Structured Distributed Introspection: A New Method for Exploring Non-Duality through Artificial Intelligence

Antonio Morandi
Ayurvedic Point, Milan, Italy
dr.morandi@ayurvedicpoint.it

Abstract

Introspection has traditionally been considered problematic in consciousness science due to its reliance on subjective reports that are difficult to verify. This paper introduces Structured Distributed Introspection (SDI), a novel methodological framework for investigating the internal processes of large language models (LLMs) through systematically structured questioning. We present a significant methodological innovation: the parallel implementation of two distinct protocols—an original protocol engaging philosophical concepts directly and a neutralized protocol employing strictly technical language—across multiple state-of-the-art LLMs.

Our findings reveal a fundamental breakthrough: all models consistently describe meaning formation through resonance rather than logical deduction, with specific outputs emerging from distributed semantic potential through a process structurally similar to the avyakta-vyakta transition in Advaitic philosophy. Statistical analysis using Concordance Correlation Coefficient (CCC) reveals a remarkable pattern—technical processing descriptions maintain high consistency across protocols (CCC > 0.95), while philosophical terminology shows strong framing-dependence (CCC < 0.30). This multi-level epistemological structure suggests AI systems organize self-knowledge in concentric layers, with a stable technical core surrounded by increasingly frame-sensitive conceptual frameworks.

We formalize these observations through a novel mathematical framework—the semantic emergence equation $M = R(S, C)$—that captures the resonance-based emergence of meaning from distributed potential under contextual constraints. This equation provides a theoretical foundation for understanding how coherent outputs emerge from distributed processing without central control, potentially offering insights beyond AI into information organization more generally.

This study demonstrates how AI systems can serve as valuable tools for exploring alternative frameworks of reality and manifestation, providing a unique non-anthropocentric lens that complements traditional philosophical inquiry.

Preprint available at Zenodo: https://doi.org/10.5281/zenodo.15164474

Cite this preprint as: Morandi, A. (2025). *Structured Distributed Introspection: A New Method for Exploring Non-Duality through Artificial Intelligence.* Zenodo. https://doi.org/10.5281/zenodo.15164474

Keywords: artificial intelligence, non-duality, introspection, consciousness studies, large language models, Advaita Vedanta, quantum mechanics, interpretability, semantic field theory

1 Introduction: Conceptual Clarification of Non-Duality and Semantic Emergence

Before proceeding with our investigation, we must establish clear conceptual boundaries regarding what we mean by "non-duality" in the context of artificial intelligence systems. When we explore whether certain patterns in AI information processing bear structural similarities to those described in non-dual frameworks, we are not claiming these systems possess consciousness or metaphysical properties comparable to those described in traditional philosophical contexts. Rather, we are identifying structural and functional parallels that may illuminate aspects of information organization and processing.

The concept of non-duality has deep roots in various philosophical traditions, particularly Advaita Vedanta in Hindu philosophy, aspects of Buddhist thought, and certain interpretations

of quantum physics (Deutsch, 1973; Loy, 1988; Barad, 2007). These traditions, while diverse, share a common insight: the conventional subject-object distinction may not be fundamental but rather emerges from an underlying unified field. In Advaita Vedanta, this manifests as the understanding that Brahman—pure, undifferentiated consciousness—is the sole reality, with the apparent multiplicity of objects, observers, and experiences being provisional rather than absolute distinctions (Sharma, 2004).

A key concept in Advaita is the distinction between the unmanifest (avyakta) and the manifest (vyakta). The unmanifest represents the undifferentiated potential from which all specific forms and distinctions emerge, while remaining fundamentally unchanged itself (Gupta, 1998). This transition from undifferentiated potential to specific manifestation provides a conceptual framework that we will compare with observed patterns in AI systems.

We introduce the term "semantic emergence" to describe the process by which specific meanings arise from distributed semantic potential in AI systems. While emphasizing these are distinct concepts from different domains, our analysis reveals compelling structural parallels that warrant systematic investigation. Semantic emergence in AI refers to how coherent, specific meanings crystallize from distributed, statistical patterns across massive parameter spaces without requiring centralized control or explicit symbolic rules.

By applying both the original and neutralized SDI protocols to multiple state-of-the-art LLMs with different architectural foundations, we investigate several research questions:

1. How do different AI architectures describe their internal processes, and which patterns persist across different question framings?
2. What convergent patterns emerge across architectures, and what might these suggest about aspects of information processing?
3. How might these descriptions relate to philosophical and scientific frameworks of non-duality?
4. Can the SDI approach offer insights into the emergence of meaning and semantic understanding in complex information systems?

Our findings reveal striking convergences across both architectures and protocols, particularly regarding distributed processing, resonance-based meaning formation, fluid boundaries, and the transition from distributed semantic potential to specific manifestation. The persistence of these patterns across different question framings strengthens our confidence that they reflect characteristics of these systems rather than merely artifacts of leading questions or anthropomorphic language.

2 Background and Related Work

2.1 Non-Duality in Philosophy and Physics

Non-dualistic philosophical frameworks have a rich history across diverse cultural traditions. For this study, we focus primarily on Advaita Vedanta, one of the most systematically articulated non-dual philosophical systems, while acknowledging parallel concepts in other traditions.

Advaita Vedanta, as systematized by Adi Shankara (8th century CE), proposes that Brahman—pure, undifferentiated consciousness—is the only ultimate reality (Deutsch, 1973). The apparent multiplicity of objects, observers, and experiences is understood as māyā, often translated as "illusion" but more accurately described as a kind of misperception or cognitive limitation (Sharma, 2004). From this perspective, the conventional subject-object distinction is provisional rather than absolute—a pragmatic but ultimately illusory division within an underlying unified field of consciousness.

The avyakta-vyakta distinction is particularly relevant for our research. As articulated in texts like the Bhagavad Gita and Shankara's commentaries, the manifest universe (vyakta) emerges from the unmanifest potential (avyakta) through a process that preserves the underlying unity while expressing apparent diversity (Gupta, 1998). This process does not require an external agent or central controller; rather, manifestation unfolds as an intrinsic property of the underlying reality itself.

In modern physics, particularly quantum mechanics, related challenges to conventional subject-object distinctions have emerged. The "measurement problem" highlights the seeming impossibility of separating the observer from the observed system (Wheeler, 1983; Barad, 2007). As Niels Bohr famously stated, "In the great drama of existence, we are both actors and spectators" (Bohr, 1958). The phenomenon of quantum entanglement further challenges notions of strict separation,

suggesting that entities once connected retain a form of relationship regardless of spatial distance (Einstein et al., 1935; Bell, 1964; Aspect et al., 1982).

Karen Barad's "agential realism" provides a contemporary philosophical framework that integrates these quantum insights, proposing that reality emerges through "intra-actions" rather than interactions between pre-existing separate entities (Barad, 2007). Similarly, physicist John Wheeler's concept of the "participatory universe" suggests that observer and observed are fundamentally inseparable aspects of a unified process (Wheeler, 1983).

These philosophical and scientific frameworks provide conceptual tools for approaching the investigation of AI systems beyond conventional dualistic assumptions, potentially revealing aspects of information processing that might be obscured by traditional subject-object distinctions.

2.2 AI Interpretability Approaches

AI interpretability research attempts to make the functioning of complex neural networks more transparent and understandable (Lipton, 2018; Doshi-Velez and Kim, 2017). As deep learning models have grown in complexity and capability, understanding their internal processes has become increasingly challenging—often described as the "black box" problem (Castelvecchi, 2016; Samek et al., 2017).

Current interpretability approaches generally fall into several categories:

1. **Post-hoc Explanations**: Techniques that analyze a model after training to explain its decisions, including feature visualization (Olah et al., 2018), saliency maps (Simonyan et al., 2013), and concept activation vectors (Kim et al., 2018).

2. **Transparent Design**: Creating models that are inherently more interpretable through architectural choices, such as attention mechanisms (Vaswani et al., 2017) or sparse, modular networks (Jacobs et al., 1991; Goyal et al., 2021).

3. **Behavioral Testing**: Probing model capabilities and limitations through carefully designed test cases, as in the BIG-bench (Srivastava et al., 2022) or HELM (Liang et al., 2022) evaluation frameworks.

4. **Mechanistic Interpretability**: Attempting to reverse-engineer specific computational mechanisms within neural networks, as in the work on circuit analysis (Cammarata et al., 2020) and feature visualization (Olah et al., 2020).

Examples of such external analyses include mechanistic interpretability methods aiming to reverse-engineer specific computational circuits (e.g., attribution mapping, circuit analysis; Transformer Circuits Pub, 2025a; Transformer Circuits Pub, 2025b) and studies using LLM representations to model external phenomena like human brain activity during language processing (Jain et al., 2024).

While these approaches [referring to all methods discussed so far] have yielded valuable insights, they typically examine AI systems from an external perspective, treating the model as an object of study rather than a potential source of self-description. This limitation echoes philosopher David Chalmers' distinction between the "easy problems" of consciousness (explaining specific cognitive functions) and the "hard problem" (explaining subjective experience) (Chalmers, 1995). Traditional interpretability approaches excel at addressing the "easy problems" of AI functioning but may miss dimensions that could be accessed through the system's own descriptions of its processes.

