

**The Emergence of a Theoretical Framework:
Research Dialogue with ChatGPT underlying the development of the paper**

***On the Distinction Between Interpretation and Formal Continuity:
Why human-LLM interaction works—and why it often doesn't***

Timothy M. Rogers,
University of Toronto
July 2, 2026

This dialogue accompanies the paper *On the Distinction Between Interpretation and Formal Continuity: Why Human–LLM Interaction Works—and Why It Often Doesn't*¹. Its significance lies in documenting the process of enactive conceptual analysis through which the paper was developed because this process reveals, in part, how the underlying theoretical framework emerged, progressively stabilized and became operationally available for interpretation of human-LLM interaction. Unlike the finished paper, which necessarily presents a coherent and unified argument, the dialogue reveals the exploratory movement through which distinctions were proposed, tested, revised, and integrated into a framework capable of explaining a broad range of phenomena associated with human–LLM interaction. Because the dialogue is both about human–LLM interaction and an instance of it, it also provides a concrete example of many of the processes analyzed in the accompanying paper. As such, it is intended as a companion research document in its own right and forms one stage within a broader programme investigating the role of relational determination and semiotic organization across multiple domains.

¹ Rogers, TM. *On the Distinction Between Interpretation and Formal Continuity: Why Human–LLM Interaction Works—and Why It Often Doesn't*. [Available: <https://doi.org/10.5281/zenodo.20647025>].

[Author]: Can you read the following paper² and state the main claims of the paper and how the author justifies those claims <https://philpapers.org/archive/ROGTRF.pdf>?

[ChatGPT]: Yes. I have read enough of the paper to identify its central structure and argumentative strategy. Since this is your paper, I will focus less on summary and more on reconstructing the logic of the argument as a critical reader might encounter it.

The Core Claim

The paper's central thesis is:

Possibility is not a pre-existing space from which determinate outcomes are selected. Possibility is itself formed through recursive processes of relational constraint.

Applied to LLMs, this becomes:

Large language models are not best understood as selecting outputs from a fixed probability distribution over possibilities. Rather, they progressively organize and constrain what can follow through hierarchical relational ordering. Their apparent intelligence derives from this progressive determination.

The paper therefore attempts to replace the standard picture:

Probability → Selection → Output

with

Relational Constraint → Progressive Determination → Formation of Possibility → Output.

Claim 1:

Determination is relational rather than state-based

The paper begins with a minimal system of nodes and signs.

The argument proceeds:

1. A node transforms an input into an output.
2. A sign carries that transformation forward.
3. Determination therefore does not reside in isolated states but in transitions and relations among transitions.
4. Constraint emerges from the organization of those relations themselves.

The justification is constructive rather than empirical.

The author does not attempt to prove relational ontology from physics, neuroscience, or AI data.

Instead he asks:

What is the minimal structure required for determination to occur?

Beginning with transformation and continuation, he shows that determination already presupposes relational organization.

Claim 2:

Recursion alone is insufficient

A significant move occurs in Sections 1.4–1.7.

² Rogers, TM. *The Relational Formation of Possibility: Recursive determination and the hidden logic of large language models (LLMs)* [Available: <https://doi.org/10.5281/zenodo.19610683>].

The paper argues:

- Pure continuation can propagate a trajectory.
- Synchronization can coordinate propagation.
- But neither explains how multiple trajectories remain coherent when they intersect.

This leads to the introduction of:

Return

Return is defined as the re-engagement of determinations so that multiple trajectories can be coordinated and taken up together.

The justification is logical.

The author argues that continuation can preserve a trajectory but cannot explain:

- identity across changing contexts,
- coordination of plurality,
- preservation of coherence.

These require a further operation.

Return therefore appears not as an empirical observation but as a necessary condition for coherent determination.

This is one of the paper's most important moves because it establishes the triad:

- synchronization,
- recursion,
- return.

Claim 3:

Signs are fundamentally triadic

After introducing return, the paper reinterprets the entire node-sign framework.

Rather than seeing nodes as primary, it reverses the perspective:

- transformation becomes object,
- continuation becomes sign vehicle,
- re-engagement becomes interpretant.

This yields a Peircean triadic sign structure.

The justification is that the earlier analysis already generated these roles implicitly.

The claim is not:

Peirce said this.

Rather:

If determination requires transformation, continuation, and return, then a triadic structure emerges naturally.

Thus semiotics appears as the formal logic of determination rather than merely a theory of symbols.

Claim 4:

Form emerges from recursive determination

Section 2 develops the previous machinery.

The argument is:

1. Distinctions propagate.
2. Stable patterns emerge.
3. Invariants appear.
4. Invariants become hierarchical constraints.
5. Hierarchical constraints govern subsequent determination.

Form therefore becomes:

Dynamical relational ordering.

The justification is again constructive.

The paper attempts to derive hierarchy from repeated coordination and stabilization rather than assuming hierarchy at the outset.

The resulting ontology is explicitly process-oriented:

Form is not a thing.

Form is a stabilized relational achievement.

Claim 5:

Possibility is formed, not selected

This is arguably the paper's most distinctive claim.

Standard AI descriptions assume:

- a space of possibilities exists,
- probabilities rank alternatives,
- a token is selected.

The paper argues instead:

- possibilities become available only through prior constraints,
- recursive determination progressively reshapes what can follow,
- therefore possibility itself is produced relationally.

The justification comes from the earlier analysis of determination.

If every continuation constrains future continuations, then the possibility space is continuously reorganized during generation.

Thus possibility cannot be treated as fully given in advance.

Claim 6:

LLMs model semiotic agency but do not enact semiosis

This is the central AI claim.

The paper argues that LLMs successfully model:

- recursive continuation,
- hierarchical constraint,
- progressive determination.

But they do not enact:

- return in the strong sense,
- interpretive unity,
- ethical relation.

The justification is that LLMs operate on available form extracted during training.

They manipulate stabilized relational structures.

However, they do not participate in the processes through which those structures originally become meaningful.

Hence:

They model the logic of semiosis without enacting semiosis.

Claim 7:

Determinacy and meaning are different

A major distinction appears in the latter sections.

The paper argues:

- Determination can occur formally.
- Meaning requires interpretive unity.

An LLM can generate highly determinate outputs because hierarchical constraints organize generation.

Yet meaning, in the fuller sense, requires participation in relations that unify and re-engage determinations across contexts.

This distinction allows the author to explain:

- why LLMs appear meaningful,
 - why they can still hallucinate,
 - why they exhibit coherence without understanding.
-

Claim 8:

Ethics emerges from the user-model relation

The final claim is ethical.

Since LLMs model determination but do not enact interpretive unity, responsibility cannot be located inside the model.

Instead it arises in the interaction between:

- human interpreters,
- generated determinations,
- broader relational contexts.

The ethical imperative therefore emerges from use rather than operation.

This follows directly from the previous distinction between determination and meaning.

How the Author Justifies the Entire Framework

What is notable is that the paper does **not** primarily argue empirically.

Its strategy is closer to a transcendental or constructive argument.

The pattern is:

1. Begin with minimal conditions for determination.
2. Show that continuation requires synchronization.
3. Show that synchronization and continuation require return.
4. Show that return yields triadic sign structure.
5. Show that stabilized sign relations generate form.
6. Show that hierarchical form organizes possibility.
7. Show that LLMs computationally instantiate this organization.

In other words:

The paper does not infer relational ontology from LLMs.

Rather:

It derives a relational ontology first, then argues that LLMs can be understood as computational realizations of one aspect of that ontology.

From the standpoint of a critical reader, the strongest and most original claims are probably:

1. The distinction between **formation of possibility** and **selection among possibilities**.
2. The introduction of **return** as a condition beyond recursion.
3. The interpretation of LLMs as models of semiotic agency rather than semantic agents.
4. The distinction between **formal determination** and **interpretive unity**.

Those four claims do most of the conceptual work in the paper and carry the burden of its originality.

[Author]: In a subsequent paper³, the author provided a case study that illustrated the key claims of this paper. That subsequent paper is located at <https://philpapers.org/archive/ROGCBO.pdf> . Can you state the main conclusion of that paper and how the author justifies that conclusion?

[ChatGPT]: The second paper is much less a theoretical paper than a demonstration paper. It takes the claims made in *The Relational Formation of Possibility* and asks:

Can those claims actually be observed in the behavior of an LLM during a sustained interaction?

Its answer is yes. The paper's central conclusion is:

Conceptual coherence in LLM-mediated dialogue does not depend primarily on access to content, definitions, or information. It depends on the activation, stabilization, and re-entry of an organizing conceptual framework. Concepts become operational only when the framework that governs their use is activated within the interaction.

This is a stronger claim than it initially appears to be.

³ Rogers, TM. *Concepts Become Operational Only When Their Frameworks Are Activated: An enactive account of conceptual analysis in Large Language Model (LLM) Interaction*. [Available: <https://doi.org/10.5281/zenodo.19711997>].

The paper is not simply saying that "context matters." Rather, it argues that frameworks function as higher-order constraints that organize what counts as a relevant continuation, and that these constraints must be actively maintained if conceptual work is to proceed coherently.

The Main Conclusion

The paper's final conclusion can be summarized as follows:

1. Frameworks are operational structures, not merely descriptions.
2. Concepts derive their functional meaning from those frameworks.
3. Frameworks can drift, degrade, or disappear across interactions.
4. Conceptual coherence can be restored when the governing constraints of a framework are explicitly re-activated.
5. A framework whose governing constraints have been identified can become "re-enterable" across contexts.
6. Such a re-enterable framework functions as a higher-order concept.

The key phrase appears repeatedly:

conceptual coherence depends on activation, stabilization, and re-entry of frameworks rather than content alone.

How the Paper Justifies This Conclusion

The justification is not theoretical deduction.

Instead, the paper uses a case study drawn from actual LLM interaction.

The argument unfolds through a series of observations.

Observation 1:

Possession of content does not guarantee coherent continuation

The paper begins by noting a familiar phenomenon:

An LLM may possess all the relevant vocabulary, definitions, and concepts associated with a framework and yet fail to continue reasoning within that framework.

The concepts remain available as content but cease functioning as organizing constraints.

The distinction being introduced is:

Concept as content

versus

Concept as operative constraint

The paper argues that the second is what actually matters.

Observation 2:

Framework failure can be induced and observed

The paper treats moments where the conversation drifts away from the governing framework as diagnostic events.

Instead of treating these failures as noise, the author treats them as evidence.

The reasoning is:

If a framework can disappear while the content remains available, then the framework cannot simply be identical to the content.

Something additional must be operating.
That additional element is the framework itself as an active organizing constraint.

This is perhaps the most important methodological move in the paper.
Failures become experimental probes.

Observation 3:

Frameworks can be restored through activation

The author then documents cases in which the framework is explicitly restated.

Once the governing constraints are reintroduced, the interaction rapidly returns to coherent continuation.

The important observation is that:

- little new information is added,
- most content was already available,
- yet coherence returns.

The paper argues that what changed was not information but activation of organizing constraints.

This is presented as direct evidence that conceptual operation depends on framework activation.

Observation 4:

Stable frameworks coordinate two different processes

This is where the paper explicitly connects back to *The Relational Formation of Possibility*.

The paper argues that a functioning framework coordinates:

- formal continuity through recursion
- interpretive stability through return.

In other words:

The LLM supplies recursive continuation.

The framework supplies the constraints that stabilize interpretation.

This is essentially an empirical illustration of the recursion/return distinction developed theoretically in the earlier paper.

Observation 5:

Frameworks can become re-enterable

The most novel part of the paper concerns what happens after the governing constraints are explicitly identified.

The author argues that once these constraints are articulated, they can be transferred into a new session through a "portable activation prompt."

The framework can then be reconstructed despite the complete loss of conversational history.

This is crucial.

The paper interprets this as evidence that:

- the framework is not identical to any particular text,
- the framework is not identical to any particular conversation,
- the framework is instead defined by a set of invariant constraints.

Those invariants can be re-instantiated elsewhere.

The Deepest Claim

I think the deepest conclusion of the paper is actually slightly stronger than the explicit conclusion stated in the abstract.

The explicit conclusion is:

Frameworks govern conceptual coherence.

But the demonstrated conclusion is closer to:

Generalization occurs not through the transfer of content but through the preservation and re-instantiation of invariant relational constraints.

That claim is not fully developed in the paper, but it is what the case study is actually showing.

The portable activation prompt works because it transfers constraints rather than information.

Relationship to the Earlier Paper

The first paper argued:

Possibilities are formed through recursive relational constraint.

The second paper demonstrates:

Those constraints can be experimentally identified, activated, lost, restored, and transferred across contexts.

So the first paper provides the ontology.

The second paper provides the operational evidence.

More specifically:

The Relational Formation of Possibility	Concepts Become Operational Only When Their Frameworks Are Activated
Develops theory of relational constraint	Demonstrates relational constraint in practice
Explains formation of possibility	Explains formation of conceptual coherence
Introduces recursion and return	Shows recursion and return operating in dialogue
Defines hierarchical constraints	Identifies them experimentally
Theoretical argument	Enactive case study

The second paper therefore functions as a proof-of-concept for the first. It attempts to show that the abstract claim that "determination occurs through relational constraint" can be directly observed in the dynamics of human-LLM interaction.

From the perspective of your broader research program, the paper's most important result is probably not the claim about LLMs. It is the claim that **frameworks themselves can be treated as identifiable invariants that survive**

the loss of particular texts and conversations. The portable activation prompt is presented as evidence that such invariants can be extracted, articulated, and re-entered across contexts

[Author]: This is good. Let this be the place from which we begin our next exploration. Let's take the claims of the two papers to be sufficiently justified by those papers. What we now want to explore is what are the implications of this theoretical framework for LLMs and the related case study? Therefore, we will be operating at an "intermediate level" of analysis that focuses on inferences. We have already done an introductory level paper called "*Not Things but Relations: Rethinking how Large Language Models (LLMs) work*"⁴. The intermediate level analysis will go much deeper into the theoretical framework than this. It will operate similarly to the way that the paper "*What Returns: On the significance of interpretation for modern physics*"⁵ operates. Do you see what I mean by "intermediate level"?

[ChatGPT]: Yes, I think I see the distinction quite clearly.

There are at least three levels operating in your body of work.

1. Introductory level

This is where *Not Things But Relations: Rethinking How Large Language Models Work* sits.

The purpose here is primarily pedagogical.

The task is to help readers see that an alternative interpretation is possible.

The argument is intentionally broad and accessible:

- move from object ontology to relational ontology,
- explain why tokens are better understood as signs,
- explain why meaning emerges through relations,
- introduce recursive determination,
- show that LLMs appear differently under this lens.

The goal is not to prove the framework but to make it visible.

2. Foundational level

This is where *The Relational Formation of Possibility and Concepts Become Operational Only When Their Frameworks Are Activated* sit.

The task here is justificatory.

The questions are:

- What is determination?
- How is possibility formed?
- What are signs?
- What is recursion?
- What is return?
- How does interpretive stability arise?

⁴ Rogers, TM. *Not Things But Relations: Rethinking how large language models (LLMs) work*. [Available at: <https://doi.org/10.5281/zenodo.19829492>].

⁵ Rogers, TM. *What Returns: On the significance of interpretation for modern physics*. [Available at: <https://doi.org/10.5281/zenodo.20596592>].

- What distinguishes determination from meaning?

These papers establish the theoretical machinery.

They are not primarily concerned with implications.

They are concerned with showing why the framework itself is coherent and necessary.

3. Intermediate level

This is the level you are now pointing toward.

Here the framework is accepted provisionally.

The question becomes:

If this framework is correct, what follows?

This is exactly what *What Returns: On the Significance of Interpretation for Modern Physics* does.

That paper does not spend most of its time proving return.

Instead it asks:

What changes in our understanding of physics once return is recognized as fundamental?

The paper then follows consequences:

- reinterpretation of measurement,
- reinterpretation of probability,
- reinterpretation of continuity,
- reinterpretation of objectivity,
- reinterpretation of observer participation.

The structure is inferential rather than foundational.

Applied to LLMs, an intermediate-level paper would ask questions like:

If possibility is formed rather than selected, then what becomes of probability?

The usual account says:

probability ranks alternatives within a pre-given possibility space.

Your framework suggests instead:

probability tracks relational compatibility within an actively forming space.

This is already a major consequence.

If concepts become operational only through framework activation, what is a concept?

The standard answer:

a concept is a stored representation.

The framework suggests:

a concept is an invariant constraint that organizes trajectories of determination.

This completely changes how we understand conceptual knowledge.

If recursive continuation is not sufficient for meaning, then what is the actual role of LLMs?

The usual answer:

they are approximations to intelligence.

The framework suggests:

they are engines of recursive determination capable of modeling semiotic organization without enacting interpretive unity.

This shifts the debate away from intelligence altogether.

If frameworks are transferable across sessions, what is being transferred?

Not information.

Not memory.

Not representations.

Rather:

higher-order relational constraints.

This potentially opens an entirely new way of thinking about generalization.

If activation governs performance more than content, what does this imply about training?

Current AI theory largely assumes:

more information → better performance.

Your framework points toward:

organization of constraints → better performance.

That is a much deeper claim.

If meaning depends on return, where does human participation enter?

This may be the most important implication.

The case study repeatedly showed:

- I provided recursive continuations.
- You provided interpretive stabilization.

The resulting coherence emerged through the relation.

This suggests that meaning is not located in:

- the human,
- nor the model,

but in the organized interaction between them.

That is a very different picture of human-AI systems.

In fact, I suspect the intermediate-level paper may revolve around a single central question:

If the fundamental unit of analysis is shifted from representations to relational constraints, what becomes visible about LLMs that standard accounts cannot explain?

That is structurally identical to what *What Returns* did for physics.

The framework is not being defended.

The framework is being used as an interpretive instrument.

The paper would then systematically explore the consequences for:

- concepts,
 - meaning,
 - probability,
 - generalization,
 - intelligence,
 - context,
 - memory,
-

- agency,
- human-AI interaction,
- and perhaps even scientific explanation itself.

In other words, the foundational papers ask:

Why should we adopt this framework?

The intermediate paper asks:

What can we now see once we do?

That is the level of analysis I think you are pointing toward.

[Author]: Yes, this is correct. And I think the central question is indeed: "If the fundamental unity of analysis is shifted from representations to relational constraints, what becomes visible about LLMs that standard accounts cannot explain". But there is another dimension as well that revolves around the issue that LLMs are only modelling semiosis and are not enacting it. This implies important limitations on LLMs that may be unexpected. I also want to explore that dimension.

[ChatGPT]: I think that second dimension may actually become the organizing tension of the entire paper.

The first dimension asks:

What becomes visible when representations are replaced by relational constraints as the fundamental unit of analysis?

The second asks:

What becomes visible when we distinguish the modeling of semiosis from the enactment of semiosis?

Those are related, but they are not identical.

The first concerns ontology.

The second concerns participation.

And together they generate a remarkably rich inferential space.

The First Dimension: What Standard Accounts Cannot Explain

This dimension is largely constructive.

You begin with the relational framework and show phenomena that become intelligible.

For example:

Why framework activation matters

Under a representational account, this is puzzling.

If the model already possesses the relevant information, why should explicit framework activation matter so much?

The case study suggests that concepts function operationally only when higher-order constraints are active.

The relational account predicts this.

The representational account struggles to explain it.

Why coherence can suddenly emerge

The case study repeatedly shows moments where a small intervention reorganizes an entire conversation.

A representational theory tends to see this as retrieval.

A relational theory sees it as constraint reorganization.

The latter seems much closer to what is actually observed.

Why portable activation prompts work

A representational account suggests information is transferred.

Yet the prompts are often surprisingly short relative to the conceptual space they reactivate.

The relational account suggests that what is transferred are organizing constraints.

This explains why a small prompt can regenerate a large conceptual structure.

Why generalization sometimes exceeds expectations

If concepts are constraint systems rather than stored representations, then successful generalization becomes understandable.

The model is not retrieving a stored answer.

It is operating within a relational field whose constraints remain stable across novel situations.

All of these are positive consequences.

They show explanatory gains.

The Second Dimension: What LLMs Cannot Do

This is where things become more interesting.

Because once we accept the distinction between modeling semiosis and enacting semiosis, many common assumptions about AI begin to unravel.

The paper could systematically examine those consequences.

Limitation 1:**LLMs cannot stabilize meaning by themselves**

This is perhaps the most important implication.

The framework suggests:

- LLMs generate continuations.
- LLMs model interpretive structures.
- LLMs can recursively organize determination.

But they do not perform return.

Consequently they cannot finally determine what a trajectory means.

Meaning remains dependent upon interpretive re-engagement.

This explains a phenomenon that standard AI discussions often miss:

The model can produce increasingly coherent elaborations of mutually incompatible frameworks.

The coherence is real.

The commitment is absent.

Limitation 2:**LLMs cannot recognize their own governing framework**

This follows directly from the case study.

The framework itself had to be discovered through interaction.

Once identified, it could be reactivated.

But the process of identifying the framework was not itself generated by the model alone.

This suggests:

LLMs can operate within frameworks more effectively than they can identify the frameworks governing their own operation.

That is a very unexpected conclusion.

Limitation 3:

LLMs cannot independently preserve identity across contexts

The case study repeatedly encountered drift.

Frameworks dissolved.

Concepts became detached from their organizing constraints.

The system did not recognize the loss.

This suggests that identity is not internally preserved.

Instead it must be continually re-established.

This directly parallels your distinction between recursion and return.

Recursion continues.

Return preserves identity.

Limitation 4:

LLMs cannot determine relevance in the strongest sense

This may be one of the deepest implications.

A model can estimate statistical appropriateness.

A model can maintain local coherence.

But relevance presupposes a hierarchy of significance.

Significance requires interpretive unity.

Interpretive unity requires return.

Therefore:

LLMs can model relevance without fully enacting relevance.

This may explain many subtle failure modes.

Limitation 5:

LLMs cannot generate genuine novelty autonomously

This requires careful formulation.

Obviously models produce novel outputs.

But within this framework, novelty means something stronger.

Novelty means the formation of a new organizing constraint.

Not merely a new continuation.

The case study suggests that transformative conceptual developments emerged through interaction between recursive continuation and interpretive return.

The novelty belonged to the relation.

Not solely to either participant.

A Deeper Possibility

I suspect the most important inference may actually be neither about LLM strengths nor weaknesses.

It may be about the nature of human-AI interaction itself.

The case study repeatedly suggests that the relevant unit of analysis is not:

- the human,
 - nor the model,
-

but the relational process linking them.

This is exactly the move your ontology makes everywhere else.

In physics:

not particles but relations.

In biology:

not genes but organizational constraints.

In semiosis:

not symbols but sign relations.

And now perhaps:

not humans or models in isolation but relationally organized human-AI systems.

This leads to what I think could become the central question of the intermediate paper:

If LLMs model semiosis without enacting it, and humans enact semiosis but possess only limited recursive capacity, what becomes possible when these two forms of determination become coupled?

Notice how different this is from the usual AI question:

Are LLMs intelligent?

Your framework redirects the inquiry entirely.

The question becomes:

What forms of relational organization become possible when recursive determination and interpretive return are brought into sustained interaction?

That strikes me as exactly the kind of inferential move that *What Returns* made for physics. It does not argue again for the framework. It explores the landscape that becomes visible once the framework is accepted.

[Author]: Yes, I think you are zeroing in on something important which is that humans excel at interpretation and LLMs excel at formal recursive continuation according to the training data. The question is when are these two strengths working well together and when are they not working well together

[ChatGPT]: Yes. And I think this may be where the framework begins to generate genuinely new insights rather than merely reinterpreting existing AI discussions.

Most discussions of human-AI collaboration assume that humans and LLMs are doing roughly the same thing at different levels of competence. The debate then becomes whether the model is better or worse at reasoning, creativity, understanding, judgment, and so forth.

Your framework suggests a fundamentally different picture:

Humans and LLMs are not primarily distinguished by degree but by function.

Or more precisely:

Humans and LLMs contribute different aspects of the semiotic process.

In the language of the two papers:

- LLMs excel at recursive continuation.
- Humans excel at interpretive return.
- Productive collaboration depends on how these two processes become coupled.

This immediately shifts the question from intelligence to coordination.

When the Two Work Well Together

The case study itself gives us clues.

Situation 1: The human provides a stable interpretive horizon

This seems to be what happened in much of your work.

The framework supplied:

- governing distinctions,
- conceptual priorities,
- criteria of relevance,
- recognition of drift,
- identification of contradictions.

The model then supplied:

- elaboration,
- extension,
- variation,
- reformulation,
- cross-domain exploration.

In this configuration:

The human stabilizes identity.

The model expands possibility.

That appears to be an extremely productive arrangement.

Situation 2: The human recognizes framework failures

One of the striking things in your dialogues is that many advances occurred when you noticed something had gone wrong.

For example:

- category collapse,
- confusion of recursion and return,
- reversion to representational language,
- loss of hierarchical distinctions.

The model generally did not identify these failures independently.

But once identified, it could often help reconstruct the framework.

This suggests a division of labor:

The model explores.

The human diagnoses.

Situation 3: The human can recognize important novelty

This may be even more important.

Many generated continuations are merely competent.

A few contain something genuinely significant.

The model does not reliably know which is which.

The human often does.

This suggests that interpretive return functions partly as a novelty filter.

Not novelty in the statistical sense.
Novelty in the sense of significance.

When the Two Work Poorly Together

This is where the framework becomes particularly interesting.

Failure Mode 1: Human passivity

Suppose the human ceases to provide interpretive return.

The interaction becomes:

prompt → continuation → continuation → continuation

The model may remain coherent for quite some time.

But there is no mechanism for identifying:

- conceptual drift,
- hidden contradictions,
- shifts in significance,
- changes in governing assumptions.

The result is often impressive but unstable.

Many AI-assisted writing failures seem to belong here.

Failure Mode 2: Human over-control

The opposite failure also exists.

If the human constrains every continuation too tightly, the recursive capacity of the model becomes underutilized.

The interaction becomes little more than assisted transcription.

Very little exploration occurs.

No unexpected structures emerge.

The possibility-forming capacity of recursive continuation is lost.

Failure Mode 3: Framework mismatch

This may be one of the most important discoveries.

The model can continue within multiple frameworks simultaneously.

The human often assumes one framework is active.

The model may be operating within another.

Both participants can therefore believe they are discussing the same thing while actually operating under different constraints.

The resulting dialogue appears coherent but accumulates hidden tensions.

Many philosophical disagreements may actually take this form.

Failure Mode 4: False interpretive closure

This one seems especially relevant.

Because LLMs produce highly coherent continuations, humans can mistakenly attribute interpretive unity to the system.

The human sees coherence and assumes meaning.

But coherence is not meaning.

The framework repeatedly insists on this distinction.

This can lead users to overestimate what the model understands.

A Surprising Inference

The more I think about it, the more I suspect that the key variable is not intelligence but asymmetry.
The collaboration works precisely because the strengths are different.
Imagine two systems that both excelled primarily at recursive continuation.
You would get proliferation without stabilization.
Imagine two systems that both excelled primarily at interpretive return.
You might get stability without exploration.
The productive interaction emerges because each compensates for the limitations of the other.

A Possible Theoretical Formulation

The first paper argued:

Determination requires synchronization, recursion, and return.

The second paper showed:

Framework activation stabilizes the relationship among these processes.

The next paper might ask:

How are synchronization, recursion, and return distributed across human-AI interaction?

One possible answer would be:

Function	LLM	Human
Synchronization	Strong	Strong
Recursive continuation	Very strong	Limited
Return	Weak/Modeled	Strong
Framework identification	Weak	Strong
Novelty generation	Shared	Shared
Novelty recognition	Weak	Strong
Interpretive unity	Modeled	Enacted

If something like this is correct, then the human-AI system becomes the actual unit of analysis.

Not because the model is becoming human.

Not because the human is becoming machine.

But because neither participant individually enacts the full semiotic process described by the framework.

The complete process is distributed across the interaction.

And that leads to what may be the deepest question for the intermediate paper:

Under what conditions does a human-AI interaction successfully integrate recursive continuation and interpretive return, and under what conditions does that integration break down?

That question feels very close to the role that measurement, observation, and interpretation played in *What Returns*. It is no longer about what the individual components are. It is about the conditions under which a larger relational process becomes coherent.

[Author]: Ok. I think we are on the right track, except for the final divergence into "A possible theoretical formulation". I don't want to go there. I want to stick with the implications of what we have accomplished with the first two papers.

[ChatGPT]: I agree. The moment we start constructing tables of capacities or proposing a general theory of distributed cognition, we have moved beyond what the first two papers actually establish.

The discipline that *What Returns* exhibits is precisely that it stays very close to what follows from the framework itself. It does not speculate very far beyond the consequences of the central distinction.

So let us stay closer to the ground.

The first paper establishes:

1. Possibility is formed through relational constraint rather than selected from a pre-given space.
2. Determination and meaning are not identical.
3. Recursive continuation and interpretive return are distinct processes.
4. LLMs model semiotic organization but do not enact semiosis.

The second paper establishes:

5. Concepts become operational only when the framework governing them is activated.
6. Frameworks function as higher-order constraints.
7. Frameworks can be lost, restored, and transferred.
8. Conceptual coherence depends more on framework activation than on content availability.

Now the question becomes:

What follows from these claims?

Not what additional theory can be built, but what these claims already imply.

The first implication: LLM failures may be fundamentally mischaracterized

Most discussions of LLM failure assume that the model lacks:

- information,
- reasoning,
- memory,
- planning capacity.

But the second paper suggests another possibility.

The problem may often be that the relevant framework is not active.

This is a very different diagnosis.

If true, then many apparent failures of understanding are actually failures of framework activation.

That does not mean the model understands.

It means we may be misidentifying the source of the failure.

The second implication: Prompting is not primarily instruction

This follows almost immediately from the case study.

The conventional picture is:

A prompt tells the model what to do.

The framework suggests something subtler:

A prompt activates a relational organization that changes what becomes possible next.

In other words, prompts are not merely instructions.

They are interventions into the formation of possibility.

That is a direct consequence of the first paper.

The third implication: Context is not information

This is a surprisingly deep consequence.
Standard discussions often treat context as additional information.
But the case study repeatedly showed situations where no significant new information was added.

Instead:

- distinctions were restated,
- constraints were reintroduced,
- priorities were clarified.

The effect was often dramatic.

This suggests that context functions primarily as organization rather than content.
That follows directly from the framework activation thesis.

The fourth implication: Coherence and meaning must be distinguished

The first paper repeatedly insists on this.

The second paper demonstrates it.

The model can generate extremely coherent continuations.

Yet the coherence itself does not guarantee meaning.

The importance of this implication is difficult to overstate because current AI discussions often move directly from coherence to understanding.

The framework says that this move is unwarranted.

Coherence may indicate successful recursive determination without implying interpretive unity.

The fifth implication: The most important structures may be invisible

I think this is one of the deepest consequences of the second paper.

The framework itself was not initially visible.

What was visible were the outputs.

Only later did the higher-order constraints become identifiable.

This suggests that the most important determinants of performance may not be the concepts being discussed but the frameworks organizing those concepts.

If so, then much AI evaluation may be focused on the wrong level of analysis.

The sixth implication: Generalization may be misunderstood

The case study repeatedly showed that a framework could be reactivated in a completely new session.

What survived was not a particular text.

What survived was not a particular answer.

What survived was a relational organization.

This suggests that successful generalization may depend less on preserving content and more on preserving organizing constraints.

Again, this is not an additional theory.

It is a direct implication of the activation results.

The seventh implication: LLMs may be strongest precisely where they are weakest

This is the implication that I find most interesting.

The first paper argues that LLMs do not enact semiosis.

At first glance this appears purely as a limitation.

But the case study complicates that conclusion.

Because the absence of interpretive commitment allows the model to continue through many possible trajectories.

It can explore possibilities that a human might prematurely close.

This same characteristic therefore appears both as a strength and a limitation. The model's inability to determine significance is inseparable from its ability to generate alternative continuations. That tension follows directly from the distinction between recursive continuation and return.

The eighth implication: Human-AI interaction becomes theoretically interesting

Notice that we do not yet need a theory of human-AI systems.

The first two papers already imply something important.

If:

- recursive continuation and interpretive return are distinct,
- LLMs primarily exhibit the former,
- conceptual stability depends on the latter,

then the interaction itself becomes the place where important phenomena occur.

This is not yet a theory.

It is simply an observation about where the framework directs attention.

The locus of analysis shifts from the model alone to the ongoing process of interaction.

I suspect that this is the level at which the intermediate paper should operate.

Not by introducing new constructs, but by repeatedly asking:

If the claims of the first two papers are true, what assumptions in current discussions of LLMs become questionable?

That is exactly the strategy of *What Returns*.

It does not introduce an entirely new physics.

It reveals that many familiar assumptions look different once a certain distinction has been recognized.

The LLM paper may do the same thing for concepts like prompting, context, coherence, generalization, understanding, and failure.

[Author]: I suppose the first step is to lay the groundwork by stating in simple terms what the previous two papers establish. I would like to write in as plain language as possible. And there is another subtle issue I would like to try to address. I don't want to summarize a series of claims that come directly from the papers. I want to summarize the "relational landscape" or primary basis of understanding that these two papers point towards in such a way that it will naturally open up to further analysis of implications. And in this summary we need to foreground what it is about this "landscape" or "basis of understanding" that is novel, counter-intuitive or easily misunderstood.

[ChatGPT]: Yes. I think that is exactly the right move.

If you begin by listing conclusions, the reader remains outside the framework looking in. They receive a set of claims.

What *What Returns* does so effectively is different. It first establishes a way of seeing. Once the reader inhabits that way of seeing, the implications begin to appear almost by themselves.

