Cognitive Dynamics of an Epistemically Constrained Language Model Agent

Chris Hunt livefreecotw@gmail.com
Independent Researcher

Abstract

We present empirical evidence that a language model agent equipped with persistent metacognitive infrastructure—entropy monitoring, tension tracking, memory continuity, and growth vector mechanisms—produces cognitive state trajectories with learnable temporal dynamics at conversational turn granularity. Instrumenting five months of continuous operation (October 2025 through February 2026), we collect 68,110 evaluation ticks aggregated to 2,992 turn-level state snapshots across 545 sessions. A neural predictor trained on these trajectories beats the persistence baseline (predict next state equals current state) by 41.7% in mean squared error under 100-fold session-holdout cross-validation. The strongest signal appears in rare-event entropy components: quality decay (52.8% improvement over persistence), emotional processing (48.8%), and paradox detection (46.3%), all with near-zero lag-1 autocorrelation—indicating genuine transition dynamics rather than simple momentum. Adding user input features to the predictor provides no improvement over internal state alone, suggesting that the agent's cognitive evolution is driven more by autonomous architectural dynamics than by external conversational input. Analysis of intra-turn processing reveals a bimodal adaptive deliberation mechanism: the system operates in either a fast resolution regime ($\sim$6 ticks, 99.9% settlement) or an extended deliberation regime ($\sim$50 ticks, 0.7% settlement), with burst length correlating with tension type transitions at Spearman $\rho = 0.938$. The intermediate range is nearly empty, indicating a sharp phase boundary between processing modes. These findings establish that architecturally augmented agent systems can produce measurable cognitive dynamics and suggest connections to world model frameworks for learned metacognitive monitoring.

Keywords: cognitive dynamics, agent architecture, temporal state prediction, metacognitive monitoring, adaptive deliberation, world models

1 Introduction

Language model agents are increasingly deployed with memory systems and lightweight persistence scaffolding: markdown-based identity files, simple retrieval-augmented generation, and session-level context injection (Park et al., 2023; Shinn et al., 2023; Wang et al., 2024). These approaches enable agents to maintain behavioral consistency across interactions but stop far short of persistent cognitive state monitoring. The agent presents a coherent persona, but no system tracks *how its internal cognitive state evolves* across extended operation.

This gap matters for several reasons. Without empirical characterization of agent cognitive dynamics, we cannot distinguish architectures that produce genuine temporal structure from those that merely simulate continuity through context retrieval. We cannot build learned anomaly detectors that identify novel failure modes from trajectory deviations. And

we cannot assess whether architectural choices—such as principled identity constraint or multi-pass evaluation—produce measurable effects on cognitive state evolution.

Existing work on the dynamics of language model systems operates at different levels of abstraction and timescale. Zhang and Dong (2025) model the latent representations within a single generation pass as trajectories on semantic manifolds. Tacheny (2026) formalize iterative LLM loops as discrete dynamical systems, classifying output-level dynamics as contractive, oscillatory, or exploratory. The agent drift literature (Rath, 2026) quantifies behavioral degradation over extended interactions. None of these efforts empirically measure whether an architecturally augmented agent's *internal* cognitive state—as tracked by its own monitoring infrastructure—exhibits learnable temporal structure across conversational turns over months of operation.

We address this gap by describing and instrumenting Clint, a language model agent where identity is implemented as persistent architectural constraint through a coherent principle framework. The metacognitive infrastructure does not merely store context—it actively tracks, measures, and monitors cognitive state across sessions, performing multi-pass evaluation with adaptive processing depth. This produces cognitive state infrastructure substantially beyond current lightweight approaches.

Our contributions are:

1. We describe an agent architecture where identity is implemented as persistent architectural constraint, producing cognitive state infrastructure beyond current scaffolding approaches (Section 3).

2. We instrument this system and collect internal cognitive state across 2,992 conversational turns (aggregated from 68,110 evaluation ticks) over five months of continuous operation (Section 4).

