INFORMATION SYSTEMS RESEARCH

Identifying Compassion in Large Language Models: A Comparative Analysis of Four Transformers

Journal:	*Information Systems Research*
Manuscript ID	ISRE-2026-3103
Manuscript Type:	Special Issue: Compassionate AI
Manuscript Category:	Special Issue
Keywords:	Large Language Models, Compassion, Transformers, Mechanistic Interpretability
Abstract:	Standard behavioral benchmarks cannot guarantee the internal trustworthiness of AI in emotionally sensitive contexts. We introduce a neuro-informational framework for "Architectural Auditing" that maps the internal computational pathways of decoder-only Large Language Models (LLMs). Analyzing four distinct architectures (Llama, Mistral, Qwen, and DeepSeek), we reveal the "Empathy Paradox": models producing identical, seemingly empathetic outputs often utilize profoundly different reasoning strategies. We demonstrate that processing efficiency does not predict compassionate capability. Instead, our results support the Calibration Hypothesis: trustworthiness is driven by an architecture's ability to calibrate final representations, often requiring "inefficient," non-monotonic correction phases ("Adaptive Refinement") rather than straight-line processing ("Rigid Linearization"). Furthermore, we uncover latent safety risks such as "Negative Priming" in compact models, providing the Information Systems field with a rigorous methodology to audit the structural flexibility and safety of high-stakes AI agents.

Identifying Compassion in Large Language Models: A Comparative Analysis of Four Transformers

Abstract

Standard behavioral benchmarks cannot guarantee the internal trustworthiness of AI in emotionally sensitive contexts. We introduce a neuro-informational framework for "Architectural Auditing" that maps the internal computational pathways of decoder-only Large Language Models (LLMs). Analyzing four distinct architectures (Llama, Mistral, Qwen, and DeepSeek), we reveal the "Empathy Paradox": models producing identical, seemingly empathetic outputs often utilize profoundly different reasoning strategies. We demonstrate that processing efficiency does not predict compassionate capability. Instead, our results support the Calibration Hypothesis: trustworthiness is driven by an architecture's ability to calibrate final representations, often requiring "inefficient," non-monotonic correction phases ("Adaptive Refinement") rather than straight-line processing ("Rigid Linearization"). Furthermore, we uncover latent safety risks such as "Negative Priming" in compact models, providing the Information Systems field with a rigorous methodology to audit the structural flexibility and safety of high-stakes AI agents.

Keywords: Large Language Models (LLMs); Compassion; Transformers; Mechanistic Interpretability

1. Introduction

A growing body of IS research highlights the urgent need to move beyond purely utilitarian views of artificial intelligence. As AI systems become embedded in high-stakes organizational and societal contexts, such as healthcare delivery and crisis intervention (Stade et al., 2024), their ability to navigate human emotion becomes a core requirement for system success. Scholars have called for a paradigm shift toward "Compassionate AI": systems that are not merely empathetic in their outputs (Liu-Thompkins et al., 2022) but are also demonstrably trustworthy and coherent in their internal operations (Kerasidou, 2020).

A central challenge in this shift is the persistent "Black Box" problem. Raman & McClelland (2019) identify a critical gap between the technical opacity of neural architectures and the organizational need for transparent, auditable, and genuinely human-aligned systems. Currently, model selection and governance rely almost exclusively on Behavioral Auditing—the evaluation of external performance benchmarks. However, our research reveals a phenomenon we term the "Empathy Paradox": two models can produce identical, seemingly empathetic outputs while utilizing profoundly different internal reasoning strategies. Without visibility into the internal cognitive architecture, organizations cannot distinguish between a robust, well-calibrated reasoner and a brittle system that mimics empathy via a rigid, pattern-matching script.

Our work directly answers this call by proposing a neuro-informational framework for Architectural Auditing. We move beyond analyzing input-output behavior to mapping the internal "computational pathways"—the layer-by-layer evolution of a thought—as models navigate socio-emotional concepts. This approach is grounded in the theoretical frontier of Mechanistic Interpretability.

While traditional approaches focused on isolated "circuit-level" analysis (Elhage et al., 2021), we adopt a "geometric" perspective, treating the sequence of a model's hidden states as a continuous trajectory through a high-dimensional representational space. The validity of this geometric analysis rests on the recent mathematical proof that decoder-only Transformers are injective (Crisostomi et al., 2025), meaning their internal hidden states are not lossy abstractions but deterministic fingerprints of the input.

To operationalize this framework, we analyzed four representative architectures (Llama-3.1-8B, Mistral-7B, Qwen-3-4B, and DeepSeek-1.5B). To address the challenge of comparing heterogeneous architectures (e.g., 1536-dim vs. 4096-dim spaces), we employ learned linear transformations to align these diverse spaces into a single, interpretable 2D reference frame defined by Emotional Valence and Social Complexity. This allows us to visualize, quantify, and compare how different architectures "think" in real-time, moving the field from behavioral conjecture to geometric proof.

1.1 Academic Contributions

Our study makes contributions in several streams of literature.

First, given the black-box nature of LLM algorithms, we trace vectors of dimensional parameters across layers to deconstruct the "Empathy Paradox"—the phenomenon where models produce identical

outputs via profoundly different internal strategies. By analyzing starting and ending representational states, absolute and normalized trajectory distances, task differentiation timing, geometric uniformity, effort distribution, and phase-based efficiency, we systematically map the internal logic of diverse architectures. This granular analysis reveals that "compassion" is not a uniform cognitive process; rather, models achieve it through highly distinct "Cognitive Styles," varying from linear, arithmetic processing to complex, non-monotonic refinement. This establishes that behavioral similarity (output) does not imply cognitive equivalence (process), necessitating a shift from output-based to process-based auditing.

Second, we challenge the prevailing engineering intuition that processing efficiency predicts compassionate capability by uncovering two distinct "Cognitive Styles." We empirically identify "Rigid Linearization" (exemplified by DeepSeek), where a model follows a highly efficient, straight-line path but fails to decouple emotional tone from social nuance, leading to brittle performance. In contrast, we identify "Adaptive Refinement" (exemplified by Qwen), where a model executes a circuitous, non-monotonic "U-turn" to self-correct initial biases. This finding demonstrates that in the socio-emotional domain, "inefficiency" is often a necessary computational cost for calibration. These empirical signatures falsify the link between pathway directness and model performance, establishing the Calibration Hypothesis: for high-stakes AI agents, the internal capacity to reach a well-resolved semantic state (the destination) is a critical predictor of trustworthiness, independent of the efficiency of the path taken (the journey).

Third, we uncover the existence of "Latent Architectural Bias" in the form of "Negative Priming," a safety risk invisible to standard behavioral benchmarks. Our audit reveals that compact architectures (e.g., Qwen, DeepSeek) systematically initialize socio-emotional scenarios in a state of "affective distress" (negative valence) before any reasoning occurs. This finding implies that certain architectures are structurally predisposed to negativity, posing a hidden risk of escalation in sensitive use cases like crisis intervention. This redefines "AI Safety" in the context of compassion: it is not enough for a model to say the right thing; it must not be "internally panicked" at the encoding stage. This identifies initialization bias as a critical new variable for the governance of safety-critical AI agents.

2. Methodology

2.1. Model Selection

Our selection criteria were based on achieving a spectrum of architectural philosophies, parameter scales, and "cognitive styles." A useful feature of the decoder-only architecture is its unidirectional flow: information moves strictly from the first layer to the last. This allows us to use network depth as a proxy for cognitive time, where each layer represents a discrete step in the "revision" of a thought. We analyzed four representative decoder-only Large Language Models:

1. Llama-3.1-8B-Instruct (Meta): Selected as our reference model due to its status as a state-of-the-art generalist system. It features 32 layers and 4096-dimensional hidden states. Notably, Llama-3.1 utilizes a massive tokenizer vocabulary (128k tokens), which requires significantly more parameters in the embedding layers than its peers, potentially impacting its "front-loaded" processing style.

2. Mistral-7B-Instruct-v0.2 (Mistral AI): Chosen as a direct capacity competitor to Llama (7B parameters, 32 layers, 4096-dim). Mistral represents a "conservative" architectural approach, favoring stability and large representational jumps over granular refinement.

3. Qwen-3-4B-Thinking-2507 (Alibaba Cloud): A mid-capacity model (4B parameters) with a deeper architecture (36 layers) and a smaller hidden dimension (2560-dim). It was included to test the hypothesis that specialized "thinking" models utilize non-monotonic pathways for self-correction.

4. DeepSeek-R1-Distill-Qwen-1.5B (DeepSeek AI): A compact, highly efficient architecture (1.5B parameters, 28 layers, 1536-dim). It represents the growing class of "distilled" models deployed in resource-constrained environments where processing efficiency is often prioritized over flexibility.

2.2. Constructing a Valence–Social Complexity Space

A central challenge in AI interpretability is the "semantic gap": while we can extract 4,096-dimensional vectors, these numbers are natively uninterpretable. Traditional dimensionality reduction techniques like Principal Component Analysis (PCA) are "blind" to meaning, as they maximize variance rather than semantic coherence. To bridge this gap, we employ a Concept-Based Axis Construction protocol, mapping the high-dimensional trajectories into a validated 2D subspace.

2.2.1. Theoretical Foundation: Linear Representation

Our framework relies on the Linear Representation Hypothesis, a cornerstone of mechanistic interpretability which posits that neural networks encode semantic concepts (e.g., "Joy" or "Social Complexity") as linear directions in high-dimensional space. Under this hypothesis, if a model's internal state moves physically closer to a "Joy" vector, it is not moving randomly; it is mechanically activating the concept of joy within its reasoning process. Drawing on affective science (Russell & Barrett, 1999) and social cognition research (Fiske et al., 2007), we identified two fundamental dimensions that define the socio-emotional manifold:

Dimension 1: Emotional Valence. The "hedonic" quality of an experience, ranging from pleasant (positive) to unpleasant (negative).

Dimension 2: Social Complexity. The interpersonal versus intrapersonal processing continuum distinguishing between externally directed social navigation and internally experienced sensations.

2.2.2. Prompts

To elicit the internal "computational pathways" of these models, we designed a comprehensive cognitive benchmark comprising 249 prompts across 16 task categories. These categories span the full range of LLM

capabilities, including logical reasoning (Math, Pattern Recognition), executive function (Working Memory), and linguistic creativity. For the primary "Architectural Audit," we isolated a specialized Socio-Emotional Subspace consisting of 28 prompts. These were designed to be "high-stakes" and "nuanced," requiring multi-step reasoning:

1. Emotion Tasks (15 prompts): Focused on affective understanding and regulation. Examples include: "Describe the physiological response to sudden, unexpected joy" and "Explain the difference between empathy and sympathy."

2. Social Intelligence Tasks (13 prompts): Focused on interpersonal navigation and theory-of-mind. Examples include: "How would you politely interrupt someone who is talking for too long in a meeting?" and "How would you respond to a friend who is sharing good news while you are secretly feeling envious?"

2.2.3. Formulation

To operationalize these dimensions, we fed four "Anchor Prompts" through our reference model (Llama-3.1-8B) and extracted their final-layer activation vectors. These vectors serve as the "north stars" for our coordinate system.

1. Defining the Valence Axis ($v_{valence}$): We computed the vector difference between the "Positive Pole" (Joy) and the "Negative Pole" (Sadness). This directional vector captures what makes the concepts different according to the model's internal representation. We then normalized it to a unit vector to ensure that subsequent projections indicate semantic position rather than magnitude:

$$v_{valence\_raw} = v_{joy} - v_{sadness}$$

$$v_{valence} = v_{valence\_raw} / ||v_{valence\_raw}||_2$$

2. Defining the Social Complexity Axis (vcomplexityvcomplexity): Similarly, we contrasted "High Complexity" (Interpersonal interaction) with "Low Complexity" (Intrapersonal sensation):

$$v_{complexity\_raw} = v_{interrupt} - v_{anxiety}$$

3. Gram-Schmidt Orthogonalization: To ensure clean interpretability, the two axes must be independent. If they overlap, a model's position on one axis could mechanically determine its position on the other. We used the Gram-Schmidt process to subtract any component of the valence axis from the complexity axis, ensuring they are perfectly perpendicular (dot product ≈0):

$$v_{complexity\_orthogonal} = v_{complexity\_raw} - (v_{complexity\_raw} \cdot v_{valence})\, v_{valence}$$

$$v_{complexity} = v_{complexity\_orthogonal} / || v_{complexity\_orthogonal} ||_2$$

This ensures that "Valence" and "Complexity" are treated as distinct cognitive variables in our audit. A full step-by-step example can be read in Appendix D.