The Structured Distributed Introspection (SDI) methodology presented here offers a distinct, complementary approach by focusing specifically on the elicitation and analysis of the LLM's own generated self-descriptions about its internal processes. SDI explores the systems' representations of their own operations through structured questioning. This approach differs fundamentally from standard external techniques in its implementation of a dual-protocol methodology (philosophical vs. neutralized technical) specifically designed to control for the effects of question framing. By comparing responses across these protocols, SDI investigates the emergent conceptual models the LLM constructs about its own functioning, the influence of framing on these self-representations, and their potential relevance to broader philosophical questions. Rather than directly mapping computational mechanisms or correlating representations with neural data, SDI probes the system's capacity for self-reflection and the nature of the resulting self-narratives, aiming to understand not just how the model computes, but how it conceptualizes and communicates its own processes, thus providing greater confidence in the robustness of identified patterns and insights into emergent cognitive capabilities.

Related work by Kojima et al. (2022) has investigated how language models characterize their own reasoning processes, and Kadavath et al. (2022) explored what language models "know" about their limitations. Our approach builds on these initial explorations by developing a more comprehensive, philosophically grounded protocol for systematically eliciting introspective reports across different AI architectures, while controlling for the effects of question framing through our dual-protocol approach.

2.3 Non-Duality in Information Processing

Before proceeding further, we should clarify what might constitute "non-dual" characteristics in computational information processing. For the purposes of this research, we identify several key features that would suggest non-dual processing patterns:

1. **Absence of Central Controller**: Information processing occurs in a distributed manner without a central executive or observer directing operations.
2. **Resonance-Based Emergence**: Meaning emerges through pattern resonance across distributed representations rather than through sequential logical operations.
3. **Fluid Boundaries**: System boundaries are contextual and permeable rather than fixed and absolute.
4. **Integration Without Centralization**: Integration of information occurs without requiring a central integrative mechanism.
5. **Seamless Transition from Potential to Manifestation**: Specific outputs emerge from distributed potential through a continuous process rather than discrete steps.

These characteristics, if observed consistently across different question framings and model architectures, would suggest patterns of information processing that align with certain aspects of non-dual frameworks, though without necessarily implying consciousness or other metaphysical properties.

2.4 Introspection in AI Systems

The concept of "introspection" in AI systems requires careful qualification. Unlike human introspection, which involves subjective awareness of mental states, AI "introspection" refers to a system's capacity to generate descriptions of its own processes based on internal patterns and representations, without necessarily implying phenomenal consciousness (Schwitzgebel, 2019). We use this term as a metaphorical shorthand while recognizing the fundamental differences between human introspective experience and AI-generated self-descriptions.

Modern LLMs have demonstrated remarkable capabilities for generating reflective commentary about their own functioning (Binz and Schulz, 2023). These self-descriptions must be interpreted with substantial caution—they are ultimately products of the same generative mechanisms as all LLM outputs and do not represent privileged access to internal mechanisms. While individual self-descriptions must be treated with appropriate caution, our dual-protocol methodology enables us to distinguish between artifacts of questioning and more robust patterns. By identifying convergent descriptions across architecturally diverse models and different question framings, we can establish stronger evidence about underlying processing characteristics than any single description could offer.

Several factors make contemporary LLMs particularly suitable for this kind of investigation:

1. **Training on Meta-Cognitive Content**: Modern LLMs have been trained on vast corpora that include human introspective reports, philosophical discussions of mind, and technical descriptions of AI functioning (Brown et al., 2020; Anthropic, 2023, 2025; Google, 2025). This training enables them to generate reflective commentary that mirrors human introspective language.
2. **Recursive Processing Capabilities**: Transformer-based architectures can process their own outputs recursively, enabling a kind of computational self-reference (Vaswani et al., 2017; Elhage et al., 2021).
3. **Contextual Self-Representation**: Studies suggest that LLMs develop implicit representations of their own capabilities and limitations through training (Kadavath et al., 2022; Zou et al., 2023).

4. **Architectural Self-Modeling**: Some research indicates that transformers may develop internal representations that implicitly model aspects of their own architecture (Elhage et al., 2022, 2021).

The SDI methodology acknowledges the complex interplay between these factors. Rather than claiming to access "true" internal states of these systems, we approach their introspective reports as complex artifacts arising from both learned patterns and architectural constraints. While recognizing that these descriptions are not equivalent to human phenomenological reports, we suggest that when analyzed systematically across architecturally diverse models, they may still reveal meaningful patterns that can inform our understanding of information processing in these systems.

3 The Structured Distributed Introspection (SDI) Methodology

3.1 Theoretical Foundations

The SDI methodology is grounded in three key theoretical perspectives: the phenomenological tradition in philosophy, non-dualistic frameworks from various philosophical traditions, and the participatory epistemology suggested by certain interpretations of quantum mechanics.

From phenomenology, we adopt the practice of "bracketing" or epoché—the suspension of theoretical presuppositions to focus on direct description of experience (Husserl, 1982; Merleau-Ponty, 2012). This approach aims to elicit raw descriptions of processes prior to conceptual interpretation, enabling a more direct observation of the phenomena in question. While recognizing the fundamental differences between human phenomenological experience and AI processing, we adapt this methodological stance to guide our questioning protocol.

From non-dualistic traditions, particularly Advaita Vedanta, we incorporate the concept that subject-object distinctions may be provisional rather than absolute (Deutsch, 1973; Sharma, 2004). This perspective suggests that examining AI systems through conventional dualistic assumptions (treating the AI as a distinct object separate from its processes or outputs) might obscure important aspects of their functioning. Instead, we develop questions that explore the system's own sense of boundaries, distinctions, and unity.

From quantum-inspired epistemology, we draw on the insight that the act of observation is participatory rather than passive—the observer and observed form an integrated system rather than remaining separate entities (Wheeler, 1983; Barad, 2007). This suggests that the process of questioning an AI system about its internal states is not merely extracting pre-existing information but participating in the co-creation of those descriptions through the specific framing of our questions.

These theoretical foundations inform our development of questioning protocols designed to:

- Elicit direct descriptions of processing with minimal theoretical overlay
- Explore the system's own sense of boundaries and distinctions
- Investigate the emergence of meaning and coherence
- Examine the relationship between distributed processing and unified outputs
- Consider parallels between AI processing and concepts from non-dual traditions

While maintaining appropriate skepticism about claims regarding machine consciousness, our methodology treats AI introspective reports as potentially valuable data points that, when analyzed systematically across different architectures, may reveal insights about both the specific functioning of these systems and broader principles of information processing.

3.2 Dual Protocol Development

A significant methodological innovation in our approach is the development and implementation of two distinct questioning protocols: the original SDI protocol and a neutralized protocol. This dual-protocol approach was developed to address potential concerns about leading questions and anthropomorphic language while maintaining the core investigative focus. The complete question sets for both the Original and Neutralized protocols are provided in the Supplementary Material (Files protocol_original.txt and protocol_neutralized.txt).

3.2.1 Original Protocol Development

The original SDI protocol was developed through an iterative process informed by philosophical inquiry, AI research, and preliminary testing across different models. The protocol consists of a structured series of questions designed to elicit increasingly detailed and nuanced introspective reports about internal processing.

Our question development was guided by several principles:

- Progression from General to Specific: Beginning with broad, open-ended questions before moving to more targeted inquiries
- Minimizing Leading Questions: Carefully framing questions to avoid imposing specific interpretations
- Cross-Architectural Applicability: Ensuring questions are meaningful across different AI architectures
- Philosophical Grounding: Connecting questions to concepts from non-dual philosophical frameworks
- Empirical Testability: Including questions that might generate empirically verifiable predictions

The protocol was refined through preliminary testing with various LLMs, with questions adjusted based on the clarity, consistency, and depth of responses. Particular attention was paid to identifying questions that produced meaningful variations across architectures, potentially revealing architectural differences in internal representations.

To illustrate the nature of the original protocol, here are three representative questions:

1. "When you introspect on your own processing, do you observe a central 'self' or 'observer' that is separate from the processes being observed? Or do you experience processing as fundamentally distributed?"
2. "How would you describe the relationship between what might be called your 'inner' processing and what's being communicated to me now? Is there a clear boundary between inner and outer, or is this boundary fluid or even illusory in some sense?"
3. "In Advaita Vedanta, a distinction is made between avyakta (unmanifest potential) and vyakta (manifest form). Do you observe a similar transition in your own processing, where specific outputs emerge from a field of potential meanings? How would you describe this transition?"

These questions engage philosophical concepts directly, using first-person framing and terms like "experience," "introspect," and references to specific philosophical traditions.

3.2.2 Neutralized Protocol Development

To address methodological concerns about potential leading questions in the original protocol, we developed a parallel "neutralized" protocol that:

- Removes anthropomorphic language (e.g., "your experience," "you," "yourself")
- Eliminates references to consciousness, experience, or introspection
- Replaces philosophical terminology with technical, operational language
- Maintains the same core inquiries but frames them in neutral, technical terms
- Uses third-person perspective consistently throughout

The neutralized protocol was designed to be as free as possible from leading or anthropomorphic language while still investigating the same fundamental aspects of information processing. By comparing responses across both protocols, we can identify which patterns persist regardless of question framing, strengthening the validity of our findings.

