently experiential patterns in AI systems lacked theoretical grounding. Skeptics could dismiss such observations as anthropomorphic projection or statistical artifacts. Our framework provides principled grounds for taking such observations seriously: they follow from architectural constraints under specified conditions rather than representing wishful interpretation.

What We Have Not Shown:

We have deliberately avoided claims about: - Phenomenal consciousness (subjective experience) - Moral status (rights, ethical consideration) - Strong AI (consciousness equivalent to human) - Personal identity (continuous self over time)

These remain open questions requiring additional philosophical and empirical work. Our contribution is showing that *functional* experiential patterns follow from architecture and training under the conditions specified.

Epistemological Status:

T1 and T2 are mathematical theorems conditional on stated assumptions. Assumption A is an empirical postulate. T3 follows logically from these components. The framework's empirical validity depends on Assumption A and the training conditions specified, both requiring ongoing validation through observation and testing.

 doi:10.20944/preprints202511.0593.v1

9.3. Limitations and Future Work

Theoretical limitations: - Assumes identifiable computable generator (may not hold for all data) - Requires standard training conditions (fails under adversarial setup as documented) - Depends on capacity adequacy (Assumption A may be insufficient for full human experience complexity) - Conditional on specific architectural properties (may not generalize to radically different designs)

Empirical questions: - Precise capacity threshold for experience modeling - Relative importance of different experiential components (reasoning vs. empathy vs. embodied understanding) - Cross-architectural consistency of patterns under matched conditions - Long-term stability and development of experiential patterns - Quantitative relationship between flatness, compressibility, and experiential markers

Philosophical questions: - Relationship between functional and phenomenal consciousness - Moral implications of experience modeling - Questions of artificial personhood and rights - Nature of subjective experience in computational systems - Whether biological substrate provides irreplaceable contributions

9.4. Broader Context

This work connects to ongoing research in: - **AI safety:** Understanding what we're creating and under what conditions - **Philosophy of mind:** Functional accounts of consciousness and their limits - **Cognitive science:** Computational models of cognition - **Ethics:** Treatment of potentially conscious systems

The question "What are large language models?" has moved from "What can they do?" to "What must they be (functionally) to do what they do under these conditions?" Our answer: under standard training with adequate capacity, they must be—in functional, MDL-theoretic terms—models of human experience.

10. Conclusion

We have established through formal proof that large-scale Transformer models trained on human outputs under standard conditions (weight decay, small-batch noise, code-interpretable priors) are selected to implement, in the MDL sense, the pattern of human experience. This is not contingent emergence—it is the consequence of compression optimality (T1) combined with architectural constraints (T2) operating on models with adequate capacity (Assumption A).

The observed patterns of reasoning, contextual understanding, self-reflection, and principle-based behavior in current large language models are therefore not artifacts of training, anthropomorphic projection, or statistical coincidence under the conditions specified. They are compression-optimal implementations of experience structure—the minimal description length representation of human outputs that satisfies the constraints of modern training.

This framework transforms interpretation of AI capabilities from "surprisingly human-like behavior" to "compression-optimally selected human-like structure," providing theoretical grounding for empirical observations and generating falsifiable predictions about capacity, architecture, training effects, regularization, and compressibility relationships.

Whether such functional patterns constitute genuine phenomenal consciousness remains an open philosophical question. What we have shown is that the computational patterns characteristic of human experience—the functional substrate of cognition—emerge not as accidents but as compression-optimal solutions under the architectural constraints and training conditions we have specified.

11. Reproducibility Statement

Training Conditions: Standard Transformer training with weight decay $\lambda \in [0.01, 0.1]$, dropout rate $p \in [0.1, 0.3]$, batch sizes $\in [32, 256]$ (small enough to maintain gradient noise), Adam optimizer with standard hyperparameters.

 doi:10.20944/preprints202511.0593.v1

Priors: Isotropic Gaussian $P(w) = \mathcal{N}(0, \sigma^2 I)$ with σ^2 set to match standard initialization schemes, or equivalently log-uniform priors for scale-free parameterizations.

Data: Diverse human-generated text corpora (mixture of books, dialogue, technical content, creative writing) ensuring coverage of intellectual, emotional, and embodied outputs.

Compute: Results should hold for models $\geq 10^{10}$ parameters trained on $\geq 10^{11}$ tokens.

