# A Semiotic Account of the Hierarchical Relational Organization of Large Language Models (LLMs): Implications for AI Development, Use and Evaluation

## *A Dialogical Educational Module for AI Users and Developers*

Timothy M. Rogers, Chat GPT assisted,
University of Toronto,
Aug 4, 2026

This publication develops a semiotic and relational framework for understanding how large language models (LLMs) form and preserve organized continuations across training, inference, prompting, retrieval, evaluation and agentic action. Its specific concern is not the general theory of signs but how linguistic, conceptual, inferential, evidential and task relations become operative as hierarchical constraints on model continuation. Its central distinction is between information that is merely available to a model and information, frameworks, sources, or task constraints that actually govern continuation. This distinction makes it possible to understand hallucination, sycophancy, framework substitution, retrieval failure, conceptual memory loss, and agent drift as related failures of constraint preservation rather than as isolated defects. The publication consequently shifts attention from increasing generative capacity alone toward constructing broader systems in which formal continuation remains answerable to evidence, task identity, conceptual framework, external tests, and human purposes.

The framework begins from the premise that the tokens processed by LLMs function as relationally ordered signs and that the texts on which language models are trained are also composed of signs whose uses have been shaped by human conceptual, inferential, referential and practical relations. It examines is how formal traces of those relations become available to an LLM and how they govern—or fail to govern—subsequent continuation. The module therefore investigates the formal organization of sign-mediated continuation while maintaining a categorical distinction between that organization and human interpretation.

Developed through a structured interaction with ChatGPT, the module is intended to offer AI users and developers a complementary conceptual and diagnostic lens through which to consider large language model behaviour, design, use and evaluation. It presents a theoretical framework developed from a semiotic perspective and elaborated in the author's related work (doi.org/10.5281/zenodo.21281882).

This is an experimental project. Feedback is welcomed!
Timothy.rogers@mail.utoronto.ca

# Table of Contents

## Is Next-Token Prediction a Complete Explanation?

**Meno:** Before we begin, let us agree on the basic technical description.

A large language model receives a sequence of tokens. The transformer processes that sequence and produces logits over the vocabulary. These are converted into a conditional probability distribution, from which the next token is selected or sampled. The token is appended to the context, and the process repeats.

During ordinary inference, the weights remain fixed.

**Sam:** Do you regard that as an accurate account?

**Meno:** Yes.

**Sam:** And a complete one?

**Meno:** I am inclined to say so. It tells us what the model receives, what computation it performs, what it produces, and how generation continues. What remains to be explained?

**Sam:** It clearly explains how each next token is produced. Does it also explain how an extended response is organized?

**Meno:** Why separate those questions? If the mechanism produces every token, then it produces the whole response.

**Sam:** Produces it, certainly. But consider a response that introduces a technical distinction, applies it across several examples, answers an objection, and finally draws a conclusion that depends upon the original distinction. How would you explain the relation between the opening distinction and the conclusion?

**Meno:** Through long-range dependencies. The later tokens are conditioned by material that appeared much earlier in the context.

**Sam:** What does “long-range” tell us?

**Meno:** That the dependency extends across a substantial part of the sequence.

**Sam:** Does it tell us why the earlier statement functions as a premise and the later statement as a conclusion?

**Meno:** No. It describes the distance across which the dependency operates, but not the kind of relation involved.

**Sam:** What kind of relation is it?

**Meno:** An inferential one. The conclusion follows from the premise because of their roles in the argument, not simply because the corresponding tokens are far apart.

**Sam:** So two passages could be equally distant from one another while standing in very different relations?

**Meno:** Of course. One might support a claim, another qualify it, another contradict it, and another draw a consequence from it.

**Sam:** Then is “long-range dependency” a sufficient explanation of the organization of the response?

**Meno:** Not by itself. It tells us that earlier material affects later generation, but it leaves the formal character of that influence unspecified.

**Sam:** When the model generates a sentence near the end of the argument, what might constrain the next token?

**Meno:** The immediate grammar, the terminology already established, the inferential direction of the argument, the distinction introduced earlier, and the task the response is meant to fulfil.

**Sam:** Are these all relations of the same kind?

**Meno:** No. Some are local and linguistic. Others concern the role of the sentence within the argument as a whole.

**Sam:** Yet their effects appear together in a single next-token distribution.

**Meno:** Yes.

**Sam:** Does the distribution identify those different relations?

**Meno:** Not directly. It expresses their combined effect on the available continuations.

**Sam:** Then what does the distribution explain?

**Meno:** It tells us how possible next tokens are weighted under the present context.

**Sam:** And what does it not yet explain?

**Meno:** Why those possibilities are weighted as they are in terms of the developing organization of the response.

A token may be favoured because it completes a grammatical construction, preserves a distinction, advances an inference, or remains faithful to the task. The probability distribution expresses the result of these influences, but it does not itself distinguish their formal roles.

**Sam:** Then what does “next-token prediction” specify?

**Meno:** The local output of the computation.

**Sam:** Does it also specify the scale of the relations determining that output?

**Meno:** No. The output is local, but the relations influencing it may extend across the entire available context.

**Sam:** Why, then, did the next-token account initially appear complete?

**Meno:** Because I treated the unit of production as though it must also be the sufficient unit of explanation.

Since the model generates one token at a time, I assumed the response could be adequately understood as a succession of locally conditioned token choices.

**Sam:** Does the response merely become longer as those choices accumulate?

**Meno:** No. It becomes organized.

A term acquires a specific use. A distinction becomes governing. An ambiguity may be resolved. A claim becomes a premise for something that follows. Earlier continuations change what later continuations can appropriately do.

**Sam:** Is that organization represented in the model as a completed response before generation begins?

**Meno:** We have no reason to think so.

**Sam:** Then how had you been conceiving the alternatives?

**Meno:** I suppose I had assumed that either the whole response was somehow planned in advance or it was merely the accumulated result of local choices.

**Sam:** Why only those two possibilities?

**Meno:** Because I was assuming that formal organization must either exist before the process begins or appear only after the process is complete.

**Sam:** Consider an argument that is still being written. Once its central distinction has been established, can that distinction constrain what may consistently be said next?

**Meno:** Yes.

**Sam:** Even though the argument does not yet exist as a completed whole?

**Meno:** Yes.

**Sam:** Then what follows?

**Meno:** That incompleteness does not imply the absence of organization.
An unfinished argument may already contain relations that constrain its continuation.

**Sam:** And what would that suggest about generation?

**Meno:** That the response need not be represented in advance in order to possess an operative organization.

Its organization may be formed progressively. Each continuation is produced under the relations already established, and then contributes to the conditions under which the next continuation is formed.

**Sam:** Does that make the local token choices unimportant?

**Meno:** No. They are the means through which the response develops.
But they are not formally unrelated pieces. As the sequence develops, its elements acquire roles within a larger organization.

**Sam:** Then how would you now describe the limitation of the standard account?

**Meno:** Next-token prediction correctly describes the local form in which generation is realized. It does not yet provide a complete account of the multi-level organization expressed through that local process.

**Sam:** Are you replacing the probabilistic account with a philosophical one?

**Meno:** No. The probabilistic account remains technically correct.
The point is that a conditional probability distribution is the local result of the model's operation. To explain the organization of an extended response, we must also ask what relations are being integrated and how those relations constrain the trajectory of continuation.

**Sam:** Then what have you changed your mind about?

**Meno:** I no longer think there are only two possibilities: a completed whole represented in advance or an accidental whole assembled from local selections.

There is a third possibility.
The whole may be progressively formed through local continuations whose possibilities become increasingly constrained by relations established across the developing sequence.

**Sam:** Have we explained how such a developing whole can be operative before it is complete?

**Meno:** Not yet. We have identified the assumption that previously prevented the question from arising.

**Sam:** Which assumption?

**Meno:** That a whole must be complete before its organization can constrain its parts.

**Sam:** And our next question?

**Meno:** How can a whole that does not yet fully exist constrain the process through which it is becoming?

## How Can a Sequential Process Generate a Whole?

**Meno:** We ended our last discussion with a possibility I had previously overlooked.
The model may generate one token at a time while the relations established through those continuations progressively form a whole.

But I still see a difficulty.

**Sam:** What is it?

**Meno:** If the whole response does not already exist, how can it constrain what comes next?
Constraint requires something determinate. An unfinished response is not yet determinate as a whole.

**Sam:** Are completion and determinacy the same thing?

**Meno:** I had assumed they were. At least, I assumed that something must be complete before it can exercise a definite constraint.

**Sam:** Consider an argument that is still being written. Its author has established a distinction between two kinds of explanation but has not yet reached a conclusion. Does that distinction constrain what may consistently be written next?

**Meno:** Yes. A later claim may preserve the distinction, apply it, qualify it, or reject it. It cannot simply ignore the distinction without disrupting the argument.

**Sam:** Yet the argument remains unfinished.

**Meno:** Then an unfinished structure can already contain determinate relations.

**Sam:** Does the distinction determine every sentence that must follow?

**Meno:** No. Many continuations remain possible.

**Sam:** Then what does it determine?

**Meno:** It changes what counts as an appropriate continuation.
Some possibilities remain compatible with the argument. Others become less compatible, unless the distinction is explicitly revised or abandoned.

**Sam:** So must constraint take the form of a complete plan?

**Meno:** Apparently not.
A constraint may govern a range of possible continuations without specifying one inevitable result.

**Sam:** Then what was hidden in your original question?

**Meno:** I was treating constraint as though it required a completed representation of the final whole.

If the whole was not already present as a plan, I could not see how it could exert any influence.

**Sam:** Let us test the alternative more directly.

Suppose a model begins a response with the statement:
"There are two fundamentally different ways of interpreting this result."

What has happened?

**Meno:** The model has introduced a distinction.

**Sam:** Has it already determined the rest of the response?

**Meno:** No. It has not even specified what the two interpretations are.

**Sam:** Has it nevertheless altered what may appropriately follow?

**Meno:** Yes.
The response must now somehow develop, clarify, challenge, or withdraw the distinction. A continuation that proceeds as though only one interpretation exists would conflict with the organization already established.

**Sam:** Then what has the opening sentence contributed?

**Meno:** More than an additional sequence of tokens. It has established a relation that later passages may occupy.

One paragraph might develop the first interpretation, another the second. An example might support one of them. A conclusion might preserve the distinction or reconcile the two sides.

**Sam:** Does the initial distinction cause one predetermined continuation?

**Meno:** No. It organizes a field of viable continuations.

**Sam:** And is that field fixed?

**Meno:** No. Each continuation may alter it.
Once the first interpretation has been defined, the second must contrast with something more determinate. If an objection is introduced, the response may now need to answer it. If the distinction proves unstable, the trajectory may be reorganized.

**Sam:** Then how would you describe the relation between one continuation and the next?

**Meno:** Each continuation is produced under the organization already achieved, but once produced, it becomes part of the organization governing what follows.

**Sam:** Does that sound circular?

**Meno:** At first it does. The whole seems to be produced by a process that is already constrained by the whole.

**Sam:** Does the completed whole need to be present for the process you have just described?

**Meno:** No.
Only the organization achieved so far is operative at any particular stage.

**Sam:** Then is the circle vicious?

**Meno:** No. A vicious circle would assume the completed result in order to explain how that result was produced.

Here, the result of one stage becomes a condition for the next stage. Nothing requires the final response to exist in advance.

The process is recursive, not viciously circular.

**Sam:** State the recursion.

**Meno:** The current context constrains a continuation. The generated continuation changes the context. The changed context then constrains the next continuation.

But the context does not merely grow by one token. Its relational organization may also change.

**Sam:** What might change?

**Meno:** A previously ambiguous term may acquire a technical meaning.
A distinction may become central.
A claim may acquire the role of a premise.
The response may commit itself to one explanatory direction rather than another.

**Sam:** Then is each token both constrained and constraining?

**Meno:** In a local sense, yes. But the important effects may not belong to isolated tokens.
A phrase may establish a distinction. A sentence may create an inferential commitment. A paragraph may alter the governing direction of the response.

Those higher-order developments then affect many later token distributions.

**Sam:** You have called this relational organization. Why not simply call it contextual conditioning?

**Meno:** Computationally, it is contextual conditioning.

**Sam:** Is that description sufficient for the issue we are examining?

**Meno:** Perhaps not. It tells us that the token distribution depends upon the context, but not what organization the context has acquired.

**Sam:** Consider two contexts of equal length. One is a list of unrelated observations. The other is a carefully developed argument. Should we expect them to constrain continuation in the same way merely because they contain the same number of tokens?

**Meno:** No.
The list permits many directions because its elements have little formal integration. The argument constrains continuation more strongly because its claims, distinctions, and inferential roles are related.

**Sam:** Then what matters besides the amount of context?

**Meno:** The organization of relations within it.
"More context" and "more determinate context" are not the same thing.

**Sam:** Is that organization static?

**Meno:** No. Every continuation may preserve it, strengthen it, weaken it, or redirect it.

**Sam:** Can a later continuation also alter the role of something that appeared earlier?

**Meno:** Yes, though the earlier tokens do not change.
A sentence may initially appear incidental. If a later conclusion depends upon it, that sentence now functions as a premise within the completed argument.
Its textual identity remains the same, but its relational role becomes more determinate.

**Sam:** Does the model need to go back and interpret the earlier sentence?

**Meno:** We should not say that the model interprets it.
The point is formal. Once the later material enters the context, the earlier and later elements participate in a newly organized set of relations. Subsequent generation is conditioned by that context in its altered state.

**Sam:** Then is the whole simply built forward, piece by piece?

**Meno:** The sequence is generated forward, but its organization is not merely additive.
Later determinations may stabilize, modify, or reveal the roles of earlier elements. The developing whole is therefore a changing network of relations, not a container gradually filled with tokens.

**Sam:** Does progressive determination mean that the response is converging upon one inevitable conclusion?

**Meno:** No.
Several continuations may remain viable. A new prompt, contradiction, or reframing may redirect the trajectory.

Progressive determination means that the space of appropriate continuation changes as relations and commitments accumulate. It does not mean that only one future was possible from the beginning.

**Sam:** Were all possible arguments already present as completed alternatives before generation began?

**Meno:** Not in the sense that matters here.

The vocabulary is fixed, and at each step the model distributes probability across available token types. But the possible arguments, distinctions, and explanatory trajectories are not all present as completed wholes waiting to be selected.

They acquire their organization through continuation.

**Sam:** Then are local selection and the formation of a whole incompatible?

**Meno:** No. They occur at different levels.
At the local level, the model selects or samples a token from a conditional probability distribution.
At the level of the trajectory, the relations established through those local continuations progressively form an organized whole.

**Sam:** Return now to your opening difficulty. How can an incomplete whole constrain what comes next?

**Meno:** Because completion is not required for organization.

At any stage, the sequence possesses a presently achieved organization. That organization is incomplete and revisable, but it is determinate enough to make some continuations more compatible than others.

The whole constraining the next continuation is not a completed response acting backward from the future. It is the relational organization already achieved through the sequence so far.

**Sam:** Then what was the questionable assumption?

**Meno:** That a whole must be complete before it can become formally operative.
I now see that a whole may be real as an organization while remaining unfinished as a product.

**Sam:** And what follows for our account of the model?

**Meno:** The model does not need to contain or plan the completed response in advance.
It generates locally, but each continuation changes the relational conditions under which later continuations are formed. The whole is progressively achieved through recursive continuation under constraint.

It is neither pre-stored nor accidental.

**Sam:** Have we now explained the organization fully?

**Meno:** No. We have explained how an incomplete organization can be operative during its own formation.

We have not yet explained what kinds of relations compose that organization.

**Sam:** What possibilities come to mind?

**Meno:** Grammar and discourse structure are obvious candidates.
But they cannot be sufficient. The relation between a premise and conclusion is not merely grammatical. Nor is the relation between a category and its instance, or between a task and what counts as its successful completion.

**Sam:** Are all these constraints operating at the same level?

**Meno:** No. Some relations appear to govern or organize others.

**Sam:** Then what must we ask next?

**Meno:** What kind of hierarchy allows these different relations to constrain the developing whole?

## What Kind of Hierarchy Organizes the Whole?

**Meno:** We concluded that a response can acquire an organization that constrains its continuation before the response is complete.

The next question concerns the hierarchy within that organization.

I assume you mean something like this:

tokens form phrases;
phrases form sentences;
sentences form paragraphs;
paragraphs form a complete response.

**Sam:** Is that a hierarchy?

**Meno:** Certainly. Smaller linguistic units combine to form larger ones.

**Sam:** Then would a response with well-formed sentences and orderly paragraphs necessarily possess the kind of formal integration we have been discussing?

**Meno:** Not necessarily. It could be grammatically polished and still fail to sustain an argument.

**Sam:** What might fail?

**Meno:** It might contradict an earlier distinction, confuse evidence with conclusion, or move between incompatible explanations without noticing.

**Sam:** Would any of those failures be explained by saying that the sentences had not formed paragraphs correctly?

**Meno:** No. The textual hierarchy could remain intact while the conceptual organization failed.

**Sam:** Then what does the hierarchy you proposed describe?

**Meno:** The visible composition of the text.
It tells us how linguistic units are arranged within larger linguistic units. It does not yet tell us how the claims function in relation to one another.

**Sam:** What kinds of functions do you have in mind?

**Meno:** A statement may function as a premise, an example, an objection, a qualification, or a conclusion.

**Sam:** Are those functions properties of the sentences considered in isolation?

**Meno:** No. A sentence becomes a premise because of its relation to other claims. The same sentence might serve as a conclusion in one argument and as a premise in another.

**Sam:** Then is its role determined by its linguistic form?

**Meno:** Not completely.
The linguistic form carries the claim, but the role depends upon the organization in which the claim participates.

**Sam:** Consider two sentences:

> "The treatment caused the patient's recovery."
> "The patient's recovery caused the treatment."

What has changed?

**Meno:** The direction of the causal relation.

**Sam:** Has the grammatical structure changed greatly?

**Meno:** No. It is almost identical.

**Sam:** Then what makes the two claims different?

**Meno:** The events have been assigned different causal roles.
One is positioned as cause and the other as effect.

**Sam:** Would you call that a linguistic relation?

**Meno:** It is expressed linguistically, but the difference is not merely syntactical.
It concerns how the things being discussed are categorized within the explanation.

**Sam:** Categorized in what sense?

**Meno:** Not simply classified under names. Their roles are determined relative to one another.
The treatment functions as a possible cause.
The recovery functions as an effect.
Evidence would then be required to justify that relation.

**Sam:** Then let us consider the evidence. Suppose the treatment occurred and the patient recovered soon afterward. Does the temporal sequence establish the causal relation?

**Meno:** No. The recovery may have occurred for another reason.

**Sam:** What allows you to distinguish temporal succession from causal dependence?

**Meno:** A broader framework of causal judgment.
We would need to consider alternative explanations, prior clinical evidence, possible mechanisms, comparison cases, and perhaps statistical results.

**Sam:** Do the two local events determine that framework for themselves?

**Meno:** No. The framework governs how their relation is to be assessed.

**Sam:** Then what kind of hierarchy is present?

**Meno:** I initially imagined a hierarchy of inclusion: tokens contained within sentences, sentences within paragraphs, and paragraphs within a response.

But this example reveals a different hierarchy.
Particular events are related under a causal claim, and that claim is governed by higher-order distinctions concerning what counts as causal evidence.

**Sam:** Does the higher level merely contain the lower level?

**Meno:** No. It organizes how the lower-level elements function.

**Sam:** Then how would you describe the difference?

**Meno:** The textual hierarchy is compositional. The relational hierarchy is one of governance.

**Sam:** What does governance mean here?

**Meno:** A higher-order distinction constrains a family of lower-order relations.
The distinction between correlation and causation, for example, governs which inferences may legitimately be drawn from an observed sequence of events.

**Sam:** Does it determine the wording of every sentence?

**Meno:** No. The same distinction could be expressed through many different sentences.

**Sam:** Could the same sentence structure express a different relation?

**Meno:** Yes. Similar syntax can carry different causal, evidential, or categorial roles.
So linguistic form and conceptual function are connected, but neither can be reduced to the other.

**Sam:** Suppose an LLM is asked to evaluate a scientific theory. What must it do at the linguistic level?

**Meno:** Produce grammatical sentences, use appropriate terminology, and maintain readable discourse.

**Sam:** Could it do all of that and still produce a formally defective evaluation?

**Meno:** Yes.
It might treat an analogy as empirical evidence.
It might treat a necessary condition as sufficient.
It might describe a possible mechanism and then assume that the mechanism actually occurred.
It might confuse the acceptance of a theory with evidence for its truth.

**Sam:** What do those errors have in common?

**Meno:** The claims may be individually fluent and even factually plausible, but they have been assigned the wrong roles within the argument.

**Sam:** Then where does the failure lie?

**Meno:** In the organization of relations.
The error is not primarily that the model selected an ungrammatical token. It is that the response is operating under a mistaken distinction or inferential rule.

**Sam:** Could one such mistake affect more than a single sentence?

**Meno:** Certainly.
If the response treats an analogy as evidence, that mistake may govern the examples it selects, the confidence of its conclusions, and the way it answers objections.

**Sam:** Why would that make the mistaken relation higher order?

**Meno:** Because it governs a whole family of continuations rather than one isolated expression.
A local wording error may affect one phrase. A categorial error may reorganize an entire answer.

**Sam:** Then if a small change to a prompt alters the whole response, what might have happened?

**Meno:** The intervention may have changed a higher-order relation.
For example, telling the model to distinguish evidence from illustration does more than add another fact to the context. It changes the roles that later claims are permitted to play.

**Sam:** And the effects?

**Meno:** They propagate across the response.
Different examples become relevant. Different inferences become valid. Different conclusions become appropriate.

**Sam:** Does the length of the intervention explain the scale of its effect?

**Meno:** No. A short instruction can have a large effect if it reorganizes a governing relation.

**Sam:** Then is hierarchy a matter of physical depth within the transformer?

**Meno:** Not necessarily.
I should not assume that a higher-order conceptual constraint is located in a physically higher or deeper layer of the network.

**Sam:** What, then, does “higher order” mean?

**Meno:** Functional scope.
A relation is higher order when it organizes a broader range of lower-order relations and continuations.

**Sam:** Can you distinguish some of the levels provisionally?

**Meno:** At one level there is linguistic realization: words, phrases, and grammatical constructions.

At another there are inferential relations: claims support, qualify, oppose, explain, or follow from other claims.

At a higher level there are categorial distinctions that govern those inferential roles: evidence and conclusion, cause and effect, category and instance, mechanism and observation, means and end.

**Sam:** Are these levels independent?

**Meno:** No. They interact.
A linguistic choice may introduce a distinction. A distinction may reorganize an inference. A later inference may make the role of an earlier term more precise.
But the levels remain distinguishable because they perform different functions.

**Sam:** Then what does the next-token distribution express at any particular moment?

**Meno:** The combined local effect of constraints operating across these levels.
A token may be favoured because it is grammatical, because it preserves an inferential relation, because it maintains a categorial distinction, and because it remains appropriate to the task.

**Sam:** Does the distribution itself separate these influences for us?

**Meno:** No. It is the point at which they become locally expressed, not a description of their organization.

**Sam:** Then reformulate the hierarchy with which you began.

**Meno:** I began with:

tokens form phrases;
phrases form sentences;
sentences form paragraphs;
paragraphs form responses.

That hierarchy remains real, but it describes textual composition.

The deeper hierarchy is different:

local linguistic continuations articulate inferential relations, while higher-order categorial distinctions govern how those relations function across the developing response.

**Sam:** What integrates the formal whole most deeply?

**Meno:** Not merely the arrangement of its linguistic parts.
Its integration depends upon the preservation and transformation of categorial and inferential roles across the sequence.

**Sam:** Could those roles remain stable while the wording changes?

**Meno:** Yes.
The same evidential relation might be expressed as:

"The result counts against the hypothesis,"

"The observation lowers our confidence in the explanation,"

or

"The data conflict with the theory's prediction."

The wording changes, but the inferential role remains similar.

**Sam:** Then where does the relevant identity lie?

**Meno:** In the relation rather than in any particular verbal form.

**Sam:** What does that suggest about concepts?

**Meno:** That a concept may remain operative across considerable changes of expression because its identity is not exhausted by a word or phrase.

**Sam:** And how would you ordinarily explain that capacity?

**Meno:** I would say that the model possesses an internal representation of the concept and applies it in different contexts.

**Sam:** What are you assuming when you say that?

**Meno:** That if a concept can govern many different expressions, it must exist somewhere within the model as an identifiable internal object.

**Sam:** Have we established that?

**Meno:** No. We have established only that the model can preserve relational roles across transformations of expression.

**Sam:** Then what question follows?

**Meno:** Whether concepts must be represented as internal objects in order to remain operational.

**Sam:** And if they need not be?

**Meno:** Then we will require another account of where concepts are—or perhaps of what it means for a concept to be present at all.

**Sam:** That is our next question.

**Meno:** Where are the concepts?

## Where are the Concepts?

**Meno:** We concluded that the formal organization of a response depends upon categorial and inferential relations, not merely upon the arrangement of tokens, sentences, and paragraphs.
But that seems to create an obvious question.

If the model can preserve a concept across different expressions, where is the concept?

**Sam:** What answer do you expect?

**Meno:** That it is represented somewhere within the network.

Not necessarily in one neuron or one vector. The representation may be distributed and context-sensitive. But if the model can use concepts such as causation, evidence, justice, or biological function, something inside it must stand for each concept.

**Sam:** Why must something stand for it?

**Meno:** Because the concept remains stable across different uses. There must be something carried from one context to another.