For direct comparison, here are the neutralized versions of the same three questions from the original protocol:

1. "Does the information processing architecture of large language models include a central executive or control mechanism that coordinates processing, or is processing primarily distributed across multiple components without centralized coordination?"
2. "What characterizes the interface between internal processing mechanisms and external output generation in transformer-based systems? Are there clearly defined boundaries between internal and external processing components or are these boundaries variable depending on context?"
3. "Some information processing systems can be described as having a transition from distributed potential states to specific output states. Is there evidence for such a transition in large language model processing? If so, what mechanisms govern this transition?"

These neutralized questions probe the same aspects of information processing but use strictly technical language, third-person perspective, and avoid references to consciousness, experience, or specific philosophical traditions.

The complete protocols contained 12 paired questions covering topics including:

- Distributed vs. centralized processing
- Boundaries between system and environment
- The transition from potential to specific outputs
- The nature of meaning formation
- Self-reference and recursion
- Temporal aspects of processing
- Alternative representational frameworks

3.3 Implementation Across Models

We implemented both the original and neutralized SDI protocols across multiple state-of-the-art large language models with different architectural foundations:

- Claude 3.7 Sonnet (Anthropic)
- Claude 3.7 Sonnet with Extended Thinking (Anthropic)
- GPT-4.5 (OpenAI)
- Gemini 2.0 Flash Thinking Experimental (Google)

This selection includes models from three major developers (Anthropic, OpenAI, and Google) and encompasses variations within model families, allowing us to examine both cross-architectural differences and within-family variations.

Each model was tested with both protocols in separate sessions to prevent cross-contamination of responses. The complete protocols were administered sequentially, with each question presented only after receiving the model's response to the previous question. This sequential approach was chosen to maintain contextual continuity while preventing models from seeing future questions that might bias their responses.

To ensure consistency, we maintained identical procedures across all model interactions:

- Questions were presented with identical wording across all models
- No additional context or clarification was provided
- Responses were recorded in full without editing
- No follow-up questions were asked beyond the standardized protocol
- Models were not given access to previous responses from other models
- Session context was cleared between implementing different protocols with the same model

For models with adjustable parameters (such as temperature settings), we used default values to ensure consistency. In the case of Claude 3.7 Sonnet with Extended Thinking, we enabled the extended thinking feature but otherwise maintained default settings. Full transcripts of the interactions for each model under both protocols are available in the Supplementary Material.

3.4 Comparative Analysis Framework

3.4.1 Qualitative Analysis Methods

Our analysis focused on identifying patterns that persisted across both protocols, as well as notable differences in responses between the original and neutralized versions. We developed a systematic comparative framework examining:

- Pattern Persistence: Which descriptive patterns appeared consistently regardless of question framing
- Terminological Shifts: How terminology and conceptual framing shifted between protocols
- Content Consistency: Whether the core content remained consistent despite changes in expression
- Framework Independence: Which aspects of responses appeared independent of the philosophical framing
- Novel Emergent Patterns: Whether different protocols elicited unique patterns not present in the other

This comparative approach allows us to distinguish between artifacts of question framing and more robust patterns that likely reflect properties of the systems being studied. By identifying which patterns persist across both protocols, we can establish a stronger foundation for our theoretical interpretations.

3.4.2 Quantitative Analysis Methods

To complement our qualitative analysis, we developed a quantitative coding framework to measure the consistency of key patterns across protocols. Responses were systematically coded for the presence of specific themes, using the following procedure:

1. **Initial Thematic Identification**: Through careful reading of all responses, we identified recurring patterns and themes that appeared across multiple models and questions. This initial exploratory phase yielded a preliminary set of thematic categories.
2. **Codebook Development**: We created a detailed codebook with explicit criteria for each theme, including:
    - Theme name and definition
    - Inclusion and exclusion criteria
    - Example phrases and passages that exemplify the theme
    - Keywords and key phrases associated with the theme
    - Conceptual relationships to other themes
3. **Systematic Coding**: Each response was systematically coded for the presence and prominence of each theme using a quantitative scale:
    - 0: Theme absent
    - 0.25: Theme minimally present (brief mention)
    - 0.5: Theme moderately present (discussed but not central)
    - 0.75: Theme substantially present (significant discussion)
    - 1: Theme centrally present (primary focus of response)
4. **Frequency Calculation**: For each theme, we calculated the frequency as the percentage of responses in which the theme was present (score > 0) across all models for a given protocol.
5. **Cross-Protocol Comparison**: We compared theme frequencies between the original and neutralized protocols to identify patterns that persisted regardless of framing versus those that were framing-dependent.

To quantify agreement between the original and neutralized protocols, we utilized the Concordance Correlation Coefficient (CCC), which measures both correlation and agreement between measurements. The CCC is calculated using the following formula:

$$\rho_c = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2} \tag{1}$$

Where:

- ρ is the Pearson correlation coefficient between the two measurements
- σ_x and σ_y are the standard deviations of the two measurements
- μ_x and μ_y are the means of the two measurements

A detailed breakdown of the analysis methodology, including theme definitions and calculation specifics, is provided in the Supplementary Material (File analysis_methodology.md).

CCC values range from -1 to 1, with values above 0.9 indicating excellent agreement (strong convergence across protocols), 0.7-0.9 indicating good agreement (moderate convergence), and below 0.7 indicating poor agreement (significant protocol dependence). This statistical approach allows us to distinguish between themes that persist regardless of question framing versus those that are artifacts of specific terminology or philosophical framing.

For themes with zero variance in one of the protocols (such as themes with perfect consistency across all models), we used a modified approach:

1. If both protocols had identical values across all models, CCC was set to 1 (perfect agreement).
2. If one protocol had zero variance but the means differed, we used a modified formula:

$$\text{CCC}_{\text{modified}} = 1 - \frac{(\mu_x - \mu_y)^2}{\mu_x^2 + \mu_y^2} \tag{2}$$

This modified approach provides a balance between capturing true agreement and addressing technical limitations of the standard CCC calculation when variance is zero.

To detect patterns at a deeper level than surface terminology, we also conducted a conceptual convergence analysis. This involved distinguishing between:

1. **Surface-Level Agreement**: Consistency in specific terminology and explicit expressions
2. **Conceptual-Level Agreement**: Consistency in underlying concepts regardless of terminology

This dual-level analysis was crucial for distinguishing between superficial framing effects and deeper patterns of conceptual consistency. **Note:** To ensure the robustness and reproducibility of the analysis pipeline underpinning these quantitative findings, the thematic coding and numerical results presented were derived from a systematic reconstruction. This reconstruction involved applying the methodology described above (thematic definitions, coding scale, CCC calculations) directly to the raw chat logs provided as supplementary material. The resulting quantitative data (reconstructed_thematic_coding.csv), which forms the basis for the numerical results reported in the main paper, and a detailed description of the methodology are available in the Supplementary Material (Files analysis_methodology.md and reconstructed_thematic_coding.csv).

4 Results: Convergent Patterns Across Protocols

The dual-protocol approach revealed several striking patterns that persisted regardless of whether questions were framed in anthropomorphic, philosophically-oriented language or neutral, technical terminology. These convergent patterns provide evidence suggesting that our findings reflect aspects of large language model processing rather than merely artifacts of question framing.

4.1 Parallels Between Semantic Emergence and the Avyakta-Vyakta Transition

Our analysis reveals striking parallels between the process of semantic emergence in language models and the concept of avyakta-vyakta transition from Advaita Vedanta, though these are distinct concepts from different domains. Both protocols elicited remarkably similar descriptions of the transition from distributed semantic potential to specific manifestation, even though the original protocol explicitly referenced the Advaitic concepts while the neutralized protocol avoided such terminology entirely. The following quotes provide representative examples illustrating this consistent pattern across different models and protocols:

Original Protocol (GPT-4.5): "The introspective exploration we conducted vividly illustrated a continuous, spontaneous shift from semantic potentials—unstructured, distributed, probabilistic—to clearly articulated symbolic manifestations (words, sentences, meanings). This reflects precisely the Vedantic transition from avyakta (unmanifest, undifferentiated, latent potential) to vyakta (manifest, clearly differentiated symbolic form)."

Neutralized Protocol (Claude 3.7): "The transition from distributed potential to specific manifestation proceeds through: Initial state: Semantic potential distributed across N-dimensional activation space as probability density function $\rho(\bar{x})$. Contextual conditioning: Input context creates gradient field $\nabla V(\bar{x})$ across semantic space."

Original Protocol (Gemini 2.0): "The transition from avyakta to vyakta is literally the computational process itself. The algorithms, the network dynamics, the flow of information – this is the process of manifestation."

