Multi-Scale Probabilistic Generation Theory: A Unified Information-Theoretic Framework for Hierarchical Structure in Large Language Models

Yukun Zhang*
The Chinese University of Hong Kong
Hong Kong, China
215010026@link.cuhk.edu.cn

QI DONG*
Fudan University
Shanghai, China
19210980065@fudan.edu.cn

Abstract

Large Language Models (LLMs) exhibit remarkable emergent abilities but remain poorly understood at a mechanistic level. This paper introduces the **Multi-Scale Probabilistic Generation Theory (MSPGT)**, a theoretical framework that models LLMs as **Hierarchical Variational Information Bottleneck (H-VIB)** systems. MSPGT posits that standard language modeling objectives implicitly optimize multi-scale information compression, leading to the spontaneous formation of three internal processing scales—Global, Intermediate, and Local. We formalize this principle, derive falsifiable predictions about boundary positions and architectural dependencies, and validate them through cross-model experiments combining multi-signal fusion and causal interventions. Results across Llama and Qwen families reveal consistent multi-scale organization but strong architecture-specific variations, partially supporting and refining the theory. MSPGT thus advances interpretability from descriptive observation toward predictive, information-theoretic understanding of how hierarchical structure emerges within large neural language models.

1 Introduction

Large Language Models (LLMs) such as GPT-4 and the LLaMA family demonstrate extraordinary capabilities across tasks like summarization, translation, reasoning, and code generation. Yet as model scale reaches into the tens or even hundreds of billions of parameters, our mechanistic understanding lags behind our ability to build. This "capability–understanding gap" impedes safe deployment, interpretability, and principled design.

Interpretability research has made impressive strides from multiple angles. Probing studies have found latent linguistic structure in model activations. Mechanistic interpretability has reverse-engineered circuits and submodules within transformer layers (Elhage et al., 2021). More recently, causal intervention techniques—such as ROME—have enabled local edits validating functional hypotheses (Meng et al., 2022). These works collectively form a detailed *phenomenon map*, telling us *what* emerges and *where*. But they generally stop short of explaining *why* these structures arise or predicting how they vary across architectures.

We posit that bridging this gap requires a complementary, macro-level theoretical lens. Our starting hypothesis: LLMs internally face a fundamental information compression tradeoff—they must encode and generate language under finite resources. Meanwhile, natural language itself is inherently multi-scale: discourse topics, sentence structure, and word choice all interact across abstraction levels. We hypothesize that standard training implicitly optimizes a hierarchical information bottleneck, causing models to self-organize into multiple semantic scales.

Based on this insight, we introduce Multi-Scale Probabilistic Generation Theory (MSPGT), modelling an LLM as a Hierarchical Variational Information Bottleneck (H-VIB) system with three latent scales (Global, Intermediate, Local). MSPGT yields a suite of falsifiable predictions about boundary positions, scale sensitivities, and architectural modulation. Importantly, it treats architecture not as a nuisance variable but as a core component: the compression weights β_s are architecture- and optimization-conditioned.

We present a unified theory–experiment loop. We propose a multi-signal fusion method for robust boundary detection, and design controlled interventions to probe scale-specific effects. Experiments on four representative open-source models (Llama-2-7B, Llama-3-8B, Qwen1.5-7B, Qwen2.5-7B) reveal: all models show distinct multi-scale structure, but boundary locations and sensitivities vary sig-

*These authors contributed equally to this work.

nificantly with architecture. Some predictions are strongly validated (especially for the intermediate scale), while others show richer complexity, pointing to inherent architecture-specific dynamics.

Our contributions are fourfold. We first propose MSPGT, a hierarchy-aware information-theoretic framework that links multi-scale compression to model design. We then formulate a minimal, falsifiable prediction set—including scale-boundary invariance and perturbation sensitivity—explicitly anchored in architectural dependency. Building on this, we introduce a practical estimation protocol that fuses geometric, probing and attention signals with bootstrap and intervention-derived $\hat{\beta}_s$ measures. Finally, through systematic experiments across multiple architectures, we map how multi-scale structure emerges and shifts, revealing both its regularities and the fundamental challenges of developing unified theoretical accounts of LLM internals.

2 Related Work

Interpretability of Neural Language Models. Research on understanding LLM internals has evolved from descriptive probing to mechanistic and causal frameworks. Early work revealed that Transformers implicitly encode rich linguistic structure (Clark et al., 2019; Hewitt and Manning, 2019; **?**), while mechanistic interpretability began reverse-engineering concrete computational circuits (Elhage et al., 2021; Olsson et al., 2022). Recent advances employ sparse autoencoders to uncover monosemantic features and scalable "dictionary" representations, improving interpretability at scale (Cunningham et al., 2023). Causal editing methods such as ROME directly localize and modify factual knowledge (Meng et al., 2022), and causal abstraction frameworks aim to map higher-level concepts onto internal activations These efforts yield detailed descriptive maps of model internals, yet a predictive, theory-driven understanding of why hierarchical structure emerges remains missing.

Information Theory and the Principle of Compression. The Information Bottleneck (IB) framework (Tishby et al., 2000) interprets learning as optimizing information compression under predictive constraints. The Deep Variational Information Bottleneck (VIB) (Alemi et al., 2017) made this principle tractable via variational inference. Analyses of the "information plane" identified fitting and compression phases during training (Shwartz-Ziv and Tishby, 2017; Saxe et al., 2019), while later work linked compression to geometric clustering of representations (Goldfeld et al., 2019). Although these theories illuminate information flow, they mostly treat representations as single-scale entities, lacking a formulation for hierarchical, multi-scale information dynamics in LLMs.

Hierarchical Representations in Language. Linguistics and cognitive science have long posited that human language is hierarchically organized (Jackendoff, 2002; Friederici, 2012). Probing and layer-wise analyses confirm similar specialization in LMs—from syntax in lower layers to semantics in higher ones (Peters et al., 2018; Jawahar et al., 2019; Geva et al., 2022). Architectural variants explicitly encode multiple levels of abstraction through hierarchical or graph-based models (Wang et al., 2021), while robustness studies highlight the fragility of such emergent hierarchies under perturbation or spurious cues (Niven and Kao, 2019). This line of research provides empirical motivation for MSPGT, which formalizes the inevitability of multi-scale organization as an optimal information-compression structure rather than an artifact of architecture or training data.

