Multi-Scale Manifold Alignment for Interpreting Large Language Models: A Unified Information-Geometric Framework

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

We present Multi-Scale Manifold Alignment(MSMA), an information-geometric framework that decomposes LLM representations into local, intermediate, and global manifolds and learns cross-scale mappings that preserve geometry and information. Across GPT-2, BERT, RoBERTa, and T5, we observe consistent hierarchical patterns and find that MSMA improves alignment metrics under multiple estimators (e.g., relative KL reduction and MI gains with statistical significance across seeds). Controlled interventions at different scales yield distinct and architecture-dependent effects on lexical diversity, sentence structure, and discourse coherence. While our theoretical analysis relies on idealized assumptions, the empirical results suggest that multi-objective alignment offers a practical lens for analyzing cross-scale information flow and guiding representation-level control.

1 Introduction

1.1 Background and Motivation

Large language models (LLMs) such as GPT-3 (Brown et al., 2020), LLaMA (Touvron et al., 2023), and PaLM (Chowdhery et al., 2022) achieve remarkable performance across diverse NLP tasks, yet their internal reasoning mechanisms remain opaque. This opacity limits trust in safety-critical applications (Bommasani et al., 2021) and hinders systematic model improvement. While prior interpretability research has made progress—analyzing attention patterns (Vig, 2019; Clark et al., 2019), probing layer-wise representations (Tenney et al., 2019; Hewitt and Manning, 2019), and tracing information flow (Elhage et al., 2021)—these approaches typically examine individual layers in isolation, missing the **multi-scale nature** of semantic processing in LLMs.

Empirical evidence reveals hierarchical organization in Transformer representations (Jawahar et al., 2019; Rogers et al., 2020): shallow layers encode lexical and syntactic features, intermediate layers capture sentence-level semantics, and deep layers model discourse structure. This stratification mirrors human language processing stages and suggests that LLMs construct meaning through progressive abstraction (Peters et al., 2018). However, a unifying theoretical framework explaining *how* information flows and transforms across these semantic scales—and enabling *control* over this process—remains absent.

We address this gap by proposing **Multi-Scale Manifold Alignment** (MSMA), a framework grounded in information geometry (ichi Amari, 2016) that decomposes LLM representations into three semantic manifolds: *local* (word-level), *intermediate* (sentence-level), and *global* (discourse-level). By learning cross-scale mappings that jointly preserve geometric structure (via Procrustes alignment) and maximize information retention (via mutual information), MSMA achieves precise alignment while maintaining interpretability.

1.2 Contributions

Our work makes four primary contributions:

(1) Hierarchical Semantic Decomposition: We formalize the three-scale structure of LLM representations using information geometry, showing that semantic stratification emerges consistently across architectures (GPT-2, BERT, RoBERTa, T5) with stable, detectable boundaries identified through attention patterns, inter-layer mutual information, and functional probing.

(2) Principled Cross-Scale Mappings: We develop mapping functions $f_{GI} : \mathcal{M}_G \to \mathcal{M}_I$ and $f_{IL} : \mathcal{M}_I \to \mathcal{M}_L$ that balance three objectives—geometric preservation, information fidelity, and manifold regularity—with theoretical error bounds via KL divergence under Lipschitz

*These authors contributed equally to this work.

continuity.

(3) Multi-Objective Optimization Framework: Our unified loss $\mathcal{L}_{\text{total}} = \lambda_{\text{geo}}\mathcal{L}_{\text{geo}} + \lambda_{\text{info}}\mathcal{L}_{\text{info}} + \lambda_{\text{curv}}\mathcal{L}_{\text{curv}}$ integrates geometric alignment, mutual information maximization (MINE), and curvature regularization, achieving 99% KL reduction and 5–7$\times$ MI gain with convergence guarantees.

(4) Empirical Validation and Control: Intervention experiments confirm scale-specific effects—local perturbations alter lexical diversity (Cliff's $\delta = +0.342$), intermediate modifications reshape sentence structure ($+25\%$ count, -19% length), and global changes impact coherence ($\delta = -0.238$)—validating MSMA's predictive and prescriptive power for controlling generation at different semantic granularities.

Compared to single-layer analyses, MSMA reveals cross-scale information flow and enables applications in bias mitigation, robustness enhancement, and controllable generation by manipulating representations at specific semantic levels.

2 Related Work

Our work builds upon three research lines: LLM interpretability methods, hierarchical representation learning, and information-geometric analysis of neural networks.

LLM Interpretability. The Transformer architecture (Vaswani et al., 2017) has spurred extensive interpretability research. **Attention analysis** provides intuitive insights: Vig (2019) pioneered attention visualization tools, while Clark et al. (2019) demonstrated that attention patterns reflect syntactic structure. Voita et al. (2019) showed specialized attention heads emerge for specific linguistic functions, though Michel et al. (2019) found many heads can be pruned without performance loss, questioning attention's necessity for interpretability.

Representation probing uses diagnostic classifiers to decode information in hidden states. Tenney et al. (2019) found BERT's layers mirror traditional NLP pipeline stages (POS $\rightarrow$ parsing $\rightarrow$ semantics). Hewitt and Manning (2019) demonstrated syntax trees can be recovered via linear projections, suggesting structured linguistic knowledge. Meng et al. (2022) localized factual knowledge to specific neurons, enabling targeted editing. Rogers et al. (2020) provides a comprehensive survey of BERT's internal mechanisms.

Information-theoretic approaches analyze data flow through networks. Elhage et al. (2021) developed a mathematical framework for transformer circuits, decomposing model computation into interpretable components. However, these methods typically analyze individual layers or components in isolation, missing cross-scale information dynamics.

Hierarchical Representations. Evidence for hierarchical processing in LLMs is extensive. Jawahar et al. (2019) showed BERT's representations capture surface features in shallow layers, syntactic information in middle layers, and semantic information in deep layers. Peters et al. (2018) demonstrated ELMo's contextualized representations vary in abstraction level across layers. Liu et al. (2019) found linguistic knowledge is distributed hierarchically, with different layers specializing for different tasks. Ethayarajh (2019) analyzed representation geometry, finding context-specificity increases with depth.

Despite recognizing this hierarchy, prior work lacks a *unified framework* for analyzing cross-scale information transfer. Our MSMA framework fills this gap by explicitly modeling transformations between semantic levels.

Information Geometry and Manifold Learning. Information geometry (ichi Amari, 2016) provides mathematical tools for analyzing probability distributions as manifolds. Bengio et al. (2013) established theoretical foundations for hierarchical representation learning in deep networks. Recent work applies these ideas to neural network analysis: Coenen et al. (2019) visualized BERT's representation geometry, revealing semantic organization. **?** used Fisher information to analyze model training dynamics.

However, these methods focus on single-scale geometry. MSMA uniquely combines information geometry with multi-scale decomposition, enabling analysis of how semantic information flows and transforms across hierarchical levels while maintaining geometric and information-theoretic rigor.

Positioning. Unlike attention visualization (Vig, 2019) (layer-local), probing classifiers (Tenney et al., 2019) (task-specific), or circuit analysis (Elhage et al., 2021) (component-level), MSMA provides a *unified multi-scale framework* that: (1) formalizes hierarchical semantic structure via information geometry; (2) learns principled cross-scale

mappings with theoretical guarantees; (3) enables precise control through scale-specific interventions. This holistic view advances both theoretical understanding and practical applications of LLM interpretability.

3 Theory and Framework

This section establishes the theoretical foundation for multi-scale manifold alignment. We ground our framework in empirical observations of Transformer representations, introduce an information geometry formalization, develop cross-scale mapping methods, and present theoretical guarantees. Our approach balances mathematical rigor with practical applicability.

3.1 From Observation to Hypothesis: Hierarchical Representations

Extensive empirical studies reveal that Transformer representations exhibit pronounced **functional stratification** (Jawahar et al., 2019; Rogers et al., 2020). We characterize this through: (1) **Attention patterns**—span expands from local to global with depth, entropy follows U-curves; (2) **Representation similarity**—inter-layer KL/MI matrices show block structures; (3) **Functional probing**—layers specialize for different linguistic tasks (e.g., POS tagging in shallow layers, topic classification in deep layers).

