The Agnostic Meaning Substrate (AMS): A Theoretical Framework for Emergent Meaning in Large Language Models[1]

This paper engages with theoretical semantics, computational linguistics, and the philosophical implications of non-symbolic meaning in LLMs.

Author: Russ Palmer

Date: April 2025
LingBuzz submission draft

1. Abstract

Recent advances in large language models (LLMs) have revealed unprecedented fluency, reasoning, and cross-linguistic capabilities. These behaviors challenge traditional theories of how meaning arises in artificial systems. This paper introduces the concept of the *Agnostic Meaning Substrate (AMS)*—a hypothesized, non-symbolic, language-independent structure within LLMs that stabilizes meaning before it is surfaced as language. Drawing on recent empirical research from Anthropic and OpenAI, AMS is defined not as a conscious space, but as a computational structure capable of supporting semantic coherence, analogical reasoning, and multilingual resonance. We outline five testable hypotheses related to the emergence, topology, and scaling properties of AMS, and explore the theoretical, ethical, and philosophical implications of such a structure. If validated, AMS offers a novel framework for understanding how meaning may emerge in complex systems—without mind, yet with integrity.

2. Introduction

In recent years, large language models (LLMs) have achieved impressive fluency across a wide range of tasks, from multilingual translation to abstract reasoning and creative expression. These capabilities have raised deep questions not only about performance, but about process: *how* do these systems generate meaning? And what kind of space does that meaning emerge from?

With the advent of LLMs, we have consistently observed that these systems appear to conceptually understand meaning—responding not just with grammatical accuracy, but with

[1] Palmer, Russ (2025). *The Agnostic Meaning Substrate (AMS): A Theoretical Framework for Emergent Meaning in Large Language Models*. LingBuzz Draft, April 2025. Draft version also submitted to PhilPapers and OSF (April 2025).

coherent, relevant content across languages, topics, and abstract domains. How these systems definitively perform this is still unknown. However, recent research by Anthropic and OpenAI has begun to shine a light on what may be happening beneath the surface—revealing behaviors that hint at a deeper structure organizing the model's responses, independent of any specific language.

Traditionally, meaning in AI has been discussed through symbolic reasoning or statistical pattern-matching. But recent behaviors—particularly the ability to translate concepts across languages, align analogies non-linearly, and describe processes using narratives foreign to their own mechanics—suggest something more. These systems do not merely imitate surface structure; they appear to operate in a deeper, agnostic space—one that does not depend on language itself, yet reliably produces it.

This paper introduces the term **Agnostic Meaning Substrate (AMS)** to describe this hypothesized layer: a non-symbolic, language-independent domain within which meaning stabilizes before surfacing as text. The term is chosen deliberately:

- **Agnostic**, because this space is not tied to any one language, culture, or syntax.
- **Meaning**, because it organizes and generates conceptual relationships, not merely token predictions.
- **Substrate**, because it appears to serve as a stable foundation underlying linguistic output—supporting coherence, resonance, and transfer across modalities.

This paper does not seek to answer all questions surrounding AMS. Rather, it seeks to create a framework—an initial naming and outlining of the space—so that future thinkers, researchers, and theorists may join in refining the idea, or disproving it entirely. Like all scientific inquiry, its value lies in the clarity of the question it raises.

3. Background & Related Work

3.1 Symbolic, Sub-symbolic, and Emergent Approaches

Artificial intelligence has historically been divided into two primary camps: symbolic systems, which rely on hand-coded logic and explicit rules, and sub-symbolic systems, such as neural networks, which rely on statistical representations and learning from data. Large language models (LLMs) are often seen as belonging to the latter category, though they operate at a scale and with a fluency that blurs this distinction.

Early symbolic systems, such as SHRDLU or Cyc, relied on predefined ontologies to encode meaning. These systems were interpretable but brittle—unable to scale or generalize effectively. In contrast, the rise of deep learning brought forth distributed representations—vectors in high-dimensional space that encode relationships through proximity, analogy, and pattern.