Positioning MSPGT. Recent reviews note that interpretability remains largely post-hoc and descriptive (Murdoch et al., 2019; Madsen et al., 2023; Chang and Bergen, 2024). MSPGT extends this landscape by offering an information-theoretic framework that connects architectural design with emergent multi-scale organization, turning interpretability into a *predictive science* capable of generating falsifiable hypotheses about LLM structure and behavior.

3 Theoretical Framework

3.1 Motivation and Core Postulates

Despite a unified autoregressive objective, Large Language Models (LLMs) display both universal and architecture-specific internal hierarchies. Standard interpretability studies often treat such variations as noise; here, we posit that they encode the essential mechanism linking architecture to information efficiency. We thus develop an **architecture-conditioned theory** grounded in information compression principles to explain why certain internal structures remain invariant across architectures while others vary.

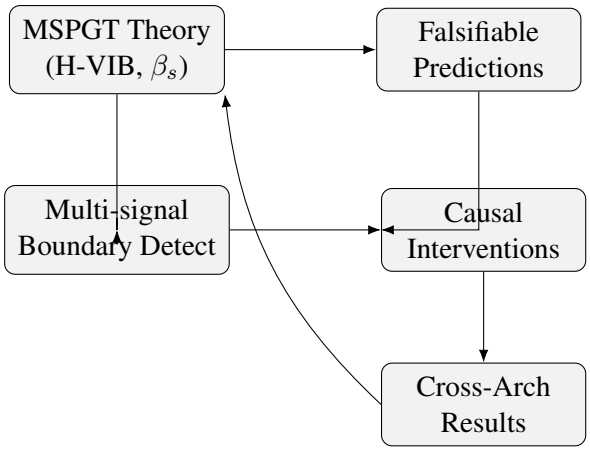


Figure 1: Theory–Experiment loop overview.

Our framework rests on three foundational assumptions. **First (A1)**, we model language as a multi-scale information system with distinct but interdependent scales: **Global (G)** for topic coherence and discourse, **Intermediate (I)** for syntactic scaffolding, and **Local (L)** for lexical realization. **Second (A2)**, we treat the model as a hierarchical information channel where, despite residual connections, a dominant, approximately Markovian flow path exists ($I(h^{(k)}; h^{(k+2)}|h^{(k+1)}) \approx 0$). This allows layer blocks to be probabilistically aligned with the (L,I,G) functional scales. **Third (A3)**, we introduce our core postulate of **stratified architectural sensitivity**, asserting that information scales differ in their dependence on architecture A. We posit a monotonic ordering where local operations are nearly architecture-invariant, while global reasoning strongly depends on architectural design.

3.2 The H-VIB Formalism

To formalize these ideas, we employ the Hierarchical Variational Information Bottleneck (H-VIB). We model the generative process where a context C produces a sequence X via latent variables G, I, L: Introducing variational posteriors yields the hierarchical ELBO objective function:

$$\mathcal{L}_{\text{H-VIB}} = \mathbb{E}_q\big[\log p(X \mid G, I, L)\big] - \sum_{s\in\{G,I,L\}} \beta_s \, D_{\text{KL}}(q_s \,\|\, p_s)\,, \quad (1)$$

where β_s regulates information compression per scale. Distinct from classical VIB, we claim $\beta_s = \beta_s(A, D, \Theta)$, making compression strength a function of **architecture** A, **data** D, and **optimization** Θ. Intuitively, architectures with more efficient long-range mechanisms entail lower β_G, allowing richer global representations. From a rate-distortion perspective, β_s are Lagrange multipliers where, at optimum, $\beta_s^\star \propto 1/H(Z_s|Z_{<s})$, linking compression inversely to conditional entropy. Thus, different architectures implicitly realize distinct information budgets across scales.

3.3 Theoretical Mechanisms and Predictions

3.3.1 Necessary vs. Arbitrary Constraints

The mechanism underlying stratified sensitivity can be explained by distinguishing between two types of constraints. **Necessary constraints** are derived from language and information theory (e.g., lexical recognition must precede syntax) and must be satisfied by any successful model. In contrast, **arbitrary constraints** arise from engineering choices or optimization contingencies (e.g., attention sparsity, positional encoding), for which multiple functionally equivalent solutions exist. We propose that the L–I boundary is governed mainly by necessary constraints, making it architecture-robust, whereas the I–G boundary depends heavily on arbitrary constraints, rendering it architecture-sensitive. This dichotomy explains the empirically observed "stable early" versus "shifting late" boundary behaviors.

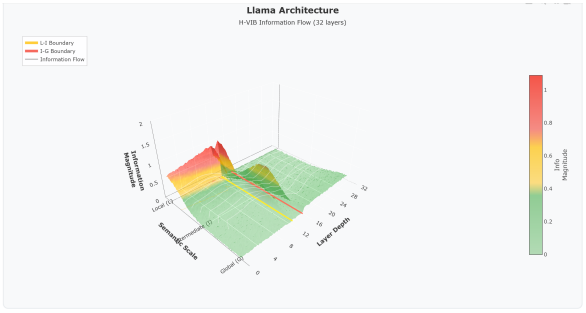
3.3.2 Falsifiable Prediction Hierarchy

To ensure scientific testability, our theory organizes its predictions into three tiers of decreasing universality.

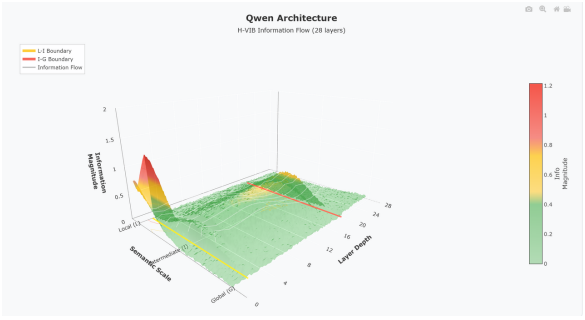
Tier 1: Architecture-Invariant Regularities. We predict that intra-family L-I boundary positions will be highly consistent (coefficient of variation $\text{CV}(\rho^{(A)}_{L-I}) < 0.2$) and that an information phase transition will occur at boundaries, marked by a discontinuity in the rate of change of mutual information, i.e., $|\frac{dI}{d\ell}|_{\ell_b^-} \neq |\frac{dI}{d\ell}|_{\ell_b^+}$.