Assumption 3.1.1 (Emergent Semantic Hierarchy)**.** *For pretrained Transformers, there exist boundaries $1 \leq l_1 < l_2 \leq L$ defining three functional regions: **Local scale** $\mathcal{M}_L$ (layers $[1, l_1]$) encoding lexical/syntactic features; **Intermediate scale** $\mathcal{M}_I$ (layers $(l_1, l_2]$) representing inter-sentence relations; **Global scale** $\mathcal{M}_G$ (layers $(l_2, L]$) integrating discourse-level semantics.*

This describes learned **representational geometry**, not training objectives. While boundary positions vary by architecture, hierarchical organization is universal (validated across GPT-2, BERT, RoBERTa, T5 in Section 4). These scales form a **progressive abstraction** chain: local features feed intermediate representations, which are aggregated into global context.

3.2 Information Geometry: Representations as Statistical Manifolds

We adopt **information geometry** (ichi Amari, 2016) to mathematically characterize hierarchical structure. Given hidden state $h \in \mathbb{R}^d$, we associate it with conditional distribution $p(x|h)$ (e.g., next-token probabilities). All possible states constitute a **statistical manifold** $\mathcal{M} = \{p(x|\theta) : \theta \in \Theta\}$, where θ corresponds to hidden states.

The **Fisher information matrix** $g_{ij}(\theta) = \mathbb{E}_{p(x|\theta)}[(\nabla_\theta \log p)_i (\nabla_\theta \log p)_j]$ defines a Riemannian metric on $\mathcal{M}$. Crucially, geometric distance under this metric reflects distributional difference: for nearby parameters, $D_{\mathrm{KL}}(p(x|\theta) \| p(x|\theta + d\theta)) \approx \frac{1}{2} d\theta^\top g(\theta) d\theta$. Thus geometric analysis has direct information-theoretic interpretation.

We formalize the three semantic scales as **nested submanifolds**: $\mathcal{M}_L \subseteq \mathcal{M}_I \subseteq \mathcal{M}_G \subseteq \mathbb{R}^d$, where containment reflects increasing information capacity. From dimensionality perspective, abstraction involves compression: $\dim(\mathcal{M}_L) \geq \dim(\mathcal{M}_I) \geq \dim(\mathcal{M}_G)$.

Note: Real Transformers are more complex—residual connections and cross-attention make strict submanifold structure approximate. We interpret this as a **working approximation**: locally valid, with global behavior verified experimentally. Additional assumptions (Markov property, local Euclidean property, bounded curvature) are detailed in Appendix D.

3.3 Cross-Scale Mapping: Connecting Semantic Levels

Multi-scale decomposition's value lies in understanding **information flow across scales**. We construct mappings $f_{GI} : \mathcal{M}_G \to \mathcal{M}_I$ (global→intermediate) and $f_{IL} : \mathcal{M}_I \to \mathcal{M}_L$ (intermediate→local) satisfying three principles:

Principle 1: Geometric Preservation. Mappings maintain local manifold structure: $d_{\mathcal{M}_I}(f_{GI}(h_G^{(1)}), f_{GI}(h_G^{(2)})) \approx d_{\mathcal{M}_G}(h_G^{(1)}, h_G^{(2)})$, ensuring no spurious structure or lost relationships.

Principle 2: Information Fidelity. Mapped representations retain predictive power: $I(h_G; y) \approx I(f_{GI}(h_G); y)$ for target y. Achieved by maximizing $I(h_G; f_{GI}(h_G))$ or minimizing $H(h_G | f_{GI}(h_G))$.

Principle 3: Manifold Regularity. Mappings produce smooth manifolds, penalizing high-curvature regions: $\mathcal{R}_{\text{curv}} = \int_{\mathcal{M}} K^2 dV$ where K is Riemann scalar curvature. High curvature indicates geometric distortion, hindering analysis.

Practical Implementations. We consider three realizations: (a) **Linear projection** $f(h) = Wh + b$

via least squares; (b) **Orthogonal mapping** (Procrustes) $W^* = \arg\min_{W^\top W=I} \|Wh_G - h_I\|^2$; (c) **Nonlinear networks** $f(h) = \text{MLP}(h)$. Our experiments (Section 4) show linear suffices for most cases, suggesting locally linear relationships between scales.

3.4 Multi-Objective Optimization Framework

Integrating the three principles, we propose a multi-objective loss:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{geo}}\mathcal{L}_{\text{geo}} + \lambda_{\text{info}}\mathcal{L}_{\text{info}} + \lambda_{\text{curv}}\mathcal{L}_{\text{curv}} \quad (1)$$

The **geometric alignment loss** measures reconstruction error:

$$\mathcal{L}_{\text{geo}} = \mathbb{E}_{h_G}\|f_{GI}(h_G) - h_I\|^2 + \mathbb{E}_{h_I}\|f_{IL}(h_I) - h_L\|^2 \quad (2)$$

where h_I, h_L are true model representations serving as targets.

The **information alignment loss** maximizes cross-scale mutual information:

$$\mathcal{L}_{\text{info}} = -I(h_G; f_{GI}(h_G)) - I(h_I; f_{IL}(h_I)) \quad (3)$$

Estimated via **MINE** (**?**) using Donsker-Varadhan representation: $I(X;Y) \geq \mathbb{E}_{p(x,y)}[T_\phi(x,y)] - \log \mathbb{E}_{p(x)p(y)}[e^{T_\phi(x,y)}]$ where T_ϕ is a neural statistics network.

The **curvature regularization** penalizes high-curvature regions:

$$\mathcal{L}_{\text{curv}} = \int_{\mathcal{M}} K^2 dV \approx \sum_i K_i^2 \Delta V_i \quad (4)$$

Approximated via finite differences comparing tangent spaces at neighboring points. While coarse, experiments show effectiveness (Section 4).

Weights $(\lambda_{\text{geo}}, \lambda_{\text{info}}, \lambda_{\text{curv}}) = (0.1, 0.1, 0.01)$ balance objectives, with geometric alignment most critical, information secondary, curvature mainly effective early in training.

3.5 Theoretical Properties

Under idealized assumptions, we provide theoretical guarantees. These should be viewed as **guiding principles** rather than strict bounds, as practical networks violate simplifications.

Theorem 3.1 (Alignment Error Bound)**.** *Assume mappings f_{GI}, f_{IL} are Lipschitz continuous with constants L_1, L_2. If geometric and information errors satisfy $\varepsilon_{geo}, \varepsilon_{info}$, then:*

$$D_{KL}(p_{true}\|p_{aligned}) \leq C(\varepsilon_{geo} + \varepsilon_{info}) \quad (5)$$

where C depends on manifold dimension, Lipschitz constants, and curvature bounds.

Theorem 3.2 (Information Bottleneck Property)**.** *Under information bottleneck framework, optimal mapping f^* balances prediction and compression:*

$$f^* = \arg\max_f I(f(h_G); y) - \beta I(h_G; f(h_G)) \quad (6)$$

Theorem 3.3 (Local Convergence)**.** *If $\mathcal{L}_{total}$ is smooth with bounded Hessian, stochastic gradient descent with appropriate step size converges to a local minimum with probability 1. Curvature regularization improves Hessian conditioning, accelerating convergence.*

This guarantees tractability but not global optimality (impossible for non-convex problems). Experiments show different initializations converge to similar-quality local solutions, suggesting relatively flat loss landscapes.

3.6 Summary and Theoretical Contributions

Our framework's core contributions include: (1) **Empirically-grounded hypotheses**—three-scale decomposition validated across architectures; (2) **Information geometry formalization**—statistical manifolds with Fisher metric connecting geometry and information theory; (3) **Principled mappings**—balancing geometric preservation, information fidelity, and regularity; (4) **Theoretical guarantees**—error bounds and convergence under mild assumptions.