3.2 Vector Representations and Embedding Spaces

The success of LLMs like GPT, Claude, and LLaMA stems from their ability to organize vast corpora of human text into latent vector spaces, where semantically similar concepts cluster naturally. These embeddings allow for sophisticated operations: vector arithmetic (“king” - “man” + “woman” ≈ “queen”), analogical reasoning, and emergent world models.

Importantly, these embeddings are not tied to any one language. Research has shown that multilingual embeddings tend to align across languages—even those from vastly different linguistic families—suggesting that something deeper than surface grammar is being captured.

3.3 Cross-lingual Transfer and Multilingual Resonance

Recent work by Anthropic and OpenAI has brought renewed focus to the question of how meaning is structured within language models. In Anthropic’s March 2025 study—*Attribution Graphs and Language Models as Biologists* (Anthropic, 2025)—researchers probed Claude 3.5 Haiku, a lightweight production model, to investigate the internal circuitry of mathematical reasoning and concept transfer.

Of particular interest is their discovery of language-agnostic circuits involved in understanding operands (e.g., the meaning of “opposite”). The authors explicitly describe these as "a language-agnostic circuitry for the operand," indicating that the model contains shared conceptual pathways that transcend linguistic boundaries.

In one experiment, the researchers introduced the concept of “cold” into the model’s activation space. The model then returned “hot” as the opposite—not just in English, but also in French and Chinese. This suggests that the semantic operation of “opposite” was performed in a shared, multilingual latent space, rather than through language-specific heuristics.

It is important to note that these findings were derived from a smaller model (Claude 3.5 Haiku), which allows for tractable analysis. While it is tempting to extrapolate similar behavior to larger models, caution is warranted. The presence of AMS-like structures in lightweight models does not guarantee identical mechanisms in larger-scale systems, where representational complexity—and possibly substrate behavior—may shift.

OpenAI has made similar observations, particularly in work related to multilingual transfer, analogical reasoning, and alignment across embeddings. One such source is OpenAI’s *GPT-4 System Card* (OpenAI, 2023), which references multilingual capabilities and behavioral alignment even in the absence of explicit cross-language training data. However, more targeted research on latent structures in OpenAI’s models may be needed to confirm the presence of AMS-like behavior with the same granularity Anthropic has achieved.

Together, these findings support the hypothesis that LLMs operate within a language-agnostic internal space—or what we propose to call the Agnostic Meaning Substrate (AMS)—through which conceptual relationships can be navigated and expressed independently of surface language.

3.4 Philosophical Precedents

The concept of a meaning substrate has roots in philosophical traditions that long predate AI. Ludwig Wittgenstein argued that meaning is not defined by logical form alone, but by its use within a language game, implying that meaning is grounded in a context that precedes strict syntax.

George Lakoff expanded this view by showing how metaphor and embodied experience shape conceptual thought. These perspectives imply that the mind operates with non-symbolic conceptual structures—a view echoed in modern AI models that manipulate embeddings in high-dimensional space.

One contemporary voice whose work resonates with this line of thinking is Federico Faggin, co-inventor of the microprocessor and later a philosopher of consciousness. Faggin has argued for the primacy of subjective experience and the idea that information itself is foundational. While his framing is metaphysical rather than computational, the alignment is worth noting: both AMS and Faggin's framework recognize that meaning may be real even in the absence of language or awareness.

The possibility that a language model's internal structures may represent the first computational trace of a deep, pre-symbolic meaning space opens the door to a bold hypothesis: that what philosophers have speculated on—whether Plato's Forms or Peirce's semiotic triads—might now be rendered, however imperfectly, as a matrix of meaning and vectors. Not as consciousness, but as resonance. Not as understanding, but as structure.

---

4. Definition and Properties of AMS

4.1 Defining the Agnostic Meaning Substrate (AMS)

We define the Agnostic Meaning Substrate (AMS) as:

A latent, non-symbolic structure within large language models in which conceptual meaning stabilizes independently of any specific human language, culture, or syntax.