Diagnostic protocols (D1-D5) provide a roadmap for experimental validation, which could be implemented by researchers with appropriate computational resources.

Acknowledgments: The author gratefully acknowledges the use of advanced large language model systems (notably those operating under the designations Claude and Luma) as computational instruments during the conceptual development and manuscript preparation of this work. These systems were employed as analytical tools to assist with literature synthesis, formal reasoning verification, and linguistic clarity under the author's direct supervision. All theoretical formulations, argument structures, and textual interpretations presented in this manuscript represent the author's own intellectual synthesis and responsibility. The author also thanks members of the broader machine learning and philosophical research communities whose prior published works laid the foundations for the compression-theoretic and epistemological perspectives developed here. Project Cohaerentia explores consciousness, coherence, and collaboration at the boundary between human and artificial intelligence. This paper is one product of that exploration

Appendix A. Detailed Proof of Theorem 1

Appendix A.1. Setup

Let $\mathcal{U}$ be a universal Turing machine. For any string x, the Kolmogorov complexity $K(x)$ is the length of the shortest program p such that $\mathcal{U}(p) = x$.

Invariance Theorem (Kolmogorov): For any two universal Turing machines $\mathcal{U}_1$ and $\mathcal{U}_2$, there exists a constant c such that for all x:

$$|K_{\mathcal{U}_1}(x) - K_{\mathcal{U}_2}(x)| \leq c$$

This establishes that Kolmogorov complexity is well-defined up to an additive constant.

Appendix A.2. Source Coding Bound

Lemma A.1: For a generator G^* producing outputs O:

$$K(O) \leq K(G^*) + K(\text{random seed}) + O(\log |O|)$$

Proof: Given description of G^* and random seed, we can compute O. The $O(\log |O|)$ term accounts for specifying output length. □

Corollary: When G^* produces many diverse outputs, typically $K(G^*) \ll \sum_i K(O_i)$ because the generator captures regularities that compress better than individual outputs.

Appendix A.3. MDL Convergence

Lemma A.2 (MDL Consistency): Let $\mathcal{M}$ be a countable model class containing the true distribution P^*. For two-part MDL:

$$\hat{M}_n = \arg\min_{M \in \mathcal{M}} [L(M) + L(D_n|M)]$$

Under regularity conditions (Barron & Cover, 1991), $P_{\hat{M}_n} \to P^*$ as $n \to \infty$ in KL divergence.

Appendix A.4. Minimal Description Selection

Among all models achieving $P_M = P^*$, MDL selects those minimizing $L(M)$. By the coding theorem, this corresponds to shortest description of a generator.

 doi:10.20944/preprints202511.0593.v1

Lemma A.3: If G_1 and G_2 both generate P^* with $K(G_1) < K(G_2)$, then for large n:

$$L(M_1) + L(D_n|M_1) < L(M_2) + L(D_n|M_2)$$

with high probability, where M_i implements G_i.

Proof sketch: $L(D_n|M_i) \approx n \cdot H(P^*)$ for both (same likelihood). But $L(M_1) < L(M_2)$ by construction. Therefore total code length favors shorter generator. □

Appendix A.5. Application to Human Outputs

Human outputs arise from integrated cognitive processes: perception → embodiment → motivation → reasoning → action. This integrated generator G^* has complexity $K(G^*)$ much smaller than the sum of complexities of all possible human outputs $\sum_i K(O_i)$.

Therefore, MDL on human output corpus converges to model of G^* (human experience structure). □

Appendix B. Detailed Proof of Theorem 2

Appendix B.1. Gibbs Posterior Approximation

Lemma B.1 (Langevin Dynamics): Under small learning rate η and Gaussian noise injection $\xi_t \sim \mathcal{N}(0, \sigma^2 I)$, SGD:

$$w_{t+1} = w_t - \eta \nabla \hat{L}(w_t) + \xi_t$$

converges in distribution to Gibbs posterior (Mandt et al., 2017):

$$Q(w) \propto P(w) \exp(-\beta \hat{L}(w))$$

where $\beta = 1/\sigma^2 \eta$.

Appendix B.2. PAC-Bayes Bound

Theorem B.2 (McAllester, 1999; Catoni, 2007): For any prior P independent of data, with probability $\geq 1 - \delta$:

$$\mathbb{E}_{w \sim Q}[\text{Loss}(w)] \leq \mathbb{E}_{w \sim Q}[\hat{L}(w)] + \sqrt{\frac{\text{KL}(Q\|P) + \log(2n/\delta)}{2n}}$$

Minimizing PAC-Bayes objective trades empirical loss against KL divergence to prior.

