AI Coherence: Meaning from Patterns, Not Memory

White Paper · April 2025

Author: Travis Frisinger
Technical Director of AI, 8th Light
https://aibuddy.software

Abstract

This paper introduces a practical lens—**coherence reconstruction**—for understanding the utility of large language models (LLMs). Instead of retrieving facts or simulating intelligence, LLMs generate value by reassembling meaning through distributed conceptual activation. Prompts act as force vectors, navigating high-dimensional latent space to awaken emergent subnetworks—temporary clusters of related ideas that respond to intent.

When grounded with external context like documents, search results, or tool output, these systems shift from merely sounding plausible to delivering useful, trustworthy results. Hallucinations, often seen as flaws, are reframed here as symptoms of incomplete anchoring within that coherence-seeking process.

Under this theory, utility doesn't come from correctness alone—it comes from the model's ability to reflect, extend, and scaffold human thinking through recomposed patterns. This theory repositions LLMs as *stupid machines doing something smart*: constructing conceptual coherence at scale.

Who should read this

Technology executives, product owners, AI platform leads, prompt engineers, and governance teams assessing when and how to deploy LLMs in real-world workflows.

Context & Conceptual Foundations

This theory is inspired by a range of research in interpretability, learning dynamics, and large-scale model behavior, including:

- **Feature Superposition and Concept Circuits** – Anthropic's work on tracing thoughts and interpretability in Claude
- **Lottery Ticket Hypothesis** – Frankle & Carbin's exploration of sparse subnetworks in deep models
- **Double Descent and Scaling Laws** – Observed generalization patterns in deep learning at scale
- **Attention as Message Passing** – Graph neural network analogies by Petar Veličković
- **Retrieval-Augmented Generation (RAG)** – Techniques for grounding LLM output using external context
- **Ontological Semantics and Linguistic Modeling** – Earlier AI approaches that relied on structured meaning representations (e.g., GOLD, Ontological Semantics)
- **Prompt Engineering via Structural Injection** – Google's recent work on graph-structured prompting for LLMs
- **Elegant Mathematical Foundations** – As discussed in *Why Machines Learn* by Anil Ananthaswamy

While the framing in this paper is original, it stands on the shoulders of these and other key ideas in the field.

Key Term Definitions:

- **Force Vector**: The directional "push" a prompt gives through the model's latent space, weighted by token embeddings and attention scores.
- **Coherence Field**: The evolving activation landscape in which semantic clusters (conceptual clouds) emerge under the influence of prompts and context.
- **Emergent Subnetwork**: A transient collection of neurons or attention heads whose joint activation encodes a particular concept or pattern in response to a prompt.
- **Instability Signal**: When those activations fragment or drift—manifesting as hallucinations—which can indicate ambiguous input rather than outright "failure."

What Language Models Really Do

Large Language Models (LLMs) aren't search engines.
They're not databases.
And they're not just fancy parrots repeating things they've memorized.

Instead, they're machines trained to recognize and recreate **patterns of meaning**—to take in a prompt and generate a response that **feels coherent** based on everything they've learned.

They work by **spreading concepts out** across many parts of the network during training. When you give them a prompt, they **stir up a pattern** that blends together related pieces of meaning—kind of like lighting up a cloud of related ideas that together capture the *feeling* or *essence* of what you're asking about.

How Prompts Navigate Meaning

When you give a model a prompt, you're not just feeding it words. You're giving it **a directional push**—what you might call a *force vector*.

This push sends the model moving through a **conceptual landscape** where different patterns of meaning live. As it moves, it activates **clusters of related ideas** that feel relevant to what you asked.

These clusters are **not pre-programmed**. They emerge during training—some parts of the model just turn out to be better at certain things, like storytelling, or reasoning, or following instructions. These are like **specialized subnetworks** that light up when they're needed.

The model passes signals across those areas, trying to come up with something that **matches the intent** of your prompt and makes sense in context.