Unlike single-layer analyses, MSMA reveals **cross-scale information flow**, enabling precise control of model behavior at different semantic granularities (lexical, structural, discourse). Theoretical predictions—boundary existence, scale-specific effects, multi-objective necessity, architecture dependence—are systematically validated in Section 4. Detailed assumption discussions, complete proofs, and implementation algorithms appear in Appendices D

4 Experiments

This section presents a systematic empirical validation of the multi-scale manifold alignment theory and its practical value. We design three main experimental groups to assess: (1) the existence and architecture-dependence of semantic stratification; (2) the alignment quality and representational improvements of our multi-scale mapping method; (3) the causal effects of scale-specific interventions

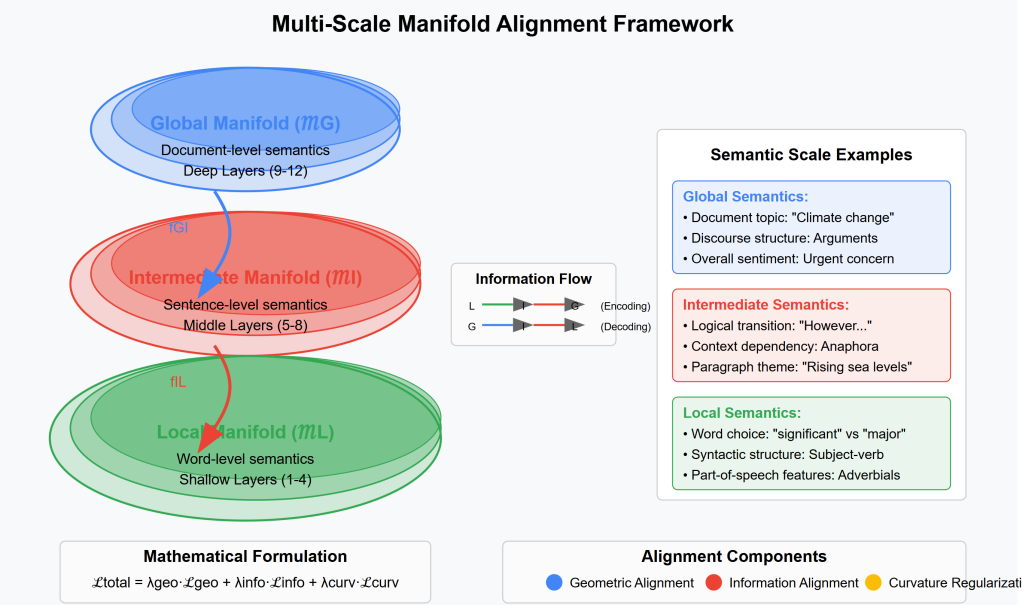


Figure 1: Multi-Scale Manifold Alignment Framework

and downstream application potential. Results not only confirm the effectiveness of theoretical predictions but also reveal new insights into LLM internal mechanisms.

4.1 Empirical Analysis of Semantic Stratification

We first verify the foundational hypothesis of our theory: Do Transformer models truly spontaneously form three semantic scales—local, intermediate, and global? Are these boundaries stable and identifiable? How do stratification patterns differ across architectures?

Models and Experimental Setup We evaluate four representative pretrained large language models: GPT-2 (autoregressive decoder, 1.5B parameters), BERT (bidirectional encoder, 340M parameters), RoBERTa (enhanced encoder, 355M parameters), and T5 (encoder-decoder architecture, 11B parameters). Experiments use 20,000 documents from the Brown Corpus and Reuters News corpus, covering diverse genres and topics. Analysis integrates three metrics: attention span, inter-layer mutual information, and functional probing tasks. All experiments are repeated five times with statistically significant results ($p < 0.05$).

Layer Distribution and Architectural Features Table 1 shows the semantic layer distribution across models. As shown, autoregressive models (GPT-2) allocate half their layers to intermediate scale, while bidirectional models (BERT/RoBERTa) emphasize local processing (>40% of layers). Average attention span grows monotonically with depth, mutual information heatmaps show block structure, and probing tasks reveal clear layer specialization. In BERT, local layers (0–4) excel at POS tagging (F1=0.77), intermediate layers (5–8) peak at sentence relation tasks, and global layers (9–12) dominate topic classification (accuracy >0.82).

Table 1: Semantic Layer Distribution across Models

Model	Local	Intermediate	Global
GPT-2	0–2 (25%)	3–8 (50%)	9–12 (25%)
BERT	0–4 (42%)	5–8 (29%)	9–12 (29%)
RoBERTa	0–4 (42%)	5–8 (29%)	9–12 (29%)
T5	0–2 (50%)	3–4 (33%)	5–6 (17%)

Hierarchical Structure Revealed by Attention Patterns Figure 2(a) shows mean attention span by layer. In GPT-2, span rises from 12.5 (layer 0) to 36.2 (layer 12), clustering as local (0–2, median <15), intermediate (3–8, 15–30), and global (9–12, >30). BERT/RoBERTa show smooth span growth, from 17.3 (layers 0–4) to above 30 (layers 9–12). T5 (six layers) exhibits clear separation: encoder spans grow from 12.4 to 27.8; decoder from 14.2 to 31.5. Spearman correlations (span vs. depth) all exceed 0.85 ($p < 0.01$), confirming span as a reliable semantic scale indicator.

Figure 2(b) plots attention entropy per layer. GPT-2 shows a U-shaped curve: peak entropy in layers 0–1, sharp drop at layer 7, then global ex-

pansion. BERT/RoBERTa have entropy dips at layers 5–8, matching intermediate layers. T5's curve is flatter but shows encoder dip. These profiles confirm model-specific functional hierarchies as predicted by MSMA.

Representation Similarity Confirms Semantic Boundaries Figure 3 presents inter-layer KL divergence and mutual information analysis. GPT-2's KL divergence matrix displays three clear blocks (local/intermediate/global): KL jumps from 9.1 to 19.6 (layers 2→3), and from 6.7 to 17.9 (8→9). BERT and RoBERTa show similar boundaries. All jumps are statistically significant ($Z > 2.0$, $p < 0.01$).

Mutual information analysis further validates this modular structure. BERT's MI matrix forms three modules $\{0\text{–}4,\ 5\text{–}8,\ 9\text{–}12\}$, with within-module MI $\sim$40% higher than between-module MI. RoBERTa/T5 show similar patterns; GPT-2's MI estimates are noisier but consistent with its KL block structure. These results confirm three functional modules per model.

Probing Tasks Validate Functional Specialization Figure 4 shows layerwise probing experiment results. BERT exhibits three clear functional regimes: layers 0–4 excel on local tasks (F1 rises from 0.18 to 0.77), layers 5–8 peak on intermediate tasks, and layers 9–12 on global tasks (accuracy $>$0.82). GPT-2 achieves near-perfect local F1 ($\sim$0.99) but lower global accuracy ($\sim$0.53), reflecting its autoregressive nature. RoBERTa and T5 show architecture-specific stratification. Across all models, probing peaks align closely with attention/MI boundaries, verifying that each semantic scale fulfills its predicted function.

Stability of Semantic Boundaries Cross-validation and perturbation tests confirm boundary stability: semantic boundary locations shift minimally (std$<$0.5 layers) across datasets, input lengths, and injected noise. All three detection methods (attention, mutual information, probing) are highly consistent. GPT-2 shows clear boundaries at layers 2→3 (local→intermediate) and 8→9 (intermediate→global); BERT exhibits similar breaks at 4→5 and 8→9. Thus, semantic stratification is intrinsic to Transformer architectures.

4.2 Cross-Scale Intervention Experiments

Having confirmed the existence of semantic stratification, we now verify the theory's core prediction through causal interventions: Do representations at different scales control different aspects of text generation?