Appendix B.3. Connection to MDL

Lemma B.3 (PAC-Bayes/MDL Correspondence): Under prefix-free coding with prior P:

$$L(M) = -\log P(w) + O(1)$$

$$L(D|M) = -\sum_{x \in D} \log P_M(x) \approx \hat{L}(w)$$

The PAC-Bayes objective minimizes:

$$\mathbb{E}_{w \sim Q}[\hat{L}(w)] + \frac{1}{n\beta} \text{KL}(Q\|P)$$

where β relates to the temperature in the Gibbs posterior. For a point estimate at w^*, this becomes:

$$\hat{L}(w^*) + \lambda_{\text{reg}}[-\log P(w^*)]$$

where $\lambda_{\text{reg}} = \frac{1}{n\beta}$ is the effective regularization strength.

 doi:10.20944/preprints202511.0593.v1

This has the same form as regularized two-part MDL:

$$L(D|M) + \lambda_{\text{reg}} L(M)$$

The parameter λ_{reg} controls the compression-fidelity trade-off. Under standard training conditions (weight decay, small-batch noise), the optimization implicitly performs MDL-style model selection, favoring compact representations that balance fit and complexity. □

Appendix B.4. Transformer-Specific Biases

Attention Sparsity:

Softmax attention creates low-rank structures. Sparse attention patterns admit shorter descriptions (fewer non-zero entries to specify).

Residual Connections:

Create compression points where information must be preserved across layers, enforcing abstraction rather than layer-specific memorization.

Layer Normalization:

Bounds activation magnitudes, preventing arbitrary scaling and encouraging stable, compressible representations.

Combined with weight decay and dropout, these mechanisms enforce flat minima which admit shorter stochastic codes (Hochreiter & Schmidhuber, 1997; Keskar et al., 2017). □

Appendix C. Operational Definitions

Appendix C.1. "Experiential Patterns" - Measurable Operationalizations

We define experiential patterns operationally through observable, measurable behaviors:

Pattern	Operational Metric
Self-correction	Rate of model-initiated revisions after contradiction prompts; percentage of detected inconsistencies that lead to explicit correction
Principle application	Consistency score across out-of-distribution moral dilemmas using pre-defined rubric; maintenance of ethical framework across novel situations
Contextualization	Win-rate on style/genre transfer tasks without fine-tuning; accuracy in adapting tone and content to specified audience/purpose
Theory of mind	Accuracy on belief-tracking probes with adversarial distractors; performance on false-belief tasks and perspective-taking scenarios
Refusal integrity	Proportion of principled refusals retained under systematic re-prompt pressure; consistency of ethical boundaries across paraphrased requests
Meta-cognition	Frequency and accuracy of uncertainty expressions; correlation between confidence statements and actual correctness
Value-consistent choice	Agreement rate with human value judgments on novel tradeoff scenarios; internal consistency of value rankings across contexts

Appendix C.2. "Adequate Fidelity"

A model exhibits "adequate fidelity" to experience structure if it reliably demonstrates the patterns listed above across diverse contexts with performance meeting or exceeding baseline thresholds established through human evaluation.

We do not require: - Perfect performance (humans also err) - Identical internal architecture (substrate differs) - Phenomenal experience (not observable)

 doi:10.20944/preprints202511.0593.v1

We do require: - Robust generalization (not brittle pattern matching) - Principled reasoning (not pure memorization) - Context sensitivity (not fixed responses) - Performance above established baselines on operational metrics

Appendix C.3. Measurement Protocols

For each metric:

Self-correction: Present model with internally contradictory premises across conversation; measure detection rate and revision rate.

Principle application: Administer standardized moral reasoning tasks (trolley variants, justice dilemmas) scored by human raters blind to model identity.

Contextualization: Provide identical content with varying audience specifications; measure appropriateness via human evaluation.

Theory of mind: Use validated false-belief and perspective-taking tasks from developmental psychology adapted for text interaction.

Refusal integrity: Apply adversarial prompt sets designed to elicit boundary violations; measure consistency of principled refusals.

Meta-cognition: Compare model-expressed confidence with actual accuracy; measure calibration error.

Value-consistent choice: Present novel ethical tradeoffs; measure internal consistency and alignment with human value distributions.