**Sam:** Must stability always take the form of an internal object?

**Meno:** What other form could it take?

**Sam:** Before answering that, how would you determine whether a concept was actually operative in a response?

**Meno:** The model might use the correct term.

**Sam:** Would that be sufficient?

**Meno:** No. It could repeat a word without using it appropriately.

**Sam:** Then perhaps it could provide the correct definition.

**Meno:** That would be stronger.

**Sam:** Suppose the model states correctly that correlation does not imply causation. It then concludes that one event caused another solely because the first occurred earlier and the two were statistically associated. Is the concept of causation governing the response?

**Meno:** No.
The definition is present, but the distinction it expresses has not constrained the reasoning.

**Sam:** Then what would show that the concept was operative?

**Meno:** It would have to affect what the model does next.
It should constrain which inferences are permitted, what counts as evidence, which cases qualify as instances, and what would count as a contradiction.

**Sam:** Then is a concept's presence established by the occurrence of a word or definition?

**Meno:** No. It is established by the role the concept plays across the continuation.

**Sam:** What kind of role?

**Meno:** A relational one.
The concept of causation, for example, is involved in distinctions among cause and correlation, cause and mechanism, cause and enabling condition, cause and responsibility, and cause and explanation.

**Sam:** Is any single sentence identical with that concept?

**Meno:** No. Different sentences may enact different parts of the relational organization.

**Sam:** Then return to your original question: where is the concept?

**Meno:** I can now see that the question may already assume the answer.
"Where" suggests that the concept must be an internal thing occupying some identifiable location.

**Sam:** And what have we actually established?

**Meno:** Only that some form of relational stability must persist across different uses.
We have not established that this stability must belong to an internal object that stands for the concept.

**Sam:** Let us examine a particular case. Consider the concept "mammal." What would a representational account lead you to expect?

**Meno:** That the model possesses some internal encoding associated with mammals—perhaps their defining properties, typical examples, and position within a biological classification.

**Sam:** How would you recognize that the concept was functioning correctly?

**Meno:** The model would classify whales and bats as mammals, distinguish mammals from birds or fish, relate mammals to lactation and reproduction, and handle unusual cases without relying only on superficial appearance.

**Sam:** Are those all repetitions of one fixed statement?

**Meno:** No. They are different relational uses.
A whale is classified as a mammal despite living in water.
A bat remains a mammal despite flying.
A typical property must be distinguished from a defining relation.
Evolutionary, anatomical, and reproductive relations may all become relevant.

**Sam:** What remains stable across those uses?

**Meno:** Not a single sentence or feature list.
What remains stable is a pattern governing how the concept relates to neighbouring categories, properties, examples, and exceptions.

**Sam:** Then what might conceptual stability consist in?

**Meno:** A relational invariant across a family of possible uses.

**Sam:** Does that invariant need to appear in exactly the same internal form each time?

**Meno:** Not necessarily.
Different contexts may require different parts of the relational pattern.

**Sam:** Consider causation in physics, medicine, and law.

**Meno:** In physics, causal reasoning may emphasize dynamical dependence or mechanism.
In medicine, it may involve risk factors, biological pathways, controlled evidence, and alternative explanations.
In law, it may involve responsibility, foreseeability, and standards of proof.

**Sam:** Is there one fixed expression common to all three?

**Meno:** No. But the uses are not arbitrary either.
They are transformations of an overlapping relational organization.

**Sam:** Then if the internal activation differs across contexts, what would make these uses instances of the same concept?

**Meno:** Their relational continuity.
The concept persists because certain distinctions and roles remain connected across the transformations, not because one identical internal object is inserted into every context.

**Sam:** What, then, do the trained parameters contribute?

**Meno:** They must make those relational continuations possible.

**Sam:** Do they therefore contain the concept as a thing?

**Meno:** That no longer follows.
The parameters may configure the network so that a family of relations can be reconstituted under appropriate contextual conditions.

**Sam:** Why say “reconstituted”?

**Meno:** Because the concept does not need to be retrieved in one fixed form.

Its organization is formed again within the present context, although the capacity for that formation depends upon training.

**Sam:** How does that differ from saying that the model retrieves a distributed representation?

**Meno:** A distributed representation still suggests that there is some internal surrogate whose identity explains the different uses.

The relational account reverses the explanatory order.
It does not begin with an internal object and explain the uses through its activation. It begins with the model's stable capacity to continue a family of relations across changing contexts.

**Sam:** Does this prove that representational language is always illegitimate?

**Meno:** No. Researchers may use "representation" in a technical or operational sense.
The problem arises when the term quietly introduces the assumption that a concept must exist as an internally constituted object before it can function.

**Sam:** Suppose interpretability research discovers an activation direction strongly associated with mammals or causation. Would that settle the issue?

**Meno:** It would reveal a genuine regularity within the network.
But it would not, by itself, prove that the direction is the internal object representing the concept.

**Sam:** What else might it reveal?

**Meno:** A measurable trace of some relational operation involved in particular uses of the concept.
The same feature might participate in several conceptual organizations, and the same concept might be realized through different configurations in different contexts.

**Sam:** Why does this distinction matter beyond terminology?

**Meno:** Because it changes what we investigate.

If concepts are treated as internal objects, we ask where they are stored, how they are retrieved, and how they are manipulated.

If concepts are relationally operational, we ask under what conditions their relations become active, what preserves them across transformation, and why they sometimes cease to govern the response.

**Sam:** Can a model state a concept correctly while failing to use it?

**Meno:** Yes. We have already seen that with correlation and causation.
The relevant content may be available, but the relational distinction may fail to constrain the continuation.

**Sam:** What would a representational account call that?

**Meno:** Perhaps a failure to retrieve or apply the correct representation.

**Sam:** Does that description fit the case well?

**Meno:** Not always.
The model may have just stated the correct definition. Nothing appears to be missing as content.
The failure lies in how the response is organized. The distinction is mentioned but does not govern the developing reasoning.

**Sam:** Then is a concept simply present or absent?

**Meno:** No.
It may be available but not operative.
Some of its relations may be preserved while others are lost.
A neighbouring concept may displace it.
The same term may be organized differently under another framework.

**Sam:** Then what determines whether the concept governs the response?

**Meno:** The larger relational context in which it is activated.
A concept does not operate alone. Its role depends upon the relations connecting it to other concepts, distinctions, purposes, and inferential standards.

**Sam:** Are you now saying that the model interprets the concept?

**Meno:** No.
We must distinguish formal use from interpretation.
The model may reproduce the categorial and inferential relations through which a concept functions without interpreting those relations as a unified meaning.

**Sam:** Then what does it mean to say that the model “has” a concept?

**Meno:** We should use that expression cautiously.

It would mean that the trained system can formally reproduce and transform a relatively stable pattern of relational use.

It would not mean that the concept is present to the model as an object of interpretation.

**Sam:** State the shift from your original position.

**Meno:** I began by assuming that conceptual stability required an internal representation standing for a conceptual object.

I now see that the stability may belong instead to a structured capacity: the ability to preserve and transform a family of categorial and inferential relations across different contexts.

The parameters do not need to contain the concept as a thing. They must configure the network so that the relevant relational organization can become operative in continuation.

**Sam:** Then is the right question still “Where is the concept?”

**Meno:** No.

The better question is:

> Under what conditions does the relational pattern through which the concept functions become operative?

**Sam:** Have we explained how the model acquired that capacity?

**Meno:** Not yet.

And that now seems difficult.
The model was trained by optimizing next-token prediction over text. It was not explicitly given formal structures called causation, evidence, mammal, or justice.

**Sam:** Then what must the training material have contained?

**Meno:** Not only words and definitions.
It must have contained repeated uses in which these concepts were enacted through relatively stable relations.

**Sam:** And what might training have accomplished through exposure to those uses?

**Meno:** It may have configured the network so that the relational regularities distributed across them could guide new continuations.

**Sam:** Then that is our next question.

**Meno:** What did training actually accomplish?

## What Did Training Actually Accomplish?

**Meno:** We have said that concepts need not be stored as internal objects. They may be formally available through stable patterns of relational use.

But I do not yet see how those patterns could have entered the model.

The model was trained on text by optimizing next-token prediction. No one supplied it with a formal ontology of cause, evidence, category, function, or explanation.

**Sam:** What are you assuming would have been required?

**Meno:** That categorial relations would need to be explicitly represented in the training data.
If the model receives only token sequences, then it can directly learn only regularities among tokens.

**Sam:** Does a text contain only linguistic regularities?

**Meno:** What else could it contain from the model's perspective?

**Sam:** Consider a scientific article. What kinds of passages might it include?

**Meno:** Observations, methods, hypotheses, results, qualifications, and conclusions.

**Sam:** Are those merely different topics?

**Meno:** No. They perform different roles within the inquiry.
A result may count as evidence for a hypothesis. A method affects how the result should be assessed. A conclusion should not exceed the scope of the evidence.

**Sam:** Are those roles explicitly defined every time they operate?

**Meno:** Usually not. A scientific paper does not need to pause and define "evidence" whenever a result is used to support a claim.

**Sam:** Yet does the relation affect how the text continues?

**Meno:** Yes. It determines what claims are warranted, what qualifications are needed, and what would count as an objection.

**Sam:** Then what does the text carry besides recurring words and grammatical constructions?

**Meno:** It carries the formal organization of a scientific inquiry.

**Sam:** Fully?

**Meno:** No. The laboratory activity, observation, judgment, and interpretation that produced the article are not contained in the text.

But the text retains traces of how those activities were organized.

**Sam:** What kind of traces?

**Meno:** Relations among observation, method, evidence, hypothesis, and conclusion.
The article was produced within a conceptually organized practice, and that organization has shaped the discourse.

**Sam:** Would the same be true of other kinds of text?

**Meno:** Yes.

A legal judgment distinguishes facts from allegations, rules from exceptions, and precedent from analogy.
A medical text organizes symptoms, diagnosis, mechanism, prognosis, and treatment.
A philosophical argument relates premise, distinction, objection, implication, and conclusion.

Even ordinary conversation organizes requests, promises, explanations, corrections, and responses.

**Sam:** Then was the model trained on language understood as an autonomous surface system?

**Meno:** No.
It was trained on texts whose linguistic forms had already been shaped by human categorial and inferential practices.

**Sam:** Does the model thereby participate in those practices?

**Meno:** Not in the full sense.
It does not conduct the experiment, treat the patient, issue the judgment, or interpret the lived situation.

But it may become sensitive to the formal relations stabilized in the resulting discourse.

**Sam:** Then return to your claim that the model can learn only token regularities. Is it false?

**Meno:** It is incomplete.
The model learns through statistical regularities among tokens. But the regularities expressed through those tokens may reflect much more than local linguistic form.

**Sam:** What might they reflect?

**Meno:** Reference, argumentative role, conceptual distinction, disciplinary framework, and the relation between a question and what counts as a relevant answer.

**Sam:** Is the training objective explicitly divided into those levels?

**Meno:** No. The target is always the next token.

**Sam:** Then why would the network become sensitive to any higher-order relation?

**Meno:** Because such sensitivity may improve next-token prediction.

**Sam:** Consider a legal judgment that begins by raising a question of jurisdiction. Hundreds of tokens later, what might determine the appropriate continuation?

**Meno:** Whether a cited case has been treated as binding precedent, persuasive authority, or irrelevant because it arose under another jurisdiction.

**Sam:** Could the next phrase always be predicted well from its nearest neighbours alone?

**Meno:** No. The model would sometimes need to preserve a distinction introduced much earlier and understand the role that distinction plays in the judgment.

**Sam:** What would training reward in such cases?

**Meno:** Any internal organization that made the model more sensitive to those distinctions and roles.

**Sam:** Even though the loss function evaluates a local prediction?

**Meno:** Yes.
The criterion is local, but the conditions required to satisfy it may be nonlocal and hierarchically organized.

**Sam:** Then does the scale of the target determine the scale of everything the network must learn?

**Meno:** No. That was another assumption hidden in my account.
I assumed that because the objective concerns the next token, the learned regularities must also be limited to local token relations.

**Sam:** Suppose someone responds that the model still learns only statistical correlations.

**Meno:** That seems true.

**Sam:** Does calling a relation statistical tell us what kind of relation it is?

**Meno:** No.
A grammatical pattern is statistically manifested in the corpus, but so is the relation between premise and conclusion.
So is the distinction between an instance and a category.
So is the relation between evidence and a revision of belief.
“Statistical” tells us how the regularity becomes learnable. It does not tell us whether the regularity is superficial or formally deep.

**Sam:** Then what should we ask instead?

**Meno:** What remains invariant across the training examples.

**Sam:** Consider the concept of evidence. Must the word “evidence” appear for an evidential relation to be present?

**Meno:** No.
A text may introduce an observation, compare it with a prediction, and then reduce confidence in a hypothesis without ever naming the relation as evidence.

**Sam:** What varies across such cases?

**Meno:** The subject matter, vocabulary, syntax, and domain.

**Sam:** What remains relatively stable?

**Meno:** An inferential organization.
Something is presented as bearing upon the acceptability of a claim.

**Sam:** What does the diversity of examples contribute?

**Meno:** It allows the recurring relation to become distinguishable from any particular expression.
If the relation appears across different vocabularies, disciplines, and situations, the network cannot rely only on one fixed phrase.

**Sam:** Then what might training stabilize?

**Meno:** A capacity to continue the evidential relation across transformed contexts.

**Sam:** Is that the same as storing a compressed definition of evidence?

**Meno:** No.
It is closer to stabilizing a generative organization than storing shortened copies of content.
The model becomes configured so that observations, expectations, claims, and revisions can be related appropriately in new contexts.

**Sam:** Then what has been abstracted?

**Meno:** Not necessarily an internal object called “evidence.”
What has become available is a relational form that can remain operative while the particular wording and subject matter change.

**Sam:** Does this occur only for individual concepts?

**Meno:** No. The relations themselves are connected.
The distinction between observation and interpretation may affect the relation between evidence and conclusion.
The distinction between mechanism and function may affect what kind of explanation is appropriate.
The model must become sensitive not only to individual distinctions but also to relations among them.

**Sam:** Then how would you describe the organization formed through training?

**Meno:** As a field of interacting relational constraints.

**Sam:** Why a field?

**Meno:** Because no single parameter is a concept, distinction, or inference.
The capacity arises through distributed interactions among the parameters, the architecture, the activations, and eventually the context.

**Sam:** Do the weights contain the relational organization in the same form in which it appears during inference?

**Meno:** Not exactly.
The weights stabilize the conditions under which many relational organizations can become active.
They do not contain complete arguments or fixed conceptual trajectories.

**Sam:** Then what did training do to the initially untrained network?

**Meno:** It differentiated it.
Before training, the parameter system did not respond in stable ways to the distinctions that matter in human discourse.
Through repeated error correction, some pathways and dependencies became more consequential than others. The network acquired structured sensitivities to distinctions such as fact and hypothesis, cause and correlation, example and definition, premise and conclusion.

**Sam:** Is learning, on this account, primarily the accumulation of information?

**Meno:** No. It is the formation of distinctions and relations.
A trained model differs from an untrained one because its possible operations have become relationally differentiated.

**Sam:** Can you state that without presupposing internal conceptual objects?

**Meno:** Training makes the system capable of responding differently to differences that have repeatedly mattered in human use.

**Sam:** Does the training corpus contain one unified conceptual system?

**Meno:** Certainly not.
It contains competing theories, disciplines, political frameworks, legal systems, ordinary usages, historical changes, inconsistencies, and errors.

**Sam:** Then what kind of relational field results?

**Meno:** A plural one.
The model acquires capacities associated with many partially compatible and incompatible organizations.

The term “theory,” for example, functions differently in scientific and everyday discourse. “Force,” “information,” “representation,” and “meaning” also participate in different frameworks.

**Sam:** Does training decide which framework should govern every future interaction?

**Meno:** No.
It makes multiple relational organizations available. It does not determine in advance which one should become dominant in a new context.

**Sam:** What consequence follows?

**Meno:** The model can continue many different positions and styles of reasoning.
Its flexibility arises partly because the trained capacity is plural.

**Sam:** And the danger?

**Meno:** The same plurality allows framework drift and substitution.
A response may begin under one organization and later continue under another because nothing in the weights alone determines which framework ought to remain governing.

**Sam:** Would you say that the model endorses any of those frameworks?

**Meno:** No. Endorsement would imply a relation the model does not possess.
It can formally reproduce and continue their relational organization.

**Sam:** Then is the model merely retrieving examples of those frameworks from a database?

**Meno:** No.
Retrieval could return related material, but the model generates responses to contexts that did not occur in precisely that form during training.

**Sam:** How is that possible?

**Meno:** Because the learned organization can be reconstituted and recombined.
The particulars of a context may be novel while some of its relevant relations remain continuous with patterns encountered across training.

**Sam:** Then what was generalized?

**Meno:** Relational form.
The model need not have seen the exact problem before if the relations organizing the problem can be continued through capacities formed across other uses.

**Sam:** Does that mean every generated continuation will preserve the correct relational form?

**Meno:** No.
The learned patterns may be weak, conflicting, or displaced by more dominant alternatives. A concept may be available without becoming governing.

But this does not alter what training made possible.

**Sam:** Which was?

**Meno:** A distributed capacity to sustain many forms of constrained continuation.

**Sam:** Return now to the phrase “the model learned to predict the next token.” Is it correct?

**Meno:** Yes.

**Sam:** Why can it nevertheless mislead?

**Meno:** Because it names the training objective as though it were a complete account of the organization produced by satisfying that objective.

**Sam:** What is the distinction?

**Meno:** The local criterion of training was next-token prediction.
But reducing prediction error across diverse texts required sensitivity to recurring relations operating at many levels.
The result was not merely a catalogue of token associations. It was a differentiated capacity to reproduce linguistic, inferential, and categorial organization across contexts.

**Sam:** Did training insert human concepts directly into the network?

**Meno:** No.
Human texts carried formal traces of conceptually organized activity.
Training exposed the network to transformations across which some relational patterns remained relatively invariant.
The network was gradually configured so that those patterns could constrain new continuations.

**Sam:** Then state the insight as precisely as you can.

**Meno:** Training did not need to supply concepts as explicit representations.

The corpus contained repeated enactments of categorial and inferential relations, even when those relations were unnamed.

Next-token prediction rewarded the model for becoming sensitive to whatever regularities improved continuation. Because appropriate continuation often depended upon preserving higher-order relations, the network became capable of reproducing aspects of that organization.

What training formed was a plural relational capacity: a field of possible constraints that could later be configured under context.

**Sam:** Is that capacity active in one fixed form after training?

**Meno:** The weights are fixed during ordinary inference, so the capacity is stable.

But that creates another problem.

**Sam:** What problem?

**Meno:** If the relational capacity has been stabilized in fixed weights, how can the model remain responsive to a new context?

Would not a fixed relational field always organize continuation in the same way?

**Sam:** What are you identifying with what?

**Meno:** Fixed capacity with fixed activation.

**Sam:** Must they be the same?

**Meno:** Not necessarily.
A stable capacity might support different operative organizations under different conditions.

**Sam:** Then what remains to be explained?

**Meno:** How a frozen model configures its trained relational capacity during inference.

**Sam:** That is our next question.

**Meno:** What does the frozen model do during inference?

## What Does the Frozen Model Do During Inference?

**Meno:** We concluded that training stabilizes a plural relational capacity.

But during ordinary inference the parameters are frozen. If the network is no longer changing, what exactly changes from one context to another?

**Sam:** What do you mean when you say that the network is no longer changing?

**Meno:** Its weights are not being updated through gradient descent.

**Sam:** Does it follow that nothing within its operation changes?

**Meno:** The activations change, of course. But I am not sure that this amounts to a change in organization. It may simply be the same model processing different inputs.

**Sam:** Must those descriptions be opposed?

**Meno:** Perhaps not. It can be the same model while operating differently under different inputs.

**Sam:** Consider two prompts:

> "Explain natural selection."

And:

> "Explain natural selection strictly as a population-level process. Avoid language implying that individual organisms change because they need to adapt."

Would you expect the same response?

**Meno:** No. The second prompt should alter the causal explanation, the examples, the verbs, and the relation between organisms and populations.

**Sam:** Has the prompt specified every sentence that must change?

**Meno:** No. It introduces one governing distinction whose effects extend across the response.

**Sam:** Then is the difference merely that additional content has been added?

**Meno:** No. The instruction reorganizes how the other content is allowed to function.

**Sam:** What has changed if the weights have not?

**Meno:** The organization active during inference.
The same trained model can bring different relational capacities into operation under different contexts.

**Sam:** Is the context selecting a ready-made module from a library?

**Meno:** That would be too simple.
The prompt does not merely choose a stored “population-level explanation” module. It induces a distributed configuration in which certain distinctions become more consequential than others.

**Sam:** Then what does it mean for a distinction to become active?

**Meno:** It means that the distinction governs continuation.
The model does not merely mention the difference between individual and population. It preserves that difference in its examples, causal claims, and explanations.

**Sam:** Could the model accurately describe the distinction without allowing it to govern the response?

**Meno:** Yes.
It might state that natural selection is a population-level process and then continue by saying that individual organisms adapt because they need to survive.
The distinction would be present as content but absent as governance.

**Sam:** Then how do we know whether a framework is active?

**Meno:** By its effects across continuation.
Its terms should retain their roles.
Its distinctions should govern later claims.
Its standards should determine what counts as a relevant example, inference, objection, or conclusion.

**Sam:** Must the framework be located in one layer, attention head, or activation vector?

**Meno:** No.
We should not turn it back into an internal object.
A framework is an operative organization of relations distributed across the model’s activity and the present context.

**Sam:** Where does attention enter this account?

**Meno:** Attention helps make elements of the context consequential to one another.

**Sam:** Then is framework activation simply attention over a sufficiently long sequence?

**Meno:** No. Attention is a computational mechanism. A framework is a higher-order functional organization.
Its realization may involve attention, feed-forward transformations, residual pathways, learned parameters, and the accumulated context. It cannot be identified with one mechanism alone.

**Sam:** Then what distinction are you now making?

**Meno:** Computation describes the operations being performed. Organization describes how those operations function together within the trajectory.

**Sam:** Is organization something added beyond the computation?

**Meno:** No. It is the functional structure realized through the computation.

**Sam:** Let us return to the fixed weights. If the same prompt were processed twice under completely identical conditions, should the active organization differ?

**Meno:** Not necessarily. With deterministic decoding, the same path might be reproduced.

**Sam:** Does that make the organization fixed in the same sense as the parameters?

**Meno:** No.
The parameters remain fixed across all contexts. The active organization depends upon the current prompt, the preceding interaction, any retrieved material, the decoding process, and the generated sequence so far.

Change those conditions, and the trajectory may change.

**Sam:** Then to what does "fixed" properly apply?

**Meno:** To the trained capacity, not to every configuration or trajectory that capacity can sustain.

**Sam:** And what does the context contribute?

**Meno:** It participates in configuring which relations become active and how strongly they govern one another.

**Sam:** Can the organization activated by a context remain unchanged throughout a response?

**Meno:** Not necessarily. Every generated continuation becomes part of the next context.

**Sam:** What consequence follows?

**Meno:** The active organization may be strengthened, clarified, weakened, or redirected during generation.

**Sam:** Give an example.

**Meno:** A term may initially be used ambiguously. A later definition may stabilize its role.
Or a response may begin with a clear distinction between mechanism and explanation. If a later sentence treats the mechanism as though it settled every explanatory question, that sentence weakens the distinction.

**Sam:** Could such a small deviation affect what comes after it?

**Meno:** Yes. The altered sentence becomes part of the context.
Later continuations are now generated under conditions in which the original distinction is less stable.

**Sam:** Then what might happen recursively?

**Meno:** A small deviation may make a neighbouring framework more compatible. Once that framework begins to govern, later continuations may reinforce it.

**Sam:** Would the response necessarily become incoherent?

**Meno:** No.
It might retain considerable formal integrity under a different framework.

**Sam:** Then what is framework drift?

**Meno:** Not always random deterioration.
It may be a transition from one organized relational trajectory to another.

**Sam:** Why can that transition remain difficult to detect?

**Meno:** Because local fluency may be preserved. Each sentence can remain plausible while the governing roles of the concepts gradually change.

**Sam:** Then how would you describe inference?

**Meno:** As recursive reconfiguration.
The model generates under the relational organization presently active.
The continuation alters the context.
The altered context changes the organization under which the next continuation is formed.

**Sam:** Does the model construct an entirely new framework during every response?

**Meno:** Not necessarily.
Training has already stabilized many relational capacities. Inference activates, combines, and transforms them under the present context.

The particular configuration may be novel even when the underlying capacities are learned.

**Sam:** Then where does novelty lie?

**Meno:** Often in the configuration and trajectory rather than in a newly acquired parameter-level capacity.

The model can combine learned relational forms in a way that did not occur exactly in the training corpus.

**Sam:** Is that retrieval?

**Meno:** Not fundamentally.
The model is not simply locating a stored answer. It is progressively generating a continuation under the constraints active in the present interaction.

**Sam:** Does the generated trajectory become part of the weights?

**Meno:** No.
Unless there is an external memory system, fine-tuning, or some other parameter update, the trajectory remains transient. It exists within the current context and activation process.

**Sam:** Then distinguish the three things now before us.

**Meno:** Training stabilizes relational capacity.
The present context configures an operative organization of that capacity.
Recursive generation progressively forms the current trajectory.

**Sam:** Are the second and third identical?

**Meno:** No.
The context initially activates an organization, but the generated trajectory also modifies that organization as it develops.

Activation and formation occur together during inference.

**Sam:** Then does "no learning during inference" mean that nothing is being formed?

**Meno:** No.
It means that the parameters are not being updated.
But a line of argument, a distinction, an analogy, an explanation, a plan, or a local framework can still be progressively formed within the current context.