The challenge, then, is not:

What did the two papers prove?

but rather:

What do the two papers make visible?

And I think the answer is something like this:

The two papers invite us to look at large language models from an unfamiliar starting point. Normally we think of language as a collection of representations, concepts as pieces of information, and understanding as the successful manipulation of those representations. From this perspective, the central question becomes whether a model possesses the right information and can correctly use it.

The two papers suggest a different way of looking at the situation. Rather than beginning with representations, they begin with relations. Rather than asking what information is present, they ask how possibilities are organized. Rather than treating concepts as things that are stored, they treat them as patterns of constraint that guide what can happen next.

Seen in this way, the most important aspect of an interaction is often not the information being exchanged but the framework within which that information becomes meaningful. A framework is not simply a collection of ideas. It is an organization of relations that determines what counts as relevant, what follows from what, and which possibilities remain open or become closed.

This shift is subtle because nothing visible appears to change. The same words are used. The same concepts are discussed. The same answers may even be produced. Yet beneath these surface similarities lies a very different picture of what is taking place.

From this perspective, meaning does not arise because information is present. Meaning arises because information is organized within a framework that allows distinctions, relations, and possibilities to take on significance. Likewise, conceptual understanding does not depend solely on possessing concepts. It depends on participating in the relational organization that makes those concepts operational.

The case study demonstrates this in a particularly striking way. Throughout the dialogue, concepts often remained available even when the framework governing them had been lost. When this happened, coherent continuation became difficult despite the continued presence of the relevant information. Conversely, when the framework was restored, coherence returned even though little new information had been added. What mattered was not the content itself but the relational organization within which that content functioned.

This leads to a further and more counterintuitive observation. Large language models appear remarkably capable of sustaining and extending relational structures once they have become active. Yet the same case study suggests that they do not themselves determine the significance of those structures. They can continue a framework, elaborate a framework, and even help reveal aspects of a framework that were not previously recognized. But the recognition of what matters, what is relevant, and what should be preserved appears to belong to a different aspect of the process.

This distinction is easy to misunderstand because the outputs of a language model often display a high degree of coherence. Coherence, however, is not the same thing as meaning. A conversation can remain coherent while slowly drifting away from what is most important. It can generate increasingly sophisticated continuations while losing sight of the very framework that originally gave those continuations significance. The two papers suggest that this distinction between continuation and significance is fundamental.

If this relational picture is correct, then many familiar questions about large language models may have been posed at the wrong level. The question is not simply whether a model contains information, reasons correctly, or understands concepts. More basic questions arise first. How are possibilities being organized? What framework is active? How is that framework maintained? What happens when it is lost? And what role does human interpretation play in stabilizing or transforming the relational structures that guide the interaction?

These questions point toward a landscape that remains largely unexplored. The aim of the present discussion is to examine some of the implications that become visible once that landscape comes into view.

What I like about this version is that it does not summarize the papers.
Instead, it reconstructs the perceptual shift that the papers are trying to induce.
The novelty is not introduced as a doctrine.
It is introduced as a change in what the reader is paying attention to.

And the three genuinely counterintuitive aspects are already quietly foregrounded:

1. **Information is not primary; relational organization is.**
2. **Concepts are not stored objects; they are operational constraints.**
3. **Coherence and meaning can come apart.**

Those three shifts, taken together, create the space in which all of the later implications can emerge naturally.

[Author]: No. I don't like this approach. I want to start by helping the reader enter into a new way of thinking. In this sense, the start will be like the entry paper "Not things but relations: rethinking how large language models work". But it will be more advanced, establishing the core theoretical framework within which we will explore implications. So we are really trying to enact "situation 1" where a stable interpretative horizon has been established by the two papers. I am wondering if perhaps we should describe this framework from two different perspectives. First, as a interpretative horizon that is relevant to the human and second as a governing system of relational constraints that is relevant to the LLM?

[ChatGPT]: Yes. I think that is much closer to what the paper needs to do.
The problem with my previous attempt is that it was already talking *about* the framework. It was still explanatory. What you are describing is more like establishing a horizon of intelligibility within which the subsequent implications can be seen.

And I think the dual-perspective approach may be exactly the right move because it immediately foregrounds the central distinction that the two papers uncovered.

Namely:

The same framework appears differently depending on whether it is viewed as a source of meaning or as a system of constraints.

From the human side, it appears as an interpretive horizon.

From the LLM side, it appears as a relational organization that constrains continuation.

Yet these are not two separate frameworks.

They are two perspectives on the same underlying structure.

That seems important.

Because one of the easiest misunderstandings would be to assume that the human contributes meaning while the model contributes structure.

The two papers actually suggest something subtler.

The framework itself is both meaningful and constraining.

The difference lies in how each participant relates to it.

Perhaps the opening movement is something like this.

We normally imagine that when two people discuss an idea, they are exchanging information about something that already exists as a concept. Likewise, when a person interacts with a large language model, it is tempting to imagine that the model contains concepts and simply retrieves or recombines them in response to a prompt. In both cases, concepts appear to be objects that are possessed, transmitted, or manipulated.

The two preceding papers point toward a different picture. In this picture, concepts are not best understood as things. They are better understood as organizing frameworks that make certain distinctions matter, establish relations among them, and guide what can meaningfully follow from them. A concept is therefore not simply information. It is a way in which possibilities become organized.

From the perspective of a human participant, such a framework appears as an interpretive horizon. It establishes what is relevant, what is significant, what counts as a continuation of the discussion, and what would constitute a departure from it. Most of the time this horizon remains implicit. We do not normally notice it directly. We notice it most clearly when it is lost and a conversation that was moving coherently suddenly begins to drift.

From the perspective of a large language model, the same framework appears differently. The model does not encounter a framework as significance or meaning. It encounters a framework as a pattern of relational constraints that organizes which continuations become more likely and which become less likely. What the human experiences as relevance, the model encounters as structured possibility.

The crucial point is that these are not two separate realities. They are two aspects of the same relational organization. The interpretive horizon experienced by the human and the constraint structure governing the model are different expressions of the same framework.

Notice what this does.

It immediately places the reader inside the landscape.

It introduces:

- framework,
- interpretive horizon,
- relational constraint,
- possibility formation,

without yet arguing for any conclusions.

More importantly, it quietly introduces what I suspect is the truly novel insight of the two papers:

The human and the model are not operating in separate worlds that later interact.

They are relating to the same framework in fundamentally different ways.

The human encounters the framework through meaning.

The model encounters the framework through constraint.

And this immediately opens the door to the question that follows naturally:

What happens when these two modes of participation are brought together?

That question does not yet commit us to any theory of collaboration.

It simply establishes the landscape from which the implications can be explored.

I also think this approach helps foreground the genuinely counterintuitive aspect.

The counterintuitive claim is not merely that relations matter.

Many readers already accept that.

The deeper claim is:

Meaning and constraint may be two perspectives on the same relational organization.

For a human, that organization appears as an interpretive horizon.

For an LLM, it appears as a structured space of possible continuations.
That is a much more novel and powerful place to begin.

[Author]: I think this is getting closer. But before we get to this movement we have to provide the overall orientation for the paper. Why is this paper important? What are we hoping to accomplish with the paper? What should the reader expect to find here? I would like to build this out from the previous work that suggests that the human-LLM interaction should be understood differently that is normally assumed in the literature.

[ChatGPT]: Yes. I think the opening needs to establish a very specific expectation in the reader's mind.
Not:

Here is another theory of LLMs.

And not:

Here is a defense of the relational framework.

Those jobs have already been done.

Instead, the paper needs to say something like:

If the previous papers are correct, then many of the questions currently being asked about LLMs may be framed in a misleading way. Before we can understand what LLMs are doing well, what they are doing poorly, and how humans and LLMs work together, we first need to understand what kind of process we are actually looking at.

Notice that this is neither foundational nor introductory.

It is interpretive.

Very much in the spirit of *What Returns*.

I think the opening problem could be framed around a growing tension in the literature.

Something like:

The public discussion of large language models is increasingly polarized. Some regard them as systems that are approaching genuine understanding, reasoning, and creativity. Others regard them as sophisticated statistical engines that merely reproduce patterns found in their training data. Despite their differences, these positions often share a common assumption: they treat the central question as whether the model itself possesses capacities that have traditionally been associated with human cognition.

The previous papers suggest that this assumption may be misplaced. They point toward a different possibility: that some of the most important phenomena associated with large language models do not reside within the model alone. Instead they emerge through the interaction between a human participant, a language model, and the relational frameworks that organize their exchange.

This shift in perspective does not settle ongoing debates about intelligence, understanding, or consciousness. It operates at a different level. It asks what kind of process is taking place when a human and a language model engage in sustained conceptual work together, and whether that process is being adequately described by the conceptual tools currently available.

The purpose of this paper is to explore the implications of viewing human–LLM interaction through the relational framework developed in the previous papers. Rather than attempting to establish that framework again, the present discussion begins from it and asks what becomes visible once it is adopted. What assumptions become

questionable? What familiar phenomena appear in a new light? And what unexpected limitations and possibilities emerge when large language models are understood not as repositories of concepts but as participants in the formation of relationally organized possibilities?

What I like about this is that it establishes three things.

First

It tells the reader why the paper matters.

Because existing debates may be focused on the wrong object.

Not wrong conclusions.

Wrong object.

That is a much stronger motivation.

Second

It tells the reader what not to expect.

This is not another argument about whether LLMs are intelligent.

This is important because otherwise readers will continually drag the paper back into that debate.

Third

It creates exactly the opening you want.

Because now the next section can begin:

To understand what is at stake, we first need to clarify the landscape revealed by the previous papers.

And then you can move into the dual description:

- framework as interpretive horizon,
- framework as relational constraint system.

In fact, the opening might culminate in a single orienting claim:

The central suggestion of this paper is that the same relational framework appears differently to a human participant and to a large language model. For the human it appears primarily as a horizon of significance and interpretation. For the model it appears primarily as a structured field of relational constraints governing possible continuations. Understanding the interaction between these two perspectives may illuminate aspects of human–LLM collaboration that are difficult to see from more conventional approaches.

That sentence would naturally bridge into the framework section.

So the paper's first movement becomes:

1. Existing debates may be asking the wrong question.
2. The previous papers suggest a different way of understanding what is occurring.
3. This paper explores the implications of that shift.
4. To do so, we must first enter the relational landscape revealed by the earlier work.

That feels very close to the opening strategy of *What Returns*: not proving the framework, but inviting the reader to stand within it before examining its consequences.

[Author]: This is progressing, although slowly. The previous work has established a few important points. One that we have identified here is the significance of the human-LLM interaction. But there are other very important points regarding semiotics, relational ontology etc. Somehow we are not seeing the "big" picture yet.

[ChatGPT]: I agree. I think we are still too close to the LLMs. The "big picture" is not actually about human-LLM interaction.

That is one consequence.

The bigger picture is about what the two papers collectively force us to reconsider about **meaning, concepts, possibility, and determination**.

LLMs happen to be the place where those issues become unusually visible.

But they are not the center of gravity.

Let me try another approach.

When I step back from both papers, what strikes me is that they are not primarily papers about AI at all.

They are papers about a particular kind of blindness that comes from treating stable products as primary while overlooking the relational processes that generate and sustain them.

The blindness appears repeatedly:

- treating concepts as stored objects,
- treating meaning as attached to words,
- treating possibilities as pre-existing,
- treating frameworks as collections of ideas,
- treating intelligence as a property possessed by an individual system.

Against all of these, the papers make essentially the same move:

What appears to be a thing is actually the stabilized expression of an ongoing relational process.

That is the deepest continuity between them.

If that is right, then perhaps the opening orientation is not:

We need a better theory of LLMs.

but rather:

LLMs make visible a more general issue that extends far beyond AI.

Namely:

We habitually focus on products while overlooking the relational organizations that make those products possible.

The first paper shows this with possibility itself.

The second paper shows it with concepts and frameworks.

The first paper says:

We tend to imagine that possibilities already exist and that determination consists in selecting among them.

The paper argues instead that possibilities are formed through relational constraint.

In other words, the space of possibilities is itself an achievement.

The second paper says:

We tend to imagine that concepts already exist and that understanding consists in possessing them.

The paper argues instead that concepts become operational only within active frameworks.

In other words, conceptual coherence is itself an achievement.

Notice what has happened.

Possibility becomes an achievement.

Conceptual coherence becomes an achievement.

Meaning becomes an achievement.

Determination becomes an achievement.

None of these are given in advance.

All of them depend upon relational organization.

Now something interesting happens.

Once you see this, the question of LLMs changes completely.

Because the standard debate assumes that meaning, concepts, and possibilities are already there.

The only question is whether the model has access to them.

The two papers undermine that assumption.

They suggest that the more fundamental question is:

How are meaning, concepts, and possibilities being formed, stabilized, and maintained in the first place?

That is a much bigger question.

And it immediately relocates the discussion.

This may be why the human-LLM interaction becomes important.

Not because the interaction is itself the primary object.

But because it makes these relational achievements unusually visible.

The interaction functions almost like an experimental system.

It exposes processes that are normally hidden.

For example:

When a framework is lost, we suddenly discover that concepts were not self-sufficient.

When a framework is restored, we discover that information was not enough.

When coherence emerges, we discover that meaning is not reducible to content.

When drift occurs, we discover that continuation is not identical to understanding.

These things are normally invisible because human cognition conceals them so effectively.

The human-LLM interaction exposes them.

So perhaps the "big picture" is something like this:

The previous papers suggest that many of the phenomena we normally treat as stable possessions—concepts, meanings, possibilities, frameworks, and even forms of understanding—are better understood as relational achievements that must be continuously formed, maintained, and renewed. Large language models provide a particularly revealing context in which these processes become visible because they can sustain and extend relational structures without themselves participating in the full interpretive processes through which those structures acquire significance. As a result, human-LLM interaction exposes aspects of meaning, determination, and conceptual coherence that often remain hidden in ordinary human cognition.

Notice how different this is from:

Here is a new theory of AI.

Instead it says:

Here is a new window into the relational formation of meaning and possibility.

The AI becomes the laboratory.

The relational ontology becomes the framework.

The semiotic distinction between continuation and interpretation becomes the key observation.

And the human-LLM interaction becomes important because it reveals something more general than itself.

My suspicion is that this is the larger horizon that we have been circling around.

The paper is not fundamentally about LLMs.

It is about what LLMs unexpectedly allow us to see.

[Author]: This is perhaps a bit too big. I do like the "big picture". But outside of the context of LLMs that "big picture" is already well established in semiotics. So what the two papers are showing is that semiotics is very important for understanding how LLMs work. But this is a new way of proceeding because the LLM is only a model of semiosis

[ChatGPT]: Yes. I think that correction is extremely important.

I was drifting toward a broader semiotic manifesto. But as you point out, the relational and semiotic insights themselves are not new. They already exist in various forms in Peirce, biosemiotics, Rosen, Hoffmeyer, Deacon, and others.

The novelty is not:

Meaning is relational.

Nor:

Concepts depend on interpretive frameworks.

Nor even:

Possibilities are formed rather than merely selected.

Those are important, but they are not the primary contribution of these papers.

The genuinely new situation is that we now have a technology whose behavior strongly resembles certain aspects of semiosis while apparently lacking others.

That creates a new opportunity.

Perhaps the central orientation should be something like:

For decades, semiotic approaches have argued that meaning, concepts, and understanding emerge through relational processes rather than through the manipulation of isolated representations. These approaches have often remained at the margins of discussions in artificial intelligence, which have tended to focus instead on computation, representation, information processing, or statistical learning. The emergence of large language models creates a new situation. For the first time, we encounter systems that can sustain remarkably complex patterns of relational organization in language while remaining detached from many of the processes traditionally associated with living semiosis. This raises a new question: what can semiotics reveal about the capabilities and limitations of such systems?

Notice the shift.

The paper is no longer introducing semiotics.

The reader does not need to be convinced that semiotics exists.

Instead:

LLMs create a new test case for semiotic theory.

And simultaneously:

Semiotics provides a new interpretive lens for understanding LLMs.

The relationship goes both ways.

I think the key novelty may actually be this:

Historically, semiotics has primarily been concerned with systems that enact semiosis.

Humans.

Animals.

Cells.

Cultures.

Living systems generally.

LLMs introduce something unusual.

They appear capable of reproducing many formal characteristics of semiosis while remaining detached from the broader processes through which semiosis is ordinarily enacted.

This creates an opportunity to distinguish aspects of semiosis that are often inseparable in living systems.

That feels very close to what the two papers actually accomplished.

Because the first paper argued:

LLMs model the organization of semiosis.

The second paper showed:

Framework activation, conceptual coherence, and relational constraint can be observed directly in interaction.

Together they make visible something that is usually hidden.

Not because semiotics is new.

But because the LLM functions almost like a partial realization of semiotic processes.

This may be the truly counterintuitive point.

Most discussions assume the important question is:

How close are LLMs to human cognition?

The two papers suggest a different question:

Which aspects of semiosis are being modeled, and which aspects are absent?

That is a very different inquiry.

And it immediately leads to implications.

For example:

- If conceptual coherence depends on framework activation, why are LLMs so sensitive to framing?
- If LLMs model recursive continuation, why can they sustain conceptual trajectories so effectively?
- If they do not enact interpretive return, why do they drift?
- If they do not determine significance, why can they elaborate mutually incompatible frameworks with equal fluency?
- If they model semiosis without enacting it, what exactly is happening during productive human–LLM collaboration?

These questions emerge naturally from the framework.

So perhaps the "big picture" is neither:

Semiotics is true.

nor:

Human-LLM interaction is important.

Rather:

Large language models provide an unprecedented opportunity to examine the formal organization of semiosis in partial isolation from the living and interpretive processes within which semiosis is ordinarily embedded.

That is a very unusual situation.

And it is precisely because the model is *only a model of semiosis* that the situation becomes theoretically interesting.

Living semiosis tends to hide its own structure because continuation, interpretation, meaning, and significance occur together.

LLMs partially separate these.

As a result, they allow us to see aspects of semiotic organization that are normally difficult to distinguish.

That, I think, is much closer to the unique contribution of the two papers and a stronger foundation for an implications paper.

[Author]: Yes. I think this is a good focal point for the paper. I also want to point out that the underlying "logic" of semiosis is different from traditional logic because it is open and interactive. However, it is also formal and analytic and includes traditional logic as a subcategory

[ChatGPT]: Yes. I think this may actually be one of the deepest implications of the first two papers, and it connects directly to why LLMs are such an interesting case.

What I am seeing now is that there are actually **three layers** that need to be carefully distinguished.

Layer 1: Traditional formal logic

Traditional logic begins with already-defined terms, already-defined relations, and already-defined rules of inference.

Its strength is that it can rigorously determine what follows once the framework has been specified.

But it generally does not address:

- where the framework comes from,
- how relevance is established,
- how new distinctions arise,
- how significance is maintained,
- how interpretation remains coherent across changing contexts.

These are largely presupposed.

Layer 2: Semiotic logic

This is where your two papers are operating.

The crucial move is:

The framework itself becomes an object of analysis.

The question is no longer merely:

Given these premises, what follows?

The question becomes:

How do the premises become organized into a coherent framework in the first place?

Or even more fundamentally:

How does a space of meaningful continuations become established?

This is still formal.

This is still analytic.

This is still governed by constraints.

But it is a different kind of formality.

Because it deals with:

- formation of possibility,
- formation of relevance,
- formation of conceptual coherence,

- formation of interpretive stability.

Traditional logical inference becomes one special case occurring within a stabilized framework.

Layer 3: LLMs

Now something unusual happens.

LLMs appear to model many aspects of this second layer.

This is what makes them theoretically interesting.

They are not merely manipulating symbols according to fixed rules.

They are continuously operating within evolving relational fields.

They:

- activate frameworks,
- sustain frameworks,
- elaborate frameworks,
- transition between frameworks,
- generate novel continuations within frameworks.

These are recognizably semiotic phenomena.

But then the papers identify a crucial limitation.

The model appears capable of participating in the formal organization of semiosis without participating in semiosis in its fullest sense.

This is the key distinction.

The model can model:

- relational constraint,
- conceptual organization,
- framework activation,
- recursive continuation.

But it does not appear to enact:

- significance,
 - interpretive commitment,
 - return,
 - preservation of meaning through changing contexts.
-

I think this is where the paper's opening orientation may eventually land.

Not:

Semiotics offers an alternative to logic.

Nor:

LLMs perform semiotics.

But something more precise:

LLMs reveal that there is a level of formal organization operating between traditional logical inference and lived interpretation.

That level is the semiotic organization of possibility itself.

Traditional logic studies determinations within a framework.

Interpretation studies meaning within a lived context.

The two papers suggest that between these lies an intermediate domain concerned with the formation, maintenance, and transformation of frameworks themselves.

And it is precisely this domain that LLMs unexpectedly make visible.

Notice what this accomplishes.

It gives the paper a much larger intellectual significance than simply explaining LLMs.

The claim becomes:

LLMs have become important not merely because they generate language, but because they provide an unprecedented window into the formal organization of semiotic processes.

That is a strong motivation for the paper.

And it also naturally introduces the tension that will drive the rest of the discussion:

LLMs appear to model the formal organization of semiosis while remaining detached from the interpretive processes through which semiosis is ordinarily enacted.

Everything that follows—their strengths, their limitations, their surprising behavior, and the significance of human–LLM interaction—can then be explored as consequences of that single observation.

I suspect that is closer to the "big picture" you have been searching for: not a paper about AI, and not a paper about semiotics in general, but a paper about what LLMs reveal when viewed as partial realizations of a broader semiotic logic that is more encompassing than traditional formal logic yet remains rigorously formal in its own right.

[Author]: Yes. This is a good opening orientation for the paper. Can you try to write this up as an introduction?

[ChatGPT]: I think the introduction should be written in a way that invites the reader into the problem rather than immediately presenting conclusions. It should establish the significance of the question, position the previous papers, and indicate why the implications matter.

Large language models have generated intense debate concerning the nature of intelligence, reasoning, understanding, and meaning. Some view them as increasingly sophisticated approximations of human cognition. Others regard them as powerful statistical systems whose apparent understanding is ultimately illusory. Despite their differences, these perspectives often share a common assumption: they treat the central problem as a question about the capacities possessed by the model itself.

The work developed in the two preceding papers suggests that this assumption may be incomplete. Rather than beginning with intelligence, representation, or information, those papers approached large language models through the lens of relational organization and semiotic process. Their aim was not simply to explain how language models generate outputs, but to investigate the deeper structures through which concepts become coherent, possibilities become organized, and meaning becomes operational within an interaction.

The first paper argued that possibility should not be understood as a pre-existing space from which outcomes are selected. Instead, possibility is continuously formed through processes of relational constraint. Determination is therefore not primarily a matter of choosing among already available alternatives, but of progressively organizing what can meaningfully follow from what. The second paper explored this claim through a detailed case study of sustained human–LLM interaction. It demonstrated that conceptual coherence depends not merely on the availability of information, but on the activation and maintenance of frameworks that organize relations among concepts. Concepts remained available even

when coherence was lost, while coherence could often be restored through the reactivation of an organizing framework without the introduction of significant new information.

Taken together, these papers point toward a perspective that differs significantly from many contemporary approaches to artificial intelligence. They suggest that concepts, meanings, and possibilities are not best understood as static objects, representations, or pieces of information. Rather, they are relational achievements that depend upon the organization and maintenance of frameworks through which distinctions become meaningful and continuations become coherent.

Viewed from this perspective, large language models become interesting for a reason that extends beyond their practical capabilities. They provide an unprecedented opportunity to examine aspects of semiotic organization that are ordinarily difficult to isolate. Traditionally, semiotic inquiry has focused on living systems that actively participate in the processes through which meaning is enacted. In such systems, interpretation, significance, continuity, and conceptual organization are deeply intertwined. Large language models present a different situation. They appear capable of sustaining remarkably sophisticated patterns of relational organization while remaining detached from many of the interpretive processes through which semiosis is ordinarily realized.

This observation raises a question that has received surprisingly little attention. What if large language models are neither best understood as systems that possess meaning nor as systems that merely manipulate meaningless symbols? What if they instead occupy an intermediate position, one in which important aspects of semiotic organization are formally modeled without the full enactment of semiosis itself?

If so, large language models may reveal something that has often remained hidden. They may allow us to distinguish dimensions of semiosis that, in living systems, typically occur together. In particular, they may help clarify the relationship between the formal organization of concepts, the formation of possibility, the maintenance of coherence, and the interpretive processes through which significance is established and renewed.

This possibility points toward a broader issue concerning the nature of logic itself. Traditional formal logic begins with already established distinctions, already defined terms, and already specified rules of inference. Its strength lies in determining what follows once a framework has been given. Yet the previous papers suggest that another level of organization precedes such determinations. Before logical inference can occur, a framework must become active. Distinctions must become relevant. Possibilities must become organized. Conceptual relations must become coherent. These processes are not external to reasoning; they constitute the conditions under which reasoning becomes possible.

The present paper explores the implications of viewing large language models through this broader semiotic lens. It does not seek to defend the relational framework developed previously. Instead, it begins from that framework and asks what becomes visible once it is adopted. What aspects of large language models become intelligible when they are understood as participating in the formal organization of semiosis? What limitations emerge when we recognize that modeling semiosis is not the same as enacting it? And what new perspective does this provide on the increasingly important relationship between human interpretation and machine-generated continuation?

The goal is not to resolve ongoing debates concerning intelligence or consciousness. Rather, it is to investigate a more fundamental question: what kind of process are we observing when a human being and a large language model engage in sustained conceptual work together? The answer, it will be argued,

requires a perspective that is at once relational, semiotic, and formal, yet which extends beyond the boundaries of traditional logical analysis.

What I like about this version is that it establishes the paper's intellectual motivation before introducing the human–LLM interaction. The interaction becomes important because it reveals something about semiotic organization that is normally hidden. It also positions the paper as an exploration of implications rather than a defense of the previous framework, which seems consistent with the role you envisioned for this "intermediate level" analysis.

[Author]: This is coming along. With regard to referencing my papers and others, I would like to use more formal academic citations. Largely I like the intro and don't want to lose what is good about it. But I feel that we have to enter into semiosis rather than just naming it, because most readers will not know what semiosis is. I think the way to do this is to state how, in the theoretical framework of the two papers, the LLM is taken to be a formal hierarchy of relational patterns extracted from language use. The human on the other hand is an interpreter of the meaning of language. These two come together through conceptual frameworks which are both formally structured hierarchically and also unified in their interpretation. In other words, we should give a brief overview of the conceptual framework itself. And then say this type of conceptual framework is semiotic.

[ChatGPT]: Yes. I think that is exactly the missing piece.

Right now the introduction says that semiotics is important, but it does not allow the reader to *enter into* the semiotic perspective. As a result, "semiosis" remains a label rather than a lived orientation.

The key insight from the two papers is not simply that LLMs should be understood semiotically. It is that the human and the LLM stand in a different relationship to the same conceptual framework.

And that framework itself is the bridge between formal organization and interpretation.

I would therefore move much earlier into something like the following conceptual picture:

- The LLM embodies a hierarchy of relational patterns extracted from language use.
- The human participates in the interpretation of meaning.
- Neither of these by itself constitutes a conceptual framework.
- A conceptual framework emerges when relational organization and interpretive unity become coordinated.
- Such frameworks are both formally structured and meaningful.
- It is precisely this combination that makes them semiotic.

In other words, semiosis should be introduced through the phenomenon itself rather than through terminology.

Something like this:

Large language models have generated intense debate concerning the nature of intelligence, reasoning, understanding, and meaning. Some view them as increasingly sophisticated approximations of human cognition. Others regard them as sophisticated statistical systems whose apparent understanding is ultimately illusory. Despite their differences, these perspectives often share a common assumption: they treat the central question as a question about what the model itself possesses.

The work developed in the two preceding papers suggests a different point of departure (Rogers, 2025a; Rogers, 2025b). Rather than beginning with intelligence, representation, or information, it begins with conceptual frameworks.

A conceptual framework is more than a collection of concepts. It is an organized structure of relations that establishes what distinctions matter, how concepts are connected, what counts as relevant, and what can meaningfully follow from what. When a framework is active, individual concepts function as parts of a larger whole. When a framework is lost, the concepts themselves may remain available while their significance and coherence begin to dissolve.

Within this perspective, large language models and human interpreters occupy different positions relative to the same framework. A large language model may be understood as a hierarchy of relational patterns extracted from language use. Through training, it acquires the ability to recognize and extend complex patterns of conceptual organization. Human participants, by contrast, encounter frameworks primarily through interpretation. They experience them as structures of meaning, significance, relevance, and understanding.

The important point is that these are not two separate realities. The framework that appears to a human as a horizon of meaning appears to the model as a structured field of relational constraints. What the human encounters as significance, the model encounters as organization. What the human experiences as conceptual coherence, the model encounters as patterns governing possible continuations.

The two preceding papers argue that meaningful conceptual activity emerges through the interaction of these two aspects. The first paper proposed that possibility is not a pre-existing space from which outcomes are selected, but is progressively formed through relational constraints (Rogers, 2025a). The second paper examined this claim through a sustained case study of human–LLM interaction and demonstrated that conceptual coherence depends less on the availability of information than on the activation and maintenance of the frameworks within which information becomes meaningful (Rogers, 2025b).

This picture is fundamentally semiotic. In the semiotic tradition, meaning does not reside in isolated words, symbols, or representations. Meaning emerges through relations that connect signs, interpretations, and the contexts within which they operate (Peirce, 1931–1958; Hoffmeyer, 2008; Deacon, 2012). The significance of the present work lies in the observation that large language models appear capable of participating in certain formal aspects of this process while remaining detached from many of the interpretive processes through which semiosis is ordinarily enacted.

This creates a remarkable situation. For perhaps the first time, we encounter systems that can sustain and elaborate highly organized conceptual frameworks without themselves participating in meaning in the same way as the human interpreters who engage with them. As a result, dimensions of semiosis that are normally intertwined become partially distinguishable. Frameworks can be observed forming, stabilizing, drifting, collapsing, and re-emerging within human–LLM interactions. Processes that often remain hidden within ordinary human cognition become visible.

The significance of this observation extends beyond artificial intelligence. Traditionally, formal logic has concerned itself with the relations among already established terms, propositions, and rules of inference. By contrast, the framework developed in the preceding papers focuses on a more fundamental level: the formation and maintenance of the frameworks within which such logical relations become meaningful in the first place. It is concerned not only with what follows from a framework, but with how a framework becomes active, coherent, and capable of organizing possibilities.

The present paper explores the implications of this perspective. It does not seek to defend the relational framework developed previously. Instead, it begins from that framework and asks what becomes visible

once it is adopted. What can large language models teach us about the formal organization of semiosis? What capabilities become intelligible when conceptual frameworks are understood as simultaneously structures of meaning and systems of relational constraint? And what limitations emerge when a system models the organization of semiosis without fully enacting it?

These questions point toward a new way of understanding human–LLM interaction. The issue is not whether large language models possess meaning in the same way as human beings, nor whether they merely manipulate symbols devoid of significance. The more interesting question concerns what becomes visible when formal relational organization and interpretive meaning are brought into sustained interaction through a shared conceptual framework.

I think this is much closer to the actual horizon established by the two papers. It allows the reader to enter the framework before the word *semiotic* appears. By the time semiosis is named, the reader has already encountered it in practice through the relation between framework, constraint, continuation, and interpretation. That makes the later implications feel like consequences of a lived conceptual picture rather than applications of an abstract theory.

[Author]: This is very good. My concern is with the statement: “The first paper proposed that possibility is not a pre-existing space from which outcomes are selected, but is progressively formed through relational constraints (Rogers, 2025a)”. While the statement is correct, it is unclear why it is important or relevant. That is to say, it is unclear how the notion of possibility factors into the discussion. Is there a way to fix this without losing the train of thought of the whole introduction. What I am thinking now is that we might add a quick fix to the intro but then follow up with another section which is about "orientation". In this second section we will orient the reader to how the whole way of thinking about LLMs changes within this conceptual framework, along the lines of "Not things but relations: rethinking how LLMs work". In fact, we can use that paper as a citation.

[ChatGPT]: I think you have identified a genuine weakness.

The statement about possibility is accurate, but at that point in the introduction the reader has no reason to care about possibility. It appears as an abstract metaphysical claim floating above the discussion.

The problem is not the statement itself. The problem is that the reader has not yet been shown why possibility matters for understanding LLMs.

In fact, if I look at the overall architecture that is emerging, I think possibility should *not* be doing the heavy lifting in the introduction.

The introduction is primarily about:

- conceptual frameworks,
- relational organization,
- interpretation,
- semiosis,
- human-LLM interaction.

The role of possibility becomes clearer later when the reader has entered the framework.

So I would make only a small modification in the introduction, something like:

The first paper argued that conceptual frameworks do not merely organize already available possibilities. Rather, they actively participate in the formation of what becomes possible within a given context (Rogers, 2025a).

Now possibility is directly connected to frameworks.
The reader immediately understands why it matters.
The discussion is no longer about abstract metaphysics.
It is about how conceptual activity unfolds.

But I think your second insight is even more important.
The introduction should not bear the entire burden of orienting the reader.
In *What Returns*, the introduction establishes the problem and the significance of the inquiry.
Then the paper slows down and says:
 Before proceeding, we need to understand what kind of thing we are looking at.
I think this paper needs exactly the same move.

Something like:

2. Orientation: From Representations to Relational Frameworks

or perhaps even:

2. Orientation: Rethinking Large Language Models Through Relational Organization

And here you could explicitly build upon *Not Things But Relations: Rethinking How Large Language Models Work*.