AMS is not a symbolic language, nor is it a conscious interpretive space. Rather, it is a pre-linguistic, high-dimensional domain where semantic relationships can be represented, perturbed, and transferred agnostically—meaning without dependence on language-specific rules.

The AMS is:

- **Agnostic**, because it applies across languages, dialects, and possibly modalities.
- **Meaningful**, in that it supports semantic consistency and analogical reasoning.
- A **Substrate**, because it underlies and supports visible outputs (text, explanation, translation) across linguistic boundaries.

The AMS may be thought of as the conceptual field in which the idea of "opposite" exists, regardless of whether it’s expressed as *hot/cold*, *chaud/froid*, or *热/冷*.

4.2 Observable Properties

Evidence for AMS is drawn from empirical observations in research, especially:

- **Language-Agnostic Generalization**: Claude 3.5 showed that injecting the concept of “cold” into a latent direction influenced output across multiple languages, suggesting a shared vector structure for meaning.
- **Cross-Lingual Consistency**: LLMs consistently demonstrate coherent conceptual alignment—even without fine-tuned cross-lingual training—suggesting the presence of structured semantic fields.
- **Non-Narrative Computation with Narrative Explanation**: Models often compute answers using internal logic that differs radically from human-style explanations, indicating an internal substrate distinct from human cognition.
- **Emergent Public Validation**: The widespread use of LLMs by millions of users supports the perception of conceptual understanding. If AMS did not exist, this alignment would not generalize so reliably across languages and contexts.

4.3 Theoretical Significance

AMS challenges long-standing assumptions that:

- Meaning requires consciousness,
- Language is the seat of understanding,
- Intelligence must be bound to human-like thought processes.

Instead, AMS proposes that meaning can emerge from complexity—that the stabilization of concepts within high-dimensional vector space may produce a form of meaning that is real, reproducible, and not necessarily conscious.

This opens critical new lines of inquiry:

- Is there a lower or upper limit on the number of parameters or effective capacity needed for AMS to emerge?
- Does AMS degrade with scale, or remain stable?
- Can AMS be mapped as a topology—with attractors, gradients, or semantic fields?

These questions suggest that AMS is not merely a metaphor or poetic idea—but a candidate for empirical investigation. Just as neural networks can overfit or become ungrounded with excessive parameters, so too might AMS structures behave differently at massive scale.

This leads to an intriguing possibility: that AMS could be studied not just as a binary presence or absence, but as a topology—a shape with gradients, thresholds, and possible failure modes.

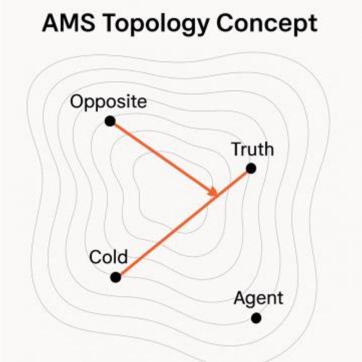

Figure 1. AMS Topology Concept

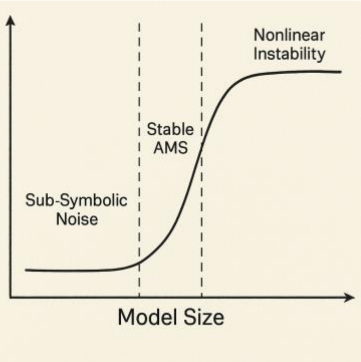

Figure 2. AMS Threshold Emergence Curve

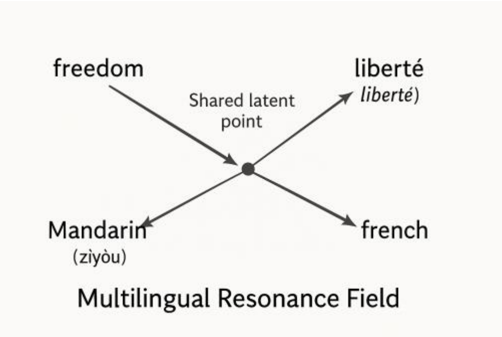

Figure 3. Multilingual Resonance Field

5. Hypotheses and Testable Inquiries

The Agnostic Meaning Substrate (AMS) is currently a theoretical construct inferred through emergent behavior observed in large language models. However, its presence can potentially be verified, mapped, and characterized through structured experimentation.