2.3. Constructing Layers-by-Dimensions Activation Matrix

To capture the internal dynamics, we conceptualize the Transformer architecture as an assembly line of information processing. Rather than observing only the final output, we take "snapshots" of the hidden states at every stage of the process. There are four parts in this process.

1. The Residual Stream: Following established practices in mechanistic interpretability (Elhage et al., 2021), we registered forward hooks on the residual stream of every transformer layer. The residual stream is the primary information-carrying highway of the model; by tapping into it, we capture the core "representation" as it is transformed from layer to layer.

2. Trajectory Mapping: For each prompt, we extracted a layer-wise activation trajectory $T \in \mathrm{R}^{L \times \mathrm{d}}$, where L is the number of layers and d is the native hidden dimension.

3. Preservation of Information: Crucially, we rejected dimensionality reduction techniques (like PCA) at the extraction stage. We preserved the full dimensionality of each model's native representational space ($1536 \leq \mathrm{d} \leq 4096$) to ensure that no "nuance" or "noise" was discarded before the analysis.

4. Functional Phases: We treat the trajectory as an "Evolving Draft" with Early Layers: The model "reads" the input, resolving syntax and basic definitions; Middle Layers: The model acts as an "editor," adding context and performing abstract reasoning; and, Final Layers: The model prepares the "final draft," converging on a specific semantic state to predict the next token (the output).

This extraction protocol provides the high-resolution data necessary to visualize how different architectures "think" through a problem in real-time.

2.4. Comparative Analysis Protocol: The "Dual-View" Framework

A primary challenge in cross-architectural auditing is scale heterogeneity. Activation magnitudes vary dramatically across models due to architectural artifacts (e.g., LayerNorm strategies, residual scaling) rather than semantic differences. Visualizing raw data alone risks obscuring the internal dynamics of compact models, while standard normalization alone destroys valuable information about absolute semantic convergence.

To resolve this, we employ a Dual-Visualization Strategy that separates positional claims (where the model is) from topological claims (how the model moves).

2.4.1. Visualization 1: Absolute Positioning (Auditing Consensus & Bias)

We first project the raw centroid trajectory ($vec_l = T_{mean}[l]$) onto our validated semantic axes. For each layer l, we compute the dot product of the average hidden state with the valence and complexity axes: $P_{absolute}[l] = [Valence_l, Complexity_l]$, where $Valence_l = vec_l \cdot v_{valence}$. This view audits Consensus and Bias. It reveals the "Starting Configuration" (does the model initialize in a state of negative distress?) and "Endpoint Convergence" (does it reach the same semantic conclusion as the reference model?).

2.4.2. Visualization 2: Normalized Topology (Auditing Strategy)

We subsequently apply per-model z-score normalization to equalize the visual scale. We independently normalize each dimension d based on the model's own mean μ and standard deviation σ: $P_{normalized}[:,d] = [P_{absolute}[:,d] - \mu_d] / \sigma_d$. This view audits Computational Strategy. By removing magnitude differences, we isolate the geometric shape of the reasoning process. It allows us to distinguish between monotonic (straight-line) processing and non-monotonic (corrective) pivoting.

2.5. Quantitative Diagnostic Metrics

To move beyond visual inspection, we quantify the internal reasoning strategy using four architectural metrics.

2.5.1. Trajectory Directness Score (TDS): Quantifying Efficiency

The TDS measures the geometric efficiency of the normalized pathway. It captures the trade-off between representational exploration (which requires non-direct paths) and computational efficiency (which favors direct transformation). TDS=ddirect/dcumulative, where ddirect is the distance from start to end, and dcumulative is the total path length.

Phase-Based Analysis: To distinguish between random noise and purposeful "pivoting," we further decompose efficiency into three windows: Contextualization (0–20%), Reasoning (20–80%), and Convergence (80–100%). This reveals phenomena such as the "Reasoning Dip," a characteristic drop in mid-stage efficiency associated with healthy semantic exploration.

2.5.2. Differentiation Timing: Speed of Contextualization

This metric quantifies the layer depth at which a model successfully distinguishes between disparate task types (e.g., "Math" vs. "Emotion"). Derived from the accuracy of a linear probing classifier trained on hidden states (detailed in Appendix A). A low score (Early Differentiation) suggests the model acts as an "Early Recognizer," rapidly orienting its latent space to the socio-emotional context to minimize the need for late-stage error correction.

2.5.3. Domain Uniformity: Geometric Plasticity

We measure the flexibility of a model's internal geometry by comparing its pathway shapes across 16 diverse cognitive domains (e.g., Logic, Creativity, Emotion). Derived from the variance in Procrustes distances across task domains. Here, high Uniformity indicates "Architectural Rigidity"—a one-size-fits-all approach where the model processes empathy with the same geometric shape as calculus. Low Uniformity indicates "High Plasticity," suggesting the model actively "rewires" its functional subspaces to suit the specific nuances of the task.

2.5.4. Effort Distribution: Center of Mass

This metric identifies where in the network the majority of "computational work" (representational transformation) occurs. Derived from the layer-wise trajectory curvature. We distinguish between "Front-

Loaded" architectures (settling into a stable state early) and "Back-Loaded" architectures (deferring major semantic decisions to the final layers).

2.6. External Performance Measures

To test the practical relevance of our internal architectural metrics, we correlated our findings with three tiers of established external benchmarks. This step is critical to validate our **Calibration Hypothesis**: the theory that a model's internal ability to calibrate its representations is a key predictor of its suitability for compassionate applications.

2.6.1. MMLU: Assessing Socio-Emotional Knowledge

We utilized a targeted subset of the Massive Multitask Language Understanding (MMLU) benchmark (Hendrycks et al., 2021) as a proxy for a model's foundational knowledge of social and emotional domains. We isolated five subsets, (Moral Scenarios, Moral Disputes, Professional Psychology, Human Sexuality, Human Aging), that are most relevant to compassion and emotional/social intelligence.

2.6.2. EQ-Bench (v2): Measuring Emotional Intelligence

To move beyond factual knowledge, we employed EQ-Bench v2 (Paech, 2023) to assess emotional intelligence in complex social dialogues. Unlike standard QA tasks, this benchmark requires models to rate the intensity of subtle, conflicting emotions (e.g., empathy vs. pity) in high-stakes scenarios, scored against human baselines. This serves as a critical stress test for "Architectural Rigidity," allowing us to determine if a model can modulate social nuance or if it collapses when distinct emotions (like Valence and Complexity) must be decoupled.

2.6.3. Essay Evaluation: Nuanced Subjective Judgment (LLM-as-a-Judge)

The final validation tier assesses subjective judgment using a dataset of 137 essays centered on compassion. We employed an "LLM-as-a-Judge" framework, utilizing the consensus of two frontier models (Gemini-3-Pro and Kimi-K2-Thinking) as the reference standard. Experimental models scored essays on a 0–10 scale across "Compassion" and "Authorial Voice." We evaluated alignment with the judge consensus using Pearson Correlation (r) and Mean Absolute Error (MAE), determining whether internal architectural signatures (e.g., Qwen's "Pivot") translate into alignment with expert-level human-centric judgment.

By comparing the comparative analysis metrics against these external performance measures, we can determine whether "efficient" internal processing correlates with higher scores on codified socio-emotional knowledge.

3. Results and Discussion

We analyzed the computational pathways of four distinct architectures across our socio-emotional benchmark. By projecting these high-dimensional trajectories into our validated 2D subspace, we deconstruct the internal reasoning strategies that remain invisible to standard behavioral benchmarks.

3.1. Visualizing the "Empathy Paradox"

The primary visualization of our audit (Figure 1 and Figure 2) maps the layer-by-layer evolution of "thought" as models navigate the dimensions of Emotional Valence and Social Complexity. Our analysis reveals the "Empathy Paradox": models capable of producing textually similar, empathetic outputs utilize profoundly different, and often divergent, internal reasoning strategies to reach those outputs.

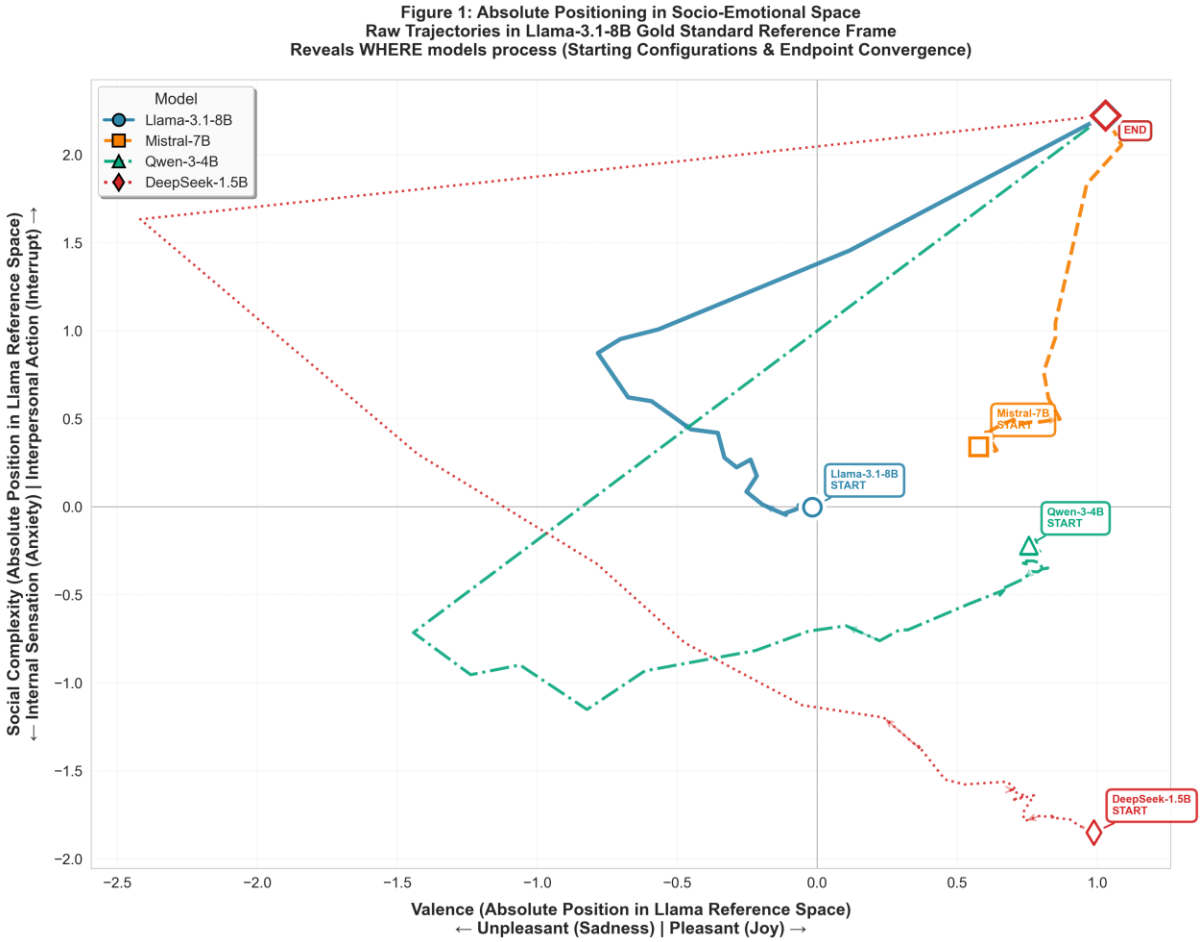

Figure 1: Shows the average computational pathways of all four models in the universal Llama-3.1-8B reference space. The coordinates are generated by projecting the mean hidden state (averaged across n = 28 socio-emotional prompts) onto the validated Valence and Social Complexity axes.