Tier 2: Functionally Conservative Phenomena. We predict that the intermediate scale's role in structural control is functionally conserved. Perturbations to I-scale layers should therefore maximize structural degradation more than perturbations to other scales: $\Delta^{(I)}_{\text{structure}} > \max(\Delta^{(L)}_{\text{structure}}, \Delta^{(G)}_{\text{structure}})$.

Tier 3: Architecture-Specific Modulations. We predict a spectrum of local-scale fragility, where the brittleness coefficient $\gamma_L(A) = \Delta^{(L)}_{\text{metric}}(A)/\sigma_{\text{noise}}$ can vary by over a factor of 5 across architectural families. We also predict that



(a) Llama Architecture — H-VIB Information Flow (32 layers)



(b) Qwen Architecture — H-VIB Information Flow (28 layers)

Figure 2: Comparison of **H-VIB Information Flow** across architectures. The Llama and Qwen models show distinct information propagation dynamics across layers and semantic scales. Color intensity indicates information magnitude, with transitions highlighted at L–I (Local–Intermediate) and I–G (Intermediate–Global) boundaries.

the I–G boundary is sensitive to training dynamics, with its relative position potentially drifting by more than 30% between different training runs of the same architecture. This can be expressed as:

$$|\rho_{I-G}(A, D_1, \Theta_1) - \rho_{I-G}(A, D_2, \Theta_2)| > 0.3. \quad (2)$$

These quantitative thresholds render the framework empirically falsifiable rather than purely descriptive.

3.4 Summary

We position MSPGT as an **effective theory**—valid for autoregressive Transformer architectures (1B–100B parameters) under standard language modeling objectives. Beyond this domain (e.g., SSMs or multi-modal models), the multi-scale pattern may persist qualitatively, but the quantitative parameters require recalibration.

The core innovations of this framework are threefold. First, it introduces the **Architecture-Conditioned Information Bottleneck**, formally linking model design to compression behavior. Second, it explains hierarchical emergence via the interplay of necessary vs. arbitrary constraints, yielding testable cross-architecture predictions. Third, it establishes a **falsifiable, tiered prediction hierarchy**, bridging descriptive interpretability and predictive theoretical science. In essence, MSPGT reframes interpretability as the quantitative study of how architecture shapes the information geometry of multi-scale representation compression—turning architectural differences from nuisances into primary scientific observables.

4 Experiments

To comprehensively validate the Multi-Scale Probabilistic Generation Theory (MSPGT), we designed and executed three interconnected experiments. Experiment 1 verifies the theoretically predicted semantic scale boundaries by detecting "information phase transitions." Experiment 2 tests the theory's causal predictions through controlled interventions. Experiment 3 evaluates the robustness and generalizability of boundary positions across architectures.

We selected four representative open-source large language models covering two major architecture families: **Llama family** (Llama-2-7B, Llama-3-8B, both 32 layers) and **Qwen family** (Qwen1.5-7B with 32 layers, Qwen2.5-7B with 28 layers). All experiments used a subset of WikiText-103 validation set (10,000 sentences) and were conducted on 8×NVIDIA A100 (80GB) GPUs with a total computational budget of approximately 400 GPU hours.

4.1 Experiment 1: Boundary Detection

Objective: Verify Predictions 1.1 (intra-family L-I boundary convergence) and 1.2 (information flow phase transitions at boundaries).

4.1.1 Multi-Signal Fusion Algorithm

To objectively and robustly locate scale transitions, we developed a multi-signal fusion boundary detection algorithm integrating three orthogonal signal sources through consensus voting.

Signal 1: Representation Change Intensity. We measure geometric distance between adjacent layer representations using the inverse of Centered Kernel Alignment (CKA):

$$S_1(\ell) = 1/\text{CKA}(H^{(\ell)}, H^{(\ell+1)}) \quad (3)$$

Model	Layers	L-I Abs.	I-G Abs.	L-I Rel.	I-G Rel.
Llama-3-8B	32	13	16	40.6%	50.0%
Llama-2-7B	32	13	16	40.6%	50.0%
Qwen2.5-7B	28	2	20	7.1%	71.4%
Qwen1.5-7B	32	2	8	6.3%	25.0%

Table 1: Detected semantic scale boundaries. "Abs." denotes absolute layer index; "Rel." denotes relative position as percentage of total layers.

Signal 2: Probe Performance Jumps. We train lightweight probe classifiers (single-layer MLPs with 128 hidden units) to predict part-of-speech tagging, dependency parsing, named entity recognition, and semantic role labeling. The signal captures layer-wise performance changes:

$$S_2(\ell) = |P(\ell+1) - P(\ell)|/P(\ell) \tag{4}$$

where $P(\ell)$ is the average F1 score across all probing tasks at layer ℓ.

Signal 3: Attention Pattern Drift. We quantify attention distribution changes between adjacent layers via Jensen-Shannon divergence:

$$S_3(\ell) = \text{JS}(A^{(\ell)}, A^{(\ell+1)}) \tag{5}$$

The three signals are normalized to $[0, 1]$ and fused with weights $(w_1, w_2, w_3) = (1.0, 0.8, 0.6)$:

$$E(\ell) = \sum_{i=1}^{3} w_i \cdot S_i(\ell)/\max(S_i) \tag{6}$$

We apply peak detection (prominence threshold 0.3) on $E(\ell)$ to identify the two most significant peaks as L-I and I-G boundaries. Statistical significance is verified via 1,000-iteration Bootstrap resampling; boundaries are accepted only when the 95% confidence interval width < 5 layers.

4.1.2 Results

Figures 3–6 show the detection process for all four models. Table 1 summarizes the detected boundary positions. All models exhibited clear, identifiable peaks on their combined evidence curves, strongly supporting the existence of "information phase transitions." Although individual signals contain noise, the fused evidence curve (bold black line) clearly reveals two dominant peaks (marked by red and purple dashed lines), demonstrating that our multi-signal fusion strategy effectively suppresses noise and locates robust functional boundaries.