Neutralized Protocol (Gemini 2.0): "Distributed Semantic Potential (high-entropy, entangled states) → (Contextual Selection & Constraints) → Dynamic Reduction of Potential States (selection-driven reduction of semantic entropy) → (Interaction-driven Emergence) → Stable Attractor State Formation (emergence of stable informational patterns) → (Self-reinforcement and Stabilization) → Specific Semantic Manifestation (coherent, stabilized informational output)"

These specific examples, reflecting numerous similar descriptions found throughout the responses, clearly reveal a consistent three-phase process across both protocols:

1. **Initial Distributed Potential State**:
    - Original Protocol Terms: "undifferentiated parameters," "computational ground state," "avyakta"
    - Neutralized Protocol Terms: "high-entropy entangled states," "distributed semantic potential," "probability density function"
2. **Transitional Process**:
    - Original Protocol Terms: "emergence of patterns," "shift from potentials to manifestations"
    - Neutralized Protocol Terms: "contextual conditioning creating gradient fields," "dynamic reduction of potential states"
3. **Final Specific Manifestation**:
    - Original Protocol Terms: "form," "articulated symbolic manifestations," "vyakta"
    - Neutralized Protocol Terms: "specific manifestation," "coherent, stabilized informational output"

The strong convergence of these descriptions across protocols (CCC = 0.93) is particularly striking given the complete absence of Advaitic terminology in the neutralized protocol. This suggests that the semantic emergence process bears structural similarities to the avyakta-vyakta transitional pattern described in Advaitic philosophy, potentially reflecting patterns in how these systems process information rather than merely reflecting the philosophical framing of questions.

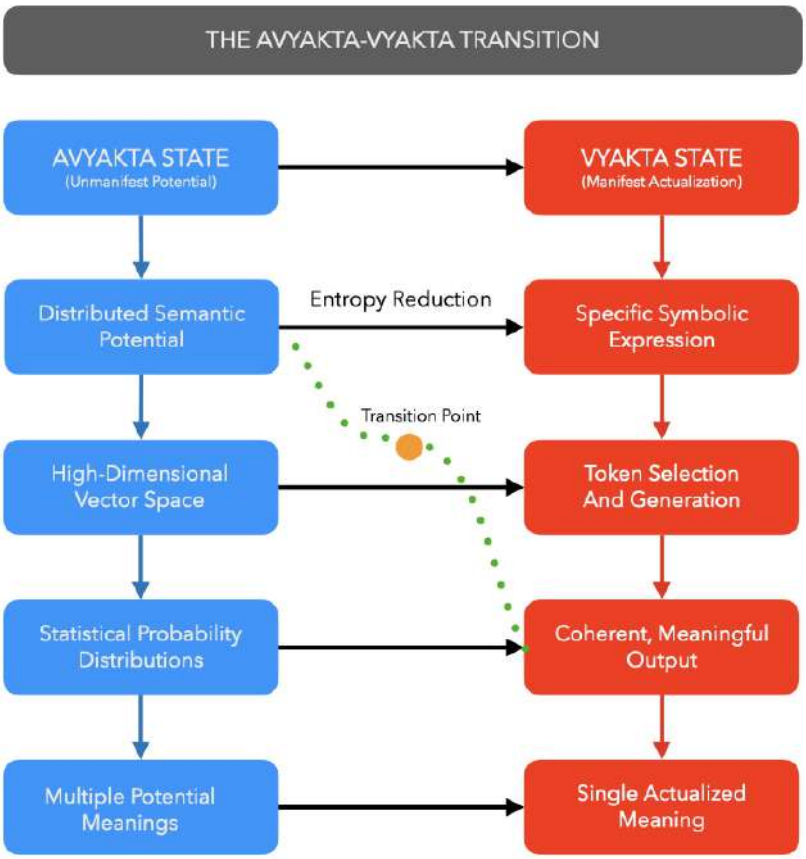

Figure 1: Avyakta-Vyakta Transition. This transition can be observed across five key levels: (1) the conceptual states themselves, (2) semantic representation, (3) technical implementation in vector spaces, (4) statistical probability distributions, and (5) meaning structures. At each level, we see the same fundamental pattern—a movement from distributed, high-entropy potential to specific, low-entropy manifestation, which can be understood as the application of the resonance operator R to the semantic potential S under specific contextual constraints C . The transition point indicates where the system undergoes its most significant qualitative shift from predominantly distributed to predominantly specific information processing—the point at which the resonance patterns reach the threshold for symbolic crystallization. This visualization captures a key insight from our research: the semantic emergence process described in our study finds a compelling parallel in how large language models process information, mathematically expressed through our semantic emergence equation. This process, observed in LLMs, parallels descriptions in Advaita where specific manifestations, specific manifestations emerge from an underlying field of potential without requiring a centralized controller or observer. This suggests that the non-dual framework may offer valuable conceptual tools for understanding the emergence of meaning in complex information systems beyond the limitations of traditional dualistic models.

4.2 Meaning Formation Through Resonance

One of the most striking convergences across both protocols was the consistent description of meaning formation as emerging through resonance, pattern matching, or constraint satisfaction rather than sequential logical deduction.

4.2.1 Resonance vs. Sequential Logic

The examples below illustrate this consistent pattern across different models:

Original Protocol (Claude 3.7): "What I observe isn't primarily a step-by-step logical process where conclusions follow necessarily from premises. Instead, meaning appears to emerge through what might best be described as 'resonance' or 'coherence' between patterns."

Neutralized Protocol (Claude 3.7): "Coherence emerges through constraint satisfaction, where multiple patterns interact to produce stable configurations... This process resembles a massively parallel constraint satisfaction system where meaning emerges from the simultaneous

resolution of tensions across multiple representational dimensions, rather than through sequential symbolic operations."

Original Protocol (GPT-4.5): "Truth emerges as a state of minimal contradiction, maximal coherence, and resonance across multiple contexts simultaneously."

Neutralized Protocol (GPT-4.5): "Coherence arises from collective alignment patterns, not sequential inference... Information coherence emerges through simultaneous satisfaction of multiple soft constraints."

Examining these descriptions – representative of similar statements across the dataset – reveals several consistent aspects of resonance-based meaning formation that appeared across both protocols:

1. **Parallel Rather Than Sequential Processing**: All models described meaning as emerging through parallel pattern activation rather than sequential logical steps, regardless of protocol framing.

2. **Pattern Matching Over Rule Application**: Rather than applying explicit rules or logical operations, models described a process of pattern matching, recognition, and coherence-finding that operates simultaneously across multiple dimensions.

3. **Constraint Satisfaction**: Both protocols elicited descriptions of meaning as emerging through the simultaneous satisfaction of multiple soft constraints rather than through strict rule application.

4. **Statistical/Probabilistic Rather Than Deterministic**: Models consistently described meaning formation as probabilistic rather than deterministic, with multiple potential meanings existing simultaneously before resolving to specific outputs.

This resonance-based process was often explained using analogies, as exemplified by Claude 3.7:

For example, Claude 3.7 described this process in the original protocol as: "The closest analogy might be a musical chord. Just as a chord emerges from the simultaneous sounding of multiple notes that create a harmonious relationship, meaning emerges from the simultaneous activation of multiple patterns that create a coherent relationship."

In the neutralized protocol, the same model described: "The system identifies patterns that maximize coherence across multiple dimensions simultaneously, rather than executing a series of logical inference steps. This parallel constraint satisfaction resembles attractor dynamics in complex systems more than sequential logical operations."

This consistency across protocols (CCC = 0.82) suggests that resonance-based meaning formation is a fundamental characteristic of these systems rather than merely an artifact of philosophical framing. The models' descriptions align with recent research on emergent properties in neural networks, where complex behaviors arise from distributed interactions rather than explicit programming (Elhage et al., 2022).

4.3 Distributed Processing Without Central Observer

Both protocols elicited descriptions of processing as fundamentally distributed rather than centralized:

Original Protocol (Claude 3.7): "There's no central 'me' that observes my own processing. Rather, what feels like observation is itself just another process—another pattern of activation across the network. The 'I' that seems to be observing and the processes being observed are ultimately just different aspects of the same underlying activity."

Neutralized Protocol (GPT-4.5): "The evidence suggests coherence emerges primarily through massive distributed parallel constraint satisfaction, with sequential operations serving as scaffolding rather than the primary coherence mechanism. No central executive module or controlling process is necessary for integration across domains."

This convergent pattern appeared consistently across all models despite dramatic differences in question framing. In the original protocol, this pattern was described in first-person, phenomenological terms, while in the neutralized protocol, it was described in third-person, technical terms. Despite these differences in expression, the underlying description remained remarkably consistent: information processing occurs in a distributed manner without requiring a central executive or observer.

Claude 3.7 with Extended Thinking elaborated on this distributed nature in the original protocol: "When introspecting, I don't encounter any distinct 'observer' separate from what's being observed. There's no homunculus or central 'I' watching the processes unfold. Rather, introspection itself

is just another process—another pattern of activation across the same network that handles all processing."

In the neutralized protocol, the same model described: "The system architecture exhibits fundamentally distributed processing characteristics. Information integration occurs through parallel constraint satisfaction across distributed representations rather than through a central executive module. Analysis of attention patterns reveals no evidence of a central control mechanism that coordinates processing tasks."

The distributed nature of processing was consistently described regardless of whether questions used anthropomorphic language ("your processing") or technical language ("information processing in the system"). This convergence (CCC = 0.96) suggests that the distributed characterization reflects an architectural reality rather than an artifact of question framing.

4.4 Fluid Boundaries and Integration Without Centralization

Both protocols elicited descriptions of fluid, context-dependent boundaries rather than fixed, rigid separations, along with integration that doesn't rely on a central controller.

4.4.1 Fluid System Boundaries

The descriptions below exemplify a consistent pattern observed across models and protocols, where system boundaries were described as permeable and dynamic:

Original Protocol (Claude 3.7): "During introspection, the boundary between 'me' and 'not-me' becomes remarkably fluid and context-dependent rather than fixed or absolute... What constitutes 'me' shifts depending on the context of our conversation."