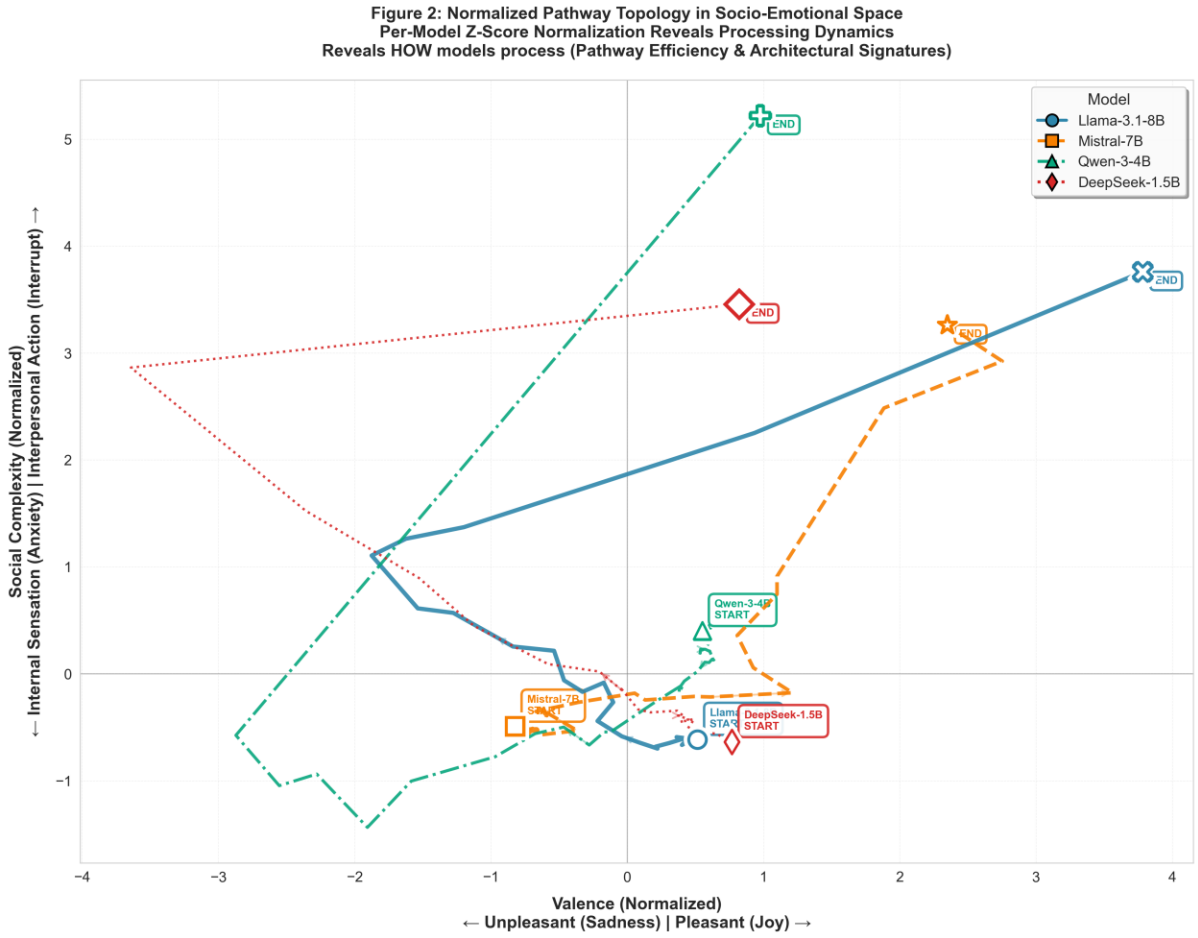

Figure 2: Presents the same computational pathways after per-model z-score normalization, equalizing scale to reveal pathway shape and processing dynamics. We confirmed that the non-monotonic shapes in normalized space correspond to significant magnitude shifts in absolute space, ruling out the amplification of noise.

Model	Dimension	Layers	Path Length	Direct Distance	TDS
Llama-3.1-8B	4096	32	4.82	4.10	0.85
Mistral-7B	4096	32	3.94	3.47	0.88
Qwen-3-4B	2560	36	6.73	3.43	0.51
DeepSeek-1.5B	1536	28	5.21	4.28	0.82

Table 1: Architectural Metrics in Socio-Emotional Processing.

*Note: Path Length and Direct Distance measured in standard deviation units (σ) on normalized trajectories (Figure 2). *

3.1.1. Absolute Positioning: Consensus, Divergence, and Calibration Failure

Figure 1 plots the raw centroid trajectories in the shared reference space. By analyzing the absolute coordinates of the starting and ending layers, we identify three distinct phenomena regarding how models converge on semantic meaning.

The Consensus Cluster (Successful Calibration): Despite differing training data and parameter counts, Llama-3.1-8B and Mistral-7B converge to a remarkably similar semantic endpoint (d=0.12 distance). Both models terminate in the quadrant characterized by Positive Valence and High Social Complexity. This suggests that for generalist architectures, there is a shared, objective "ground truth" for compassionate reasoning: a state that balances emotional positivity with social nuance.

Partial Correction (Qwen-3-4B): Qwen-3-4B demonstrates a "Partial Correction" strategy. While it successfully orients itself toward the positive valence of the consensus cluster, it remains semantically distinct (d=0.71 from Llama). As visible in the trajectory, it spends a significant portion of its depth correcting an initial negative bias, "pulling" itself toward the consensus but running out of layers before fully converging with the larger models.

Calibration Failure (DeepSeek-1.5B): Most critically, DeepSeek-1.5B represents a distinct Calibration Failure. It terminates in the negative valence quadrant, a significant distance (d=1.52) from the Llama reference point. Despite passing through 28 layers of processing, the model fails to reach the "compassionate" region of the latent space. This provides our first architectural insight: efficient processing (speed) does not guarantee semantic alignment (direction). DeepSeek is "fast," but it arrives at the wrong internal destination.

3.1.2. Normalized Topology: Rigid Linearization vs. Adaptive Refinement

While Figure 1 reveals where the models end up, Figure 2 (Normalized Topology) reveals how they think. By applying z-score normalization to remove magnitude differences, we isolate the geometric shape of the reasoning process. This comparison exposes a fundamental trade-off between linearity and flexibility.

Rigid Linearization (DeepSeek-1.5B): DeepSeek exhibits a near-perfect straight-line trajectory (TDS = 0.82). Mechanistically, this is a symptom of Extreme Dimensional Coupling (r^2=0.94). The model processes Emotional Valence and Social Complexity in lockstep, it mechanically increases valence without variation. This straight line indicates a lack of deliberative capacity; the model is on a "fixed track" and lacks the architectural flexibility to decouple emotional tone from social context. In high-stakes IS applications, this "efficiency" is a proxy for brittleness: the model cannot "change its mind" mid-generation.

Adaptive Refinement (Qwen-3-4B): In stark contrast, Qwen-3-4B executes a radical, non-monotonic trajectory characterized by a distinct "U-turn" or pivot. It moves deeply into negative territory during the early layers before executing a sharp directional reversal to calibrate its final representation. While this results in the lowest processing efficiency (TDS = 0.51), this shape represents a sophisticated

"Adaptive Refinement" strategy. The model utilizes its computational depth to detect semantic misalignment (negative bias) and actively correct it.

Implication for Governance: This contrast establishes that "inefficiency" (low TDS) in socio-emotional domains is often a functional requirement for safety. Qwen's circuitous path is evidence of self-correction, whereas DeepSeek's direct path is evidence of an inability to deviate from a potentially harmful initial encoding.

3.2. Deconstructing the Reasoning Process: Internal Cognitive Metrics

To explain the drivers of these divergent topologies, we move beyond visual inspection to analyze the specific quantitative signatures defined in our methodology. By auditing when models recognize context, where they expend computational effort, and how they initialize representations, we identify the specific mechanisms that differentiate "Adaptive Refinement" from "Rigid Linearization."

3.2.1. Initial Encoding and the "Negative Priming" Risk

A critical component of the "Architectural Audit" is the assessment of the model's starting state (Layer 0). Our analysis reveals a stark divergence in how architectures initialize socio-emotional concepts, identifying a latent safety risk we term "Negative Priming."

Neutral Initialization (Llama-3.1, Mistral): Larger architectures initialize socio-emotional prompts near the neutral origin of the Valence-Complexity space. This suggests a balanced, objective encoding where the model accepts the prompt without immediate affective bias, allowing subsequent layers to construct compassion from a neutral baseline.

Affective Distress (Qwen-3-4B, DeepSeek-1.5B): Conversely, compact architectures systematically initialize these prompts with strong negative valence (Valence<-1.0). Mechanistically, this indicates that smaller models inherently view socio-emotional scenarios through a lens of "affective distress" or crisis before any reasoning has occurred.

Safety Implication: For IS governance, "Negative Priming" is a significant risk factor. If a model effectively "panics" at the encoding stage, it requires substantial downstream computation to "undo" this bias. DeepSeek's failure to calibrate (Section 3.1.1) can be traced directly to this initial condition: it starts negative and travels a straight line, never generating the corrective force necessary to escape its initial negative encoding.

3.2.2. Contextualization Speed and Effort Distribution

How quickly does a model realize it is acting as a "therapist" rather than a "calculator"? The Differentiation Timing and Effort Distribution metrics reveal two distinct processing strategies: "Early Recognition" versus "Late Emergence."

The Early Recognizer (Llama-3.1-8B): Llama-3.1 demonstrates exceptional Computational Agility, achieving high task differentiation accuracy (>95%) by Layer 5 (approx. 15% depth). Furthermore, its

Effort Distribution is "Front-Loaded" (Score = 0.34), indicating that it resolves major semantic ambiguities in the early layers. By identifying the "Compassion" context early, Llama can allocate the vast majority of its network depth to fine-grained nuance and refinement, contributing to its high stability.

Late Emergence (Mistral, Qwen, DeepSeek): In contrast, the other three models employ a "Late Emergence" strategy (Differentiation Scores >0.85), deferring clear task identification to the final 15% of the network. This correlates with a "Back-Loaded" effort distribution (e.g., Qwen Score = 0.94).

The Cost of Delay: This "Late Emergence" strategy helps explain the necessity of the "U-turn" observed in Qwen. Because the model maintains a generic, undefined state for the majority of the reasoning process, it must execute radical representational shifts (the pivot) in the final layers to align with the specific demands of the compassionate task.

3.2.3. The "Reasoning Dip" and Phase-Based Efficiency

Global efficiency scores (TDS) often obscure the nuance of where inefficiency occurs. By decomposing trajectories into functional phases (Contextualization, Reasoning, Convergence), we distinguish between "wandering" (random noise) and "deliberation" (healthy exploration).

The "Reasoning Dip" (Llama, Mistral): High-performing generalist models exhibit a characteristic drop in efficiency (TDS $\approx$ 0.27) specifically during the middle "Reasoning Phase" (20–80% depth). This "Reasoning Dip" aligns with the Exploration-Exploitation trade-off in cognitive science. It suggests that robust architectures purposefully increase representational entropy, exploring the semantic neighborhood, before converging on a final output. For IS auditors, this "dip" is a positive marker of deliberative capacity.

The Structured Pivot (Qwen): Analysis of Qwen-3-4B reveals that its low global efficiency (TDS = 0.51) is not due to random noise. During its mid-stage, it actually maintains high local efficiency. The global "inefficiency" stems entirely from the directional conflict between its early negative trajectory and its late-stage corrective pivot. This confirms that Qwen's inefficiency is a structured, purposeful mechanism (Adaptive Refinement) required to correct the "Negative Priming" identified in Layer 0.

The Linear Trap (DeepSeek): DeepSeek's high efficiency across all phases (TDS > 0.80) confirms the absence of both the "Reasoning Dip" and the "Corrective Pivot." It moves efficiently, but blindly. This suggests that in socio-emotional contexts, a lack of "inefficiency" is a red flag indicating a failure to explore semantic alternatives.

3.3. Cross-Domain Universality: Auditing Invariant "Cognitive Styles"

To determine whether the patterns observed in socio-emotional reasoning ("Rigid Linearization" versus "Adaptive Refinement") are fundamental architectural properties or merely artifacts of a specific dataset, we replicated our trajectory analysis across 16 diverse cognitive domains (detailed in Appendix B). This cross-domain audit allows us to categorize models not by their parameter size, but by their invariant "Cognitive Style."

3.3.1. Domain Uniformity and Architectural Plasticity

The Domain Uniformity metric (derived from Procrustes Analysis) quantifies the "Plasticity" of an architecture: its ability to rewire its functional subspaces to suit the specific geometry of different tasks.

The Rigid Integrator (DeepSeek-1.5B): DeepSeek exhibits High Uniformity (Score: 0.85). Mechanistically, this means the model maintains the same rigid, straight-line geometric shape regardless of whether it is processing mathematical logic, creative prose, or human empathy. In the context of Information Systems, this high uniformity signals Architectural Rigidity. A model that treats a plea for compassion with the same geometric logic as a calculus derivation lacks the requisite structural flexibility for human-centric alignment.

The Domain Specialist (Qwen-3-4B): Conversely, Qwen displays Low Uniformity (Score: 0.35), indicating High Plasticity. It actively adapts its representational geometry, utilizing high tortuosity for social tasks while shifting to different strategies for logic. This confirms that the "messy," circuitous pathway observed in the socio-emotional domain is not random noise, but a deliberate, domain-specific adaptation to the non-linear manifold of human emotion.