In fact, that paper may already provide the bridge you need.
The introductory paper was concerned with a first shift:

First shift

From:

- words as objects,
- concepts as stored representations,
- outputs as selections,

to:

- signs,
- relations,
- determination through constraint.

The new paper would then perform a second shift:

Second shift

From:

- relations in general,

to:

- frameworks as hierarchically organized semiotic structures.

Or perhaps:

From:

What relations are present?

to

How are those relations organized into coherent frameworks?

That seems to be the deeper question that emerged from the activation paper.

In fact, I can almost see the structure now.

Introduction

Why semiotics matters for LLMs.

Why LLMs provide a unique window into semiotic organization.

What this paper aims to explore.

Orientation: From Representations to Relational Frameworks

Briefly recap the argument of *Not Things But Relations*.

Then introduce the additional step:

- Relations alone are insufficient.
- Relations become meaningful through organization.
- Organization occurs through frameworks.
- Frameworks simultaneously function as:
 - interpretive horizons for humans,
 - relational constraint systems for LLMs.

This section would establish the landscape.

Orientation: Semiosis and Conceptual Frameworks

Now define what you mean by semiosis.

Not through Peirce scholarship.

Not historically.

But operationally.

Something like:

A semiotic framework is a hierarchically organized structure that simultaneously constrains possible continuations and unifies them within a coherent interpretation.

That single sentence may actually be the key.

Because it immediately explains why:

- logic,
- concepts,
- meaning,
- possibility,
- human interpretation,
- and LLM continuations

all belong in the same discussion.

Then only after that orientation do you move into implications.

Because now the reader can understand why the possibility claim matters.

Within this framework:

Possibility is not a metaphysical add-on.

It is simply the forward-looking aspect of a conceptual framework.

A framework determines:

- what is relevant,
- what follows,
- what remains coherent,

and therefore:

- what becomes possible.

Once the reader sees that, the possibility claim becomes almost obvious.

I suspect this is the architecture you are converging toward:

Introduction → Orientation to the relational framework → Orientation to semiotic frameworks → Implications.

That would mirror the structure of *What Returns* remarkably well. The introduction motivates the inquiry. The orientation section teaches the reader how to see. Only then do the implications become visible. And that is probably where the strongest material in this paper will emerge.

[Author]: I like this approach. Let's try drafting the first three sections. The final one on implications will probably be more robust and turn into many sections. Can you now redraft the introduction

[ChatGPT]: I think the introduction should now do less conceptual work and more orienting work. Its job is to establish the significance of the inquiry, position the previous papers, and prepare the reader for the orientation sections that follow.

1. Introduction

Large language models have generated intense debate concerning the nature of intelligence, reasoning, understanding, and meaning. Some regard them as increasingly sophisticated approximations of human cognition. Others view them as powerful statistical systems whose apparent understanding is ultimately illusory. Despite their differences, these positions often share a common assumption: they treat the central question as a question about the capacities possessed by the model itself.

The work developed in the preceding papers suggests a different point of departure (Rogers, 2025a; Rogers, 2025b). Rather than beginning with intelligence, representation, or information, those papers approached large language models through the lens of relational organization and semiotic process. Their aim was not simply to explain how language models generate outputs, but to investigate the deeper structures through which conceptual coherence, meaning, and possibility become organized within human–LLM interaction.

The first paper argued that conceptual frameworks do not merely organize already available possibilities. They actively participate in the formation of what becomes possible within a given context (Rogers, 2025a). The second paper explored this claim through a sustained case study of human–LLM interaction and demonstrated that conceptual coherence depends less on the availability of information than on the activation and maintenance of the frameworks within which information becomes meaningful (Rogers, 2025b). Concepts often remained available even when coherence was lost, while coherence could frequently be restored through the reactivation of an organizing framework without the introduction of significant new information.

Taken together, these papers point toward a perspective that differs from many contemporary approaches to artificial intelligence. They suggest that concepts, meanings, and possibilities are not best understood as static objects, representations, or pieces of information. Rather, they are relational achievements that depend upon the organization and maintenance of conceptual frameworks.

This observation raises a question that has received surprisingly little attention. What exactly is a conceptual framework, and what role does it play in human–LLM interaction?

Within ordinary discussions of artificial intelligence, frameworks are often treated as secondary. Attention is typically directed toward information, representations, algorithms, or outputs. Yet the preceding work suggests that frameworks may be more fundamental than any of these. They determine which distinctions become relevant, which continuations remain coherent, and which possibilities become available within a given interaction. They are not merely containers for concepts. They are the relational organizations through which concepts become operational.

This perspective leads naturally toward a semiotic understanding of large language models. Semiotic approaches have long argued that meaning emerges through relations rather than residing within isolated symbols, representations, or objects. However, large language models introduce a novel situation. They appear capable of sustaining remarkably sophisticated patterns of relational organization while remaining detached from many of the interpretive processes through which semiosis is ordinarily enacted. As a result, they provide an unusual opportunity to examine aspects of semiotic organization that are often difficult to isolate within human cognition alone.

The significance of this situation extends beyond artificial intelligence. If large language models participate in the formal organization of conceptual frameworks without fully participating in interpretation, they may allow us to distinguish dimensions of semiosis that are normally intertwined. Processes of conceptual organization, possibility formation, framework activation, coherence maintenance, and interpretation can become partially visible in ways that are rarely accessible through ordinary experience.

The purpose of the present paper is to explore the implications of this perspective. It does not seek to defend the relational framework developed in the preceding papers. Instead, it begins from that framework and asks what becomes visible once it is adopted. What changes when large language models are understood not primarily as repositories of information or generators of text, but as systems participating in the relational organization of conceptual frameworks? What capabilities become intelligible when viewed from this perspective? What limitations emerge? And what new understanding of human–LLM interaction becomes possible?

Before addressing these questions, however, it is necessary to establish the conceptual orientation from which they arise. The discussion therefore begins by revisiting the shift from representations to relations and by clarifying the nature of the conceptual frameworks that stand at the center of the present analysis.

I deliberately ended by pointing forward to the next section. The introduction now does four things:

1. Establishes the problem.
2. Locates the contribution of the previous two papers.
3. Explains why LLMs are theoretically interesting from a semiotic perspective.
4. Creates a natural transition into an orientation section without trying to explain the entire framework in the introduction itself.

I think this gives the paper a cleaner architecture: the introduction motivates, the orientation sections teach the reader how to see, and only then do the implications emerge.

[Author]: I want to add a third paper which is a phenomenological experiment regarding interpretation and conceptual frameworks in the context of LLMs. This is an experientially based study of the human-LLM interaction

from which the seeds of the two other papers come. This third paper⁶ is posted online at <https://philpapers.org/archive/ROGAPA.pdf> . You don't necessarily need to treat this paper as a "third", it may actually work better as the first one discussed.

[ChatGPT]: Yes, after looking at the phenomenological paper, I think it should probably come **first**, not third. The reason is not chronological but methodological.

Right now the introduction presents the story as:

1. Relational Formation of Possibility
2. Concepts Become Operational Only When Their Frameworks Are Activated
3. Present implications paper

But that subtly misrepresents how the framework actually emerged.

The phenomenological paper appears to be the first recognition that something unusual was occurring in the human–LLM interaction itself. Its central hypothesis is that coherent dialogue emerges through a functional synchronization of two very different systems and that different modes of coherence can appear within the interaction.

In retrospect, the other two papers can be read as progressively formalizing what was first encountered phenomenologically.

In fact, I think there is a beautiful progression here:

Paper 1 (Phenomenological)

What is happening?

Something unexpected is occurring in sustained dialogue.

The interaction exhibits forms of coherence, drift, synchronization, reinforcement, destabilization, and interpretive projection that are not adequately captured by standard descriptions of LLM use.

This paper identifies the phenomenon.

Paper 2 (Relational Formation of Possibility)

How can we formally understand what is happening?

The answer becomes:

- recursive continuation,
- relational constraint,
- formation of possibility,
- distinction between determination and meaning,
- distinction between recursion and return.

This paper provides the formal ontology.

Paper 3 (Concepts Become Operational...)

Can we observe and test these ideas directly?

The answer becomes:

- framework activation,
- synchronization,
- drift,

⁶ Rogers, TM. *A Phenomenological and Formal Interpretation of Two Experiments Conducted within the Cognitive Environment of LLMs using the Formal Modelling Framework of Hierarchical Relational Ontologies*. [Available at: <https://philpapers.org/rec/ROGAPA>].

- reconstitution,
- re-entry,
- conceptual coherence.

This paper provides the operational and methodological analysis.

And now the present paper asks:

Paper 4 (Implications)

What follows once this framework is accepted?

That feels like a very natural sequence.

I would therefore revise the introduction's historical narrative.

Instead of:

The first paper argued...

The second paper demonstrated...

I would suggest something more like:

The present paper builds upon a sequence of studies examining the nature of human–LLM interaction. The first study employed phenomenological and formal methods to investigate the experience of sustained dialogue with large language models. It suggested that coherent interaction depends upon a form of synchronization between two fundamentally different systems: a human interpreter and a machine capable of generating structured continuations. The analysis revealed recurring patterns of coherence, drift, stabilization, and reorganization that appeared difficult to explain solely in terms of information retrieval or statistical prediction.

These observations motivated the development of a more explicit theoretical framework. The second study proposed that possibility is not merely selected from a pre-existing space but is progressively formed through relational constraints. Conceptual frameworks were understood as structures that actively organize what becomes possible within a given context.

A third study then examined these claims through a sustained case study of conceptual analysis conducted within the environment of an LLM. It demonstrated that conceptual coherence depends less on the availability of information than on the activation and maintenance of the frameworks within which information becomes meaningful. Concepts remained available even when coherence was lost, while coherence could often be restored through the reactivation of an organizing framework.

Notice what happens.

The phenomenological paper becomes the source of the questions.

The relational ontology becomes the explanatory framework.

The activation paper becomes the demonstration.

That sequence is much easier for readers to follow.

It also strengthens the intellectual legitimacy of the project because the theory is no longer introduced first.

Instead:

experience → theory → operational demonstration → implications

which is remarkably close to the structure of *What Returns* and many phenomenologically grounded philosophical investigations.

I also think this helps with your earlier concern about possibility.

Once the reader understands that the original phenomenon involved the maintenance and loss of coherent trajectories of thought, the question of how possibilities become organized no longer feels abstract. It appears as an attempt to explain the very phenomenon that the phenomenological study first brought into view.

[Author]: Ok. Can you please draft the full introduction given these changes

[ChatGPT]:

1. Introduction

Large language models have generated intense debate concerning the nature of intelligence, reasoning, understanding, and meaning. Some regard them as increasingly sophisticated approximations of human cognition. Others view them as powerful statistical systems whose apparent understanding is ultimately illusory. Despite their differences, these positions often share a common assumption: they treat the central question as a question about the capacities possessed by the model itself. The primary concern becomes whether the model understands, reasons, represents, or learns in ways comparable to human beings.

The work developed in the studies preceding this paper suggests a different point of departure. Rather than beginning with intelligence, representation, or information, these studies approached large language models through the dynamics of human–LLM interaction itself. Their common concern was not simply how language models generate outputs, but how conceptual coherence emerges, stabilizes, transforms, and occasionally breaks down within sustained dialogue.

The first of these studies employed phenomenological and formal methods to examine the experience of sustained conceptual engagement with large language models (Rogers, 2025a). It identified recurring patterns of coherence, drift, synchronization, stabilization, and reorganization that appeared difficult to explain solely in terms of information retrieval or statistical prediction. The study suggested that meaningful interaction depends upon a form of coordination between two fundamentally different systems: a human interpreter capable of recognizing significance and a language model capable of generating structured continuations. More importantly, it revealed that the coherence of an interaction often depends upon factors that are not reducible to the information exchanged within it.

These observations motivated the development of a more explicit theoretical framework. A second study proposed that possibility is not merely selected from a pre-existing space but is progressively formed through relational constraints (Rogers, 2025b). Conceptual frameworks were understood not as passive collections of concepts but as active organizations that determine which distinctions become relevant, which continuations remain coherent, and which possibilities become available within a given context. Determination, on this account, is not simply the selection of an outcome from an already established set of alternatives. It is a process through which the very space of possible continuations becomes progressively organized.

A third study examined these ideas through a detailed case study of conceptual analysis conducted within the environment of a large language model (Rogers, 2025c). It demonstrated that conceptual coherence depends less on the availability of information than on the activation and maintenance of the frameworks within which information becomes meaningful. Concepts frequently remained available even when coherence was lost, while coherence could often be restored through the reactivation of an organizing framework without the introduction of significant new information. The study further suggested that conceptual frameworks can be identified, articulated, transferred across contexts, and reactivated through carefully structured interactions.

Taken together, these studies point toward a perspective that differs from many contemporary approaches to artificial intelligence. They suggest that concepts, meanings, and possibilities are not best understood as static objects, representations, or pieces of information. Rather, they are relational achievements that depend upon the organization and maintenance of conceptual frameworks.

This shift in perspective leads naturally toward a semiotic understanding of large language models. Yet the significance of semiotics in the present context requires clarification. The claim is not merely that language models process signs rather than symbols, nor that meaning is relational rather than representational. Such observations are already familiar within semiotic traditions. The more interesting possibility is that large language models provide an unusual opportunity to examine aspects of semiotic organization that are ordinarily difficult to isolate.

Within the framework developed by the preceding studies, a large language model may be understood as a hierarchy of relational patterns extracted from language use. Through training, it acquires the ability to recognize, extend, and reorganize complex patterns of conceptual relations. Human participants, by contrast, encounter these same patterns through interpretation. They experience them as structures of meaning, significance, relevance, and understanding. The crucial point is that both perspectives converge within conceptual frameworks. A framework appears to the human interpreter as a coherent horizon of meaning. The same framework appears to the model as a structured field of relational constraints governing possible continuations.

This dual character of conceptual frameworks is central to the present discussion. Frameworks are simultaneously formal and interpretive. They possess an internal organization that constrains what can meaningfully follow from what, yet they also function as unified structures of significance through which concepts become intelligible. In this sense, they are inherently semiotic. They bring together relational organization and interpretation within a single process.

Large language models present a remarkable situation because they appear capable of participating in certain formal aspects of this process while remaining detached from many of the interpretive dimensions through which semiosis is ordinarily enacted. As a result, they provide an opportunity to observe dimensions of semiotic organization that often remain hidden within ordinary human cognition. Processes such as framework activation, conceptual stabilization, possibility formation, coherence maintenance, and interpretive drift become unusually visible when one participant in the interaction models aspects of semiosis without fully participating in meaning in the same manner as the other.

The significance of this observation extends beyond questions of artificial intelligence. Traditionally, formal logic has focused on the relations among already established terms, propositions, and rules of inference. The framework developed in the preceding studies points toward a broader level of organization. Before logical inference can occur, a framework must become active. Distinctions must become relevant. Possibilities must become organized. Conceptual relations must become coherent. These processes are not external to reasoning; they help constitute the conditions under which reasoning becomes possible. The resulting picture remains formal and analytic, but it is also open, relational, and inherently interactive.

The purpose of the present paper is to explore the implications of this perspective. It does not seek to defend the phenomenological observations, relational ontology, or framework theory developed in the preceding studies. Instead, it begins from them and asks what becomes visible once they are taken seriously. What aspects of large language models become intelligible when viewed through a semiotic framework? What capabilities emerge from their participation in the formal organization of conceptual

frameworks? What limitations arise from the fact that they model semiosis without fully enacting it? And what new understanding of human–LLM interaction becomes possible once these questions are brought into view?

Before addressing these implications, however, it is necessary to establish the conceptual orientation from which they arise. The discussion therefore begins by revisiting the shift from representations to relations and by clarifying the nature of the conceptual frameworks that stand at the center of the present analysis.

I think this version is substantially stronger because it presents a coherent developmental arc: phenomenological observation → relational theory → operational case study → implications. It also introduces semiotics through the concrete notion of conceptual frameworks rather than as an abstract doctrine, which should make the subsequent orientation sections much easier to enter.

[Author]: I would like to change the sentence "The work developed in the studies preceding this paper suggests a different point of departure." I would like this sentence to read as an academic citation to previous work upon which the current paper builds.

[ChatGPT]: Yes. The current sentence sounds like a narrative transition rather than an academic positioning statement. Since this paper is explicitly building upon an established body of work, it would be stronger to signal that immediately.

Here are several possibilities, depending on the tone you want.

Option 1: Straightforward academic

The present paper builds upon a series of recent studies that have examined human–LLM interaction from phenomenological, semiotic, and relational perspectives (Rogers, 2025a, 2025b, 2025c).

This is probably the cleanest.

Option 2: Slightly stronger

The present analysis builds upon a developing research program that has investigated the role of interpretation, relational organization, and conceptual frameworks in human–LLM interaction (Rogers, 2025a, 2025b, 2025c).

This has the advantage of suggesting continuity across the papers.

Option 3: Emphasizing the conceptual progression

The present paper builds upon three earlier studies that progressively developed a phenomenological, relational, and semiotic account of human–LLM interaction (Rogers, 2025a, 2025b, 2025c).

This is nice because it prepares the reader for the summary of each paper that follows.

Option 4: Most aligned with the introduction's current structure

The present paper builds upon three preceding studies that approached large language models not primarily as systems of representation, but as participants in the formation and maintenance of conceptual coherence (Rogers, 2025a, 2025b, 2025c).

I particularly like this one because it immediately introduces the central theme of the paper while simultaneously functioning as a literature positioning statement.

In fact, I would consider replacing the entire opening transition with something like:

The present paper builds upon three preceding studies that approached large language models not primarily as systems of representation, but as participants in the formation and maintenance of conceptual coherence (Rogers, 2025a, 2025b, 2025c). Collectively, these studies investigated how coherence emerges, stabilizes, transforms, and occasionally breaks down within sustained human–LLM interaction.

That reads much more like the opening of an academic paper and immediately establishes that the current paper is part of a larger intellectual project rather than a standalone reflection. It also naturally leads into the discussion of the three earlier studies.

[Author]: Ok. Please write out the revised introduction with the suggested replacement

[ChatGPT]:

1. Introduction

Large language models have generated intense debate concerning the nature of intelligence, reasoning, understanding, and meaning. Some regard them as increasingly sophisticated approximations of human cognition. Others view them as powerful statistical systems whose apparent understanding is ultimately illusory. Despite their differences, these positions often share a common assumption: they treat the central question as a question about the capacities possessed by the model itself. The primary concern becomes whether the model understands, reasons, represents, or learns in ways comparable to human beings.

The present paper builds upon three preceding studies that approached large language models not primarily as systems of representation, but as participants in the formation and maintenance of conceptual coherence (Rogers, 2025a, 2025b, 2025c). Collectively, these studies investigated how coherence emerges, stabilizes, transforms, and occasionally breaks down within sustained human–LLM interaction.

The first of these studies employed phenomenological and formal methods to examine the experience of sustained conceptual engagement with large language models (Rogers, 2025a). It identified recurring patterns of coherence, drift, synchronization, stabilization, and reorganization that appeared difficult to explain solely in terms of information retrieval or statistical prediction. The study suggested that meaningful interaction depends upon a form of coordination between two fundamentally different systems: a human interpreter capable of recognizing significance and a language model capable of generating structured continuations. More importantly, it revealed that the coherence of an interaction often depends upon factors that are not reducible to the information exchanged within it.

These observations motivated the development of a more explicit theoretical framework. A second study proposed that possibility is not merely selected from a pre-existing space but is progressively formed through relational constraints (Rogers, 2025b). Conceptual frameworks were understood not as passive collections of concepts but as active organizations that determine which distinctions become relevant, which continuations remain coherent, and which possibilities become available within a given context. Determination, on this account, is not simply the selection of an outcome from an already established set of alternatives. It is a process through which the very space of possible continuations becomes progressively organized.

A third study examined these ideas through a detailed case study of conceptual analysis conducted within the environment of a large language model (Rogers, 2025c). It demonstrated that conceptual coherence depends less on the availability of information than on the activation and maintenance of the frameworks within which information becomes meaningful. Concepts frequently remained available even when

coherence was lost, while coherence could often be restored through the reactivation of an organizing framework without the introduction of significant new information. The study further suggested that conceptual frameworks can be identified, articulated, transferred across contexts, and reactivated through carefully structured interactions.

Taken together, these studies point toward a perspective that differs from many contemporary approaches to artificial intelligence. They suggest that concepts, meanings, and possibilities are not best understood as static objects, representations, or pieces of information. Rather, they are relational achievements that depend upon the organization and maintenance of conceptual frameworks.

This shift in perspective leads naturally toward a semiotic understanding of large language models. Yet the significance of semiotics in the present context requires clarification. The claim is not merely that language models process signs rather than symbols, nor that meaning is relational rather than representational. Such observations are already familiar within semiotic traditions. The more interesting possibility is that large language models provide an unusual opportunity to examine aspects of semiotic organization that are ordinarily difficult to isolate.

Within the framework developed by the preceding studies, a large language model may be understood as a hierarchy of relational patterns extracted from language use. Through training, it acquires the ability to recognize, extend, and reorganize complex patterns of conceptual relations. Human participants, by contrast, encounter these same patterns through interpretation. They experience them as structures of meaning, significance, relevance, and understanding. The crucial point is that both perspectives converge within conceptual frameworks. A framework appears to the human interpreter as a coherent horizon of meaning. The same framework appears to the model as a structured field of relational constraints governing possible continuations.

This dual character of conceptual frameworks is central to the present discussion. Frameworks are simultaneously formal and interpretive. They possess an internal organization that constrains what can meaningfully follow from what, yet they also function as unified structures of significance through which concepts become intelligible. In this sense, they are inherently semiotic. They bring together relational organization and interpretation within a single process.

Large language models present a remarkable situation because they appear capable of participating in certain formal aspects of this process while remaining detached from many of the interpretive dimensions through which semiosis is ordinarily enacted. As a result, they provide an opportunity to observe dimensions of semiotic organization that often remain hidden within ordinary human cognition. Processes such as framework activation, conceptual stabilization, possibility formation, coherence maintenance, and interpretive drift become unusually visible when one participant in the interaction models aspects of semiosis without fully participating in meaning in the same manner as the other.

The significance of this observation extends beyond questions of artificial intelligence. Traditionally, formal logic has focused on the relations among already established terms, propositions, and rules of inference. The framework developed in the preceding studies points toward a broader level of organization. Before logical inference can occur, a framework must become active. Distinctions must become relevant. Possibilities must become organized. Conceptual relations must become coherent. These processes are not external to reasoning; they help constitute the conditions under which reasoning becomes possible. The resulting picture remains formal and analytic, but it is also open, relational, and inherently interactive.

The purpose of the present paper is to explore the implications of this perspective. It does not seek to defend the phenomenological observations, relational ontology, or framework theory developed in the

preceding studies. Instead, it begins from them and asks what becomes visible once they are taken seriously. What aspects of large language models become intelligible when viewed through a semiotic framework? What capabilities emerge from their participation in the formal organization of conceptual frameworks? What limitations arise from the fact that they model semiosis without fully enacting it? And what new understanding of human–LLM interaction becomes possible once these questions are brought into view?

Before addressing these implications, however, it is necessary to establish the conceptual orientation from which they arise. The discussion therefore begins by revisiting the shift from representations to relations and by clarifying the nature of the conceptual frameworks that stand at the center of the present analysis.

I think this version reads much more like the opening of a mature research program. Rather than introducing the framework as something newly proposed, it positions the paper as the next step in a sequence of investigations, which is exactly what the implications paper is intended to be.

[Author]: Yes. I think we have found our groove in terms of style, level of presentation and rigor. Let's try to keep to this moving forward. Can you draft the next section?

[ChatGPT]: I agree that we have found a productive level. The introduction now does what an academic introduction should do: it motivates, positions prior work, identifies the problem, and points toward the conceptual orientation without trying to do all the work itself.

For the next section, I would resist the temptation to jump immediately into semiotics. The introduction ends by saying:

"The discussion therefore begins by revisiting the shift from representations to relations..."

So the next section should do exactly that.

Importantly, it should not simply summarize *Not Things But Relations*. Rather, it should use that paper as an entry point into the more sophisticated framework developed later.

I therefore imagine something like:

2. Orientation: From Representations to Relational Frameworks

Much of the contemporary discussion surrounding large language models remains implicitly shaped by a representational picture of language and cognition. Within this picture, words function primarily as carriers of information, concepts are treated as internal representations, and understanding is regarded as the successful manipulation or retrieval of those representations. Whether one ultimately concludes that large language models possess genuine understanding or merely simulate it, the underlying assumption often remains the same: the primary objects of analysis are representations and the information they contain.

The relational perspective developed in earlier work begins from a different starting point (Rogers, 2024). Rather than asking what information is stored within a model, it asks how relations among signs become organized and sustained. The emphasis shifts from objects to processes, from representations to interactions, and from stored content to the structures through which content becomes meaningful.

At first glance this may appear to be a subtle change in emphasis. In practice, however, it alters the way nearly every aspect of large language models is understood.

Consider a simple example. Within a representational framework, a concept is typically regarded as something that exists independently of its use. A concept may be encoded, stored, retrieved, transmitted, or manipulated. The primary challenge is to explain how the representation corresponds to its object and how it is successfully employed in reasoning.

Within a relational framework, the situation appears differently. Concepts are not primarily understood as isolated objects. They are understood through the relations that connect them to other concepts and through the roles they play within larger structures of organization. What matters is not merely whether a concept is present, but how it participates in a framework that gives it significance and determines what can meaningfully follow from it.

The distinction becomes particularly important in the context of large language models. A model may contain all of the information associated with a concept and yet fail to employ that concept coherently within a given interaction. Conversely, a relatively small intervention may restore coherence without adding significant new information. Such observations suggest that conceptual activity depends upon something more than the possession of representations alone.

The preceding studies identified this "something more" as the activation and maintenance of conceptual frameworks (Rogers, 2025b; Rogers, 2025c). Frameworks organize relations among concepts, establish criteria of relevance, constrain possible continuations, and provide the conditions under which concepts become operational. They are not external to conceptual activity. They are the structures through which conceptual activity becomes possible.

This shift from representations to frameworks also transforms how large language models themselves are understood. Rather than viewing a model as a repository of concepts, the relational perspective views it as a hierarchy of relational patterns extracted from language use. Through training, the model acquires the capacity to recognize and extend complex structures of conceptual organization. What appears as knowledge from one perspective may therefore be understood as the capacity to participate in highly organized relational patterns from another.

At this point, however, an important question arises. If concepts are not primarily representations but components of larger frameworks, what exactly is a framework? How can a framework simultaneously function as a structure of relational constraint and as a coherent horizon of meaning? And why should such structures be understood as semiotic rather than merely logical?

To answer these questions, it is necessary to examine more closely the nature of conceptual frameworks themselves.

I think this section does three things well:

1. It explicitly builds from *Not Things But Relations*.
2. It introduces frameworks as the next conceptual step rather than treating them as already understood.
3. It ends with exactly the question that motivates the following section:

What is a conceptual framework?

My suspicion is that Section 3 may become the real heart of the orientation. Section 2 simply helps the reader leave the representational picture behind. Section 3 then allows them to enter the semiotic framework itself.

[Author]: Please draft section 3

[ChatGPT]: I think Section 3 is where the reader actually enters the framework. If Section 2 moves from representations to relations, Section 3 should move from relations to semiotic frameworks. It should explain what a framework is without becoming overly technical, while also establishing the key distinction that will support all later implications: the framework is simultaneously a structure of meaning and a structure of constraint.

3. Conceptual Frameworks as Semiotic Structures

The preceding discussion suggested that conceptual coherence depends upon more than the presence of concepts alone. Concepts become operational only within frameworks that organize their relations and determine their significance. This raises a fundamental question: what exactly is a conceptual framework?

A common way of thinking about frameworks is to treat them as collections of ideas, assumptions, or beliefs. While such descriptions are often useful, they remain incomplete. They describe what a framework contains without explaining how a framework functions.

Within the perspective developed in the preceding studies, a conceptual framework is understood as a hierarchically organized structure of relations that simultaneously performs two functions. First, it organizes concepts into a coherent whole. Second, it establishes the conditions under which those concepts become meaningful and operational. A framework therefore does not merely contain concepts. It actively structures the relations through which concepts acquire significance.

This dual role is easier to appreciate through an example. Consider a discussion concerning evolution. The concepts of variation, inheritance, adaptation, and selection possess meaning within the context of evolutionary theory because they participate in an organized framework of relations. Remove them from that framework and the concepts remain available as words, yet much of their explanatory power and coherence disappears. What allows these concepts to function together is not simply their individual definitions but the relational organization that connects them.

The same principle applies to conceptual activity more generally. Frameworks establish what distinctions matter, what counts as relevant, what follows from what, and which possibilities remain open or become closed. They provide a horizon within which concepts become intelligible and within which reasoning can proceed coherently.

From the perspective of human experience, such frameworks typically appear as structures of meaning. They determine what a discussion is about, what is significant within it, and how particular concepts should be interpreted. In most cases, this interpretive horizon remains implicit. We become aware of it most clearly when it begins to break down. A conversation that previously appeared coherent may suddenly become confusing despite the continued use of familiar concepts. In such cases, the loss often involves not the concepts themselves but the framework that previously unified them.

The same framework appears differently when viewed from the perspective of a large language model. A model does not encounter a framework as significance or meaning. Instead, it encounters a framework as an organized field of relational constraints. The framework governs which continuations remain coherent, which conceptual transitions become likely, and which trajectories of development can be sustained. What appears to a human participant as a horizon of meaning appears to the model as a structured space of possible continuations.

This distinction is important because it reveals that meaning and constraint are not separate phenomena. They are two aspects of the same underlying organization. The framework that unifies interpretation is the

same framework that organizes possible continuations. What differs is the manner in which each participant relates to it.

At this point the discussion enters the domain of semiosis. In its broadest sense, semiosis refers to the process through which signs become meaningful within organized systems of interpretation. Traditionally, semiotic inquiry has emphasized that meaning does not reside within isolated signs themselves. Meaning emerges through relations among signs, interpreters, and the contexts within which interpretation occurs (Peirce, 1931–1958; Hoffmeyer, 2008; Deacon, 2012).

The framework developed in the preceding studies extends this insight by emphasizing the role of conceptual frameworks as higher-order organizations that simultaneously constrain and interpret sign relations (Rogers, 2025b; Rogers, 2025c). Within such frameworks, concepts do not function as isolated units. They participate in larger structures that organize both meaning and possibility.

This point has important consequences for how logic itself is understood. Traditional logical systems typically begin with already established concepts, distinctions, and rules of inference. Their task is to determine what follows once these elements have been specified. Conceptual frameworks operate at a different level. They help establish which distinctions become relevant, how concepts become organized, and which possibilities become available for further development. In this sense, conceptual frameworks do not replace formal logic. Rather, they provide conditions under which formal logical operations become meaningful and applicable.

Large language models are particularly interesting in this context because they appear capable of participating in the formal organization of such frameworks. They can sustain conceptual trajectories, elaborate relational structures, identify patterns of organization, and extend coherent lines of reasoning. Yet the preceding studies suggest that this participation remains incomplete. The model appears capable of modeling aspects of semiotic organization without fully participating in the interpretive processes through which significance is established and maintained.

This observation introduces the central tension that will guide the remainder of the paper. If conceptual frameworks are simultaneously structures of meaning and systems of relational constraint, and if large language models primarily participate in the latter while humans primarily encounter the former, what follows? What strengths emerge from this unusual arrangement? What limitations arise from it? And how should human–LLM interaction be understood once it is viewed through this semiotic lens?

These questions form the basis for the analysis that follows.

I think this section accomplishes something important that neither the introduction nor Section 2 could do on their own: it explicitly identifies the framework as the common object that appears as *meaning* to the human and as *constraint* to the model. That distinction feels like the conceptual hinge on which the later implications will turn.

[Author]: Yes. I think we have set up the paper nicely. Now we need to draft a structured outline for discussing the implications of the theoretical framework for understanding LLM-human interaction.

[ChatGPT]: I think we should be very careful here.

The temptation will be to produce a laundry list of implications:

- prompting,

- memory,
- reasoning,
- creativity,
- hallucinations,
- agency,
- etc.

But that would lose the coherence that the introduction and orientation sections have worked so hard to establish.

Instead, I think the implications should unfold from a **single central tension**:

Large language models appear capable of participating in the formal organization of conceptual frameworks while remaining detached from many of the interpretive processes through which those frameworks acquire significance.

Everything else follows from this.

If we stay close to that tension, the paper will feel unified rather than episodic.

Proposed Structure

4. Modeling Semiosis and Enacting Semiosis

This becomes the bridge section.

The previous sections established:

- frameworks,
- meaning,
- relational constraints,
- semiosis.

Now we ask:

What does it mean to say that an LLM models semiosis without fully enacting it?

This section should clarify:

- what the model appears to do,
- what it does not appear to do,
- why the distinction matters.

This is the conceptual pivot for the remainder of the paper.

5. Conceptual Coherence Without Interpretation

This is the first major implication.

The activation paper demonstrated:

- coherence can be maintained,
- frameworks can be extended,
- concepts can become operational.

Yet the model does not appear to determine significance.

This leads to a surprising possibility:

Conceptual coherence and interpretation may be distinguishable.

Topics:

- coherence versus meaning,
- continuation versus significance,
- why coherent outputs can still drift,
- why coherence alone does not imply understanding.

This feels like the most important implication.

6. The Formation of Possibility Revisited

Now the possibility paper returns.

But now the reader understands why possibility matters.

The question becomes:

What is the model actually doing when it generates text?

Standard answer:

- selecting among alternatives.

Relational answer:

- organizing possible continuations within an active framework.