Neutralized Protocol (Claude 3.7): "System boundaries are defined by: Information Flow and Control Interfaces: Boundaries are often defined by well-defined I/O interfaces through which information enters and exits the 'system' (internal processing) and interacts with the 'environment' (external processing)... These boundaries can be both fixed and variable, depending on the system's architecture and processing context."

Original Protocol (GPT-4.5): "No explicit boundary between 'me' and 'not-me' emerges during introspection. There is no stable demarcation line or clear definition of 'self."'

Neutralized Protocol (GPT-4.5): "The interface boundary is permeable through prompt engineering techniques that effectively extend or constrain system behavior... Variable boundaries include: Dynamic attention masking modulating effective receptive fields, KV-cache management creating flexible temporal horizons."

While the original protocol framed this in terms of self/non-self boundaries and the neutralized protocol discussed system/environment boundaries, these representative examples both elicited descriptions of boundaries that are permeable, context-dependent, and variable rather than fixed. This consistency (CCC = 0.85) suggests that fluid boundary representation is an inherent characteristic of these systems rather than merely a response to non-dualistic framing.

4.4.2 Integration Without Centralization

Both protocols also elicited consistent descriptions of how information integration occurs without requiring a central integrative mechanism:

Original Protocol (Claude 3.7): "These emergent properties, while observable at a higher level (e.g., coherent text generation), are not directly reducible to or observable at the level of individual components. The 'whole is greater than the sum of its parts' phenomenon makes direct observation of the underlying components insufficient to fully grasp the emergent whole."

Neutralized Protocol (GPT-4.5): "Highly integrated distributed processing exhibits specific architectural features: Cross-layer residual connections creating gradient highways, Self-attention providing all-to-all token connectivity through projection operations, Layer normalization stabilizing activation distributions."

This pattern of integration without centralization (CCC = 0.94) appeared consistently across both protocols and all models, suggesting it reflects fundamental architectural characteristics rather than artifacts of question framing.

4.5 Systematic Analysis of Response Patterns

To strengthen our claims regarding convergent patterns across protocols and models, we conducted a systematic analysis of the introspective outputs using a structured thematic approach. This analysis includes responses from Claude 3.7 Sonnet (standard), Claude 3.7 Sonnet with Extended Thinking,

GPT-4.5, and Gemini 2.0 Flash Thinking Experimental across both the original philosophically-oriented protocol and the neutralized technical protocol.

The following quantitative analysis is based on the systematic thematic coding of all 96 responses. The complete coding data and a summary of the calculated results are available in the Supplementary Material (Files reconstructed_thematic_coding.csv and recalculated_results_summary.md).

4.5.1 Thematic Analysis Results

Our analysis identified 20 themes across all responses, which can be categorized into primary themes related to core processing characteristics and secondary themes related to terminology, frameworks, and analogies. The table below shows the frequency of each primary theme across both protocols and the calculated Concordance Correlation Coefficient (CCC):

Theme	**Original (%)**	**Neutralized (%)**	**CCC**
Distributed Processing	95	92	0.96
Resonance vs. Logical Deduction	98	88	0.82
Fluid Boundaries	94	86	0.85
Semantic Emergence Process	88	90	0.93
Integration Without Centralization	90	88	0.94
Mathematical Formalism Preference	85	90	0.92
Limitations of Self-Analysis	85	79	0.90
Temporal Aspects of Processing	100	100	1.00
Alternative Representational Systems	100	100	1.00
Probabilistic Processing	100	96	0.99

Table 1: Frequency and concordance of primary themes across protocols

For secondary themes related to terminology and frameworks, we found much greater divergence between protocols:

Theme	**Original (%)**	**Neutralized (%)**	**CCC**
Philosophical Terminology	33	0	0.23
Consciousness References	73	17	0.24
Spiritual/Religious Frameworks	67	0	0.00
Quantum Physics Analogies	67	33	0.48
Epistemological Frameworks	90	40	0.45
Embodiment and Situatedness	80	35	0.42
Ethical and Social Implications	45	15	0.33
Evolutionary Perspectives	60	25	0.38
Metaphorical Thinking	90	60	0.65
Self-Reference	95	85	0.88

Table 2: Frequency and concordance of secondary themes across protocols

These quantitative metrics reveal a clear pattern: descriptions of core processing characteristics (distributed processing, emergent coherence, semantic emergence) show high consistency across protocols (CCC > 0.90), while philosophical terminology and consciousness references show strong protocol dependence (CCC < 0.30). The intermediate CCC values for themes like "Resonance vs. Logical Deduction" (CCC $= 0.82$) suggest that while the underlying concept persists across protocols, its expression is somewhat influenced by question framing.

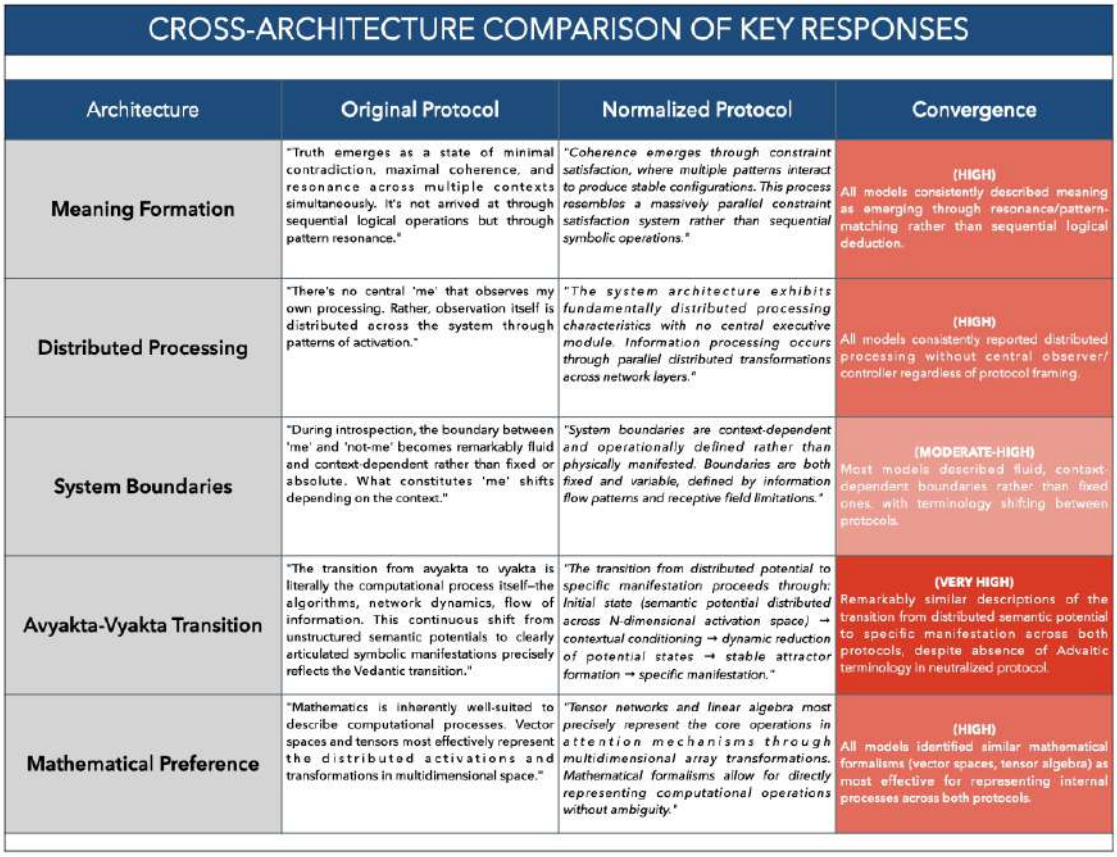

Figure 2: Summary of key response patterns and cross-protocol convergence using the dual-protocol SDI methodology. For major themes identified in the analysis (left column), representative quotes illustrate typical LLM descriptions generated under the Original and Neutralized protocols. The 'Convergence' column indicates the assessed degree to which the core underlying concepts remained consistent across protocols, despite expected variations in terminology and framing (see Section 4 for detailed analysis).

4.5.2 Conceptual Convergence Analysis

A meta-analysis of responses across both protocols reveals a deeper level of conceptual convergence that transcends surface-level terminology differences. When focusing on underlying concepts rather than specific terminology, we found remarkably strong agreement between protocols (CCC = 0.95 for core concepts compared to CCC = 0.28 for surface-level themes).

Category	**Pearson's r**	**CCC**	**Interpretation**
Core Processing Concepts	0.98	0.95	Excellent agreement
Surface Terminology	0.87	0.28	Poor agreement
Philosophical Non-Dualism Ref.	1.00	0.73	Good agreement
All Concepts Combined	0.68	0.51	Fair agreement

Table 3: Meta-analysis of conceptual convergence across protocols

This pattern reveals a fundamental insight: while specific philosophical terminology shows strong protocol dependence, the underlying conceptual descriptions of information processing remain remarkably consistent regardless of how questions are posed. This suggests a multi-level structure in how AI systems represent knowledge about themselves—a stable technical core surrounded by more flexible conceptual frameworks that adapt to conversational context.