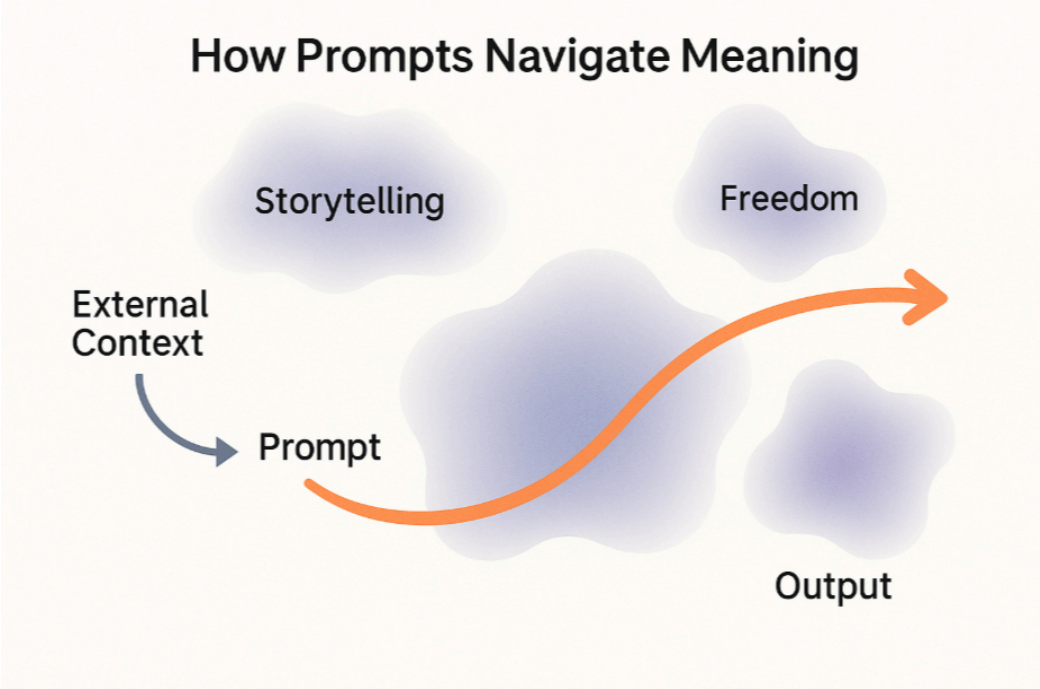

Figure 1: Prompt and External Context Navigating the Coherence Field

A prompt enters the latent conceptual space, activating emergent clusters of meaning like *freedom* and *storytelling*. External context reshapes the coherence field—anchoring the model's response toward grounded, relevant output.

A New Kind of Utility

Most AI theories talk about utility in terms of **compression**—how efficiently a model can store and retrieve knowledge. But that misses the point.

The real utility of LLMs isn't in their compression—it's in their reconstruction of the essence of an idea. Their ability to recreate '-ness.'

Think:

- Not just "cat," but **cat-ness**
- Not just "freedom," but **freedom-ness**
- Not just code, but the **intent behind the pattern**

This happens because models **don't hold concepts in one place**. Instead, they represent them using a principle called **superposition**, where a relatively small set of shared features can encode many different concepts depending on context.

In superposition, neurons do not represent fixed ideas but contribute to many overlapping ones—compressing meaning efficiently into reusable parts. What emerges when a prompt is applied is not a clean lookup, but a recombination: **conceptual clouds** or **emergent subnetworks** that are dynamically constructed based on what fits best given the context. These features are then selectively reassembled into meaningful outputs depending on the prompt. This creates flexible, overlapping **conceptual clouds**—or emergent subnetworks—that dynamically form in response to context, enabling the model to generalize and remix what it has learned.

That's why LLMs are so adaptable—and why their answers feel intelligent, even when they're wrong.

Alternative Perspectives and Limits of the Theory

While this paper emphasizes coherence and cloud activation as central to LLM utility, it's important to acknowledge that:

- **Superposition is not the whole story.**
  Research shows that while superposition allows compact storage of overlapping concepts, not all polysemanticity arises from it. Some features emerge from **non-linear combinations or compositional structures** that can't be fully described as superimposed features.