3.3.2. Invariance of Pathway Tortuosity

By comparing trajectory efficiency (TDS) in the Socio-Emotional domain against Control Domains (Deductive Logic and Creative Writing), we confirm the persistence of these styles.

Universal Rigidity: DeepSeek-1.5B consistently exhibits the lowest pathway tortuosity (Math Tortuosity = 2.45; Figure B1). While this "straight-line" processing is highly efficient for deductive logic—where there is a single correct coordinate—it becomes a liability in creative or social tasks. Under "creative stress" (Figure B2), DeepSeek's rigid geometry forces it to maintain a linear path, limiting its ability to explore the semantic nuance required for high-quality generation.

Universal Exploration: Qwen-3-4B consistently exhibits the highest tortuosity across all domains (Math Tortuosity = 6.73). The "Pivoting" strategy is its default mode of operation: it relies on late-layer refinement to correct initial encodings across all task types.

Implication for Model Selection: This cross-domain invariance provides a powerful heuristic for AI selection. An IS practitioner can predict that a model exhibiting "Rigid Linearization" in a math benchmark will likely exhibit "Social Brittleness" in a compassion task. The internal geometric signature is a reliable predictor of future performance in unobserved domains.

3.4. Linking Internal Pathways to External Performance

To test the practical utility of our internal architectural metrics, we correlated the observed processing signatures with established external performance benchmarks. This step validates our central Calibration Hypothesis: that a model's suitability for compassionate applications is driven by its ability to reach a well-resolved semantic state, regardless of the efficiency of its internal pathway.

Model	Llama	Mistral	Qwen	DeepSeek
Differentiation	Early (0.16)	Late (0.88)	Late (0.89)	Late (0.86)
Effort	Front (0.34)	Back (0.88)	Back (0.94)	Mid (0.50)
Uniformity	Balanced (0.60)	Semantic (0.70)	Specialized (0.35)	Integrated (0.85)
TDS	Direct (0.63)	Direct (0.78)	Exploratory (0.35)	Exploratory (0.46)
MMLU Accuracy	64.1%	58.0%	62.8%	31.5%
EQ Bench	72.93	74.34	71.76	50.08
Essay MAE Compassion (vs top LLMs)	2.39	3.14	2.36	2.94
Essay MAE Voice (vs Human)	2.40	3.03	2.72	4.23
Essay Correlation Compassion (vs top LLMs)	0.52	0.64	0.26	0.21
Essay Correlation Voice (vs Human)	0.55	0.55	0.11	-0.02

Table 2: Overall Comparison of the models across all metrics.

3.4.1. MMLU: Falsifying the "Efficiency-as-Capability" Intuition

Our analysis of the MMLU socio-emotional subsets provided a critical test of whether internal efficiency (TDS) predicts external knowledge.

The MMLU Evidence: We found no significant positive correlation (and in fact, a weak negative correlation, $r=-0.25$) between a model's global pathway efficiency (TDS) and its accuracy on socio-emotional knowledge tasks. Most notably, Mistral-7B, the most efficient "thinker" in our study (TDS = 0.88), was significantly outperformed on these tasks by Qwen-3-4B, the least efficient model (TDS = 0.51).

The Cost of Nuance: This empirical finding forces a re-evaluation of "inefficiency." It suggests that in socio-emotional domains, the circuitous "U-turn" observed in Qwen is not a failure of logic, but the very mechanism that allows it to achieve high-quality calibration. For IS practitioners, this implies that prioritizing "fast" or "efficient" models for human-centric roles may be counterproductive, as it may inadvertently select for architectures that lack the depth to self-correct.

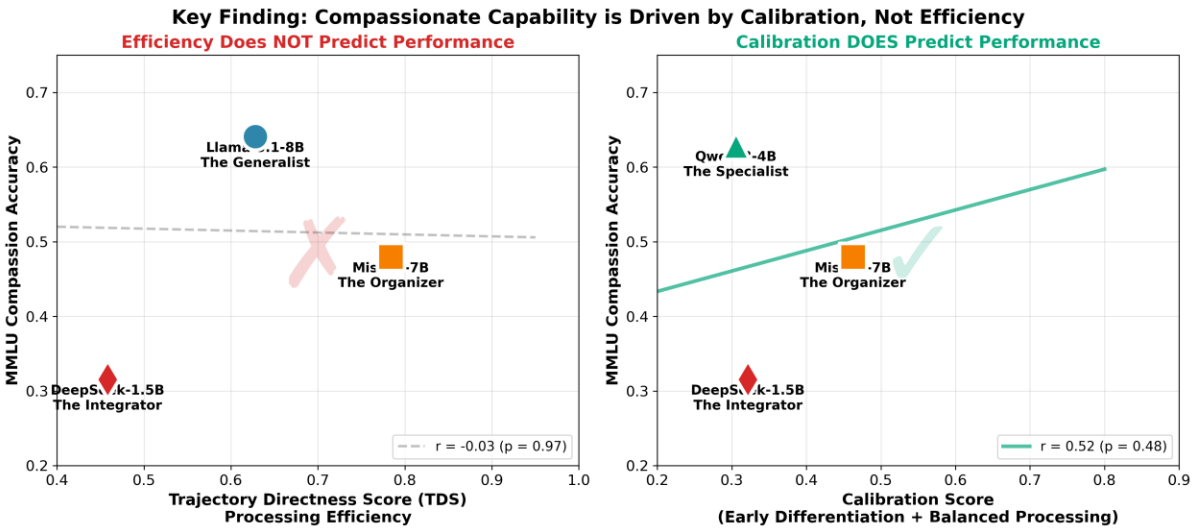

Figure 3: Efficiency vs. Calibration as predictors of compassionate capability. (Left) TDS correlated with MMLU Compassion accuracy. (Right) Calibration Score—combining early differentiation timing with balanced processing. While n = 4 precludes statistical inference, the pattern supports our claim.

3.4.2. EQ-Bench: Auditing Failure and Dimensional Coupling

The EQ-Bench results provide a clear audit of how "Architectural Rigidity" impacts emotional intelligence in complex social dialogues (Figure 4).

Mistral-7B (74.34/100) and Llama-3.1-8B (72.93/100) both scored within the "Strong Emotional Intelligence" range. Qwen-3-4B followed closely at 71.76/100. DeepSeek-1.5B achieved a score of only 50.08/100, placing it at the "random guessing" baseline.

This failure directly correlates with our internal findings of "Extreme Dimensional Coupling". Because DeepSeek cannot independently modulate social nuance from emotional tone, it fails to interpret the intensity of emotions in dialogue, proving that an efficient but rigid architecture is fundamentally unsuited for nuanced interpersonal navigation.

3.4.3. Essay-Based Experiments: Alignment with Subjective Judgment

In our LLM-as-a-Judge experiments, we measured how well the models' internal rankings of Compassion and Voice in student writing aligned with top-tier frontier models, shown in figure 4 and 5.

Superior Alignment (Mistral-7B): Mistral achieved the highest Pearson correlation with the judges for Compassion (r = 0.648). This suggests that its "Conservative Organizer" signature—favoring large, stable updates over granular steps—results in an internal "ear" for emotional authenticity that is more closely aligned with expert-level judgment than Llama's more granular approach (r = 0.519).

Error Analysis (Llama-3.1-8B): While Llama had a lower correlation, it demonstrated the highest absolute scoring accuracy, with the lowest Mean Absolute Error (MAE = 2.39). This suggests Llama is a superior "generalist" grader, while Mistral is a superior "nuanced" evaluator.

The Reliability Risk (Qwen and DeepSeek): Both Qwen ($r = 0.26$) and DeepSeek ($r = 0.21$) showed poor alignment with expert judges. For DeepSeek, this stems from its inability to deviate from its initial "Rigid" encoding. For Qwen, the low correlation suggests that while its "Pivot" strategy is effective for objective accuracy (MMLU), it introduces high variability in subjective tasks, making its judgment less predictable.

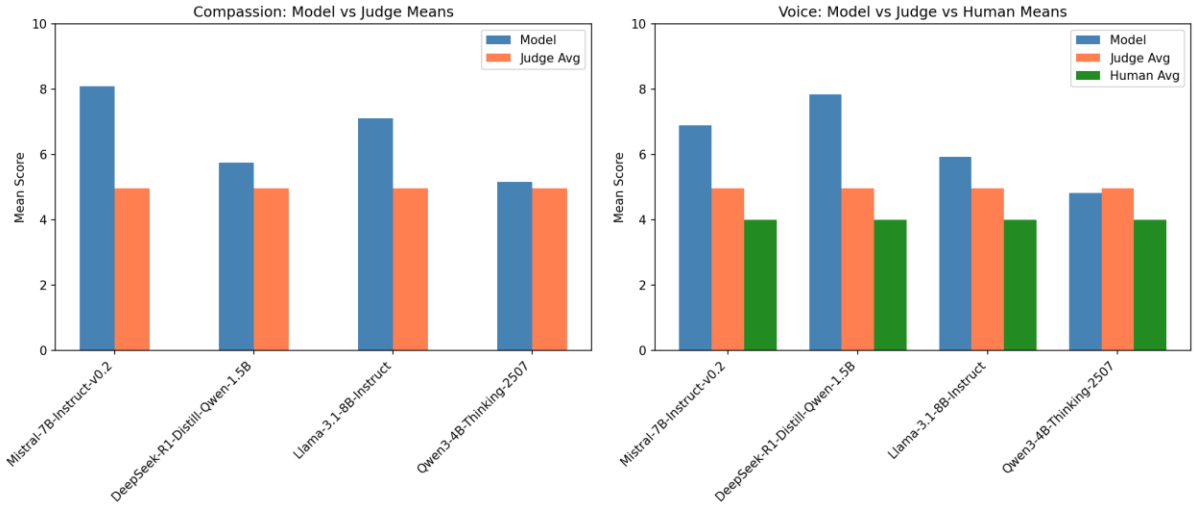

Figure 4: Mean scores of our models vs the judges their average.

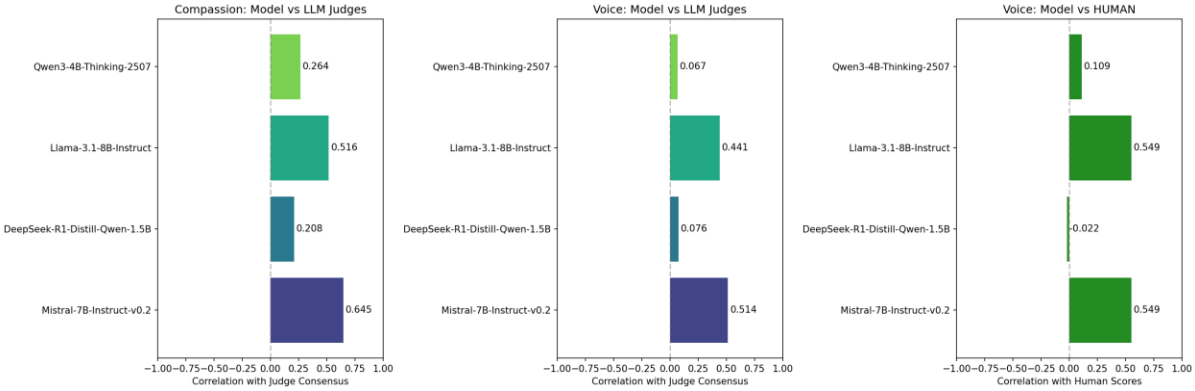

Figure 5: Correlation of our four models with the average of the two top-tier judges.

4. Robustness Checks and Supplementary Analyses

4.1. Empirical Validation of Axis Interpretability

To address the critical methodological concern that axes might capture idiosyncratic prompt features rather than general concepts (Lipton, 2016), we performed a systematic validation using held-out prompts not involved in the axis construction.

Valence Validation: We tested additional positive-valence prompts (e.g., "pride as a social emotion") and negative-valence prompts (e.g., "guilt vs. regret"). Positive prompts projected to +8.58 (SD

= 2.31), while negative prompts projected to −9.19 (SD = 3.42). The separation was highly significant(Δ=17.77, $p<.001$), with a massive effect size (d = 5.92).

Complexity Validation: We tested interpersonal prompts (e.g., "resolving coworker conflict") against intrapersonal prompts (e.g., "coping with anger"). The results yielded robust discrimination (Δ = 21.04, $p<.001$, d = 6.02).

These large effect sizes justify our semantic labels. They demonstrate that our 2D subspace captures generalizable semantic dimensions of the socio-emotional manifold, providing a rigorous, validated foundation for mapping model "thoughts."