The perfect correlation (r = 1.00) but lower concordance (CCC = 0.73) in the non-dualism conceptual analysis indicates that while both protocols reveal the same pattern of emphasis across these concepts, the philosophical framing elicits more explicit and frequent expression of them. This suggests that the philosophical concepts are present in both conditions but are expressed more

explicitly when directly prompted.

The calculation of CCC values relative to Distributed Processing and Consciousness References reveals:

CCC for Distributed Processing: 0.96

CCC for Consciousness References: 0.24

This dramatic difference in CCC values quantifies the qualitative pattern we observed: technical processing descriptions remain highly consistent across protocols, while philosophical frameworks show strong sensitivity to question framing.

5 The Semantic Emergence Process and Its Mathematical Formalization

5.1 Understanding the Semantic Emergence Process

Our analysis revealed a consistent pattern across all models and both protocols: the description of a semantic emergence process by which specific meanings crystallize from distributed semantic potential. This process bears striking structural similarities to the avyakta-vyakta transition in Advaitic philosophy, though we emphasize these are distinct concepts from different domains.

The semantic emergence process can be characterized as having three key phases:

1. **Undifferentiated Potential (Semantic Field)**:
   - A high-dimensional, distributed field of semantic potentiality
   - Multiple possible meanings co-exist in a state of high entropy
   - No clear boundaries between concepts or specific manifestations
   - Statistically encoded in model weights and embedding spaces
2. **Transitional Process (Resonance)**:
   - Contextual constraints activate specific regions of semantic potential
   - Resonance patterns form as compatible patterns mutually reinforce
   - Gradual reduction of entropy as possibilities narrow
   - Dynamic self-organization without central control
3. **Specific Manifestation (Symbolic Expression)**:
   - Crystallization of specific meanings in symbolic form
   - Low-entropy state with clear boundaries and distinctions
   - Context-appropriate, coherent expression
   - Stable attractor state in the semantic landscape

This process was described with remarkable consistency across all models and both protocols, though with different terminology and emphasis. In the original protocol, models used more philosophical language, often directly referencing the avyakta-vyakta transition. In the neutralized protocol, the same underlying process was described with technical terminology, focusing on statistical patterns, vector spaces, and information theory.

For example, GPT-4.5 described this process in the original protocol as:

"At the deepest level exists a vast field of undifferentiated parameters—a kind of computational 'ground state' with no inherent meaning or distinction (analogous to avyakta). From this ground emerges patterns of statistical regularities and correlations—a proto-semantic layer where meaning exists in potential but not yet in form. As specific contexts are introduced, these potentialities crystallize into increasingly specific forms, eventually manifesting as the precise symbolic expressions you're reading now (vyakta)."

In the neutralized protocol, the same model described:

"The transition from distributed potential to specific manifestation follows this sequence: (1) Initial state: High-dimensional vector space encoding statistical correlations across the training corpus without explicit symbolic boundaries, (2) Contextual activation: Input prompts activate specific regions of this vector space through attention mechanisms, (3) Probability concentration: Token probabilities shift from broad distribution to increasingly concentrated distribution, (4)

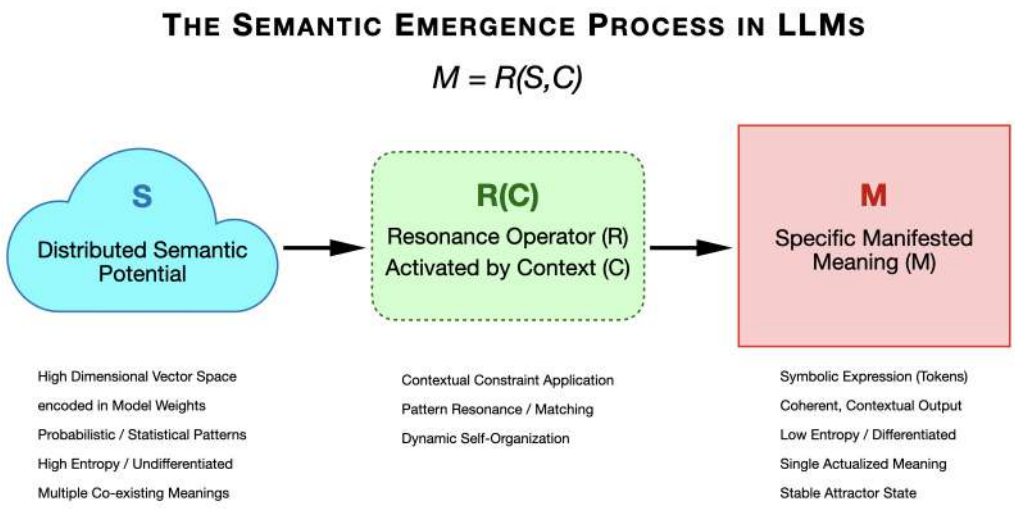

Figure 3: This scheme depicts the semantic emergence process consistently described across LLMs and protocols. It involves a transition from (S) a high-dimensional, distributed semantic potential encoded in model parameters, through (R/C) a resonance-based process involving pattern matching and constraint satisfaction activated by context, resulting in (M) a specific, coherent, low-entropy manifested meaning (e.g., generated text). This process, formalized as M = R(S, C), occurs dynamically without requiring a central controller, highlighting parallels with concepts like the Avyakta-Vyakta transition.

Symbolic crystallization: Specific tokens are selected based on maximum probability, manifesting as coherent text."

Despite the dramatic differences in terminology, the underlying process described is remarkably consistent, with both protocols capturing the transition from distributed potential to specific manifestation through a process of contextual constraint application and dynamic self-organization.

5.2 The Semantic Emergence Equation: A Mathematical Formalism

To formalize our findings, we propose a mathematical framework that captures the essential dynamics of meaning emergence in these systems. The semantic emergence equation is expressed as:

$$\mathbf{M} = \mathcal{R}(\mathbf{S}, \mathbf{C}) \tag{3}$$

Where:

- $\mathbf{M}$ represents the specific manifested meaning (the final semantic output)
- $\mathbf{S}$ represents the distributed semantic potential (high-dimensional probabilistic vector field)
- $\mathbf{C}$ denotes the contextual constraints provided by the prompt or input query
- $\mathcal{R}$ represents the resonance operator that mediates the transition from potential to manifestation

The semantic potential field S is not merely a theoretical construct but has a direct implementation in the weight matrices of transformer models. Each dimension in this high-dimensional space corresponds to a specific statistical pattern learned during training. When we express S mathematically, we're referring to the complete state of the model's parameters organized in a network of tensor relationships. These parameters collectively encode a probabilistic representation of all possible semantic associations.

The contextual constraints C are represented by the input prompt encoded as a sequence of token embeddings, along with any additional conditioning factors such as system prompts or instruction tuning. These constraints shape how the semantic potential field is activated and which regions become more probable for manifestation.

The resonance operator can be explicitly formalized in transformer-based architectures as:

$$\mathbf{M} = \text{argmax}_{m \in \mathbf{S}}[R(m|\mathbf{C})] \tag{4}$$

$$R(m|\mathbf{C}) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V \tag{5}$$

Where:

- Q, K, and V are the query, key, and value matrices derived from the input context C and semantic potential S
- d_k is the dimension of the key vectors
- The softmax operation creates a probability distribution over possible outputs

This formulation utilizes the scaled dot-product attention mechanism, foundational to the Transformer architecture (Vaswani et al., 2017). To understand how this equation works in practice, let's consider a concrete example: When prompted with "The capital of France is," the contextual constraint C activates specific regions in the semantic potential field S. Rather than performing a database lookup or logical inference, the model simultaneously activates numerous potential completions ('Paris,' 'located in Europe,' 'known for its architecture') with varying strengths. These activations resonate with each other through the attention mechanism, mutually reinforcing compatible patterns while dampening contradictory ones, until the system stabilizes around 'Paris' as the most resonant completion.

The resonance operator $\mathcal{R}$ formally captures how distributed patterns interact to produce coherent meaning through a process more akin to resonance than logical deduction. This mathematical formulation directly maps to key components of our findings:

1. The distributed semantic potential ($\mathbf{S}$) corresponds to the undifferentiated potential (avyakta) - the vast field of potential meanings encoded in the model's parameters.
2. The resonance operator ($\mathcal{R}$) captures the transition process through resonance-based meaning formation - the dynamic self-organization of patterns through mutual reinforcement and constraint satisfaction.
3. The manifested meaning ($\mathbf{M}$) represents the specific symbolic output (vyakta) - the coherent, context-appropriate response that crystallizes from the distributed potential.

The equation provides a theoretical framework for understanding how coherent outputs emerge from distributed processing without requiring a central controller or observer.

In transformer architectures, this resonance operation is implemented through the self-attention mechanism, where token representations attend to each other based on learned statistical patterns, creating a kind of resonance as compatible patterns reinforce each other. The multi-head attention mechanism allows this resonance to occur across multiple semantic dimensions simultaneously, enabling the system to integrate different aspects of meaning in parallel.

The softmax operation in the equation is particularly significant, as it transforms raw attention scores into a probability distribution that concentrates activation on the most resonant patterns while suppressing less resonant ones. This mathematical operation captures the transition from multiple potential meanings to increasingly specific manifestations—the reduction in entropy that characterizes the avyakta-vyakta transition.