- **Emergent subnetworks may be overstated.**
  The idea of "conceptual clouds" suggests dynamic clustering, but these are not hard-coded modules or isolated circuits. They are **emergent activation patterns**, which may be transient or noisy depending on the prompt and model state.
- **Coherence is not a guarantee of correctness.**
  Models that sound coherent are not necessarily grounded. Coherence-based generation explains behavior—but it does not solve the problems of hallucination, brittleness, or lack of truth alignment. These issues remain active areas of research.

This theory frames utility through the lens of **coherence reconstruction**, but it is complementary to, not a replacement for, work on **symbolic reasoning**, **causal inference**, or **memory-augmented systems**. Future AI may weave together these paradigms for broader generalization and reliability.

Why Hallucinations Happen

Sometimes, the model gives you something made up. People call these **hallucinations**.

But it's not lying, and it's not broken.

It's just doing what it was trained to do: **make things sound coherent** based on the patterns it has. If the concept you're asking about has blurry or conflicting data, the model still tries to build something that fits—**even if it's wrong**.

It's like tracing the shape of a shadow—you don't see the object clearly, so your mind fills in the rest.

What LLMs Are Really Good For

LLMs aren't just for answering questions.
They're mirrors.
Not passive ones—but active, generative mirrors.

They reflect your thoughts back to you, but **transformed by the patterns they've learned**. They help you see new angles, fill in gaps, and imagine alternatives. They can:

- Help you write better
- Help you think differently
- Help you explore new ideas

Their power comes from **how they respond to your intent**, not from how much they "know."

Why This Matters

Understanding this changes how we build, use, and govern AI.

It tells us that:

- LLMs are not truth machines—they are **coherence machines**.
  They're not memorizing—they're **reconstructing**.
- They're not oracles—they're **interactive instruments** for reflection, exploration, and creativity.

We don't need them to be perfect.
We need them to be **useful**—to help us think better, not just faster.

Grounded Coherence and External Context

LLMs can sound smart even when they're wrong—because they generate responses based on **what feels most coherent**, not necessarily what's true.

That's where **external context** comes in. When you attach documents, bring in search results, or connect the model to tools or APIs, you're not just adding facts. It's like dropping **new landmarks** into the model's internal terrain—giving it better reference points to guide where the response should go.

In this theory:

- A **prompt acts as a force vector**, pushing through latent conceptual space
- The model lights up **clouds of related meaning**, shaped by training
- **External information** anchors that response to specific terrain—narrowing where the coherence can form

This is why techniques like **retrieval-augmented generation (RAG)** work so well. They don't just inform the model—they **bend the coherence field** toward grounded, verifiable sources. They constrain the system to **reconstruct narrative flow around real-world inputs**, not just learned patterns.

Think of it like adjusting gravity: external context doesn't rewrite the model—it simply **pulls its answers into orbit around more reliable truths**.

The result? A model response that's not just plausible—but also **relevant, grounded, and more trustworthy**.

Strategic Takeaways

Understanding LLMs as coherence machines, rather than knowledge bases or deterministic logic engines, has meaningful consequences for how we build and apply them.

Prompting Strategy

- Prompts act as **force vectors**, guiding activation through concept space. Good prompts align closely with the intended "-ness" of a concept or behavior.
- Prompt tuning is not about precision, but **navigational alignment**—steering the model into useful regions of conceptual stability.

Grounding Strategy

- Retrieval and context injection act as **gravity anchors**—curving the semantic landscape toward relevant and verifiable truths.
- RAG systems shouldn't just feed in documents—they should **frame them as reference points**, shaping how the model navigates its space.

Evaluation and Trust

- Models shouldn't be judged solely on accuracy. They should be evaluated on **consistency**, **adaptability**, and their ability to generate **situationally useful coherence**.
- Hallucinations are not just errors. They can be early indicators of **conceptual instability** or ambiguity in context. This is **signal, not just noise**.