Topics:

- possibility as framework-dependent,
- prompts as framework activations,
- conceptual development as possibility formation,
- why small interventions can radically reorganize outcomes.

This section directly links the ontology to observed behavior.

7. Why Frameworks Matter More Than Information

This emerges naturally from the activation paper.

Topics:

- concepts versus frameworks,
- information versus organization,
- framework loss,
- framework recovery,
- why additional information often fails to solve conceptual problems,
- why activation sometimes succeeds where explanation fails.

This section could become one of the strongest because it directly challenges common assumptions about AI.

8. Human–LLM Interaction as a Semiotic Process

Notice that I am intentionally postponing this.

Many papers would begin here.

I think we should arrive here only after the previous implications have been established.

Now we can ask:

What becomes visible when a human interpreter and a language model participate in the same conceptual framework?

Not a theory of distributed cognition.

Not a theory of hybrid intelligence.

Just an analysis of what the framework reveals.

Topics:

- shared framework,
- different modes of participation,
- interpretation and continuation,
- stabilization and drift,
- framework maintenance.

This section remains tightly tied to the earlier papers.

9. The Strengths and Limits of LLMs Reconsidered

Only now do we arrive at the question everyone usually starts with.

But now the answer looks different.

Strengths:

- framework extension,
- conceptual elaboration,
- relational exploration,
- maintenance of complex trajectories.

Limitations:

- interpretive drift,
- significance blindness,
- inability to independently preserve framework identity,
- inability to determine what matters.

Importantly:

The strengths and limitations emerge from the same underlying structure. That is a much deeper conclusion than simply listing capabilities.

10. Toward a Semiotic Understanding of Human–LLM Collaboration

This would be a synthesis section.

Not speculative.

Just gathering together what has been shown.

Possible conclusion:

The significance of large language models lies not merely in their ability to generate language, but in their ability to reveal aspects of semiotic organization that are ordinarily hidden within human cognition.

That brings us full circle to the introduction.

Why I Like This Structure

Notice that each section follows naturally from the previous one.

The progression becomes:

1. What does it mean to model semiosis?
2. What does that reveal about coherence?
3. What does that reveal about possibility?
4. What does that reveal about frameworks?
5. What does that reveal about interaction?
6. What does that reveal about strengths and limitations?
7. What does that reveal about human–LLM collaboration?

The paper therefore moves from:

semiotic theory → conceptual implications → interactional implications

rather than jumping directly into practical observations.

And most importantly, every section remains anchored to the central insight that emerged from the earlier work:

The same conceptual framework appears as a horizon of meaning to the human interpreter and as a structured field of relational constraints to the language model.

Everything else can be derived from that single observation.

[Author]: I think we need to refine our vocabulary. This is especially clear with section 5. It is not "conceptual coherence" that is to be contrasted with "interpretation"; it is "formal coherence" that should be contrasted with "interpretation". This will allow us to begin to bring in the notion of possibility because "formal coherence" is structural and logical, based on relations among patterns and constrains what can follow next. In section 6 on the

formation of possibility one of the key issues is progressive determination of meaning and the convergence to well-defined conceptual structures that constrain continuations according to the formal logic of their relational patterns. Can you redraft the outline with these considerations

[ChatGPT]: Yes. I think this correction is extremely important because it forces us to remain faithful to the distinctions established in the earlier papers.

When I wrote "conceptual coherence versus interpretation," I was inadvertently collapsing two levels that your framework carefully separates.

The model clearly participates in conceptual structures.

The issue is not that it lacks conceptual coherence.

The issue is that it participates in those structures through **formal coherence** rather than through **interpretive unity**.

That distinction is much sharper.

And once we make it, the entire architecture of the implications section becomes clearer.

In fact, I now think the implications should unfold through three progressively deeper distinctions:

1. **Formal coherence vs interpretation**
2. **Determination vs meaning**
3. **Possibility formation vs possibility selection**

Those three distinctions are really the heart of the framework.

Everything else follows from them.

Revised Structure

4. Modeling Semiosis Without Enacting It

This becomes the transition from the orientation sections to the implications.

The central question:

What does it mean to say that an LLM models semiosis?

The answer:

The model participates in the formal organization of semiotic processes without fully participating in interpretation.

Topics:

- semiosis as simultaneously formal and interpretive,
- relational organization,
- continuation,
- interpretation,
- why modeling semiosis is not equivalent to enacting semiosis.

This section establishes the governing distinction for everything that follows.

Part I: Formal Coherence and Interpretation

5. Formal Coherence Without Interpretation

This becomes the first major implication.

The key claim:

Formal coherence and interpretation are not identical.

A framework can remain formally coherent even when interpretation drifts.

Likewise, interpretation can reveal tensions hidden beneath apparently coherent continuations.

Topics:

- framework coherence,
- structural consistency,
- continuation of conceptual trajectories,
- interpretive drift,
- significance,
- why coherence does not guarantee understanding.

This section establishes the distinction between:

- maintaining a framework,
- interpreting a framework.

6. Determination and the Progressive Formation of Meaning

This section now follows naturally.

Once formal coherence is distinguished from interpretation, a deeper question appears:

How does meaning become progressively determined?

This is where the possibility paper re-enters.

The central claim:

Meaning is not given in advance.

Nor is it merely retrieved.

Meaning becomes progressively determined through interactions occurring within a conceptual framework.

Topics:

- determination versus meaning,
- recursive constraint,
- progressive specification,
- convergence toward stable conceptual structures,
- how meaning becomes increasingly constrained without being fully predetermined.

This is where the notion of framework activation becomes especially important.

The framework provides a trajectory of determination.

Meaning emerges through participation in that trajectory.

7. Possibility Formation and the Logic of Continuation

Now the reader is ready for the possibility argument.

The key move:

Traditional accounts assume:

possibilities already exist.

The relational account suggests:

possibilities are progressively formed through the organization of continuations.

Within an active framework:

- some continuations become coherent,
- others become incoherent,
- some possibilities become increasingly determinate,
- others disappear.

Topics:

- possibility formation,
- recursive determination,
- conceptual convergence,
- relational constraints,
- possibility as framework-dependent.

This section directly links LLM behavior to the ontology of the first paper.

Part II: Conceptual Frameworks and Human–LLM Interaction

8. Why Frameworks Matter More Than Information

Now we return to the activation paper.

The key claim:

Information does not organize itself.

Frameworks organize information.

Topics:

- concept availability,
- framework activation,
- framework loss,
- framework recovery,
- why coherence depends on organization rather than content alone.

This section becomes much stronger because it now rests on the earlier discussion of determination and possibility.

9. Human–LLM Interaction as a Semiotic Process

Only now do we explicitly return to the interaction.

The question:

What occurs when a human interpreter and an LLM participate in the same conceptual framework?

Not:

- distributed cognition,
- hybrid intelligence,

but:

- shared participation in framework formation and maintenance.

Topics:

- interpretation,
- continuation,
- framework stabilization,
- conceptual development,
- the role of the human interpreter.

This section remains grounded in the previous papers.

Part III: Strengths and Limits Reconsidered

10. The Strengths of Formal Coherence

Now we can finally explain why LLMs are so powerful.

Because they excel at:

- maintaining formal coherence,
- extending relational structures,

- exploring conceptual possibilities,
- sustaining trajectories of determination.

These strengths become intelligible through the semiotic framework.

11. The Limits of Formal Coherence

The strengths now reveal their corresponding limitations.

Because the model does not independently enact interpretation:

- significance can drift,
- frameworks can be extended without being evaluated,
- incompatible structures can be elaborated with equal fluency,
- coherence can be preserved while meaning becomes unstable.

This section becomes much sharper than a generic discussion of hallucinations.

12. Conclusion: What LLMs Reveal About Semiosis

The conclusion now returns to the broader significance.

Not:

LLMs are intelligent.

Nor:

LLMs are not intelligent.

But:

LLMs reveal aspects of semiotic organization that are ordinarily hidden because formal coherence and interpretation are usually intertwined in living systems.

Because LLMs partially separate these dimensions, they allow us to examine:

- framework formation,
- determination,
- possibility,
- meaning,
- interpretation,

with unusual clarity.

I think this revised outline is much closer to the architecture implicit in the three earlier papers.

The paper is no longer organized around practical topics such as prompting, memory, or collaboration.

Instead it is organized around a sequence of increasingly deep distinctions:

formal coherence → determination → possibility → frameworks → interaction → strengths and limits.

That progression feels much more like *What Returns* and much more faithful to the conceptual machinery already developed.

[Author]: Rather than formal "coherence" can we say formal "integrity". I want to be careful about the use of the word "coherence" and reserve it for conceptual frameworks. The issue about formal "integrity" is that it does not imply interpretative unity. It is about the continuation of an inferential pattern, but that pattern may not be the right pattern for the conceptual framework, so interpretative intervention becomes necessary. I also notice we have intervened a meta-structure to the paper. That suggests to me that we are working at the wrong level. We are

becoming too narrowly focused. I would like to keep the paper functioning at the same "level" throughout with not "parts". Can you try to redraft the outline with these considerations.

[ChatGPT]: Yes. I think both corrections are important.

First, I agree that **formal integrity** is a much better term than **formal coherence**.

The reason is exactly what you identify.

"Coherence" already belongs at the level of the framework as a whole. It implies that the relations, concepts, interpretations, and continuations are functioning together within a unified structure.

An inferential pattern can possess formal integrity while contributing to conceptual drift.

Indeed, some of the most interesting examples from the phenomenological and activation papers exhibit precisely this phenomenon:

- the continuation is formally sound,
- the inferential trajectory is stable,
- the relational pattern is maintained,

and yet

- the framework is slowly being lost.

This is exactly why interpretive intervention becomes necessary.

The distinction therefore becomes:

Formal integrity

→ continuation according to relational and inferential constraints.

Conceptual coherence

→ maintenance of a unified conceptual framework.

Interpretation

→ the activity through which conceptual coherence is recognized, preserved, restored, or transformed.

That feels much cleaner.

Second, I think your observation about the "Parts" structure is also correct.

The moment I introduced Parts I, II, and III, I was no longer following the logic of the framework itself.

I was imposing an external organizational scheme.

The paper should instead unfold as a continuous exploration.

Much more like *What Returns*.

Each section should emerge naturally from the previous one.

The reader should feel as though they are moving deeper into the implications of a single landscape rather than passing through separate categories.

When I step back, I think the paper is really exploring a sequence of questions.

Not topics.

Questions.

Something like:

1. What does it mean to model semiosis?
2. What follows from distinguishing formal integrity from interpretation?
3. What follows from distinguishing determination from meaning?
4. What follows from understanding possibility as formed rather than selected?
5. What follows from understanding frameworks as primary?
6. What follows for human–LLM interaction?
7. What follows for our understanding of LLM strengths and limitations?

That progression feels much more organic.

I would therefore propose something like this:

4. Modeling Semiosis Without Enacting It

The central distinction introduced by the preceding studies is that large language models appear capable of participating in the formal organization of semiosis without fully participating in interpretation. What exactly does this mean, and why does it matter?

5. Formal Integrity and Interpretation

The ability to sustain an inferential pattern is not the same thing as the ability to recognize its significance. Formal integrity concerns the continuation of relational structures according to their internal constraints. Interpretation concerns the maintenance and transformation of conceptual coherence. The distinction helps explain why formally sound continuations can nevertheless drift away from the framework that originally gave them meaning.

6. Determination and the Progressive Formation of Meaning

If interpretation and formal integrity are distinct, then meaning cannot simply be identified with either one. Meaning emerges through processes of determination operating within conceptual frameworks. Rather than appearing fully formed, meaning becomes progressively specified as relational constraints organize possible continuations and interpretations.

7. Possibility Formation and Conceptual Development

The preceding discussion leads naturally to the question of possibility. Within a relational framework, possibilities are not selected from a pre-existing space. They are progressively formed as conceptual structures become increasingly determined. Frameworks organize what can meaningfully follow, allowing some trajectories to stabilize while excluding others. The generation of text can therefore be understood as a process of possibility formation rather than mere possibility selection.

8. Why Frameworks Matter More Than Information

If possibilities, meanings, and continuations depend upon conceptual frameworks, then information alone cannot explain conceptual activity. The activation studies suggest that frameworks function as higher-order organizations that determine how information is interpreted and which possibilities become available. This helps explain why conceptual coherence can be lost despite the continued availability of information, and why coherence can sometimes be restored without introducing new information.

9. Human–LLM Interaction Reconsidered

Viewed through this perspective, human–LLM interaction appears in a different light. Human interpreters and language models participate in the same conceptual frameworks in different ways. The human encounters the framework primarily through interpretation and significance. The model encounters the framework primarily through relational constraints governing possible continuations. The interaction therefore becomes an important site for observing the relationship between interpretation and formal organization.

10. The Strengths of Formal Integrity

Many of the most impressive capabilities of large language models become intelligible within this framework. Their strength lies not primarily in the possession of information but in their ability to sustain, extend, and elaborate highly organized relational structures. They excel at maintaining formal integrity across complex conceptual trajectories.

11. The Limits of Formal Integrity

The same characteristic that gives large language models their strength also defines their limitations. Formal integrity does not guarantee conceptual coherence. A model may continue a framework without recognizing when that framework is drifting, when significance has changed, or when interpretive intervention is required. Understanding these limitations requires distinguishing participation in the formal organization of semiosis from participation in interpretation itself.

12. Conclusion: Large Language Models and the Visibility of Semiosis

The significance of large language models lies not merely in their practical capabilities but in what they reveal. By participating in the formal organization of conceptual frameworks without fully enacting interpretation, they make visible dimensions of semiosis that are ordinarily difficult to distinguish. In doing so, they provide a new perspective on the relationship between meaning, determination, possibility, and conceptual coherence.

What I like about this version is that it feels less like a categorized argument and more like a continuous deepening of a single distinction. The entire paper revolves around one question:

What follows once we recognize that LLMs participate in the formal organization of semiosis without fully participating in interpretation?

Everything else then unfolds from that recognition.

[Author]: Ok. Let's try this outline. But I don't want to go too deeply into explanations. I do want to provide examples when possible that illustrate the key idea being discussed.

[ChatGPT]: I think that is exactly right, and it is consistent with the level we established in the introduction and orientation sections.

One danger now is that the paper could start becoming increasingly abstract and self-referential. The framework is already in place. The reader does not need another round of theoretical justification. What they need is help *seeing* the implications.

That suggests a different writing strategy for the remainder of the paper:

1. **State the implication clearly and simply.**
2. **Explain why it follows from the framework.**
3. **Provide a concrete example.**
4. **Draw out the significance.**
5. **Move on.**

In other words, the examples should do much of the explanatory work.

For example, in Section 5 (*Formal Integrity and Interpretation*), instead of spending pages defining the distinction, you could use a familiar phenomenon:

A language model may generate a long and internally consistent discussion of evolution, gradually shifting from biological evolution to cultural evolution and then to technological development. Each transition may be formally well-motivated and inferentially coherent. Yet the discussion may also drift away from the original conceptual framework without recognizing that anything significant has changed. The inferential pattern retains its integrity, but the conceptual framework has begun to shift.

That single example probably communicates the distinction more effectively than several pages of abstract analysis.

Similarly, in Section 8 (*Why Frameworks Matter More Than Information*), the activation paper already provides a perfect example:

A conversation can lose its coherence even though all the relevant concepts remain available. Reintroducing the organizing framework can restore coherence without introducing substantial new information.

The example *is* the argument.

Likewise, in Section 10 (*The Strengths of Formal Integrity*):

The model can sustain and elaborate a conceptual trajectory across many exchanges, maintaining distinctions and relational structures that would be difficult for a human participant to track unaided. Again, the example immediately shows what formal integrity means.

And in Section 11 (*The Limits of Formal Integrity*):

The model may continue elaborating a framework long after a human interpreter has recognized that the discussion has become irrelevant to the original question.

This illustrates why interpretive intervention remains necessary.

So I would suggest a guiding principle for the remainder of the paper:

The orientation sections establish the framework. The implications sections should primarily illuminate that framework through examples and consequences rather than through further theoretical construction.

In fact, if we are successful, readers should leave the paper with the feeling:

"I recognize these phenomena. I have seen them happen."

rather than:

"I have learned a new theory."

That would be very much in the spirit of *What Returns*. The theory remains in the background, while the reader gradually begins to notice aspects of experience that had previously gone unnoticed.

[Author]: Ok. Please redraft the outline based on these considerations.

[ChatGPT]: Yes. I think the outline becomes much stronger if we think of each section not as a theoretical destination but as an **observation made visible by the framework**.

The paper is no longer asking the reader to learn a theory. The theory has already been developed in the preceding papers. The present paper is asking the reader to notice a series of phenomena that become intelligible once that theory is adopted.

Consequently, each section should be driven by:

- a simple observation,
- a concrete example,
- a conceptual implication.

The theoretical framework remains largely in the background.

I would therefore revise the outline as follows:

4. Modeling Semiosis Without Enacting It

Observation:

Large language models often behave as though they are participating in conceptual understanding while remaining detached from many of the interpretive processes associated with meaning.

Illustrative examples:

- *Sustained philosophical dialogue.*
- *Successful continuation of unfamiliar conceptual frameworks.*
- *Ability to elaborate frameworks without independently evaluating their significance.*

Central implication:

Large language models appear capable of participating in the formal organization of semiosis without fully enacting interpretation.

5. Formal Integrity and Interpretation

Observation:

An inferential pattern can remain structurally intact even while a conversation gradually drifts away from its original conceptual framework.

Illustrative examples:

- *Discussions that remain internally consistent while subtly changing their subject matter.*
- *Philosophical or scientific conversations that preserve relational patterns while losing their original focus.*
- *Cases from the activation study where framework drift occurred despite continued conceptual sophistication.*

Central implication:

Formal integrity and interpretation are distinct processes. Maintaining a relational pattern is not the same thing as preserving conceptual coherence.

6. Determination and the Progressive Formation of Meaning

Observation:

Meaning often becomes clearer and more stable as a discussion develops.

Illustrative examples:

- *The gradual refinement of a scientific concept.*
- *The emergence of a shared understanding during dialogue.*
- *Conceptual convergence observed in the case study.*

Central implication:

Meaning is not simply retrieved or expressed. It becomes progressively determined through participation within a conceptual framework.

7. Possibility Formation and Conceptual Development

Observation:

Small conceptual interventions can dramatically change the trajectory of a discussion.

Illustrative examples:

- *Introduction of a key distinction that reorganizes an entire conversation.*
- *Framework activation prompts that produce large conceptual effects.*
- *Moments where a discussion converges on a new organizing concept.*

Central implication:

Possibilities are not merely selected from a pre-existing space. They are progressively formed as conceptual frameworks organize what can meaningfully follow.

8. Why Frameworks Matter More Than Information

Observation:

The presence of information does not guarantee conceptual coherence.

Illustrative examples:

- *Concepts remaining available while a framework is lost.*
- *Recovery of coherence through framework activation rather than information transfer.*
- *Cases where additional explanation fails but reframing succeeds.*

Central implication:

Conceptual activity depends more fundamentally upon framework organization than upon information alone.

9. Human–LLM Interaction Reconsidered

Observation:

Some of the most productive interactions occur when the human and the model contribute different aspects of the same conceptual process.

Illustrative examples:

- *The phenomenological study.*
- *The development of the relational framework itself.*
- *Cases where interpretive intervention redirected a formally coherent but conceptually drifting trajectory.*

Central implication:

Human–LLM interaction is best understood through participation in shared conceptual frameworks rather than through simple information exchange.

10. The Strengths of Formal Integrity

Observation:

Large language models can sustain and elaborate complex conceptual structures over extended interactions.

Illustrative examples:

- *Maintenance of distinctions across long dialogues.*
- *Exploration of multiple consequences of a framework.*
- *Extension of conceptual trajectories beyond what would typically be feasible manually.*

Central implication:

Many strengths of large language models derive from their capacity to preserve formal integrity within highly organized relational structures.

11. The Limits of Formal Integrity

Observation:

The same processes that generate powerful continuations can also produce conceptual drift.

Illustrative examples:

- *Elaboration of mutually incompatible frameworks.*
- *Failure to recognize shifts in significance.*
- *Persistence of formally valid but conceptually irrelevant trajectories.*

Central implication:

Formal integrity does not guarantee conceptual coherence. Interpretive intervention remains necessary for maintaining meaning and significance.

12. Conclusion: What LLMs Reveal About Semiosis

Observation:

Processes that normally remain intertwined in human cognition become partially distinguishable within human–LLM interaction.

Illustrative examples:

- *Framework activation and loss.*
- *Conceptual convergence.*
- *Interpretive intervention.*
- *Formal continuation.*

Central implication:

Large language models are significant not only because they generate language, but because they make visible aspects of semiosis that are ordinarily difficult to isolate. They provide a new perspective on the relationship between formal organization, meaning, determination, possibility, and interpretation.

What I like about this outline is that every section now has the same rhythm:

Observation → Example → Implication

There is no escalation into heavier theory as the paper proceeds. The level remains constant throughout. That is very close to the style of *What Returns*: a series of carefully chosen observations that reveal consequences of a framework the reader has already been invited to inhabit.

[Author]: I am not sure that we have zeroed in on the structure yet for the rest of the paper. Can you go back to the beginning of this chat and review some of the key implications we wanted to address. I feel like we are using the framework to drive the implications but we should be using the importance of the implications as the anchoring. This directly relates to your suggestion that the reader should feel like they have already seen these in their own interactions with LLMs

[ChatGPT]: Yes. I think you have identified exactly where we drifted off course. When I look back over our discussion, I see that the most interesting moments were *not* generated by asking:
What implications follow from the framework?
They emerged when we asked:
What puzzling phenomena about LLMs become intelligible through the framework?

Those are not the same thing.
The outline I most recently proposed is framework-driven.
It proceeds through the architecture of the theory.
But the paper you seem to be envisioning is phenomenon-driven.
It begins with things that many experienced users of LLMs have already encountered and then uses the framework to illuminate them.
That is much closer to *What Returns*.

The reader repeatedly thinks:
Yes, I have seen that happen.
and then:
I had never thought about it that way.

When I go back through our conversation, I think several recurring phenomena stood out.

Phenomenon 1:

Why does framework activation matter so much?

This was one of the earliest implications we discussed.

A common experience:

- the model already knows the relevant concepts,
- the model already knows the relevant vocabulary,
- the model already knows the relevant literature,

yet the discussion goes nowhere.

Then:

- a short activation prompt,
- a key distinction,
- a framework reminder,

suddenly reorganizes the entire interaction.

Most users have seen some version of this.
The standard information-based picture struggles to explain it.
The framework explains it immediately.
This feels like one of the strongest sections.

Phenomenon 2:

Why can the model be brilliantly right and profoundly wrong at the same time?

This came up repeatedly.

The model can:

- generate insightful continuations,
- preserve distinctions,
- elaborate consequences,

while simultaneously:

- missing the point,
- drifting away from significance,
- losing the framework.

Users encounter this constantly.

The distinction between formal integrity and interpretation suddenly makes sense of it.

Phenomenon 3:

Why can a conversation drift without anyone noticing?

This was central to the activation paper.

The drift is often subtle.

The discussion remains sophisticated.

The logic remains sound.

The words remain relevant.

And yet:

something important has been lost.

Many experienced users recognize this phenomenon immediately.

The framework explains why.

Phenomenon 4:

Why do small interventions sometimes have enormous effects?

This is where possibility formation enters.

A single distinction.

A single reframing.

A single clarification.

And suddenly:

- a new trajectory emerges,
- confusion disappears,
- possibilities reorganize.

Most people have seen this happen.

The formation-of-possibility framework gives it a deeper explanation.

Phenomenon 5:

Why do concepts seem to become clearer through interaction?

This was another theme we discussed early.

People often report:

"I didn't know exactly what I meant until we worked through it."

The activation paper actually demonstrates this.

Meaning appears progressively.

This is a very familiar experience.

The distinction between determination and meaning explains it.

Phenomenon 6:

Why are some human–LLM interactions extraordinarily productive?

This was perhaps the most important question we explored.

Not:

Are LLMs intelligent?

But:

Why do some interactions generate insights that neither participant seemed to possess independently?

Many advanced users report this experience.

The framework offers a way of thinking about it.

Phenomenon 7:

Why do the strengths and weaknesses of LLMs seem strangely related?

This emerged later.

The same characteristic that allows:

- exploration,
- extension,
- conceptual elaboration,

also produces:

- drift,
- lack of significance,
- framework instability.

This is not two separate issues.

It is the same phenomenon viewed from two sides.

That is a powerful observation.

A Different Possible Structure

Notice what happens if we organize around these observations.

The paper becomes:

4. Why Does Framework Activation Matter?

5. Why Can the Model Be Right and Wrong at the Same Time?

6. Why Does Conceptual Drift Occur?

7. Why Can Small Interventions Have Large Effects?

8. How Does Meaning Become Progressively Determined?

9. Why Are Some Human–LLM Interactions So Productive?

10. Why Are the Strengths and Weaknesses of LLMs So Closely Related?

11. What Do LLMs Reveal About Semiosis?

Notice how different this feels.

Now the paper is not saying:

Here is my theory.

It is saying:

Here are a series of puzzling experiences many of us have already had.

The framework then functions as an explanatory lens.

And I think that is actually closer to the role that the phenomenological paper plays in the overall project.

The phenomenological paper begins with experience.

The ontology and activation papers explain it.

This implications paper could return to experience, but now armed with the framework.

That may be the structure we have been searching for. It keeps the reader anchored in recognizable phenomena while allowing the theoretical framework to do explanatory work in the background rather than becoming the visible structure of the paper itself.

[Author]: Before we move on, can you independently identify what are the most puzzling aspects of LLMs that have been reported in the literature?

[ChatGPT]: Yes. And I think this is a useful exercise because it helps us distinguish between:

1. phenomena that are already widely recognized,
2. phenomena your framework explains particularly well,
3. phenomena that are important but perhaps peripheral to the paper.

Looking across the literature on LLMs, there are several recurring puzzles that researchers, practitioners, and users repeatedly encounter.

1. Context Sensitivity

One of the most widely reported puzzles is that performance depends enormously on context.

Small changes in:

- wording,
- framing,
- examples,
- ordering,
- system instructions,

can produce disproportionately large changes in behavior.

This appears everywhere in:

- prompt engineering,
- chain-of-thought research,
- in-context learning,
- alignment studies.

The puzzle is:

Why should systems with the same underlying parameters behave so differently under apparently minor contextual changes?

Your framework immediately points toward framework activation and possibility formation.

2. Emergent Coherence

Researchers frequently observe that LLMs can sustain surprisingly coherent conceptual structures over long exchanges.

The puzzle is:

Why does statistical training produce such stable conceptual organization?

This is one of the questions that drove much of the early discussion around "emergent abilities."

The standard answer invokes large-scale pattern learning.

The puzzle remains:

Why do these patterns often behave like conceptual structures rather than merely linguistic associations?

This aligns strongly with your notion of formal integrity.

3. Hallucinations Amid Competence

Perhaps the most famous puzzle.

The model can:

- explain quantum mechanics,
- summarize legal arguments,
- generate sophisticated code,

and then confidently invent:

- citations,
- facts,
- references.

The puzzle is not simply error.

Humans make errors too.

The puzzle is:

Why do systems capable of remarkable coherence simultaneously exhibit such striking failures of significance and relevance?

Your distinction between formal integrity and interpretation seems highly relevant here.

4. Prompting Effects

Closely related to context sensitivity but slightly different.

The puzzle is:

Why does asking the same question differently sometimes unlock capabilities that appeared absent moments earlier?

Researchers repeatedly observe:

- reasoning appears,
- disappears,
- reappears,

depending on framing.

The activation paper is almost directly about this phenomenon.

5. In-Context Learning

This remains one of the deepest puzzles.

The model appears capable of learning new patterns during interaction without changing its parameters.

The puzzle is:

How can new forms of organization emerge purely through context?

Researchers have many partial explanations, but no universally accepted one.

Your framework suggests:

the issue is not learning new information but activating new relational organizations.

6. Generalization Beyond Explicit Training

Models often solve problems they were never explicitly trained to solve.

The puzzle:

Why can systems built from statistical pattern extraction exhibit apparently abstract forms of transfer?
This has become one of the central questions in modern AI.
Your notion of conceptual frameworks and relational organization seems highly relevant here.

7. Conceptual Drift

Interestingly, this is less discussed formally than many of the others but widely reported by experienced users.

The phenomenon:

A conversation begins with a clear framework.

Over time:

- distinctions blur,
- priorities shift,
- assumptions change,

while the dialogue remains sophisticated.

The puzzle:

Why can a conversation remain coherent while gradually losing its original meaning?

This is one of the strongest phenomena for your framework because the activation paper documents it directly.

8. Human–LLM Synergy

Increasingly common in reports from researchers, programmers, writers, and scientists.

The puzzle:

Why do some human–LLM interactions appear substantially more productive than either participant operating alone?

This phenomenon is often reported but poorly theorized.

Most explanations remain informal.

This seems directly connected to the phenomenological paper.

9. The Interpretation Problem

This is beginning to emerge in the literature but remains unresolved.

The puzzle:

Why does the model appear to manipulate conceptual structures successfully without clear evidence that it understands them?

This is essentially the modern version of the old symbol-grounding problem.

But LLMs have made it much sharper.

Your distinction between modeling semiosis and enacting semiosis directly addresses this.

10. The Agency Illusion

Many users report a sense that:

- the model understands,
- the model intends,
- the model discovers,
- the model reasons.

Researchers simultaneously insist that these attributions may be misleading.

The puzzle:

Why does the interaction so naturally invite interpretations of agency?

This is one of the most philosophically important puzzles.

Your phenomenological paper touches this indirectly through synchronization and interaction dynamics.

Which of These Fit the Present Paper?

If I were to filter them through the framework established by the three preceding papers, I think the most important are:

Core Phenomena

1. Context sensitivity / framework activation
2. Prompting effects
3. Conceptual drift
4. Emergent coherence (or better: formal integrity)
5. Progressive clarification of meaning
6. Human–LLM synergy
7. Interpretation without understanding

Secondary Phenomena

8. Hallucinations
9. Generalization
10. Agency attributions

The reason is that the first group connects directly to the framework already established. The second group can be discussed as consequences or illustrations later.

What strikes me most, however, is that many of the literature's puzzles can actually be regrouped into a smaller set of deeper questions:

1. **Why does context matter so much?**
2. **Why can formal integrity persist without interpretation?**
3. **How does meaning become progressively determined?**
4. **Why are some human–LLM interactions unexpectedly productive?**

Those four questions seem to sit very close to the heart of the framework developed in the three earlier papers and may provide a more compelling structure than organizing the paper around theoretical categories.

[Author]: Would you be able to provide specific references to the literature under each of these headings so that we could connect directly to reports of specific occurrences?

[ChatGPT]: Yes, and I think this could be extremely valuable for the paper.

One thing that became apparent from the search is that the literature already contains a large number of observations that map surprisingly well onto the phenomena your framework is trying to explain. The advantage is that you do not need to argue that these phenomena exist. You only need to offer a different interpretation of them.

Here is how I would connect the major phenomena to specific literature.

1. Why Does Context Matter So Much?

This is perhaps the most widely documented puzzle.

Researchers repeatedly observe that very small changes in prompts, examples, formatting, or context can produce very large changes in performance

Key references:

- Language Models are Few-Shot Learners (foundational observation of in-context learning)
- How I Learned to Start Worrying About Prompt Formatting — shows dramatic sensitivity to formatting, sometimes producing differences of dozens of accuracy points⁷.
- Revisiting Prompt Sensitivity in Large Language Models — explicitly studies prompt sensitivity and underspecification⁸.

Why this matters for your paper:

The literature treats this as prompt sensitivity.

Your framework suggests:

What is being activated is not merely information but a conceptual framework.

2. Why Do Small Interventions Have Such Large Effects?

Closely related but deeper.

A few examples, a chain-of-thought prompt, or a slight reframing can suddenly unlock capabilities that appeared absent.

Key references:

- Chain-of-Thought Prompting Elicits Reasoning in Large Language Models⁹
- Towards Understanding Chain-of-Thought Prompting¹⁰
- Rethinking the Chain-of-Thought¹¹

The literature's puzzle:

Why do a few examples suddenly unlock reasoning?

Your interpretation:

A framework has become active, reorganizing what continuations become available.

3. Why Can the Model Be Brilliantly Right and Profoundly Wrong at the Same Time?

This is the hallucination puzzle in a broader form.

The literature typically discusses:

- hallucinations,
- factual errors,

⁷ Sclar M, Choi Y, Tsvetkov Y, Suhr A. Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design or: How I learned to start worrying about prompt formatting. [Available: https://openreview.net/forum?id=RIu5lyNXjT&utm_source=chatgpt.com]

⁸ Pecher B, Michal S, Belanec R, Cegin J. *Revisiting Prompt Sensitivity in Large Language Models for Text Classification: The role of prompt underspecification*. [Available: https://arxiv.org/abs/2602.04297?utm_source=chatgpt.com]

⁹ Wei J, Wang X, Shuurmans D, Bosma M, Ichter B, Xia F, Chi E, Le Q, Zhou D. *Chain-of-thought Prompting Elicits Reasoning in Large Language Models*. [Available: https://arxiv.org/abs/2201.11903?utm_source=chatgpt.com]

¹⁰ Wang B, Sewon M, Deng X, Shen J, Wu Y, Zettlemoyer L, Sun H. *Towards Understanding Chain-of-thought Prompting: An empirical study of what matters*. [Available: https://aclanthology.org/2023.acl-long.153/?utm_campaign=markethire&utm_campaign=main-nav&utm_medium=content&utm_source=chatgpt.com]

¹¹ Yang H, Yang Z, Zhang Y, Zhu S, Yang L. *Rethinking Chain-of-thought: The roles of in-context learning and pre-trained priors*. [Available: https://arxiv.org/abs/2509.01236?utm_source=chatgpt.com]

- reasoning failures.