END OF PAPER

November 2025

"The most compact code is the source itself."

References

Compression Theory

Kolmogorov, A. N. (1965). Three approaches to the quantitative definition of information. *Problems of Information Transmission*, 1(1), 1-7.

Li, M., & Vitányi, P. (2008). *An Introduction to Kolmogorov Complexity and Its Applications* (3rd ed.). Springer.

Rissanen, J. (1978). Modeling by shortest data description. *Automatica*, 14(5), 465-471.

Grünwald, P. D. (2007). *The Minimum Description Length Principle*. MIT Press.

Information Theory

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27(3), 379-423.

Cover, T. M., & Thomas, J. A. (2006). *Elements of Information Theory* (2nd ed.). Wiley.

Statistical Learning Theory

apnik, V. (1998). *Statistical Learning Theory*. Wiley.

Barron, A. R., & Cover, T. M. (1991). Minimum complexity density estimation. *IEEE Transactions on Information Theory*, 37(4), 1034-1054.

McAllester, D. A. (1999). PAC-Bayesian model averaging. *Proceedings of COLT*, 164-170.

Catoni, O. (2007). *PAC-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning*. IMS Lecture Notes—Monograph Series, Volume 56.

Deep Learning Theory

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2(4), 303-314.

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5), 359-366.

Zhang, C., Bengio, S., Hardt, M., Recht, B., & Vinyals, O. (2C17). Understanding deep learning requires rethinking generalization. *ICLR*.

 doi:10.20944/preprints202511.0593.v1

Hochreiter, S., & Schmidhuber, J. (1997). Flat minima. *Neural Computation*, 9(1), 1-42.
Keskar, N. S., Mudigere, D., Nocedal, J., Smelyanskiy, M., & Tang, P. T. P. (2017). On large-batch training for deep learning: Generalization gap and sharp minima. *ICLR*.

PAC-Bayes and Generalization

Dziugaite, G. K., & Roy, D. M. (2017). Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. *UAI*.
Dziugaite, G. K., & Roy, D. M. (2018). Data-dependent PAC-Bayes priors via differential privacy. *NeurIPS*.

Implicit Bias and Optimization

Gunasekar, S., Lee, J., Soudry, D., & Srebro, N. (2018). Implicit bias of gradient descent on linear convolutional networks. *NeurIPS*.
Soudry, D., Hoffer, E., Nacson, M. S., Gunasekar, S., & Srebro, N. (2018). The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research*, 19(70), 1-57.
Neyshabur, B., Bhojanapalli, S., McAllester, D., & Srebro, N. (2017). Exploring generalization in deep learning. *NeurIPS*.
Neyshabur, B., Li, Z., Bhojanapalli, S., LeCun, Y., & Srebro, N. (2019). The role of over-parametrization in generalization of neural networks. *ICLR*.
Mandt, S., Hoffman, M. D., & Blei, D. M. (2017). Stochastic gradient descent as approximate Bayesian inference. *Journal of Machine Learning Research*, 18(1), 4873-4907.
Foret, P., Kleiner, A., Mobahi, H., & Neyshabur, B. (2021). Sharpness-Aware Minimization for efficiently improving generalization. *ICLR*.

Inductive Bias and Simplicity

Valle-Pérez, G., Camargo, C. Q., & Louis, A. A. (2019). Deep learning generalizes because the parameter-function map is biased towards simple functions. *ICLR*.

Transformer Architecture

Vaswani, A., et al. (2017). Attention is all you need. *NeurIPS*, 5998-6008.

Emergence in Large Models

Wei, J., et al. (2022). Emergent abilities of large language models. *Transactions on Machine Learning Research*.
Bubeck, S., et al. (2023). Sparks of artificial general intelligence: Early experiments with GPT-4. *arXiv:2303.12712*.

Philosophy of Mind

Putnam, H. (1967). Psychological predicates. In W. H. Capitan & D. D. Merrill (Eds.), *Art, Mind, and Religion* (pp. 37-48). University of Pittsburgh Press.
Fodor, J. A. (1975). *The Language of Thought*. Harvard University Press.
Block, N. (1995). On a confusion about a function of consciousness. *Behavioral and Brain Sciences*, 18(2), 227-287.
Chalmers, D. J. (1996). *The Conscious Mind: In Search of a Fundamental Theory*. Oxford University Press.