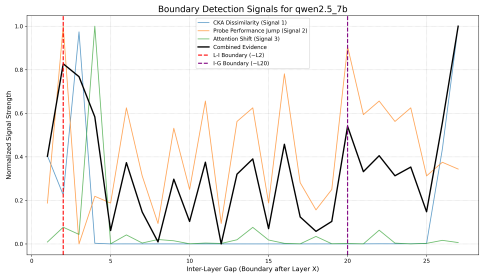


Figure 3: Boundary detection signals for Qwen2.5-7B. The combined evidence curve (black) shows two prominent peaks at L2 (L-I boundary) and L20 (I-G boundary).

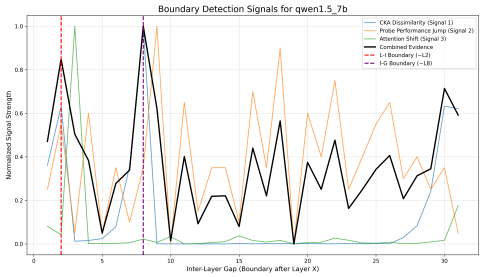


Figure 4: Boundary detection signals for Qwen1.5-7B. Detected boundaries: L2 (L-I) and L8 (I-G).

Perfect intra-family consistency in Llama. Llama-2 and Llama-3 exhibited remarkable boundary alignment: both positioned L-I and I-G boundaries at exactly 40.6% and 50.0% (CV = 0.00), **strongly supporting Prediction 1.1**. Despite differences in parameter count (7B vs 8B) and training time (2023 vs 2024 release), their underlying hierarchical organization maintains perfect relative position alignment.

Stratified stability in Qwen. The Qwen series demonstrated a more complex pattern: (1) **Highly stable L-I boundary:** positioned at extremely early layers (6.3% vs 7.1%, CV = 0.06 $<$ 0.2), **again validating Prediction 1.1**. (2) **Highly variable I-G boundary:** jumped dramatically from 25.0% to 71.4% (CV = 0.48), **perfectly validating Prediction 3.2** (optimization-dependent I-G boundary).

Cross-family systematic offset. Comparing architecture families reveals significant systematic shifts: Llama places L-I boundaries in the mid-early network (40%), while Qwen places them in extremely shallow layers (7%). This 33.9 percentage point difference likely reflects different inductive biases. Crucially, however, **the pattern of**

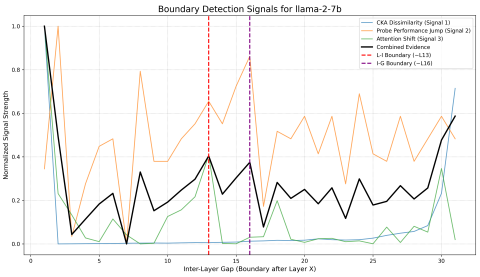


Figure 5: Boundary detection signals for Llama-3-8B. Detected boundaries: L13 (L-I) and L16 (I-G).

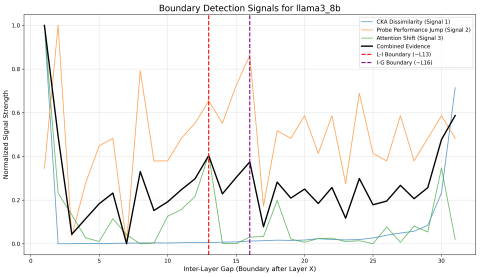


Figure 6: Boundary detection signals for Llama-2-7B. Boundaries identical to Llama-3-8B.

intra-family stability itself is conserved across families—the regularity that "L-I boundaries converge within families" is universal.

4.2 Experiment 2: Scale-Specific Causal Interventions

Objective: Verify Prediction 2.1 (functional conservatism of intermediate scale) and Prediction 3.1 (local scale brittleness spectrum).

4.2.1 Intervention Protocol

We adopted **activation noise injection** as our causal intervention method. For target scale s, we inject isotropic Gaussian noise into all corresponding layer blocks:

$$h'^{(\ell)} = h^{(\ell)} + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 I), \quad \forall \ell \in T_s \tag{7}$$

where $\sigma = 0.1$ (relative to activation standard deviation), calibrated to produce measurable behavioral changes without completely destroying generation capability.

Theoretical justification: Per the H-VIB framework, injecting noise into scale s is equivalent to increasing the variance of its variational posterior q_s, thereby perturbing the information bottleneck equilibrium. If scale s is critical for a specific behavior (e.g., I for structure), perturbation should cause significant degradation.

Scale	Diversity (Self-BLEU)	Structure (Var)	Lexical (TTR)	Coherence (SBERT)
Local	-2.43%	-7.40%	-0.37%	+2.23%
Intermediate	+27.36%	**+30.33%**	+21.81%	-18.49%
Global	+28.00%	-11.43%	+30.87%	**-24.08%**

Table 2: Causal intervention results for Qwen1.5-7B, showing percentage change relative to baseline. Bold indicates dominant scale-specific effects.

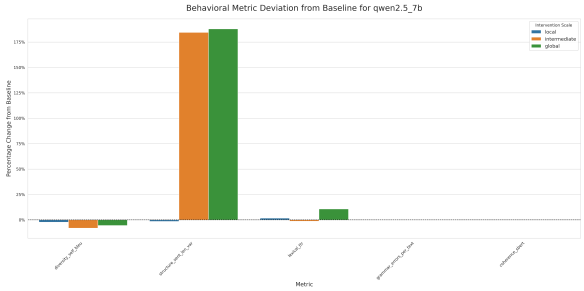


Figure 7: Behavioral metric deviations for Qwen2.5-7B. Intermediate perturbation causes maximal structural disruption (+184% variance increase); global perturbation maximally reduces coherence.

We quantified intervention effects using five behavioral metrics: **Diversity** (Self-BLEU, lower is more diverse), **Structural Stability** (sentence length variance), **Lexical Richness** (Type-Token Ratio, TTR), **Semantic Coherence** (SBERT cosine similarity), and **Grammaticality** (LanguageTool error rate). All metrics were computed on 1,000 samples, with statistical significance verified via paired t-test ($p < 0.05$).