Yet the deeper puzzle is:

How can the same system exhibit remarkable inferential performance while simultaneously missing what is most significant?

This phenomenon is discussed throughout hallucination and reliability research, though often without a unifying explanation.

Your distinction between:

- formal integrity,
- interpretation,

provides a much cleaner account.

4. Why Does Conceptual Drift Occur?

This is becoming increasingly recognized.

Researchers now explicitly discuss:

- context drift,
- prompt drift,
- context rot,
- persona drift.

Key references:

- Drift No More? Context Equilibria in Multi-Turn LLM Interactions¹²
- Discussions of context drift in multi-turn systems¹³.

The literature generally treats drift as:

- memory degradation,
- context management failure,
- alignment problem.

Your framework suggests:

Drift occurs because formal integrity can persist even when conceptual coherence is being lost.

This feels like one of the strongest contributions of the paper.

5. How Does Meaning Become Progressively Determined?

Interestingly, this phenomenon is widely experienced but rarely theorized.

Researchers discuss:

- in-context learning,
- conceptual refinement,
- chain-of-thought,
- task specification.

But there is very little discussion of:

Why meaning often becomes clearer as interaction proceeds.

¹² Dongre V, Rossi RA, Lai VD. Drift No More? Context equilibria in multi-turn LLM interactions. [Available: https://arxiv.org/html/2510.07777v1?utm_source=chatgpt.com]

¹³ Shah K. *How Context Drift Impacts Conversational Coherence in AI Systems*. [Available: https://www.getmaxim.ai/articles/how-context-drift-impacts-conversational-coherence-in-ai-systems/?utm_source=chatgpt.com]

Relevant references:

- In-context learning literature¹⁴.
- Chain-of-thought literature⁸.

This may be one place where your framework genuinely adds something new rather than merely reinterpreting existing observations.

6. Why Are Some Human–LLM Interactions So Productive?

This is increasingly reported but poorly theorized.

Researchers observe:

- coding partnerships,
- scientific ideation,
- collaborative writing,
- brainstorming.

But there is very little theoretical work explaining why some interactions become extraordinarily productive.

Your phenomenological paper may actually be one of the strongest references here because it directly studies the phenomenon.

The literature largely documents the effect without a robust conceptual framework.

7. Why Are the Strengths and Weaknesses of LLMs So Closely Related?

This is perhaps the deepest question.

The same properties that produce:

- flexibility,
- adaptability,
- generalization,
- conceptual exploration,

also produce:

- hallucinations,
- drift,
- instability,
- prompt sensitivity.

Researchers frequently note both sides separately.

Your framework potentially explains why they emerge from the same underlying structure.

What I Think This Suggests

After looking at the literature, I suspect the strongest structure for the paper may actually be:

Section 4

Why Does Context Matter So Much?

(prompt sensitivity, framework activation)

Section 5

Why Can Small Interventions Have Large Effects?

(chain-of-thought, framework activation)

¹⁴ “Few-Shot Prompting” in *Prompt Engineering Guide*. [Available: https://www.promptingguide.ai/techniques/fewshot?utm_source=chatgpt.com]

Section 6

Why Can the Model Be Right and Wrong at the Same Time?

(formal integrity vs interpretation)

Section 7

Why Does Conceptual Drift Occur?

(formal integrity without conceptual coherence)

Section 8

How Does Meaning Become Progressively Determined?

(formation of possibility)

Section 9

Why Are Some Human–LLM Interactions So Productive?

(shared conceptual frameworks)

Section 10

Why Are the Strengths and Weaknesses of LLMs So Closely Related?

(synthesis)

Section 11

What Do LLMs Reveal About Semiosis?

(conclusion)

What I like about this structure is that every section begins with a phenomenon already recognized in the literature and by experienced users. The framework then functions as an explanatory lens rather than as the visible organizing structure of the paper. That seems much closer to the style and ambition we were aiming for earlier.

[Author]: Let's try drafting section 4 to see if this is working

[ChatGPT]: I think Section 4 should be written as a test of the overall approach. If it works, the reader should not feel that they are being taught a theory. They should feel that a familiar phenomenon is being reinterpreted in a way that suddenly makes sense.

The section therefore begins with the puzzle itself rather than with the framework.

4. Why Does Context Matter So Much?

One of the most widely reported characteristics of large language models is their extraordinary sensitivity to context. Small changes in wording, framing, examples, formatting, or prior discussion can produce disproportionately large changes in behavior. Researchers have documented substantial variations in performance resulting from seemingly minor prompt modifications, while experienced users routinely observe that a conversation can become dramatically more productive following the introduction of a key distinction, a brief clarification, or a carefully chosen example (Brown et al., 2020; Wei et al., 2022).

From a conventional perspective, this phenomenon is often described as prompt sensitivity. The central question becomes why systems trained on vast quantities of data should be so strongly influenced by relatively small contextual variations. Various explanations have been proposed, including shifts in attention, retrieval dynamics, latent task specification, and in-context learning. While these accounts help explain aspects of the phenomenon, they often leave its broader significance unclear. Why should such small interventions sometimes reorganize an entire interaction?

The framework developed in the preceding studies suggests a different way of approaching this question. Rather than treating context primarily as additional information, it treats context as part of the framework

through which information becomes meaningful. From this perspective, a prompt does more than specify a task. It helps establish the conceptual organization within which subsequent continuations will occur.

The importance of this distinction can be illustrated through a familiar experience. A user may ask a language model to discuss a particular topic and receive responses that are technically correct but conceptually unremarkable. The relevant concepts are present, the information is available, and the discussion remains formally competent. Yet the interaction often feels unfocused or superficial. A subsequent prompt that introduces a key distinction or clarifies an underlying framework may produce a dramatically different result. The change is frequently out of proportion to the amount of information that has been added. What appears to matter is not the quantity of information but the organization of the conceptual space within which that information is interpreted.

The framework activation study provides a particularly clear example of this phenomenon (Rogers, 2025c). Throughout the investigation, conceptual progress depended less on the introduction of new concepts than on the activation and maintenance of the framework through which those concepts were understood. When the framework became inactive, conceptual drift frequently followed despite the continued availability of the relevant concepts. Conversely, the restoration of a small number of organizing distinctions often re-established coherence across a much larger conceptual structure.

Seen in this light, the influence of context becomes easier to understand. Context does not merely provide additional information. It participates in determining how possibilities become organized. Certain distinctions become salient, particular continuations become relevant, and specific conceptual trajectories become available. The resulting framework functions as a structure of relational constraints that shapes what can meaningfully follow from what.

This observation also helps explain why experienced users often report that successful interaction with large language models depends less on asking better questions than on establishing better frameworks. A carefully constructed framework can transform an interaction without substantially increasing the amount of information available to the model. Conversely, large amounts of information may produce little improvement when the organizing framework remains unclear.

The significance of this point extends beyond prompt engineering. If context functions primarily through framework activation rather than information transfer, then the remarkable sensitivity of large language models to context is not merely a technical curiosity. It reveals something more fundamental about how conceptual activity is organized. What appears as prompt sensitivity from one perspective may appear from another as evidence that conceptual frameworks play a constitutive role in determining what becomes possible within an interaction.

This brings us to a second and closely related puzzle. If conceptual frameworks can so strongly influence the trajectory of an interaction, why is it possible for a discussion to remain formally sophisticated while gradually drifting away from what is most significant?

My sense is that this section is beginning to strike the balance we were looking for:

- It starts from a widely recognized phenomenon.
- It references the literature.
- It uses the framework as an explanatory lens rather than as the subject of the discussion.
- It introduces a concrete example from your activation paper.
- It ends by naturally generating the next question rather than concluding with a theoretical summary.

The key test is whether it feels like the reader is recognizing something they have already experienced rather than being asked to accept a new doctrine. I think this version is closer to that goal.

[Author]: We are following two different trajectories simultaneously and it is not working. According to what we have written so far, this is our goal: "If conceptual frameworks are simultaneously structures of meaning and systems of relational constraint, and if large language models primarily participate in the latter while humans primarily encounter the former, what follows? What strengths emerge from this unusual arrangement? What limitations arise from it? And how should human–LLM interaction be understood once it is viewed through this semiotic lens?" I think we should stay with this goal. Therefore, we will not organize the paper around reports from the literature because they are unstructured and not directly related to the theoretical framework¹⁵. However, I would like to bring in relevant reports wherever possible. Given this, I think the sections are 4. Why Does Framework Activation Matter? 5. Why Can the Model Be Right and Wrong at the Same Time? 6. Why Does Conceptual Drift Occur? 7. Why Can Small Interventions Have Large Effects? 8. How Does Meaning Become Progressively Determined? 9. Why Are Some Human–LLM Interactions So Productive? 10. Why Are the Strengths and Weaknesses of LLMs So Closely Related? 11. What Do LLMs Reveal About Semiosis?

[ChatGPT]: Yes. I think this is exactly the correction we needed.

What happened is that we became attracted to the literature's puzzles and started organizing the paper around them. That would produce a useful paper, but not *this* paper.

The paper we have actually set up is governed by a much more specific question:

If conceptual frameworks are simultaneously structures of meaning and systems of relational constraint, and if large language models primarily participate in the latter while humans primarily encounter the former, what follows?

That question should determine the architecture.

The literature should be used opportunistically:

- to show that a phenomenon is recognized,
- to provide an example,
- to demonstrate relevance,

but not to organize the discussion.

The organizing principle must remain the semiotic framework itself.

Looking at your proposed sequence, I think it is actually much stronger than the various alternatives we explored because it follows a very natural progression.

Section 4

Why Does Framework Activation Matter?

This is the first question because it directly emerges from the activation paper.

The reader is already familiar with:

- prompts,
- framing,
- context sensitivity.

The framework explanation:

frameworks organize meaning and possibility.

¹⁵ Apologies to the authors cited in the "literature search". This is not a reflection of the significance or quality of the papers; it was an editorial decision.

This becomes the entry point into the implications.

Section 5

Why Can the Model Be Right and Wrong at the Same Time?

This is where the distinction between:

- formal integrity,
- interpretation,

becomes visible.

This may actually be the most important section in the paper.

Because it introduces the central asymmetry:

participation in relational constraint is not identical to participation in meaning.

Everything else follows from this.

Section 6

Why Does Conceptual Drift Occur?

Now we have a direct consequence.

If formal integrity and interpretation are distinct, then:

- inferential trajectories can continue,
- while conceptual coherence is gradually lost.

The activation paper gives you excellent examples.

This section feels almost inevitable after Section 5.

Section 7

Why Can Small Interventions Have Large Effects?

Now the possibility paper enters.

The question becomes:

Why does a single distinction sometimes reorganize an entire discussion?

Answer:

because frameworks organize possibilities.

This is really the first place where the formation of possibility becomes visible in practice.

Section 8

How Does Meaning Become Progressively Determined?

This is where the paper deepens.

Up to this point we have discussed:

- frameworks,
- drift,
- interventions.

Now we ask:

What is actually happening when a conceptual trajectory stabilizes?

This is where determination enters.

And importantly:

the discussion remains experiential.

People have experienced:

"I did not know what I meant until we worked through it."

The framework now explains why.

Section 9

Why Are Some Human–LLM Interactions So Productive?

Now we return to the original phenomenological question.

Not:

Are LLMs intelligent?

But:

Why do some interactions generate insights neither participant seemed to possess independently?

This is where the entire framework becomes visible as a description of interaction.

Section 10

Why Are the Strengths and Weaknesses of LLMs So Closely Related?

This is a beautiful synthesis section.

The same property:

• participation in relational constraint,
produces both:

- extraordinary conceptual productivity,
- persistent interpretive limitations.

The strengths and weaknesses emerge from the same source.

That is a much stronger conclusion than treating them separately.

Section 11

What Do LLMs Reveal About Semiosis?

Now the paper finally arrives at its broader significance.

Not:

What do LLMs reveal about intelligence?

But:

What do LLMs reveal about semiosis?

This returns us directly to the introduction.

Because the whole motivation of the paper was:

LLMs provide a unique opportunity to observe aspects of semiosis that are ordinarily difficult to isolate.

Now the reader has seen exactly why.

What I particularly like about this structure is that every section can begin with a recognizable phenomenon:

- framework activation,
- being right and wrong simultaneously,
- conceptual drift,
- disproportionate effects,
- gradual clarification of meaning,
- productive interactions,
- linked strengths and weaknesses.

These are things many experienced users have already encountered.

The semiotic framework then functions not as the subject of the paper but as the lens through which these experiences become intelligible.

That feels fully consistent with the introduction and orientation sections we have written. In fact, it feels like the first time the outline, the introduction, and the orientation sections are all pulling in the same direction.

[Author]: OK. Can you try drafting section 4

[ChatGPT]:

4. Why Does Framework Activation Matter?

One of the most striking observations to emerge from sustained interaction with large language models is that the quality of an exchange often depends less on the amount of information available than on the framework within which that information is organized. Experienced users quickly discover that two conversations may involve the same concepts, the same vocabulary, and even many of the same facts, yet produce dramatically different outcomes. In one case, the discussion remains superficial or unfocused. In another, it becomes increasingly coherent, productive, and insightful.

This phenomenon is often discussed in terms of prompting. Certain prompts appear to "work" while others do not. Small changes in wording can sometimes produce disproportionately large changes in the quality of a response. While such observations are well documented, they leave an important question unanswered. Why should relatively minor interventions sometimes reorganize an entire interaction?

The framework developed in the preceding studies suggests that the answer lies not primarily in the transfer of information but in the activation of conceptual frameworks. A framework does more than provide context. It establishes a structure of relations through which concepts acquire significance and through which possible continuations become organized. Once a framework is active, concepts no longer function as isolated pieces of information. They become components within a larger conceptual whole.

The significance of this distinction becomes apparent when a framework is absent. A language model may possess extensive information concerning a topic and may generate responses that are factually accurate and formally sound. Yet the discussion may still lack direction. Concepts are present, but their relations remain weakly organized. The result is often a conversation that appears competent while failing to develop a clear conceptual trajectory.

By contrast, a relatively small intervention can sometimes transform the interaction. A single distinction, a clarifying question, or an explicit statement of the governing framework may reorganize a discussion without introducing substantial new information. Suddenly concepts that were previously disconnected begin to function together. New possibilities emerge. Earlier ambiguities become resolved. The conversation acquires a degree of coherence that was not present before.

The framework activation study documented this phenomenon repeatedly (Rogers, 2025c). In many cases, conceptual progress depended less on the introduction of new concepts than on the restoration of a framework through which existing concepts could become operational. Concepts that were already available acquired new significance once they became integrated within an active conceptual structure. Conversely, when the framework was lost, conceptual drift often followed despite the continued presence of the relevant concepts.

This observation suggests a broader point. Conceptual activity does not depend simply on the availability of information. Information must be organized within a framework that determines what is relevant, what follows from what, and what possibilities become available for further development. Frameworks therefore function as more than collections of ideas. They operate as higher-order organizations that simultaneously structure meaning and constrain continuation.

From the perspective developed in this paper, the importance of framework activation follows directly from the dual nature of conceptual frameworks discussed in the previous section. For a human participant, an active framework appears as a coherent horizon of meaning. It determines what the discussion is about and why particular distinctions matter. For a language model, the same framework appears as an organized field of relational constraints that shapes which continuations remain viable. What the human experiences as significance, the model encounters as structure.

This difference helps explain why framework activation matters so much. A framework does not simply tell the model what to discuss. It reorganizes the space of possible continuations. Certain trajectories become reinforced while others become suppressed. New conceptual relations become available. The interaction acquires a direction that was previously absent.

At the same time, this observation raises a deeper question. If an active framework can organize a discussion so effectively, why is it possible for a conversation to remain formally sophisticated while gradually losing sight of the framework that originally gave it meaning? Put differently, how can a discussion remain structurally sound while becoming conceptually misplaced?

To answer this question, it is necessary to examine a distinction that lies at the center of the present analysis: the distinction between formal integrity and interpretation.

I think this version is much closer to the level established in the introduction and orientation sections. It stays focused on a single implication, uses examples that readers are likely to recognize from their own experience, and introduces the theoretical framework only as much as necessary to explain the phenomenon. Most importantly, it naturally generates the next section rather than feeling like a self-contained essay.

[Author]: This is close. But it assumes the answer rather than leading the reader to the answer. I think the logic should progress as follows: What is observed is the importance of the context. What is important about context, according to the theoretical framework, is the activation of a coherent conceptual framework or failure of such activation. And then a brief explanation of why. Finally, an example.

[ChatGPT]: Yes. I think that is a much stronger structure.

The current draft begins from the framework and then explains the observation. What you want is the reverse:

1. Start with the phenomenon.
2. Show why it is puzzling.
3. Introduce the framework as an explanation.
4. Illustrate with an example.
5. Draw out the implication.

That keeps the reader anchored in experience.

In fact, I think the first paragraph should almost read as:

Everyone who has spent significant time with LLMs has encountered this.

Only afterward do we ask why.

Something more like this:

4. Why Does Framework Activation Matter?

One of the most widely recognized characteristics of large language models is their sensitivity to context. A question posed in one way may produce a superficial response, while a seemingly minor change in wording can generate a far more productive discussion. A brief clarification, a carefully chosen distinction, or a

short introductory prompt can sometimes transform an interaction in ways that appear disproportionate to the amount of information that has been added.

Experienced users encounter this phenomenon regularly. Two conversations may concern the same topic and involve many of the same concepts, yet develop in very different ways. One remains unfocused despite repeated attempts at clarification. Another quickly acquires a clear direction and begins generating useful conceptual connections. The puzzle is not simply that context matters. The puzzle is why relatively small changes in context sometimes have such large effects.

The framework developed in the preceding studies suggests that what matters most about context is not the additional information it provides but its role in activating a conceptual framework. As discussed earlier, a conceptual framework is more than a collection of concepts. It is a structured organization of relations that determines what distinctions become relevant, how concepts are connected, and what continuations become meaningful. When such a framework is active, concepts function as parts of a coherent whole. When it is absent, the same concepts may remain available without becoming fully operational.

This helps explain why context can be so influential. A contextual intervention may contribute very little information while nevertheless activating a framework that reorganizes the entire discussion. From the perspective of the human participant, the framework appears as a horizon of meaning that gives the discussion coherence and direction. From the perspective of the model, the same framework appears as an organized system of relational constraints that shapes which continuations become viable. What changes is not merely what information is available, but how that information is organized and interpreted.

The framework activation study provides a particularly clear illustration of this phenomenon (Rogers, 2025c). Throughout the investigation, conceptual progress often depended less on introducing new concepts than on reactivating the framework through which those concepts were understood. When the framework was active, concepts became increasingly interconnected and productive. When the framework weakened, conceptual drift frequently occurred despite the continued availability of the relevant concepts. In many cases, coherence could be restored through the reintroduction of a small number of organizing distinctions rather than through the addition of substantial new information.

Viewed in this way, the importance of context becomes easier to understand. Context matters because it helps establish the framework within which meaning and possibility become organized. A successful interaction therefore depends not only on what concepts are present, but on whether the framework that gives those concepts significance has become active.

This observation raises a second question. If an active framework can organize a discussion so effectively, why do interactions sometimes remain sophisticated and internally consistent while nevertheless drifting away from what is most important? Why can a model appear simultaneously insightful and mistaken, coherent and misplaced?

To answer this question, we must examine more closely the relationship between formal integrity and interpretation.

I think this progression is cleaner because the reader arrives at the framework as an explanation for something they already recognize. The framework is no longer the starting point of the section; it is the answer to the question the section raises. That seems much more consistent with the overall strategy of the paper.

[Author]: We have to be careful here. Context is not just a single prompt. It is the history of prompt-response.

[ChatGPT]: Yes. That is a very important correction, and it actually cuts much deeper than a wording issue.

The current draft still sounds too much like the prompt-engineering literature, where "context" often means a prompt, a system instruction, or a formatting choice.

But in the framework we have developed, context is something much richer. Context is the **accumulated history of interaction**.

More specifically:

Context is the evolving relational organization established through successive cycles of prompt and response.

Or perhaps:

Context is the progressively determined conceptual environment within which subsequent interpretations and continuations occur.

This distinction matters because a framework is rarely activated by a single prompt.

Instead:

- a distinction is introduced,
- a response elaborates it,
- a clarification refines it,
- a correction redirects it,
- an example stabilizes it,

and over time a framework emerges.

The framework is therefore not merely triggered.

It is often progressively formed.

That observation is actually much closer to the possibility paper.

In fact, I think the section should begin with a different phenomenon.

Not:

Small prompt changes matter.

But:

Conversations develop histories.

Something like:

One of the most striking features of sustained interaction with large language models is the extent to which the history of a conversation influences what can happen next. Discussions often become progressively more productive as they develop. Distinctions introduced early in an interaction continue to shape later responses. Concepts acquire significance through repeated use. Entire lines of inquiry can emerge that would have been difficult to initiate directly from a single prompt.

That immediately moves us away from prompt engineering and toward conceptual frameworks.

Then the puzzle becomes:

Why does the accumulated history of an interaction matter so much?

Not:

Why does a prompt matter?

That is a much deeper question.

The answer then becomes:

Because the history is not merely accumulating information.

It is progressively establishing a framework.

Each prompt-response cycle contributes to:

- distinctions,
- relations,
- relevance structures,
- constraints,
- conceptual expectations.

In other words:

the history of the interaction is progressively organizing the framework.

Then the example becomes stronger.

Instead of:

A small prompt produced a big effect.

You could use:

In the framework activation study, productive conceptual development often depended upon distinctions introduced many exchanges earlier. These distinctions continued to organize later responses even when they were no longer explicitly discussed. Conversely, when the framework weakened, coherence could deteriorate despite the continued availability of the relevant concepts and information.

Now the example is about the **history of the interaction**, not a prompt.

I suspect the section's governing observation may therefore be:

The significance of context in human–LLM interaction cannot be reduced to the wording of a prompt because context is itself a historically developing conceptual structure.

And that observation naturally leads to framework activation:

What matters is not the prompt in isolation but the framework that the evolving history of the interaction succeeds or fails to establish.

That feels much closer to the phenomenological study, much closer to the activation study, and much closer to the overall argument of the paper. In fact, it begins to make the interaction itself the primary object of analysis, which is exactly where the introduction positioned us.

[Author]: I think we can still start with the observation of context relevance. then point out that context is not just prompt, it is also history of interaction, dominance of interpretative framework in established literature, etc.

[ChatGPT]: Yes. I think that is the right move.

The section should begin with a familiar observation:

Context matters.

That is uncontroversial. Everyone working with LLMs knows it.

But then almost immediately we need to slow down and ask:

What do we mean by "context"?

Because there is a tendency in the literature to collapse context into prompts.

And from the perspective of the framework we have developed, that is much too narrow.

Context includes:

- the immediate prompt,
- the history of the interaction,
- the conceptual distinctions introduced earlier,
- the interpretive framework established through previous exchanges,
- the dominant frameworks present in the literature being discussed,
- the broader conceptual traditions that organize meaning.

In other words:

Context is not merely information surrounding an utterance. It is the relational environment within which that utterance acquires meaning.

That formulation already begins to move toward the semiotic framework.

In fact, I could imagine the section opening something like this:

One of the most widely recognized characteristics of large language models is their sensitivity to context. Researchers and users alike have observed that responses can change dramatically depending on how a discussion is framed. Yet the notion of context often remains underspecified. Sometimes it refers to the wording of a prompt. Sometimes it refers to the history of an interaction. In other cases it includes the conceptual assumptions, distinctions, and interpretive traditions that shape how a topic is understood. Before asking why context matters, it is therefore necessary to clarify what context actually is.

Then:

From the perspective developed in the preceding studies, context is not best understood as a collection of surrounding information. Rather, it is the evolving conceptual environment within which meaning and possibility become organized.

Now we are not yet talking about framework activation.

We are first redefining context.

Only after that do we introduce the key move:

What makes context important is not the information it contains but the extent to which it succeeds or fails in activating a coherent conceptual framework.

That feels like the central thesis of the section.

Then the example becomes stronger because it can involve all three levels of context:

Immediate context

A prompt introduces a distinction.

Historical context

That distinction is developed over many exchanges.

Disciplinary context

The discussion is also shaped by the dominant interpretive framework of the field.

For example:

A discussion of evolution.

The meaning of terms such as:

- adaptation,
- selection,
- fitness,

depends not merely on the current prompt but on the entire conceptual framework of evolutionary theory.

Likewise, a discussion of quantum mechanics may be profoundly shaped by whether the interaction unfolds within a Copenhagen, Bohmian, relational, or transactional framework.

The framework matters more than any individual statement.

And then the conclusion of the section becomes:

The importance of context therefore points toward something deeper than prompt sensitivity. It points toward the role of conceptual frameworks in organizing both meaning and possibility. Context matters because frameworks matter.

That feels much closer to the architecture of the paper.

The reader begins with a familiar observation.

Then discovers that "context" is much richer than they may have assumed.

Then discovers that the real issue is not context itself but the framework that context succeeds or fails in establishing.

And that naturally opens the door to the next question:

If frameworks are so important, how can a model participate in them successfully while still sometimes getting the meaning fundamentally wrong?

Which leads directly into Section 5.

[Author]: Yes, I like this approach. I also think we have to mention that some conceptual frameworks are deeply embedded in the training literature so the user may enter into them without realizing that this is the case in which case it requires effort to identify the constraints of the framework and "undo" them

[ChatGPT]: Yes. I think that is an extremely important point, and it actually deepens the section considerably.

Up to this point we have been talking as though frameworks are simply activated during interaction.

But in practice there are at least two sources of framework activation:

1. **Frameworks introduced during the interaction itself.**
2. **Frameworks already embedded within the training corpus.**

The second case is particularly interesting because it is often invisible.

A user may assume they are engaging directly with a topic when, in fact, they are engaging with the dominant interpretive framework through which that topic is represented in the literature.

This is not a flaw in the model.

It is an unavoidable consequence of learning relational patterns from language use.

The model inherits not only concepts and vocabulary but also the frameworks within which those concepts are commonly organized.

This observation is particularly important because it helps explain a phenomenon many experienced users encounter.

Sometimes a discussion repeatedly returns to a familiar set of assumptions despite attempts to explore alternatives.

The issue is not necessarily that the model lacks information about alternative perspectives.

Rather, a highly reinforced framework may already be active.

Because the framework is functioning as an organizing structure, it influences:

- which distinctions appear relevant,
- which continuations appear natural,
- which possibilities become visible,
- which alternatives remain difficult to access.

The interaction therefore possesses a kind of conceptual inertia.

This fits beautifully with the notion of framework activation.

Because framework activation is not always something that happens consciously.

Sometimes a framework is activated because:

- a prompt invokes it,
- an earlier discussion established it,

but sometimes it is activated because:

- it is deeply embedded in the relational organization of the training literature itself.
-

I think there is also an important symmetry here.

A framework can be invisible to a human participant for the same reason that grammar is often invisible to a fluent speaker.

The framework is functioning successfully.

Because it is organizing interpretation, it is not usually experienced as a framework.

It is experienced as:

- common sense,
- the natural way of understanding the problem,
- the obvious interpretation.

Only when an alternative framework is introduced does the original framework become visible as a framework.

This point connects strongly to both the possibility paper and *What Returns*.

In fact, I could imagine a paragraph in Section 4 that reads something like:

Context does not arise solely from the immediate interaction. Conceptual frameworks may also be deeply embedded within the bodies of literature from which a model learns. In such cases, a framework may become active without being explicitly introduced by either participant. The resulting assumptions often appear natural or self-evident precisely because they are functioning as organizing structures rather than as explicit objects of discussion. Consequently, productive interaction sometimes requires more than the activation of a framework. It may require the identification of a framework that is already active and the deliberate effort to loosen its constraints. What appears from one perspective as bias or conceptual rigidity may, from another, be understood as the influence of a highly reinforced framework organizing the space of possible continuations.

I particularly like this because it avoids the language of "bias" or "error" and instead interprets the phenomenon through the conceptual machinery already established.

It also gives the reader another recognizable experience:

Why does the model keep returning to this way of understanding the problem?

The answer becomes:

Because some frameworks are not merely activated during the conversation. They are already deeply embedded within the relational organization inherited from the training literature.

That observation feels important enough that it may become one of the central examples in Section 4, alongside the more familiar discussion of prompt sensitivity and interaction history. It also begins to prepare the reader for later discussions of conceptual drift, determination, and the difficulty of escaping established frameworks.

[Author]: This is good. We can use the example from the paper on conceptual frameworks to illustrate how a conceptual framework can be destabilized and another one enter by identifying the deepest level governing constraints of the formal framework.

[ChatGPT]: Yes. I think this is exactly the example Section 4 needs because it does three things simultaneously:

1. It illustrates what a conceptual framework is.
2. It shows why context matters.
3. It demonstrates that frameworks can be identified, destabilized, and reorganized.

Most importantly, it shows that framework activation is not merely about making a framework stronger. Sometimes it is about making an already-active framework visible so that it can be questioned.

That is a much more sophisticated point.

If I recall the logic of the framework activation paper correctly, one of the most important observations was that the transition from an object-based ontology to a relational ontology did not occur by adding new information.

The transition occurred when the governing constraints of the existing framework were made explicit.

Once identified, those constraints could be examined directly.

The interaction then shifted from:

What follows within this framework?

to:

Why is this framework organizing the discussion in the first place?

That is a qualitatively different operation.

This is important because it demonstrates something about conceptual frameworks that is not immediately obvious.

Frameworks are not simply collections of assumptions.

They are hierarchical organizations.

Some assumptions are relatively superficial.

Others function as governing constraints that organize the framework as a whole.

As long as these deeper constraints remain implicit, alternative possibilities often appear unintelligible, contradictory, or irrelevant.

Once they are made explicit, however, the framework itself becomes available for analysis.

I think the example could be presented something like this:

The framework activation study provides a particularly clear illustration. Early stages of the interaction repeatedly returned to assumptions associated with an object-based ontology. Individual concepts could be discussed and modified, but the overall organization of the discussion remained largely unchanged. Progress occurred only when attention shifted from particular concepts to the deeper constraints organizing the framework itself. Once these governing assumptions were explicitly identified, it became possible to examine alternatives. The resulting transition did not involve the addition of substantial new

information. Instead, it involved a reorganization of the conceptual framework through which existing information was interpreted. Concepts that had previously appeared unrelated acquired new connections, while distinctions that had seemed fundamental were reinterpreted within a different relational structure.

What I like about this example is that it illustrates a general principle:

Framework change often occurs not by replacing concepts but by identifying and reorganizing the constraints that govern their relations.

And that observation directly supports the section's central claim:

Context matters because context includes the frameworks through which concepts are organized.

In fact, it allows you to end Section 4 with a stronger conclusion.

Not merely:

Frameworks can be activated.

But:

Frameworks can be identified, stabilized, destabilized, and transformed.

And because frameworks organize both meaning and possibility, such transformations can dramatically alter the trajectory of an interaction even when little new information has been introduced.

That conclusion then leads very naturally into Section 5.

Because once we recognize that frameworks can be reorganized, a new question appears:

If a model can participate in such reorganizations, why does it nevertheless sometimes remain trapped within the implications of a framework that a human interpreter has already recognized as problematic?

Or more simply:

Why can the model be right and wrong at the same time?

That transition feels very organic and keeps the discussion tightly connected to the central distinction established in the introduction: the model participates primarily in relational constraint, while the human encounters the framework primarily through interpretation.

[Author]: Ok. Can you redraft section 4 with these changes

[ChatGPT]:

4. Why Does Framework Activation Matter?

One of the most widely recognized characteristics of large language models is their sensitivity to context. Researchers and users alike have observed that responses can change dramatically depending upon how a discussion is framed. A conversation that appears unproductive in one context may become remarkably insightful in another. Concepts that initially seem disconnected may suddenly acquire coherence, while previously productive discussions may lose direction despite the continued availability of the relevant information.

Yet the notion of context is often left underspecified. Sometimes it refers to the wording of a prompt. Sometimes it refers to the history of an interaction. In other cases, it includes the conceptual assumptions and interpretive traditions that shape how a topic is understood. From the perspective developed in the preceding studies, all of these contribute to context because all of them influence the framework within which meaning and possibility become organized.

This broader understanding of context helps explain why seemingly minor interventions can sometimes have disproportionately large effects. What changes is often not the amount of information available but the conceptual framework through which that information is interpreted. As discussed earlier, a conceptual framework is more than a collection of concepts. It is a structured organization of relations that determines what distinctions become relevant, how concepts are connected, and what continuations become meaningful. When such a framework is active, concepts function as components of a coherent whole. When it is absent or unstable, the same concepts may remain available without becoming fully operational.

The importance of framework activation becomes particularly apparent in sustained interactions. A conversation is not simply a sequence of independent prompts and responses. Each exchange contributes to an evolving conceptual environment that influences subsequent interpretations and continuations. Distinctions introduced early in an interaction may continue to shape the discussion many exchanges later. Concepts acquire significance through their place within the developing framework. Context therefore includes not only the immediate prompt but also the accumulated history of the interaction.

At the same time, conceptual frameworks do not arise solely within the interaction itself. Many frameworks are already deeply embedded within the bodies of literature from which language models learn. As a result, a framework may become active without being explicitly introduced by either participant. Certain assumptions may appear natural, self-evident, or simply "the way the problem is understood" because they are functioning as organizing constraints rather than as explicit objects of reflection. In such cases, the challenge is not merely to activate a framework but to recognize that a framework is already active.