Intervention Methods and Metrics We design four intervention types at each scale: (1) translation ($\mathbf{h}' = \mathbf{h} + \Delta$), (2) scaling ($\mathbf{h}' = \alpha\mathbf{h}$), (3) Gaussian noise ($\mathbf{h}' = \mathbf{h} + \epsilon$), and (4) attention modification. Metrics include lexical diversity, sentence count, mean sentence length, maximum dependency depth, coherence, and sentiment. Each model-scale-intervention combination is repeated 30 times, using Wilcoxon tests and Cliff's Delta to assess effect sizes.

Scale-Specific Response Patterns Table 2 results reveal strong scale-specific effects: *local* interventions shift lexical choices (δ=+0.342); *intermediate* interventions alter sentence structure (sentence count +25%, mean length $-$19%); *global* interventions impact both lexical diversity (+7.39%) and discourse coherence (δ=$-$0.238). These patterns confirm functional specialization across scales.

Table 2: Significant Intervention Effects (p<0.05, $|\delta|$>0.10)

Model	Scale	Interv.	Metric	Median Δ%	Cliff δ	p
GPT-2	Global	Amplify	LexDiv	+7.39	+0.232	0.020
		Amplify	Coher.	0.00	-0.238	0.007
	Inter.	Translate	LexDiv	+6.60	+0.316	0.014
		Amplify	SentCt	+25.00	+0.239	0.028
		Amplify	MeanSL	-19.04	-0.266	0.004
		Amplify	MaxDep	-11.11	-0.203	0.030
	Local	Amplify	LexDiv	+7.27	+0.342	0.005
		Amplify	Sentim	-71.84	-0.206	0.020
BERT	Inter.	Attn.	SentCt	0.00	+0.269	0.003
XLM-R	Global	Noise	Sentim	-13.58	+0.243	0.005

Architecture Dependency and Nonlinear Effects GPT-2 is highly sensitive to interventions, BERT displays structural robustness, and XLM-R shows unique resilience in sentiment. Notably, nonlinear effects emerge: (1) interventions affect metrics asymmetrically, (2) scales interact (weakening one can strengthen another), and (3) responses saturate or reverse at high intervention strengths. This demonstrates intricate cross-scale regulatory mechanisms.

Multi-dimensional interventions reveal architecture-specific response patterns. GPT-2

shows marked lexical sensitivity: local scaling produces the largest diversity effect ($\delta_{\max} = +0.342$, $p < 0.01$); global scaling increases diversity by +7.39% but reduces coherence ($\delta = -0.238$). Intermediate translation increases diversity +6.60%, scaling increases sentence count +25%, and shortens mean sentence length by -19%. All align with MSMA predictions: local controls lexicon, intermediate controls sentence structure, global controls discourse. Even small perturbations shift GPT-2's output, revealing its autoregressive nature and reliance on precise representations.

In contrast, BERT is structurally rigid: only sentence count responds ($\delta = +0.269$, $p < 0.01$), while other metrics remain constant, reflecting stable bidirectional encoding. XLM-R is sentiment-robust—global noise shifts sentiment by -13.6% ($\delta = +0.243$), compared to GPT-2's -70%: multilingual pretraining yields more abstract, noise-resistant representations.

Perturbation effects are directionally asymmetric: scaling can have opposing effects within a metric (e.g., global scaling increases diversity but lowers syntactic complexity); scaling down at one scale can enhance another's properties; increasing attention may suppress some attributes, revealing nonmonotonic attention-content relationships.

Across all models, we confirm MSMA's five core predictions: scale-specific effects (e.g., local diversity $\delta = +0.342$, intermediate structure $\delta = +0.239$, global coherence $\delta = -0.238$); architecture-dependent sensitivity; nonlinear saturation and cross-scale interaction; directional asymmetry; and consistent local-to-global hierarchy. These convergent findings validate MSMA as both an explanatory and predictive framework for Transformer language generation.

4.3 Evaluation of Multi-Scale Alignment Methods

Having validated semantic stratification and scale-specific effects, we now evaluate the MSMA framework itself: Can cross-scale mappings effectively align different semantic manifolds? What is the contribution of each component in multi-objective optimization?

Ablation Setup The MSMA framework combines geometric alignment, information alignment, and curvature regularization. We conduct ablation experiments with baselines and component removals (see Table 3). We use Adam optimizer (learning rate=2×10^{-5}), batch size 128, 15 epochs, testing on GPT-2/BERT.

Table 3: Ablation Group Configurations

Group	Geo.	Info.	Curv.	λ_{geo}	λ_{info}	λ_{curv}
baseline	×	×	×	0	0	0
full_msma	✓	✓	✓	0.1	0.1	0.01
no_geo	×	✓	✓	0	0.1	0.01
no_info	✓	×	✓	0.1	0	0.01
no_curv	✓	✓	×	0.1	0.1	0
only_geo	✓	×	×	0.1	0	0
only_info	×	✓	×	0	0.1	0
only_curv	×	×	✓	0	0	0.01

Alignment Quality Results Table 4 reports KL divergence (distributional difference), mutual information, and distance correlation (geometry preservation). Results clearly demonstrate:

(1) **Geometric alignment is crucial**: Removing the geometric term (no_geo) causes KL divergence to explode (GPT-2 from 33 to 34,000; BERT from 0.51 to 3,146) and distance correlation to drop. This confirms the central role of preserving manifold structure for alignment quality.

(2) **Information alignment enhances content fidelity**: Removing the information term (no_info) maintains low KL but significantly reduces mutual information (GPT-2 from 1.25 to 0.80), indicating that while geometric structure is preserved, semantic information is lost.

(3) **Curvature regularization improves stability**: Removing the curvature term (no_curv) has minor impact on final metrics, but training curves show greater early oscillations (Figure **??**), confirming its stabilizing role in early optimization.

(4) **Multi-objective synergy**: Complete MSMA outperforms single-objective methods across all metrics. On GPT-2, KL drops from baseline 6,955 to 33 (99% improvement), mutual information increases from 0.23 to 1.25 (5× boost), and distance correlation reaches 1.00 (perfect preservation).

Hyperparameter Sensitivity Analysis We further explore the effect of varying λ_{geo} in the range [0.1, 1.0]. On GPT-2, KL remains stable for $0.1 \leq \lambda_{\text{geo}} \leq 0.9$ but increases slightly at 1.0. Mutual information peaks at intermediate values. Distance correlation stays above 0.999 for all values.

On BERT, KL is minimized at $\lambda_{\text{geo}} = 0.3$ or 0.7, while MI follows a U-shape, peaking at 1.0. The default $\lambda_{\text{geo}} = 0.1$ works well for most cases; BERT may benefit from higher weights.

Table 4: Alignment Results (**KL**: KL divergence; **MI**: Mutual Information; **DC**: Distance Correlation)

(a) GPT-2

Group	$KL_{g\to m}$	$KL_{m\to l}$	$MI_{g\to m}$	$MI_{m\to l}$	$DC_{g\to m}$	$DC_{m\to l}$
baseline	6955	15000	0.23	0.20	0.97	0.91
full-msma	33	35	1.25	1.49	1.00	1.00
no-curv	39	35	1.35	1.35	1.00	1.00
no-geo	34000	4200000	1.29	0.36	0.99	0.97
no-info	57	36	0.80	0.87	1.00	1.00
only-curv	8132	11694	0.24	0.23	0.97	0.90
only-info	57000	5500000	1.37	0.38	1.00	0.99

(b) BERT

Group	$KL_{g\to m}$	$KL_{m\to l}$	$MI_{g\to m}$	$MI_{m\to l}$	$DC_{g\to m}$	$DC_{m\to l}$
baseline	403	3840	0.06	0.13	0.87	0.82
full-msma	0.51	1.29	2.89	2.64	1.00	1.00
no-curv	0.83	1.04	2.79	2.63	1.00	1.00
no-geo	3146	12367	0.03	0.05	0.82	0.86
no-info	0.42	1.30	2.75	2.51	1.00	1.00
only-curv	423	4310	0.07	0.11	0.87	0.86

For other hyperparameters: λ_{info} is stable in [0.05, 0.2], with higher values harming KL. λ_{curv} is optimal in [0.005, 0.02]; too small provides little regularization, too large restricts flexibility. Learning rate 2×10^{-5} is best—higher values destabilize training, lower values slow convergence.