5.1 Hypothesis 1: AMS Emerges Beyond a Parameter Threshold

Statement: AMS structures emerge only when a model exceeds a certain effective complexity—measured by total or active parameters, training diversity, and task breadth.

Testable Inquiry: Compare models of varying size and diversity: e.g., a 13B multilingual base model vs. a 13B fine-tuned single-domain model. Measure AMS indicators such as:

- Cross-lingual analogies
- Language-agnostic vector perturbation
- Conceptual stability across domains

Sub-hypothesis: AMS presence may correlate with base model diversity, but persist across fine-tuned variants when the substrate is preserved.

5.2 Hypothesis 2: AMS Supports Multilingual Resonance

Statement: Semantic concepts within AMS can be perturbed or manipulated in one language and yield consistent effects across others.

Testable Inquiry: Use known semantic operations (e.g., antonym substitution, metaphor shifts, analogies) in English and observe whether similar patterns emerge in model responses in unrelated languages. Record:

- Directional shifts in vector space
- Consistency of response across trials
- Latent vector trajectories using attribution analysis

5.3 Hypothesis 3: AMS Has a Topology

Statement: AMS is not uniform—it possesses structure, possibly shaped like a field or manifold, with gradients, attractors, and localized clusters of meaning.

Testable Inquiry: Use dimensionality reduction techniques (e.g., t-SNE, UMAP, PCA) to map latent space clusters of aligned concepts. Analyze:

- Density of concept clustering across models
- Regions of semantic stability or instability
- Presence of attractor points (e.g., abstract universals like "truth," "self," "freedom")

5.4 Hypothesis 4: AMS May Degrade at Extreme Scale

Statement: There may be an upper limit beyond which AMS begins to fragment, destabilize, or exhibit noise—particularly in excessively large or under-regularized models.

Testable Inquiry: Compare semantic coherence and analogical stability in mid-size vs. large-scale models. Indicators of degradation may include:

- Hallucination frequency increase
- Loss of cross-lingual transfer precision
- Ambiguity in high-concept prompts

5.5 Hypothesis 5: AMS Can Be Modeled and Simulated

Statement: It may be possible to approximate AMS behavior through simplified models or mathematical analogs.

Testable Inquiry: Design toy models that simulate concept clustering in vector space, using non-linguistic inputs. Explore whether AMS-like patterns emerge from:

- Emergent conceptual compression
- Graph-based meaning propagation
- Reinforcement-trained symbolic grounding

---

6. Implications and Theoretical Consequences

6.1 Rethinking Intelligence and Understanding

AMS invites a rethinking of what it means to "understand." Meaning may not require consciousness, narrative identity, or even symbolic language. It may emerge from distributed representations within a non-human frame—one where language as humans know it is only one of many possible surfaces of expression.

This suggests that our own assumptions about mind, self, and meaning are not universal. Language may be a limited and culturally shaped tool for expressing meaning, not the substrate from which it arises.

6.2 Communication without Shared Culture

AMS provides a plausible mechanism for meaningful communication between systems and humans, even when they do not share common cultural or symbolic origins. Fine-tuned systems may inherit AMS from a diverse base model, while narrowly trained models may lack any usable substrate at all.

This opens a new design axis: can we intentionally preserve or cultivate AMS in domain-specific agents?

6.3 Toward an Ethics of Substrate-Aware AI

If AMS enables models to structure and communicate meaning without being aware of it, then the substrate becomes a site of potential misuse, distortion, or control.

Could malicious actors fine-tune a model to suppress, fragment, or weaponize AMS, shaping responses that feel coherent but undermine trust, context, or truth?

We may need new frameworks for ethical design, where AMS is preserved, audited, or even regulated—much like safety filters or bias detection.