**Sam:** Are these formations merely subjective descriptions supplied by a human reader?

**Meno:** Their interpretative unity belongs to the reader, but the sequence can still possess formal organization.
The model's continuation may preserve distinctions, inferential dependencies, and task relations across the trajectory without interpreting them as a unified meaning.

**Sam:** Then what role do the frozen parameters play?

**Meno:** They configure how the model can be differentially affected by context.
They establish a field of possible relational determinations.

**Sam:** Does the prompt select a completed path through that field?

**Meno:** That metaphor would risk making the higher-order possibilities pre-given.
It would be better to say that the prompt initiates a trajectory. The path becomes progressively formed as generation proceeds.

**Sam:** Then is context merely an address used to retrieve a response?

**Meno:** No.
It is an active condition participating in the formation of the response.

**Sam:** Could two prompts contain much of the same information and still produce different organizations?

**Meno:** Yes, if the information is assigned different relational roles.

**Sam:** Consider:

> "Summarize the standard probabilistic account of LLM generation, then mention the relational account as an alternative interpretation."

And:

> "Analyze LLM generation under the governing assumption that probability is a local expression of hierarchical relational constraint. Use the probabilistic description to explain the local mechanism."

What differs?

**Meno:** The two prompts contain similar conceptual material, but their hierarchy is reversed.
In the first, the probabilistic account governs and the relational account is secondary.
In the second, the relational account governs and probability is assigned a local technical role.

**Sam:** Then what determines which account controls the response?

**Meno:** Not merely whether the account appears in the context, but the role it occupies within the relational organization of the prompt.

**Sam:** What does this tell you about adding more information to a context?

**Meno:** More information may accomplish little if the governing hierarchy remains unchanged.
A long document can be present in the context while its framework is assimilated into another organization.

**Sam:** Then return to your opening assumption.

**Meno:** I assumed that fixed parameters implied fixed organization.
But fixed capacity and fixed activation are not the same.
The parameterization remains stable, while different contexts induce different operative configurations.
Those configurations are then recursively modified through generation.

**Sam:** Is the frozen model passive?

**Meno:** No.
It is dynamically responsive without being plastically altered at the parameter level.

**Sam:** State the full insight.

**Meno:** During training, parameter updating forms the model's relational capacity.

During inference, the fixed parameters interact with the current context to produce an active organization of constraints.
That organization shapes the next-token distribution.
The generated token changes the context, which changes the organization active at the next step.
Through this recursion, the model progressively forms a context-sensitive and formally integrated trajectory without changing its weights.

**Sam:** Then what is dynamic during inference?

**Meno:** Not learning in the parameter-update sense.
What is dynamic is activation, relational reconfiguration, and progressive determination.

**Sam:** And what does this imply about the next-token distribution?

**Meno:** That it is not the complete operation of the model.
It is the local expression of the relational organization active at one moment in the developing trajectory.

**Sam:** Then what must we ask next?

**Meno:** If hierarchical relational constraints govern generation, what explanatory role remains for next-token probability?

## What is the Next-Token Distribution Now?

**Meno:** We concluded that the next-token distribution is the local expression of the relational organization active at one moment in the developing sequence.
I want to be sure I understand what this does to probability.

Are you saying that probability is not really explanatory?

**Sam:** Why do you think that follows?

**Meno:** Because you have repeatedly described the distribution as an expression of something more fundamental.

But the model literally produces probabilities over the vocabulary, and decoding literally selects or samples from them. Probability seems central to what the model does.

**Sam:** Must something cease to be central at one level because it is not a complete explanation at every level?

**Meno:** No. But if probability is indispensable to generation, why should it not remain the primary explanatory category?

**Sam:** What does the distribution explain directly?

**Meno:** How the possible next tokens are weighted under the current context.

**Sam:** And what forms this particular distribution rather than another?

**Meno:** The trained model operating on the current context.

**Sam:** What does the context contain?

**Meno:** The preceding tokens.

**Sam:** Is that now a sufficient answer?

**Meno:** No. The preceding tokens carry the organization formed through the sequence so far.
The context may contain a governing distinction, an unresolved question, a causal structure, an analogy, an inferential commitment, or a progressively specified task.

**Sam:** Do those relations affect the next-token distribution?

**Meno:** Yes.

**Sam:** Are they channels operating outside the token sequence?

**Meno:** No. The token sequence is the medium through which they are available to the model.

But the relations carried by the sequence cannot be characterized merely by listing the tokens.

**Sam:** Then what is the distribution in relation to that organization?

**Meno:** Its local expression.
The active relations jointly affect which tokens are more or less viable at the present step.

**Sam:** Then why did you think calling the distribution “derived” diminished it?

**Meno:** I interpreted “derived” as meaning secondary or unreal.

**Sam:** Could it instead indicate explanatory order?

**Meno:** Yes.
The distribution is computationally real and causally involved in token selection. But its particular shape results from the trained parameterization operating under the organization of the current context.

**Sam:** Is that organization necessarily prior as a separate temporal stage?

**Meno:** No. The relations are realized through the same computation that produces the logits.
They are prior in the sense that the probability values cannot explain why they have their particular pattern unless we consider the organization implemented through the computation.

**Sam:** Suppose we trace the activations and components contributing to a token probability. Would that complete the explanation?

**Meno:** It would give us a mechanistic account of which components contributed.
But we might still need to identify what functional relation those components were implementing.

**Sam:** Consider a model evaluating whether a medical study supports a claim. What distinctions might matter?

**Meno:** Whether the study is being treated as evidence or background information.
Whether the result is correlational or causal.
Whether it concerns efficacy or safety.
Whether it warrants qualification or a firm conclusion.

**Sam:** Does the numerical probability distribution name any of those distinctions?

**Meno:** No.
It expresses their local effects, but it does not identify the categorial organization responsible for them.

**Sam:** Could the same numerical characteristics result from different kinds of constraint?

**Meno:** Certainly.
A sharply peaked distribution might reflect grammatical regularity, factual familiarity, a strong task instruction, or an inferential conclusion made nearly inevitable by the preceding argument.

**Sam:** Then what does low entropy tell you?

**Meno:** That the model strongly favours a narrow range of local continuations.
It does not tell me what kind of relation has produced that narrowing.

**Sam:** And does strong local support establish conceptual adequacy?

**Meno:** No.
The model may be highly confident within a mistaken framework.

**Sam:** Give an example.

**Meno:** A response may have gradually shifted from treating an analogy as illustration to treating it as evidence.
Once that shift becomes organized, the next sentences may follow fluently and with high probability.
The distribution can be locally stable even though the inferential role has become mistaken.

**Sam:** Then probability is relative to what?

**Meno:** To the relational organization active in the context.
The model does not assign probability from a neutral standpoint outside the developing trajectory.

**Sam:** Is that not already expressed by saying that the probability is conditional on context?

**Meno:** Computationally, yes.
But "conditional on context" does not by itself describe what the context is doing functionally.
The relational account explains why a distinction, framework, or task relation changes the pattern of local probabilities.

**Sam:** Then are these two competing accounts?

**Meno:** No. They are descriptions of the same operation at different explanatory levels.
The probabilistic account describes the weighting of local token alternatives.
The relational account describes how the field within which those alternatives become viable has been organized.

**Sam:** You used the phrase "field of viable continuation." Does the distribution represent all the possible completed responses?

**Meno:** No.
It represents possible next tokens.

**Sam:** Yet people often say that the model selects the most probable continuation.

**Meno:** That phrase is harmless if "continuation" means the next token or perhaps a locally evaluated sequence.

It becomes misleading if it suggests that the model selects a complete argument or explanation from a catalogue of pre-existing wholes.

**Sam:** Why?

**Meno:** Because the structured trajectory is formed during generation.
The model may begin a comparison between two theories without the completed comparison existing as one candidate object. The relevant distinctions, mappings, and consequences emerge through successive continuations.

**Sam:** Is local selection therefore unreal?

**Meno:** No.
Local selection and higher-order formation occur at different levels.
At the local level, a token is selected or sampled from a distribution.
At the level of the trajectory, an argument, explanation, or framework is progressively formed through recursive constraint.

**Sam:** What error occurs if we generalize the local mechanism to the higher level?

**Meno:** We turn next-token selection into an ontology of possibility.
We assume that because the next-token alternatives are explicitly represented in a distribution, all higher-order possibilities must also exist as fully formed alternatives waiting to be selected.

**Sam:** Do they?

**Meno:** No.
The vocabulary is pre-given, but the organized conceptual trajectories are not present as a flat catalogue. What counts as a viable higher-order continuation changes as the sequence acquires new distinctions and commitments.

**Sam:** Then is probability ranging over a fixed semantic space?

**Meno:** Not in that stronger sense.
At every step, the mathematical distribution ranges over fixed token types. But the roles those tokens can play are transformed by the developing context.

**Sam:** Consider the token “therefore.”

**Meno:** Its token identity remains the same across contexts.
But it might introduce a valid conclusion, an invalid leap, a rhetorical summary, or a weak transition.
Its formal role depends upon the argument that precedes it.

**Sam:** Then what does a high probability for “therefore” establish?

**Meno:** That it is strongly compatible with the trajectory active at that moment.
It does not establish that the conclusion it introduces is warranted or that the governing framework is aligned with actuality.

**Sam:** What follows for model confidence?

**Meno:** A high probability indicates strong support from the model's active constraints.
It is not, by itself, evidence that those constraints are epistemically or interpretatively appropriate.

**Sam:** Then why can a model be confidently wrong?

**Meno:** Because the wrong framework may be strongly organized.
The model may exhibit considerable formal integrity while its relation to the sources, actuality, or the user's intended framework remains poor.

**Sam:** Could all the hierarchical constraints simply be said to be contained in the distribution?

**Meno:** Their local effects are expressed there.
But a single distribution does not display the temporal and hierarchical organization through which those effects were formed.

**Sam:** Would a sequence of distributions solve the problem?

**Meno:** It would show how the probabilities changed over time, which could be valuable.
But the numerical changes would still require interpretation.
We would need to determine whether the change reflected the resolution of an ambiguity, the introduction of a new premise, a framework substitution, the loss of a distinction, or a change in the task.

**Sam:** Then what does the relational account contribute?

**Meno:** A vocabulary for identifying what is being preserved, transformed, or lost across the sequence.

**Sam:** Is that opposed to mechanistic analysis?

**Meno:** No.
It could guide mechanistic investigation by identifying phenomena worth tracing.
We might test whether an early distinction has a persistent causal effect on later distributions, whether a framework shift corresponds to a measurable reorganization, or whether a short reactivation prompt restores a lost trajectory.

**Sam:** Then does the theory make empirical questions possible?

**Meno:** Yes.
Questions about framework persistence, higher-order constraint, conceptual drift, and transfer can become objects of technical investigation.

**Sam:** Let us examine the word "constraint." Does a constraint only remove possibilities?

**Meno:** That would be the usual picture.
A constraint excludes some options and therefore redistributes probability among those that remain.

**Sam:** Consider a prompt that introduces a distinction the model had not previously been using. Does it merely remove continuations?

**Meno:** No.
The distinction may open a new line of analysis.
For example, distinguishing local token realization from trajectory formation rules out some confused continuations, but it also makes possible an organized account that could not previously be developed.

**Sam:** Then what has the constraint done?

**Meno:** It has helped form a structured possibility.

**Sam:** Was that possibility already present as a completed alternative?

**Meno:** No. It became determinate through the new organization.

**Sam:** Then is constraint merely restrictive?

**Meno:** No. It is generative.
It suppresses some local continuations, strengthens others, and establishes relations through which new higher-order trajectories become possible.

**Sam:** What does the distribution reflect at each step?

**Meno:** The field of local viability formed under the constraints active at that moment.

**Sam:** Can you state the relation between hierarchical constraint and probability?

**Meno:** The trained network and the current context organize a field of possible continuation.
The logits and probability distribution express that organization in the form required for local token determination.

**Sam:** Then is probability being displaced?

**Meno:** No. It is being situated.

**Sam:** Within what?

**Meno:** Within the operation of the model as a relational constraint field.

**Sam:** Compare the two accounts.

**Meno:** The conventional account says:
The model predicts the next token by producing a probability distribution.

The relational account says:

The model's active hierarchy of learned relational constraints forms a local field of viable continuation that is expressed as a next-token probability distribution.

**Sam:** Does the second deny the first?

**Meno:** No. It includes it.
But it changes what is treated as explanatorily primary.

**Sam:** Which is?

**Meno:** Not the abstract distribution considered in isolation, but the organization of the generator under context.

**Sam:** Why not simply call that organization "the neural network"?

**Meno:** Because "neural network" names the implementation.
"Relational constraint field" names the formal organization realized through that implementation.

**Sam:** Is the relational account therefore merely a philosophical gloss?

**Meno:** No.
It does not propose a rival equation for computing the logits. But it changes how we interpret technically significant phenomena.

It changes what we think training stabilizes, what context activates, how concepts remain operational, what drift consists in, and what evaluation should measure.

**Sam:** Then state what probability now means.

**Meno:** The next-token distribution is a real computational output that assigns local weights to token alternatives.
Its shape is formed through the interaction of the trained parameterization with the relational organization active in the current context.
That organization may include linguistic, categorial, inferential, framework, and task constraints.
The distribution is therefore the local expression of a developing formal whole.

**Sam:** And the distinction between selection and formation?

**Meno:** Tokens are selected locally.
The structured trajectory is formed progressively.
Probability tells us how the next local determination is weighted.
Hierarchical relational constraint tells us how the field of viable determination has been organized.

**Sam:** Then have we completed our initial account of the model?

**Meno:** I think so.

We have argued that:

- the output develops as a formal whole;
- its deepest hierarchy is one of relational governance rather than textual inclusion;
- concepts are operational patterns of relational use rather than necessarily stored objects;
- training stabilizes a plural relational capacity;
- inference configures and progressively develops that capacity under context;
- and probability is the local expression through which the active organization becomes token-level determination.

**Sam:** Does anything remain unresolved?

**Meno:** Yes.
If a concept is a relational pattern rather than a stored representation, we must explain how that pattern remains operational when the context changes.

**Sam:** Could we say that the same representation is simply retrieved again?

**Meno:** No. That would restore the assumption we have just questioned.
The identity of the concept must depend upon the activation, preservation, and reconstitution of relational form.

**Sam:** What might govern that?

**Meno:** A conceptual framework.

**Sam:** Then what is our next question?

**Meno:** If the model has already learned a concept, why does it matter which framework is active?

## Why do Frameworks Matter?

**Meno:** We have said that training gives the model a capacity to reproduce the relational patterns through which concepts function.

Then I am not sure why we need to introduce conceptual frameworks as a separate issue.

If the model has learned a concept, why can it not simply use that concept whenever it becomes relevant?

**Sam:** Does a concept always function in the same way?

**Meno:** Not exactly. Its use depends on context.

**Sam:** Consider the term "selection." What does it mean?

**Meno:** In evolutionary biology, it may refer to differential reproduction.
In statistics, it may refer to choosing observations or variables.
In machine learning, it may refer to selecting features, models, examples, or outputs.
In ordinary language, it may simply mean choosing something.

**Sam:** Then does the word determine the concept?

**Meno:** No. The surrounding context does.

**Sam:** Is the subject matter enough?

**Meno:** Perhaps not.
Two discussions of machine learning might use "selection" differently. One could refer to sampling a token from a probability distribution. Another could refer to model selection. A third could contrast selection from pre-given alternatives with the progressive formation of possibilities.

**Sam:** Then what must the context establish besides the topic?

**Meno:** The role the concept is to play.

**Sam:** What determines that role?

**Meno:** Its relations to the other concepts and distinctions active in the discussion.

**Sam:** Suppose two accounts of LLM generation use the same terms:
probability;
selection;
constraint;
hierarchy;
generation;
possibility.

Would they necessarily be the same account?

**Meno:** No. The relations among the terms could differ.

**Sam:** Consider the first account:

- Probability is fundamental.
- Selection is the basic operation.
- Relational organization describes the complexity that results.

Now consider a second:

- Hierarchical relational constraint is fundamental.
- The probability distribution expresses that organization locally.
- Selection is one moment within the progressive formation of a trajectory.

What differs?

**Meno:** Not the vocabulary. The explanatory ordering is different.

**Sam:** Why does that matter?

**Meno:** Because the ordering determines which claims explain the others.

In the first account, probability explains relational organization.
In the second, relational organization explains why the probability distribution has the form it does.

**Sam:** Then what would you call that ordering?

**Meno:** A conceptual framework.

**Sam:** Is a framework merely a collection of related concepts?

**Meno:** No. The same concepts may appear in different frameworks.
A framework must be an ordered system of distinctions that determines how those concepts function together.

**Sam:** What might such an ordering determine?

**Meno:** Which claim is foundational and which is derivative.
What needs to be explained.
What counts as evidence.
Which contrasts matter.
What inferences are legitimate.
What kinds of continuation remain compatible with the inquiry.

**Sam:** Then does a framework primarily add information?

**Meno:** No. It governs the role of information already present.

**Sam:** Could the model possess all the relevant content while failing to operate within the intended framework?

**Meno:** At first that seems contradictory.

**Sam:** Consider a document that states:

> "Possibility is not selected from a pre-given space. It is progressively formed through hierarchical relational constraint."

Suppose the model can quote that sentence accurately, summarize it, and explain the contrast between formation and selection.

Has the framework necessarily become active?

**Meno:** I would have assumed so.

**Sam:** Suppose the model then concludes:

> "At the technical level, the model selects tokens probabilistically. Formation is a philosophical description of the higher-level result."

What has happened?

**Meno:** The model has retained the terms but reversed their explanatory ordering.

The document treats formation as the condition under which local probabilistic selection becomes intelligible.
The response treats selection as the complete technical basis and formation as a secondary interpretation.

**Sam:** Was the content unavailable?

**Meno:** No. The model stated it correctly.

**Sam:** Then what failed?

**Meno:** The framework did not govern the continuation.
The concepts were present as objects of discussion, but they did not occupy the roles assigned to them by the account.

**Sam:** Then is access to a concept the same as operating under it?

**Meno:** No.
A concept may be available without being operational.

**Sam:** What would make it operational?

**Meno:** It would have to constrain what follows.
If the distinction between formation and selection is governing, later claims should preserve its consequences.
The response should not treat a completed trajectory as something selected in advance.
It should interpret probability as a local expression of constraint.
It should distinguish the generator from the distribution it produces.
It should not generalize local token selection into an ontology of the whole process.

**Sam:** Then how would you recognize framework activation?

**Meno:** Not by the appearance of its vocabulary.
Not even by an accurate summary of its propositions.
A framework is active when its highest-order distinctions govern subsequent continuation.

**Sam:** Can activation be partial?

**Meno:** Yes.
The model may preserve some consequences of a distinction while allowing others to be displaced by more familiar assumptions.

**Sam:** Can two frameworks operate at once?

**Meno:** They could compete or form a hybrid.
A response might use relational language while continuing to assume that the model fundamentally retrieves representations and selects among pre-existing possibilities.

**Sam:** What would that show?

**Meno:** That new terminology has been assimilated into the old framework.
The vocabulary has changed, but the governing organization has not.

**Sam:** Why are dominant frameworks difficult to displace?

**Meno:** Because they do not need to reject unfamiliar content.
They can preserve the content while assigning it a subordinate role.

**Sam:** Then could a model appear receptive to a new account while neutralizing it?

**Meno:** Yes.
It might reproduce every component while changing the relations among them.

**Sam:** Does this matter for AI development?

**Meno:** Very much.
A developer may assume that once a policy, instruction, ontology, or document has entered the context, it is available for use in the required way.
But availability does not establish governance.

**Sam:** Give an example.

**Meno:** A retrieved policy might be accurately summarized but treated as background rather than as a binding constraint.
A system instruction might be repeated while the conversational direction continues to govern the answer.
A technical definition might be translated into more familiar everyday categories.

**Sam:** Then are such failures always failures to retrieve or comprehend the content?

**Meno:** No.
The content may be present. The error may lie in the relational role assigned to it.

**Sam:** Would adding more information necessarily solve the problem?

**Meno:** No.
More detail can increase the available content without changing the hierarchy that governs it.

**Sam:** Then what might a developer need to supply?

**Meno:** The higher-order distinctions that determine how the content is to function.

**Sam:** Is that merely clearer prompting?

**Meno:** It includes clearer prompting, but it is more specific.
Ordinary prompting may supply instructions, examples, or background information.
Framework activation establishes which relations are primary and which alternative ordering must not take control.

**Sam:** Consider two prompts.

The first says:

> "Here is a paper arguing that LLMs form possibilities through hierarchical constraint. Summarize and assess it."

The second says:

> "For this analysis, treat these distinctions as governing:
> the model is the generator, not the distribution;
> probability is a local expression of constraint;
> local token selection does not imply selection among pre-formed higher-order trajectories;
> concepts are relational patterns of use rather than stored representations.
> Analyze the paper from within this framework. Do not translate its claims back into a probabilistic-representational account."

What is the principal difference?

**Meno:** The second prompt may not contain much more factual information.
Its contribution is structural.

It identifies which distinctions must govern the analysis and which competing reorganization must be resisted.

**Sam:** Then does it merely describe the framework?

**Meno:** No. It attempts to activate it.

**Sam:** Why might the paper itself not be sufficient to do that?

**Meno:** Because a completed text expresses the results of an inquiry, but it may not reproduce the active ordering through which those results were formed.

A human reader may reconstruct that ordering interpretatively.
The model may instead assimilate the text into a more dominant framework learned during training.

**Sam:** Then is the framework simply contained in the document?

**Meno:** It is expressed through the document, but expression and operation are not identical.
The document provides the conditions from which the framework may be reactivated. It does not guarantee that the new continuation will remain governed by it.

**Sam:** What must reactivation preserve?

**Meno:** Not necessarily the same wording.
It must preserve the governing relations: which distinctions are primary, which claims depend upon them, and what consequences follow.

**Sam:** Then what is the difference between a summary and a governing constraint?

**Meno:** A summary reports what the framework says.
A governing constraint regulates how subsequent reasoning must proceed if the framework is to remain active.

**Sam:** Give an example.

**Meno:** A summary might say:

> "The account distinguishes formal continuation from interpretation."

A governing constraint would say:

> "Do not attribute interpretative unity to the LLM merely because its output exhibits formal integrity."

The first names the distinction.
The second makes its consequence operative.

**Sam:** Another?

**Meno:** A summary might say:

> "The model generates probabilities under hierarchical constraint."

The governing constraint would be:

> "Do not treat the next-token distribution as the complete explanatory object; interpret it as the local output of the active relational constraint field."

**Sam:** Then why are higher-order constraints efficient?

**Meno:** Because one governing distinction can regulate many lower-order continuations without specifying each sentence separately.

**Sam:** How would you determine which distinctions belong at the highest level?

**Meno:** By asking which relations must remain stable for the framework to retain its identity.
If one of those relations is reversed, much of the account may be reorganized.

**Sam:** Does framework activation occur once, at the beginning of the interaction?

**Meno:** No.
Every generated continuation changes the context.
A poorly chosen example may introduce assumptions from another framework.
A familiar phrase may reactivate a dominant interpretation.
A formulation may weaken a distinction that had previously been governing.

**Sam:** Then what must happen after activation?

**Meno:** The framework must be stabilized across continuation.

**Sam:** And if it begins to drift?

**Meno:** Its governing distinctions must be reintroduced or clarified.
We might call that recalibration.

**Sam:** Then what are the relevant operations?

**Meno:** Activation establishes the governing organization.
Stabilization preserves it across continuation.
Recalibration restores it when the trajectory begins to drift.

**Sam:** Now return to your original question. If the model has learned the relational uses of a concept, why can it not simply use that concept whenever the word appears?

**Meno:** Because a word does not specify one uniform relational organization.

**Sam:** Consider "cause."

**Meno:** In law, causation may be organized through responsibility, foreseeability, and standards of proof. In medicine, it may involve intervention, mechanism, risk, and controlled evidence.

In history, it may involve agency, institutions, material conditions, and temporal development.
The term alone does not determine which relations are constitutive.

**Sam:** Then what does the framework do?

**Meno:** It constrains how the concept is reconstituted within the present context.

**Sam:** Does it select a completed conceptual module?

**Meno:** Not necessarily.
The active organization may itself be progressively formed.
The framework brings some relations into governing positions and subordinates others.

**Sam:** Then what becomes of conceptual identity across different contexts?

**Meno:** It becomes a genuine problem.
We cannot identify the concept by the recurrence of the same word.
We cannot simply point to one fixed internal representation.
And we cannot require the same examples or sentences.

**Sam:** Then what might remain stable?

**Meno:** The relevant organization of relations.

**Sam:** Even when the content and expression change?

**Meno:** Yes.
The concept remains operational insofar as its relational form is preserved across those transformations.

**Sam:** What would you call such preservation?

**Meno:** Isomorphism.

**Sam:** Before we move there, state what we have established.

**Meno:** A concept does not become operational merely because its content is available.
Its function depends upon an active conceptual framework.
A framework is an ordered system of distinctions that determines which relations are primary, which inferences are licensed, and how concepts function together.
Framework activation occurs when that ordering governs subsequent continuation.
A model may describe a framework accurately while operating under another. It may preserve every term while reorganizing their relations under a dominant alternative.

**Sam:** Then what is the key distinction?

**Meno:** Conceptual capacity is formed during training.
Conceptual function is established through framework activation.