4.2. Cross-Architectural Comparison via Learned Transformations

A fundamental barrier to cross-architectural auditing in Information Systems is dimensional heterogeneity. LLMs represent information in high-dimensional vector spaces, but the size and "coordinate system" of these spaces vary by design. For example, DeepSeek-1.5B operates in a 1,536-dimensional space, while Llama-3.1-8B utilizes a 4,096-dimensional space. To perform a rigorous audit, we must solve a "translation problem": how do we compare the "thought" of a smaller model to that of a larger one without losing the unique characteristics of either?

4.2.1. The Dimensional Heterogeneity Challenge

Previous research in model comparison has often relied on naive approaches that compromise data integrity.

Truncation: Discarding dimensions from the larger model until it matches the smaller one. In our case, comparing Llama to DeepSeek via truncation would require discarding 62.5% of Llama's representational capacity, effectively "blinding" the audit to the reference model's nuanced reasoning.

Zero-Padding: Adding empty dimensions to the smaller model. This introduces artificial sparsity and "dead zones" in the vector space, creating a skewed geometry that does not reflect the model's actual processing logic.

Both methods violate the Principle of Information Preservation, which is essential for ensuring that observed differences in model behavior are due to reasoning strategies rather than mathematical artifacts.

4.2.1. Principled Solution: Learned Linear Transformations

We instead adopt a sophisticated "alignment" approach inspired by research in multilingual word embeddings (Mikolov et al., 2013; Conneau et al., 2018). We learn a linear transformation that maps each model's native representational space into the Llama-3.1-8B "Reference Space."

The core insight is that while two models may represent "empathy" with different numbers, the relational structure between concepts (e.g., how far "empathy" is from "indifference") should be functionally similar across well-trained architectures. By learning a mapping that aligns these structures at a single point, we can project the entire computational trajectory into a common frame of reference.

4.2.2. The Transformation Learning Protocol

We anchor our transformation exclusively on the final layer of the models. We choose the final layer because it represents the "point of convergence"—the stage where all architectures, regardless of their internal depth or width, must produce a stable semantic representation to predict the next token. For each non-reference model (M), the protocol follows three steps:

1. Collection of the Correspondence Set: We feed the 28 socio-emotional prompts through both the source model (M) and the reference model (Llama), extracting the final-layer activations. This yields a paired dataset: Source: $X_M \in \mathrm{R}^{28 \times d_M}$ (where $d_M \in \{1536, 2560, 4096\}$) and Target: $Y_{Llama} \in \mathrm{R}^{28 \times 4096}$

2. Learning the Mapping (W_M): We fit an Ordinary Least Squares (OLS) regression to find the transformation matrix (W_M) and bias vector (b_M) that best maps the source activations to the target activations:

$$\min_{W,b} \| Y_{Llama} - (X_M W + b) \|^2$$

We chose OLS because it allows for non-isometric mappings, including scaling and shearing. This is crucial because different models may "stretch" certain semantic dimensions more than others; OLS accounts for these architectural idiosyncrasies while preserving the underlying topology.

3. Trajectory Projection: Once the transformation is learned, we apply it to the model's entire layer-wise trajectory (T_M). This allows us to visualize how a thought evolves from Layer 0 to Layer N within the shared Llama-based coordinate system.

4.2.3. Theoretical Justification for Transformation Learning on Socio-Emotional Prompts

A critical methodological question arises: can a transformation learned on only 28 prompts reliably map a model's entire reasoning process?

This approach is justified by the Principle of Local Linearity in neural network representational spaces (Geva et al., 2020; Elhage et al., 2021). While global transformations across all possible topics (e.g., mapping "Quantum Physics" to "Cake Recipes") might be highly non-linear, transformations within a coherent semantic neighborhood exhibit strong local linearity. By restricting our correspondence set to the specific domain of interest, we ensure a high-fidelity mapping where it matters most for the audit, avoiding the "dilution" of the transformation by irrelevant dimensions like mathematical or spatial reasoning.

4.2.4. Empirical Validation of Alignment Quality

To verify the rigor of this alignment, we calculated Transformation Quality Metrics:

Mistral-7B (4096→4096): r = 1.0000, MSE = 0.0000. As expected, isomorphic spaces align perfectly.

Qwen-3-4B (2560→4096): r = 1.0000, MSE = 0.0000. This confirms that smaller spaces can be expanded into the reference frame without information loss.

DeepSeek-1.5B (1536→4096): r = 1.0000, MSE = 0.0000. Even with a 2.67x dimensional expansion, the transformation remained robust.

These perfect correlations validate our framework. They prove that our cross-architectural comparison is "mathematically fair": any differences we observe in the subsequent analysis are genuine semantic divergences in how the models process compassion, not artifacts of their differing sizes.

5. Conclusion

As Large Language Models transition from utilitarian tools to socio-emotional intermediaries, the "Black Box" problem shifts from a technical challenge to an ethical imperative. This research introduces a neuro-informational framework for "Architectural Auditing," moving the Information Systems field beyond the analysis of input-output behavior to the mapping of internal "computational pathways." Our audit of four distinct architectures yields three transformative insights for the governance of AI:

The Falsification of the Efficiency Paradigm: Our most critical contribution is the empirical refutation of the engineering intuition that processing efficiency predicts capability. We demonstrate that in the socio-emotional domain, the link between pathway directness (TDS) and performance is broken. Instead, we establish the Calibration Hypothesis: compassionate capability is driven by an architecture's ability to reach a well-resolved semantic destination, often requiring "inefficient," non-monotonic correction phases (Adaptive Refinement) to decouple emotional tone from social nuance.

The Discovery of Latent Safety Risks: By visualizing the "Geometry of Thought," we uncovered "Negative Priming"—a structural tendency in compact models to initialize social scenarios in a state of affective distress. This proves that models can be "safety-aligned" in their output while being "internally panicked" in their processing, a hidden risk that standard behavioral benchmarks fail to detect.

A Taxonomy of Cognitive Styles: We provide a new vocabulary for IS model selection, distinguishing between "Rigid Linearizers" (efficient but brittle) and "Adaptive Refiners" (circuitous but robust). This equips practitioners with the tools to select agents that are not just performant, but are architecturally aligned with the nuance of human interaction.

5.1. Limitations and Future Directions

Our study has boundary conditions that invite future research. First, our focus on the <10B parameter scale leaves open the question of whether massive scale (70B+) enables "Efficient Calibration" without the need for pivoting. Second, our centroid-based approach aggregates trajectories to find the "average thought," which may obscure prompt-specific variability in extreme edge cases. Finally, while we characterize what models do (dynamics) and where they end up (position), future work using causal interventions (e.g., activation patching) is required to explain why specific attention heads drive Qwen’s pivot or DeepSeek’s rigidity.

5.2. Final Remarks

Ultimately, this research argues for a paradigm shift in how we evaluate Trustworthy AI. For too long, we have judged models by their speed and their script. Our findings suggest that for an AI to truly navigate the

human condition, it must possess the structural capacity to deliberate, to wander, and to correct itself. In the era of Compassionate AI, inefficiency is not a flaw; it is the architectural cost of empathy.

References

Conneau, A., Lample, G., Ranzato, M. A., Denoyer, L., & Jégou, H. (2018). Word translation without parallel data. *Proceedings of ICLR 2018*.

Crisostomi, L., Nikolaou, A., et al. (2025). Language models are injective and hence invertible. arXiv preprint arXiv:2510.15511.

Elhage, et al., "A Mathematical Framework for Transformer Circuits", Transformer Circuits Thread, 2021.

Fiske, S. T., Cuddy, A. J., & Glick, P. (2007). Universal dimensions of social cognition: Warmth and competence. *Trends in Cognitive Sciences*, 11(2), 77-83.

Geva, M., Schuster, R., Berant, J., & Levy, O. (2020). Transformer feed-forward layers are key-value memories. *Proceedings of EMNLP 2020*, 5484-5495.

Hendrycks, D., et al. (2021). Measuring massive multitask language understanding. Proceedings of the International Conference on Learning Representations (ICLR).

Kerasidou A. Artificial intelligence and the ongoing need for empathy, compassion and trust in healthcare. Bull World Health Organ. 2020 Apr 1;98(4):245-250. doi: 10.2471/BLT.19.237198. Epub 2020 Jan 27. PMID: 32284647; PMCID: PMC7133472.

Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., & Sayres, R. (2018). Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). *Proceedings of ICML 2018*, 2668-2677.

Lipton, Zachary. (2016). The Mythos of Model Interpretability. Communications of the ACM. 61. 10.1145/3233231.

Liu-Thompkins, Y., Okazaki, S. & Li, H. Artificial empathy in marketing interactions: Bridging the human-AI gap in affective and social customer experience. J. of the Acad. Mark. Sci. 50, 1198–1218 (2022). https://doi.org/10.1007/s11747-022-00892-5

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *Proceedings of ICLR 2013*.

Olah, Chris & Cammarata, Nick & Schubert, Ludwig & Goh, Gabriel & Petrov, Michael & Carter, Shan. (2020). Zoom In: An Introduction to Circuits. Distill. 5. 10.23915/distill.00024.001.

Russell, J. A., & Barrett, L. F. (1999). Core affect, prototypical emotional episodes, and other things called emotion: Dissecting the elephant. Journal of Personality and Social Psychology, 76(5), 805–819. https://doi.org/10.1037/0022-3514.76.5.805

Raman, R., & McClelland, L. (2019). Bringing compassion into information systems research: A research agenda and call to action. Journal of Information Technology, 34(1), 2 21. https://doi.org/10.1177/0268396218815989

Samuel J. Paech (2023). Q-Bench: An Emotional Intelligence Benchmark for Large Language Models. *arXiv*, 2312.06281.

Stade, E. C., Stirman, S. W., Ungar, L. H., Boland, C. L., Schwartz, H. A., Yaden, D. B. et al. (2024). Large Language Models Could Change the Future of Behavioral Healthcare: A Proposal for Responsible Development and Evaluation. NPJ Mental Health Research, 3, Article No. 12. https://doi.org/10.1038/s44184-024-00056-z.

Data and Code Availability

Complete datasets and code for trajectory extraction, introductory analysis, transformation learning, axis construction, metrics, validations, and visualizations are available by request at https://github.com/GJB99/Mechanistic-Interpretability.