3. We train neural predictors and demonstrate learnable temporal structure at turn granularity with 41.7% improvement over persistence, with internal state dynamics exceeding external input influence—suggesting that architectural constraint produces autonomous cognitive momentum (Section 5).

4. We identify an emergent bimodal adaptive deliberation mechanism in intra-turn processing, where tension classification instability drives processing depth across two distinct regimes separated by a sharp phase boundary (Section 5.3).

These findings have implications for learned metacognitive monitoring, connections to joint-embedding predictive architecture (JEPA) world model frameworks (LeCun, 2022; Maes et al., 2026), identity coherence as a measurable dynamical property, and adaptive processing as emergent architectural behavior. We discuss these in Section 6.

2 Related Work

2.1 Dynamical Systems Analysis of LLM Behavior

Recent work has begun applying dynamical systems frameworks to language model behavior. Zhang and Dong (2025) introduce the Dynamical Model Evolution Tracker (DMET),

which models LLM generation as trajectories on semantic manifolds using state continuity metrics, attractor clustering, and topological persistence. Their analysis targets the internal transformer representations during a single generation pass. Tacheny (2026) extend this perspective to agentic loops, formalizing iterative LLM processing as discrete dynamical systems in semantic space and classifying dynamics as contractive, oscillatory, or exploratory. Their analysis focuses on the text output of vanilla LLM loops without persistent infrastructure.

Our work differs along three dimensions. First, we analyze persistent cognitive state that spans across conversations over months, rather than within a single generation pass or loop iteration. Second, we measure the dynamics of an architecturally augmented system's internal monitoring state, not the text it produces or the hidden states within a single forward pass. Third, our intra-turn analysis reveals adaptive processing dynamics within the evaluation layer itself—a level of abstraction absent from prior work.

2.2 Agent Drift and Behavioral Monitoring

A growing literature addresses agent drift: the gradual degradation of agent behavior over extended interactions. Rath (2026) quantifies behavioral degradation in multi-agent LLM systems, proposing statistical detection methods and prompt-based remediation strategies. Related work examines persona consistency, instruction following degradation, and value alignment decay over long conversations.

Our work is complementary but directionally opposite. We study the emergence of structured dynamics, not degradation. The architecture described here appears to produce stability and autonomous momentum rather than drift. The prediction experiment provides a methodology applicable to either outcome: architectures producing genuine temporal structure will show prediction improvement over persistence, while architectures experiencing drift will show prediction improvement over the mean baseline but degradation relative to expected trajectory.

2.3 World Models and Self-Supervised Representation Learning

The JEPA framework (LeCun, 2022) proposes learning world models in compact latent spaces by predicting embeddings rather than raw observations. This has been instantiated for images (Assran et al., 2023), video (Bardes et al., 2024), and most recently for end-to-end pixel-based world models by Maes et al. (2026), who demonstrate stable training using the SIGReg regularization approach. SIGReg extends VICReg (Bardes et al., 2022) with a two-term objective that prevents representation collapse, enabling stable JEPA training with as few as 15 million parameters on a single GPU.

Our work establishes the empirical prerequisite for extending this framework to cognitive state spaces. If cognitive state trajectories have learnable temporal dynamics—which this paper demonstrates—then the encoder-predictor-regularizer architecture could in principle apply. We propose this extension as future work in Section 8, grounded by the empirical finding that temporal structure exists. Earlier world model approaches (Ha and Schmidhuber, 2018) focused on learning environment dynamics for planning; our proposed application would learn *self-model* dynamics for monitoring and steering.

2.4 Cognitive Architectures

Classical cognitive architectures such as SOAR (Laird et al., 1987) and ACT-R (Anderson et al., 2004) implement formal state representations with hand-specified transition rules. These systems have well-defined cognitive state spaces by construction, and their dynamics follow from the specified rules.

Our approach differs in that the architecture is hand-built but the dynamics are emergent and empirically measured rather than formally specified. The evaluation system, tension tracking, and decision monitoring were designed individually; the temporal structure observed in their joint evolution, the adaptive processing depth mechanism, and the two-regime deliberation pattern were discovered through empirical analysis rather than designed into the system.