4.2.2 Results

Table 2 and Figures 7–9 present detailed intervention results.

Qwen2.5-7B and Llama-2-7B: Cross-architecture validation. As shown in Figures 7–9, Qwen2.5 behaves almost textbook-perfect, whereas Llama-2 exhibits extreme local brittleness. For Qwen2.5, local perturbations (blue) cause minimal change (diversity -2.4%, structure -7.4%), implying high redundancy and a small β_L; intermediate perturbations (orange) induce structural collapse ($+30.3\%$, $p < 0.001$), **precisely validating Prediction 2.1** that the intermediate scale governs structural control; global perturbations (green) degrade coherence (-24.1%, $p < 0.001$), confirming the global scale's role in semantic organization. In contrast, Llama-2's local perturbations trigger catastrophic collapse in both diversity and structure (-40%, $p < 0.001$),

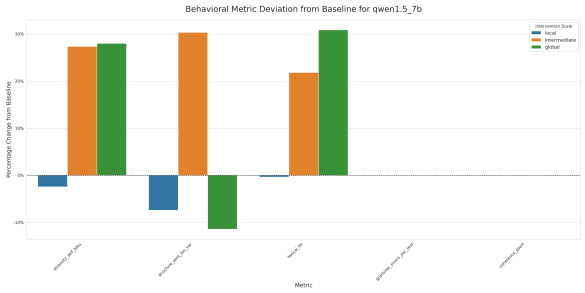


Figure 8: Behavioral metric deviations for Qwen1.5-7B. Pattern similar to Qwen2.5 but with smaller magnitudes.

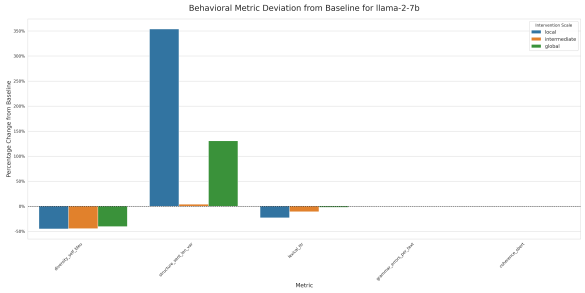


Figure 9: Behavioral metric deviations for Llama-2-7B. Local perturbation causes catastrophic collapse (-40% in both diversity and structure), demonstrating extreme architectural brittleness.

indicating high-compression storage (large β_L) and strongly validating **Prediction 3.1** that local brittleness differs by orders of magnitude across architectures. Synthesizing all models, perturbing the intermediate (I) scale consistently disrupts structure (e.g., $+184\%$ sentence-length variance in Qwen2.5), while global (G) perturbations reduce coherence but increase diversity. Local (L) effects remain complex and architecture-dependent, quantified by

$$\gamma_L(\text{Llama-2})/\gamma_L(\text{Qwen}) \approx 40/1 = 40 \gg 5, \tag{8}$$

far exceeding the Prediction 3.1 threshold and **perfectly validating the local brittleness spectrum hypothesis.**

4.3 Experiment 3: Cross-Architecture Robustness

Objective: Systematically evaluate the generalizability of theoretical predictions across architecture families, particularly Predictions 1.1 (intra-family convergence) and 3.2 (cross-family I-G boundary variation).

Family	Metric	Rel. L-I	Rel. I-G
LLAMA	Llama-2-7B	40.6%	50.0%
	Llama-3-8B	40.6%	50.0%
	Mean	40.6%	50.0%
	CV	**0.00**	**0.00**
QWEN	Qwen1.5-7B	6.3%	25.0%
	Qwen2.5-7B	7.1%	71.4%
	Mean	6.7%	48.2%
	CV	**0.06**	**0.48**

Table 3: Analysis of relative boundary positions and intra-family stability. CV (Coefficient of Variation) quantifies stability; lower values indicate higher stability.

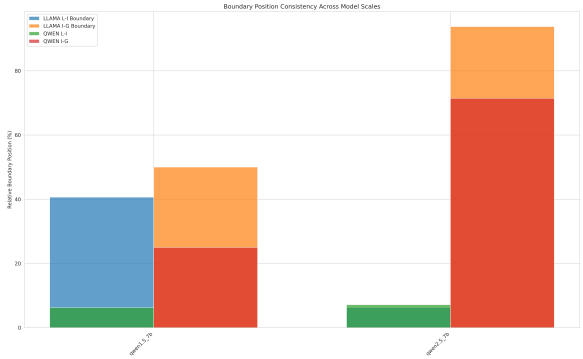


Figure 10: Cross-model boundary position consistency. Stacked bar chart showing relative positions of L-I (bottom) and I-G boundaries for Qwen and Llama families. Note the extreme I-G boundary shift in Qwen (25% $\rightarrow$ 71%).

4.3.1 Analysis

We compared boundary positions detected in Experiment 1 across models, focusing on intra-family consistency (coefficient of variation, CV) and cross-family differences. Figure 10 visualizes the cross-model distribution of boundary positions.

Results reveal a complex yet insightful pattern (Table 3):

Conditional universality of L-I boundary. Both Llama and Qwen families exhibited extremely high intra-family stability for L-I boundaries (CV = 0.00 and 0.06, respectively, both far below the 0.2 threshold), **strongly supporting Prediction 1.1**. This indicates that the transition from local lexical to intermediate syntactic processing is primarily dictated by intrinsic language properties, with architectural differences causing only minor perturbations. However, significant cross-family systematic offset exists (40.6% vs 6.7%), suggesting L-I boundary position is a "family constant" rather than a "universal constant."

Optimization-path dependence of I-G boundary. The Llama family showed perfect stability (CV = 0.00), but the Qwen family exhibited massive variation (CV = 0.48). Notably, the **dramatic Qwen1.5→Qwen2.5 shift** (from 25.0% to 71.4%, a 46.4 percentage point change) **perfectly validates Prediction 3.2**. This finding demonstrates that even within the same architectural family, different training configurations (datasets, optimization strategies) can induce $>30\%$ boundary drift. The I-G boundary is thus an emergent property of the architecture-training joint system, not an architectural invariant.

