Conceptual Frameworks and the Technical Interpretation of Token Generation in Large Language Models

Timothy M Rogers, ChatGPT assisted
University of Toronto,
April 25, 2026

Abstract:
This paper argues that the standard probabilistic description of large language models (LLMs) is formally correct but conceptually incomplete. The incompleteness arises from a conflation between the distributions produced by the model and the generative mechanism that produces them. Drawing on the distinction developed in *Concepts Become Operational Only When Their Frameworks Are Activated* (Rogers, 2026), I show that this conflation reflects a deeper dependence of technical explanation on conceptual framework. By distinguishing efficient causality (stepwise token generation) from formal causality (the hierarchical constraint structure encoded in the model), I introduce the notion of the model as a *conditional probability distribution generator (CPDG)*. This distinction reveals that token generation is governed by a distributed, multi-scale constraint system that is obscured by flat probabilistic interpretations. The result is not a rejection of probabilistic descriptions, but a clarification of their limits and the conditions under which they remain explanatorily adequate.

1. Introduction

Large language models (LLMs) are typically described as systems that perform *next-token prediction* using conditional probability distributions. This description is mathematically precise and widely adopted. However, it leaves unresolved a persistent ambiguity: whether the probability distribution itself constitutes the explanatory basis of the model's operation, or whether it is the output of a deeper generative structure.

In *Concepts Become Operational Only When Their Frameworks Are Activated* (Rogers, 2026), I argued that conceptual coherence in the environment of an LLM depends on the activation of a governing framework that determines how distinctions are applied in practice. The present paper extends that claim to the technical interpretation of LLMs. I argue that the dominant probabilistic framework introduces a *level conflation* that obscures the internal organization of the model. This conflation gives the appearance of explanatory completeness while excluding the very categories required to describe the model's governing structure.

2. The Dominant Probabilistic Framework

The standard description of LLMs represents the model as estimating:

$$p(t_{n+1} \mid t_1, t_2, \ldots, t_n)$$

Under this framework:

- token generation is understood as *sequential prediction*
- explanation is framed in terms of *probability, selection, and likelihood*
- the model is treated as a system that selects the next token from a distribution conditioned on prior tokens

This formulation is formally exact, as it follows from the chain rule of probability. However, it introduces a conceptual flattening:

The conditional probability distribution is treated as if it were the causal mechanism itself.

3. The Level Conflation

The flattening arises from the collapse of two distinct explanatory categories:

- *Efficient causality*
  → the stepwise generation of tokens during inference
- *Formal causality*
  → the parameterized, hierarchical constraint structure encoded in the model

In the dominant framework, these are implicitly identified:

The act of sampling from a conditional distribution is taken to fully explain the system's behavior.

This identification obscures the fact that:

- the distribution is *produced* by a structured generator
- the generator encodes *multi-scale constraints* that shape all outputs
- these constraints are not reducible to local token-to-token relations

As a result, the generator disappears into its output, and the hierarchical organization of the model is rendered invisible.

4. The Model as a Conditional Probability Distribution Generator

To resolve this conflation, the model must be understood not as a distribution, but as a *conditional probability distribution generator (CPDG).*

That is, a function:

$$f_\theta : (t_1, \ldots, t_n) \mapsto p_\theta(\cdot \mid t_1, \ldots, t_n)$$

This distinction restores the separation between:

- the *generator* (formal structure)
- the *distribution* (stepwise output)

Under this interpretation:

- token generation remains sequential (efficient causality)
- but is governed by a *distributed, hierarchical constraint system* (formal causality)

This system arises from:

- layered neural parameterization
- attention-based nonlocal coupling
- learned regularities across multiple scales

5. Hierarchical Constraint and Multi-Scale Organization

The CPDG framework reveals that:

- tokens are generated *one at a time,* but not *in isolation*
- each token is conditioned by a state encoding:
  - long-range dependencies

- compositional structure
- global coherence constraints

These constraints operate across multiple scales:
- local (syntax, collocation)
- intermediate (phrases, semantic units)
- global (discourse structure, conceptual coherence)

The result is a system in which *local generation is governed by global organization.* This organization is a property of the model's parameterization and would arise under training on any sufficiently structured sequential domain, not only natural language.

6. Framework Dependence of Technical Explanation
The distinction between CPDG and distribution is not merely terminological. It reflects a deeper principle:

> *Technical explanations are framework-dependent because frameworks determine what distinctions are available.*

Within the dominant probabilistic framework:
- the CPDG distinction appears unnecessary
- hierarchical constraint remains unarticulated

Within the activated framework:
- the same system exhibits a layered causal structure
- probability becomes a *derived description*, not the primary explanatory category

Thus, the apparent sufficiency of the dominant framework depends on the prior exclusion of formal-causal distinctions.

7. Implications
Recognizing the CPDG structure has several implications:

1. *Clarification of causation*
   → separates stepwise generation from governing structure
2. *Recovery of hierarchy*
   → makes multi-scale organization explicit
3. *Reinterpretation of probability*
   → shifts from primary mechanism to constrained output
4. *Alignment with empirical behavior*
   → explains long-range coherence and structured generation

8. Conclusion
The standard probabilistic account of LLMs is not incorrect, but it is incomplete. By collapsing formal causality into efficient causality, it obscures the hierarchical constraint structure that governs token generation.

The distinction introduced here—between the model as a generator and the distributions it produces—clarifies this limitation and demonstrates that:

> *the adequacy of a technical explanation depends on the conceptual framework within which it is formulated.*

This supports the broader claim of Rogers (2026): concepts become operational only when their governing frameworks are activated. In the present case, activating the appropriate framework reveals that token generation is not merely a sequence of probabilistic selections, but the unfolding of a structured, constraint-governed process.

References

Rogers, T. M. (2026). *Concepts Become Operational Only When Their Frameworks Are Activated: An Enactive Account of Conceptual Analysis in Large Language Model (LLM) Interaction.* Zenodo. https://zenodo.org/records/19711998 .