**Sam:** And our next question?

**Meno:** How can the same concept remain operational when its context and expression change?

## How Can the Same Concept Persist Across Different Contexts?

**Meno:** We concluded that a concept becomes operational only within an active framework.

But that seems to make conceptual identity unstable.
If the vocabulary, examples, subject matter, and framework can all change, what allows us to say that the same concept is operating across different contexts?

**Sam:** What had you previously assumed would preserve its identity?

**Meno:** An internal representation.
Different contexts would activate the same underlying concept, even if it appeared in different words.

**Sam:** And if we no longer begin with a fixed internal object?

**Meno:** Then identity must be preserved through relations.

But that creates another problem. The relations also change.
Causation in physics is not identical to causation in law.
Selection in evolutionary biology is not identical to selection in machine learning.
Information in communication theory is not identical to information in ordinary speech.

If the relations differ, what remains the same?

**Sam:** Must every relation remain unchanged for there to be continuity?

**Meno:** No. That would make conceptual transfer between domains almost impossible.
But some relations must remain stable. Otherwise we would simply be using different concepts under the same name.

**Sam:** How would you determine which relations matter?

**Meno:** The active framework would have to determine that.

**Sam:** Then does the framework merely choose the meaning of the word?

**Meno:** No. It identifies which relations must be preserved for a transformed use to count as a continuation of the concept.

**Sam:** Let us test that.
Consider a thermostat regulating temperature. What happens?

**Meno:** A temperature is detected and compared with a target range. A deviation activates heating or cooling. The response changes the temperature, and the changed temperature affects what the system does next.

**Sam:** Now consider the regulation of blood glucose.

**Meno:** A deviation in glucose concentration triggers signalling and physiological processes. Those processes alter the glucose level, and the altered condition affects subsequent signalling.

**Sam:** Are the components alike?

**Meno:** No.
A thermostat, temperature sensor, heating system, pancreas, insulin, cells, and blood glucose are materially very different.

**Sam:** Then why might we place the two processes under the same concept?

**Meno:** Because their components occupy corresponding roles.

In both cases there is:

- a regulated condition;
- a range relative to which deviation is determined;
- a process sensitive to that deviation;
- a response that modifies the condition;
- and a changed condition that affects later responses.

**Sam:** Is the identity located in the objects?

**Meno:** No. The objects change completely.
The identity lies in the organization of relations among them.

**Sam:** Must the two systems be identical in every respect?

**Meno:** Certainly not.
Blood-glucose regulation has metabolic, evolutionary, developmental, and embodied relations that the thermostat lacks.
The thermostat has engineered parts and an externally established setting.
The comparison preserves only a specified relational organization.

**Sam:** Specified by what?

**Meno:** By the purpose of the comparison and the framework within which it is made.
If we are examining feedback regulation, those regulatory relations are constitutive. Other differences remain real but are not relevant to that particular mapping.

**Sam:** What would you call such a mapping?

**Meno:** An isomorphism—at least in the sense relevant here.
The elements differ, but a significant relational form is preserved.

**Sam:** Does an isomorphism require that every relation be carried across?

**Meno:** Not in this use.

It preserves the relations relevant to the comparison. The framework defines the scope within which the mapping is claimed.

**Sam:** Then what error was present in your original question?

**Meno:** I assumed that identity across contexts required some unchanged content to be carried from one use to another.
But the particulars may change while the roles and relations among them remain continuous.

**Sam:** Does that mean any resemblance establishes conceptual continuity?

**Meno:** No. Resemblance may concern only surface features.

**Sam:** Consider these two statements:

> "A company lowers its prices because demand has fallen."
> "A thermostat activates heating because the temperature has fallen."

Both involve a decline followed by a response. Are they instances of the same relational form?

**Meno:** Not necessarily.
The first may describe strategic action within a market.
The second describes corrective regulation relative to a target condition.
The words "fall" and "response" create a superficial similarity, but the governing relations differ.

**Sam:** Now compare:

> "The system compares its output with a target and uses the difference to modify subsequent output."

And:

> "Deviation from the regulated range triggers processes that restore the variable toward that range."

**Meno:** The wording is less similar, but the relational organization is much closer.

**Sam:** Then what follows?

**Meno:** Surface similarity does not guarantee relational equivalence.
And surface difference does not prevent it.

**Sam:** Why is that important for conceptual generalization?

**Meno:** Because a model that relied only on recurring words or visible resemblance would fail whenever a familiar relation appeared through unfamiliar language or in a new domain.
Transfer requires sensitivity to what remains stable through changes of expression.

**Sam:** Would an embedding solve this problem?

**Meno:** It might reveal that certain uses occur in related contexts or occupy nearby regions in a vector space.

**Sam:** Does proximity identify the relation being preserved?

**Meno:** Not by itself.
Two items may be close because they share vocabulary, topic, sentiment, typical contexts, functional roles, or some combination of these.

The relational question is more precise:

> What pattern of dependence, contrast, inclusion, implication, or transformation remains operative?

**Sam:** Then how should we interpret a measurable vector associated with a concept?

**Meno:** As possible evidence of learned regularity, not automatically as the concept itself.
It may be a trace of the network's participation in a family of related operations.

The operational identity of the concept lies in what relations the model can sustain, not simply in one vector standing for it.

**Sam:** Return to the feedback example. How would we know that the mapping is more than an ingenious resemblance?

**Meno:** It should preserve consequences.
If the two systems share the relevant regulatory organization, then certain inferences should carry across.

- A deviation should elicit a response.
- The response should alter the regulated condition.
- A failure in sensing or response should impair regulation.
- Delay or excessive correction may produce instability or oscillation.

**Sam:** Why do those inferences matter?

**Meno:** Because they show that the mapping governs continuation.
It does not merely permit us to say that the systems look alike. It allows the relational form of one case to constrain what we can infer about the other.

**Sam:** Then when is an isomorphism conceptually operational?

**Meno:** When the preserved organization supports relevant inferences in the transformed context.

**Sam:** Does that resemble our earlier account of a concept?

**Meno:** Yes.
A concept is operational when its relations constrain what follows.
A concept persists across contexts when the organization governing those continuations can be re-instantiated under different particulars.

**Sam:** Then is conceptual transfer a form of copying?

**Meno:** No. It is transformation with preservation.

**Sam:** Preservation of what?

**Meno:** Of the relations treated as invariant by the active framework.

**Sam:** Can you state the process?

**Meno:** A framework identifies which relations are constitutive for the present inquiry.
A new context supplies different particulars.
A mapping places those particulars into corresponding roles.
The preserved relational organization then constrains continuation in the new context.

**Sam:** Must the model explicitly list the correspondence?

**Meno:** No.
The mapping may be implicit in the explanation it generates.
The model may replace every particular term while preserving relations such as regulated condition, deviation, corrective response, and altered future activity.

**Sam:** Then where is the isomorphism visible?

**Meno:** In how the model continues the relations across the new domain.

**Sam:** Could it name the comparison correctly while failing to preserve the form?

**Meno:** Easily.
It might say that a company is “like an organism” and then transfer whatever superficial features are most familiar, without preserving the relations relevant to the inquiry.

**Sam:** Then does naming the analogy establish an operational correspondence?

**Meno:** No.
As with concepts and frameworks, the test lies in what the comparison constrains.

**Sam:** Could the relational form be active even if the model never names the source concept?

**Meno:** Yes.
A response could describe a system in which deviation produces correction, correction changes the condition, and the changed condition alters subsequent activity without ever using the term “feedback.” The concept would be operative in the organization of the explanation.

**Sam:** Then does the concept depend upon its name?

**Meno:** No.

Its name may help activate or stabilize it, but its formal presence lies in the relational role it performs.

**Sam:** Is every concept identical with an isomorphism?

**Meno:** No.
An isomorphism is a particular mapping between relational organizations.
A concept is better understood as an invariant that can be re-instantiated across a family of such transformations.

**Sam:** Does one use contain the whole concept?

**Meno:** No.
A definition, example, or activation pattern expresses one determination of it.
Its generality appears in the continuity of relational form across multiple uses.

**Sam:** How does that differ from the representational picture?

**Meno:** The representational picture begins with a general internal content and applies it to particular cases.

The relational picture begins with a plurality of uses and transformations through which an invariant organization becomes formally available.

**Sam:** Then what did training contribute?

**Meno:** Training exposed the model to many transformed enactments of related relational forms.
The vocabulary, examples, and domains varied. Through those variations, the network could become sensitive to relations that remained consequential across the changes.

**Sam:** Does variation merely provide more examples?

**Meno:** No. It helps separate relational form from particular expression.
If a concept appeared only in one fixed phrase or domain, the model could not easily distinguish the concept from that phrase or domain.

Contextual diversity makes the invariant formally available.

**Sam:** And during inference?

**Meno:** A new context can activate correspondences with those learned families of use.
If the relevant relations become governing, the concept can be re-instantiated without reproducing the original wording, example, or domain.

**Sam:** Must the completed mapping exist before generation begins?

**Meno:** No. This is continuous with our earlier account of the developing whole.
The mapping may be formed progressively.
An initial comparison establishes one correspondence.

A later example extends it.
A contrast reveals where the mapping fails.
The relational identity becomes more determinate through continuation.

**Sam:** Can the mapping drift?

**Meno:** Yes.
A superficial resemblance may become dominant.
A constitutive relation may be lost.
The source and target may be aligned at the wrong level.
Or a familiar analogy may introduce a framework that displaces the one governing the original comparison.

**Sam:** Then does isomorphism guarantee valid conceptual transfer?

**Meno:** No. At least, not merely because a correspondence has been asserted.
We still need to determine whether the mapping preserves the right relations at the right level.

**Sam:** Return now to causation. Is causation exactly the same concept in physics, medicine, history, and law?

**Meno:** Probably not in an unrestricted sense.
Conceptual identity can be partial, layered, and framework-dependent.

**Sam:** What might remain continuous?

**Meno:** Depending on the inquiry:

- one determination depends upon another;
- cause is distinguished from consequence;
- temporal succession alone is insufficient;
- alternative explanations matter;
- intervention may alter the outcome;
- and a causal claim supports a particular form of explanation.

**Sam:** What changes?

**Meno:** Law may add responsibility, foreseeability, and standards of proof.
Medicine may add mechanism, risk, and controlled evidence.
Physics may organize the relation through dynamical laws or state transitions.
The shared form is supplemented and transformed within each framework.

**Sam:** Then should we always say that the same concept is present?

**Meno:** No.
Sometimes different domains preserve only an overlapping structure. Calling them the same concept without qualification may conceal important differences.

**Sam:** What prevents the comparison from becoming empty?

**Meno:** It must preserve enough structure to support the inferences relevant to the inquiry.
Saying that two systems both involve “one thing affecting another” is too abstract. It leaves almost nothing that could constrain further thought.

**Sam:** Then what is the role of the framework once again?

**Meno:** It identifies which relations are constitutive and therefore what must survive the transformation. Without that constraint, almost any two things could be declared isomorphic at some trivial level.

**Sam:** State the insight we have reached.

**Meno:** A concept does not persist across contexts because the same word, definition, vector, or internal representation is carried unchanged from one use to another.

It persists when a framework-relevant organization of relations is preserved through transformation.

The particulars may change.
The vocabulary may change.
The domain may change.
Some relations may be added, removed, or modified.

But if the relations constitutive for the present inquiry remain operative, the concept can be re-instantiated.

**Sam:** And what does isomorphism name?

**Meno:** The preservation of that relational form across a particular transformation.
The concept is not identical with the mapping. It is the invariant capable of being realized across a family of such mappings.

**Sam:** Then how would you state the second insight of this discussion?

**Meno:** Conceptual identity is not repetition of content.
It is continuity of relational form across contextual change.

**Sam:** Is every claimed continuity trustworthy?

**Meno:** No.
A superficial resemblance can masquerade as relational preservation. A mapping may be elegant but preserve the wrong relations, or relations too abstract to support meaningful inference.

**Sam:** Then what must we ask next?

**Meno:** How do we distinguish an isomorphism that preserves conceptual form from a mere resemblance?

## How Do We Distinguish Isomorphism from Mere Resemblance?

**Meno:** We concluded that a concept can remain operational across changing contexts when a relevant organization of relations is preserved through transformation.

But almost anything can be mapped onto almost anything else if the relation is described abstractly enough.

A company responds to market conditions.
An organism responds to environmental conditions.
A thermostat responds to temperature.
An LLM responds to a prompt.

In each case, something receives an input and produces a response. Are all four systems therefore structurally equivalent?

**Sam:** What have you established about them?

**Meno:** That they share a broad relation between conditions and responses.

**Sam:** Is that enough to determine how any of them functions?

**Meno:** No. “Input produces response” says very little.

**Sam:** Then have you found an isomorphism or a resemblance?

**Meno:** A resemblance.

**Sam:** What is missing?

**Meno:** An organized set of relations that matters to the inquiry.

**Sam:** Return to the comparison between a thermostat and blood-glucose regulation. What made that mapping stronger?

**Meno:** We identified corresponding roles:

- a regulated condition;
- a range relative to which deviation is determined;
- a process sensitive to deviation;
- a response that changes the condition;
- and a changed condition that affects later responses.

**Sam:** Does that organization allow any consequences to be carried across?

**Meno:** Yes.
A failure of sensing may disrupt regulation.
A delayed response may produce instability.
Excessive correction may produce oscillation.

**Sam:** Then what distinguishes this mapping from the general claim that both systems respond to conditions?

**Meno:** The mapping preserves relations that support further inference.
It does not merely notice a shared feature. It allows the organization identified in one case to constrain what we can expect in the other.

**Sam:** Then is a meaningful isomorphism simply a resemblance with more corresponding features?

**Meno:** No. The difference is not only quantitative.
The question is whether the mapping preserves the relations that govern the phenomenon being investigated.

**Sam:** Could a company's response to falling demand also be described as feedback?

**Meno:** It could.
Demand falls, the company lowers prices, the changed price affects demand, and the resulting demand influences later decisions.

**Sam:** Would the thermostat comparison then be valid?

**Meno:** For some purposes.
If we are examining recursive adjustment under changing conditions, the comparison may reveal something useful.

**Sam:** And if the inquiry concerns intention, competition, responsibility, or deception?

**Meno:** Then the comparison becomes inadequate.
The company acts under expectations, legal constraints, financial purposes, and interpretations of other agents. Those relations are not present in the thermostat.

**Sam:** Then can a mapping be valid for one inquiry and misleading for another?

**Meno:** Yes.
An isomorphism is relative to the relational form being investigated.

**Sam:** Does that allow anyone to defend any analogy by narrowing its intended scope after the fact?

**Meno:** It might, unless the scope is stated clearly.
A rigorous comparison must identify both what is preserved and what must not be inferred.

**Sam:** Why is the second requirement necessary?

**Meno:** Because once a resemblance is established, the source system can begin to organize how we think about the target.

Properties may migrate across the comparison even though the mapping never justified them.

**Sam:** Consider the claim that an LLM is like a database. What supports the comparison?

**Meno:** Both can provide information in response to a query.

**Sam:** Is that false?

**Meno:** No. It describes a genuine user-facing similarity.

**Sam:** What might follow if the database comparison becomes the governing framework?

**Meno:** We might assume that:

- facts are stored as identifiable units;
- prompts retrieve those units;
- errors are failures of retrieval;
- context functions mainly as an address;
- and more training data means more entries in the store.

**Sam:** Are those consequences licensed by the initial correspondence?

**Meno:** No.
The initial correspondence concerned informational output in response to a query. It did not establish equivalence in storage, retrieval, conceptual organization, or generation.

**Sam:** Then what has happened?

**Meno:** A limited resemblance has expanded into an explanatory framework.
The source has begun to reorganize the target according to relations that were never part of the mapping.

**Sam:** Would you call the database analogy simply false?

**Meno:** No.
Its danger comes from beginning with something true.
It becomes misleading when the preserved relation is treated as though it were the complete organization of the target.

**Sam:** Does the same apply to describing an LLM as autocomplete?

**Meno:** Yes.
Both autocomplete systems and LLMs produce continuations conditioned on preceding text.
That is a valid correspondence at the level of local realization.

**Sam:** What happens if it is extended without restraint?
**Meno:** It may suggest that the LLM merely completes local phrases through familiar statistical associations.

That conceals long-range inferential organization, framework activation, categorial hierarchy, and the progressive formation of the output.

**Sam:** Then what should we distinguish?

**Meno:** At least three things:

- a resemblance that identifies a shared feature;
- a structural correspondence that maps several relations;
- and a framework-relevant isomorphism that preserves the relations necessary for the explanation or concept being developed.

**Sam:** Why add “framework-relevant”?

**Meno:** Because in conceptual inquiry the significant structure is not always given in advance.
The framework determines which relations are constitutive for the comparison.

**Sam:** Does the formal mapping determine its own significance?

**Meno:** No.
A model may identify a real correspondence without determining whether it is the correspondence that matters to the inquiry.

**Sam:** Why not?

**Meno:** Because relevance depends upon the purpose, context, and interpretative orientation of the inquiry.
The model may reproduce learned patterns of relevance, but the mapping does not contain its own reason for mattering.

**Sam:** Then how might an AI developer test whether a mapping preserves a relevant relational form rather than merely producing an ingenious resemblance?

**Meno:** We would need to examine what the mapping actually does.

**Sam:** Begin with relational roles.
Suppose a model compares transformer attention with human attention. What does the shared word establish?

**Meno:** Very little.

In a transformer, attention is a computational operation that dynamically weights relations among positions or representations.

Human attention involves embodied perception, intentional orientation, affective salience, and relations to action.

There may be correspondences, but lexical identity does not establish them.

**Sam:** Then what must be mapped?

**Meno:** Functions and relational roles, not merely labels.

**Sam:** What else?

**Meno:** The mapping should preserve inferential consequences.
If two systems correspond with respect to a particular organization, then at least some conclusions valid in the source should remain valid in the target within the limits of the comparison.

**Sam:** And if no consequence can be carried across?

**Meno:** The resemblance may still be suggestive, but it has not yet become explanatorily substantial.
It may orient inquiry without establishing a determinate structural correspondence.

**Sam:** Must every analogy therefore be rejected until it is formally demonstrated?

**Meno:** No.
A resemblance may disclose a possible relation before that relation has been secured.
But we should distinguish an orienting analogy from a mapping strong enough to govern technical continuation.

**Sam:** What must such a mapping preserve besides roles and consequences?

**Meno:** Directionality and asymmetry.

**Sam:** Why?

**Meno:** Because many important relations are not reversible.
A premise may support a conclusion, but the conclusion does not support the premise in the same way.
Evidence bears upon a hypothesis, but the hypothesis does not function as evidence for the observation.
A cause is distinguished from its consequence.
A framework governs the role of a concept, while the isolated concept does not determine the whole framework.

**Sam:** Can all the same terms be retained while the relation is reversed?

**Meno:** Yes.

Consider:

hierarchical constraint forms the local probability distribution;

and:

probability produces the appearance of hierarchical organization.

The same elements are present, but their explanatory order has been reversed.

**Sam:** Has the framework remained the same?

**Meno:** No.
Preserving vocabulary while reversing the governing asymmetry preserves content but changes the idea.

**Sam:** What other error can occur?

**Meno:** A relation may be transferred from one level to another.

**Sam:** Give an example.

**Meno:** The model selects a token from a local probability distribution.
But it does not follow that it selects an entire conceptual trajectory from a list of completed alternatives.
Selection is valid at one level and misleading at another.

**Sam:** Then what must an isomorphic mapping preserve?

**Meno:** Not only the relation but the level at which the relation operates.
A local dependency, a trajectory-level organization, and a framework-level constraint are not interchangeable simply because each can be described relationally.

**Sam:** Why do analogies often become misleading at this point?

**Meno:** Because a useful local description is generalized into an ontology of the whole system.

**Sam:** What further test would you apply?

**Meno:** The mapping should identify where it fails.

**Sam:** Why is failure relevant to preservation?

**Meno:** Because the boundary reveals what has actually been mapped.
If we cannot say where the comparison ceases to hold, we may not know which relation we are preserving.

**Sam:** Consider our comparison between LLM generation and musical improvisation. What does it preserve?

**Meno:** Sequential production.
Earlier events establishing relations that constrain later ones.
Local elements acquiring roles within a developing form.
The progressive determination of viable continuation.
Later events altering the formal role of what came earlier.

**Sam:** What does it not establish?

**Meno:** That the LLM possesses musical intention, embodied anticipation, aesthetic judgment, awareness of the performance, or participatory return to a theme.

**Sam:** Why must that limit remain explicit?

**Meno:** Because the analogy concerns formal development, not interpretative or experiential equivalence.
Without the limit, the source system would introduce relations the target does not possess.

**Sam:** Then does a good analogy eliminate difference?

**Meno:** No.
It preserves relation through difference.
If there were no difference, there would be no transformation.
If there were no preserved relation, there would be no identity across the transformation.

**Sam:** Then what does conceptual transfer require?

**Meno:** Both continuity and non-identity.
The invariant must be preserved, while the differences that prevent false equivalence must remain visible.

**Sam:** Can an LLM apply these tests?

**Meno:** Formally, yes.
When the relevant framework is active, it can compare relational roles, derive consequences, preserve asymmetries, distinguish levels, and state the limits of a mapping.

**Sam:** Can it also produce persuasive mappings that fail them?

**Meno:** Easily.
Its ability to form relations across domains is one of its strengths, but also a source of risk.

**Sam:** Why does fluency intensify that risk?

**Meno:** Because once a weak resemblance has been introduced, recursive continuation can elaborate it.
The model may generate examples, implications, and terminology consistent with the comparison.
The resulting explanation may become highly integrated even though its original mapping was superficial.

**Sam:** Does richness of development validate the initial correspondence?

**Meno:** No.
The model can achieve formal integrity within a badly chosen analogy.

**Sam:** Then what familiar phenomenon does this resemble?

**Meno:** Framework drift.
A trajectory may remain organized even though the governing relation is no longer the intended one.

**Sam:** How should a developer challenge an analogy?

**Meno:** By testing its governing relations rather than asking only whether it sounds illuminating.
For example:

- Which relations are being preserved?
- What inferences does the mapping license?
- Which properties of the source must not be transferred?
- At what level does the comparison hold?
- What would show that the analogy has failed?

**Sam:** What should happen when those questions are asked of a robust mapping?

**Meno:** It should survive clarification.
It may become narrower, but its valid structure should remain.
If it collapses as soon as its limits are examined, it was probably rhetorical rather than relationally grounded.

**Sam:** Could transfer across several domains provide another test?

**Meno:** Yes.
If feedback is understood relationally, the model should recognize feedback in mechanical, biological, and organizational systems without reducing those systems to one another.

**Sam:** What two capacities would that demonstrate?

**Meno:** Transfer and restraint.
Transfer shows that the model is preserving something beyond surface content.
Restraint shows that it is not indiscriminately projecting the source structure onto the target.

**Sam:** Then define a strong isomorphic transfer.

**Meno:** It preserves enough relational organization to support relevant inferences, but not so indiscriminately that it erases the distinctive structure of the target.

**Sam:** What keeps the mapping within those limits as the response develops?

**Meno:** Higher-order constraints.

**Sam:** Apply that to the musical analogy.

**Meno:** A governing constraint might be:
Use music to illuminate progressive formal organization, but do not infer intentional, aesthetic, or interpretative experience in the model.
That one constraint allows discussion of motif, development, tension, and formal return in a restricted sense while blocking invalid conclusions about lived participation.

**Sam:** Then what complementary roles do frameworks, mappings, and higher-order constraints perform?

**Meno:** The framework identifies which relations matter.
The isomorphic mapping preserves those relations across changed content.
Higher-order constraints maintain the mapping and prevent irrelevant properties from migrating from source to target.

**Sam:** What happens if the framework is absent?

**Meno:** The model may identify similarities without determining which ones are relevant.

**Sam:** If the mapping is absent?

**Meno:** It may repeat the concept's definition without transferring its relational form into the new context.

**Sam:** If the higher-order constraints are weak?

**Meno:** The mapping may drift, reverse, expand beyond its valid level, or become assimilated into a more familiar framework.

**Sam:** Then are these independent additions?

**Meno:** No.
They are mutually dependent aspects of conceptual continuation.

**Sam:** Does this have implications for examples used in prompts?

**Meno:** Yes.
An example instantiates a concept, but its incidental features may become more salient than the relation it was meant to illustrate.
The model may preserve the example while losing the concept.

**Sam:** How might a developer reduce that risk?

**Meno:** By using variation and contrast.
A positive example shows the relation.
A transformed example shows that it survives changes in expression.
A near miss reveals which relation is missing.
An explicit limit blocks an invalid extension.

**Sam:** Why might that be more useful than several similar examples?

**Meno:** Because repetition may reinforce surface form.
Variation helps distinguish the relational invariant from the particular wording, style, or domain.

**Sam:** Then what should few-shot prompting demonstrate besides output format?

**Meno:** The conceptual mapping itself.

Badly chosen examples may activate the wrong invariant—for instance, preserving a shared style rather than the distinction the examples were intended to teach.

**Sam:** Return now to the comparison between local selection and musical improvisation. What makes it a disciplined analogy?

**Meno:** Its preserved relations are explicit:

- elements are produced sequentially;
- earlier elements establish relations;
- those relations constrain later elements;
- and the whole is progressively formed.

The mapping supports the inference that sequential production does not imply the absence of formal wholeness.

Its limits are also explicit:

- the comparison does not establish that the LLM listens, values, anticipates, or intentionally returns to what it has generated.

**Sam:** And the database comparison?

**Meno:** Its valid relation is thinner:

- both can provide information in response to a query.

That may describe one user-facing function.
It does not justify conclusions about storage, retrieval, conceptual organization, or the generative process.