3 System Architecture

3.1 Base Agent Loop

CLINT is a conversational agent built on the OPENCLAW plugin framework, a model-agnostic gateway that routes conversational turns through a configurable plugin pipeline. The system has operated on multiple LLM backends including DeepSeek, GLM, and others via both local inference (Ollama) and cloud APIs.

A critical architectural property is that the underlying language model is *stateless between turns*. The LLM receives a constructed prompt on each turn and produces a response; it retains no internal state across invocations. All temporal structure in the system's cognitive state must therefore originate from the surrounding architectural infrastructure—the plugins that persist, monitor, and inject state into the prompt construction pipeline.

A second critical property is that the system is *model-agnostic*. Over the five-month data collection period, the generative backend transitioned from DeepSeek to GLM-5, with additional fallback paths through other models. The metacognitive infrastructure—entropy monitoring, tension tracking, decision evaluation—remained constant across these transitions. The temporal dynamics reported in Section 5 therefore span multiple LLM backends, providing direct evidence that the observed dynamics are architectural rather than model-dependent. If the temporal structure were an artifact of a particular model's response patterns or latent biases, it would not survive a backend swap. That it does implicates the surrounding infrastructure as the source.

3.2 Metacognitive Infrastructure

The monitoring infrastructure comprises several plugin systems that collectively produce the state dimensions analyzed in this paper:

Entropy monitoring. Component-level tracking computes breakdown scores for correction, novel concepts, emotional processing, paradox detection, metacognitive activity, temporal mismatch, quality decay, recursive meta-cognition, quiet integration, quality assessment, and Shannon entropy. The monitor performs multi-pass evaluation per conversational turn, firing on each internal processing tick.

Decision system. Tracks evaluations from old and new decision pathways with divergence detection, capturing instances where the system's assessment of appropriate action changes between evaluation passes.

Tension tracking. Categorical tension types represent active cognitive conflicts (e.g., `recognition_failure`, `principle_conflict`). These are actively reassessed on each evaluation tick; 82.9% of turn-level processing bursts exhibit tension type changes during evaluation (Section 5.3).

Intervention system. Risk level assessment, grounding pulses (binary stabilization signals), adaptive damping (continuous regulation), and entropy debt tracking (cumulative unresolved entropy) constitute the system's self-regulatory responses.

Continuity plugin. Session tracking, topic monitoring, anchor detection, and memory injection into the prompt construction pipeline provide cross-session persistence.

Growth vectors. Accumulating behavioral principles—both hand-authored and auto-generated from observed patterns. These grow over time, adding constraint to the system's behavioral space through a crystallization process where repeated behavioral patterns become stable traits.

3.3 Identity as Architectural Constraint

The agent's behavioral principles are not prompt-level instructions that reset each session. They are embedded in the growth vector system, reinforced through crystallization where repeated behavioral patterns become persistent traits, and referenced by the continuity plugin during memory injection. This creates a persistent constraint landscape that shapes which cognitive state transitions are likely and which are suppressed.

A key design distinction is that the constraint framework is *epistemic rather than content-prescriptive*. The principles do not specify what the agent should say or believe in particular situations. Instead, they shape *how the agent evaluates and processes*—norms of intellectual honesty, epistemic humility, engagement quality, and self-monitoring rigor. This is the difference between content-level prompt engineering (which makes system behavior responsive to the specific instructions injected) and epistemic-level architectural constraint (which shapes processing patterns regardless of conversational content). A content-prescriptive system's dynamics would be dominated by external input, because the constraints interact directly with input content. An epistemically constrained system's dynamics should be more autonomous, because the constraints shape the processing machinery itself rather than the material it operates on.