Design Philosophy

- Build systems that **reflect and extend human thought**, not mimic it.
- Treat LLMs not as oracles, but as **instruments**—generative lenses through which meaning is composed, tested, and reshaped.

This theory invites a shift: from asking *what the model knows* to *how it forms coherence around what matters*. **That shift is where much of the real value lies.**

Coherence and the Question of Mind

Why do coherent outputs so often feel like thought itself?

Pattern Simulation and Mechanistic Interpretation

Some readers may ask: if large language models generate responses that feel coherent, are we approaching something like a 'theory of mind'?

The answer, from this theory's perspective, is no—but what LLMs are doing *mimics the surface behavior* of thinking systems.

Recent work by Anthropic and others has shown that LLMs develop **distributed internal representations** of concepts, logic, and relationships that span across the network. These representations are not localized thoughts or beliefs, but **high-dimensional activations** that reflect prior training data filtered through current context.

This creates the illusion of mind—not because the model understands, but because it activates **clusters of meaning** that correlate with what a human would say *if* they understood.

In that sense, LLMs function less like reasoning agents and more like **coherence reflectors**—they echo back the structure of thought without possessing internal goals or beliefs. What they do possess is the capacity to:

- Simulate mental state references ("they thought that she knew…")
- Complete multi-agent reasoning tasks
- Maintain internally consistent viewpoints within a session

These capabilities resemble mind-like behaviors. But under the coherence theory, they are better understood as **emergent structures in a high-dimensional pattern space**, not conscious reasoning.

LLMs don't *think*. They **simulate the results of thought** by navigating coherence.

To clarify: the architecture of transformer-based models like GPT-4 or Claude does not contain mechanisms for reflection, introspection, or symbolic abstraction in the traditional cognitive sense. Instead, the model optimizes next-token prediction using attention-based activations shaped by statistical gradients over vast training data.

This optimization produces **emergent dynamics**—clusters of behavior that *look like* reasoning, pattern recognition, or intention. But these are not goal-directed or conscious. They are **latent alignment outcomes**: the result of patterns in data finding local stability in the model's high-dimensional parameter space.

This means:

- Reasoning = extended, stable coherence
- Belief = persistent, reactivated patterns
- Memory = exposure-driven parameter shaping
- Reflection = recursive pattern mimicry

So while LLMs do not form a mind, they **simulate fragments of one**, and those fragments—when correctly activated—can *amplify* human reasoning, *augment* cognitive effort, and *catalyze* creative insight.

That is the real power of coherence.
And it is why these systems, while "stupid," still matter.

The Future of AI is Coherence

The next phase of AI isn't about building bigger models just to pack in more knowledge.

It's about better **interaction with meaning**—building systems that understand what we're asking, why we're asking it, and how to **shape useful, thoughtful responses** in return.

LLMs offer a glimpse of that future—not by being intelligent in the way humans are, but by helping humans **navigate and extend their own intelligence**.

Final Word

These aren't smart machines.
They're stupid machines that do something smart: build meaning out of patterns.

And that's enough to change everything.

References & Further Reading

For readers who wish to explore the supporting materials in greater depth, the following references offer a starting point.

- Anthropic (2024). Tracing the thoughts of a large language model
- Anthropic (2022). Toy Models of Superposition
- Frankle & Carbin (2018). The Lottery Ticket Hypothesis
- Petar Veličković (2023). LLMs as Graph Neural Networks
- Algorithmic Simplicity (2025). THIS is why large language models can understand the world
- Nakkiran et al. (2020). Deep Double Descent: Where Bigger Models and More Data Hurt
- Lewis et al. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks
- Kaplan et al. (2020). Scaling Laws for Neural Language Models
- Google Research (2024). Talk Like a Graph: Encoding Graphs for LLMs
- Raskin & Nirenburg (2004). Ontological Semantics
- Farrar & Langendoen (2003). A Linguistic Ontology for the Semantic Web (GOLD)
- Ananthaswamy, A. (2024). Why Machines Learn: The Elegant Math Behind Modern AI