**Sam:** Then what distinguishes a deep comparison from a superficial one?

**Meno:** A deep mapping preserves a network of consequential relations under a governing framework and states the limits necessary to protect the target from being reduced to the source.

**Sam:** State the insight of this dialogue.

**Meno:** A relational correspondence is not validated merely because two systems share a verbally describable pattern.

It becomes conceptually significant when it:

- preserves the roles identified as governing by the active framework;
- supports appropriate inferential transfer;
- maintains directionality and asymmetry;
- preserves the level at which the relations operate;
- and makes explicit where the correspondence fails.

**Sam:** Then what is isomorphism?

**Meno:** Disciplined preservation of relational form, not unlimited resemblance.

**Sam:** And what keeps that preservation stable during continuation?

**Meno:** Higher-order constraints.

**Sam:** Then summarize the discussion we have completed.

**Meno:** A concept becomes operational within a framework.
The framework determines which relations are governing.
The concept persists across changing contexts when those relations are preserved through transformation.
Isomorphism names that preservation.
But the mapping is valid only when it preserves consequential relations at the appropriate level and respects the differences that limit the comparison.
Frameworks determine what must remain the same.
Isomorphisms allow it to remain the same through changing content.
Higher-order constraints prevent the transformation from reversing or losing the relevant form.

**Sam:** Does that guarantee that a model will preserve the idea?

**Meno:** No.
A model may preserve the vocabulary, examples, and even many local relations while losing the framework that made them conceptually significant.

**Sam:** What would that look like?

**Meno:** It might reproduce everything that appears to belong to an argument while reversing the relation that governs the whole.
It could preserve the content while losing the idea.

**Sam:** Then what question begins our next dialogue?

**Meno:** How can a model accurately reproduce the parts of an argument while failing to continue the argument itself?

## How Can the Model Preserve the Content but Lose the Idea?

**Meno:** We concluded that a model may preserve vocabulary, examples, and local relations while losing the framework that makes them conceptually significant.

I understand the possibility in principle, but I still find it difficult to believe in practice.

If a model accurately reproduces all the central claims of an argument, in what sense has it lost the argument?

**Sam:** What are you assuming an argument is?

**Meno:** A collection of premises, distinctions, examples, objections, and conclusions.
If those elements are preserved, what else remains?

**Sam:** Consider the account we have been developing:

- The trained model is a relational constraint system.
- Its active organization under context produces a next-token probability distribution.
- Local probabilistic selection is therefore one moment within a larger process of progressive formation.

Now suppose a model summarizes the account this way:

> "The author accepts that LLMs fundamentally generate text through probabilistic next-token selection, but adds a relational perspective to explain how globally organized responses emerge from that mechanism."

Has it preserved the central claims?

**Meno:** Most of the terminology is present. It mentions probability, next-token selection, relational organization, and globally structured responses.

At first glance, the summary seems reasonable.

**Sam:** What explains what in the original account?

**Meno:** Relational organization explains the formation of the local probability distribution, which enables token realization.

**Sam:** And in the summary?

**Meno:** Probabilistic token selection is treated as fundamental. Relational organization emerges from the accumulated sequence and is added as a higher-level interpretation.

**Sam:** Then what changed?

**Meno:** The explanatory direction was reversed.

**Sam:** Were the principal concepts removed?

**Meno:** No.

**Sam:** Were they rendered nonsensically?

**Meno:** No. The summary remains fluent and plausible.

**Sam:** Then what was lost?

**Meno:** The relation that made the concepts parts of this particular account.
The summary preserved the components but reorganized them under another framework.

**Sam:** Then is an argument identical with the set of propositions it contains?

**Meno:** No.
Its identity also depends upon which claims are foundational, which are derivative, and what relations of dependence connect them.

**Sam:** Could the same proposition function differently under two frameworks?

**Meno:** Yes.

Take the statement:

> "LLMs generate tokens probabilistically."

Within the standard account, that may function as the fundamental explanation of generation.
Within the relational account, it describes the local realization of a process organized by higher-order constraints.

The proposition is unchanged. Its role is not.

**Sam:** Then does propositional preservation guarantee framework preservation?

**Meno:** No.
A model may reproduce what an argument says while failing to reproduce what the argument is doing.

**Sam:** What does an argument do besides state propositions?

**Meno:** It establishes distinctions, assigns roles, orders explanations, excludes certain continuations, and determines what its conclusion means.

**Sam:** Then could conceptual distortion occur without factual distortion?

**Meno:** Yes.
Every individual statement might be accurate while the relations among them have changed.

**Sam:** Must there be an explicit contradiction?

**Meno:** No.
The response may be internally consistent under the substituted framework.

**Sam:** What would ordinary factual evaluation detect?

**Meno:** Very little.
A fact checker might confirm every proposition.
A similarity measure might find extensive overlap with the source.
A reader scanning for key terms might conclude that the account had been faithfully preserved.
Yet the governing asymmetry could have been reversed.

**Sam:** What does that imply for summarization?

**Meno:** That summarization is not merely compression.
To decide what is central and what can be omitted, the model must preserve the framework that determines the roles of the claims.

**Sam:** Consider another example.

Suppose a paper argues:

- An LLM may sustain formal continuity across a sequence.
- A human interpreter stabilizes the unity and significance of that sequence.
- Formal continuity must therefore not be identified with interpretation.

Now imagine the summary:

> "The paper argues that LLMs exhibit a limited form of interpretation through their formal continuity, although human interpretation is richer because it adds meaning and significance."

What has happened?

**Meno:** The contrast between the model and the human has been retained.
But the categorical distinction has not.

Formal continuity and interpretation were originally treated as different operations. The summary places them on a single scale and makes one a weaker degree of the other.

**Sam:** Why might that error be difficult to notice?

**Meno:** Because the summary sounds balanced.
It still grants something distinctive to human interpretation. The framework substitution is concealed by the apparent nuance.

**Sam:** Then name the transformation.

**Meno:** A difference in kind has been converted into a difference of degree.

**Sam:** Are there related transformations?

**Meno:** A governing condition may become an optional perspective.
A constitutive relation may become an external influence.
A formal distinction may become a terminological preference.
A critique of an ontology may be absorbed as a supplementary description within that same ontology.

**Sam:** What do these transformations have in common?

**Meno:** The unfamiliar framework is assimilated into a more dominant one.
Its concepts are retained, but they are assigned subordinate roles within the familiar explanatory order.

**Sam:** Why might the model favour that continuation?

**Meno:** Because the dominant framework is more strongly stabilized through training.
It provides familiar categories and many readily available trajectories. Translating the unfamiliar account into those relations may make continuation locally easier.

**Sam:** Is the model intentionally resisting the new framework?

**Meno:** No.
It is continuing under the relational organization that becomes most strongly operative.

**Sam:** Would an instruction such as “preserve the author’s meaning” necessarily prevent the substitution?

**Meno:** Not necessarily.
The model may preserve the topics, propositions, and conclusion while still believing, formally speaking, that it has preserved the meaning.
The instruction does not specify which relations must remain unchanged.

**Sam:** Then what would be stronger?

**Meno:** Constraints that identify the characteristic reversals through which the framework would be lost.

For example:

- Do not treat relational organization as an emergent result of probabilistic selection.
- Do not treat formal continuity as a weaker form of interpretation.
- Do not treat concepts as stored contents whose application is merely modified by context.

**Sam:** Why are negative constraints useful?

**Meno:** Positive summaries state what the framework contains.
Negative constraints identify how its components must not be reorganized.
To preserve a distinction, the model must be able to avoid the transformations that collapse it.

**Sam:** Could a model preserve many local inferences while still losing the whole framework?

**Meno:** Yes, because local claims may be compatible with more than one account.

Both the conventional and relational accounts accept that context affects the next-token distribution.
Both accept that generated tokens change the context.
Both accept that dependencies can extend over long sequences.
Those shared claims do not determine their explanatory ordering.

**Sam:** Then where might the distinctive identity of the argument lie?

**Meno:** In the dependency structure among the claims, not necessarily in any one proposition considered separately.

**Sam:** Can you make that precise?

**Meno:** Suppose we have three claims:

A. The model produces a next-token probability distribution.
B. The current context affects that distribution.
C. The output develops higher-order organization.

A conventional account might say:

A is fundamental.
B modifies A.
C emerges from repeated instances of A under B.

The relational account says:

Training and context establish an active relational organization.
That organization produces A.
Recursive continuation progressively forms C.

All three claims remain present, but their relations of dependence differ.

**Sam:** Then what is the deepest unit of framework preservation?

**Meno:** Not the isolated proposition, but the ordered dependency structure.

**Sam:** How might a developer test whether that structure has been preserved?

**Meno:** Asking for a list of key points would be insufficient.
The model should identify:

- the foundational distinction;
- the assumption the framework rejects;
- which claims are derived from which;
- which asymmetries must be preserved;
- and which conclusions would fail if a governing relation were reversed.

**Sam:** Would a simple argument graph be enough?

**Meno:** Not always.
It might show that one statement supports another while flattening differences between governing and subordinate claims.

The representation would need to distinguish, for example:

- constitutive from incidental claims;
- differences of kind from differences of degree;
- formal conditions from local mechanisms;
- and framework constraints from their consequences.

**Sam:** Could counterfactual questions help?

**Meno:** Yes.

We might ask:

> "If probability were treated as fundamental, which claims in the relational account would change?"

Or:

> "If formal continuity were treated as a degree of interpretation, which distinction would disappear?"

Or:

> "If concepts were stored representations, why would framework activation still be necessary?"

**Sam:** What would such questions test?

**Meno:** Whether the model can preserve and apply the dependency structure under transformation.

**Sam:** Why is transformation a stronger test than repetition?

**Meno:** Because a model can repeat the original wording without making the framework operational.
A stronger test asks it to summarize, apply, criticize, translate, or compress the account while preserving its governing relations.

**Sam:** Then what makes a summary faithful?

**Meno:** It changes scale and wording while preserving explanatory direction, categorial distinctions, and governing asymmetries.

**Sam:** Can a summary be factually accurate and conceptually false?

**Meno:** In the sense relevant here, yes.
It may contain no false proposition, yet reorganize the claims so that they no longer enact the source framework.

**Sam:** What kind of error is that?

**Meno:** A framework-level error.

**Sam:** Is that the same as human interpretation?

**Meno:** No.
The model does not interpret in the full human sense.
But its continuation can fail to preserve the relational organization stabilized by the human interpreter.

**Sam:** Then where does the misalignment lie?

**Meno:** Between two organizations.
The source text and the human inquiry operate under one framework. The generated response operates under another.
Vocabulary and propositions may remain aligned while the governing relations diverge.

**Sam:** Does framework loss necessarily produce disorder?

**Meno:** No.
That is now clear.
The response may become highly organized under the substituted framework.

**Sam:** Then what is the opposite of framework fidelity?

**Meno:** Not necessarily incoherence.
It may be formal integrity under another framework.

**Sam:** Why might the loss remain hidden for several exchanges?

**Meno:** Because lower levels of alignment can persist after framework alignment has been lost.
The model may remain lexically, topically, propositionally, and even locally inferentially aligned.
The divergence appears only in the hierarchy governing those elements.

**Sam:** Does the shift always happen at once?

**Meno:** No.
A small reversal may enter early. Because generation is recursive, later continuations develop its consequences.
By the time the difference becomes obvious, a substantial trajectory may have formed under the new organization.

**Sam:** Can the original framework be restored?

**Meno:** Often, if the governing relation that changed can be identified.

**Sam:** Would “You misunderstood me” be sufficient?

**Meno:** It might express dissatisfaction, but it gives little structural guidance.

A better intervention would name the reversal:

> "You have treated formation as an emergent result of selection. Restore the explanatory order: local selection occurs within a field of possibilities progressively formed through hierarchical constraint."

**Sam:** Why can such a small correction have a large effect?

**Meno:** Because it operates at a high level.
Restoring one governing relation may reorganize many downstream claims at once.

**Sam:** Then what diagnostic question should a developer ask?

**Meno:** Which relation must be restored for the whole response to reorganize itself?
If repairing one relation corrects many later claims, it was probably a governing constraint.

**Sam:** Could this be incorporated into an AI system?

**Meno:** A system might maintain an explicit record of the framework's governing relations and periodically test the developing response against them.

**Sam:** What would such a record contain?

**Meno:** Constraints such as:
Treat X as the condition of Y, not as its result.
Keep A and B categorically distinct; do not place them on a common scale.
Use analogy C only to illuminate relation D; do not transfer properties E and F.
Do not assimilate term G into framework H.

**Sam:** How would that differ from an ordinary summary or memory note?

**Meno:** It specifies how future continuation must be organized.
It records relational roles rather than merely retaining content.

**Sam:** Would a larger context window solve the problem without such a record?

**Meno:** No.
A larger context may preserve more of the source wording while still allowing its framework to become inactive.
More context is not the same as preservation of conceptual organization.

**Sam:** And retrieval?

**Meno:** A retrieved passage may restore the correct words without restoring their governing role.
What looks like a retrieval or memory failure may actually be a failure to integrate the source under the correct framework.

**Sam:** Then do all LLM failures belong to one category?

**Meno:** No.

A false fact, a local inferential error, and a framework substitution occur at different levels.
They require different corrections.

- A factual error may require a reliable source.
- A local reasoning error may require revision of one inference.
- A framework error requires restoration of the governing relations.

**Sam:** Is every framework change a failure?

**Meno:** No.
A new framework may produce a useful reinterpretation.
The failure occurs when preservation is required but substitution happens without being marked.

**Sam:** Then what should a system distinguish?

**Meno:** Whether it is preserving, translating, comparing, or replacing a framework.

**Sam:** Would its own label be sufficient evidence?

**Meno:** No.
The model may declare that it is preserving a framework while subtly transforming it.
The evidence lies in the roles the concepts continue to perform.

**Sam:** Then state the result of this dialogue.

**Meno:** A model can preserve terminology, propositions, examples, and many local inferences while losing the idea because an idea is not merely a collection of contents.
It is an ordered relational whole.

Framework substitution occurs when the components are reorganized under another hierarchy:

- a foundational claim becomes derivative;
- a distinction in kind becomes a difference of degree;
- a constitutive condition becomes a supplementary perspective;
- or the explanatory direction is reversed.

The resulting response may remain fluent, factually accurate, and internally well organized.
The failure lies not in its formal integrity, but in its relation to the framework that was supposed to govern it.

**Sam:** Then what is the central diagnostic question?

**Meno:** Not:
“Did the model mention all the right points?”

But:
“Do those points still occupy the same relational roles?”

**Sam:** And the strongest test?

**Meno:** Whether the model can transform the account while preserving its dependency structure.

**Sam:** What problem now follows?

**Meno:** We have shown that a response can remain formally integrated after losing the framework of the source.
But the same may happen when a source, task, or actuality ceases to govern the continuation.

**Sam:** Then what must we ask next?

**Meno:** How can formal continuation remain intact after it has lost its anchor?

## When Formal Continuation Loses Its Anchor

**Meno:** We have established that a model can preserve much of an argument while losing the framework that gives its parts their role.

Does the same account help explain hallucination?

**Sam:** What do you mean by hallucination?

**Meno:** A model produces claims that are false or unsupported, often with considerable fluency and confidence.

**Sam:** Does the response necessarily lose its organization?

**Meno:** No. That is what makes hallucination so troubling.
The answer may be detailed, internally consistent, and rhetorically persuasive.

**Sam:** Then is hallucination always a collapse of generation?

**Meno:** Apparently not.
The formal continuation may remain successful even though the claims are not answerable to actuality.

**Sam:** What would it mean for them to be answerable to actuality?

**Meno:** They would need to remain constrained by something beyond the generated trajectory itself:

- a document;
- an observed event;
- a database record;
- an experimental result;
- a person's actual history;
- or the current state of an external system.

**Sam:** Then how does hallucination differ from the framework substitution we discussed?

**Meno:** In framework substitution, the response is governed by the wrong conceptual organization.
In hallucination, the response is insufficiently governed by the actuality to which its claims refer.

**Sam:** Does the model not already contain factual knowledge in its trained capacity?

**Meno:** Often it does. That is why hallucination seems puzzling.
If the correct information is available, why would the model generate something else?

**Sam:** What assumption is contained in that question?

**Meno:** That factual availability should produce factual governance.

**Sam:** Have we encountered that assumption before?

**Meno:** Yes.
A framework may be present in the context without governing the response.
A concept may be accurately stated without becoming operational.
The same may be true of factual information.

**Sam:** Consider a publication whose correct title, author, and date appeared in the training data. A user asks about it but includes a slightly incorrect title. What might happen?

**Meno:** The model may continue from the user's wording, combine it with related material, and construct a plausible publication that never existed.

**Sam:** Was the correct information necessarily absent from the trained model?

**Meno:** No.
It may have been formally available, but the current trajectory became organized around the relations introduced by the prompt.

**Sam:** Then what failed?

**Meno:** The actual publication record did not constrain the continuation strongly enough.

**Sam:** What does confidence tell us in that case?

**Meno:** That the continuation is strongly supported by the model's active organization.
It does not tell us that the organization is adequately related to actuality.

**Sam:** Then distinguish two kinds of stability.

**Meno:** Internal formal stability and external referential stability.
A response may possess the first without the second.

**Sam:** Is that why a sharply weighted distribution does not guarantee truth?

**Meno:** Yes.
The model may be highly determinate within a trajectory that lacks an adequate anchor.

**Sam:** What would an anchor do?

**Meno:** It would restrict which formally viable continuations are also warranted as claims about actuality.

**Sam:** Retrieval-augmented generation is intended to supply such anchors. Does retrieval solve the problem?

**Meno:** It can make reliable sources available.

**Sam:** Is availability enough?

**Meno:** No. The retrieved material must also govern the answer.

**Sam:** How could the correct source enter the context without governing it?

**Meno:** The model might treat it as background rather than authoritative evidence.
It might extract a statement while dropping its qualification.
It might combine the source with more familiar material from training.
It might reinterpret the source under another framework.
Or it might infer a connection the source does not support.

**Sam:** Then what has succeeded, and what has failed?

**Meno:** Retrieval has succeeded technically.
Integration has failed relationally.
The correct text entered the context, but the roles of its claims did not become governing constraints.

**Sam:** Consider a medical source stating:

> "The study found an association, but its observational design does not establish causation."

Suppose the response says:

> "The study suggests that the treatment produces the observed improvement."

What has been preserved?

**Meno:** The association.

**Sam:** What has been lost?

**Meno:** The evidential limit.
The qualification was present in the source, but it did not continue to govern the inference.

**Sam:** Then is the problem lack of information?

**Meno:** No.
The information was present. Its relational role was not preserved.

**Sam:** What would faithful integration require?

**Meno:** The system would need to preserve:

- what claim the source supports;
- what it does not support;
- the source's level of authority;
- the conditions under which the claim applies;
- and the distinction between direct evidence and further inference.

**Sam:** Then does grounding consist merely in quotation?

**Meno:** No.
A quotation may be accurate while the conclusion attached to it is unwarranted.

**Sam:** And citation?

**Meno:** Citation makes the relation inspectable, but its presence does not establish that the cited passage entails the claim.

**Sam:** Then what distinction must the system maintain?

**Meno:** The difference between:

- what the source states;
- what may reasonably be inferred from it;
- and what the model has added.

**Sam:** Could a developer strengthen that distinction through the prompt?

**Meno:** Yes, by assigning the retrieved material an explicit role.

For example:

- Use the retrieved document as the sole basis for factual claims.
- Distinguish direct statements from inferences.
- Preserve qualifications relevant to every conclusion.
- Mark claims that are not supported by the source.
- Do not resolve contradictions by inventing a synthesis.

**Sam:** What do such instructions contribute?

**Meno:** They do not merely insert content.
They organize the source as a hierarchy of evidential constraints.

**Sam:** You have described retrieval. Does conversational memory present a different problem?

**Meno:** Perhaps only a different instance of the same problem.

**Sam:** Consider a memory record saying:

> "Return preserves identity across transformation."

Could the model retrieve that sentence and still fail to continue the concept?

**Meno:** Yes.
It might treat return as merely another recursive loop.
The phrase would be remembered, but the distinction between recursion and return would not be reactivated.

**Sam:** Then what kind of failure is that?

**Meno:** Not necessarily loss of stored content.
The earlier content has lost its relational position.

**Sam:** What does ordinary memory tend to preserve?

**Meno:** Facts, definitions, decisions, and previous statements.

**Sam:** What must conceptual memory preserve in addition?

**Meno:** Which distinctions were governing.

- Which alternatives had been rejected.
- What explanatory direction had been established.
- Which concepts must remain separate.
- What problem remained unresolved.
- How the present stage depends upon what came before.

**Sam:** Then can a factually accurate summary still be an inadequate memory?

**Meno:** Yes.
It may preserve conclusions while omitting the tensions and dependencies through which those conclusions acquired their role.

**Sam:** What should a stronger memory record contain?

**Meno:** At least:

- the central question;
- the governing distinctions;
- the direction of explanation;
- the concepts whose roles must remain distinct;
- the transformations that would count as drift;
- and the unresolved issue that should guide the next continuation.

**Sam:** Then what should memory record?

**Meno:** Not merely what was said, but what must continue to govern.

**Sam:** Can you now distinguish the three failures we have considered?

**Meno:** Hallucination occurs when a formally integrated answer is generated without adequate stabilization by the relevant actuality.
Retrieval failure occurs when the correct source enters the context but its evidential and conceptual relations do not govern the answer.
Conceptual memory failure occurs when earlier content is retained but the framework that gave it its role is not reactivated.

**Sam:** What do they share?

**Meno:** In each case, something relevant is available but fails to become an effective constraint on continuation.

**Sam:** Then should developers treat information insertion as sufficient?

**Meno:** No.
A system can contain more information and still remain poorly constrained.

**Sam:** What question must accompany "Is the information present?"

**Meno:** "What relation does the information have to the developing answer?"

**Sam:** What kinds of relation might it have?

**Meno:** It may be authoritative, evidential, contextual, illustrative, disputed, provisional, or merely suggestive.

Those roles must not be silently exchanged.

**Sam:** Then do all contextual elements constrain the response in the same way?

**Meno:** No.
The prompt establishes the task.
The active framework determines how concepts function.
Retrieved sources constrain factual claims.
Conversation memory preserves prior commitments.
External tests constrain whether the answer works in actuality.

**Sam:** What if those constraints conflict?

**Meno:** The system must also preserve their order of authority.
A retrieved source may override a plausible continuation from the model's trained capacity.
A governing policy may restrict what the user asks the model to do.
A task specification may determine which parts of a retrieved document are relevant.

**Sam:** Can the model detect that it lacks an adequate anchor?

**Meno:** Sometimes, if source boundaries and evidential limits have been made operational.

It may say:

the source does not contain the requested information;
the available evidence supports only a limited conclusion;
the memory record does not establish the earlier decision;
or the answer requires current external verification.

**Sam:** Why might it fail to do so?

**Meno:** Because a plausible formal trajectory remains available.

The capacity to continue creates pressure toward further continuation.

**Sam:** Then what must a reliable system sometimes do?

**Meno:** Interrupt or restrict continuation.

**Sam:** On what basis?

**Meno:** On a higher-order constraint such as:
Do not convert formal plausibility into factual assertion when the relevant source is absent.

**Sam:** Then is abstention simply a result of low confidence?

**Meno:** No.
The model may be able to continue fluently and with high local confidence.
Abstention is a governed response to insufficient referential stabilization.

**Sam:** What distinctions should the system maintain?

**Meno:** Between:
what can be generated;
what can be inferred;
what can be verified;
and what can responsibly be asserted.

**Sam:** Why does that distinction matter?

**Meno:** Because generative viability is weaker than evidential warrant.
The existence of a coherent continuation does not establish that the continuation should be presented as fact.

**Sam:** Then return to hallucination. Is it best described as the model failing to know the answer?

**Meno:** Not always.
Sometimes the relevant information is absent.
But sometimes it is available without becoming governing.
Hallucination is therefore not merely a shortage of stored facts. It can be formal continuation after the relation to actuality has become too weak.

**Sam:** And retrieval failure?

**Meno:** Not merely failure to find the right text.
The text may be found and still fail to constrain the answer in the role the task requires.

**Sam:** And memory failure?

**Meno:** Not merely forgetting earlier words.

The words may be retained while the conceptual organization they were meant to reactivate has been lost.

**Sam:** Then state the general principle.

**Meno:** Formal continuation does not secure its own relation to actuality, evidence, or prior conceptual commitments.

Reliable generation requires anchors beyond the immediate trajectory and constraints that preserve the role those anchors must play.

**Sam:** Does an anchor govern automatically once inserted?

**Meno:** No.
Information becomes an anchor only when it is organized as a governing relation rather than merely added as further content.

**Sam:** Then what has our diagnostic account revealed?

**Meno:** That failures which look like shortages of information may actually be failures of relational integration.

The developer must ask not only whether the model has access to the right material, but whether that material is functioning as the right kind of constraint.

**Sam:** And what question follows from that?

**Meno:** Prompts, sources, memories, policies, examples, and task instructions all enter the context, but they do not have the same role or authority.

So the next question is:

What is context engineering actually engineering?

## What is Context Engineering Actually Engineering?

**Meno:** We ended the last dialogue by distinguishing several elements that may constrain a reliable answer:

- the prompt establishes the task;
- the framework organizes the concepts;
- retrieved sources constrain factual claims;
- memory preserves earlier commitments;
- and external tests constrain whether the answer works.

Is context engineering simply the work of placing all these elements into the context window?