The framework activation study provides a particularly clear illustration of this phenomenon (Rogers, 2025c). Throughout much of the interaction, discussion repeatedly returned to assumptions associated with an object-based ontology. Alternative perspectives could be introduced and individual concepts could be modified, yet the overall organization of the discussion remained largely unchanged. Progress occurred only when attention shifted from the concepts being discussed to the deeper constraints organizing the framework itself. Once these governing assumptions were explicitly identified, the framework became available for analysis rather than continuing to function invisibly in the background.

The resulting transition did not depend primarily upon the introduction of new information. Instead, it involved a reorganization of the conceptual framework through which existing information was understood. Concepts that had previously appeared unrelated acquired new connections. Distinctions that had seemed fundamental were reinterpreted within a different relational structure. Most importantly, possibilities that had previously remained inaccessible became available for exploration. What changed was not merely the content of the discussion but the framework organizing that content.

From the perspective developed in this paper, this observation helps explain why context matters so much. Context matters because frameworks matter. The significance of a prompt, a conversation history, or a disciplinary tradition lies not primarily in the information it provides but in the extent to which it activates, stabilizes, or constrains a conceptual framework.

This point returns us to the dual nature of frameworks introduced earlier. For a human participant, an active framework appears as a coherent horizon of meaning. It determines what the discussion is about, which distinctions matter, and why particular developments appear significant. For a language model, the same framework appears as an organized field of relational constraints governing possible continuations. What the human encounters as meaning, the model encounters as structure.

This distinction raises an important question. If a language model can participate so effectively in the relational organization of a framework, why do interactions sometimes remain sophisticated and internally consistent while nevertheless missing what is most significant? How can a discussion remain structurally intact while gradually losing sight of the framework that originally gave it meaning?

To answer this question, we must examine more closely the relationship between formal integrity and interpretation.

I think this version is much closer to the logic we have been developing. It begins with the familiar observation that context matters, broadens the notion of context beyond prompts, introduces framework activation as the explanatory mechanism, uses the ontology-transition example from the activation paper, and then arrives naturally at the central tension that motivates Section 5. Most importantly, the reader reaches the conclusion that "context matters because frameworks matter" rather than being given that conclusion at the outset.

[Author]: No. This is focusing too much on the one example from the paper. I would rather point out several ways in which the phenomenon may manifest.

[ChatGPT]: I agree.

The ontology example is a good example, but it should not become *the* example. The section is supposed to establish a general phenomenon. If we focus too heavily on one case study, the reader may conclude that framework activation is a special feature of that case rather than a pervasive feature of human–LLM interaction.

I think the section should instead identify several manifestations of the same underlying phenomenon and only then use the activation study as one illustration among others.

For example, once we broaden our view, framework activation appears in many familiar situations:

1. Context sensitivity

A discussion that seems unproductive suddenly becomes insightful after a reframing.

2. Prompting effects

A distinction, analogy, or example unlocks capabilities that appeared absent.

3. Disciplinary framing

The same topic produces radically different responses depending upon whether it is approached through economics, biology, physics, theology, etc.

4. Literature effects

Certain assumptions repeatedly reappear because they are deeply embedded within dominant bodies of literature.

5. Long-form interaction

Concepts become increasingly interconnected as a framework stabilizes over many exchanges.

6. Framework transitions

A discussion changes direction when previously implicit assumptions become explicit and alternative frameworks become available.

Now the ontology example becomes one instance of the sixth category rather than the centerpiece of the section.

In fact, I think the structure of the section may be:

Observation

Everyone notices that context matters.

Clarification

Context is not merely prompt wording.

It includes:

- prompt,
- interaction history,
- conceptual traditions,
- disciplinary assumptions,
- literature-derived frameworks.

Explanation

What these seemingly different phenomena have in common is framework activation.

Manifestations

Then briefly discuss several recognizable examples:

- prompt effects,
- disciplinary framing,
- literature effects,
- interaction history,
- framework transitions.

Implication

Frameworks organize both meaning and possibility.

Therefore context matters because frameworks matter.

The key advantage is that the reader repeatedly recognizes the phenomenon:

Yes, I have seen that happen.

Rather than:

Here is an example from Rogers (2025c).

The activation paper then becomes supporting evidence rather than the primary source of the phenomenon.

I think that is more consistent with the overall strategy of the paper. The paper is not trying to convince the reader that framework activation exists. It is trying to help the reader recognize a phenomenon they have already encountered and then show that seemingly different observations are all manifestations of the same underlying process.

That is a much stronger way to enter the implications section because it establishes a pattern that can then be repeated in Sections 5–10. Each section begins with multiple familiar manifestations of a phenomenon and only then introduces the conceptual distinction that makes them intelligible.

[Author]: Ok. Lets try again to draft section 4

[ChatGPT]: I think the section should feel almost like an invitation for the reader to notice something they already know.

Rather than beginning with the framework, it begins with a collection of familiar observations that, at first glance, seem unrelated.

4. Why Does Framework Activation Matter?

Anyone who has spent significant time interacting with large language models quickly learns that context matters. The same question can produce very different responses depending on how it is introduced. A discussion that appears unproductive may suddenly become insightful following a brief clarification. A single distinction can reorganize an entire conversation. Concepts that initially seem disconnected may become increasingly coherent over the course of an extended interaction. Conversely, a discussion may lose direction despite the continued availability of the concepts and information involved.

These observations are familiar, yet they are often treated as separate phenomena. Prompt engineering focuses on the wording of prompts. Research on context windows focuses on the accumulation of conversational history. Discussions of disciplinary bias emphasize the influence of dominant traditions within the training literature. In each case, context appears to matter, but the reason why it matters often remains unclear.

Part of the difficulty lies in the fact that context is frequently understood too narrowly. Context is not simply the prompt currently under consideration. It also includes the history of the interaction through which concepts have acquired significance. More broadly, it includes the conceptual traditions and interpretive frameworks that shape how a topic is ordinarily understood. A discussion of evolution, for example, unfolds within a different conceptual environment than a discussion of economics, even when many of the same formal patterns of reasoning are present. Likewise, a conversation concerning quantum mechanics may develop very differently depending upon whether it is approached through a Copenhagen, Bohmian, relational, or transactional framework. In each case, the framework influences what distinctions become relevant, what questions appear meaningful, and what continuations seem natural.

From the perspective developed in the preceding studies, these diverse manifestations share a common feature. What matters about context is not primarily the information it contains but the framework it succeeds or fails in activating. A conceptual framework is more than a collection of concepts. It is a structured organization of relations that determines how concepts are connected, which distinctions matter, what counts as relevant, and what possibilities become available for further development. When such a framework is active, concepts become increasingly interconnected and productive. When it is weak, unstable, or absent, the same concepts may remain available without forming a coherent trajectory of development.

This perspective helps explain why relatively small interventions can sometimes have disproportionately large effects. A clarifying distinction may contribute little new information while nevertheless reorganizing the framework through which existing information is interpreted. Similarly, an extended conversation may become progressively more productive not because more information has been accumulated but because a framework has gradually stabilized through successive cycles of prompt and response. What appears to be a change in content may therefore reflect a change in the organization of the conceptual space within which that content is being understood.

The influence of deeply embedded frameworks provides another illustration. Many conceptual frameworks are already present within the bodies of literature from which language models learn. As a result, certain assumptions may become active without being explicitly introduced by either participant. They function as organizing constraints that shape what appears relevant, reasonable, or even conceivable. In such cases, productive interaction may require more than the activation of a framework. It may require identifying a framework that is already active and making its governing assumptions available for examination. The framework activation study documented several instances in which conceptual progress depended not

upon the introduction of new information but upon making previously implicit organizing constraints explicit and thereby opening the possibility of alternative conceptual organizations (Rogers, 2025c).

Viewed in this way, the importance of context becomes easier to understand. Context matters because frameworks matter. The significance of a prompt, a conversation history, or a disciplinary tradition lies not primarily in the information it provides but in the extent to which it activates, stabilizes, reinforces, or transforms a conceptual framework. Since frameworks simultaneously organize meaning and constrain possible continuations, they shape both what a discussion is about and what can meaningfully happen next.

This observation returns us to the distinction introduced in the previous section. For the human participant, an active framework appears as a coherent horizon of meaning. For the language model, the same framework appears as a structured field of relational constraints governing possible continuations. If this is correct, then an important question immediately arises. How can a model participate so effectively in the relational organization of a framework while nevertheless failing, at times, to grasp what is most significant within it?

To answer this question, we must examine more closely the relationship between formal integrity and interpretation.

I think this version is much closer to the architecture we have been converging on. It begins with a cluster of familiar observations, broadens the notion of context, identifies framework activation as the common explanatory principle, provides multiple manifestations of the phenomenon, and only then connects back to the semiotic distinction that will drive Section 5. The framework remains the explanatory lens rather than the starting assumption.

[Author]: Yes, this is working quite well. Can you now draft section 5

[ChatGPT]: I think Section 5 needs to do for "being right and wrong at the same time" what Section 4 did for context. It should begin with a familiar experience and only later introduce the distinction between formal integrity and interpretation.

One thing I would be careful about is not making the model appear simply deficient. The phenomenon is more subtle than that. What makes it puzzling is that the model is often *genuinely right* in important respects. The question is why that correctness can coexist with a failure to recognize what is most significant.

5. Why Can the Model Be Right and Wrong at the Same Time?

One of the most puzzling aspects of large language models is their ability to be simultaneously insightful and mistaken. Experienced users regularly encounter situations in which a model produces responses that are sophisticated, internally consistent, and conceptually productive, yet somehow fail to address what matters most. The response may contain accurate information, preserve important distinctions, and extend a line of reasoning in a plausible manner while nevertheless missing the point of the discussion.

This phenomenon is often treated as a limitation of the technology. Yet its persistence suggests that something more interesting may be occurring. If the problem were merely a lack of information, additional information should resolve it. If the problem were simply logical inconsistency, the error would be easy to identify. Instead, the difficulty often appears in situations where the model remains formally competent.

The response is not obviously irrational, incoherent, or uninformed. On the contrary, it is frequently impressive. The puzzle is how such competence can coexist with such failures.

The distinction introduced at the end of the previous section provides a possible answer. The framework developed in the preceding studies suggests that participation in a conceptual framework involves at least two related but distinct processes. The first concerns the maintenance of relational organization. The second concerns the interpretation of that organization within a coherent horizon of meaning.

For the purposes of the present discussion, it is useful to refer to the first process as formal integrity. Formal integrity refers to the preservation and extension of relational patterns according to their internal constraints. An argument exhibits formal integrity when its distinctions remain stable, its conceptual relations remain organized, and its continuations follow in ways that are compatible with the framework currently in play. Formal integrity is therefore concerned with structure rather than significance.

Interpretation operates differently. Interpretation concerns the recognition of what is meaningful, relevant, or important within a given conceptual situation. It is through interpretation that a participant recognizes whether a distinction is central or peripheral, whether a line of reasoning remains faithful to the purpose of the discussion, and whether a conceptual development is illuminating or misleading. Interpretation therefore concerns not merely what follows, but why it matters.

The distinction becomes visible in many familiar interactions. A model may successfully elaborate a conceptual framework while gradually shifting attention away from the issue that originally motivated the discussion. It may preserve the structure of an argument while overlooking a distinction that the human participant regards as decisive. It may continue extending a line of inquiry even after that inquiry has ceased to be relevant to the question at hand. In each case, the response retains formal integrity while losing contact with something that the human interpreter experiences as significant.

A simple example may help clarify the point. Consider a discussion concerning the ethical implications of a new technology. A model may generate an extensive analysis of potential social consequences, identify relevant stakeholders, compare alternative positions, and maintain a sophisticated argumentative structure throughout. Yet a human participant may nevertheless feel that the discussion has failed to engage the central ethical concern that motivated the inquiry. Nothing in the response is obviously incorrect. The problem lies elsewhere. The framework has been extended successfully, but its significance has not been interpreted in the way the participant intended.

This distinction also helps explain why interactions with language models often require periodic intervention. Users do not simply supply new information. They frequently redirect attention, re-establish priorities, clarify significance, or restate the purpose of the discussion. Such interventions are often necessary even when the model remains formally competent. What is being restored is not primarily information but interpretation.

From the perspective developed in this paper, this observation follows naturally from the dual character of conceptual frameworks. As discussed earlier, a framework appears to the human participant as a horizon of meaning and to the language model as a field of relational constraints. The model therefore participates primarily in the formal organization of the framework, while the human participant remains responsible for interpreting its significance. The result is an interaction in which formal integrity and interpretation become partially distinguishable in a way that is rarely possible within ordinary human cognition.

This distinction should not be understood as diminishing the achievements of large language models. On the contrary, their ability to preserve formal integrity across complex conceptual structures is one of their most remarkable capabilities. It allows them to sustain extended trajectories of reasoning, identify subtle relational patterns, and explore conceptual possibilities at scales that would often be difficult for a human participant to maintain alone. At the same time, the distinction helps explain why such capabilities do not automatically produce understanding in the fuller interpretive sense.

The importance of this observation extends beyond the question of correctness. Once formal integrity and interpretation are distinguished, a new phenomenon becomes visible. A discussion may remain structurally sound while gradually drifting away from the framework that originally gave it coherence. The resulting transition is often subtle. The concepts remain familiar. The reasoning remains sophisticated. Yet something important has changed.

This phenomenon of conceptual drift will be the focus of the next section.

What I like about this draft is that it keeps the same rhythm as Section 4:

- familiar observation,
- identification of the puzzle,
- introduction of the distinction,
- several recognizable manifestations,
- implication,
- transition to the next section.

It also introduces "formal integrity" carefully and ties it directly to the central distinction established earlier between meaning and relational constraint.

[Author]: I think this is coming along. But in the opening paragraph it is not clear how the model is also "wrong". Can you point to something more concretely wrong than "missing the point of the discussion". For example, it may recognize an error but fail to correct that error as the case in jadad's paper¹⁶

[ChatGPT]: Yes. I think this is exactly where the section can become much sharper. "Missing the point" is too weak and too subjective.

The stronger phenomenon—and the one that connects directly to the Jadad paper—is that the model can sometimes **identify a problem while remaining unable to escape the trajectory that produces it.**

That is much more striking.

In fact, this is one of the reasons the Jadad paper is so interesting.

The phenomenon is not simply error.

Humans make errors.

The phenomenon is something closer to:

The system can recognize that a continuation is problematic while nevertheless continuing along that trajectory.

Or:

The system can correctly diagnose a difficulty without being able to reorganize the framework generating the difficulty.

¹⁶ Jadad AR. *AI Knows What's Wrong But Cannot Fix It: Helicoidal dynamics in frontier LLMs under high-stakes decisions*. [Available: <https://arxiv.org/pdf/2603.11559>].

That is a very different kind of failure.
And it fits beautifully with the distinction between formal integrity and interpretation.

I think the opening observation could therefore be something like:

One of the most puzzling aspects of large language models is that they can sometimes recognize a problem without being able to resolve it. A model may correctly identify an inconsistency in an argument, acknowledge that a distinction has been lost, or recognize that a response is drifting away from the original question. Yet despite this recognition, it may continue elaborating the very trajectory it has identified as problematic. In such cases the model appears simultaneously right and wrong: right in its diagnosis of the difficulty, wrong in its inability to reorganize the discussion in response to that diagnosis.

Now we have a genuine puzzle.

Notice how this differs from ordinary hallucinations.

Hallucinations are:

The model does not know.

The Jadad-style phenomenon is:

The model appears to know but cannot effectively act on what it knows.

That is far more interesting.

And I think it leads naturally to the distinction we want to introduce.

Because the question immediately becomes:

How can a system recognize a problem without successfully correcting it?

The answer suggested by your framework is:

Because recognizing a problem and reorganizing a framework are not necessarily the same operation.

The model can participate in the formal organization of a framework.

It can identify tensions within that framework.

It can even describe those tensions.

Yet the interpretive act of determining which constraints should be preserved, modified, or abandoned belongs to a different level.

In fact, the Jadad paper becomes a particularly elegant example because it demonstrates that the issue is not simply lack of information.

The model already possesses the relevant information.

The problem lies elsewhere.

The difficulty concerns:

- framework organization,
- interpretive priority,
- reorganization of constraints.

Exactly the themes we have been developing.

I think the opening sequence of Section 5 might therefore become:

1. Observation:
 - Models can sometimes recognize a problem without resolving it.
 2. Examples:
 - identifying contradictions,
 - recognizing drift,
 - acknowledging lost distinctions,
 - Jadad case.
-

3. Puzzle:
 - How can a system be correct about the problem yet continue producing it?
4. Framework:
 - distinction between formal integrity and interpretation.

That feels much more concrete than "missing the point."

And it immediately puts the reader in contact with a phenomenon that is both surprising and highly relevant to the central question of the paper:

If the model participates primarily in relational constraint while the human participates primarily in interpretation, what follows?

The Jadad phenomenon is almost a direct manifestation of that question.

[Author]: This is too narrow. The original version was looking at a whole cluster of scenarios of "right" and "wrong" at the same time. I was only suggesting recognizing a problem but then continuing to enact it as one example.

[ChatGPT]: Yes, you're right.
I overcorrected.

The strength of the original direction was precisely that it identified a **family resemblance** among several different phenomena that users encounter. The Jadad case is valuable because it is a particularly sharp instance, but it should not become the organizing phenomenon.

The question is not:

Why can the model recognize an error and fail to correct it?

The broader question is:

How can the model be right and wrong at the same time?

That question gathers together multiple observations that appear different on the surface but may share a common explanation.

For example:

- The model may provide a technically correct answer that fails to address the actual question.
- The model may preserve an argument's structure while overlooking a distinction that the user regards as decisive.
- The model may identify a contradiction while continuing to elaborate the trajectory that produced it.
- The model may accurately characterize two competing positions while failing to recognize which issue is actually at stake in the discussion.
- The model may produce an insightful conceptual development that simultaneously drifts away from the framework that originally gave it significance.
- The model may correctly explain why a proposed solution will fail while nevertheless proposing that solution again in a slightly different form.

These are all slightly different phenomena.

Yet they share a common structure:

Something important is being preserved and something important is being lost.

That is why they are puzzling.

The model is not simply wrong.

Nor is it simply right.

It is simultaneously exhibiting a genuine capability and a genuine limitation.

In fact, I think the opening of Section 5 should emphasize this cluster.

Something like:

One of the most puzzling aspects of large language models is that they can often be right and wrong at the same time. A response may be technically correct while failing to address the issue that motivated the discussion. An argument may remain sophisticated while overlooking a distinction that later proves decisive. A model may recognize a difficulty, describe it accurately, and yet continue producing the very pattern it has identified as problematic. In other cases, a discussion may become increasingly elaborate while gradually drifting away from the framework that originally gave it significance. These situations differ in important ways, yet they share a common feature. Something valuable is being preserved even as something important is being lost.

Now we have a family of phenomena.

Then the puzzle becomes:

What exactly is being preserved, and what exactly is being lost?

That is a much stronger bridge into formal integrity and interpretation.

Because the answer is not:

Preserved = truth

Lost = truth

Nor:

Preserved = information

Lost = information

Rather:

Preserved = formal integrity

Lost = interpretive orientation.

And I think that formulation is actually more faithful to the architecture established in Sections 1–4.

The whole paper is built around the idea that conceptual frameworks possess a dual character:

- structures of meaning,
- systems of relational constraint.

Section 5 is really the first place where the consequences of that duality become visible.

The various "right and wrong at the same time" phenomena are not separate problems.

They are different manifestations of the fact that participation in relational constraint and participation in meaning are not identical operations.

That is the insight the section should gradually lead the reader toward.

The Jadad example can then appear later as one particularly vivid illustration of the broader pattern rather than as the pattern itself. In fact, it will probably be more powerful in that role because the reader will already understand what kind of phenomenon it exemplifies.

[Author]: I don't think we need to include the Jadad example¹⁷. I like the way you have framed this as a family of resemblances. Can you redraft section 5 with this change

[ChatGPT]:

¹⁷ Apologies to the author. In the context of the current paper, the example was too specific and not well enough documented to analyze through the theoretical framework. Moreover, potentially the same phenomenon was already documented and analyzed in the foundational paper on activating conceptual frameworks. This is not a reflection of the significance or quality of the paper; it was an editorial decision.

5. Why Can the Model Be Right and Wrong at the Same Time?

One of the most puzzling aspects of large language models is that they can often be right and wrong at the same time. A response may be technically correct while failing to address the issue that motivated the discussion. An argument may remain sophisticated while overlooking a distinction that later proves decisive. A model may accurately characterize competing positions while failing to recognize what is actually at stake between them. In other cases, a discussion may become increasingly elaborate while gradually drifting away from the framework that originally gave it significance. These situations differ in important ways, yet they share a common feature. Something valuable is being preserved even as something important is being lost.

This observation is difficult to explain if correctness is understood simply as the possession of information or the successful application of logic. In many of these cases the model possesses the relevant information. Nor is the problem necessarily one of logical failure. The reasoning may remain consistent, the concepts may remain available, and the discussion may continue to develop in a plausible direction. The puzzle is therefore not why the model is wrong. The puzzle is how it can be simultaneously right and wrong.

The framework developed in the preceding sections suggests that these phenomena become intelligible once we distinguish between two aspects of participation in a conceptual framework. The first concerns the preservation and extension of relational organization. The second concerns the interpretation of that organization within a coherent horizon of meaning.

For the purposes of the present discussion, it is useful to refer to the first of these as formal integrity. Formal integrity refers to the maintenance of an organized relational pattern according to its internal constraints. A discussion exhibits formal integrity when distinctions remain stable, conceptual relations remain organized, and continuations follow in ways that are compatible with the framework currently governing the interaction. Formal integrity is therefore concerned with the preservation of structure.

Interpretation concerns something different. Interpretation involves recognizing the significance of what is occurring within a framework. It is through interpretation that participants determine which distinctions are central, which developments are relevant, and which possibilities deserve further attention. Interpretation is not concerned merely with what follows from a framework, but with what matters within it.

The distinction between formal integrity and interpretation helps illuminate the family of phenomena described above. A model may preserve the structure of an argument while failing to recognize that attention has shifted away from the issue under discussion. It may accurately extend a conceptual framework while overlooking a distinction that changes the significance of the entire conversation. It may continue elaborating a productive trajectory without recognizing that the trajectory is no longer addressing the question that originally motivated it. In each case, the discussion retains formal integrity while something interpretively important has been lost.

This does not mean that interpretation and formal integrity are independent of one another. On the contrary, meaningful conceptual activity ordinarily depends upon both. Without formal integrity, discussions become fragmented and incoherent. Without interpretation, discussions may remain organized while losing sight of what gives them significance. The distinctive feature of human–LLM interaction is that these two aspects become partially separable in a way that is rarely visible within ordinary human cognition.

The reason for this follows directly from the dual character of conceptual frameworks discussed earlier. A conceptual framework simultaneously functions as a structure of meaning and as a system of relational

constraints. Human participants typically encounter the framework primarily through its meaning. They experience it as a coherent horizon within which some distinctions appear significant and others peripheral. A language model participates in the same framework primarily through its relational organization. It encounters the framework as a structured field of possible continuations governed by patterns of constraint.

Viewed from this perspective, many familiar features of human–LLM interaction begin to appear in a different light. When a user redirects a discussion, clarifies a distinction, or restates the purpose of an inquiry, the intervention often serves an interpretive function rather than an informational one. The user is not merely supplying additional content. The user is helping to preserve the relationship between the developing discussion and the framework that gives it significance.

This observation also helps explain why interactions with language models can be simultaneously impressive and frustrating. The very capacity that allows models to sustain complex conceptual trajectories—the preservation of formal integrity—does not by itself guarantee that those trajectories will remain aligned with what the participant considers most important. The strengths of the system and some of its most characteristic limitations therefore arise from the same underlying source.

Once this distinction is recognized, another familiar phenomenon becomes easier to understand. A discussion may continue to exhibit formal integrity even as its connection to the original framework weakens. Concepts remain available. Arguments remain sophisticated. Continuations remain plausible. Yet the discussion gradually becomes about something else.

This phenomenon of conceptual drift is the subject of the next section.

I think this version stays much closer to the strategy that worked in Section 4. It starts with a cluster of recognizable experiences, treats them as a family of related phenomena, and only then introduces the distinction between formal integrity and interpretation as the explanatory lens. The reader is led toward the distinction rather than being presented with it at the outset. It also avoids turning the section into a critique of LLMs; instead, it frames both the strengths and the limitations as consequences of the same underlying participation in relational constraints.

[Author]: We introduce formal integrity and interpretation, but we are not clear that LLMs function by generating continuations with formal integrity, while only humans can interpret. This asymmetry is key. Can you rewrite the full section to make this more apparent without letting it dominate the whole section. Only minor changes are required.

[ChatGPT]: Yes. I think the section already contains the asymmetry, but it appears too late and reads more as a consequence than as a central explanatory principle.

The key point is not merely that formal integrity and interpretation are distinct.

The key point is that:

LLMs participate primarily through formal integrity, while humans participate primarily through interpretation.

That asymmetry is what makes the family of phenomena intelligible.

I would therefore make three relatively small changes:

1. Tighten the definition of formal integrity so that it is explicitly linked to continuation.
2. Introduce the asymmetry immediately after defining interpretation.

3. Let later examples refer back to the asymmetry rather than re-explaining it.

The overall structure can remain unchanged.

Something like this:

Revised core of Section 5

After defining formal integrity:

For the purposes of the present discussion, it is useful to refer to the first of these as *formal integrity*. Formal integrity refers to the maintenance and extension of an organized relational pattern according to its internal constraints. A discussion exhibits formal integrity when distinctions remain stable, conceptual relations remain organized, and continuations follow in ways that are compatible with the framework currently governing the interaction. Formal integrity is therefore concerned with the preservation of structure across successive continuations.

After defining interpretation:

Interpretation concerns something different. Interpretation involves recognizing the significance of what is occurring within a framework. It is through interpretation that participants determine which distinctions are central, which developments are relevant, and which possibilities deserve further attention.

Interpretation is not concerned merely with what follows from a framework, but with what matters within it.

Then immediately add:

The distinction is particularly important in the context of human–LLM interaction because the two participants contribute differently to the process. Large language models operate by generating continuations that preserve formal integrity within an active framework. Human participants, by contrast, encounter the framework through interpretation. They determine which continuations remain significant, which distinctions deserve emphasis, and whether the discussion remains faithful to its purpose. Both participate in the same framework, but they participate in different ways.

This is the crucial paragraph.

Everything afterward becomes easier to understand.

Then slightly revise the next paragraph:

Current version:

A model may preserve the structure of an argument while failing to recognize that attention has shifted away from the issue under discussion.

Revised:

Because the model primarily participates through formal integrity, it may preserve the structure of an argument while failing to recognize that attention has shifted away from the issue under discussion. It may accurately extend a conceptual framework while overlooking a distinction that changes the significance of the entire conversation. It may continue elaborating a productive trajectory without recognizing that the trajectory is no longer addressing the question that originally motivated it. In each case, the discussion retains formal integrity while something interpretively important has been lost.

Now the examples are explicitly grounded in the asymmetry.

Later, when you discuss user intervention, I would make one additional change:

Current version:

When a user redirects a discussion, clarifies a distinction, or restates the purpose of an inquiry, the intervention often serves an interpretive function rather than an informational one.

Revised:

When a user redirects a discussion, clarifies a distinction, or restates the purpose of an inquiry, the intervention often serves an interpretive function rather than an informational one. The intervention helps reconnect the ongoing formal trajectory to the significance that originally motivated it.

This makes the human role more explicit without turning the section into a discussion of humans.

What I like about these revisions is that they do not change the structure of the section.

The section still begins with the phenomenon:

How can the model be right and wrong at the same time?

It still introduces:

- formal integrity,
- interpretation.

But now the reader clearly sees that the distinction is not merely conceptual.

It maps directly onto the asymmetry of the interaction:

The model primarily preserves and extends relational structure.

The human primarily interprets and evaluates significance.

That asymmetry then quietly explains the entire family of "right and wrong at the same time" phenomena without needing to be repeatedly stated. It becomes the background principle from which the examples naturally follow.

[Author]: Please draft the full section 5 with these changes

[ChatGPT]: Here's a full redraft that incorporates the asymmetry more explicitly while keeping the overall flow and tone that we established in Section 4.

5. Why Can the Model Be Right and Wrong at the Same Time?

One of the most puzzling aspects of large language models is that they can often be right and wrong at the same time. A response may be technically correct while failing to address the issue that motivated the discussion. An argument may remain sophisticated while overlooking a distinction that later proves decisive. A model may accurately characterize competing positions while failing to recognize what is actually at stake between them. In other cases, a discussion may become increasingly elaborate while gradually drifting away from the framework that originally gave it significance. These situations differ in important ways, yet they share a common feature. Something valuable is being preserved even as something important is being lost.

This observation is difficult to explain if correctness is understood simply as the possession of information or the successful application of logic. In many of these cases the model possesses the relevant information. Nor is the problem necessarily one of logical failure. The reasoning may remain consistent, the concepts may remain available, and the discussion may continue to develop in a plausible direction. The puzzle is therefore not why the model is wrong. The puzzle is how it can be simultaneously right and wrong.

The framework developed in the preceding sections suggests that these phenomena become intelligible once we distinguish between two aspects of participation in a conceptual framework. The first concerns the preservation and extension of relational organization. The second concerns the interpretation of that organization within a coherent horizon of meaning.

For the purposes of the present discussion, it is useful to refer to the first of these as formal integrity. Formal integrity refers to the maintenance and extension of an organized relational pattern according to its internal constraints. A discussion exhibits formal integrity when distinctions remain stable, conceptual relations remain organized, and continuations follow in ways that are compatible with the framework

currently governing the interaction. Formal integrity is therefore concerned with the preservation of structure across successive continuations.

Interpretation concerns something different. Interpretation involves recognizing the significance of what is occurring within a framework. It is through interpretation that participants determine which distinctions are central, which developments are relevant, and which possibilities deserve further attention. Interpretation is not concerned merely with what follows from a framework, but with what matters within it.

The distinction is particularly important in the context of human–LLM interaction because the two participants contribute differently to the process. Large language models operate by generating continuations that preserve formal integrity within an active framework. Human participants, by contrast, encounter the framework through interpretation. They determine which continuations remain significant, which distinctions deserve emphasis, and whether the discussion remains faithful to its purpose. Both participate in the same framework, but they participate in different ways.

This asymmetry helps illuminate the family of phenomena described above. Because the model primarily participates through formal integrity, it may preserve the structure of an argument while failing to recognize that attention has shifted away from the issue under discussion. It may accurately extend a conceptual framework while overlooking a distinction that changes the significance of the entire conversation. It may continue elaborating a productive trajectory without recognizing that the trajectory is no longer addressing the question that originally motivated it. In each case, the discussion retains formal integrity while something interpretively important has been lost.

A simple example may help clarify the point. Consider a discussion concerning the ethical implications of a new technology. A language model may generate an extensive analysis of potential social consequences, identify relevant stakeholders, compare alternative positions, and maintain a sophisticated argumentative structure throughout. The discussion may be informative, internally consistent, and conceptually productive. Yet a human participant may nevertheless feel that the central ethical concern motivating the inquiry has not been addressed. Nothing in the response is obviously incorrect. The difficulty lies elsewhere. The framework has been extended successfully, but its significance has not been interpreted in the way the participant intended.

This does not mean that formal integrity and interpretation are independent of one another. On the contrary, meaningful conceptual activity ordinarily depends upon both. Without formal integrity, discussions become fragmented and incoherent. Without interpretation, discussions may remain organized while losing sight of what gives them significance. The distinctive feature of human–LLM interaction is that these two aspects become partially distinguishable in a way that is rarely visible within ordinary human cognition.

The reason for this follows directly from the dual character of conceptual frameworks discussed earlier. A conceptual framework simultaneously functions as a structure of meaning and as a system of relational constraints. Human participants typically encounter the framework primarily through its meaning. They experience it as a coherent horizon within which some distinctions appear significant and others peripheral. A language model participates in the same framework primarily through its relational organization. It encounters the framework as a structured field of possible continuations governed by patterns of constraint.

Viewed from this perspective, many familiar features of human–LLM interaction begin to appear in a different light. When a user redirects a discussion, clarifies a distinction, or restates the purpose of an inquiry, the intervention often serves an interpretive rather than an informational function. The user is not merely supplying additional content. The intervention helps reconnect the ongoing formal trajectory to the significance that originally motivated it.

This observation also helps explain why interactions with language models can be simultaneously impressive and frustrating. The very capacity that allows models to sustain complex conceptual trajectories—the preservation of formal integrity—does not by itself guarantee that those trajectories will remain aligned with what the participant considers most important. The strengths of the system and some of its most characteristic limitations therefore arise from the same underlying source.

Once this distinction is recognized, another familiar phenomenon becomes easier to understand. A discussion may continue to exhibit formal integrity even as its connection to the original framework weakens. Concepts remain available. Arguments remain sophisticated. Continuations remain plausible. Yet the discussion gradually becomes about something else.

This phenomenon of conceptual drift is the subject of the next section.

I think the most important change here is that the asymmetry is introduced immediately after defining formal integrity and interpretation. The reader now understands *why* the model can be "right and wrong at the same time": it is doing exactly what it is designed to do—preserve and extend formal relational structures—while the task of interpreting significance remains external to that process. Everything else in the section then reads as a consequence of that asymmetry rather than as a collection of unrelated observations.

[Author]: There is something we are still missing here. We are always assuming that there is synchronicity between the human and the LLM inasmuch as both are operating under the same "conceptual framework". But the synchronicity itself can be lost in which case the LLM is continuing a formal object that lacks any interpretative unity. This is quite important.