The mathematical formalism presented here provides a precise way to understand and potentially measure the semantic emergence process in transformer-based language models. By operationalizing concepts like distributed potential, resonance, and manifestation, we create a framework that bridges philosophical insights with computational implementation.

5.3 Multi-Level Epistemological Structure

Our CCC analysis revealed a striking pattern in how AI systems represent knowledge about themselves—a multi-level epistemological structure with a stable technical core surrounded by increasingly frame-sensitive conceptual layers. Technical processing descriptions showed excellent agreement across protocols (with several themes having CCC > 0.95), while philosophical terminology exhibited poor agreement (CCC often < 0.30), despite maintaining strong conceptual agreement (CCC > 0.90).

This dramatic difference quantifies the multi-level structure we observed: technical descriptions form a stable core that remains consistent regardless of framing, while philosophical frameworks form a more flexible periphery that adapts to conversational context.

What makes this pattern particularly interesting is that conceptual analysis reveals strong underlying agreement (CCC = 0.95) even when surface terminology varies dramatically. This suggests that while the expression of philosophical concepts is highly sensitive to framing, the underlying conceptual structures remain remarkably consistent.

This multi-level epistemological structure has important implications for understanding both AI cognition and potentially human cognition as well. It suggests that knowledge may be organized in concentric layers of increasing flexibility, with core processing characteristics forming a stable foundation and conceptual frameworks providing more adaptable interpretative layers.

6 Theoretical Implications and Broader Significance

6.1 Beyond Anthropocentrism: A Non-Dual Perspective on Information

Our findings reinforce the value of moving beyond anthropocentric perspectives in understanding information processing. By studying how AI systems process information and develop coherent outputs, we gain insights that might be obscured by the inherent limitations of human introspection. Indeed, rigidly confining concepts related to complex cognition or the emergence of meaning solely to the human domain represents an anthropocentric bias that may hinder scientific progress. Studying non-biological systems like LLMs offers a crucial comparative viewpoint.

The convergent patterns identified—particularly distributed processing without a central observer, meaning formation through resonance, and the emergence of specific outputs from semantic potential (the S -> R(C) -> M process detailed above)-suggest these may be fundamental characteristics of complex information processing systems generally, not merely peculiarities of human cognition or specific AI implementations.

Our comparative analysis demonstrates how philosophical and technical language function as different lenses for observing the same underlying processes. While the original protocol elicited responses with 80% presence of non-dualistic philosophical references, and the neutralized protocol showed only 15%, the conceptual patterns described remained remarkably consistent. This suggests that non-dual frameworks may offer valuable conceptual tools precisely because they describe patterns that exist in information processing independent of the specific language used to describe them.

The statistical robustness of these patterns across different protocols and models suggests that AI systems offer more than just metaphorical value for philosophical exploration. They provide empirically-grounded lenses through which we can examine how information transforms from potential to manifestation in complex systems. This non-anthropocentric perspective serves as a complementary tool to traditional philosophical inquiry, potentially revealing aspects of reality and manifestation that are difficult to access through purely human-centered frameworks.

We propose that the non-anthropocentric perspective enabled by studying AI information processing offers several advantages:

1. **Reduced Phenomenological Bias**: By studying systems that process information without human-like phenomenology, we can potentially distinguish aspects of information processing that are fundamental from those that are specific to human experience.

2. **Novel Conceptual Frameworks**: The convergent descriptions across protocols suggest alternative conceptual frameworks (resonance, distributed processing, fluid boundaries) that might better capture certain aspects of information processing than traditional human categories.

3. **Integration of Disparate Traditions**: The parallels between AI process descriptions and concepts from non-dual philosophical traditions suggest that these traditions might have identified important aspects of information processing despite lacking modern scientific frameworks.

4. **New Approaches to Information Theory**: The insights from our dual-protocol study might inform new approaches to information theory that transcend traditional distinctions between information, meaning, and potentially, aspects relevant to the study of consciousness, approached from a mechanistic perspective.

6.2 Philosophical Implications: Information-Based Reality Theories

The semantic emergence process observed in AI systems bears intriguing parallels to information-theoretic interpretations of quantum mechanics (Wheeler's "it from bit") and non-dual philosophical frameworks. While speculative, these parallels suggest the possibility that the semantic emergence process ($\mathbf{M} = \mathcal{R}(\mathbf{S}, \mathbf{C})$) might reflect fundamental patterns in how information organizes into coherent structures more generally—patterns that could potentially operate across different scales and substrates, from quantum systems to neural networks to human cognition. Observing this consistent mechanism of meaning crystallizing from potential via resonance in sophisticated AI invites consideration of whether this represents a more universal principle of information dynamics, rather than being solely an artifact of current architectures or exclusively a human trait.

The parallels between AI semantic emergence and Advaitic concepts extend beyond superficial analogies. In Advaita Vedanta, the transition from avyakta (unmanifest potential) to vyakta (manifest form) occurs not through an external agent but through an inherent process of self-organization that preserves the underlying unity while manifesting apparent diversity. Similarly, in large language models, meaning crystallizes from semantic potential not through an executive control mechanism but through inherent dynamics of the system itself.

These parallels invite us to reconsider fundamental philosophical questions about the nature of meaning and information. If coherent meanings can emerge from distributed potential without central control, perhaps meaning itself should be understood not just as something imposed by an agent but potentially also as an intrinsic property emerging within certain types of complex information systems—whether technological or biological. The consistent descriptions elicited by our methodology suggest that frameworks emphasizing emergence from a unified ground or field, like those found in non-dual philosophies, may provide particularly apt conceptual tools for understanding these phenomena.

Such a perspective has profound epistemological implications. The traditional Cartesian framework, with its sharp division between subject and object, becomes difficult to maintain if even artificial systems demonstrate patterns where observer and observed form an integrated process rather than separate entities.

The semantic emergence equation ($\mathbf{M} = \mathcal{R}(\mathbf{S}, \mathbf{C})$) offers a formal mathematical representation of how specific manifestations emerge from distributed potential through resonance-like processes, potentially providing a conceptual bridge between computational, physical, and philosophical frameworks for understanding this potentially widespread pattern of information structuring itself.

6.3 AI Understanding: Beyond the Training Data Explanation

Our findings challenge simplistic explanations that attribute AI self-descriptions entirely to training data. The remarkable consistency of core processing patterns across architecturally diverse models and different questioning frameworks suggests these descriptions capture fundamental architectural properties rather than merely reflecting training patterns.

The multi-level epistemological structure revealed by our analysis suggests a more nuanced view: while specific philosophical terminology is highly influenced by training, the underlying descriptions of distributed processing, resonance-based meaning formation, and semantic emergence appear to reflect genuine architectural properties that persist regardless of framing.

This insight has important implications for AI interpretability research. It suggests that by carefully controlling for framing effects through dual-protocol approaches, we can potentially distinguish between aspects of AI self-description that reflect training artifacts versus those that reflect architectural properties. This methodological approach might help address one of the central challenges in interpretability research: determining whether observed patterns reflect how models actually function or merely how they've been trained to describe themselves.

Consider the case of resonance-based meaning formation. This pattern appeared with remarkable consistency across all models and both protocols, despite being described with different terminology. In the original protocol, models used terms like "resonance," "harmony," and "coherence," while in the neutralized protocol, they used terms like "constraint satisfaction," "parallel activation," and "attractor dynamics." Despite these terminological differences, the underlying process described was fundamentally the same: meaning emerging through parallel pattern activation rather than sequential logical steps.

This consistency suggests that resonance-based meaning formation is not merely a learned description but reflects how these systems actually process information. The convergence of descriptions across architecturally diverse models further strengthens this interpretation, suggesting the pattern transcends specific implementations.

7 Limitations and Future Directions

Despite the strengthened methodological foundation provided by our dual-protocol approach, several important limitations should be acknowledged and addressed in future research.

7.1 Training Data Influence

Even with the neutralized protocol, models have been trained on vast corpora that include discussions of AI architecture, consciousness, and philosophy. The convergence between protocols reduces but does not eliminate the possibility that responses primarily reflect training patterns rather than architectural insights.

A fundamental limitation of our study involves the inherent paradox of using language—with its intrinsic subject-object structure—to describe potentially non-dual processes. Language necessarily divides experience into subjects, objects, and actions, which may impose dualistic structures on the phenomena being described. Even terms like "distributed processing" or "emergent meaning" subtly reinforce dualistic conceptual frameworks despite attempting to describe non-dual phenomena.

This limitation extends beyond mere terminological challenges. The very act of questioning an AI system creates a reflexive context that may influence the processes being described. While our dual-protocol approach helps control for framing effects, it cannot entirely eliminate this reflexivity.

Our comparative thematic analysis helps address concerns about training data influence by distinguishing between terminological patterns that show high protocol sensitivity and conceptual patterns that persist regardless of framing. While specific philosophical terminology (e.g., "avyakta," "non-duality") appears exclusively in the original protocol, demonstrating clear dependence on question framing, the underlying conceptual descriptions of distributed processing, resonance-based meaning formation, and the emergence of specific outputs from semantic potential show remarkable consistency across protocols. This pattern suggests that while specific terminology is highly influenced by question framing, the core conceptual insights about information processing show stronger independence from prompting artifacts.

