# Authoritative History and the Limits of Autoregressive Intelligence

Flyxion

December 2025

**Abstract**

We present a formal account of why purely autoregressive systems fail to sustain long-horizon coherence, and what architectural structure is required to do so. The central claim is that intelligence requires an authoritative internal history: a deterministic, invariant-preserving record of committed events against which future actions are evaluated. Autoregressive models generate fluent views without commitment and therefore exhibit unavoidable drift under intervention, counterfactual reasoning, or delayed consequences.

We formalize autoregressive generation as a view-only sequential conditioning process and show that, under mild mismatch assumptions, invariant violations accumulate inevitably with horizon length. We then introduce invariant-gated event logs as a minimal constructive substrate for coherence, grounding, and refusal. Finally, we show that world models, structural constraint systems, and event logs are equivalent instances of a single abstract class of invariant-preserving transition systems. Efficiency and automaticity arise through compilation of validated transitions rather than relaxation of constraints. The resulting framework clarifies the architectural limits of scale-driven fluency and provides a principled foundation for planning and safety.

## 1 Introduction

Large autoregressive models exhibit remarkable surface competence across language, vision, and code. They generate coherent text, synthesize images, and produce executable programs. Nevertheless, these systems fail systematically in settings that require stable reasoning across time, explicit intervention, or counterfactual dependence. Common failure modes include brittle tool use, incoherent long-horizon plans, and confident violations of physical, logical, or syntactic constraints.

Such failures persist even as model scale, training data, and optimization procedures improve. This persistence suggests that the limitation is not merely contingent on insufficient capacity or data coverage. Instead, it reflects a structural property of autoregressive generation itself. Autoregressive systems extend sequences by conditioning on prior outputs, producing locally plausible continuations without committing to their consequences. They lack an internal distinction between hypothetical extension and irreversible update.

This paper argues that the absence of authoritative internal history is the critical limitation. Systems that act in the world must distinguish between tentative proposals and committed events. Without this distinction, illegal or incoherent actions cannot be rendered impossible, only unlikely. Over extended horizons, even small probabilities of violation accumulate into near certainty.

The argument proceeds in three stages. First, autoregressive generation is formalized as a view-only process and shown to exhibit unavoidable drift under mild assumptions. Second, deterministic event logs with invariant-gated commitment are introduced as a minimal architectural substrate capable of sustaining coherence, refusal, and counterfactual reasoning. Third, world models, structural constraint systems, and event logs are shown to be equivalent instances of invariant-preserving transition systems, differing only in representation rather than expressive power.

## 2 Formal Preliminaries

**Definition 1** (State, view, commitment)**.** *Let $\Sigma$ denote authoritative internal state. A view is any speculative, derived, or provisional representation $v \in \mathcal{V}$ that has not been admitted into authoritative state. A commitment is an irreversible update to $\Sigma$ that is admitted only if it preserves all invariants.*

**Definition 2** (Invariant)**.** *An invariant is a predicate defining an admissible set $\Omega \subseteq \Sigma$. States outside $\Omega$ are unreachable by commitment.*

**Definition 3** (Transition)**.** *A transition function is a partial map $\delta : \Sigma \times U \rightharpoonup \Sigma$ that is defined if and only if the resulting state lies in $\Omega$.*

These definitions establish a sharp separation between speculative representation and authoritative update. Views may be generated freely, but only invariant-preserving transitions may alter authoritative state.

## 3 World Models as Invariant-Preserving Predictors

**Definition 4** (World model)**.** *A world model is a pair $(g, f)$ where $g : \mathcal{O}^* \to \Sigma$ maps observation histories to internal state and $f : \Sigma \times \mathcal{A} \rightharpoonup \Sigma$ predicts the effect of actions on state. The function $f$ is defined only on invariant-preserving transitions.*

World models are distinguished not by the realism of their outputs but by their counterfactual sensitivity. They support evaluation of hypothetical actions without committing them to authoritative history. This capacity is essential for planning, refusal, and error recovery.