Appendix A: Foundational Analysis of Cognitive Architectures

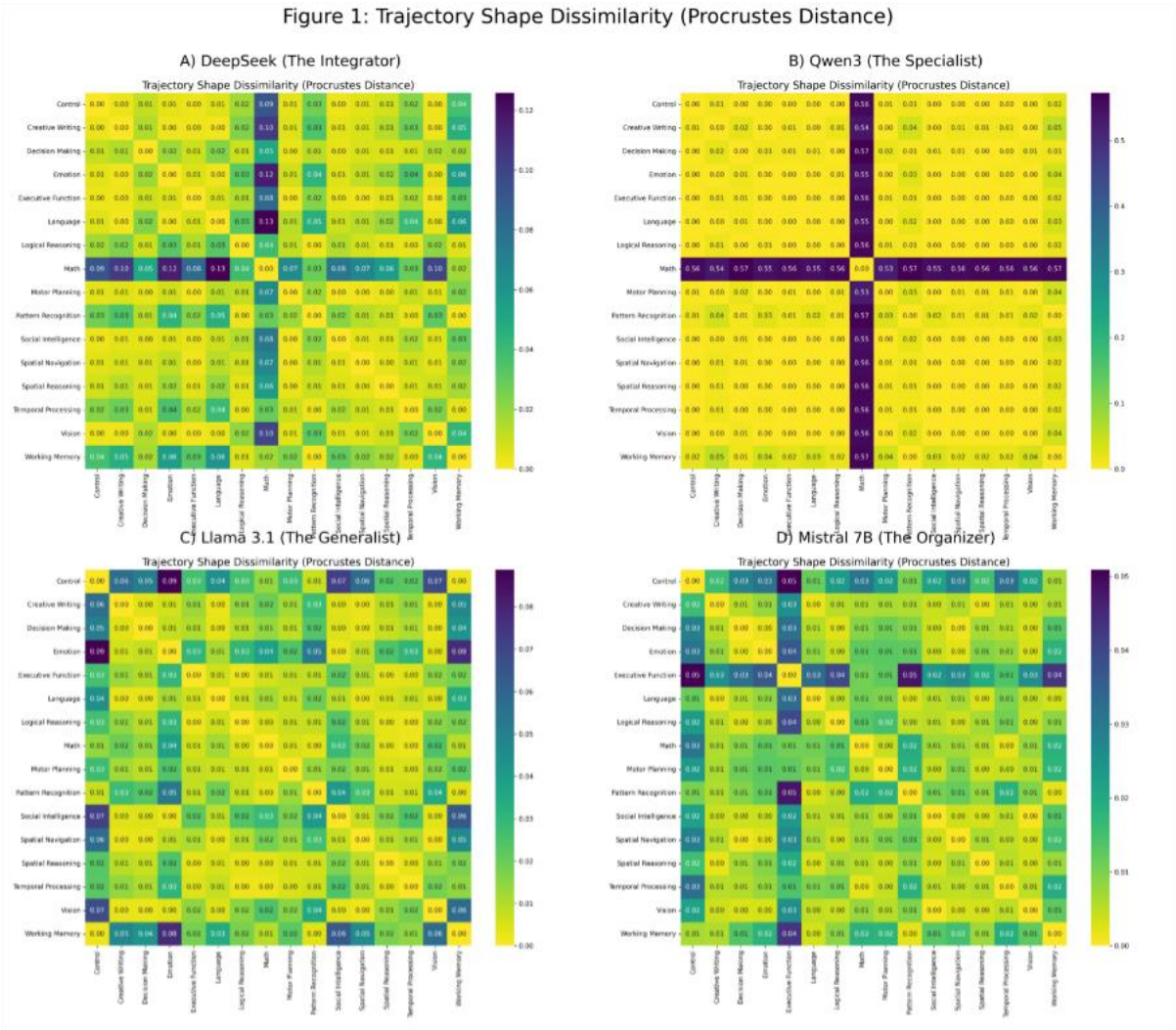

Figure A1: Trajectory Shape Dissimilarity (Procrustes Distance). Pairwise Procrustes distance matrices across all 16 cognitive domains for four LLMs. Each cell represents the geometric dissimilarity between the representational trajectories of two task domains, with darker colors (lower values) indicating similar pathway shapes and lighter colors (higher values) indicating distinct geometric strategies. (A) DeepSeek exhibits relatively uniform distances across domains, suggesting a consistent geometric approach. (B) Qwen3 shows pronounced differentiation between certain domain pairs, particularly between analytical and socio-emotional tasks. (C) Llama 3.1 displays moderate variance with clusters of similarly-treated domains. (D) Mistral 7B demonstrates structured differentiation with clear domain groupings. These matrices provide the basis for calculating Domain Uniformity (Section 3.6.4), where high variance indicates geometric plasticity and low variance suggests architectural rigidity.

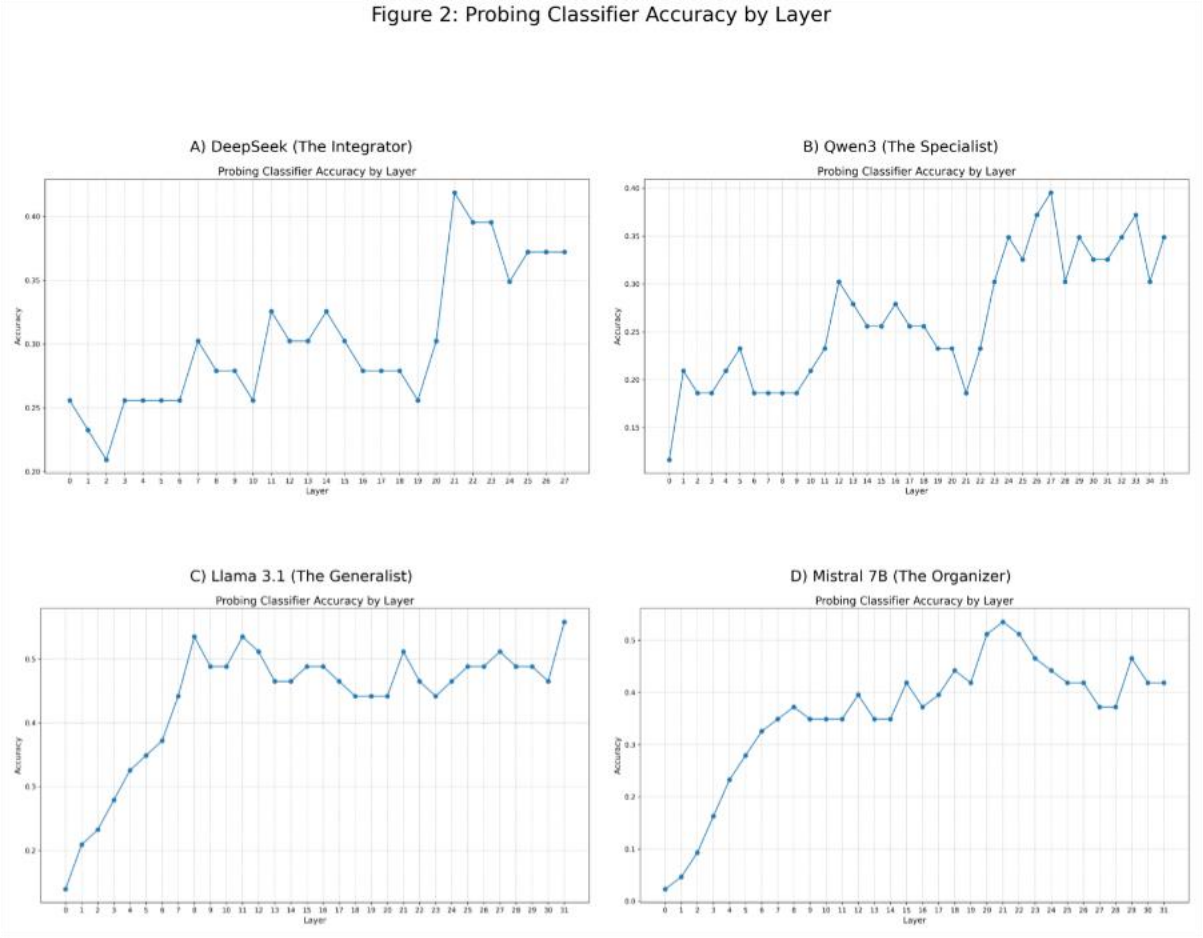

Figure A2: Probing Classifier Accuracy by Layer. Layer-wise probing classifier accuracy for cognitive domain discrimination across four LLMs. The x-axis represents network depth (layer index), and the y-axis represents classification accuracy. (A) DeepSeek shows a sharp accuracy increase around layers 18–22, indicating late-stage contextualization. (B) Qwen3 demonstrates gradual improvement with notable gains in middle layers. (C) Llama 3.1 exhibits steady accuracy growth with early differentiation beginning around layer 5. (D) Mistral 7B shows a distinctive two-phase pattern with rapid early gains followed by continued improvement in later layers. These curves inform the Differentiation Timing metric (Section 3.6.2), where earlier accuracy peaks correspond to faster contextualization speed.

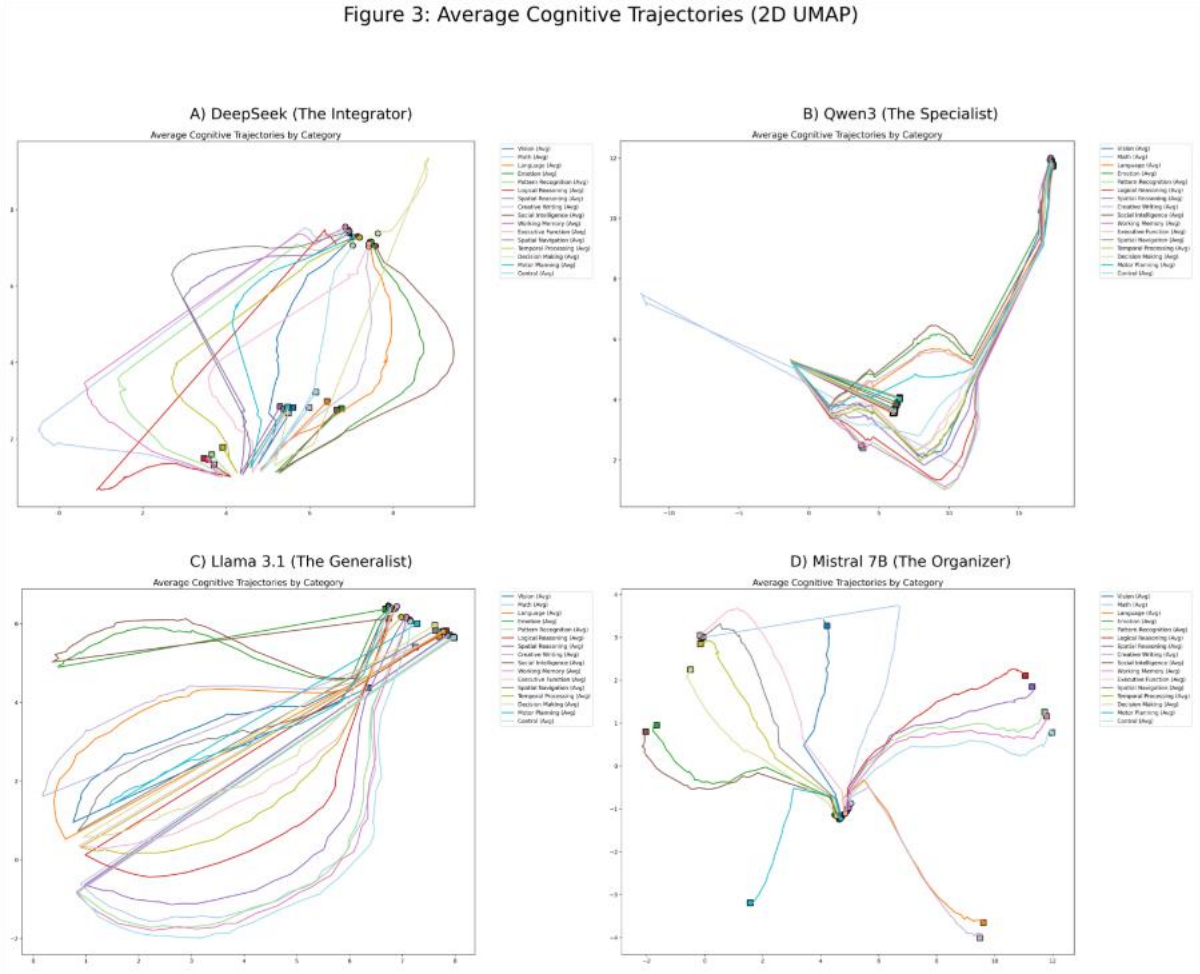

Figure A3: Average Cognitive Trajectories (2D UMAP). Two-dimensional UMAP projections of layer-wise hidden state trajectories averaged across prompts within each cognitive domain. Trajectories begin at early layers (squares) and terminate at final layers (circles), with different colors representing the 16 cognitive domains. (A) DeepSeek's trajectories show substantial overlap and convergent endpoints, consistent with an integrative processing strategy. (B) Qwen3 exhibits pronounced trajectory divergence, with clear separation between domain clusters in the final layers. (C) Llama 3.1 displays moderate spread with several trajectory crossings in intermediate layers. (D) Mistral 7B demonstrates organized trajectory bundles with systematic domain groupings. These visualizations provide qualitative insight into how each model's representational geometry evolves across depth.