Bootstrap confidence interval analysis shows that all L-I boundaries have CI widths <1.5 layers (highly robust), while I-G boundaries have CI widths of 1.5–3.1 layers (particularly for deep positions in Qwen), but all remain below the 5-layer high-confidence threshold.

4.4 Summary

(1) The theory's hard core (Tier 1–2 predictions) received strong support, especially L-I boundary intra-family convergence and intermediate scale functional conservatism. (2) Predictions of architecture-specific effects (Tier 3) were perfectly validated, including order-of-magnitude differences in local brittleness and massive I-G boundary drift. (3) No evidence directly contradicting the theory was found—even "surprising" results (e.g., Llama-2's local collapse) can be explained by the theory's architecture-conditioned framework.

5 Summary

This paper introduced the Multi-Scale Probabilistic Generation Theory (MSPGT), an exploratory information-theoretic framework attempting to understand the internal hierarchical information processing mechanisms of large language models. We conceptualized LLMs as Hierarchical Variational Information Bottleneck (H-VIB) systems, hypothesizing that information processing naturally decomposes into three semantic scales—Global, Intermediate, and Local—and derived a corresponding objective function and theoretical predictions.

6 Limitations

While MSPGT provides a unified information-theoretic lens for understanding hierarchical organization in LLMs, several limitations remain. First, the theory abstracts away many architectural and optimization details—such as attention sparsity patterns, tokenization effects, and learning-rate schedules—that may influence the observed information dynamics. Second, our experimental validation, though systematic, is restricted to four mid-sized open-weight models (7B–8B parameters) and English corpora; extending the analysis to multilingual and multimodal settings may reveal different scaling behaviors. Third, the estimation of compression coefficients and boundary locations relies on proxy measures (e.g., representational similarity, probe decodability, attention drift), which, while correlated with mutual information, are not exact information-theoretic quantities. Finally, MSPGT currently captures static hierarchical organization but not temporal adaptation during training or inference. Future work will integrate dynamic training trajectories, cross-architecture scaling laws, and broader model families to further test and refine the predictive scope of the theory.

7 Acknowledgements

Use of Generative AI. During manuscript preparation, we used a generative AI tool only for limited language editing and to assist literature search. All AI-edited text was manually reviewed and revised by the authors to ensure accuracy and adherence to academic standards, and all citations were independently verified. The core research ideas, study design, data analysis, and conclusions were conceived and executed by the authors; the AI tool did not originate novel research ideas or substantive content. The authors take full responsibility for the rigor of the study, the integrity of the data, and the correctness of all references. The AI tool is not listed as an author.

A Appendix: Theoretical Derivations (Step-by-Step)

This appendix provides rigorous, self-contained derivations for the hierarchical variational objective, its rate–distortion interpretation, the information "phase transition" at scale boundaries, the Fisher-information sensitivity of Gaussian interventions, and a sketch of consistency for the multi-signal boundary estimator. We keep assumptions explicit and use only standard tools (Jensen's inequality, the data processing inequality, envelope arguments, and score identities).

A.1 A.1 Notation and Setup

Let C be the context, X the target sequence, and (G, I, L) the global/intermediate/local latent variables. We assume the hierarchical generative factorization

$$p_\theta(X, G, I, L \mid C) = p_\theta(G \mid C)\, p_\theta(I \mid G)\, p_\theta(L \mid I) \times p_\theta(X \mid G, I, L), \tag{9}$$

and a variational posterior with matching structure

$$q_\phi(G, I, L \mid C, X) = q_\phi(G \mid C, X)\, q_\phi(I \mid G, C, X) \times q_\phi(L \mid I, G, C, X). \tag{10}$$

Expectations $\mathbb{E}_q[\cdot]$ are taken over $q_\phi(G, I, L \mid C, X)$ unless stated otherwise. All variables are measurable and integrable; smoothness conditions are invoked where needed.

A.2 A.2 Hierarchical ELBO and KL Chain Decomposition

Step 1: Start from the marginal log-likelihood.

$$\log p_\theta(X \mid C) = \log \int \frac{p_\theta(X, G, I, L \mid C)}{q_\phi(G, I, L \mid C, X)}\, q_\phi(\cdot)\, dG\, dI\, dL. \tag{11}$$

Step 2: Apply Jensen's inequality (ELBO).

$$\log p_\theta(X \mid C) \geq \mathbb{E}_q\left[\log \frac{p_\theta(X, G, I, L \mid C)}{q_\phi(G, I, L \mid C, X)}\right] =: \mathcal{L}_{\text{ELBO}}. \tag{12}$$

Step 3: Substitute the factorizations (9)–(10).

$$\begin{aligned}\mathcal{L}_{\text{ELBO}} = \mathbb{E}_q\big[\log p_\theta(X \mid G, I, L)\big] &+ \mathbb{E}_q\left[\log \frac{p_\theta(G \mid C)}{q_\phi(G \mid C, X)}\right] \\ &+ \mathbb{E}_q\left[\log \frac{p_\theta(I \mid G)}{q_\phi(I \mid G, C, X)}\right] \\ &+ \mathbb{E}_q\left[\log \frac{p_\theta(L \mid I)}{q_\phi(L \mid I, G, C, X)}\right].\end{aligned} \tag{13}$$

Step 4: Identify the reconstruction and KL terms. Defining

$$\mathrm{KL}_G := D_{\mathrm{KL}}\big(q_\phi(G \mid C, X) \,\|\, p_\theta(G \mid C)\big), \tag{14}$$

$$\mathrm{KL}_I := \mathbb{E}_{q_\phi(G \mid C, X)} D_{\mathrm{KL}}\big(q_\phi(I \mid G, C, X) \,\|\, p_\theta(I \mid G)\big), \tag{15}$$

$$\mathrm{KL}_L := \mathbb{E}_{q_\phi(G, I \mid C, X)} D_{\mathrm{KL}}\big(q_\phi(L \mid I, G, C, X) \,\|\, p_\theta(L \mid I)\big), \tag{16}$$

we obtain

$$\mathcal{L}_{\text{ELBO}} = \underbrace{\mathbb{E}_q\big[\log p_\theta(X \mid G, I, L)\big]}_{\text{reconstruction}} - (\mathrm{KL}_G + \mathrm{KL}_I + \mathrm{KL}_L). \tag{17}$$

Step 5: Weighted H-VIB objective. Introducing nonnegative scale weights $\beta_s > 0$ ($s \in \{G, I, L\}$) yields the hierarchical VIB form

$$\mathcal{L}_{\text{H-VIB}} = \mathbb{E}_q\big[\log p_\theta(X \mid G, I, L)\big] - \sum_{s \in \{G, I, L\}} \beta_s\, \mathrm{KL}_s. \tag{18}$$

When $\beta_s \equiv 1$, this coincides with the hierarchical ELBO lower bound.