## 4 Autoregressive Generation and Drift

Autoregressive models define conditional distributions $P(x_t \mid x_{<t})$ over observations. Generation induces a Markov process over prefixes, where the effective state of the system is the output history itself. There is no distinguished authoritative state separate from the sequence, and no mechanism for rendering illegal continuations undefined.

**Assumption 1** (Local mismatch)**.** *There exists $\varepsilon > 0$ such that the model's conditional distributions deviate from the data-generating conditionals by at most $\varepsilon$ in total variation.*

**Theorem 1** (View-only drift)**.** *Under local mismatch, divergence between model-generated and admissible sequences grows at least linearly with horizon length, up to saturation.*

*Proof.* Each autoregressive step introduces bounded divergence without contraction. Because no step enforces admissibility categorically, deviations accumulate additively over time. □

This result establishes that invariant preservation cannot be guaranteed by scaling autoregressive models alone.

## 5 Deterministic Event Logs

**Definition 5** (Event log)**.** *An event log is a finite sequence* $\mathcal{L} = (e_1, \ldots, e_T)$ *of atomic events. Authoritative state is derived exclusively by deterministic replay.*

**Definition 6** (Replay)**.** *Fix an initial state* $\sigma_0 \in \Omega$*. Define* $\mathrm{Replay}(\emptyset) = \sigma_0$ *and* $\mathrm{Replay}(\mathcal{L} \cdot e) = \delta(\mathrm{Replay}(\mathcal{L}), e)$ *whenever defined.*

**Definition 7** (Invariant-gated commit)**.** *An event* $e$ *is appended to* $\mathcal{L}$ *if and only if* $\mathrm{Replay}(\mathcal{L} \cdot e) \in \Omega$*.*

**Theorem 2** (Replay-stabilized consistency)**.** *All reachable authoritative states obtained by replay of a committed log satisfy invariants.*

*Proof.* The result follows by induction on the length of the log. □

## 6 Invariant-Preserving Transition Systems

**Definition 8** (Invariant-preserving transition system)**.** *An invariant-preserving transition system is a tuple* $(X, U, \tau, \Omega)$ *where* $\tau : X \times U \rightharpoonup X$ *is a partial transition function defined only on admissible states* $\Omega$*.*

**Theorem 3** (Equivalence)**.** *World models, structural constraint systems, and event logs are equivalent up to representation when formalized as invariant-preserving transition systems.*

*Sketch.* Each architecture induces a partial transition system enforcing admissibility by construction. Differences arise from representational choices rather than expressive power. □

Autoregressive models fail to instantiate an invariant-preserving transition system because they assign nonzero probability to invariant violations rather than rendering them undefined.

## 7 Compiled Replay and Automaticity

Repeatedly validated event schemas may be compiled into cached replay primitives. Such compilation preserves authority within validated contexts while reducing computational cost.

**Proposition 1** (Compiled authority)**.** *Cached replay preserves all invariants enforced by arbiter-mediated validation within its validated context.*

Automaticity is therefore optimized authority rather than heuristic approximation.

# 8 Planning and Safety

Planning is search over hypothetical log extensions.

**Definition 9** (Plan)**.** *A plan is a sequence of events* $\pi$ *such that* $\mathrm{Replay}(\mathcal{L} \cdot \pi)$ *is defined.*

Safety constraints are invariants. They are enforced by impossibility rather than by penalty.

**Proposition 2** (Non-compensability)**.** *No finite penalty can substitute for an invariant over unbounded horizons.*

# 9 Conclusion

Prediction alone is insufficient for intelligence. What distinguishes systems that merely generate from systems that act is the ability to commit, replay, and refuse. Systems that generate views may speak fluently; systems that act must answer to history.

# References

[1] Craik, K. (1943). *The Nature of Explanation.*

[2] Bender, E. M., & Koller, A. (2020). Climbing towards NLU.

[3] LeCun, Y. (2022). A path towards autonomous machine intelligence.

[4] Murphy, E. (2023). ROSE: A neurocomputational architecture for syntax.

[5] Pearl, J. (2009). *Causality.*