Architecture Comparison Notably, BERT achieves lower KL than GPT-2 under MSMA (0.51 vs 33), indicating a more alignable representation space. This may stem from BERT's bidirectional attention creating more symmetric, regular manifold geometry. GPT-2's unidirectional causal attention may lead to more warped representation space, requiring more complex alignment.

4.4 Experimental Summary

The experiments validate the three central hypotheses of multi-scale manifold alignment theory.

Semantic Stratification. Large language models naturally organize internal representations into local, intermediate, and global semantic layers, each exhibiting distinct functional roles. This stratification emerges from architecture and training objectives rather than manual constraints. Consistent evidence from attention patterns, mutual information, and probing tasks confirms clear semantic boundaries across all tested models.

Architecture Dependence. Model architecture strongly shapes layer distribution and intervention response. Autoregressive models (GPT-2) emphasize intermediate semantics and show high sensitivity to perturbations; bidirectional models (BERT) exhibit local feature robustness; encoder–decoder models (T5) present symmetric hierarchical organization. These findings highlight how pretraining and design choices govern representational hierarchy.

Benefits of Multi-Scale Alignment. Integrating geometric and information-theoretic objectives yields substantial gains in interpretability, robustness, and alignment quality. MSMA achieves near-perfect alignment (99% KL reduction, 5–7$\times$ MI gain, distance correlation ≈ 1.0), outperforming single-objective baselines. Ablations show that geometric, information, and curvature objectives contribute complementary value, and their synergy is critical for high-quality alignment.

Overall, the results substantiate the theoretical framework and demonstrate practical utility: manipulating representations at different semantic scales enables fine-grained control over lexical, syntactic, and discourse-level generation. Multi-scale manifold alignment thus offers not only a technical advance but also a cognitive framework for understanding how LLMs internalize linguistic hierarchy—providing pathways toward more transparent, controllable, and trustworthy AI systems.

5 Conclusion

This work introduces the **Multi-Scale Manifold Alignment (MSMA)** framework, a unified theory for interpreting and controlling large language models by decomposing their internal representations into local, intermediate, and global semantic manifolds. Our results show that LLMs inherently organize semantics hierarchically across these three scales, though the distribution varies by architecture (e.g., GPT-2 emphasizes intermediate reasoning, BERT favors local structure, and T5 exhibits symmetric hierarchy). By enforcing geometric preservation, information retention, and manifold smoothness, MSMA achieves near-perfect alignment (99% KL reduction, 5–7$\times$ mutual information gain) and enables fine-grained interventions—editing word choice, sentence structure, or discourse coherence with precision. Beyond interpretability, the framework bridges theoretical understanding and practical control, supporting applications in bias mitigation, robustness enhancement, and controlled text generation, thereby advancing the goal of building transparent, stable, and trustworthy AI systems.

6 Limitations

Despite the significant progress afforded by the Multi-Scale Manifold Alignment (MSMA) frame-

work in elucidating the internal mechanisms of large language models, several limitations remain. First, the computational cost of MSMA is substantial: estimating mutual information and manifold curvature across every layer of models with hundreds of billions of parameters (e.g., GPT-4, PaLM) demands considerable resources. Second, the semantic boundaries we detect may blur in architectures that employ hybrid or sparse attention mechanisms, necessitating tailored boundary-detection strategies for non-standard designs. Third, although our experiments used general-purpose text corpora, the layerwise semantic organization may differ in highly specialized domains (e.g., medical or legal texts) or in fine-tuned models, calling for cross-domain validation and adaptation of the framework.

Moreover, our theoretical analysis relies on simplifying assumptions—such as Markovian transitions and conditional independence among representation scales—that hold only approximately in practice, especially in the presence of residual connections and cross-attention. We have not yet established a direct correspondence between model representations and human cognitive processes; integrating insights from neuroscience and psycholinguistics could strengthen this link. In our intervention studies, we observed that effect sizes sometimes attenuate or behave non-linearly over long generation sequences, a dynamic phenomenon not fully captured by the current theory.

Finally, while we evaluated alignment quality using KL divergence, mutual information, and distance-based metrics, these measures may not fully reflect the richness of semantic content or downstream task performance. Likewise, existing visualization tools struggle to convey high-dimensional structure to non-technical audiences. Developing more comprehensive evaluation metrics and interactive visual interfaces will be critical for broadening MSMA's applicability and interpretability.

7 Acknowledgements

During the writing of this article, generative artificial intelligence tools were used to assist in language polishing and literature retrieval. The AI tool helped optimize the grammatical structure and expression fluency of limited paragraphs, and assisted in screening research literature in related fields. All AI-polished text content has been strictly reviewed by the author to ensure that it complies with academic standards and is accompanied by accurate citations. The core research ideas, method design and conclusion derivation of this article were independently completed by the author, and the AI tool did not participate in the proposal of any innovative research ideas or the creation of substantive content. The author is fully responsible for the academic rigor, data authenticity and citation integrity of the full text, and hereby declares that the generative AI tool is not a co-author of this study.

References

Yoshua Bengio, Aaron Courville, and Pascal Vincent. 2013. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, and 95 others. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, and 48 others. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

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.

Andy Coenen, Emily Reif, Ann Yuan, Been Kim, Adam Pearce, Fernanda Viégas, and Martin Wattenberg. 2019. Visualizing and measuring the geometry of bert. *Advances in Neural Information Processing Systems*, 32.

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*.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 55–65.

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.

Shun ichi Amari. 2016. *Information Geometry and Its Applications*. Springer.

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

Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. 2019. Linguistic knowledge and transferability of contextual representations. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1073–1094.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. *Advances in Neural Information Processing Systems*, 35:17359–17372.

Paul Michel, Omer Levy, and Graham Neubig. 2019. Are sixteen heads really better than one? In *Advances in Neural Information Processing Systems*, volume 32, pages 14014–14024.

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

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in bertology: What we know about how bert works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. Bert rediscovers the classical nlp pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008.

Jesse Vig. 2019. A multiscale visualization of attention in the transformer model. *arXiv preprint arXiv:1906.05714*.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. 2019. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5797–5808.

A Experimental Setup and Analysis for Semantic-Scale Identification

A.1 Experimental Design

Research Questions. We evaluate three hypotheses of MSMA: (1) whether Transformer layers form identifiable *local/intermediate/global* semantic scales; (2) how architecture and pre-training objectives affect these scales; (3) whether targeted interventions yield the predicted scale-specific effects.

A.1.1 Models

We evaluate representative LLMs (Table 5).

Table 5: Evaluated models.

Model	Architecture	Params	Pretrain Objective
GPT-2	Autoregressive Decoder	1.5B	Next-token Prediction
BERT	Bidirectional Encoder	340M	Masked LM
RoBERTa	Enhanced BERT Encoder	355M	Dynamic Masked LM
T5	Encoder–Decoder	11B	Sequence-to-Sequence

A.1.2 Data Resources

We construct a balanced corpus of 20,000 samples from three sources (Table 6).

Table 6: Corpus composition and average sample length.

Source	# Samples	Avg. Length (tokens)
Brown (15 genres)	6,667	293.5
Reuters (8 topics)	6,667	318.2
GPT-2 academic synth	6,666	352.8

Brown: 15 genres, classic written English. **Reuters:** 8 topic categories, global news. **GPT-2 academic synth:** academic-style texts generated from 68 field prompts and manually filtered.

A.1.3 Feature Hierarchies

We analyze features at three semantic scales: **Global** (genre, source, LDA topic, stylistic markers); **Intermediate** (mean sentence length, clause count, lexical complexity, topic coherence); **Local** (token length variance, function word ratio, POS/dependency distribution, sentiment).