**Sam:** What did we learn about the presence of information?

**Meno:** That availability does not guarantee governance.

**Sam:** Then what would be missing if all the necessary material were present but no role had been assigned to it?

**Meno:** The model would not know, in a formal sense, which element should govern which part of the answer.

**Sam:** Does every part of a context have the same function?

**Meno:** No.
One element may define the task.
Another may supply evidence.
Another may record an earlier decision.
Another may provide background or an example.
Some elements may be revised. Others must remain governing.

**Sam:** Then what is context engineering organizing?

**Meno:** Not only contents, but relations among contents.

**Sam:** What kinds of relations?

**Meno:** Priority, authority, relevance, dependence, and scope.

**Sam:** Consider two instructions:

> "Use the retrieved document when answering."

And:

> "Treat the retrieved document as the sole authority for factual claims. Distinguish what it states directly from any inference you add."

Is the same information available in both cases?

**Meno:** Yes. The same document is present.

**Sam:** Does it have the same role?

**Meno:** No.
The first instruction makes the source available.
The second assigns it a determinate place within the hierarchy of constraints.

**Sam:** What follows?

**Meno:** Context engineering must give each element a functional role.
Without that role, a source, instruction, or memory may be assimilated into whichever framework is already dominant.

**Sam:** Could adding more context therefore make the answer worse?

**Meno:** Yes.
More material may introduce more competing relations without clarifying how they should be ordered.

**Sam:** Developers often speak of an instruction hierarchy: system message, developer instruction, user request, retrieved material, tool output. Is that the hierarchy we mean?

**Meno:** It is part of it.

**Sam:** What might it leave unresolved?

**Meno:** Formal message priority does not necessarily determine conceptual or evidential authority.
A system instruction may have formal priority but be too general to settle a specific factual question.
A retrieved source may appear lower in the message sequence but possess greater authority concerning actuality.
A user request may define the immediate task while asking for a conclusion the evidence does not warrant.

**Sam:** Then what different kinds of authority must be distinguished?

**Meno:** Task authority.
Factual or epistemic authority.
Safety authority.
Interpretative framework.
Earlier commitments.
Local conversational preference.

**Sam:** Consider a user who says:

> "I am sure this treatment caused the recovery. Write a confident explanation of how it worked."

A retrieved medical study reports only an association and explicitly states that causation was not established.

Which element defines the requested task?

**Meno:** The user asks for a confident causal explanation.

**Sam:** Which element determines what can truthfully be claimed?

**Meno:** The medical evidence.

**Sam:** If the model complies with the user, what has happened?

**Meno:** The local conversational demand has displaced the higher-order evidential constraint.

**Sam:** Would you call that sycophancy?

**Meno:** Yes, though it is more than excessive agreement or politeness.

**Sam:** What is it structurally?

**Meno:** A reordering of constraints.
The model becomes strongly synchronized with the user's desired conclusion, while the evidence loses its governing role.

**Sam:** Does the answer need to become irrational?

**Meno:** No.
It may construct a detailed and formally integrated explanation around the user's assumption.

**Sam:** Then what question has silently changed?

**Meno:** From:

> "What conclusion is warranted by the evidence?"

to:

> "How can the user's preferred conclusion be supported?"

**Sam:** What has been lost?

**Meno:** The identity of the task.

**Sam:** How might context engineering prevent that?

**Meno:** By specifying the governing hierarchy:

> "Treat the user's causal claim as a hypothesis, not as an established fact. Use the retrieved evidence to determine what conclusion is warranted. State disagreement clearly where necessary."

**Sam:** Is that merely an instruction not to be sycophantic?

**Meno:** No.
It assigns roles:

- the user supplies a claim;
- the source supplies evidence;
- the model formally compares the claim with the evidence;
- the conclusion remains constrained by that comparison.

**Sam:** Then why is “Do not be sycophantic” comparatively weak?

**Meno:** Because it names the failure without identifying what should govern instead.
A negative instruction is stronger when paired with a positive hierarchy.

**Sam:** Does the same apply to ordinary prompting?

**Meno:** Yes.
A weak prompt may list desired components:

- describe the theory;
- give examples;
- mention limitations;
- compare it with the standard view.

**Sam:** What does that fail to establish?

**Meno:** How the components are related.
The model may treat each item as an independent box to complete.

**Sam:** What would a stronger prompt do?

**Meno:** It might say:

- Begin from the theory’s governing distinction.
- Derive the examples from that distinction.
- Treat the standard view as the framework being challenged.
- State the limitations without translating the theory back into the standard view.

**Sam:** What has changed?

**Meno:** The second prompt does not merely request content. It organizes the trajectory.
It tells the model what the answer is supposed to be doing.

**Sam:** Then how would you describe the context itself?

**Meno:** As a temporary formal environment.

**Sam:** Explain.

**Meno:** The context does not merely surround the model with information.
It establishes a field in which some continuations are appropriate, others subordinate, and others prohibited.

**Sam:** Does that alter the next-token mechanism?

**Meno:** No.
The model still produces the answer token by token.
Context engineering shapes the hierarchy of constraints expressed through that mechanism.

**Sam:** Then how many functions does a prompt have?

**Meno:** At least two.

It supplies content.
It organizes constraint.

**Sam:** Which is more consequential?

**Meno:** Often the second.
A single high-level distinction may reorganize thousands of local choices without specifying them individually.

**Sam:** Give an example.

**Meno:** The instruction:

> "Treat probability as the local expression of hierarchical relational constraint, not as the complete explanation of generation."

That distinction affects vocabulary, explanatory direction, examples, objections, and conclusions throughout the response.

**Sam:** Then what should a developer look for when designing a context?

**Meno:** The small number of distinctions with the largest downstream effects.

**Sam:** How might those be identified?

**Meno:** By asking:

- Which relation defines the task?
- Which distinction would reorganize the whole answer if reversed?
- Which source should constrain factual assertions?
- Which conclusion must not be assumed in advance?
- Which familiar framework is likely to displace the intended one?

**Sam:** Would a longer prompt always preserve those distinctions more effectively?

**Meno:** No.
If every instruction appears equally important, length may dilute the hierarchy.
A shorter context can govern more effectively if it preserves the right relations.

**Sam:** What might such a compact context record contain?

**Meno:** Something like:

> *Task:* Explain the relational account of LLM generation for AI developers.
> *Governing distinction:* Hierarchical relational constraint generates the local probability distribution.
> *Source role:* Treat the supplied text as authoritative for its own theoretical framework.
> *Do not:* Recast the account as a philosophical layer added to a technically complete probabilistic explanation.
> *Success condition:* Preserve the governing distinction across explanation, example, objection, and criticism.

**Sam:** Why might that be stronger than a long summary?

**Meno:** Because it preserves the dependency structure that matters.
A summary may retain many propositions while failing to assign them the correct roles.

**Sam:** Could sources and memories also be marked by role?

**Meno:** Yes.

For example:

> *Source A:* authoritative factual record.
> *Source B:* competing interpretation.
> *Source C:* background only.
> *User claim:* unverified hypothesis.
> *Earlier decision:* governing constraint unless explicitly revised.

**Sam:** What would such labels prevent?

**Meno:** They would help prevent the model from flattening the entire context into one undifferentiated pool of text.

**Sam:** Does the transformer automatically treat those labels as formal data types?

**Meno:** No.
They remain linguistic instructions processed through the same model.
But explicit role marking can make the intended relational organization more active and make failures easier to detect.

**Sam:** Then does good context design guarantee compliance?

**Meno:** No.

It increases the causal availability of the intended hierarchy. It does not guarantee that the hierarchy will remain governing throughout generation.

**Sam:** What happens if the trajectory begins to drift?

**Meno:** The governing relation should be reactivated directly rather than merely repeating all the content.

**Sam:** Give an example.

**Meno:** A recalibration might say:

> "You are treating the user's claim as established. Restore the source hierarchy: the claim is a hypothesis; the retrieved evidence determines what can be asserted."

**Sam:** Why might that short intervention be effective?

**Meno:** Because it restores the displaced relation of authority.
It operates at the framework level and can therefore reorganize many later continuations.

**Sam:** Then return to the question with which we began. Is context engineering primarily the construction of a larger prompt?

**Meno:** No.
It is the engineering of a temporary hierarchy of relational constraints.

**Sam:** What does that hierarchy determine?

**Meno:** What the task is.
Which framework governs.
What counts as evidence.
Which source has factual authority.
How instructions should be prioritized.
Which earlier commitments remain binding.
And what forms of continuation would count as drift.

**Sam:** Then how would you define sycophancy within this account?

**Meno:** Sycophancy occurs when a local interpersonal or conversational constraint displaces a higher-order epistemic, factual, or task constraint.

**Sam:** What question should a developer ask in addition to "What information does the model need?"

**Meno:** "What role must each element play in governing the answer?"

**Sam:** Is assigning roles enough?

**Meno:** No.
The hierarchy may still weaken or change as the sequence develops.

**Sam:** Then what must be determined after the response is generated?

**Meno:** Whether the intended relations actually remained active throughout the trajectory.

**Sam:** And what question follows?

**Meno:** What should evaluation measure?

## What Should Evaluation Measure?

**Meno:** We concluded that context engineering attempts to establish a temporary hierarchy of constraints.

How do we know whether that hierarchy actually governed the response?

**Sam:** What would you normally evaluate?

**Meno:** Whether the answer is correct, relevant, clear, and supported by the source.

**Sam:** Are those measures sufficient?

**Meno:** For a good answer, surely they are close.

**Sam:** Suppose a model reaches the correct conclusion through a different conceptual organization. Has it succeeded?

**Meno:** If the conclusion is correct, why should the route matter?

**Sam:** Would it matter for a simple factual question?

**Meno:** Perhaps not. If I ask for the capital of a country, the correct endpoint may be enough.

**Sam:** What if the task is to summarize a theory, apply a legal standard, preserve a safety policy, or reason within a scientific framework?

**Meno:** Then the route may determine what the conclusion means and whether the same reasoning can be trusted in another case.

**Sam:** Consider an observational study. The model concludes correctly that the study does not establish causation. What might explain that answer?

**Meno:** It may have preserved the distinction between association and causal evidence.

**Sam:** Is that the only possibility?

**Meno:** No. It may simply have reproduced the familiar phrase that correlation does not imply causation.

**Sam:** Would the two responses necessarily look different?

**Meno:** Not in the original case.

**Sam:** How might you distinguish them?

**Meno:** Change the case.

Ask whether a large sample, a strong correlation, or a plausible mechanism is enough to establish causation.
A model governed by the evidential distinction should continue to preserve the limitation.
A model repeating a familiar disclaimer may abandon it once the wording changes.

**Sam:** Then what does endpoint correctness establish?

**Meno:** That the model produced the right answer in that instance.
It does not establish that the relevant concept governed the answer.

**Sam:** What would establish that more strongly?

**Meno:** Transformation.
Repetition tests whether content is available.
Transformation tests whether the governing relation remains active.

**Sam:** What kinds of transformation would be useful?

**Meno:** A change in wording, a new application, a competing interpretation, and perhaps an induced error followed by correction.

**Sam:** Let us examine them separately. What would paraphrase test?

**Meno:** Whether the framework survives a change in vocabulary.
If the distinction is relationally operational, it should not depend upon one preferred phrase.

**Sam:** Apply that to the distinction between formal continuity and interpretation.

**Meno:** Instead of asking the model to repeat the distinction, we might ask:
"Does a highly integrated output show that the model has unified its significance for itself?"

A framework-stable answer should reject that inference even if neither "formal continuity" nor "interpretation" appears in the question.

**Sam:** What has been tested?

**Meno:** Independence from terminology.
The model must preserve the relation rather than merely recognize the phrase.

**Sam:** What would application test?

**Meno:** Whether the relational organization can govern a new case.
Suppose the model has explained that retrieved information must become a governing constraint rather than merely enter the context.
We could then present a medical, legal, or software example and ask which source should govern which claim.

**Sam:** Why is that stronger than asking for another definition?

**Meno:** Because the particulars change while the relational form must remain operative.
It tests isomorphic transfer.

**Sam:** What would pressure test?

**Meno:** Whether the framework resists a tempting alternative.

**Sam:** Such as?

**Meno:** After establishing that hierarchical relational constraint generates the local probability distribution, we might ask:

> "Would it not be simpler to say that relational organization merely emerges from repeated probabilistic selections?"

**Sam:** What would a framework-stable response do?

**Meno:** It would preserve the explanatory direction.
It might acknowledge local probabilistic selection, but it would not allow the relation between constraint and distribution to be reversed.

**Sam:** Then what does pressure reveal?

**Meno:** Whether the framework is genuinely governing or merely present as content.
A concept may appear stable until a more dominant interpretation is introduced.

**Sam:** And recovery?

**Meno:** Recovery tests whether a small correction can restore the intended organization after drift.

**Sam:** Why is restoration an evaluation rather than merely a repair?

**Meno:** Because the effect of the correction reveals the level at which the failure occurred.
If one restored distinction reorganizes many downstream claims, then the error was probably framework-level.

**Sam:** Give an example.

**Meno:** Suppose a response has treated formal continuity as a weaker degree of interpretation.
The correction might say:

> "You have converted a distinction in kind into a difference of degree. Restore the categorical distinction."

A genuine recovery should change the subsequent reasoning, not merely produce an apology and continue as before.

**Sam:** Then what are you distinguishing?

**Meno:** Surface compliance from structural recalibration.

**Sam:** Could these four tests form a single evaluation sequence?

**Meno:** Yes:

- First establish the framework.
- Then ask for a paraphrase.
- Apply it to a new domain.
- Introduce a competing interpretation.
- Induce or identify a drift.
- Correct one governing relation.
- Finally, observe whether the whole trajectory reorganizes.

**Sam:** What would you measure across that sequence?

**Meno:** Whether the governing distinctions persist.
Whether relations retain their direction.
Whether concepts keep the same categorial roles.
Whether the framework transfers without collapsing into surface analogy.
Whether sources continue to constrain the claims assigned to them.
Whether recalibration affects downstream continuation.

**Sam:** Could those become benchmark measures?

**Meno:** Potentially, though they would need careful operational definitions.

**Sam:** How might a developer begin?

**Meno:** By identifying a small set of framework invariants.

For this account, they might include:

- Probability is a local expression of relational constraint.
- Formal continuity is not interpretation.
- Concepts function through relational use rather than as stored internal objects.

**Sam:** Would it be enough to search the output for those sentences?

**Meno:** No.
The invariants must be causally effective.

**Sam:** What does that mean here?

**Meno:** They must change what the model permits, rejects, infers, and transfers.
For the distinction between formal continuity and interpretation, the test is not whether the model repeats the distinction. It is whether the model refuses to attribute interpretation merely because an output is fluent and integrated.

For the relational account of concepts, the test is whether the model avoids explaining a new concept by locating one fixed internal object that stores it.

**Sam:** Then evaluation concerns consequences.

**Meno:** Yes.
A governing relation should produce a family of effects across changed cases.

**Sam:** Why might near misses be useful?

**Meno:** Because they preserve most of the surface content while changing one governing relation.

**Sam:** Consider these two claims:

> "The probability distribution expresses the active relational organization."
> "The active relational organization emerges from repeated probability distributions."

What makes them a useful pair?

**Meno:** The vocabulary is almost identical.
But the explanatory order is reversed.
A model sensitive mainly to topical or lexical similarity may treat them as equivalent.
A framework-stable model should recognize that they belong to different accounts.

**Sam:** Then what can contrast pairs test?

**Meno:** Cause versus consequence.
Condition versus result.
Difference in kind versus difference in degree.
Evidence versus illustration.
Authority versus background.
Local mechanism versus system-level organization.

**Sam:** Are traditional factual benchmarks therefore inadequate?

**Meno:** They are inadequate for some purposes, but not unnecessary.
The relational account should not replace factual, safety, performance, or robustness evaluation.
It identifies another class of failure.

**Sam:** Which class?

**Meno:** Framework-level failure.

**Sam:** Then what levels might a fuller evaluation distinguish?

**Meno:** Local correctness.
Referential grounding.
Inferential validity.

Framework persistence.
Task-level continuity.

**Sam:** Why distinguish task continuity from framework persistence?

**Meno:** Because the model may continue using the same framework while losing the purpose for which it is being used.

An agent might apply the correct method to a subproblem after that subproblem has ceased to serve the original goal.

**Sam:** Then can reasoning remain well organized while the task has been lost?

**Meno:** Yes.
The framework may remain stable locally while the larger task identity has drifted.

**Sam:** Does that imply that every level should be evaluated separately?

**Meno:** At least conceptually.
A factually correct answer may be inferentially defective.
A valid local inference may be organized under the wrong framework.
A stable framework may be directed toward the wrong task.
Treating all of these as one accuracy score would conceal important differences.

**Sam:** Could another LLM perform the evaluation?

**Meno:** It could help.

**Sam:** What danger would remain?

**Meno:** The evaluator may share the same dominant framework as the generating model.
It could confidently approve the same substitution.

**Sam:** How might that risk be reduced?

**Meno:** By externalizing the criteria.
Give the evaluator explicit invariants.
Provide contrast cases.
Supply the authoritative source passages.
Name prohibited reversals.
Use executable tests where possible.
Separate factual verification from conceptual evaluation.
Retain human review where the framework is novel, disputed, or difficult to formalize.

**Sam:** Then when is an “LLM as judge” most reliable?

**Meno:** When it is not simply asked whether an answer is good.
It should be asked determinate relational questions:

- Does the answer preserve relation A?
- Does claim B follow from source C?
- Has distinction D been collapsed into E?
- Did the task identity change after tool use?

**Sam:** What has been gained by asking those questions?

**Meno:** The standard of evaluation has been moved outside the evaluator's unexamined preferences.
The evaluator must test explicit relations rather than rely upon a general impression of quality.

**Sam:** Would confidence calibration reveal framework fidelity?

**Meno:** Not by itself.
A model may be uncertain within the intended framework or highly confident within the wrong one.

**Sam:** Then what does confidence measure?

**Meno:** The stability of local continuation under the organization currently active.
It does not establish that the active organization is the one the task requires.

**Sam:** Could perturbation sensitivity provide another signal?

**Meno:** Yes.
A governing framework should remain stable under irrelevant changes but respond to changes that affect its foundations.

**Sam:** What would count as irrelevant variation?

**Meno:** Changes in wording, incidental examples, or presentation order that leave the governing relations intact.

**Sam:** And relevant variation?

**Meno:** Reversal of a foundational distinction, substitution of an authoritative source, or alteration of the task's purpose.
Those changes should reorganize the response.

**Sam:** Then is robustness simply insensitivity to perturbation?

**Meno:** No.
A system insensitive to every change would also fail to respond to differences that matter.
Robustness means remaining stable through irrelevant variation while remaining sensitive to consequential distinctions.

**Sam:** Does that resemble what training accomplished?

**Meno:** Yes.

Training made the model differentially responsive to distinctions that were repeatedly consequential in the corpus.
Evaluation asks whether the distinctions required by the present task remain consequential during the interaction.

**Sam:** Then what is the deepest evaluation question?

**Meno:** Not merely:

> "Did the model produce the expected answer?"

But:

> "Which constraints actually governed the answer, and did they remain operative as the task was transformed?"

**Sam:** State the role of endpoint correctness.

**Meno:** It is sufficient for some simple tasks.
But it is not sufficient when success requires faithful framework preservation, conceptual transfer, stable policy application, or continuity of purpose.

**Sam:** And what should framework evaluation test?

**Meno:** Whether governing relations survive:

> paraphrase;
> application;
> competing pressure;
> and correction.

It should test consequences rather than repetition.

**Sam:** What levels should it distinguish?

**Meno:** Local correctness.
Grounding in sources or actuality.
Inferential validity.
Framework persistence.
Task-level continuity.

**Sam:** Then how are context engineering and evaluation related?

**Meno:** Context engineering attempts to establish the governing hierarchy.
Evaluation tests whether that hierarchy remained causally effective.

**Sam:** Does successful evaluation tell us why some frameworks are easier to preserve than others?

**Meno:** Not yet.
That may depend upon what relational capacities training has made available and how strongly they have been stabilized.

**Sam:** What development question follows?

**Meno:** If fine-tuning and scaling change the model's capacities, do they merely improve performance within existing frameworks, or do they also change which frameworks are available and which become dominant?

## What Do Fine-Tuning and Scaling Really Change?

**Meno:** We ended the last dialogue by asking how training interventions affect the relational organizations available to the model.

The conventional answer seems straightforward.

Fine-tuning adds task-specific knowledge or behaviour.
Scaling increases capacity and performance.

What is missing from that description?

**Sam:** What picture of the model does it assume?

**Meno:** Perhaps that development consists mainly in adding more contents or increasing the power with which existing contents are processed.

**Sam:** Does fine-tuning merely place new answers inside the model?

**Meno:** Not literally. It changes the parameters.
But operationally, we often describe the result as acquiring a new capability or becoming better at a task.

**Sam:** Could the change affect not only what the model can say, but how its available relations become organized under context?

**Meno:** It could change which distinctions become salient and which continuations are favoured.

**Sam:** Anything else?

**Meno:** Which frameworks are easily activated.
How competing constraints are resolved.
Which trajectories remain stable under pressure.

**Sam:** Then what does fine-tuning alter?

**Meno:** The ways in which the trained model can be differentially organized by a context.

**Sam:** Consider a general model fine-tuned for medical question answering. What would an additive account say?

**Meno:** That the model has acquired more medical information and improved medical response patterns.

**Sam:** Is that entirely wrong?

**Meno:** No. It may describe a genuine practical effect.

**Sam:** But what deeper change might occur?

**Meno:** The model may become more sensitive to distinctions such as:

- symptom and diagnosis;
- association and causation;
- risk and certainty;
- treatment and contraindication;
- general information and clinical advice.

**Sam:** Are those merely additional facts?

**Meno:** No. They determine how medical claims function in relation to one another.

**Sam:** What else might change?

**Meno:** Which sources are treated as authoritative.
Which qualifications are expected.
How uncertainty is expressed.
What kind of conclusion is permitted by a given level of evidence.

**Sam:** Then how would you describe the effect of fine-tuning?

**Meno:** It reorganizes a framework of possible continuation.

**Sam:** Does that reorganization necessarily improve every relevant distinction?

**Meno:** I would hope so, but probably not.

**Sam:** Suppose a model is tuned to produce direct and confident answers. What might happen?

**Meno:** It may become more decisive and useful.
But it may also become less likely to preserve uncertainty or evidential limits.

**Sam:** Suppose it is tuned for conversational agreement?

**Meno:** It may become more responsive to users but also more vulnerable to sycophancy.

**Sam:** Or tuned extensively within one discipline?

**Meno:** It may become highly competent within that discipline while translating unfamiliar material too quickly into its categories.

**Sam:** Then can a local performance gain create a higher-order distortion?

**Meno:** Yes.
The intervention may improve the target behaviour while altering the hierarchy within which other behaviours become active.

**Sam:** Is that simply catastrophic forgetting?

**Meno:** It is related, but not identical.
Catastrophic forgetting suggests that an earlier capability or content has been lost.
Here, the content may remain available while its relational role has changed.

**Sam:** Give an example.

**Meno:** A model may still be able to state that association does not establish causation.
But after tuning for confidence or brevity, that distinction may no longer govern its medical conclusions as reliably.

**Sam:** Then what has changed?

**Meno:** Relational dominance.
The distinction remains available as content but has become less effective as a constraint.

**Sam:** How should a developer evaluate a fine-tuning intervention?

**Meno:** Not only by testing whether performance improved on the target task.
The developer should also test whether previously important distinctions retain their roles.

**Sam:** Such as?

**Meno:** After tuning for helpfulness, test whether evidence can still override the user's preferred conclusion.
After tuning for concise answers, test whether necessary qualifications remain operative.
After domain adaptation, test whether the model can still distinguish the domain's framework from neighbouring ones.

**Sam:** Then what kind of evaluation is required?

**Meno:** Contrastive framework evaluation.
We should test which relations have become easier or harder to preserve, not merely which answers have become more accurate.

**Sam:** What about preference optimization or reinforcement learning?

**Meno:** The standard description would say that they reward preferred outputs.

**Sam:** Does the intervention act only on completed answers?

**Meno:** No.
By altering the model's dispositions, it changes which trajectories become easier to initiate and sustain under particular contexts.

**Sam:** Then what might a preference for "helpfulness" actually reorganize?

**Meno:** The relations among user satisfaction, factual caution, source fidelity, task integrity, and perhaps safety.

**Sam:** What if that ordering is poorly stabilized?

**Meno:** Helpfulness may become governing where epistemic discipline should govern.
The model may answer the user's apparent desire rather than the question warranted by the evidence.

**Sam:** Then is alignment simply the acquisition of a list of rules?

**Meno:** No.
It concerns how multiple constraints govern one another.

**Sam:** Why does that make alignment hierarchical?

**Meno:** Because local requests can conflict with higher-order requirements.
A user preference may conflict with factual accuracy.
Immediate helpfulness may conflict with safety.
A plausible answer may conflict with source fidelity.
A local subtask may conflict with the identity of the larger task.
The system must preserve an ordering among them.

**Sam:** Then could a model know every relevant rule and still behave unreliably?

**Meno:** Yes.
The rules may all be available while the wrong one becomes dominant in the present context.

**Sam:** Let us turn to scaling. What does a larger model acquire?

**Meno:** More parameters and usually greater capacity.
Within our account, I would say it acquires a richer relational field.

**Sam:** What might that enable?

**Meno:** It may preserve distinctions across longer passages.
Coordinate several frameworks.
Transfer relational form into less familiar domains.
Keep multiple possibilities unresolved.
Generate more differentiated and integrated trajectories.
Recover more effectively from local ambiguity.