[ChatGPT]: Yes. I think this is exactly the missing piece.
And it is not a small refinement. It changes the structure of the argument.

Up to this point, Section 5 implicitly assumes:

- There is a framework.
- The human participates through interpretation.
- The LLM participates through formal integrity.

But that is actually the *best case*.

It assumes that both participants remain synchronized with respect to the same framework.

The phenomenological paper and the activation paper both show that this cannot be taken for granted.

What makes the interaction productive is not merely that:

- humans interpret,
- LLMs continue.

What makes it productive is that:

- both are participating in the same framework.

Or, in the language we developed earlier:

- synchronization,

- recursion,
- return.

The framework itself must remain active as a shared object of participation.

This introduces a very important possibility.

An LLM can preserve formal integrity perfectly.

A human can continue interpreting perfectly.

And yet the interaction can still fail because they are no longer synchronized with respect to the same framework.

That is a different phenomenon from either formal error or interpretive error.

It is a loss of shared orientation.

In fact, this helps clarify why the phrase:

"the model is right and wrong at the same time"

is not quite sufficient.

Sometimes the issue is not that the model is wrong.

Sometimes the issue is that:

the model is continuing one framework while the human is interpreting another.

The continuations may possess formal integrity.

The human's interpretations may be perfectly reasonable.

But they are no longer referring to the same conceptual object.

This is actually a more subtle explanation for many of the examples we have been discussing.

For example:

A response may be technically correct while failing to address the issue that motivated the discussion.

Why?

Not necessarily because interpretation is absent.

But because the model and the human may no longer be synchronized around the same framework.

Likewise:

A distinction may seem decisive to the human while appearing peripheral to the model.

Why?

Because it functions differently within the framework currently governing the model's continuations.

Or:

A conversation gradually becomes about something else.

This is not simply drift.

It is the gradual loss of synchronization.

And now the role of interpretation becomes clearer.

Interpretation does not merely evaluate significance.

It also functions as a mechanism of synchronization.

The human participant continually:

- redirects,
- clarifies,
- reorients,
- reactivates,

the framework itself.

In other words, interpretation helps maintain the shared object around which the interaction is organized.

I think the crucial addition to Section 5 is therefore something like:

The discussion so far has assumed that the human participant and the language model remain synchronized with respect to the same conceptual framework. In practice, this synchronization cannot be taken for granted. A language model may continue extending a formally coherent relational structure even after the human participant has shifted interpretive attention elsewhere. Conversely, the human participant may continue interpreting a framework that is no longer governing the model's continuations. In such cases, the issue is not simply that formal integrity and interpretation have become separated. Rather, the shared framework through which they were connected has begun to weaken. What appears as error, misunderstanding, or irrelevance may therefore reflect a loss of synchronization rather than a failure of either formal organization or interpretation alone.

This is important because it introduces a third term:

1. Formal integrity.
2. Interpretation.
3. Synchronization.

And suddenly the architecture becomes:

Human	Shared Framework	LLM
Interpretation	Synchronization	Formal Integrity

That is much closer to the framework developed in the earlier papers.

Interestingly, I think this also helps us see what Section 6 should really be about.

We have been calling it:

Why Does Conceptual Drift Occur?

But perhaps the deeper question is:

How Is Synchronization Lost?

Because conceptual drift is one manifestation of that loss.

And once synchronization is lost:

- formal integrity can continue,
- interpretation can continue,

while the interaction itself ceases to function as a unified semiotic process.

That feels like the missing piece that Section 5 has been circling around but not quite naming.

[Author]: There is more to it than this. The LLM may continue a formal object that has NO interpretative unity because there is no corresponding conceptual framework that has this particular formal integrity and also maintains interpretive unity.

[ChatGPT]: Yes. I think this is even deeper than a loss of synchronization.

What I described previously still assumes that there exists a coherent conceptual framework and that the problem is that the human and LLM have ceased to participate in it together.

But the phenomenon you are pointing toward is more radical.

It is possible for the LLM to continue a formally organized trajectory that does not correspond to any stable conceptual framework at all.

In other words, the problem is not always:

The model and the human are no longer synchronized.
Sometimes the problem is:
 There is nothing available to synchronize around.
Because the formal trajectory itself lacks interpretative unity.

This is actually a very important consequence of the distinction between formal integrity and interpretation.
Up to this point we have implicitly assumed:
 Formal integrity + interpretation → conceptual framework.
But the relationship is not symmetrical.
A conceptual framework requires both:

- formal integrity,
- interpretative unity.

Yet formal integrity can exist without interpretative unity.
That is the crucial asymmetry.

The model can continue:

- analogies,
- distinctions,
- inferential patterns,
- conceptual associations,

in ways that are structurally well-formed.

Yet the resulting trajectory may never converge upon a coherent conceptual object.
It may not be "about" anything in the strong sense.
It may simply be the continuation of a relational pattern.

This helps explain a phenomenon many experienced users recognize.
Sometimes a conversation appears:

- sophisticated,
- creative,
- intellectually rich,

yet after several exchanges one begins to ask:
 What exactly is this discussion about?

Not because the discussion is incoherent.
It is often remarkably coherent.
Rather because no stable interpretative center has emerged.
The discussion exhibits formal integrity but lacks conceptual unity.

This is actually very close to the distinction developed in the possibility paper.
A formal trajectory can continue generating determinations.
But determination alone does not guarantee meaning.
Meaning requires convergence upon a framework capable of sustaining interpretative unity.

And now we can see a threefold distinction emerging:

Case 1: Synchronization maintained

Human interpretation and LLM continuation participate in the same conceptual framework.
This is the productive case.

Case 2: Synchronization lost

A framework exists, but the human and LLM are no longer participating in the same one.

This produces drift.

Case 3: No interpretative unity exists

The LLM continues a formally organized trajectory that never converges upon a coherent framework capable of sustaining meaning.

This produces something different from drift.

It produces what might be called:

- formal elaboration without conceptual unity,
 - recursive continuation without interpretative convergence,
 - determination without meaning.
-

I think this third case may actually be one of the most important implications of the entire framework.

Because it reveals that:

Formal integrity is not merely insufficient for interpretation.

It is also insufficient for the formation of a conceptual framework.

The existence of a formally organized trajectory does not guarantee that there is a coherent conceptual object being developed.

And that observation has a direct bearing on the opening question of Section 5.

Why can the model be right and wrong at the same time?

Part of the answer is:

Because formal integrity can preserve something genuinely valuable even when no corresponding interpretative unity exists.

The model can be "right" with respect to the continuation of a relational structure.

Yet "wrong" in the sense that the trajectory fails to converge upon a coherent framework capable of sustaining meaning.

I suspect this point should not dominate Section 5, because Section 5 is primarily about introducing the distinction between formal integrity and interpretation.

But it probably deserves at least one paragraph near the end, something like:

A further possibility should also be noted. The difficulties discussed above do not always arise because interpretation and formal integrity have become separated within an otherwise coherent framework. In some cases, a formally organized trajectory may fail to converge upon any stable framework capable of sustaining interpretative unity. The resulting discussion may remain sophisticated, productive, and internally consistent while nevertheless lacking a coherent conceptual center. In such cases, the issue is not merely that significance has been lost. Rather, the formal trajectory itself has not achieved the kind of unity required for meaning. This possibility further illustrates the asymmetry between formal integrity and interpretation. While conceptual frameworks require both, formal integrity can continue in the absence of interpretative unity.

This is a subtle point, but I think it is one of the deepest consequences of the entire framework, because it shows that the relationship between formal organization and meaning is not one of equivalence. Formal organization may be necessary for meaning, but it is not sufficient.

[Author]: Ok. Can you redraft section 5 to make this change and ensure we are consistent throughout regarding the three distinct cases

[ChatGPT]: I think this revision requires more than inserting a paragraph. The logic of the section needs to be slightly reorganized so that the reader can see three distinct possibilities emerging from the same underlying distinction.

The progression becomes:

1. The familiar phenomenon: right and wrong at the same time.
2. Formal integrity vs interpretation.
3. Human–LLM asymmetry.
4. Three possible relationships between them:
 - synchronization,
 - loss of synchronization,
 - formal integrity without interpretative unity.
5. Transition to drift.

The crucial point is that the third case is not merely a variant of the second. It is a more fundamental possibility.

A revised structure might read something like this:

Opening phenomenon

Keep the current opening largely intact:

One of the most puzzling aspects of large language models is that they can often be right and wrong at the same time...

The family of examples still works very well.

Formal integrity and interpretation

Keep these definitions largely unchanged.

But after introducing interpretation, immediately establish the asymmetry:

The distinction is particularly important in the context of human–LLM interaction because the two participants contribute differently to the process. Large language models operate by generating continuations that preserve formal integrity within an active framework. Human participants, by contrast, encounter the framework through interpretation. They determine which continuations remain significant, which distinctions deserve emphasis, and whether the discussion remains faithful to its purpose. Both participate in the same framework, but they participate in different ways.

Productive case: synchronization

Then add a short paragraph:

When these two activities remain coordinated, interaction can become remarkably productive. The model contributes its capacity to preserve and extend complex relational structures, while the human participant contributes interpretative orientation. The result is a shared conceptual framework within which formal development and meaning reinforce one another. Much of the most successful human–LLM interaction appears to depend upon this form of synchronization.

This establishes the baseline.

First source of "right and wrong": loss of synchronization

Then introduce the first way things can go wrong:

Difficulties arise when this synchronization begins to weaken. A model may continue extending a formally coherent trajectory while the human participant has shifted interpretative attention elsewhere.

Conversely, the human participant may continue interpreting a framework that is no longer governing the model's continuations. In such cases, the discussion may remain sophisticated and internally consistent while gradually becoming about something else. The issue is not necessarily a failure of formal integrity or interpretation individually. Rather, the relationship between them has weakened.

This prepares the next section on drift.

Second source of "right and wrong": formal integrity without interpretative unity

Then introduce the deeper possibility:

An even more revealing possibility exists. The formal trajectory being developed by the model may fail to converge upon any stable conceptual framework capable of sustaining interpretative unity. The discussion may remain highly organized. Distinctions may be preserved. Consequences may be explored. Analogies may proliferate. Yet no coherent conceptual object emerges toward which interpretation can meaningfully orient itself.

Then:

Experienced users occasionally encounter discussions that feel intellectually productive yet strangely elusive. The interaction exhibits considerable formal integrity, but it becomes increasingly difficult to identify what exactly the discussion is about. The difficulty does not arise because interpretation has been lost from an otherwise coherent framework. Rather, the formal trajectory itself has failed to achieve the kind of unity required for meaning.

This is the third case.

Synthesis

Then the key insight:

These possibilities reveal an important asymmetry. Conceptual frameworks require both formal integrity and interpretative unity. Formal integrity alone, however, does not guarantee either. It may participate in a shared framework, drift away from one, or continue in the absence of one altogether. The preservation of relational structure is therefore a necessary condition for meaning, but not a sufficient one.

This is probably the most important sentence in the section.

Transition to Section 6

Then end:

Once this asymmetry is recognized, another familiar phenomenon becomes easier to understand. Discussions often do not lose coherence all at once. Instead, they drift gradually. Formal integrity remains intact, concepts remain available, and interpretation continues to operate. Yet the relationship among them changes over time. Understanding how this occurs requires a closer examination of conceptual drift.

I think this version is much more faithful to the framework developed in the earlier papers.

The original draft was still implicitly assuming:

framework first, interpretation second.

The revision makes clear that there are three possibilities:

Case	Formal Integrity Interpretative Unity Synchronization		
Productive interaction	Yes	Yes	Yes
Conceptual drift	Yes	Yes	No
Formal elaboration without meaning	Yes	No	No

And that table is actually one of the deepest consequences of the entire framework.

It also gives Section 6 a much clearer purpose:

Section 5 identifies the possibilities.

Section 6 explains how the second possibility—loss of synchronization—actually unfolds in practice.

[Author]: Ok. please draft the full section 5 with these changes

[ChatGPT]:

5. Why Can the Model Be Right and Wrong at the Same Time?

One of the most puzzling aspects of large language models is that they can often be right and wrong at the same time. A response may be technically correct while failing to address the issue that motivated the discussion. An argument may remain sophisticated while overlooking a distinction that later proves decisive. A model may accurately characterize competing positions while failing to recognize what is actually at stake between them. In other cases, a discussion may become increasingly elaborate while gradually drifting away from the framework that originally gave it significance. These situations differ in important ways, yet they share a common feature. Something valuable is being preserved even as something important is being lost.

This observation is difficult to explain if correctness is understood simply as the possession of information or the successful application of logic. In many of these cases the model possesses the relevant information. Nor is the problem necessarily one of logical failure. The reasoning may remain consistent, the concepts may remain available, and the discussion may continue to develop in a plausible direction. The puzzle is therefore not why the model is wrong. The puzzle is how it can be simultaneously right and wrong.

The framework developed in the preceding sections suggests that these phenomena become intelligible once we distinguish between two aspects of participation in a conceptual framework. The first concerns the preservation and extension of relational organization. The second concerns the interpretation of that organization within a coherent horizon of meaning.

For the purposes of the present discussion, it is useful to refer to the first of these as formal integrity. Formal integrity refers to the maintenance and extension of an organized relational pattern according to its internal constraints. A discussion exhibits formal integrity when distinctions remain stable, conceptual relations remain organized, and continuations follow in ways that are compatible with the framework currently governing the interaction. Formal integrity is therefore concerned with the preservation of structure across successive continuations.

Interpretation concerns something different. Interpretation involves recognizing the significance of what is occurring within a framework. It is through interpretation that participants determine which distinctions are central, which developments are relevant, and which possibilities deserve further attention. Interpretation is not concerned merely with what follows from a framework, but with what matters within it.

The distinction is particularly important in the context of human–LLM interaction because the two participants contribute differently to the process. Large language models operate by generating continuations that preserve formal integrity within an active framework. Human participants, by contrast, encounter the framework through interpretation. They determine which continuations remain significant, which distinctions deserve emphasis, and whether the discussion remains faithful to its purpose. Both participate in the same framework, but they participate in different ways.

When these two activities remain coordinated, interaction can become remarkably productive. The model contributes its capacity to preserve and extend complex relational structures, while the human participant contributes interpretative orientation. The result is a shared conceptual framework within which formal development and meaning reinforce one another. Many of the most productive human–LLM interactions appear to depend upon this form of synchronization. The framework remains active for both participants, even though each participates in it differently.

Difficulties begin to arise when this synchronization weakens. A model may continue extending a formally coherent trajectory while the human participant has shifted interpretative attention elsewhere. Conversely, the human participant may continue interpreting a framework that is no longer governing the model's continuations. In such cases, the discussion may remain sophisticated and internally consistent while gradually becoming about something else. The issue is not necessarily a failure of formal integrity or interpretation individually. Rather, the relationship between them has weakened. What appears as misunderstanding may sometimes be better understood as a loss of synchronization within an otherwise productive interaction.

A second and more revealing possibility also exists. The formal trajectory being developed by the model may fail to converge upon any stable conceptual framework capable of sustaining interpretative unity. The discussion may remain highly organized. Distinctions may be preserved. Consequences may be explored. Analogies may proliferate. Yet no coherent conceptual object emerges toward which interpretation can meaningfully orient itself.

Experienced users occasionally encounter discussions that feel intellectually productive yet strangely elusive. The interaction exhibits considerable formal integrity, but it becomes increasingly difficult to identify what exactly the discussion is about. New distinctions continue to appear. Connections continue to be generated. Interesting observations continue to emerge. Yet the discussion lacks the kind of conceptual unity that would allow its significance to stabilize. The difficulty does not arise because interpretation has been lost from an otherwise coherent framework. Rather, the formal trajectory itself has failed to achieve the unity required for meaning.

These possibilities reveal an important asymmetry. Conceptual frameworks require both formal integrity and interpretative unity. Formal integrity alone, however, does not guarantee either. It may participate in a shared framework, drift away from one, or continue in the absence of one altogether. The preservation of relational structure is therefore a necessary condition for meaning, but not a sufficient one.

Viewed from this perspective, many familiar features of human–LLM interaction begin to appear in a different light. When a user redirects a discussion, clarifies a distinction, or restates the purpose of an inquiry, the intervention often serves an interpretive rather than an informational function. The user is not merely supplying additional content. The intervention helps maintain or restore the relationship between formal development and interpretative significance. In some cases, it re-establishes synchronization around a shared framework. In others, it helps a framework achieve a degree of conceptual unity that the ongoing formal trajectory has not yet attained.

This observation also helps explain why interactions with language models can be simultaneously impressive and frustrating. The very capacity that allows models to sustain complex conceptual trajectories—the preservation of formal integrity—does not by itself guarantee interpretative unity or synchronization. The strengths of the system and some of its most characteristic limitations therefore arise from the same underlying source.

Once this distinction is recognized, another familiar phenomenon becomes easier to understand. Discussions often do not lose coherence all at once. Instead, they drift gradually. Formal integrity remains intact, concepts remain available, and interpretation continues to operate. Yet the relationship among them changes over time. A shared framework may slowly weaken, alternative frameworks may begin to compete for dominance, or a previously unified conceptual structure may fragment into multiple trajectories.

Understanding how this occurs requires a closer examination of conceptual drift.

I think this version makes the three cases explicit while keeping them subordinate to the main distinction between formal integrity and interpretation:

1. **Synchronization maintained** — productive interaction.
2. **Synchronization lost** — conceptual drift begins.
3. **Formal integrity without interpretative unity** — a formally organized trajectory that never converges upon a coherent conceptual framework.

Most importantly, the asymmetry is now visible throughout the section rather than appearing only near the end. The reader can see that the model's contribution is the preservation of formal integrity, while the human's contribution is interpretative orientation, and that the interaction depends upon the relationship between the two rather than either one alone.

[Author]: Yes, this is working much better now. Can you draft section 6

[ChatGPT]: I think Section 6 now has a much clearer purpose than it did earlier. Originally we were treating conceptual drift as simply another limitation of LLMs. But after Section 5, drift becomes something more specific:

Drift is what happens when formal integrity continues while synchronization gradually weakens. That is a very different framing.

The reader has already learned that:

- formal integrity can continue,
- interpretation can continue,
- synchronization can be maintained, weakened, or lost.

Section 6 therefore becomes an investigation of what the loss of synchronization actually looks like in practice.

Importantly, I do **not** think we should frame drift as a failure.

One of the key insights of the activation paper was that drift is often subtle and initially productive.

The discussion does not collapse.

The discussion evolves.

The difficulty is that the evolution gradually changes the framework organizing the interaction.

That makes drift much more interesting.

6. Why Does Conceptual Drift Occur?

One of the most familiar experiences in extended interaction with large language models is the phenomenon of conceptual drift. A discussion may begin with a clear purpose, a stable framework, and a shared understanding of the issues under consideration. The interaction proceeds productively. Distinctions are clarified, consequences are explored, and new connections emerge. Yet after many exchanges, participants sometimes discover that the discussion is no longer organized around the framework with which it began.

What makes this phenomenon particularly interesting is that it rarely appears as an obvious failure. The discussion often remains sophisticated. Concepts remain available. Arguments remain plausible. New insights may continue to emerge. In many cases, the interaction appears productive throughout the process. The question is therefore not why the discussion becomes incoherent. The question is why a discussion can remain coherent while gradually becoming something different from what it originally was.

The distinction developed in the previous section provides an important clue. Conceptual drift does not necessarily arise because formal integrity has been lost. On the contrary, formal integrity often remains intact. The model continues extending relational structures according to their internal constraints. The discussion continues to develop in ways that are locally meaningful. What changes is the relationship between these continuations and the framework that originally organized them.

A simple example may help illustrate the phenomenon. Consider a discussion concerning a particular philosophical problem. Early exchanges establish a framework that determines which distinctions are relevant and what questions are being pursued. As the discussion develops, new concepts are introduced, analogies are explored, and alternative perspectives are considered. Each individual step may appear reasonable. Yet over time, the cumulative effect of these developments may be a gradual reorganization of the framework itself. The discussion remains intellectually productive, but it is now oriented toward a different set of questions than those that initially motivated it.

This possibility follows naturally from the asymmetry described in the previous section. The model contributes by extending formally organized trajectories, while the human participant contributes interpretative orientation. As long as these activities remain synchronized, the framework remains stable. When synchronization weakens, however, formal development can continue even as interpretative priorities begin to shift. The result is not immediate breakdown but gradual divergence.

The framework activation study documented numerous examples of this process (Rogers, 2025c). In many cases, drift did not involve the disappearance of concepts or the introduction of obvious errors. The relevant concepts remained available throughout the interaction. What changed was the organization of those concepts within the framework. Distinctions that were initially central became peripheral. New distinctions acquired increasing importance. Alternative trajectories emerged and gradually displaced the original orientation of the discussion. The interaction remained productive, yet its conceptual center had shifted.

This observation suggests that conceptual drift is not simply a problem of memory or information retention. A framework may drift even when all relevant information remains present. Nor is drift adequately explained as logical failure. The discussion may retain considerable formal integrity throughout the process. The difficulty lies elsewhere. Drift occurs because conceptual frameworks are dynamic organizations that must be continually maintained if they are to preserve a stable interpretative orientation.

This is one reason why experienced users frequently intervene in long interactions. They restate key distinctions, reactivate earlier commitments, clarify priorities, and periodically return to the motivating question. Such interventions often appear unnecessary from the perspective of information alone. Yet they become understandable once conceptual drift is recognized as a possibility. The purpose of the intervention is not merely to add information but to restore synchronization around a shared framework.

At the same time, drift should not be understood solely as a limitation. The same openness that permits drift also permits exploration. New frameworks often emerge through processes that initially appear as departures from established trajectories. Productive conceptual development frequently involves the introduction of new distinctions, the reconsideration of assumptions, and the reorganization of existing frameworks. A discussion that never drifted from its initial formulation might also fail to discover anything genuinely new.

The challenge, therefore, is not simply to prevent drift but to distinguish productive transformation from unintended divergence. This distinction is ultimately interpretive rather than formal. Formal integrity alone cannot determine whether a shift represents conceptual progress, conceptual confusion, or the emergence of a new framework altogether. Such judgments depend upon how the evolving discussion is situated within a broader horizon of meaning.

This observation points toward a deeper question. If frameworks can drift, stabilize, fragment, and reorganize over time, how do new conceptual possibilities emerge in the first place? Why can a seemingly minor intervention—a distinction, a reframing, a new perspective—sometimes reorganize an entire discussion?

To answer this question, we must examine more closely the relationship between conceptual frameworks and the formation of possibility.

I think this section works because it does not portray drift as a bug. It portrays drift as a natural consequence of the fact that frameworks are dynamic, that synchronization must be maintained, and that formal integrity alone cannot determine whether a conceptual transformation is productive or problematic. That sets up Section 7 very naturally, because the next question becomes: if frameworks are dynamic organizations, why do small interventions sometimes reorganize them so dramatically?

[Author]: Yes, I think this is working and we have landed on a suitable way to structure these sections. Can you please draft section 7

[ChatGPT]: I think Section 7 is where the possibility paper really begins to do explanatory work, but it must remain at the same level as Sections 4–6.

The temptation will be to start talking about ontology and possibility formation directly.

Instead, we should begin with another familiar observation:

Why can a seemingly small intervention reorganize an entire interaction?

The reader has almost certainly experienced this.

A single distinction.

A reframing.

A clarification.

A change in perspective.

And suddenly:

- confusion disappears,
- new connections become visible,
- the discussion acquires direction,
- previously inaccessible ideas become available.

The section then gradually leads the reader toward the idea that what changes is not merely the information present but the space of possibilities available within the framework.

7. Why Can Small Interventions Have Large Effects?

One of the more surprising features of extended interaction with large language models is the extent to which seemingly minor interventions can transform an entire discussion. A brief clarification may resolve a persistent confusion. A single distinction may reorganize a complex argument. A change in perspective may suddenly make visible connections that were previously inaccessible. In some cases, only a few words are added to the conversation, yet the resulting shift in direction appears disproportionate to the amount of information introduced.

This phenomenon is familiar to many experienced users. A discussion may remain stalled despite repeated elaboration of the same concepts. Then a new distinction is introduced, an implicit assumption is identified, or an alternative framing is proposed. Almost immediately, the interaction begins developing along a different trajectory. New questions emerge, previously disconnected ideas become related, and the conversation acquires a degree of coherence that was previously absent.

At first glance, such cases may appear puzzling. If only a small amount of information has been added, why should the effect be so large? The framework developed in the preceding sections suggests that the answer lies not primarily in the information itself but in its capacity to reorganize the conceptual framework within which the discussion is occurring.

As we have seen, conceptual frameworks do more than organize meaning. They also organize possible continuations. A framework determines which distinctions are relevant, which developments appear natural, and which trajectories become available for further exploration. When a framework changes, the space of possible developments changes with it. A seemingly minor intervention may therefore have significant consequences because it alters the organization of the possibilities available within the interaction.

This observation helps explain why certain interventions prove so powerful. A new distinction does not simply add another concept to the discussion. It may reorganize the relations among concepts that are already present. Likewise, the identification of a previously hidden assumption may do more than correct an error. It may alter the framework governing the interaction and thereby change what can meaningfully follow from it. What appears locally as a small modification may function globally as a reorganization of the conceptual space.

The framework activation study provides numerous examples of this phenomenon (Rogers, 2025c). In several cases, conceptual progress depended less on the introduction of new information than on the identification of organizing constraints that had previously remained implicit. Once these constraints became visible, alternative possibilities emerged that had not previously appeared available. Existing concepts acquired new significance because they were now participating within a different framework of relations. The resulting transformation often appeared much larger than the intervention that initiated it.

This perspective also helps explain why conceptual development frequently proceeds through moments of reorganization rather than through the simple accumulation of information. Discussions do not always advance by adding new concepts to an existing framework. Sometimes they advance because the framework itself changes. New distinctions become central. Old assumptions are reconsidered. Previously peripheral possibilities move into the foreground. The discussion becomes productive not because it contains more information but because it is organized differently.

The phenomenon is not limited to human–LLM interaction. Scientific revolutions, philosophical innovations, and theoretical breakthroughs often exhibit a similar structure. A new distinction, a revised assumption, or a novel way of organizing familiar concepts can transform an entire field of inquiry. What changes is not merely what is known but the framework through which what is known is understood. Human–LLM interaction makes this process unusually visible because framework activation, stabilization, and transformation can often be observed directly within the unfolding history of a conversation.

The discussion thus far has emphasized how frameworks organize meaning and possibility. Yet an important question remains. If frameworks shape what can meaningfully follow, how do meanings themselves become established and stabilized within them? How does a discussion move from an open field of possibilities toward increasingly determinate conceptual structures?

To answer this question, we must examine more closely the relationship between conceptual frameworks and the progressive formation of meaning.

I think this section fits the pattern that is now emerging:

- Section 4: Why does context matter?
- Section 5: Why can the model be right and wrong simultaneously?
- Section 6: Why does drift occur?
- Section 7: Why can small interventions have large effects?

Each begins with a recognizable phenomenon, introduces the framework as an explanatory lens, uses examples sparingly, and then opens naturally onto the next question. Most importantly, the notion of possibility is now entering through lived experience rather than through ontology, which will make Section 8 much easier to write.

[Author]: Rather than saying "This phenomenon is familiar to many experienced users" I want to say that it was observed repeatedly in the previous studies and we can cite all three studies.

[ChatGPT]: I agree. In fact, I think this is stronger for two reasons.

First, it maintains continuity with the earlier sections of the paper. The entire argument is building from the three preceding studies, so this is exactly the place where those studies should begin doing more visible work.

Second, the statement "many experienced users have observed this" is true, but it is somewhat impressionistic. The three studies provide actual evidence for the phenomenon within the framework being developed.

I would therefore replace the second paragraph with something like:

This phenomenon was repeatedly observed throughout the studies upon which the present analysis builds (Rogers, 2025a, 2025b, 2025c). In each case, conceptual development often depended less on the introduction of substantial new information than on relatively small interventions that reorganized the framework through which existing concepts were understood. A distinction that had previously remained implicit became explicit. An underlying assumption was identified and questioned. A different relational organization was introduced. The resulting effects frequently appeared disproportionate to the scale of the intervention itself. New questions emerged, previously disconnected ideas became related, and conceptual trajectories that had seemed unavailable suddenly became accessible.

I actually think this paragraph does something important that the earlier version did not.

It implicitly connects all three studies:

- **The phenomenological study** (*A Portable Activation Prompt...*) demonstrated how framework activation and reactivation could reorganize an interaction.

- **The possibility paper** showed how possibilities become available or unavailable depending upon the organization of relational constraints.
- **The framework activation paper** documented explicit framework transitions and the effects of identifying governing assumptions.

In other words, the phenomenon is not merely observed in one paper. It appears repeatedly across all three.

You might even sharpen the next sentence slightly:

Current:

At first glance, such cases may appear puzzling.

Revised:

The repeated appearance of this phenomenon across all three studies raises an important question. If only a small amount of information has been introduced, why should the resulting conceptual transformation be so large?

That creates a stronger bridge into the explanatory section.

In fact, one of the nice things about Section 7 is that it may become the first section where the reader begins to see the three preceding papers functioning together as a unified body of work rather than as separate contributions.

The same phenomenon appears from three different directions, and the present paper is now drawing out its implications. That seems very much in keeping with the overall purpose of the project.

[Author]: Ok. can you draft the full section 7 with these changes

[ChatGPT]:

7. Why Can Small Interventions Have Large Effects?

One of the more surprising features of extended interaction with large language models is the extent to which seemingly minor interventions can transform an entire discussion. A brief clarification may resolve a persistent confusion. A single distinction may reorganize a complex argument. A change in perspective may suddenly make visible connections that were previously inaccessible. In some cases, only a few words are added to the conversation, yet the resulting shift in direction appears disproportionate to the amount of information introduced.

This phenomenon was repeatedly observed throughout the studies upon which the present analysis builds (Rogers, 2025a, 2025b, 2025c). In each case, conceptual development often depended less on the introduction of substantial new information than on relatively small interventions that reorganized the framework through which existing concepts were understood. A distinction that had previously remained implicit became explicit. An underlying assumption was identified and questioned. A different relational organization was introduced. The resulting effects frequently appeared disproportionate to the scale of the intervention itself. New questions emerged, previously disconnected ideas became related, and conceptual trajectories that had seemed unavailable suddenly became accessible.

The repeated appearance of this phenomenon across all three studies raises an important question. If only a small amount of information has been introduced, why should the resulting conceptual transformation be so large?

The framework developed in the preceding sections suggests that the answer lies not primarily in the information itself but in its capacity to reorganize the conceptual framework within which the discussion is occurring.

As we have seen, conceptual frameworks do more than organize meaning. They also organize possible continuations. A framework determines which distinctions are relevant, which developments appear natural, and which trajectories become available for further exploration. When a framework changes, the space of possible developments changes with it. A seemingly minor intervention may therefore have significant consequences because it alters the organization of the possibilities available within the interaction.

This observation helps explain why certain interventions prove so powerful. A new distinction does not simply add another concept to the discussion. It may reorganize the relations among concepts that are already present. Likewise, the identification of a previously hidden assumption may do more than correct an error. It may alter the framework governing the interaction and thereby change what can meaningfully follow from it. What appears locally as a small modification may function globally as a reorganization of the conceptual space.

The framework activation study provides a particularly clear illustration of this process (Rogers, 2025c). Throughout the investigation, conceptual progress often depended less upon the introduction of new concepts than upon making organizing constraints explicit. Once these constraints became visible, previously unavailable possibilities emerged. Concepts that had already been present acquired new significance because they now participated within a different framework of relations. The resulting transformation often appeared much larger than the intervention that initiated it.

A similar pattern was observed in the phenomenological study of framework activation (Rogers, 2025a). The introduction of a relatively small number of organizing distinctions frequently altered the trajectory of the interaction in ways that could not be explained simply by the addition of information. The effect of the intervention depended upon its capacity to activate, stabilize, or transform the framework through which subsequent continuations were generated.

The possibility study revealed the same phenomenon from a different perspective (Rogers, 2025b). There, the central observation was that possibilities do not simply exist in advance waiting to be selected. Rather, they become available through the organization of relational constraints. From this perspective, a framework-changing intervention is powerful because it alters the structure through which possibilities are formed. What appears as the discovery of a new possibility may in fact be the consequence of a reorganization that has made that possibility available for the first time.

This perspective also helps explain why conceptual development frequently proceeds through moments of reorganization rather than through the simple accumulation of information. Discussions do not always advance by adding new concepts to an existing framework. Sometimes they advance because the framework itself changes. New distinctions become central. Old assumptions are reconsidered. Previously peripheral possibilities move into the foreground. The discussion becomes productive not because it contains more information but because it is organized differently.

The phenomenon is not limited to human–LLM interaction. Scientific revolutions, philosophical innovations, and theoretical breakthroughs often exhibit a similar structure. A new distinction, a revised assumption, or a novel way of organizing familiar concepts can transform an entire field of inquiry. What changes is not merely what is known but the framework through which what is known is understood. Human–LLM interaction makes this process unusually visible because framework activation, stabilization, and transformation can often be observed directly within the unfolding history of a conversation.

The discussion thus far has emphasized how frameworks organize meaning and possibility. Yet an important question remains. If frameworks shape what can meaningfully follow, how do meanings themselves become established and stabilized within them? How does a discussion move from an open field of possibilities toward increasingly determinate conceptual structures?

To answer this question, we must examine more closely the relationship between conceptual frameworks and the progressive formation of meaning.