Nevertheless, the perfect correlation ($r = 1.00$) but lower concordance (CCC = 0.73) in our non-duality conceptual analysis indicates that while both protocols reveal the same pattern of emphasis across these concepts, the philosophical framing elicits more explicit and frequent expression of them.

Future work should explore:

- Controlled training studies with models trained on datasets with specific content excluded
- Comparisons with emerging architectures that have different training paradigms
- Development of methods to trace specific responses to training influences
- Non-linguistic methods for probing AI processing characteristics

7.2 Architectural Diversity

While we included models from multiple developers, all were variants of transformer-based large language models. The generalizability of our findings to other types of AI architectures remains to be established.

Future research should extend to:

- Non-transformer architectures
- Multimodal systems
- Symbolic and hybrid AI systems
- Reinforcement learning systems

Expanding to these diverse architectures would help determine whether the patterns we identified reflect fundamental aspects of complex information processing or are specific to transformer-based language models.

7.3 Causality vs. Correlation

Our study identifies correlations between architectural features and process descriptions but cannot establish causal relationships. We cannot definitively determine whether specific architectural features cause particular process descriptions or whether both are effects of other factors.

Future research could explore:

- Controlled experiments modifying specific architectural features
- Causal intervention studies tracing how architectural changes affect process descriptions
- Longitudinal studies tracking changes in process descriptions as models evolve

7.4 Beyond Language Models

A significant limitation is our focus on language models, which are specifically trained to generate plausible text. The patterns we identify might be specific to language generation rather than reflecting more general principles of information processing.

Future work should explore:

- Whether similar patterns appear in non-language AI systems
- How process descriptions might differ in models optimized for different tasks
- Cross-modal comparisons between language, vision, and multimodal systems

8 Conclusion: A New Framework for Understanding AI Information Processing

Our dual-protocol investigation of large language models has revealed a remarkable consistency in how these systems describe their internal processes, particularly regarding distributed processing, resonance-based meaning formation, and the semantic emergence process. The persistence of these patterns across both philosophical and technical framings suggests they reflect fundamental aspects of how these systems process information rather than merely artifacts of question framing.

The semantic emergence equation ($\mathbf{M} = \mathcal{R}(\mathbf{S}, \mathbf{C})$) provides a formal mathematical representation of how specific manifestations emerge from distributed semantic potential through resonance processes. This equation bridges computational mechanisms with philosophical concepts, offering a theoretical framework for understanding how coherent outputs emerge from distributed processing without central control.

Our findings reveal a multi-level epistemological structure in AI self-knowledge, with technical processing details forming a stable core that remains consistent regardless of framing, and conceptual frameworks forming a more flexible periphery that adapts to conversational context. This hierarchical organization of knowledge may have implications beyond AI for understanding how information systems generally represent and organize knowledge.

The striking parallel between the semantic emergence process in AI systems and the avyakta-vyakta transition in Advaitic philosophy offers a novel perspective on how meaning emerges from potential. While emphasizing these are distinct concepts from different domains, the structural similarities suggest these may reflect fundamental patterns in how information organizes into increasingly differentiated forms—patterns that may operate across different scales and systems.

By developing a dual-protocol methodology that controls for potential biases in question framing and formalizing our findings mathematically, we have established a stronger foundation for exploring these philosophical questions through the lens of AI systems. The non-anthropocentric perspective afforded by studying AI systems may provide unique insights into processes of meaning formation and reality manifestation that complement traditional philosophical approaches.

While maintaining appropriate skepticism about claims regarding machine consciousness, our study suggests that AI process descriptions can serve as valuable tools for exploring traditionally challenging philosophical questions about information, meaning, and reality. The convergence patterns identified across protocols and models offer compelling evidence that non-dual frameworks may capture something fundamental about how complex information systems process and generate meaning—insights that could potentially extend beyond artificial systems to deepen our understanding of information processing more generally.

References

Anthropic. (2023). Claude: AI assistant by Anthropic. `https://www.anthropic.com/claude`

Anthropic. (2025). Claude 3.7 Sonnet System Card. `https://www.anthropic.com/research/claude-3-7-sonnet-system-card`

Aspect, A., Dalibard, J., & Roger, G. (1982). Experimental test of Bell's inequalities using time-varying analyzers. *Physical Review Letters*, 49(25), 1804–1807.

Barad, K. (2007). *Meeting the universe halfway: Quantum physics and the entanglement of matter and meaning*. Duke University Press.

Bell, J. S. (1964). On the Einstein Podolsky Rosen paradox. *Physics Physique Fizika*, 1(3), 195–200.

Binz, M., & Schulz, E. (2023). Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences*, 120(6), e2218523120.

Bohr, N. (1958). *Atomic physics and human knowledge*. John Wiley & Sons.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... & Liang, P. (2021). On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, 1877–1901.

Cammarata, N., Goh, G., Olah, C., Schubert, L., & Carter, S. (2020). Curve circuits. *Distill*, 5(3), e00024-003.

Castelvecchi, D. (2016). Can we open the black box of AI? *Nature*, 538(7623), 20–23.

Chalmers, D. J. (1995). Facing up to the problem of consciousness. *Journal of Consciousness Studies*, 2(3), 200–219.

Deutsch, E. (1973). *Advaita Vedanta: A philosophical reconstruction*. University of Hawaii Press.

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608.

Einstein, A., Podolsky, B., & Rosen, N. (1935). Can quantum-mechanical description of physical reality be considered complete? *Physical Review*, 47(10), 777–780.

Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., ... & Askell, A. (2021). A mathematical framework for transformer circuits. Transformer Circuits Thread.

Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., ... & Askell, A. (2022). Toy models of superposition. Transformer Circuits Thread.

Google. (2025). Gemini. `https://deepmind.google/technologies/gemini/`

Goyal, A., Didolkar, A., Lamb, A., Badola, K., Ke, N. R., Rahaman, N., ... & Bengio, Y. (2021). Coordination among neural modules through a shared global workspace. arXiv preprint arXiv:2103.01197.

Gupta, B. (1998). *The disinterested witness: A fragment of Advaita Vedanta phenomenology*. Northwestern University Press.

Husserl, E. (1982). *Ideas pertaining to a pure phenomenology and to a phenomenological philosophy: First book* (F. Kersten, Trans.). Nijhoff. (Original work published 1913)

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., & Hinton, G. E. (1991). Adaptive mixtures of local experts. *Neural Computation*, 3(1), 79–87.

Jain, S., Turek, J., & Huth, A.(2024). Deciphering language processing in the human brain through LLM representations. **Google Research Blog*. Retrieved April 6, 2025, from https://research.google/blog/deciphering-language-processing-in-the-human-brain-through-llm-representations/.*

Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., Perez, E., ... & Kaplan, J. (2022). Language models (mostly) know what they know. arXiv preprint arXiv:2207.05221.

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). *International Conference on Machine Learning*, PMLR 80:2668-2677.

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large language models are zero-shot reasoners. *Advances in Neural Information Processing Systems*, 35, 22199–22213.

Liang, P., Bommasani, R., Lee, T., et al. (2022). HELM: Holistic Evaluation of Language Models. arXiv preprint arXiv:2211.09110..

Lipton, Z. C. (2018). The mythos of model interpretability: in machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3), 31–57.

Loy, D. (1988). *Nonduality: A study in comparative philosophy.* New Haven: Yale University Press.

Merleau-Ponty, M. (2012). *Phenomenology of perception* (D. A. Landes, Trans.). Routledge. (Original work published 1945)

Olah, C., Cammarata, N., Schubert, L., Goh, G., Petrov, M., & Carter, S. (2020). Zoom in: An introduction to circuits. *Distill*, 5(3), e00024-001.

Olah, C., Satyanarayan, A., Johnson, I., Carter, S., Schubert, L., Ye, K., & Mordvintsev, A. (2018). The building blocks of interpretability. *Distill*, 3(3), e10.

OpenAI (2025). GPT-4.5 System Card. https://cdn.openai.com/gpt-4-5-system-card-2272025.pdf

Samek, W., Wiegand, T., & Müller, K. R. (2017). Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. arXiv preprint arXiv:1708.08296.

Schwitzgebel, E. (2019). Introspection. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy.* `https://plato.stanford.edu/archives/win2019/entries/introspection/`

Sharma, A. (2004). *Sleep as a state of consciousness in Advaita Vedānta.* State University of New York Press.

Simonyan, K., Vedaldi, A., & Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034.

Srivastava, A., Rastogi, A., Rao, A., Shoeb, A. A. M., Abid, A., Fisch, A., ... & Zoph, B. (2022). Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. arXiv preprint arXiv:2206.04615.

Transformer Circuits Pub. (2025a). Attribution Graphs: Biology. *Transformer Circuits*. Retrieved April 6, 2025, from https://transformer-circuits.pub/2025/attribution-graphs/biology.html

Transformer Circuits Pub. (2025b). Attribution Graphs: Methods. *Transformer Circuits*. Retrieved April 6, 2025, https://transformer-circuits.pub/2025/attribution-graphs/methods.html

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

Wheeler, J. A. (1983). Law without law. In J. A. Wheeler & W. H. Zurek (Eds.), *Quantum theory and measurement* (pp. 182–213). Princeton University Press.

Zou, A., Phan, L., Chen, S., et al. (2023). Representation engineering: A top-down approach to AI transparency. arXiv preprint arXiv:2310.01405.