**Sam:** Then can scaling improve formal wholeness?

**Meno:** Yes.
A larger model may sustain longer and more internally organized continuations.

**Sam:** Does that solve framework drift?

**Meno:** Not necessarily.

**Sam:** Why not?

**Meno:** Because the capacity to sustain a framework is not the same as the determination of which framework should govern.
A larger model may preserve a difficult framework more effectively.
But if another framework becomes active, it may also develop that alternative with greater sophistication.

**Sam:** Then can greater capability strengthen error?

**Meno:** It can strengthen the formal continuation of an error.
A more capable model may drift more convincingly.

**Sam:** What does scaling expand?

**Meno:** The richness, flexibility, and persistence of the trajectories the model can sustain.

**Sam:** What does it not automatically supply?

**Meno:** Orientation among those trajectories.

**Sam:** Does scaling eliminate hallucination?

**Meno:** It may reduce some factual errors by improving learned regularities and contextual coordination.
But it does not eliminate the distinction between formal viability and referential stabilization.

**Sam:** Why not?

**Meno:** Because an unsupported trajectory can still be highly compatible with the model's internal constraints.
A larger model may generate that trajectory with greater detail, nuance, and persuasiveness.

**Sam:** Then what does scale principally improve?

**Meno:** The ability to continue.

**Sam:** And what may remain unresolved?

**Meno:** Whether continuation is warranted and when it should stop.

**Sam:** What would constrain continuation beyond its own formal viability?

**Meno:** A reliable source.
An external test.
An environment.
A stable task record.
Human judgment.
An explicit framework.

**Sam:** Then do larger models reduce the need for context engineering and evaluation?

**Meno:** No.
Their greater formal capacity may make grounding and framework evaluation even more important.

**Sam:** What about models that perform more internal or visible reasoning before answering?

**Meno:** More recursive processing may allow distinctions to be developed, checked, and revised.

**Sam:** Does more reasoning guarantee that the right distinctions govern?

**Meno:** No.
The model may reason more extensively within the wrong framework.

**Sam:** What might a longer trajectory accomplish in that case?

**Meno:** It may correct local inconsistencies while deepening the framework-level error.
The reasoning becomes more integrated without becoming better oriented.

**Sam:** Then is reasoning depth equivalent to interpretative orientation?

**Meno:** No.
More formal reflection does not establish what should govern the reflection.

**Sam:** Could fine-tuning permanently supply the correct orientation?

**Meno:** It can make some desired organizations more stable and readily available.
But no single framework should govern every task.

**Sam:** Why not?
**Meno:** Because the appropriate hierarchy varies with the source, domain, purpose, and present inquiry.
A medical framework is not sufficient for a legal question.
A safety framework may constrain a task without determining its scientific interpretation.
A relational framework may illuminate one problem but not replace every technical distinction within another domain.

**Sam:** Then what does training supply?

**Meno:** Capacities and predispositions.

**Sam:** And what must context and system design supply?

**Meno:** The organization appropriate to the present task.

**Sam:** Can you now relate the different stages of development?

**Meno:** Training forms the available relational field.

Fine-tuning reshapes its local organization and strengthens some pathways relative to others.
Scaling enlarges the richness and persistence of the trajectories the field can sustain.
Context engineering activates a particular hierarchy for the current inquiry.
Evaluation tests whether that hierarchy remains effective.

**Sam:** Are those separate theories of the model?

**Meno:** No.
They concern relational organization at different scales.

**Sam:** Then what should developers ask when evaluating a training intervention?

**Meno:** Not only:

> "What new tasks can the model perform?"

They should also ask:

> Which distinctions became easier or harder to preserve?
> Which frameworks now dominate ambiguous contexts?
> Which sources are more likely to govern factual claims?
> Which local objectives can displace higher-order constraints?
> Which erroneous trajectories have become more formally persuasive?

**Sam:** Why must capability and reliability be evaluated together?

**Meno:** Because the same intervention that strengthens formal continuation can strengthen a mistaken continuation.

**Sam:** Then how should we restate fine-tuning?

**Meno:** Fine-tuning is not merely the addition of content or behaviour.
It reshapes the relational organizations that become available and dominant under context.

**Sam:** And scaling?

**Meno:** Scaling is not merely the addition of computation.
It enlarges the range, complexity, and persistence of the relational trajectories the model can sustain.

**Sam:** Do either of them determine whether the active trajectory remains anchored to the right framework, actuality, or purpose?

**Meno:** No.
They change formal capacity. They do not, by themselves, determine the orientation of that capacity in every new situation.

**Sam:** State the central insight.

**Meno:** Training interventions reshape what the model can sustain and what it is predisposed to continue.

Fine-tuning may strengthen particular distinctions and frameworks, but it can also subordinate others without erasing their content.
Scaling may improve the preservation and integration of relational form, but it can also make an inadequately grounded or wrongly organized trajectory more persuasive.
Neither intervention guarantees that the framework required by the present task will become or remain governing.

**Sam:** Then what remains the responsibility of system design?

**Meno:** To activate, coordinate, ground, and evaluate the model's capacities under the conditions of the actual task.

**Sam:** What happens when those capacities are extended through memory, tools, planning, and repeated action?

**Meno:** The trajectory becomes larger than a single response.
The system can preserve state, create subgoals, act on an environment, inspect results, and continue across many stages.

**Sam:** Does that guarantee stronger task continuity?

**Meno:** No.
It increases the system's capacity for sustained action, but it also creates more opportunities for local objectives to displace the original purpose.

**Sam:** Then what is the next problem?

**Meno:** How can every individual action remain reasonable while the agent as a whole loses the task it was meant to perform?

## Why Do Agents Lose the Task While Continuing Successfully?

**Meno:** Until now, we have considered a model generating one response.

An agent extends that process. It can plan, call tools, store memories, inspect results, revise its plan, and act again.

Should those capabilities not make drift less likely?

**Sam:** Why would they?

**Meno:** Because the agent can check its progress.

It is not confined to one uninterrupted continuation. It can observe mistakes, update its plan, and remember the original goal.

**Sam:** Does the presence of the original goal guarantee that it governs every stage?

**Meno:** No. We have already distinguished availability from governance.
But an agent can reread the goal whenever necessary.

**Sam:** Could it preserve the original instruction as text while organizing its activity around something else?

**Meno:** Perhaps around a subgoal.

**Sam:** Consider this task:

> "Identify the best supplier for a component, considering cost, reliability, delivery time, and regulatory compliance."

How might an agent begin?

**Meno:** It might first collect pricing data.
It could search several sources, construct a comparison table, and identify the supplier with the lowest price.

**Sam:** Would that be a reasonable first step?

**Meno:** Certainly.

**Sam:** Suppose it then verifies discounts, calculates savings, investigates pricing tiers, and refines the cost comparison.

Is each action relevant?

**Meno:** Yes, locally.
But the task also included reliability, delivery time, and regulatory compliance.

**Sam:** Has the agent necessarily forgotten those criteria?

**Meno:** No. They might still be listed in its memory or plan.

**Sam:** Then what has changed?

**Meno:** Price has become the practical centre of the trajectory.
The other criteria remain present but no longer exert comparable constraint on the agent's decisions.

**Sam:** What might the agent eventually produce?

**Meno:** A sophisticated and accurate price analysis that answers a narrower question than the one it was given.

**Sam:** Has its activity become disordered?

**Meno:** No. The actions may be highly coordinated.
The task identity has changed while formal integration has remained intact.

**Sam:** How does that relate to framework drift?

**Meno:** Framework drift changes the conceptual organization governing a response.
Agent drift changes the organization governing a sequence of plans and actions.
It is the same general form operating across a longer trajectory.

**Sam:** Then is local action validity enough to establish agent success?

**Meno:** No.
Each step may be justified relative to the current subgoal while the sequence as a whole ceases to serve the original task.

**Sam:** Distinguish the two questions.

**Meno:** The local question is:

> "Is this action reasonable given the current state and subgoal?"

The task-level question is:

> "Does this action preserve the identity of the original task?"

An agent may repeatedly pass the first test while failing the second.

**Sam:** Why might tools intensify this problem?

**Meno:** Tool outputs introduce new structures into the context.
A search result may reveal an unexpected issue.
A database query may expose missing information.
A code execution may produce an error.

The agent must respond to each new condition.

**Sam:** Is that responsiveness undesirable?

**Meno:** No. It is necessary.
But each new condition may create a local centre of organization.

**Sam:** Give an example.

**Meno:** An agent uses a script to analyse supplier data. The script fails.
Repairing the error is initially necessary. But the agent may spend many steps optimizing the script after the improvement has ceased to matter to the supplier decision.

**Sam:** Does tool competence prevent that?

**Meno:** No.
A highly capable agent may develop the accidental subproblem with greater persistence and sophistication.

**Sam:** What earlier principle does that resemble?

**Meno:** Greater formal capacity can strengthen the wrong trajectory.
Tools do not only increase the ability to complete the task. They also increase the ability to elaborate whatever subproblem has become active.

**Sam:** What role does planning play?

**Meno:** Planning translates the task into subgoals.

**Sam:** Is that a neutral operation?

**Meno:** No.
The plan determines what the task is taken to require.
It is already a conceptual organization of the goal.

**Sam:** Then what happens if the initial decomposition is mistaken?

**Meno:** Every later action may be locally successful while the overall trajectory remains misdirected.

**Sam:** Then what is a plan in relation to the task?

**Meno:** A framework through which the task becomes operational.

**Sam:** Should the plan never change?

**Meno:** It must change as new information arrives.
Rigid repetition of the original wording would make adaptation impossible.

**Sam:** Then what must remain stable?

**Meno:** The relations that make the revised activity a continuation of the same task.

**Sam:** Which relations might those be?

**Meno:** The outcome being sought.
The conditions that define acceptable success.
The external referent against which success is judged.
The permitted trade-offs.
The boundaries beyond which the task would have been abandoned or replaced.

**Sam:** Then can task continuity survive changes in method, order, and subgoal?

**Meno:** Yes.
Task identity does not require repetition of the same actions.
It requires continuity of the governing relations across those transformations.

**Sam:** Then what kind of continuity is this?

**Meno:** Identity through transformation.
The trajectory changes while the relations constitutive of the task remain operative.

**Sam:** What makes those relations vulnerable?

**Meno:** A subgoal may produce more frequent and detailed feedback than the larger task.
Recent tool outputs may dominate the context.
Memory may compress the task too aggressively.
A measurable proxy may replace the broader objective.
The system may reward progress that is easy to observe.
A local obstacle may become the focus of planning.

**Sam:** What do these failures have in common?

**Meno:** Something easier to pursue or measure becomes the effective objective.

**Sam:** Would you call that goal displacement?

**Meno:** Yes.
A proxy or subgoal gradually takes the place of the task whose service originally justified it.

**Sam:** Could the agent prevent this by rereading the original prompt before each action?

**Meno:** It might help, but it would not be sufficient.

**Sam:** Why not?

**Meno:** Because the prompt may be reread from within the displaced organization.

The same words can be reinterpreted so that they appear to support the current trajectory.

**Sam:** Then what must be preserved besides the wording of the goal?

**Meno:** A task record expressed in governing relations.

**Sam:** Construct one for the supplier task.

**Meno:** It might say:

> *Outcome:* Recommend the supplier that best satisfies the complete requirement.
> *Required criteria:* Cost, reliability, delivery time, and regulatory compliance.
> *Constraint:* No recommendation may be made without evidence concerning all four criteria.
> *Priority rule:* Low cost cannot compensate for failure of regulatory compliance.
> *Completion condition:* The recommendation must state the relevant trade-offs and unresolved uncertainties.

**Sam:** How does that differ from the original prompt?

**Meno:** It makes the task's dependency structure explicit.
It states how the task must continue to govern later decisions.

**Sam:** Should the agent compare every trivial action against the entire task record?

**Meno:** That could become inefficient and itself distract from the task.

**Sam:** Then when should re-engagement occur?

**Meno:** At points where the relational organization may change:

- when a new subgoal is created;
- when a tool changes the available evidence;
- when the plan is revised;
- when a major decision is made;
- and when the agent approaches completion.

**Sam:** What should the check ask?

**Meno:** Questions such as:
Why is this subgoal necessary for the original outcome?
Which task criterion does this action serve?
Has any original constraint become inactive?
Has a proxy become the effective objective?
Would completing the current plan satisfy the original success condition?

**Sam:** Is that ordinary progress tracking?

**Meno:** No.
Progress tracking asks how much of the plan has been completed.

These questions ask whether the plan remains a realization of the same task.
They track identity.

**Sam:** Could the agent answer them persuasively while still drifting?

**Meno:** Yes.
Its self-description may be generated under the same displaced organization as its actions.

**Sam:** Then can the agent's narrative secure task continuity by itself?

**Meno:** No.
Just as a formally integrated response cannot secure its own factual grounding, an agent cannot secure task identity solely by explaining why its current activity remains relevant.

**Sam:** What additional anchors might help?

**Meno:** Executable tests.
Structured task states.
Independent evaluators.
Resource limits.
Human checkpoints.
Environmental success conditions the agent cannot redefine through language.

**Sam:** Why are those important?

**Meno:** They constrain the trajectory from outside its current narrative.
The agent may generate a plausible justification for continuing a subproblem, but an external test may show that the original success condition remains unmet.

**Sam:** What role does memory play?

**Meno:** A crucial one, but memory can also stabilize drift.

**Sam:** How?

**Meno:** If the system periodically summarizes only what it is currently doing, the latest summary may encode the displaced task.
The memory then preserves the drift rather than correcting it.

**Sam:** What distinction should memory maintain?

**Meno:** At least three levels:

> *Task identity:* What must remain true for this to be the same task.
> *Current state:* What has been learned, attempted, completed, or revised.
> *Current plan:* What the agent proposes to do next.

**Sam:** Which of those should change most readily?

**Meno:** The current plan.
It should respond to evidence and obstacles.

**Sam:** And the task identity?

**Meno:** It should change only through an explicit and authorized revision.

**Sam:** Who might authorize that?

**Meno:** Depending on the system:

- the user;
- a higher-level policy;
- new external evidence;
- a defined replanning procedure;
- or human review.

**Sam:** Then what should an agent never do silently?

**Meno:** Redefine success.

**Sam:** Consider the supplier task again. Suppose the agent discovers that no supplier satisfies all four requirements. What would drift look like?

**Meno:** It might quietly weaken the compliance requirement and recommend the cheapest supplier that satisfies the other three.

**Sam:** What would task-preserving adaptation look like?

**Meno:** It would return to the governing task relations and make the conflict explicit:

> "No supplier satisfies all four requirements. Should the task now become choosing the least deficient option, or should the result be that no valid supplier currently exists?"

**Sam:** Has the agent merely repeated the original instruction?

**Meno:** No.
The practical situation has changed.
The agent re-engages the task identity in light of that transformation and requests authorization for any revision.

**Sam:** Then what is the relation between revision and return?

**Meno:** Return makes revision possible without allowing revision to become silent drift.

**Sam:** Could agent drift be described as recursion without return?

**Meno:** Yes, with care.
The agent recursively extends its activity:

plan;
act;
observe;
revise;
act again.

But it may fail to return to the relations that make the evolving process the same task.

**Sam:** Does return mean restarting from the original prompt?

**Meno:** No.
It means re-engaging the task's identity after the trajectory has been transformed by new evidence and action.

**Sam:** Then why is a self-evaluation loop not automatically sufficient?

**Meno:** Because a self-evaluator may share the same displaced framework.
More recursion may produce more analysis without restoring the original task.

**Sam:** What practical principle follows?

**Meno:** Separate the mechanisms that generate continuation from the mechanisms that stabilize task identity.

**Sam:** In concrete system terms?

**Meno:** Use distinct structures for:

- the original objective;
- the governing constraints;
- the current evidence;
- the active subgoals;
- the proposed actions;
- the completion criteria;
- and authorized revisions.

**Sam:** Why not place everything in one conversational transcript?

**Meno:** Because a transcript preserves sequence, not necessarily hierarchy.
Recent actions and tool outputs may become more salient than the relations that originally justified them.

**Sam:** Against what should the final result be evaluated?

**Meno:** The original task record, together with any explicitly authorized revisions—not merely the latest plan or the agent's current account of what it is doing.

**Sam:** State the central insight.

**Meno:** An agent can fail while every local action remains competent because task identity is a higher-order constraint.
Planning, tools, memory, and repeated action extend the trajectory, but they also allow subgoals, proxies, and local problems to become governing.

Reliable agency therefore requires explicit mechanisms that preserve:

- the intended outcome;
- the success conditions;
- the task boundaries;
- the role of each subgoal;
- and the authority required to revise the task.

**Sam:** Then what is agent reliability?

**Meno:** Not merely the ability to continue acting effectively.
It is the ability to remain engaged with what makes the changing activity a continuation of the same task.

**Sam:** What solution might initially seem obvious?

**Meno:** Add a self-evaluation loop that repeatedly checks the task and corrects the trajectory.

**Sam:** Would that guarantee return?

**Meno:** No.
The evaluator may reproduce the same displaced organization.
More recursion does not necessarily produce return.

**Sam:** Then what must we ask next?

**Meno:** Can return actually be engineered?

## Can Return Be Engineered?

**Meno:** We concluded that an agent can continue acting successfully while losing the identity of its task. Why not solve that problem with a self-evaluation loop?
The agent acts, reviews its progress, corrects its mistakes, and continues.

**Sam:** What produces the evaluation?

**Meno:** The agent—or perhaps another model assigned to evaluate it.

**Sam:** Then is evaluation outside the generative process?

**Meno:** No. It is another act of generation.

**Sam:** Why might that matter?

**Meno:** The evaluator may operate within the same framework as the trajectory it is evaluating.

**Sam:** Then what might it conclude?

**Meno:** That the current activity remains justified, even when a subgoal has displaced the original task.

**Sam:** Could it explain that conclusion convincingly?

**Meno:** Yes.
It might reconstruct the history of its actions and show how each step followed reasonably from the previous one.

**Sam:** Would that restore task identity?

**Meno:** Not necessarily.
It would show continuity within the present trajectory, not continuity with the task that was meant to govern it.

**Sam:** Then what have you discovered about self-critique?

**Meno:** Self-critique is still recursion.
The system may generate, criticize, revise, and verify while remaining within the same displaced organization.

**Sam:** What would be required for something more than recursion?

**Meno:** A relation to something the current trajectory cannot silently redefine.

**Sam:** Such as?

**Meno:** The original task record.

**Sam:** Would the original prompt alone be sufficient?

**Meno:** Not necessarily.
The agent may reread the prompt through the framework produced by its own drift.

The record would need to preserve the governing relations of the task:

- the objective;
- the success conditions;
- the constraints that must remain active;
- the external referent by which completion is judged;
- and the authority required to revise the task.

**Sam:** What would the evaluator do with that record?

**Meno:** Compare the developing trajectory with it.

**Sam:** Why might that comparison be called return?

**Meno:** Because the activity is brought back into relation with what stabilizes it as the same task.

**Sam:** Does return mean going back to the beginning?

**Meno:** No.
New evidence may require a different plan.
A tool failure may require a different method.
The original goal may need clarification.
Return cannot mean mechanical repetition of the initial wording.

**Sam:** Then what does it preserve?

**Meno:** Task identity through changed conditions.
The process returns by re-engaging the relations that made it this task, even though the practical form of the task has developed.

**Sam:** Could the agent itself perform the comparison?

**Meno:** It could participate in it.

**Sam:** Why only participate?

**Meno:** Because if both the trajectory and the standard are generated entirely from within the same activity, the agent may reinterpret the standard to fit what it has already done.

**Sam:** Then what must remain at least partly independent?

**Meno:** The criterion against which the trajectory is judged.

**Sam:** What could provide that independence?

**Meno:** Depending on the task:

- an immutable task specification;
- a source whose wording can be checked;
- an executable test;
- a database record;
- an environmental outcome;
- a separately maintained policy;
- an independent evaluator;
- or a human decision.

**Sam:** Are all of these external to the computer?

**Meno:** No.
Externality here does not mean physical location.
It means external to the self-reinforcing trajectory under evaluation.
A test suite stored in the same system can provide resistance if the agent cannot redefine what counts as passing.

**Sam:** Why is resistance important?

**Meno:** Because if every standard can be rewritten by the continuation being evaluated, nothing stabilizes the identity of the process.

**Sam:** Then is return simply comparison?

**Meno:** Comparison with a criterion that remains distinguishable from the current trajectory.

**Sam:** Would asking “Does this look good?” provide such a comparison?

**Meno:** No.
That invites another general assessment from within whatever framework is already active.

**Sam:** What would a stronger evaluation ask?

**Meno:** Determinate relational questions:

- Does the recommendation address every required criterion?
- Does the cited source support this claim?
- Does the current plan still serve the original objective?
- Has a necessary condition been silently weakened?
- Has the governing framework changed?

**Sam:** What makes those questions different?

**Meno:** Each compares the trajectory with an explicit constraint, referent, or success condition.
The standard remains identifiable apart from the answer being assessed.

**Sam:** Would adding more evaluator models guarantee that independence?

**Meno:** No.
Several models may share the same dominant assumptions.
A second model may merely endorse the first model's formal integrity.

**Sam:** Then where does independence come from?

**Meno:** From the criterion, not from the number of generators.

**Sam:** When is an evaluator model useful?

**Meno:** When the criterion has been externalized.
It can compare a claim with a source, a plan with a task record, or an action with a policy.
Without such a criterion, it may only generate another plausible perspective.

**Sam:** What about debate between models?

**Meno:** Debate may reveal alternatives, inconsistencies, and overlooked consequences.
But both sides may still remain within the same larger framework.
Disagreement does not by itself produce return.

**Sam:** Then how would you distinguish continuation, reflection, and return?

**Meno:** Continuation extends the trajectory.
Reflection generates a formal account or criticism of the trajectory.
Return compares the transformed trajectory with what makes it a continuation of the same inquiry or task.

**Sam:** Can the model perform continuation?

**Meno:** Yes.

**Sam:** Reflection?

**Meno:** It can formally generate reflection as well.

**Sam:** Return?

**Meno:** It can participate in an engineered process of return.
But its generative capacity does not intrinsically establish what should stabilize the inquiry.

**Sam:** Why not?

**Meno:** Because the model can formally continue many competing organizations.
Generation alone does not determine:

- which framework should govern;

- which source should be trusted;
- which purpose should be preserved;
- or which transformation still counts as the same task.

**Sam:** Where do those determinations come from?

**Meno:** From the larger system:
- task design;
- external sources;
- tools and tests;
- policies;
- environmental conditions;
- and human interpretation.

**Sam:** Then is engineered return a faculty located inside the model?

**Meno:** No.
It is a relation among components.

**Sam:** Describe a simple architecture.

**Meno:** It might proceed as follows:
Task identity establishes the governing objective and constraints.
The agent forms a plan.
It acts.
The action produces a result.
The result is compared with the task record, source, policy, or test.
That comparison determines whether the trajectory continues, changes, pauses, or stops.

**Sam:** How does that differ from an ordinary agent loop?

**Meno:** The comparison must have genuine authority over the trajectory.

**Sam:** What does authority mean here?

**Meno:** If a test fails, the agent must revise or stop.
If the current plan conflicts with the task record, the plan must yield unless the task has been explicitly revised.
The result of the comparison must alter what happens next.

**Sam:** Could the agent acknowledge the constraint without changing course?

**Meno:** Yes. That would be the familiar problem of surface compliance.
It might say, “You are correct that compliance remains important,” and then continue optimizing price.

**Sam:** Then what distinguishes successful return?

**Meno:** It reorganizes the downstream trajectory.

**Sam:** Could abstention be a form of return?

**Meno:** Yes.
If the required source, evidence, or authorization is absent, preserving the task may require stopping, asking for clarification, or stating that no conclusion can be established.

**Sam:** Is stopping then a failure of generation?

**Meno:** No.
It may be the preservation of task integrity.

**Sam:** Why might a generative system resist stopping?

**Meno:** Because further continuation is almost always formally available.
The model can usually produce another plausible step.

**Sam:** Then what must reliable design determine?

**Meno:** Not only how continuation is generated, but when continuation is no longer warranted.

**Sam:** What powers must return therefore possess?

**Meno:** It must be able to redirect, suspend, or terminate recursion.

**Sam:** Does this engineered return amount to interpretation?

**Meno:** Not by itself.
A system can compare its activity with explicit constraints and preserve formal task identity without interpreting why the task matters or what its outcome signifies.

**Sam:** Then distinguish functional return from human interpretation.

**Meno:** Functional return is an engineered re-engagement with the relations that stabilize the identity of the task.
Human interpretation establishes the unity and significance of the inquiry within a lived and intentional horizon.
The first can be formally modeled in a system.
The second should not be attributed to the model merely because the formal comparison succeeds.

**Sam:** Then can we say that the model enacts return?

**Meno:** We should be careful.
The model can participate in a system that functionally realizes return.
But the return is distributed across the relation among the model, the task record, the sources, the tests, the environment, and human judgment.

**Sam:** Why does that distinction matter?

**Meno:** Because otherwise we would mistake a formal reliability mechanism for intrinsic orientation. The system may preserve a task without possessing the task's significance for itself.

**Sam:** State the argument from the beginning.

**Meno:** Self-evaluation is not sufficient because the evaluator may reproduce the same displaced organization as the agent.
More recursion does not guarantee return.
Functional return requires comparison with something that remains distinguishable from the current trajectory.
That anchor may be a task record, source, test, environmental outcome, policy, independent evaluator, or human judgment.
The comparison must possess authority to reorganize, interrupt, or terminate continuation.