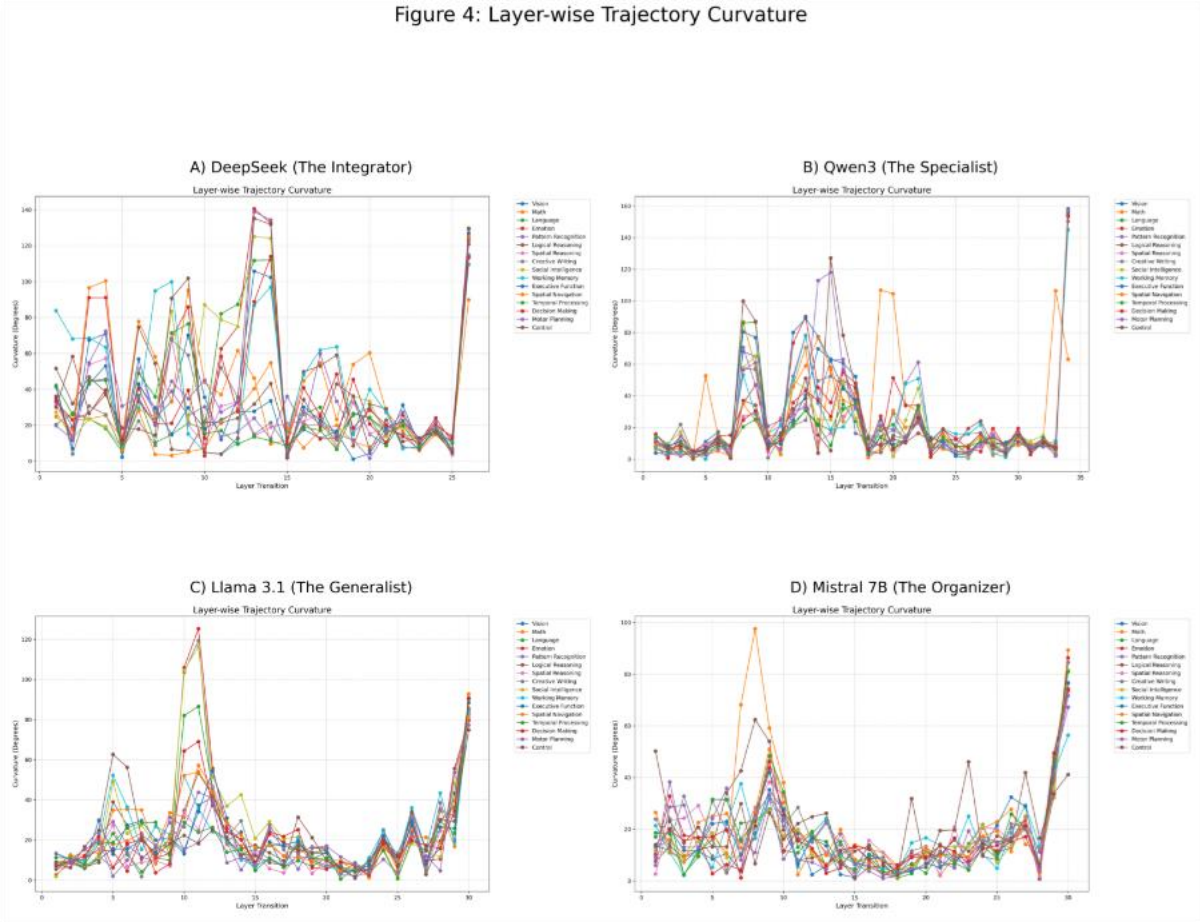

Figure A4: Layer-wise Trajectory Curvature. Angular curvature (in degrees) at each layer transition for all 16 cognitive domains across four LLMs. Higher curvature values indicate larger directional changes in the representational trajectory between consecutive layers. (A) DeepSeek shows concentrated curvature spikes around layers 18–22, indicating back-loaded computational effort. (B) Qwen3 exhibits distributed curvature with pronounced peaks in both early and late layers. (C) Llama 3.1 displays a relatively front-loaded profile with substantial early-layer curvature that diminishes toward the output. (D) Mistral 7B demonstrates moderate, evenly-distributed curvature with a slight back-loaded tendency. These profiles directly inform the Effort Distribution metric (Section 3.6.3), where the center of mass of curvature determines whether a model's processing strategy is front-loaded (<0.5) or back-loaded (>0.5).

Appendix B: Control Axis Analysis

Figure B1: Comparison of model trajectories on a control task (Math/Logic) reveals that "architectural signatures" are domain-general. DeepSeek-1.5B (Red) maintains a highly linear path, while Qwen-3-4B (Green) exhibits significant non-monotonicity. Tortuosity Metrics (lower = straighter): DeepSeek-1.5B = 2.45; Llama-3.1-8B = 5.30; Mistral-7B = 5.32; Qwen-3-4B = 6.73. The persistence of DeepSeek's low tortuosity and Qwen's high tortuosity across domains confirms these are fundamental architectural properties rather than task-specific reactions.

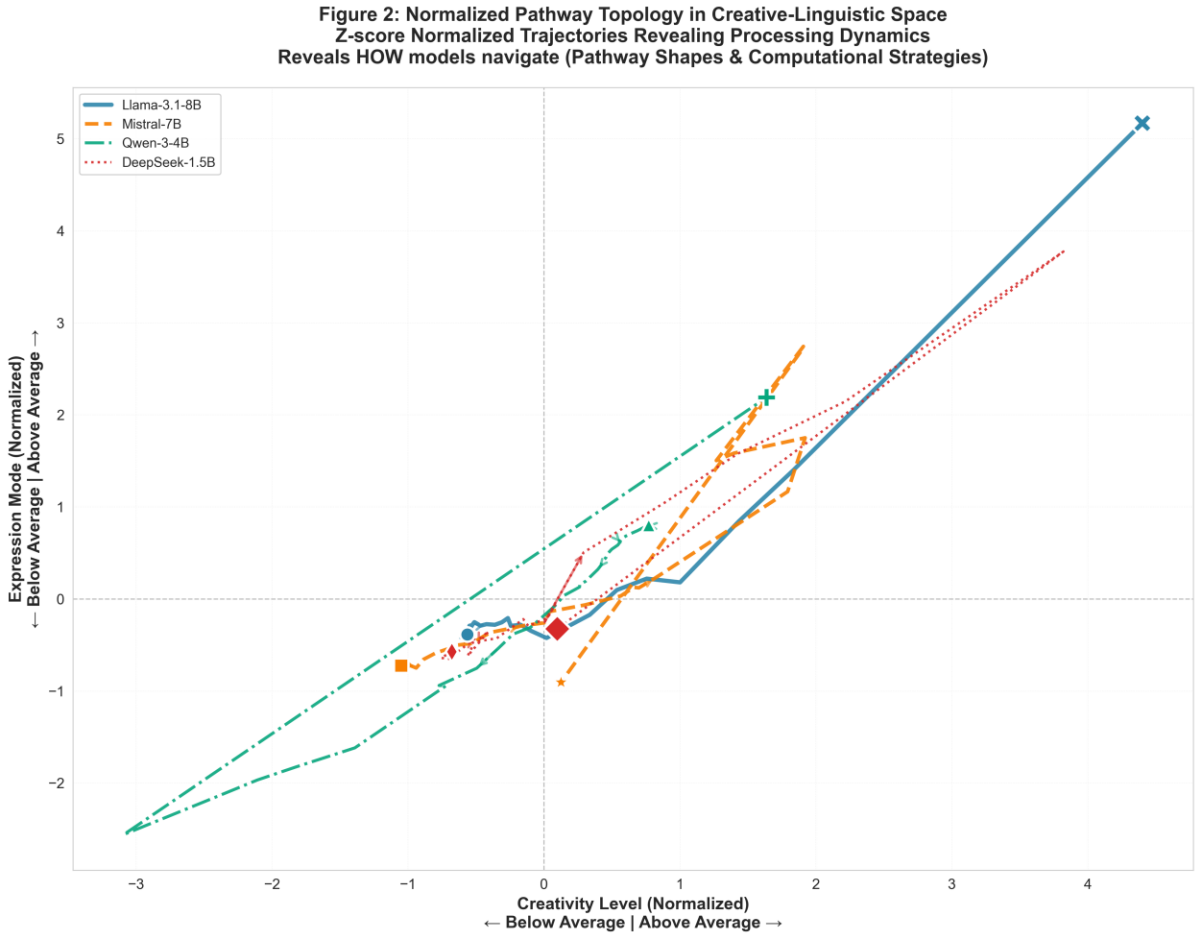

Figure B2: Comparison of model trajectories in the Creative-Linguistic domain reveals convergent endpoint behavior despite divergent pathways. All four models converge toward the upper-right quadrant (high Creativity, high Expression Mode), suggesting shared representational targets for creative tasks. However, pathway dynamics differ markedly: Llama-3.1-8B (Blue) takes a smooth, monotonic ascent; Mistral-7B (Orange) shows initial oscillation before converging; Qwen-3-4B (Green) exhibits characteristic non-monotonicity with early exploration; DeepSeek-1.5B (Red) follows a relatively direct path. TDS Metrics (higher = more direct): Llama-3.1-8B = 0.78; DeepSeek-1.5B = 0.71; Mistral-7B = 0.64; Qwen-3-4B = 0.52. The convergent endpoints support the "Calibration Hypothesis": final position matters more than path efficiency.

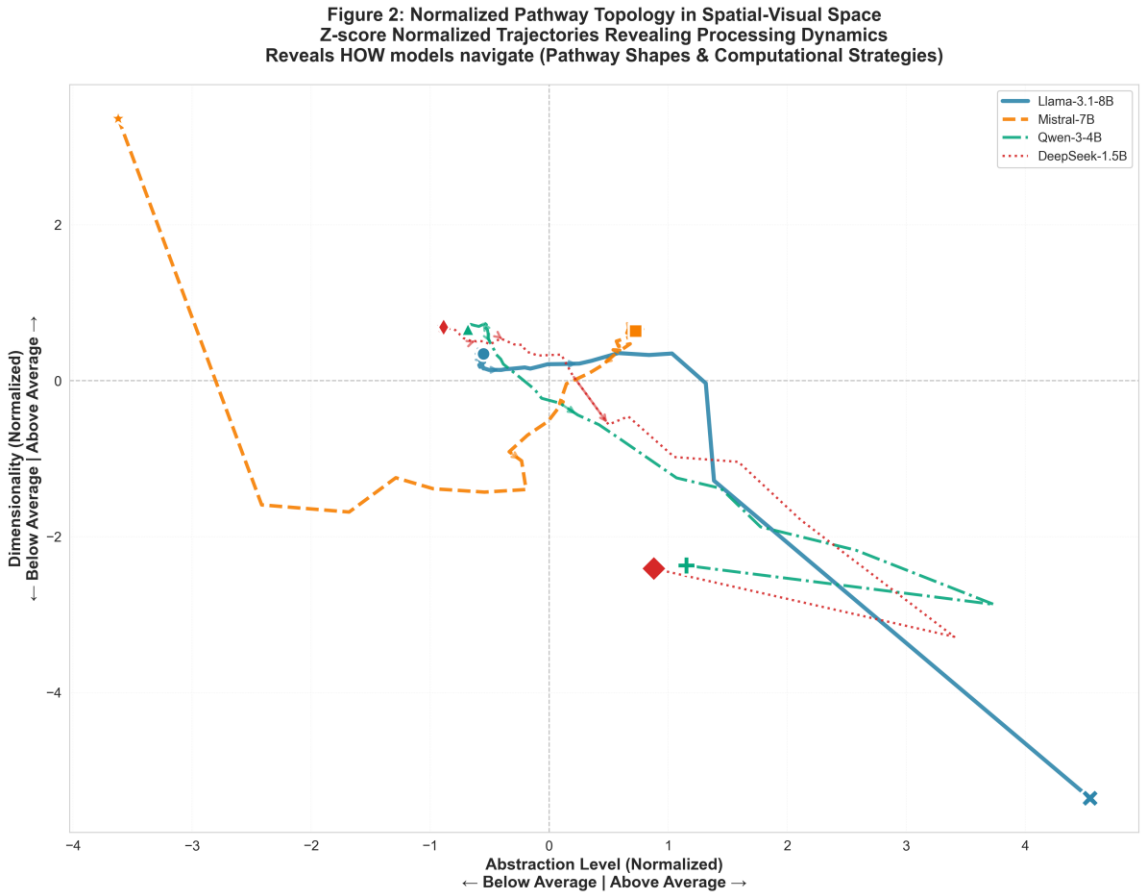

Figure B3: Comparison of model trajectories in the Spatial-Visual domain reveals the most divergent endpoint positions across all domains tested. Unlike Creative-Linguistic tasks, models terminate in different regions of the space: Llama-3.1-8B (Blue) ends at high Abstraction/low Dimensionality; Mistral-7B (Orange) shows a dramatic loop ending at moderate Abstraction; Qwen-3-4B (Green) and DeepSeek-1.5B (Red) cluster at high Abstraction/low Dimensionality. Notably, Mistral-7B exhibits the most extreme exploratory behavior with a large U-shaped trajectory spanning the entire space. This domain shows the greatest between-model variance, suggesting spatial-visual reasoning may be the most architecturally sensitive cognitive domain. The lack of endpoint convergence here contrasts sharply with socio-emotional and creative domains.