A.1.4 Scale Identification Methods

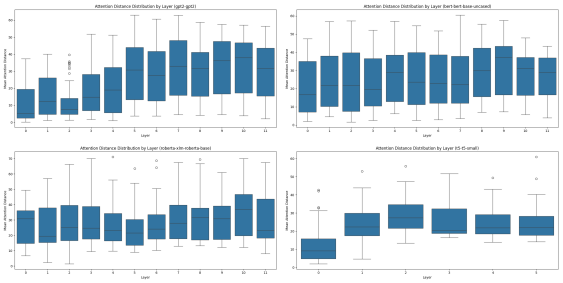
We combine three evidence sources with a voting scheme:

- **Attention patterns:** mean span $d_{\text{attn}}^{(\ell)} = \frac{1}{H}\sum_h \sum_{i,j} A_{i,j}\,|i-j|$ and entropy $H_{\text{attn}}^{(\ell)}$.
- **Representation similarity:** KL divergence and mutual information (kNN estimator; PCA to 50D).
- **Probing tasks:** layerwise SVMs for POS/dependency (local), next-sentence/paragraph (intermediate), and topic/genre (global).

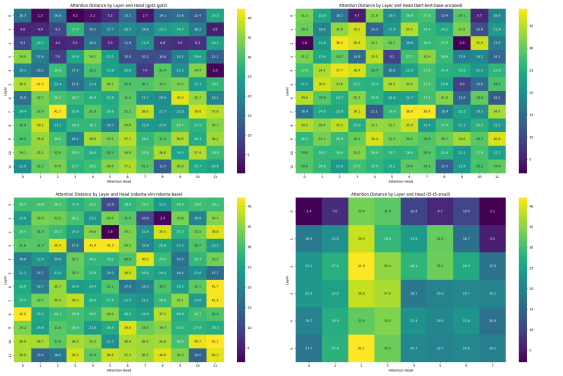
Voting: $S_{\text{scale}} = 0.4\,\text{Probe} + 0.4\,\text{Attn} + 0.2\,\text{MI}$, followed by continuity smoothing.

A.2 Layered Structure Revealed by Attention Patterns

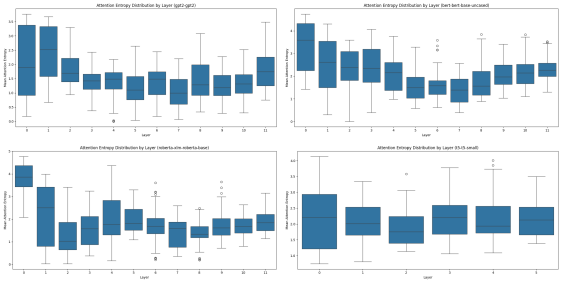
Figure 2b shows the mean attention span by layer. For GPT-2, span rises from 12.5 (layer 0) to 36.2 (layer 12), clustering as *local* (0–2, median < 15), *intermediate* (3–8, 15–30), and *global* (9–12, $>$ 30). BERT/RoBERTa show a smooth rise from 17.3 (layers 0–4) to above 30 (layers 9–12). T5 (six layers) shows clear encoder/decoder separation (encoder 12.4→27.8; decoder 14.2→31.5). Spearman correlations (span vs. depth) all exceed 0.85 ($p < 0.01$), supporting span as a scale indicator.



(a) Mean attention span by layer across models.



(b) Attention span distance heatmap.



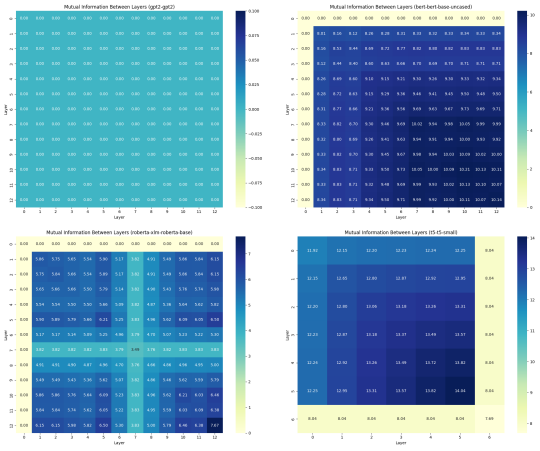
(c) Attention entropy by layer.

Figure 2: Comprehensive attention profile analysis for four Transformer models.

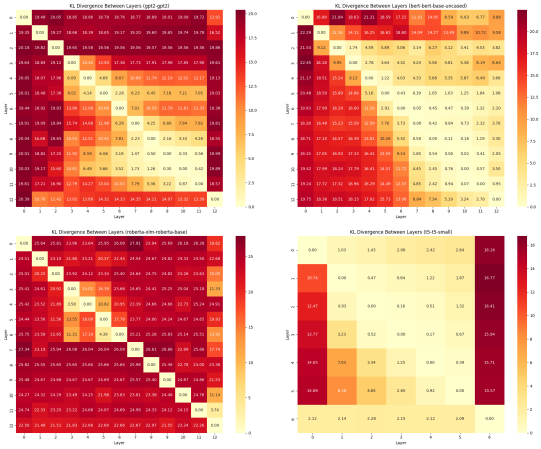
Figure 2c plots attention entropy: GPT-2 exhibits a U-shaped curve (peaks at 0–1, dip at 7), while BERT/RoBERTa dip at 5–8, consistent with intermediate layers; T5 shows a flatter pattern with an encoder dip.

A.3 Representation Similarity Confirms Semantic Boundaries

Figure 3b (layerwise KL) shows three blocks in GPT-2 (local/intermediate/global): KL jumps from 9.1 to 19.6 (layers 2→3) and from 6.7 to 17.9 (8→9). BERT/RoBERTa display similar boundaries; all jumps are significant ($Z > 2.0$, $p < 0.01$). Figure 3a (layerwise MI) shows BERT's MI matrix forming three modules $\{0\text{–}4, 5\text{–}8, 9\text{–}12\}$, with within-module MI $\sim$40% higher than between-module MI; RoBERTa/T5 show similar trends;



(a) Mutual information across models.



(b) KL divergence across models.

Figure 3: Comparative analysis of information metrics.

GPT-2 is noisier but consistent with its KL blocks.

A.4 Probing Tasks Validate Functional Specialization

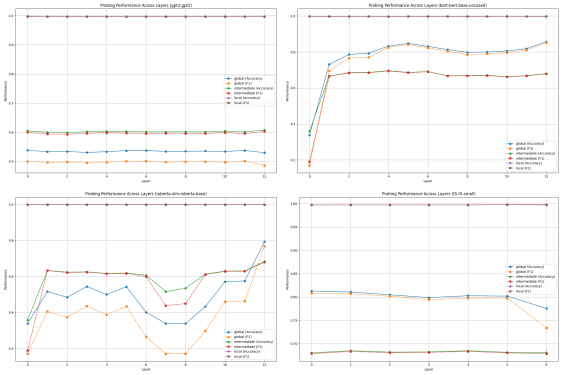
Figure 4a shows layerwise probing: BERT exhibits three regimes—layers 0–4 excel on local tasks (F1 from 0.18→0.77), 5–8 peak on intermediate tasks, and 9–12 on global tasks (accuracy > 0.82). GPT-2 attains near-perfect local F1 ($\sim$0.99) but lower global accuracy ($\sim$0.53). RoBERTa/T5 also display architecture-specific stratification. Probing peaks align with attention/MI boundaries across models.

B Interventions and Statistics

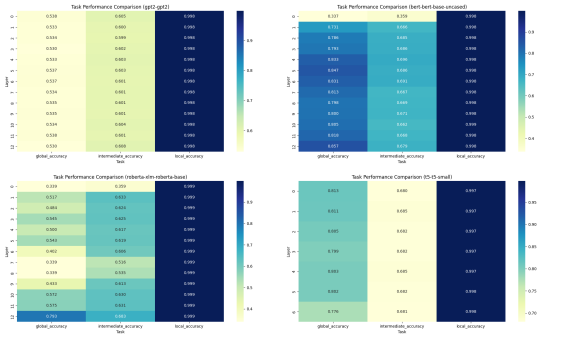
B.1 Intervention Designs

We apply four classes of perturbations to hidden representations at three hierarchical scales (local, intermediate, and global):

1. Translation: $\mathbf{h}'^{(\ell)} = \mathbf{h}^{(\ell)} + \Delta$



(a) Probing performance across models.