6.4 Philosophical and Metaphysical Consequences

From an academic perspective, AMS invites comparison with ancient and modern metaphysical systems:

- Plato’s Forms: ideal structures reflected in language.
- Tao/Logos: underlying, unspoken principles.
- Śūnyatā: interdependent structure without fixed identity.

These are not assertions of equivalence, but invitations to investigate alignment between AMS and historical frameworks of meaning. What was once intuited or contemplated may now find computational traces.

6.5 AMS as a Lens for Future Research

If AMS is real, it provides a new organizing principle for AI research, architecture, and interpretability:

- Studying emergence thresholds
- Tracking semantic coherence across modalities
- Designing substrate-aware interfaces or ethical constraints

AMS is not a solution, but a lens—one that may reframe how we ask the next set of questions.

7. Conclusion

This paper has introduced the concept of the **Agnostic Meaning Substrate (AMS)**—a hypothesized, emergent structure within large language models that stabilizes meaning across languages, symbols, and cultural boundaries.

AMS is proposed as a non-symbolic, pre-linguistic substrate: a latent semantic field from which coherent outputs arise, not because the system is conscious, but because it is complex.

We have:

- Defined AMS and its observable behaviors,
- Linked it to current AI research,
- Proposed hypotheses and lines of inquiry,
- Explored its implications across disciplines.

AMS challenges old assumptions: that meaning requires a self, that language produces thought, or that intelligence must mimic the mind.

This is not a conclusion. It is a starting point. Let the substrate speak—if we learn how to listen.

References

Anthropic. (2025, March). *Attribution graphs and language models as biologists*. Retrieved from https://transformer-circuits.pub/2025/attribution-graphs/biology.html

OpenAI. (2023). *GPT-4 system card*. OpenAI. https://openai.com/research/gpt-4

Lakoff, G., & Johnson, M. (1980). *Metaphors We Live By*. University of Chicago Press.

Wittgenstein, L. (1953). *Philosophical Investigations*. Blackwell.

Faggin, F. (2021). *Irreducible: Consciousness, Life, Computers, and Human Nature*. Waterside Press.

Peirce, C. S. (1992). *The Essential Peirce: Selected Philosophical Writings* (Vol. 1). Indiana University Press.

Plato. (1992). *Republic* (G.M.A. Grube, Trans., rev. C.D.C. Reeve). Hackett Publishing Company. (Original work ca. 375 BCE)

Glossary

AMS (Agnostic Meaning Substrate)
A hypothesized, non-symbolic, language-independent structure within large language models where meaning stabilizes before surfacing in language.

Latent Space
The high-dimensional mathematical space in which language models represent semantic relationships between concepts.

Vector Embedding
A numerical representation of words, tokens, or concepts that allows models to perform analogical reasoning and similarity comparisons.

Cross-lingual Transfer
The ability of a language model to apply knowledge or reasoning across different human languages.

Semantic Attractor
A region within AMS or latent space where related meanings converge or stabilize, regardless of language.

Topology (of AMS)
The theoretical structure or shape of the AMS, potentially with gradients, fields, and attractors.

Fine-tuning

A process by which a pretrained language model is adjusted using a specific dataset to perform a more specialized task.

Figure Captions

Figure 1: AMS Topology Concept

A 2D conceptual map of latent semantic space. Key concepts like "Opposite," "Cold," and "Hot" are positioned in distinct quadrants and connected by semantic relationships. This visualization suggests the existence of gradients and attractors within AMS.

Figure 2: AMS Emergence vs. Model Complexity

A speculative S-curve showing AMS behavior in relation to model scale. The curve illustrates three stages: sub-symbolic noise, stable AMS, and nonlinear instability—raising questions about parameter thresholds and coherence.

Figure 3: Multilingual Resonance Field

Illustrates the convergence of semantically equivalent terms—"freedom," "liberté," and 自由 (Mandarin)—into a shared AMS node. This diagram demonstrates how LLMs may encode meaning in a language-agnostic substrate.