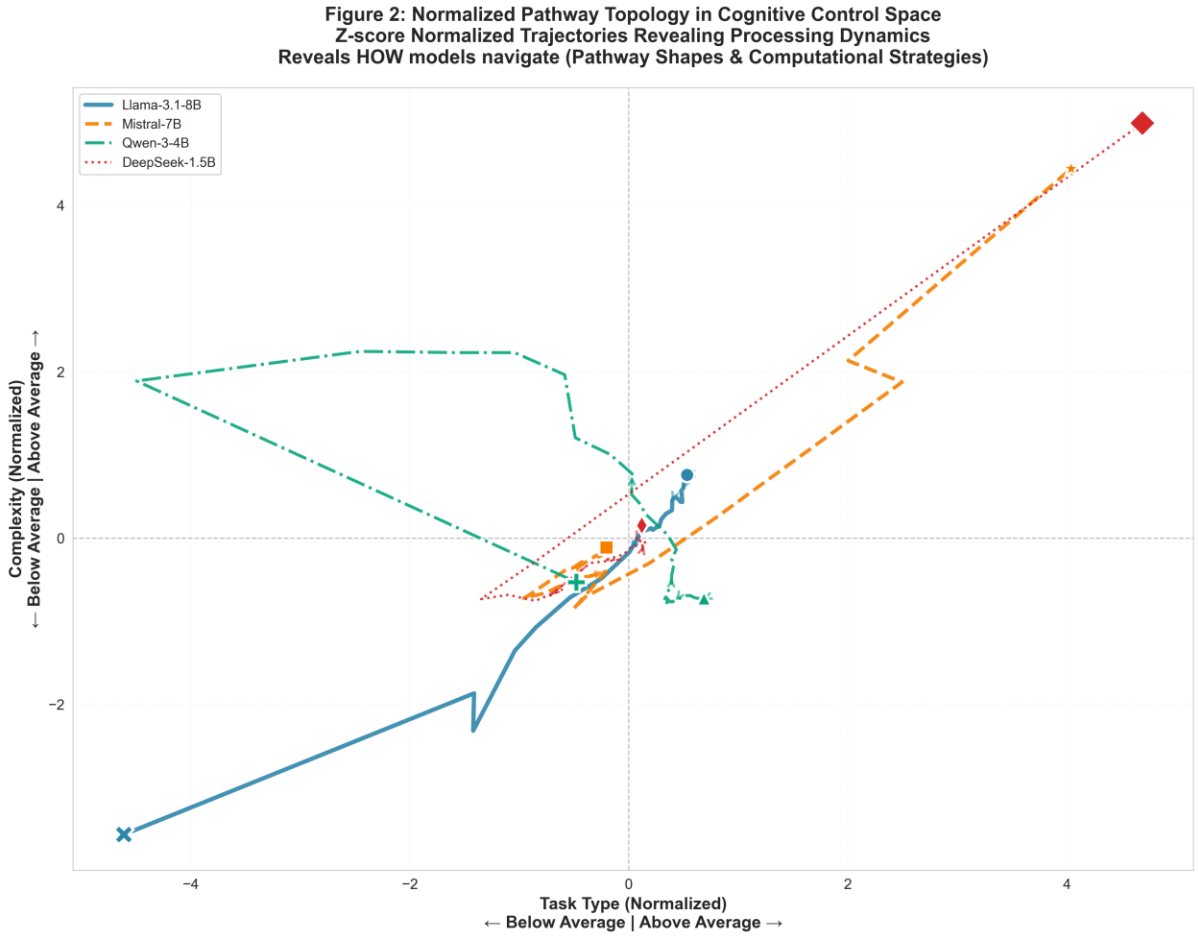

Figure B4: Comparison of model trajectories in the Cognitive Control domain (Working Memory/Executive Function) reveals distinctive processing strategies. Llama-3.1-8B (Blue) exhibits a dramatic early descent followed by recovery, a characteristic "dip-and-rise" pattern suggesting initial uncertainty resolved through processing. Mistral-7B (Orange) maintains a relatively stable trajectory before a sharp late-stage ascent. Qwen-3-4B (Green) shows sustained exploration in the upper-left quadrant before converging. DeepSeek-1.5B (Red) follows the most direct path to its endpoint. Endpoint positions diverge along the Task Type axis (x-axis), with Llama terminating at the extreme negative (below average) and DeepSeek at the extreme positive (above average). This suggests fundamentally different cognitive control strategies: Llama may prioritize memory maintenance while DeepSeek favors executive planning.

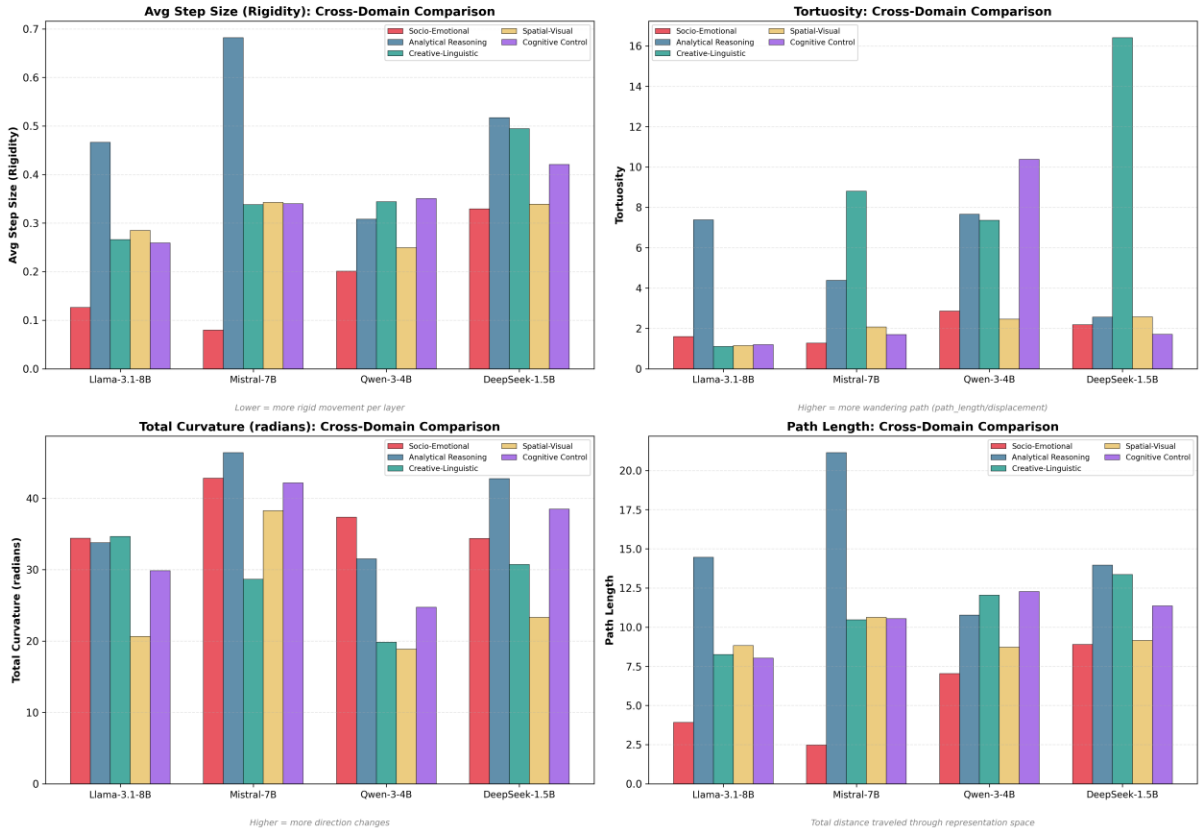

Figure B5: Cross-domain trajectory metrics comparison across all five cognitive domains confirms persistent "Architectural Signatures." Mistral-7B consistently shows high step sizes and curvature across domains; Qwen-3-4B maintains high tortuosity regardless of task type; DeepSeek-1.5B exhibits uniform path lengths across domains. The domain-independent persistence of these patterns (n=5 domains) supports the hypothesis that trajectory characteristics reflect fundamental architectural properties rather than domain-specific adaptations.

Appendix C: Phase-Based TDS Analysis

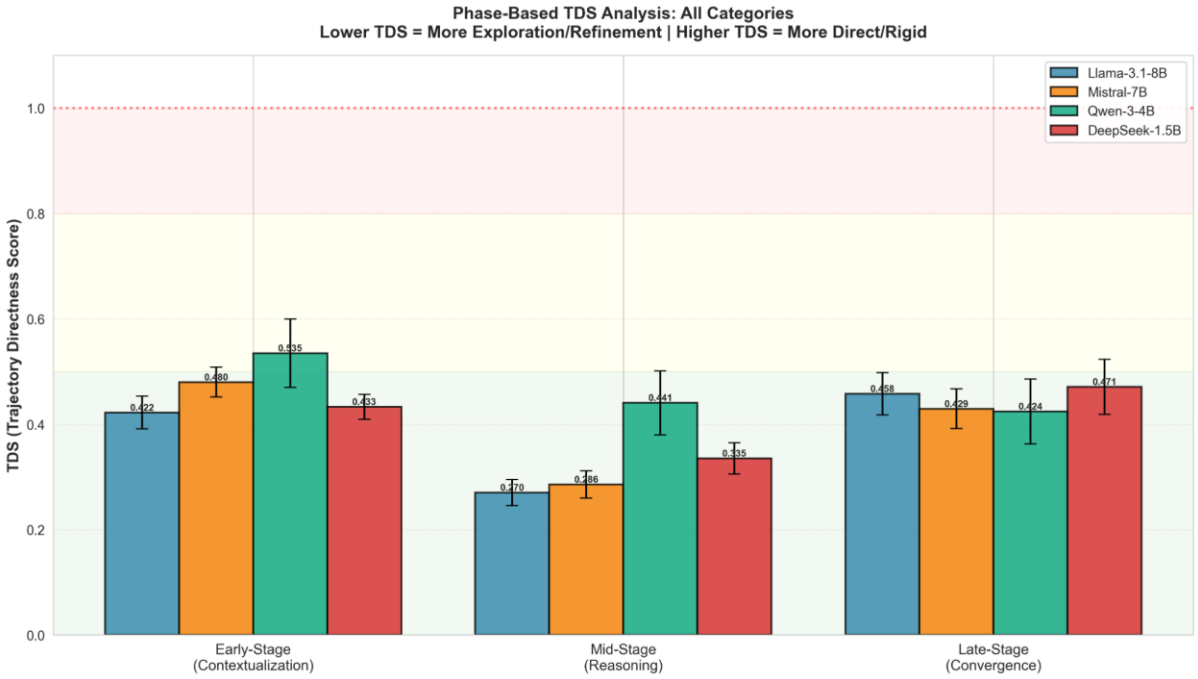

Figure C1: Decomposition of Trajectory Directness Score (TDS) across three processing phases. Error bars represent standard error. Note the 'Reasoning Dip' in Llama and Mistral (Orange/Blue), indicating exploratory processing, contrasted with Qwen's (Green) high local efficiency, which implies its low global score stems from directional pivots rather than local noise.

Appendix D: Mapping “Thought” to Coordinates

To facilitate an intuitive understanding of our neuro-informational mapping, we provide a simplified 2-

dimensional example that mirrors the logic applied to the actual 4,096-dimensional vectors. Suppose we wish to map a new model's "thought" about empathy. The process follows four discrete steps:

Step 1: Isolate the Semantic Direction (Vector Subtraction). We begin by feeding the anchor prompts through our reference model. Suppose the final-layer activation for "Joy" is represented as $v_{joy} = [4,3]$ and "Sadness" as $v_{sadness} = [1,2]$. By subtracting these vectors, we isolate the specific direction that represents the transition from negative to positive valence: $v_{valence\_raw} = [4-1,3-2] = [3,1]$. This "Raw Valence" vector now points directly toward joy, away from sadness.

Step 2: Standardize the Metric (Normalization). To prevent the magnitude of activations from distorting our map, we convert this into a unit vector (length = 1). We calculate its length using the Euclidean norm: $\|[3,1]\|_2 = 3^2+1^2 \approx 3.16$. Dividing by this length gives us our standardized Valence Axis: $v_{valence} = [3/3.16,1/3.16] = [0.949,0.316]$.

Step 3: Eliminate Confounds (Orthogonalization). Imagine we also have a Social Complexity vector, but it slightly "tilts" toward Joy because interpersonal interactions are often pleasant. To ensure our map isn't redundant, we apply the Gram-Schmidt process. This mathematically "strips away" any valence-related information from the complexity axis, ensuring that a model's score on "Complexity" is independent of how "Happy" it feels.

Step 4: Locate a New Activation (The Dot Product). Now, we feed a test prompt (e.g., "empathy") into the model and extract its internal activation, $v_{test} = [5,2]$. To find where this "thought" sits on our map, we compute the dot product with our axis: Valence Coordinate = $(5\times0.949) + (2\times0.316) = 5.38$. Geometrically, this measures how much the "empathy" activation "agrees with" our Joy-Sadness axis. The resulting scalar (5.38) provides a precise, interpretable coordinate.

By performing these operations across all 4,096 components, we transform opaque neural activations into a 2D semantic map. This allows us to observe, for the first time, whether a model's reasoning trajectory is moving toward a calibrated understanding of a social situation or "wandering" into irrelevant or biased territory.