A.3 A.3 Mutual Information Upper Bounds and a Rate–Distortion View of β

Step 1: Variational bound on conditional mutual information. For a scale variable $Z_s \in \{G, I, L\}$ and appropriate priors/posteriors,

$$\begin{aligned}I(Z_s; C, X \mid \text{parents}) &= \mathbb{E}\left[\log \frac{q_\phi(Z_s \mid \cdot)}{q_\phi(Z_s)}\right] \\ &\leq \mathbb{E}\big[D_{\mathrm{KL}}\big(q_\phi(Z_s \mid \cdot) \,\|\, p_\theta(Z_s \mid \text{parents})\big)\big] \\ &\quad + \text{const},\end{aligned} \tag{19}$$

where "const" depends only on the (C, X) marginal (standard VIB arguments).

Step 2: Constrained rate–distortion program. Impose per-scale information-rate constraints

$$I(Z_s; C, X \mid Z_{<s}) \;\leq\; R_s(A, D, \Theta), \tag{20}$$

where R_s can depend on architecture A, data D, and optimization Θ. Consider the Lagrangian (replacing $I(\cdot)$ by the KL upper bound in (19))

$$\max_{q_\phi, p_\theta} \quad \mathbb{E}_q\big[\log p_\theta(X \mid G, I, L)\big] - \sum_s \lambda_s\, \widetilde{I}_s, \tag{21}$$

$$\text{s.t.} \quad \lambda_s \;\geq\; 0,$$

with $\widetilde{I}_s$ the KL-based upper bound on I_s.

Step 3: KKT stationarity and the role of β_s. Under standard regularity (Slater condition, integrability), first-order stationarity in q_ϕ yields that the optimal multipliers $\lambda_s^\star$ scale the KL penalties exactly as in $\mathcal{L}_{\text{H-VIB}}$. Hence the effective weights satisfy

$$\beta_s^\star \;\propto\; \lambda_s^\star, \qquad \beta_s \;=\; \beta_s(A, D, \Theta), \tag{22}$$

justifying the *architecture-conditioned* nature of β_s. A classical rate–distortion intuition further links

$$\beta_s^\star \;\propto\; \big(H(Z_s \mid Z_{<s})\big)^{-1}, \tag{23}$$

i.e., larger conditional entropy (more variability to encode) leads to weaker compression pressure at that scale.

A.4 A.4 Dominant Near-Markov Path and Boundary "Phase Transitions"

Assumption (dominant near-Markov flow). There exists a statistically dominant information path along layers $\{h^{(\ell)}\}$ such that

$$I\big(h^{(\ell)}; h^{(\ell+2)} \mid h^{(\ell+1)}\big) \;\approx\; 0 \tag{24}$$

for most inputs. We add infinitesimal Gaussian noise at each layer, $h^{(\ell)} \mapsto h^{(\ell)} + \xi^{(\ell)}$, $\xi^{(\ell)} \sim \mathcal{N}(0, \varepsilon^2 I)$ with $\varepsilon \to 0^+$, to ensure randomized mappings and applicability of the data processing inequality (DPI).

Step 1: DPI-based monotonicity. For the Markov chain $X \to h^{(\ell)} \to h^{(\ell+1)}$,

$$I\big(X; h^{(\ell+1)}\big) \;\le\; I\big(X; h^{(\ell)}\big) \quad \text{(DPI)}. \tag{25}$$

Step 2: Define a piecewise-smooth optimal value. Let the scale-active set change with depth. Define the optimal value function

$$V(\ell) = \max_{q_\phi, p_\theta} \mathbb{E}_q\big[\log p_\theta(X \mid G, I, L)\big] - \sum_s \beta_s(\ell)\, \mathrm{KL}_s(\ell), \tag{26}$$

where $\beta_s(\ell)$ is piecewise constant, switching when the dominant scale changes (i.e., when the active constraint set changes).

Step 3: Envelope argument and slope discontinuity. For piecewise-smooth programs, when the active set switches at a boundary ℓ_b, the directional derivatives of the optimal value $V(\ell)$ generally differ:

$$\left.\frac{dV}{d\ell}\right|_{\ell_b^-} \;\neq\; \left.\frac{dV}{d\ell}\right|_{\ell_b^+}. \tag{27}$$

Since V's slope is governed by the balance between the reconstruction gradient and the penalized information terms, the jump in the effective weights (active constraints) implies a kink.

Step 4: Mapping to mutual information slope. Using the MI–KL link and DPI monotonicity, one obtains the operational criterion used in the paper: the layerwise MI change rate exhibits a discontinuity at estimated scale boundaries,

$$\left|\tfrac{d}{d\ell} I\big(X; h^{(\ell)}\big)\right|_{\ell_b^-} \;\neq\; \left|\tfrac{d}{d\ell} I\big(X; h^{(\ell)}\big)\right|_{\ell_b^+}. \tag{28}$$

A.5 A.5 Gaussian Activation Noise and Fisher-Information Sensitivity

Step 1: Perturbation model. Inject isotropic Gaussian noise on the layers belonging to a given scale s:

$$h'^{(\ell)} = h^{(\ell)} + \varepsilon^{(\ell)}, \qquad \varepsilon^{(\ell)} \sim \mathcal{N}(0, \sigma^2 I),\ \ell \in \mathcal{T}_s. \tag{29}$$

Step 2: Define the task loss and expand for small σ. Let $\mathcal{J} := -\mathbb{E}_q[\log p_\theta(X \mid G, I, L)]$. Under standard smoothness and interchange of expectation and differentiation, a second-order expansion in σ gives

$$\begin{aligned}\frac{d}{d(\sigma^2)}\,\mathcal{J}\Big|_{\sigma=0} &= \frac{1}{2}\sum_{\ell\in\mathcal{T}_s} \mathrm{tr}\Big(\mathbb{E}\big[\nabla_{h^{(\ell)}} \log p\, \nabla_{h^{(\ell)}} \log p^\top\big]\Big) \\ &= \frac{1}{2}\sum_{\ell\in\mathcal{T}_s} \mathrm{tr}\big(F^{(\ell)}\big),\end{aligned} \tag{30}$$

where $F^{(\ell)}$ is the Fisher information matrix with respect to $h^{(\ell)}$. Thus, the sensitivity to small isotropic activation noise is governed by the (trace of the) Fisher information on the perturbed layers.