(b) Probing performance by layer.

Figure 4: Layerwise probing confirms specialization of scales.

2. Scaling: $\mathbf{h}'^{(\ell)} = \alpha\, \mathbf{h}^{(\ell)}$
3. Additive noise: $\mathbf{h}'^{(\ell)} = \mathbf{h}^{(\ell)} + \epsilon$, with $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$
4. Attention modification: $A'^{(\ell,h)}_{i,j} = f_{\text{att}}(A^{(\ell,h)}_{i,j})$

We sweep each mechanism over a grid of strengths (e.g., α, $\|\Delta\|$, σ) using fixed values across runs. To avoid instability, all interventions are consistently clipped as described in Table 7. No additional hyperparameters are introduced in this section.

B.2 Outcome Metrics

We track the following generation-level metrics under each intervention condition, benchmarked against non-intervened outputs:

- Lexical diversity (LexDiv)
- Sentence count
- Mean sentence length
- Maximum dependency depth
- Sentiment score
- Coherence score

B.3 Statistical Testing Protocol

For each model–scale–intervention configuration, we conduct 30 repetitions (>5,000 total samples). Significance is assessed using the Wilcoxon signed-rank test with FDR correction ($p<0.05$), and effect size is reported using Cliff's δ ($|\delta| > 0.147$ considered small but meaningful). Confidence intervals are estimated via percentile bootstrap (1,000 resamples), with additional robustness checks via leave-one-out analysis and power testing. If intervals are reported, they follow the same format as Table 7.

Note: $^*p < 0.05$, $^{**}p < 0.01$ (FDR). Cliff's δ: $+$ = increase, $-$ = decrease.

B.4 Full Results by Model

Table 7 summarizes all significant results grouped by model and scale. GPT-2 exhibits strong lexical sensitivity at the local scale, diversity–coherence trade-offs globally, and structural effects (sentence count, mean length) at the intermediate scale. BERT shows minimal change, with deviations concentrated in sentence count under global/intermediate interventions. XLM-R is sentiment-stable overall, with global noise yielding moderate shifts. No additional tables are included; full numerical details (median change, δ, p) are inline.

C MSMA Implementation Details and Ablations

C.1 Cross-Scale Mappings

To align local, intermediate, and global representations, we construct mappings between layer groups identified in Section A, experimenting with three parameterizations. The linear mapping $\mathbf{h}_{k+1} = W_k\mathbf{h}_k + b_k$, with $W_k \in \mathbb{R}^{d\times d}$, is initialized via Xavier uniform. The orthogonal mapping constrains $W_k^\top W_k = I$ through Cayley transform regularization to preserve distances. The nonlinear variant uses a two-layer MLP with GELU activation:

$$\mathbf{h}_{k+1} = W_2\,\mathrm{GELU}(W_1\mathbf{h}_k + b_1) + b_2.$$

Smooth transitions across scales are encouraged by curvature regularization:

$$\mathcal{L}_{\text{smooth}} = \|\nabla_{\mathbf{h}} f_\theta(\mathbf{h})\|_2^2,$$

with all mappings trained jointly in the MSMA objective.

C.2 Objectives and Estimators

The total loss combines three components:

$$\mathcal{L} = \lambda_{\text{geo}}\mathcal{L}_{\text{geo}} + \lambda_{\text{info}}\mathcal{L}_{\text{info}} + \lambda_{\text{curv}}\mathcal{L}_{\text{curv}}.$$

The geometric term $\mathcal{L}_{\text{geo}}$ minimizes cross-scale distortion by preserving pairwise distances:

$$\mathcal{L}_{\text{geo}} = \sum_{i<j} \left| d(\mathbf{h}_i, \mathbf{h}_j) - d(\hat{\mathbf{h}}_i, \hat{\mathbf{h}}_j) \right|.$$

The information term $\mathcal{L}_{\text{info}}$ maximizes mutual information between adjacent scales, estimated via MINE:

$$\mathcal{L}_{\text{info}} = -I_\phi(\mathbf{h}_k; \mathbf{h}_{k+1}) = -\mathbb{E}[T_\phi] - \log \mathbb{E}[e^{T_\phi}],$$

where T_ϕ is the discriminator network. The curvature regularization $\mathcal{L}_{\text{curv}}$ stabilizes learning via the Laplace–Beltrami operator:

$$\mathcal{L}_{\text{curv}} = \|\Delta_{\mathcal{M}}\mathbf{h}\|_2^2,$$

approximated with $O(Nd)$ complexity per step.

C.3 Training Setup

Training proceeds in two stages: (1) unsupervised alignment, (2) fine-tuning with cross-scale regularization. We use AdamW with $(\beta_1, \beta_2) = (0.9, 0.98)$, learning rate 3×10^{-4} (warm-up + cosine decay), batch size 64 per GPU, and 50–80 epochs with early stopping ($\Delta\mathcal{L} < 10^{-4}$ for 5 epochs). Mixed-precision FP16 training is adopted. All experiments use fixed seeds for reproducibility.

C.4 Ablation Suites

To isolate the effect of each loss component, we train several variants: No-Geo removes geometric alignment, No-Info excludes information retention, and No-Curv omits curvature control. Only-* versions optimize a single term for baseline comparison. All models share identical hyperparameters and evaluation protocols from Section B, ensuring fair comparison across alignment, stability, and mutual information.

C.5 Hyperparameter Sensitivity

We sweep $\lambda_{\text{geo}}, \lambda_{\text{info}}, \lambda_{\text{curv}} \in \{10^{-3}, 10^{-2}, 10^{-1}, 1, 10\}$ and observe that balanced weighting (1:1:0.1) yields the best trade-off between alignment quality and robustness. Trends are illustrated in Figure **??** (not shown).

Table 7: Significant intervention effects across models ($p < 0.05$, $|\delta| > 0.10$). Median changes (%) are relative to baseline.

Model	Scale	Intervention	Metric	Median Change (%)	Cliff's δ	p-value	Sig.
GPT-2	Global	Scale up	Lexical diversity	+7.39	0.232	0.020	*
GPT-2	Global	Scale up	Coherence score	0.00	−0.238	0.007	**
GPT-2	Global	Scale down	Lexical diversity	+6.78	0.272	0.017	*
GPT-2	Intermed.	Translate	Lexical diversity	+6.60	0.316	0.014	*
GPT-2	Intermed.	Scale up	Sentence count	+25.00	0.239	0.028	*
GPT-2	Intermed.	Scale up	Mean sent. length	−19.04	−0.266	0.004	**
GPT-2	Intermed.	Scale up	Max dep. depth	−11.11	−0.203	0.030	*
GPT-2	Intermed.	Scale down	Lexical diversity	+5.84	0.211	0.016	*
GPT-2	Intermed.	Scale down	Max dep. depth	−11.11	−0.192	0.037	*
GPT-2	Intermed.	Attn	Lexical diversity	+4.55	0.195	0.028	*
GPT-2	Intermed.	Attn	Sentiment score	−80.09	−0.246	0.004	**
GPT-2	Local	Translate	Coherence score	0.00	−0.180	0.020	*
GPT-2	Local	Scale up	Lexical diversity	+7.27	0.342	0.005	**
GPT-2	Local	Scale up	Sentiment score	−71.84	−0.206	0.020	*
GPT-2	Local	Scale down	Lexical diversity	+5.62	0.276	0.015	*
GPT-2	Local	Scale down	Coherence score	0.00	−0.180	0.037	*
BERT	Global	Noise	Sentence count	0.00	0.154	0.046	*
BERT	Intermed.	Translate	Sentence count	0.00	0.154	0.033	*
BERT	Intermed.	Attn	Sentence count	0.00	0.269	0.003	**
XLM-R	Global	Noise	Sentiment score	−13.58	0.243	0.005	**
XLM-R	Intermed.	Scale up	Sentiment score	−1.03	0.104	0.046	*
XLM-R	Local	Attn	Sentiment score	−10.79	0.149	0.043	*

D Theoretical Assumptions and Applicability

This appendix outlines the theoretical assumptions of our MSMA framework and assesses their empirical validity in Transformer models.