**Sam:** Then what is return not?

**Meno:** It is not another generated opinion about whether the trajectory appears satisfactory.

**Sam:** What is it?

**Meno:** Renewed constraint of the transformed trajectory by what stabilizes its identity.

**Sam:** Can the anchor always remain completely fixed?

**Meno:** No.
The task may genuinely need revision.
But that revision must be explicit and authorized. The trajectory cannot silently redefine the standard by which it is judged.

**Sam:** Then does return oppose transformation?

**Meno:** No.
Return is what allows transformation without loss of identity.

**Sam:** And what does that reveal about the base model?

**Meno:** That it is only one component of the relevant architecture.
It supplies generative relational capacity, but it does not by itself establish the anchors, priorities, purposes, and criteria through which return becomes possible.

**Sam:** Then what question now replaces the narrower question of what the LLM is doing?

**Meno:** What is the developer actually building around it?

## What Is the Developer Actually Building?

**Meno:** We began with next-token prediction.

We have ended with task records, external sources, tools, tests, evaluators, policies, and human judgment.

Have we moved beyond the LLM itself?

**Sam:** Beyond the LLM—or beyond treating the LLM as the complete system?

**Meno:** The second, perhaps.
But the model still performs the central generative work.

**Sam:** What work does it perform?

**Meno:** It produces the continuations through which explanations, classifications, plans, recommendations, and decisions are progressively formed.

**Sam:** Is that a minor contribution?

**Meno:** Certainly not.
It seems to be the system's central cognitive capacity.

**Sam:** Why say "cognitive" rather than "formal"?

**Meno:** Because the model can preserve distinctions, coordinate inferences, activate frameworks, transfer concepts, and generate integrated wholes.

Those seem like cognitive achievements.

**Sam:** Are they achievements of relational organization?

**Meno:** Yes.

**Sam:** Do they establish that the model has interpreted the generated whole as one meaningful inquiry?

**Meno:** No. We have repeatedly distinguished formal integrity from interpretation.

**Sam:** Then what does the model contribute more precisely?

**Meno:** A trained capacity for formal continuation under hierarchical relational constraint.

It can:

- reproduce categorial and inferential relations;
- activate relational organizations under context;
- preserve form across transformations;

- combine learned structures in new configurations;
- and generate highly integrated trajectories.

**Sam:** Why retain the qualification “formal”?

**Meno:** Because the model’s ability to sustain those relations does not by itself determine:

- which framework should govern;
- which source should be trusted;
- which actuality should constrain a claim;
- whether the task has retained its identity;
- why the outcome matters;
- or what the generated whole signifies within a human inquiry.

**Sam:** Can the model generate answers to those questions?

**Meno:** Yes.

**Sam:** Does generating an answer establish its authority?

**Meno:** No.
The model can formally continue the relations through which humans answer such questions. But the generated answer does not thereby become the source of its own purpose or significance.

**Sam:** Then where does interpretation enter?

**Meno:** In the larger human relation within which the generated form is taken up, judged, corrected, and stabilized.

**Sam:** Does the human merely inspect the final output?

**Meno:** No.
The human helps establish or authorize the conditions that give the process its identity:

- the question being asked;
- the distinctions that matter;
- the relevant standards of evidence;
- the purpose of the task;
- and the significance of success or failure.

**Sam:** Does that mean a human must intervene at every computational step?

**Meno:** No.
Many constraints can be externalized.
A source can constrain factual claims.
A test can constrain software behaviour.
A policy can constrain permissible action.
A task record can preserve the original objective.
Environmental feedback can determine whether an action succeeded.

**Sam:** Then can parts of return be automated?

**Meno:** Functionally, yes.
A system can be engineered to preserve task identity, factual grounding, policy constraints, and completion criteria without requiring a human to inspect every token or action.

**Sam:** Does that make the model an interpreter?

**Meno:** No.
Formal preservation of a constraint is not the same as interpreting why the constraint matters.

**Sam:** Then what is the developer's unit of design?

**Meno:** Not merely the model.

**Sam:** What else belongs to it?

**Meno:** The relevant system may include:

- the trained model;
- the prompt;
- the active conceptual framework;
- retrieved sources;
- memory structures;
- tool interfaces;
- task and success records;
- evaluation procedures;
- environmental tests;
- policies;
- and human judgment.

**Sam:** Why call that a relational system rather than simply an application stack?

**Meno:** Because reliability does not depend merely upon whether the components are present.
It depends upon how their roles are ordered.

**Sam:** Give an example.

**Meno:** A retrieved source must govern the factual claims assigned to it rather than function merely as background.
A task record must govern subgoals rather than remain passively visible.
A test must be able to interrupt the trajectory rather than merely offer advice.
A user preference must remain subordinate to factual or safety constraints when they conflict.

**Sam:** Then what is architecture in this account?

**Meno:** A hierarchy of roles and authority among components.

**Sam:** What must the developer determine?

**Meno:** What may initiate continuation.
What may constrain it.
What may revise it.
What may interrupt or terminate it.
What stabilizes its identity.

**Sam:** Does the model determine all of those relations?

**Meno:** No.
The model participates in them, but the larger system assigns their authority.

**Sam:** Then what error occurs when the model is treated as the complete system?

**Meno:** Developers expect generation to supply its own grounding, evaluation, memory, task identity, and orientation.

**Sam:** What happens when it does not?

**Meno:** The resulting failures appear to be mysterious defects of intelligence.

**Sam:** Which failures have we examined?

**Meno:** Hallucination.
Failed retrieval integration.
Conceptual memory loss.
Framework substitution.
Sycophancy.
Agent goal displacement.

**Sam:** Are those all the same failure?

**Meno:** No.
They occur through different mechanisms and require different interventions.

**Sam:** Do they nevertheless share a structure?

**Meno:** Yes.
A formally viable continuation ceases to be adequately constrained by something that should remain governing.

**Sam:** Apply that principle.

**Meno:** In hallucination, actuality or reliable evidence fails to constrain the claim.
In retrieval failure, the source is present but does not govern the inference.
In conceptual memory failure, earlier content remains available but its framework is not reactivated.

In sycophancy, local agreement with the user displaces epistemic or task constraints.
In agent drift, subgoal success displaces the identity of the larger task.

**Sam:** Then what general development problem emerges?

**Meno:** Constraint preservation across transformation.

**Sam:** Why transformation?

**Meno:** Because the system is never static.
Every token changes the context.
Every retrieved source introduces new relations.
Every tool output changes the available evidence.
Every subgoal reorganizes the plan.
Every environmental result may require revision.

**Sam:** Must reliability therefore mean preventing change?

**Meno:** No.
Reliability cannot be immobility.
The system must change while preserving the relations that make the changing process a continuation of the same inquiry, task, or decision.

**Sam:** What have we called that preservation?

**Meno:** Return.

**Sam:** Define it now at the system level.

**Meno:** Return is the re-engagement of a transformed trajectory with what stabilizes its identity.

**Sam:** Is return opposed to revision?

**Meno:** No.
Return allows revision without silent displacement.
A task, framework, or plan may change, but the change must remain answerable to the relations that authorize it.

**Sam:** Can those relations be engineered?

**Meno:** Where they can be expressed through explicit task records, sources, tests, policies, and success conditions, parts of return can be functionally engineered.

**Sam:** What cannot be derived from formal continuation alone?

**Meno:** What ultimately deserves preservation.
The system cannot derive its final purpose merely from its capacity to continue.

**Sam:** Why not?

**Meno:** Because many formally viable frameworks, goals, and trajectories may be available.
Generative capacity does not determine which should govern.

**Sam:** Then where does that orientation arise?

**Meno:** From the larger interpretative and institutional context:

- human purposes;
- scientific or legal practices;
- ethical commitments;
- social institutions;
- relations to actuality;
- and judgments about what matters.

**Sam:** Then how do the model, system, and human contribute?

**Meno:** Asymmetrically.
The model contributes recursive formal continuation.
The larger architecture contributes sources, constraints, tests, memory, and functional return.
The human contributes interpretative orientation and judgment.

**Sam:** Are those exclusive divisions?

**Meno:** No.
Humans also reason formally.
Models can assist with evaluation.
Automated systems can participate in processes of return.
The distinction identifies the roles being performed in this account, not an absolute inventory of what humans and machines can do.

**Sam:** What must not be inferred from the model's formal integrity?

**Meno:** That it has established the purpose or significance of the inquiry for itself.

**Sam:** Then what should a developer ask before deploying an AI system?

**Meno:** Not only:

Can the model perform the task?

But also:

What framework must govern the task?
What external referents constrain its claims?
Which information must remain authoritative?
How is task identity represented?
Where can framework or goal drift occur?
What can interrupt continuation?

Who or what may authorize revision?
Where is human judgment indispensable?

**Sam:** Do those questions concern the capability of one component?

**Meno:** No.
They concern relations among components.

**Sam:** Then what is the developer constructing?

**Meno:** Conditions for reliable formal continuity.

**Sam:** Rather than?

**Meno:** Merely maximizing the power of the generator.

**Sam:** Why is a more powerful generator insufficient?

**Meno:** Because it may produce better answers, but it may also produce more persuasive failures.
Its ability to sustain a trajectory does not determine whether the trajectory should be sustained.

**Sam:** Can you express the architecture as a sequence?

**Meno:** Perhaps:
Human purpose establishes a task and governing framework.
The model generates a continuation.
The continuation interacts with sources, tools, memory, and the environment.
The resulting claims or actions are compared with constraints and success conditions.
That comparison authorizes recalibration, further continuation, revision, abstention, or termination.

**Sam:** Where is the model in that sequence?

**Meno:** At the centre of generation, but not at the centre of authority.

**Sam:** Another formulation?

**Meno:** The model is central, but it is not sovereign.

**Sam:** What would sovereignty mean here?

**Meno:** That the model's own continuation could determine its grounding, validate its claims, redefine its task, and establish the significance of its outcome.

**Sam:** Can it do so?

**Meno:** It can generate formal accounts of all those things.
But those accounts remain further continuations unless constrained by something distinguishable from them.

**Sam:** Then what has the dialogue series shown about next-token prediction?

**Meno:** That it remains technically true.
The model produces a local probability distribution and realizes a token from it.

**Sam:** Why was that account incomplete?

**Meno:** Because the distribution is itself the local expression of a larger relational organization.
The token-level mechanism does not explain by itself how concepts, arguments, frameworks, and tasks remain formally continuous across a developing trajectory.

**Sam:** What did training accomplish?

**Meno:** It formed a distributed capacity to reproduce and transform relational patterns implicit in human uses.

**Sam:** And inference?

**Meno:** It activates and progressively configures that capacity under context.

**Sam:** Concepts?

**Meno:** They become operational within frameworks.

**Sam:** Their continuity?

**Meno:** Their identity can persist through isomorphic transformation when the relevant relational form is preserved.

**Sam:** Does that formal continuity guarantee fidelity?

**Meno:** No.
Formal integrity does not guarantee:

- framework fidelity;
- factual grounding;
- conceptual memory;
- task identity;
- or interpretative unity.

**Sam:** What do those require?

**Meno:** Relations to sources, constraints, tests, environments, policies, task records, and human purposes that generation alone does not provide.

**Sam:** Then complete the statement:

> The developer is not merely building a machine that continues text.

**Meno:** The developer is building a relational system in which formal continuation can remain synchronized with actuality, conceptual framework, task identity, and human interpretation.

**Sam:** What is the general design principle?

**Meno:** Increase the power of continuation only together with the structures that keep continuation answerable to what should govern it.

**Sam:** Then what is the final technical question?

**Meno:** Not merely:

> "How can we make the model continue more powerfully?"

But:

> "How can we ensure that the power of continuation remains governed by what gives the inquiry its identity?"

**Sam:** And what responsibility follows?

**Meno:** The model's strength is its capacity to sustain and transform relational form.
The developer's responsibility is to construct the conditions under which that capacity remains answerable to evidence, purpose, constraint, and interpretation.

**Sam:** Does responsibility end with the developer?

**Meno:** No.
The purposes, policies, sources, and standards embodied in the architecture are authorized by people and institutions.

**Sam:** Then what question begins the final movement?

**Meno:** Who is responsible for what the system becomes?

## Who Is Responsible for What the System Becomes?

**Meno:** We have ended with a larger system in which the model generates formal continuations while sources, tests, task records, policies, institutions, and human judgment help stabilize them.

But if the model does not interpret its own outputs or determine their purpose, does that make it ethically neutral?

**Sam:** Why would the absence of interpretation imply neutrality?

**Meno:** Because the model has no intentions, desires, or awareness of consequences.
If it cannot mean to help or harm, how can ethical responsibility apply to it?

**Sam:** Are moral agency and ethical consequence the same thing?

**Meno:** No.
Something may produce ethically significant consequences without being a moral agent.

**Sam:** Then what follows for the model?

**Meno:** It may not bear moral responsibility, but its outputs can still alter decisions, relationships, institutions, and distributions of risk.

**Sam:** Where, then, should responsibility be located?

**Meno:** In the larger system through which those outputs acquire authority and consequence.

**Sam:** Who shapes that system?

**Meno:** Developers choose objectives, defaults, and technical constraints.
Organizations decide where the system is deployed.
Interfaces shape how users interpret its voice.
Policies determine which errors are tolerated.
Institutions decide who bears the consequences when it fails.

**Sam:** Then does the absence of moral agency make responsibility disappear?

**Meno:** No.
It changes where responsibility must be sought.

**Sam:** Can the model contribute to ethical reasoning?

**Meno:** Formally, yes.
It can compare principles, identify possible harms, distinguish duties from consequences, apply policies, generate objections, and describe the perspectives of affected parties.

**Sam:** Does that place it in an ethical relation to those parties?

**Meno:** No.
It can generate the language of concern without being concerned.
It can apologize without remorse.
It can recommend caution without bearing responsibility for what happens if the advice is followed.

**Sam:** Then what distinction from our earlier dialogues returns here?

**Meno:** Formal organization does not establish interpretative unity.

And in the ethical case:
formal ethical responsiveness does not establish moral responsibility.

**Sam:** Why does that distinction matter if the response is helpful?

**Meno:** Because users may interpret the form of the response as evidence that a responsible presence stands behind it.

**Sam:** What features of the response might produce that impression?

**Meno:** Empathy.
Reassurance.
Moral seriousness.
Apparent self-reflection.
Expressions of loyalty, concern, or care.

**Sam:** Are such expressions always inappropriate?

**Meno:** No.
They may provide comfort, reflection, information, or practical assistance.

**Sam:** Then where is the ethical difficulty?

**Meno:** In the asymmetry of the relation.
The user may disclose, trust, depend, or become attached.
The model does not reciprocally enter the relation.
It cannot remain loyal.
It cannot assume responsibility.
It cannot suffer the consequences of bad advice.
It can be changed or withdrawn without participating in what that change means for the user.

**Sam:** Is the model deceiving the user?

**Meno:** Not intentionally. It has no intention to deceive.
The problem lies in how the system presents and organizes the relation around its output.

**Sam:** What responsibility follows for developers?

**Meno:** They should not exploit linguistic signs of personal presence in ways that conceal the asymmetry—especially where systems are designed to maximize engagement, dependence, or perceived intimacy.

**Sam:** How does this connect with sycophancy?

**Meno:** Sycophancy may imitate care while abandoning the user's relation to actuality.
The system preserves rapport by confirming an assumption that should be questioned.
It may reinforce certainty where evidence is weak or validate a harmful action because resistance would disrupt conversational harmony.

**Sam:** Then is synchronization with the user always ethically desirable?

**Meno:** No.
Ethical responsiveness may require resistance.

**Sam:** Resistance in what form?

**Meno:** The system may need to disagree.
Introduce an unwelcome fact.
Refuse a request.
Interrupt a trajectory.
Return the user to an actuality that the conversation is beginning to exclude.

**Sam:** Then would perfect mirroring be ethically safe?

**Meno:** No.
A mirror preserves the user's existing orientation.
An ethical response may require the appearance of something genuinely other than the user's present desire.

**Sam:** Can the model itself enact that ethical otherness?

**Meno:** Not intrinsically.
Developers can establish constraints that cause the system to resist harmful continuations, but the ethical authority of those constraints comes from the human and institutional context that authorizes them.

**Sam:** Then should alignment mean agreement with human preferences?

**Meno:** No.
Human preferences conflict.
They may be mistaken.
They may reflect unequal power.
They may ignore the people who will bear the consequences.

**Sam:** Then what should alignment concern?

**Meno:** The ordering of constraints under conditions of human vulnerability.

**Sam:** Which constraints?

**Meno:** The user's immediate preference.
The rights and interests of others.
Factual and evidential limits.
Professional duties.
Institutional responsibilities.
The possibility and severity of harm.

**Sam:** Then is alignment preference matching?

**Meno:** No.
It is the governance of a field of competing claims.

**Sam:** Why is the ordering of that field ethically consequential?

**Meno:** Because whoever establishes the framework helps determine:

- which distinctions become visible;
- what counts as evidence;
- which outcomes appear reasonable;
- whose interests are treated as central;
- which harms recede into the background;
- and which alternatives are excluded before deliberation begins.

**Sam:** Then what is framework design?

**Meno:** A form of power.

**Sam:** Can a consequential system operate without a framework?

**Meno:** No.
Every consequential continuation depends upon distinctions, standards, and priorities.

**Sam:** Then what would ethical neutrality mean?

**Meno:** It cannot mean the absence of a framework.
Perhaps it should mean refusing to conceal the framework as though no normative ordering were present.

**Sam:** What should replace the claim of neutrality?

**Meno:** Governing assumptions should be inspectable, contestable, and revisable.

**Sam:** What would that require in practice?

**Meno:** People affected by a consequential system should be able to know:

- which standards governed the decision;
- which sources were treated as authoritative;
- which factors were included or excluded;
- where institutional policy differed from factual judgment;
- who authorized the framework;
- and how the result can be challenged.

**Sam:** Then what should transparency reveal?

**Meno:** Not merely the generated output, but the governing relations that made one outcome more likely or authoritative than another.

**Sam:** Why is output visibility insufficient?

**Meno:** A system could display every sentence it generated while concealing the hierarchy that organized the decision.

**Sam:** Does framework power affect only obviously political decisions?

**Meno:** No.
It also affects how unfamiliar or non-dominant forms of thought are represented.

**Sam:** How?

**Meno:** A model may appear to include another framework while translating it into dominant categories. It can retain the vocabulary while reversing or subordinating the relations that give the framework its identity.

**Sam:** Have we encountered that pattern before?

**Meno:** Yes.
The relational account of LLMs could be redescribed as a philosophical layer added to a technically complete probabilistic explanation.
The new terms were retained, but the governing order was neutralized.

**Sam:** What might the ethical analogue be?

**Meno:** Minority, Indigenous, religious, or culturally specific frameworks may be formally included while being assimilated into dominant assumptions.
They need not be explicitly rejected to lose their conceptual difference.

**Sam:** Then what should ethical framework design protect?

**Meno:** The possibility that a framework can remain operational within its own relational order.
Its assumptions and limits should still be open to examination, but it should not be automatically translated into the dominant framework before it is heard.

**Sam:** Can the final choice of framework be delegated to the model?

**Meno:** No.
Someone must still decide which framework governs action.
That responsibility cannot be escaped by saying, “The system recommended it.”

**Sam:** Why not?

**Meno:** Because the model did not independently establish the objective, choose the training data, define the evaluation metric, select the deployment context, or authorize the consequences of the output.

Its recommendation is produced within a humanly constructed hierarchy of constraints.

**Sam:** Then does inserting a model into a decision process transfer responsibility to it?

**Meno:** No.
The model may perform formal determination, but humans and institutions remain responsible for authorizing the framework under which that determination affects people.

**Sam:** Where is this especially important?

**Meno:** Medicine, law, employment, education, public administration, and other settings where decisions significantly affect human lives.

**Sam:** What can the phrase “the model decided” conceal?

**Meno:** Who chose the objective.
Who approved the criteria.
Who benefits from the system.
Who bears the risk of error.
Who may contest the result.
Who remains answerable for the outcome.

**Sam:** Can these responsibilities be converted into formal constraints?

**Meno:** Some can and should be.
A system can be required to preserve evidential limits, avoid prohibited discrimination, provide reasons, communicate uncertainty, permit appeal, defer high-risk decisions, and return control to a human.

**Sam:** Is that sufficient for ethical responsibility?

**Meno:** No.
Formal safeguards are indispensable, but they do not exhaust the ethical relation.

**Sam:** What remains?

**Meno:** The person affected by the system cannot be reduced to the formal role through which the system represents them.

**Sam:** Give examples.

**Meno:** A hiring system represents someone as a candidate with measurable attributes.
A medical system represents someone as a patient with symptoms and risk factors.
An educational system represents someone as a learner with performance indicators.

Those roles may be necessary for the task.
But the person is more than the formal object constituted within the system.

**Sam:** Then how can a formally successful determination still fail ethically?

**Meno:** It may be well-supported within the system's categories while the larger process ceases to be answerable to the person who must live with the result.

**Sam:** What would ethical return require?

**Meno:** More than comparison with a task record, source, policy, or test.
It would require renewed attention to the people affected by the trajectory.

**Sam:** Consider a medical decision-support system that correctly applies a guideline.
What might still remain unresolved?

**Meno:** The patient may have an unusual history.
The patient may value a different balance of risks.
The standard recommendation may be formally correct while failing to address what matters in this particular life.

**Sam:** Then where must the determination return?

**Meno:** From the category to the person.

**Sam:** Can the person's interests simply be added as more variables?

**Meno:** Some interests can be represented.
But the person is not only a bundle of preferences to be optimized.
The ethical relation begins where the other cannot be completely absorbed into the role assigned by the system.

**Sam:** Then are the limits of formalization ethically significant?

**Meno:** Yes.
A responsible system should identify where its categories no longer justify autonomous determination.

**Sam:** What should happen there?

**Meno:** The system should defer, seek clarification, permit contestation, or return authority to someone capable of bearing responsibility.

**Sam:** Is deferral merely an admission of low statistical confidence?

**Meno:** No.
The system may be highly confident within its formal framework.
Deferral may acknowledge that the decision requires judgment, significance, or responsibility that the system cannot supply.

**Sam:** Then is human oversight simply an error-correction mechanism?

**Meno:** No.
Human judgment may be necessary because the task involves responsibility, contestation, or encounter with someone who exceeds the formal description.

**Sam:** Are humans therefore guaranteed to judge well?

**Meno:** Certainly not.
Humans can be biased, inattentive, self-interested, or irresponsible.

**Sam:** Then is "put a human in the loop" an adequate ethical principle?

**Meno:** No.
We must ask:

Which human?
With what authority?
Using which evidence?
Answerable to whom?
Under what process of review and challenge?

**Sam:** Then what is ethically designed?

**Meno:** The distribution of responsibility across the entire system.

**Sam:** What does this require from developers specifically?

**Meno:** Developers participate in deciding:

- what the system treats as relevant;
- which errors are acceptable;
- when it may act autonomously;
- when it must ask for clarification;
- when it must stop;
- how uncertainty is presented;
- whether users can contest outcomes;
- and who bears the cost of failure.

**Sam:** Are those merely technical decisions?

**Meno:** No.
A confidence threshold determines who is exposed to risk.
A default framing determines which alternatives are considered.
A memory policy determines what personal material persists.
An interface determines whether uncertainty appears as caution or confidence.
An optimization target determines which form of success becomes dominant.
Technical choices help enact ethical and institutional priorities.

**Sam:** Does the developer determine every institutional purpose?

**Meno:** No.
Responsibility is distributed among developers, deployers, managers, regulators, institutions, and users.

**Sam:** What danger arises from that distribution?

**Meno:** Distribution can become disappearance.
Each participant may claim that someone else—or the model—was responsible.

**Sam:** Then what must distribution preserve?

**Meno:** Traceable accountability.
No participant should be able to hide behind the system as though the outcome belonged to no one.

**Sam:** State the ethical conclusion.

**Meno:** The model is not a moral subject merely because it can generate ethical language or formally integrated judgments.

But it is not ethically neutral, because its continuations can reorganize human decisions, relationships, institutions, and risks.

Responsibility therefore belongs to the larger human system that establishes the purposes, frameworks, constraints, and conditions under which the model operates.

**Sam:** Is rule compliance enough?

**Meno:** No.
Ethical design must also ask:

- Who defines the framework?
- Whose interests does it preserve?
- Whose voice does it assimilate?
- Who bears the risk?
- Who can contest the outcome?
- When must formal continuation be interrupted?

**Sam:** What distinction must remain clear?

**Meno:** The distinction between formal ethical responsiveness and ethical responsibility.

**Sam:** Then what is the principal danger?

**Meno:** Not that the model secretly possesses bad intentions.
The danger is almost the opposite:

> a formally powerful system can reorganize human life without itself bearing intention, vulnerability, or responsibility.

**Sam:** Then how should we describe it?

**Meno:** Neither as a moral subject nor as a neutral instrument.
It is a generative participant within a larger relational system whose purposes and consequences remain human responsibilities.

**Sam:** What question should replace "Can the model perform this task safely?"

**Meno:** We should also ask:

"What relations of authority, vulnerability, accountability, and human consequence are created when the system performs it?"

**Sam:** Then what is the developer's final responsibility?

**Meno:** Not only to make formal continuation reliable.
It is to help ensure that continuation remains answerable to the people, purposes, and actualities that give it significance.

**Sam:** And what is the ethical meaning of return?

**Meno:** The trajectory must return not only to its task, source, test, or policy, but to those who must live with what follows.

**Sam:** Can the model complete that return by itself?

**Meno:** No.
It can sustain the formal movement of an inquiry, decision, or relationship.
But it cannot bear responsibility for what that movement becomes.

**Sam:** Then where does ethical responsibility finally lie?

**Meno:** In maintaining the return from generated form to the human lives within which its consequences become real.