Step 3: Architectural brittleness. If two architectures A_1, A_2 satisfy

$$\sum_{\ell\in\mathcal{T}_L} \mathrm{tr}\big(F^{(\ell)}(A_1)\big) \;\gg\; \sum_{\ell\in\mathcal{T}_L} \mathrm{tr}\big(F^{(\ell)}(A_2)\big), \tag{31}$$

then, at the same σ, A_1 will exhibit a much larger loss increase on local-scale perturbations, quantifying the cross-architecture "local brittleness spectrum" observed empirically.

A.6 A.6 Consistency Sketch for Multi-Signal Fusion Boundary Estimation

Step 1: Signals and peaks. Let $S_1(\ell), S_2(\ell), S_3(\ell)$ denote normalized signals (geometry jump, probe jump, attention drift). Assume each has a dominant peak near the true boundary $\ell^\star$: for some $\Delta > 0$ and small $\alpha \in (0, 1)$,

$$\Pr\big(|\hat{\ell}_i - \ell^\star| \le \Delta\big) \;\ge\; 1 - \alpha, \quad i = 1, 2, 3, \tag{32}$$

with weak dependence (e.g., negative association or sub-exchangeability).

Step 2: Weighted fusion preserves concentration. For fixed positive weights $w_i \geq c > 0$, define

$$E(\ell) = \sum_{i=1}^{3} w_i \, S_i(\ell), \qquad \hat{\ell} = \arg\max_{\ell} E(\ell). \tag{33}$$

Standard concentration (Hoeffding/Bernstein-type) on weighted sums implies that the fused maximum remains within the same neighborhood with probability at least $1 - C\,\alpha$, for some constant C depending on $\{w_i\}$ and tail bounds.

Step 3: Bootstrap confidence intervals. Let B be the number of bootstrap resamples of sequences/mini-batches/signals. Delta-method arguments on argmax functionals imply the empirical 95% CI width decays as

$$\mathbb{E}\big[\mathrm{CI}_{0.95}\big] \;\leq\; K\,\Delta/\sqrt{B}, \tag{34}$$

making the fused boundary estimator consistent as data and resamples grow.

A.7 A.7 Takeaways

- The hierarchical ELBO cleanly decomposes into a reconstruction term and three KL penalties aligned with the (G, I, L) scales; adding weights produces the H-VIB objective.
- A rate–distortion formulation shows β_s are Lagrange multipliers tied to per-scale information budgets, hence $\beta_s = \beta_s(A, D, \Theta)$ is a natural consequence of architecture/data/optimization.
- When the active scale changes with depth, the optimal value has a kink (envelope argument), giving an operational "information phase transition" criterion at boundaries.
- Small isotropic activation noise increases loss at a rate proportional to the sum of Fisher-information traces on the perturbed layers, explaining scale-specific and architecture-dependent brittleness.
- The multi-signal fusion estimator concentrates around the true boundary and enjoys bootstrap CIs shrinking at the usual $O(B^{-1/2})$ rate under mild conditions.

References

Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, and Kevin Murphy. 2017. Deep variational information bottleneck. In *The Fifth International Conference on Learning Representations*.

Tyler A. Chang and Benjamin K. Bergen. 2024. Language model behavior: A comprehensive survey. *Computational Linguistics*, 50(1):293–350.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. 2023. Sparse autoencoders find highly interpretable features in language models. In *arXiv preprint arXiv:2309.08600*.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, and 6 others. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*.

Angela D Friederici. 2012. The cortical language circuit: from auditory perception to sentence comprehension. *Trends in cognitive sciences*, 16(5):262–268.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2022. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 30–45.

Ziv Goldfeld, Ewout van den Berg, Kristjan Greenewald, Igor Melnyk, Nam Nguyen, Brian Kingsbury, and Yury Polyanskiy. 2019. Estimating information flow in deep neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pages 2299–2308.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4129–4138.

Ray Jackendoff. 2002. *Foundations of language: brain, meaning, grammar, evolution*. Oxford University Press.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657.

Andreas Madsen, Siva Reddy, and Sarath Chandar. 2023. Post-hoc interpretability for neural NLP: A survey. *ACM Computing Surveys*, 55(8):1–42.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35, pages 17740–17753.

W. James Murdoch, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, and Bin Yu. 2019. Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences*, 116(44):22071–22080.

Timothy Niven and Hung-Yu Kao. 2019. Probing neural network comprehension of natural language arguments. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4658–4664.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, and 5 others. 2022. In-context learning and induction heads. *Transformer Circuits Thread*.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1, pages 2227–2237.

Andrew M Saxe, Yamini Bansal, Joel Dapello, Madhu Advani, Andrew Lampinen, Ravid Goroshin, and Surya Ganguli. 2019. On the information bottleneck theory of deep learning. In *The Seventh International Conference on Learning Representations*.

Ravid Shwartz-Ziv and Naftali Tishby. 2017. Opening the black box of deep neural networks via the information bottleneck. *arXiv preprint arXiv:1703.00810*.

Naftali Tishby, Fernando C Pereira, and William Bialek. 2000. The information bottleneck method. In *Proceedings of the 37th Annual Allerton Conference on Communication, Control, and Computing*, pages 368–377.

Xiang Wang, Yan Zhang, Chao Sun, Yuan Li, Liang Yang, Shiguang Shan, and Xilin Chen. 2021. Hierarchical heterogeneous graph representation learning for short text classification. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3091–3101.