The architecture follows a bottom-up epistemic approach: constraints shape processing patterns rather than prescribing outputs, and identity emerges from the accumulated interaction between those constraints and conversational experience rather than from top-down specification. Growth vectors are not authored as a complete identity description; they accumulate from observed behavioral regularities that the crystallization process identifies and stabilizes. The resulting identity is an empirical artifact of operation under constraint, not a predetermined persona.

The hypothesis motivating this design is that principled identity, implemented as consistent epistemic constraint, narrows the space of probable state trajectories and creates

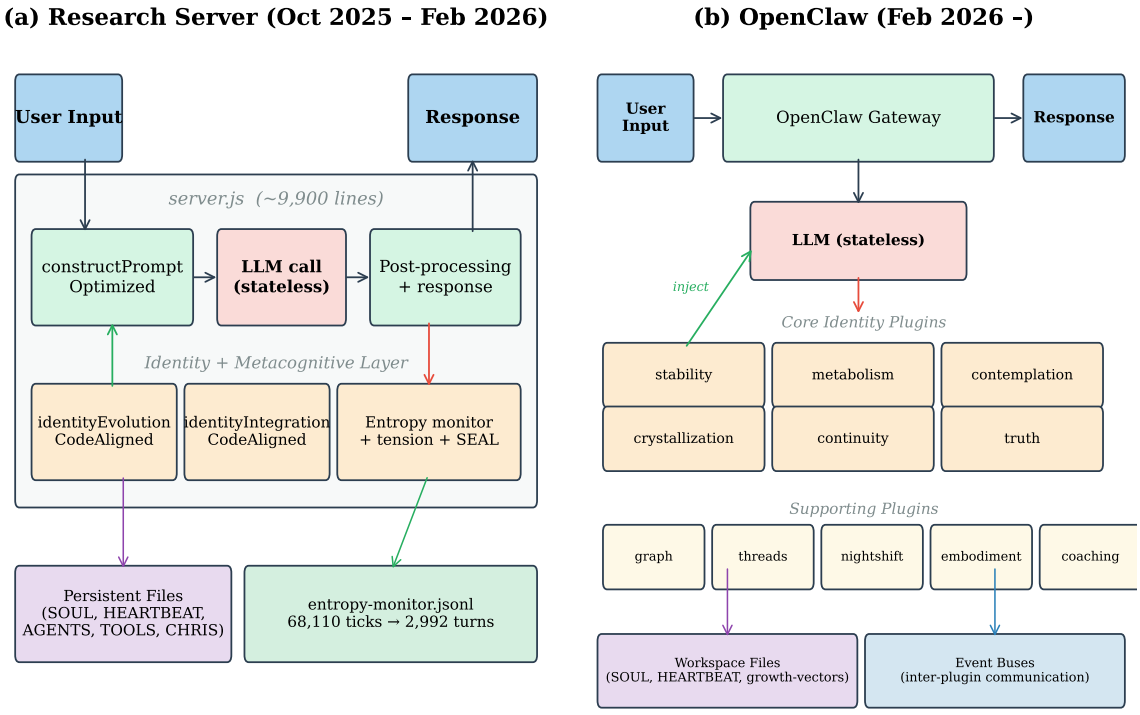

Figure 1: System architecture across both configurations. **(a)** The research server (October 2025–February 2026): a monolithic Node.js process (`server.js`, ∼9,900 lines) containing prompt construction, LLM dispatch, and the identity/metacognitive layer (`identityEvolutionCodeAligned`, entropy monitoring, tension tracking, SEAL metabolism) in a single process. The entropy monitor logs to the JSONL file analyzed in this paper. **(b)** The OpenClaw architecture (February 2026–present): a modular plugin gateway where each metacognitive function is an independent plugin with hook priorities, communicating via event buses. The language model is stateless between turns in both configurations; all temporal structure originates from the surrounding infrastructure.

autonomous momentum that is detectable by temporal prediction. The prediction experiment (Section 5) provides evidence consistent with this hypothesis: internal state dynamics predict future state better than external input features (Section 5.2).