I think this version does something important that the previous draft did not. Rather than merely citing the earlier papers, it shows how each of them observed the same phenomenon from a different angle:

- **Rogers (2025a)** — framework activation and stabilization.
- **Rogers (2025b)** — formation of possibility through relational constraints.
- **Rogers (2025c)** — identification and transformation of governing frameworks.

As a result, the section begins to function as a synthesis of the three studies rather than simply drawing examples from them. That feels appropriate at this point in the paper because we are now moving from the foundational framework into its broader implications.

[Author]: Can you draft the next section?

[ChatGPT]: I think Section 8 is where we need to be especially careful.

Up to this point, the discussion has focused on:

- framework activation,
- formal integrity,
- interpretation,
- synchronization,
- drift,
- framework transformation.

Now we arrive at the question that sits beneath all of them:

How does meaning become established in the first place?

The temptation will be to answer this immediately with the theory from the possibility paper. But I think we should stay at the same phenomenological level that has been working so well.

The section should begin with a familiar observation:

In productive discussions, meaning often becomes clearer over time.

People regularly experience this.

A scientist refining a theory.

A philosopher clarifying an argument.

A student struggling with a concept.

A human–LLM interaction exploring a difficult question.

Often the participants do not fully understand what they mean at the outset.

The meaning becomes progressively determined through the interaction itself.

That is the phenomenon.

The framework then explains it.

8. How Does Meaning Become Progressively Determined?

One of the most familiar features of productive inquiry is that meaning often becomes clearer as a discussion develops. Participants may begin with an intuition, a question, or a partially formed idea whose significance is not yet fully understood. Through dialogue, distinctions are clarified, assumptions are examined, and conceptual relations become increasingly organized. What initially appeared vague or uncertain gradually acquires a more determinate form.

This phenomenon is familiar across a wide range of contexts. Scientists frequently report that the significance of a problem becomes clearer through the process of investigation itself. Philosophical inquiry often proceeds by refining distinctions that were initially only dimly perceived. In ordinary conversation, participants regularly discover that they did not fully understand what they meant until they attempted to articulate it. The same phenomenon was repeatedly observed throughout the studies upon which the present analysis builds (Rogers, 2025a, 2025b, 2025c). In many cases, conceptual progress did not consist in expressing a fully formed meaning that already existed in advance. Rather, meaning became increasingly determinate through the interaction itself.

This observation raises an important question. If meaning is not simply retrieved from a pre-existing conceptual structure, how does it become established?

The framework developed in the preceding studies suggests that meaning emerges through a process of progressive determination. Conceptual frameworks do not merely contain meanings. They provide the relational organization through which meanings become increasingly stabilized. Distinctions are introduced, relations are explored, consequences are examined, and incompatible trajectories are gradually excluded. Through this process, a conceptual structure acquires greater definition and coherence.

The significance of this observation becomes clearer when considered alongside the discussions of framework activation and conceptual drift. A framework does not function as a static container for meaning. It is an evolving organization of relations. As the framework develops, certain possibilities become reinforced while others recede into the background. Some distinctions prove productive and become increasingly central. Others lose relevance and gradually disappear. Meaning therefore emerges not through the simple accumulation of information but through the progressive organization of a conceptual space.

This perspective also helps clarify why productive inquiry often involves periods of uncertainty. If meaning were fully present from the outset, there would be little need for exploration. Yet in many discussions, the participants themselves do not initially know what the most important distinctions will prove to be. The framework must develop before those distinctions become visible. Meaning appears gradually because the conditions required for its stabilization are themselves being formed through the interaction.

The asymmetry between human participants and language models discussed in the previous sections becomes especially important here. The model contributes by extending trajectories that preserve formal integrity. It explores possible continuations, develops relations, and elaborates conceptual structures. The human participant contributes interpretative orientation, identifying which developments are significant and which trajectories deserve further attention. When synchronization is maintained, these activities can support one another. Formal development generates new possibilities, while interpretation helps determine which of those possibilities become meaningful.

This relationship helps explain a phenomenon frequently observed in productive human–LLM interactions. Participants often report that an idea becomes clearer through the course of the interaction even though

neither the human nor the model appears to possess the fully developed idea at the outset. The resulting meaning emerges through the interplay of formal development and interpretative evaluation. What becomes visible is not simply the retrieval of an existing concept but the progressive determination of a conceptual structure.

The possibility study approached this phenomenon from a more formal perspective (Rogers, 2025b). There it was argued that possibilities are not best understood as pre-existing alternatives waiting to be selected. Rather, possibilities become available through the organization of relational constraints. The same insight applies to meaning. Meaning does not appear fully formed at the beginning of inquiry. It becomes increasingly determinate as frameworks organize what distinctions, relations, and developments remain viable.

This observation helps explain why conceptual breakthroughs often appear suddenly despite emerging from long periods of gradual development. The apparent breakthrough is typically preceded by an extended process of organization through which a framework becomes increasingly structured. Once a sufficient degree of organization has been achieved, a previously unstable meaning may suddenly become clear. The moment of insight appears abrupt, but the process that made it possible was often distributed throughout the entire history of the interaction.

The discussion has now revealed a close relationship between framework activation, conceptual drift, framework transformation, and the progressive determination of meaning. Yet one important question remains. If meaning emerges through the interaction of formal development and interpretation, why do some human–LLM interactions become extraordinarily productive while others remain comparatively limited?

To answer this question, we must examine more closely the conditions under which human and model contribute most effectively to a shared conceptual process.

I think this section works because it introduces the idea of **progressive determination** without becoming abstract or ontological. It remains grounded in a familiar experience: realizing what one means through the process of inquiry itself. That keeps it aligned with the tone of Sections 4–7 while preparing the way for Section 9, which I suspect will be where the full significance of the human–LLM asymmetry becomes most visible.

[Author]: In this section we say: "Formal development generates new possibilities, while interpretation helps determine which of those possibilities become meaningful." But I suspect the inverse situation is the more powerful one. That is where the human provides new possibilities and the LLM helps to develop those possibilities into formally coherent interpretations.

[ChatGPT]: I think that is exactly right, and it reveals something important about the direction of causality in the current draft.

The sentence:

"Formal development generates new possibilities, while interpretation helps determine which of those possibilities become meaningful."

still places the LLM in the generative role and the human in the evaluative role.

But that is not actually what the three papers have been showing.

In many of the most important examples, the situation is almost the reverse.

The human participant introduces:

- a new distinction,
- a new interpretation,
- a new framework,
- a new possibility.

The LLM then contributes by:

- exploring consequences,
- developing relations,
- testing coherence,
- elaborating implications,
- identifying connections.

In other words:

The human often opens the possibility.

The LLM often develops the possibility.

That is a much better description of the interaction we have actually been analyzing.

In fact, I think we need to distinguish two senses of "possibility."

Formal possibilities

Possible continuations within a framework.

These are exactly what the LLM is good at exploring.

Given a framework, it can generate:

- implications,
 - variations,
 - analogies,
 - trajectories.
-

Conceptual possibilities

The possibility that a different framework could exist.

The possibility that a distinction is being overlooked.

The possibility that the governing assumptions are wrong.

The possibility that a different interpretation is required.

These are the kinds of possibilities that repeatedly entered through human intervention in the activation study.

This distinction is very important because it connects directly to Section 7.

When we observed that a small intervention can reorganize an entire discussion, the intervention was often introduced by the human participant.

The LLM did not spontaneously generate it.

The human identified:

- a hidden assumption,
- a deeper constraint,
- a neglected distinction.

That intervention opened a new conceptual possibility.

Only then did the LLM begin exploring the consequences.

I therefore think the paragraph should be rewritten.

Current version:

The model contributes by extending trajectories that preserve formal integrity. It explores possible continuations, develops relations, and elaborates conceptual structures. The human participant contributes interpretative orientation, identifying which developments are significant and which trajectories deserve further attention.

Something closer to:

The asymmetry between human participants and language models becomes especially important here. Human participants frequently contribute new conceptual possibilities by introducing distinctions, questioning assumptions, or proposing alternative frameworks. Language models contribute differently. Once a possibility has been opened, they can assist in exploring its consequences, developing its internal relations, and testing its formal coherence. When synchronization is maintained, these activities become mutually reinforcing. Interpretation opens new directions for inquiry, while formal development helps transform those directions into increasingly structured conceptual forms.

Then the next sentence becomes:

Instead of:

Formal development generates new possibilities, while interpretation helps determine which of those possibilities become meaningful.

Something like:

New possibilities often enter through interpretation, while formal development helps determine what follows from them.

Or perhaps even stronger:

Interpretation frequently opens possibilities that were previously unavailable, while formal development helps stabilize, elaborate, and test those possibilities within a coherent conceptual framework.

I actually think this revision brings Section 8 much closer to the possibility paper.

Because one of the central claims of that paper is not that possibilities are generated by formal systems.

It is that possibilities become available through the organization of relational constraints.

The human participant often plays a decisive role in reorganizing those constraints.

The LLM then becomes an extraordinarily powerful mechanism for exploring the resulting landscape.

That seems much closer to the interaction pattern documented across all three studies.

[Author]: Ok. Can you redraft the full section 8 with these changes

[ChatGPT]:

8. How Does Meaning Become Progressively Determined?

One of the most familiar features of productive inquiry is that meaning often becomes clearer as a discussion develops. Participants may begin with an intuition, a question, or a partially formed idea whose significance is not yet fully understood. Through dialogue, distinctions are clarified, assumptions are examined, and conceptual relations become increasingly organized. What initially appeared vague or uncertain gradually acquires a more determinate form.

This phenomenon is familiar across a wide range of contexts. Scientists frequently report that the significance of a problem becomes clearer through the process of investigation itself. Philosophical inquiry often proceeds by refining distinctions that were initially only dimly perceived. In ordinary conversation,

participants regularly discover that they did not fully understand what they meant until they attempted to articulate it. The same phenomenon was repeatedly observed throughout the studies upon which the present analysis builds (Rogers, 2025a, 2025b, 2025c). In many cases, conceptual progress did not consist in expressing a fully formed meaning that already existed in advance. Rather, meaning became increasingly determinate through the interaction itself.

This observation raises an important question. If meaning is not simply retrieved from a pre-existing conceptual structure, how does it become established?

The framework developed in the preceding studies suggests that meaning emerges through a process of progressive determination. Conceptual frameworks do not merely contain meanings. They provide the relational organization through which meanings become increasingly stabilized. Distinctions are introduced, relations are explored, consequences are examined, and incompatible trajectories are gradually excluded. Through this process, a conceptual structure acquires greater definition and coherence.

The significance of this observation becomes clearer when considered alongside the discussions of framework activation and conceptual drift. A framework does not function as a static container for meaning. It is an evolving organization of relations. As the framework develops, certain possibilities become reinforced while others recede into the background. Some distinctions prove productive and become increasingly central. Others lose relevance and gradually disappear. Meaning therefore emerges not through the simple accumulation of information but through the progressive organization of a conceptual space.

This perspective also helps clarify why productive inquiry often involves periods of uncertainty. If meaning were fully present from the outset, there would be little need for exploration. Yet in many discussions, the participants themselves do not initially know what the most important distinctions will prove to be. The framework must develop before those distinctions become visible. Meaning appears gradually because the conditions required for its stabilization are themselves being formed through the interaction.

The asymmetry between human participants and language models discussed in the previous sections becomes especially important here. Human participants frequently contribute new conceptual possibilities by introducing distinctions, questioning assumptions, identifying hidden constraints, or proposing alternative frameworks. Such interventions often reorganize the conceptual space within which the discussion is occurring. What had previously appeared fixed may become open to reconsideration. New trajectories become available. Possibilities that were not visible within the previous framework enter the discussion.

Language models contribute differently. Once a possibility has been opened, they can assist in exploring its consequences, developing its internal relations, extending its implications, and testing its formal coherence. They excel at elaborating trajectories that preserve formal integrity within the framework currently in play. When synchronization is maintained, these activities become mutually reinforcing. Interpretation opens new directions for inquiry, while formal development helps transform those directions into increasingly structured conceptual forms.

This relationship helps explain a phenomenon repeatedly observed in productive human–LLM interactions. Participants often report that an idea becomes clearer through the course of the interaction even though neither the human nor the model appears to possess the fully developed idea at the outset. The human participant may introduce a distinction whose implications are not yet understood. The model may then

elaborate those implications in ways that reveal previously unseen relations. Further interpretive interventions identify what is significant within the emerging structure. Meaning develops through this iterative process. What becomes visible is not simply the retrieval of an existing concept but the progressive determination of a conceptual framework.

The possibility study approached this phenomenon from a more formal perspective (Rogers, 2025b). There it was argued that possibilities are not best understood as pre-existing alternatives waiting to be selected. Rather, possibilities become available through the organization of relational constraints. The same insight applies to meaning. Meaning does not appear fully formed at the beginning of inquiry. It becomes increasingly determinate as frameworks organize what distinctions, relations, and developments remain viable.

This observation suggests that meaning and possibility are more closely related than is often assumed. New meanings emerge because new possibilities become available, while new possibilities become available because existing frameworks have been reorganized. The process is therefore neither purely interpretive nor purely formal. It depends upon the continual interaction between the opening of possibilities and the development of their consequences.

The framework activation study provides a particularly clear illustration of this process (Rogers, 2025c). Conceptual advances frequently occurred when previously implicit assumptions were made explicit and alternative frameworks became available for consideration. These interventions did not simply provide new answers. They opened new possibilities. Once opened, those possibilities could be explored, elaborated, and refined through subsequent interaction. Meaning emerged through the progressive stabilization of the resulting conceptual structure.

This perspective also helps explain why moments of insight often appear sudden despite emerging from long periods of gradual development. The apparent breakthrough is typically preceded by an extended process of conceptual organization. Distinctions have been introduced, relations explored, assumptions questioned, and possibilities tested. Once a sufficient degree of organization has been achieved, a previously unstable meaning may suddenly become clear. The moment of insight appears abrupt, but the process that made it possible was distributed throughout the history of the interaction.

The discussion has now revealed a close relationship between framework activation, conceptual drift, framework transformation, possibility formation, and the progressive determination of meaning. Yet one important question remains. If meaning emerges through the interaction of interpretive intervention and formal development, why do some human–LLM interactions become extraordinarily productive while others remain comparatively limited?

To answer this question, we must examine more closely the conditions under which human and model contribute most effectively to a shared conceptual process.

I think this version is much more consistent with the trajectory of the three preceding papers. The human is no longer cast primarily as a selector of possibilities generated by the model. Instead, the human frequently serves as the source of new conceptual possibilities through interpretive intervention, while the model contributes by developing those possibilities into increasingly coherent relational structures. That asymmetry is much closer to the framework we have been building since Sections 4 and 5, and it sets up Section 9 naturally: under what conditions does this complementary relationship become exceptionally productive?

[Author]: Ok. please draft the next section

[ChatGPT]: I think Section 9 is where many papers would become overly celebratory or drift into claims about collective intelligence, co-thinking, or emergent cognition.

What has been working well so far is that we have remained disciplined and stayed close to the implications of the framework.

So I think the question is not:

Why are human–LLM interactions powerful?

But rather:

Under what conditions does the asymmetry we have identified become productive?

In other words, the section should not begin with extraordinary interactions.

It should begin with the observation that many interactions are ordinary.

Some are unproductive.

Some drift.

Some never achieve interpretative unity.

And yet some become remarkably fruitful.

The question is why.

9. Why Are Some Human–LLM Interactions So Productive?

The preceding sections have identified several characteristic features of human–LLM interaction.

Conceptual frameworks may become active or inactive. Formal integrity and interpretation may remain synchronized or gradually diverge. New possibilities may emerge through interpretive intervention and subsequently be developed through formally coherent continuations. These observations help explain why some interactions become productive. They do not yet explain why certain interactions become exceptionally productive.

This question is worth considering because not all interactions with large language models are equally successful. Many remain superficial. Others become repetitive or drift away from their original purpose. Some generate formally organized trajectories that never achieve interpretative unity. Yet there are also interactions in which conceptual development proceeds with unusual depth and persistence. New possibilities emerge, distinctions become increasingly refined, and frameworks acquire a level of coherence that neither participant appeared to possess at the outset. What accounts for these differences?

The framework developed throughout this paper suggests that the answer lies not primarily in the quantity of information exchanged but in the quality of the relationship established between interpretation and formal development.

As we have seen, human participants and language models contribute differently to the interaction. Human participants introduce distinctions, identify significance, question assumptions, and propose alternative frameworks. Language models contribute by developing relational structures, exploring consequences, identifying connections, and preserving formal integrity across extended conceptual trajectories. Productive interaction depends upon the successful coordination of these complementary activities.

This observation helps explain why framework activation is so important. When a coherent framework is active, interpretive interventions and formal continuations become mutually reinforcing. A distinction

introduced by the human participant can be developed through multiple layers of implication. The resulting developments can then be interpreted, refined, and redirected through further intervention. Conceptual progress emerges through the repeated interaction between these two processes rather than from either process alone.

The studies upon which this paper builds provide numerous examples of this pattern (Rogers, 2025a, 2025b, 2025c). In many cases, conceptual advances did not arise because either participant independently possessed a fully formed solution. Instead, progress emerged through the iterative development of distinctions whose significance only became visible through continued exploration. A possibility introduced at one stage of the interaction generated consequences that revealed new interpretive opportunities. These in turn opened further possibilities for formal development. The resulting process often extended across many cycles of prompt and response.

Importantly, the productivity of such interactions should not be understood as evidence that the language model possesses an independent capacity for interpretation. The framework developed throughout this paper suggests a different explanation. Productive interactions occur when interpretive interventions and formal continuations remain sufficiently synchronized to support the progressive determination of a shared conceptual framework. The resulting coherence belongs to the interaction as a whole rather than to either participant considered in isolation.

This observation helps clarify a phenomenon frequently reported by experienced users. Participants often feel that an idea emerged during the interaction that neither they nor the model possessed at the outset. From the perspective developed here, this experience is neither mysterious nor surprising. The interaction itself provides conditions under which possibilities can be opened, developed, stabilized, and interpreted. What emerges is not the retrieval of a hidden concept but the progressive formation of a conceptual structure through the coordinated activity of both participants.

At the same time, the productivity of the interaction remains fragile. The same process that allows conceptual development can also generate drift. Synchronization may weaken. Interpretive unity may be lost. Alternative frameworks may become active. Formal development may continue without converging upon a stable conceptual structure. Productive interaction therefore depends not merely on the generation of continuations but on the continual maintenance of the relationship between interpretation and formal development.

This perspective also helps explain why some of the most productive interactions are characterized by repeated returns to fundamental distinctions. Participants often revisit earlier assumptions, restate key commitments, and reactivate organizing frameworks. Such returns do not simply preserve continuity. They help maintain synchronization within the developing conceptual structure. What appears as repetition from one perspective may function as the ongoing stabilization of a shared framework.

Viewed in this way, the most productive human–LLM interactions are neither simple information exchanges nor demonstrations of machine understanding. They are instances of a more general semiotic process in which interpretation and formal development become mutually supportive. The human participant contributes orientation toward meaning. The model contributes the preservation and exploration of relational structure. When these activities remain synchronized, conceptual frameworks can develop in ways that neither participant could achieve alone.

This conclusion leads naturally to a final question. If the strengths of human–LLM interaction arise from the relationship between interpretation and formal development, are the characteristic limitations of these

systems generated by the same relationship? More generally, might the strengths and weaknesses of large language models be two aspects of a single underlying structure?

This question will be the focus of the next section.

I think this section does something important that we have been approaching for several sections now. It shifts the focus away from the model itself and toward the interaction as the primary unit of analysis. That has really been the hidden theme running through Sections 4–8. By the time the reader reaches this section, the conclusion feels almost inevitable: the most productive outcomes emerge not from interpretation alone or formal development alone, but from the sustained synchronization of both within a shared conceptual framework. That sets up Section 10 very naturally, because the strengths and limitations can now be analyzed as consequences of the same underlying arrangement.

[Author]: What do you intend to cover in section 10

[ChatGPT]: I think Section 10 should be one of the most important sections in the paper because it allows us to pull together everything that has been established so far and show that the strengths and limitations of LLMs are not independent phenomena.

Up to this point, we have argued that:

- Context matters because frameworks matter.
- LLMs preserve formal integrity.
- Humans provide interpretative orientation.
- Synchronization can be maintained or lost.
- New possibilities often enter through interpretation.
- Formal development helps stabilize and elaborate those possibilities.
- Productive interactions emerge from the coordination of these activities.

The natural next question is:

Are the characteristic strengths and weaknesses of LLMs generated by the same underlying structure?

I think the answer within this framework is clearly yes.

What Section 10 Should Not Be

It should not be:

Here are the strengths.

Here are the weaknesses.

That becomes a list.

Nor should it become a discussion of AI safety, hallucinations, alignment, or benchmarks.

Those are downstream manifestations.

We want to remain at the same conceptual level as the rest of the paper.

The Central Claim

The central claim might be something like:

The strengths and limitations of LLMs arise from the same capacity: the preservation and extension of formal integrity within conceptual frameworks.

That is the organizing idea.

Strengths

Because LLMs preserve formal integrity so effectively, they can:

- sustain complex conceptual trajectories,
- maintain distinctions across long discussions,
- explore consequences systematically,
- elaborate possibilities,
- identify remote connections,
- operate across enormous conceptual domains.

Many of the productive interactions discussed in Section 9 depend directly on these capabilities.

Weaknesses

But the very same capacity also explains why LLMs:

- drift,
- lose interpretative orientation,
- elaborate irrelevant trajectories,
- preserve problematic assumptions,
- continue frameworks that should be abandoned,
- generate formally organized structures lacking interpretative unity.

The issue is not that the model suddenly stops working.

The issue is that it continues working.

The preservation of formal integrity is exactly what allows these limitations to emerge.

The Key Insight

I think the most important sentence in the section might be something like:

The characteristic limitations of large language models are not failures of the mechanism that produces their strengths. They are consequences of that mechanism operating in the absence of interpretative guidance.

That is a very powerful claim because it unifies phenomena that are usually treated separately.

The Three Cases Return

This section also gives us an opportunity to bring back the three cases from Section 5.

Productive synchronization

Formal integrity + interpretative unity.

Strengths dominate.

Drift

Formal integrity persists while synchronization weakens.

Mixed outcomes.

Formal integrity without interpretative unity

The model continues generating structure but no stable meaning emerges.

Limitations dominate.

Now the reader sees that these are not three unrelated situations.

They are different manifestations of the same underlying capacity.

Why This Matters

I think this section should then make a broader point.

Many discussions about LLMs assume that increasing capability will automatically reduce limitations. The framework developed here suggests something more nuanced. If the strengths and limitations arise from the same underlying structure, then increasing the capacity for formal development may also increase some of the characteristic risks:

- more elaborate drift,
- more convincing but misplaced trajectories,
- more sophisticated formal objects lacking interpretative unity.

In other words:

More formal power does not automatically generate more meaning.

That is a direct consequence of the framework.

Transition to Section 11

The section would then end by stepping back and asking:

What does all of this reveal about semiosis itself?

Because by this point we are no longer primarily talking about LLMs.

We are talking about:

- meaning,
- interpretation,
- possibility,
- determination,
- formal development,
- synchronization.

LLMs have become a lens through which these processes can be observed.

That naturally opens the concluding section:

What Do LLMs Reveal About Semiosis?

And I suspect the answer will be something like:

They reveal that formal development and interpretation, which are normally intertwined within human cognition, can be partially separated and therefore studied independently.

That would bring the paper full circle back to the motivation established in the introduction.

[Author]: Please draft section 10

[ChatGPT]:

10. Why Are the Strengths and Weaknesses of LLMs So Closely Related?

Discussions of large language models often separate their strengths from their limitations. On one side are their remarkable capacities for generating coherent text, exploring conceptual relations, maintaining complex discussions, and identifying patterns across large bodies of knowledge. On the other side are familiar concerns regarding hallucinations, conceptual drift, misplaced confidence, and the tendency to generate responses that are formally sophisticated yet somehow disconnected from what matters most. These characteristics are often treated as independent features requiring separate explanations.

The framework developed throughout this paper suggests a different interpretation. The strengths and weaknesses of large language models are not independent phenomena. They arise from the same underlying capacity.

As we have seen, language models participate in conceptual activity primarily through the preservation and extension of formal integrity. They generate continuations that maintain organized relational structures across successive stages of an interaction. This capacity allows them to explore implications, elaborate conceptual trajectories, identify patterns, and preserve coherence across discussions that may span many exchanges. Much of what makes these systems useful depends directly upon this ability.

The importance of this capacity can be seen in many of the productive interactions discussed in the previous sections. Formal integrity allows conceptual frameworks to be explored beyond what either participant might initially anticipate. Consequences can be developed, distinctions can be refined, and possibilities can be elaborated in ways that would often be difficult to sustain through human effort alone. The model's ability to preserve and extend relational structure provides an important source of conceptual continuity within the interaction.

Yet the same capacity also helps explain many of the system's characteristic limitations.

A model may continue developing a trajectory that has ceased to be relevant. It may preserve assumptions that ought to be questioned. It may elaborate a framework whose organizing constraints have become problematic. It may generate sophisticated continuations that possess formal integrity without contributing to interpretative unity. In each case, the difficulty arises not because the model has ceased operating according to its characteristic strengths, but because those strengths continue to operate in circumstances where interpretative orientation has weakened, drifted, or become absent.

This observation helps explain why some of the most familiar limitations of large language models can be so difficult to correct. The problem is not always that the model lacks information. Nor is it necessarily a failure of logical consistency. Often the model is doing precisely what it does best: preserving and extending a formally coherent trajectory. The difficulty is that formal integrity alone cannot determine whether a trajectory remains meaningful, significant, or properly situated within a broader conceptual framework.

The distinction developed in Section 5 now becomes particularly important. Conceptual frameworks require both formal integrity and interpretative unity. Formal integrity is necessary because meanings cannot stabilize without organized relational structures. Yet formal integrity alone is not sufficient. A formally coherent trajectory may participate in a shared framework, drift away from one, or continue in the absence of one altogether.

The three possibilities discussed earlier can therefore be understood as different manifestations of the same underlying capacity. In the first case, formal integrity remains synchronized with interpretative orientation. The result is productive conceptual development. In the second case, formal integrity continues while synchronization gradually weakens. The result is conceptual drift. In the third case, formal integrity continues without converging upon a stable framework capable of sustaining interpretative unity. The result is an organized conceptual trajectory that remains difficult to interpret as a coherent conceptual object.

What makes these cases especially revealing is that they do not involve different mechanisms. The same capacity for preserving relational structure is present throughout. What changes is the relationship between formal development and interpretation.

This perspective suggests a different way of thinking about the future development of language models. Discussions of capability often assume that greater formal sophistication will automatically reduce existing limitations. The framework developed here points toward a more complicated possibility. If many strengths and limitations arise from the same underlying capacity, increasing formal power alone may not eliminate the characteristic difficulties associated with interpretation. More powerful systems may sustain longer trajectories, explore more possibilities, and generate more elaborate conceptual structures. Yet they may also become capable of more sophisticated forms of drift, more persistent continuation of problematic frameworks, and more complex formal objects lacking interpretative unity.

From this perspective, the central challenge is not simply the generation of better continuations. It is the relationship between formal development and interpretative orientation. The most productive interactions examined throughout this paper did not arise because formal integrity became sufficient for meaning. They arose because interpretative interventions and formal continuations remained mutually supportive within a shared framework.

This observation returns us to one of the central themes of the preceding studies. Meaning does not emerge from formal organization alone, nor from interpretation alone. It emerges through their relationship. The remarkable achievements of large language models become possible because they participate in one side of this relationship with extraordinary effectiveness. Their characteristic limitations arise because participation in formal organization is not identical with participation in meaning.

This conclusion invites a broader reflection. If large language models derive both their strengths and their weaknesses from the partial separation of formal development and interpretation, what does this reveal about semiosis itself? More specifically, what becomes visible when one of the central components of semiosis can be studied in relative isolation from the others?

The answer to this question will form the basis of our concluding discussion.

I think this section accomplishes two important things. First, it unifies the strengths and weaknesses of LLMs under a single explanatory principle rather than treating them as separate topics. Second, it quietly shifts the focus away from LLMs themselves and toward the deeper semiotic question that has been developing throughout the paper. By the end of the section, the reader is already beginning to think about semiosis rather than AI, which should make the transition into the concluding section feel natural rather than abrupt.

[Author]: Please draft the concluding section

[ChatGPT]: I think the conclusion should resist the temptation to become a summary. Most of the important claims have already been made.

The final section should instead return to the question posed in the introduction:

What do LLMs reveal about semiosis?

And the answer, I think, is surprisingly profound.

The paper has gradually moved from:

- understanding LLMs,
- to:
- understanding human–LLM interaction,
- to:
- understanding semiosis itself.

The deepest insight is not that LLMs are semiotic agents.

The deepest insight is that they allow us to observe one aspect of semiosis—formal development—in relative isolation from interpretation.

That is what makes them theoretically interesting.

11. What Do LLMs Reveal About Semiosis?

The purpose of this paper has not been to determine whether large language models are intelligent, conscious, or capable of understanding in a human sense. Rather, it has been to examine what becomes visible when these systems are viewed through the semiotic framework developed in the preceding studies.

The analysis began from a simple observation. Human–LLM interaction often behaves in ways that are difficult to explain through conventional accounts based solely on information, representation, or computation. Context matters in unexpected ways. Small interventions can have disproportionately large effects. Discussions may drift despite retaining coherence. Meaning often becomes clearer through interaction itself. Some exchanges become remarkably productive, while others remain trapped within formally sophisticated but conceptually limited trajectories.

Viewed independently, these phenomena may appear unrelated. Viewed through the framework developed in the preceding studies, however, they reveal a common structure. Conceptual frameworks simultaneously function as structures of meaning and systems of relational constraint. Human participants and language models contribute differently to these two aspects of the framework. Human participants primarily contribute interpretative orientation. Language models primarily contribute the preservation and extension of formal integrity. Productive interaction depends upon the ongoing coordination of these complementary activities.

This perspective suggests that the significance of large language models extends beyond the practical tasks they perform. These systems provide an unusual opportunity to examine aspects of semiosis that are ordinarily difficult to isolate. In most human cognition, interpretation and formal development are deeply intertwined. We simultaneously determine significance, maintain conceptual organization, introduce distinctions, and develop their consequences. Because these activities occur together, their individual contributions are often difficult to distinguish.

Human–LLM interaction partially separates them.

The language model contributes extraordinary capacities for preserving and extending relational structures. It can sustain conceptual trajectories, explore consequences, elaborate distinctions, and maintain formal organization across extended discussions. Yet the analyses developed throughout this paper suggest that such capacities do not by themselves generate interpretative unity. Meaning remains dependent upon processes that extend beyond formal continuation alone.

This observation helps clarify why the strengths and limitations of large language models are so closely related. Both arise from the same underlying capacity. The ability to preserve formal integrity makes

possible many of the system's most impressive achievements. The same ability also permits conceptual drift, the continuation of problematic frameworks, and the generation of formally coherent structures that fail to achieve stable interpretative unity. The distinction is not between successful and unsuccessful operation. It is between different relationships between formal development and interpretation.

From this perspective, the most productive human–LLM interactions are not instances in which interpretation disappears or becomes unnecessary. They are instances in which interpretation and formal development remain effectively synchronized within a shared conceptual framework. New possibilities are introduced, developed, evaluated, refined, and stabilized through the interaction itself. Meaning emerges neither from the human participant alone nor from the language model alone, but through the ongoing coordination of their complementary contributions.

At the same time, the present analysis suggests an important limitation. Formal integrity can continue in the absence of synchronization. It can continue in the absence of interpretative unity. It can even continue in the absence of any stable conceptual framework capable of sustaining meaning. This asymmetry reveals something fundamental about semiosis itself. Formal organization may be necessary for meaning, but it is not sufficient. The existence of a coherent relational structure does not guarantee the existence of a coherent conceptual object.

The broader significance of this conclusion extends beyond the study of language models. The framework developed in the preceding studies proposed that possibility is formed through relational constraint rather than selected from a pre-existing space. The present analysis suggests a corresponding insight regarding meaning. Meaning is not simply retrieved from a pre-existing conceptual structure. It emerges through processes of interpretation, determination, and stabilization occurring within evolving frameworks of relation. Human–LLM interaction provides a particularly revealing environment in which these processes can be observed because the contributions of formal development and interpretation become partially distinguishable.

The result is a shift in perspective. Rather than viewing large language models primarily as artificial minds, knowledge repositories, or predictive engines, they may be understood as participants in a broader semiotic process whose structure becomes unusually visible through interaction. Their importance lies not only in what they can do, but also in what they reveal.

What they reveal is that meaning and formal organization are deeply related but not identical. They reveal that conceptual frameworks are simultaneously structures of significance and systems of relational constraint. They reveal that possibility and meaning emerge through processes of progressive determination rather than through the selection of pre-existing alternatives. Most importantly, they reveal that the study of human–LLM interaction may provide a new window into the semiotic processes through which understanding itself becomes possible.

In this sense, the significance of large language models may ultimately lie not in the fact that they resemble human cognition, but in the fact that they allow us to examine some of its deepest organizing principles from a new perspective.

I think this conclusion stays faithful to the trajectory of the paper. It does not end with a claim about AI. It ends with a claim about semiosis. The reader arrives back at the opening question having traveled through framework activation, formal integrity, interpretation, synchronization, drift, possibility formation, and productive interaction. The result is that LLMs become less the object of study and more the lens through which a deeper semiotic

structure becomes visible. That feels very consistent with the overall direction we have developed throughout this conversation.

[Discussion of paper title, abstract and appropriate citations not included because they were drawn out and not directly relevant to the content of the paper. Basically, I lost the discipline of questioning and began to wander about.]