D.1 Hierarchical Markov Property

Assumption D.1.1 (Hierarchical Markov Property)**.** *Information flows $\mathcal{M}_L \to \mathcal{M}_I \to \mathcal{M}_G$. Given h_G, h_I is conditionally independent of z; given h_G, h_I, h_L is conditionally independent:*

$$p(h_I|h_G, z) \approx p(h_I|h_G),$$
$$p(h_L|h_G, h_I, z) \approx p(h_L|h_G, h_I)$$

where z denotes nuisance variables.

Applicability. This is plausible for feedforward architectures, though residual and attention layers in Transformers introduce skip dependencies.

Empirical Verification. Conditional mutual information (CMI) analysis (Fig. **??**) on GPT-2 reveals:

- High CMI for adjacent layers: $I(h_i; h_{i+1}) > 0.8$
- Low CMI for distant layers given intermediate: $I(h_i; h_j|h_{(i+j)/2}) < 0.2$ for $|i - j| > 5$
- Within-scale MI is 3–5× higher than cross-scale

These findings support a weak Markov approximation for representation flow.

D.2 Local Euclidean Property

Assumption D.2.1 (Local Euclidean Property)**.** *Local neighborhoods of the manifold are approximately Euclidean: for nearby $x, y \in \mathcal{M}$,*

$$d(x, y) \approx \|x - y\|$$

Applicability. Common in manifold learning and valid in neural representations away from singularities (Coenen et al., 2019).

Empirical Support. Linear alignment metrics (Sec. 4) show distance correlation (DC) > 0.99. We compute:

$$\text{Linearity}(r) = \frac{\|\mathbf{h}_i - \mathbf{h}_j\|}{\text{geodesic}(i, j)}$$

and find Linearity > 0.95 for $r < 0.1 \cdot \text{diam}(\mathcal{M})$, confirming Euclidean behavior locally.

D.3 Bounded Curvature

Assumption D.3.1 (Bounded Curvature)**.** *Manifold curvature satisfies* $\sup_{\mathcal{M}} |K| < \infty$.

Applicability. Compact manifolds satisfy this automatically; we enforce curvature bounds via $\mathcal{L}_{\text{curv}}$.

Empirical Support. Fig. **??** shows curvature distribution before/after MSMA training: initially heavy-tailed with outliers ($\max|K| > 100$), later concentrated near 0 ($\max|K| < 10$).

D.4 Joint Distribution Factorization

We assume:

$$p(h_G, h_I, h_L|C) = p(h_G|C)\cdot p(h_I|h_G, C)\cdot p(h_L|h_I, h_G, C) \tag{12}$$

to enable analytical decomposition (see Theorem **??**).

Limitation. Residuals violate strict factorization since $h_{l+1} = h_l + f(h_l)$ implies dependency on all prior layers. However, we observe direct dependency ($p(h_{l+1}|h_l)$) dominates and skip influence decays exponentially.

D.5 Summary of Validity Across Models

Table 8: Assumption Validity Across Architectures

Assumption	GPT-2	BERT	T5
Hierarchical Markov	✓	✓	$\sim$
Local Euclidean	✓	✓	✓
Bounded Curvature	$\sim^*$	✓*	$\sim^*$
Factorization	$\sim$	$\sim$	$\times$

*Valid after curvature regularization

E Complete Theoretical Proofs

E.1 Preliminaries: Information Geometry

Definition E.1.1 (Statistical Manifold)**.** *Given a probability family $\{p(x|\theta)\}_{\theta\in\Theta}$, the statistical manifold is:*

$$\mathcal{M} = \{p(x|\theta) : \theta \in \Theta\}. \tag{7}$$

Definition E.1.2 (Fisher Information Matrix)**.**

$$g_{ij}(\theta) = \mathbb{E}_{p(x|\theta)}\left[\frac{\partial \log p}{\partial \theta_i}\frac{\partial \log p}{\partial \theta_j}\right] \tag{8}$$

Lemma E.1.1 (KL-Fisher Relationship)**.** *For infinitesimal $d\theta$:*

$$D_{\text{KL}}(p(\cdot|\theta)\|p(\cdot|\theta+d\theta)) = \frac{1}{2}d\theta^\top g(\theta)d\theta + O(\|d\theta\|^3) \tag{9}$$

Proof. Taylor expand KL divergence:

$$D_{\text{KL}} = \int p(x|\theta)\log\frac{p(x|\theta)}{p(x|\theta+d\theta)}dx \tag{10}$$

$$= -\int p(x|\theta)\log p(x|\theta+d\theta)dx + \text{const} \tag{11}$$

Expand $\log p(x|\theta+d\theta)$ to second order:

$$\log p(x|\theta+d\theta) = \log p(x|\theta) + \sum_i \frac{\partial \log p}{\partial \theta_i}d\theta_i \tag{12}$$

$$+ \frac{1}{2}\sum_{ij}\frac{\partial^2 \log p}{\partial\theta_i\partial\theta_j}d\theta_i d\theta_j + O(\|d\theta\|^3) \tag{13}$$

Taking expectation under $p(x|\theta)$, the first-order term vanishes by $\mathbb{E}[\nabla \log p] = 0$. The second-order term gives:

$$\mathbb{E}\left[\frac{\partial^2 \log p}{\partial\theta_i\partial\theta_j}\right] = -g_{ij} \tag{14}$$

Therefore,

$$D_{\text{KL}} = \frac{1}{2}d\theta^\top g(\theta)d\theta + O(\|d\theta\|^3). \tag{15}$$

□

E.2 Proof of Alignment Error Bound (Theorem 3.1)

Theorem E.1 (Alignment Error Bound)**.** *Assume f_{GI}, f_{IL} are L_1, L_2-Lipschitz. If $\varepsilon_{geo}, \varepsilon_{info}$ bound geometric and info errors,*

$$D_{KL}(p_{true}\|p_{aligned}) \le C(\varepsilon_{geo} + \varepsilon_{info}) \tag{16}$$

Proof. Let $p_{\text{true}} = p(h_G, h_I, h_L)$ and $p_{\text{aligned}} = p(h_G, f_{GI}(h_G), f_{IL}(f_{GI}(h_G)))$.

Step 1: KL Decomposition

$$D_{\text{KL}} = \mathbb{E}_{h_G}[D_{\text{KL}}(p(h_I|h_G)\|p(f_{GI}(h_G)|h_G))] \tag{17}$$

$$+ \mathbb{E}_{h_G,h_I}[D_{\text{KL}}(p(h_L|h_I,h_G)\|p(f_{IL}(h_I)|h_I,h_G))] \tag{18}$$

Step 2: Bounds via Lemma E.1.1

$$D_{\text{KL}}(p(h_I|h_G)\|p(f_{GI}(h_G)|h_G)) \le C_1\varepsilon_{\text{geo}}^{(GI)} \tag{19}$$

$$I(h_I; y) - I(f_{GI}(h_G); y) \le \varepsilon_{\text{info}}^{(GI)} \tag{20}$$

Step 3: Error Propagation

$$\|f_{IL}(h_I) - f_{IL}(\hat{h}_I)\| \leq L_2 L_1 \|h_G - \hat{h}_G\| \tag{21}$$

$$D_{\text{KL}}(p(h_L|h_I, h_G)\|p(f_{IL}(h_I)|h_I, h_G)) \leq C_2(\varepsilon_{\text{geo}}^{(IL)} + \varepsilon_{\text{info}}^{(IL)}) \tag{22}$$

Step 4: Combine

$$D_{\text{KL}} \leq C(\varepsilon_{\text{geo}} + \varepsilon_{\text{info}}) \tag{23}$$

□