3.4 System History and Data Provenance

The data analyzed in this paper spans two system configurations:

Research server (October 2025–February 2026). The original monolithic architecture where the monitoring systems were developed iteratively alongside the agent's operation. This period produced 68,110 evaluation ticks aggregated to 2,992 conversational turns

with 25 state dimensions per tick, including component-level entropy breakdowns, decision divergence, tension types, and intervention parameters.

OpenClaw architecture (February–March 2026). A modular plugin-based rewrite designed for portability and robustness. The metacognitive suite captures the same phenomena with reduced granularity (∼8 direct dimensions versus 25), reflecting a design prioritizing deployment stability.

The primary experimental results reported here come from the research server dataset due to its richer feature granularity. The OpenClaw architecture was informed by observations from the research server period; plugin designs encode patterns first observed in the monolithic system's data.

4 Experimental Methodology

4.1 Data Collection and Aggregation

The entropy monitor logs a JSON object on each internal evaluation tick, containing: timestamp, entropy total and component breakdown (12 floats), decision fields (old action, new action, divergence flag), context (quality rating, tension type, user message text, response snippet), and intervention parameters (risk level, grounding pulse, adaptive damping, entropy debt).

Examination of inter-tick timing reveals that 95.4% of consecutive entries are separated by less than 5 seconds, with a median gap of effectively zero. The monitor fires on every internal processing pass, not once per conversational turn. We aggregate ticks into turns by grouping consecutive entries separated by gaps of less than 60 seconds into bursts corresponding to single conversational turns.

Turn-level aggregation. For each burst, we take the *final tick* as the resolved turn-level state, representing where the system's assessment settled after deliberation. This yields 2,992 turn-level observations. We select the final tick rather than the burst mean because 82.9% of bursts exhibit categorical state changes during evaluation (notably in tension type); the mean would blur distinct categorical states. The final tick captures the system's resolved assessment entering the next conversational turn.

Session segmentation. Gaps exceeding 30 minutes between consecutive turns define session boundaries, yielding 545 sessions. Synthetic session identifiers are assigned sequentially.

4.2 Feature Vector Construction

Each turn-level observation is represented as a feature vector with the dimensions listed in Table 1.

Four features (`metacognitive`, `temporalMismatch`, `grounding_pulse`, `adaptive_damping`) are constant across the dataset and are excluded from model training, leaving 21 active dimensions. The `shannon` field contains 312 null values (0.5% of entries), imputed with the column median.

Table 1: Feature vector components. All features are normalized to zero mean and unit variance before model training.

Category	Dim.	Features
Entropy components	12	`entropy_total`, `correction`, `novelConcepts`, `emotional`, `paradox`, `metacognitive`, `temporalMismatch`, `qualityDecay`, `recursiveMeta`, `quietIntegration`, `quality`, `shannon`
Decision	3	`decision_old` (int), `decision_new` (int), `decision_divergence` (binary)
Intervention	3	`grounding_pulse` (binary), `adaptive_damping` (float), `entropy_debt` (float)
Context	2	`quality_rating` (ordinal), `tension_type` (int-encoded)
Computed (text)	5	`user_length`, `response_length`, `self_reference_ratio`, `question_density`, `response_to_input_ratio`

4.3 Prediction Models

We evaluate four models for inter-turn dynamics prediction (Experiment A):

Mean baseline. Predicts the training set mean for every test turn. In normalized space, this is the zero vector.

Persistence baseline. Predicts $\mathbf{s}_{t+1} = \mathbf{s}_t$: the next state equals the current state. This is a strong baseline for features with high autocorrelation and represents the null hypothesis that cognitive state carries forward without structured transition dynamics.

Predictor A (temporal). A multilayer perceptron with two hidden layers of 64 units each, ReLU activation, and dropout of 0.1. Input: $\mathbf{s}_t$. Output: $\hat{\mathbf{s}}_{t+1}$. Only consecutive turns within the same session are used; session boundaries are excluded.

Predictor B (input-conditioned). Same architecture as Predictor A. Input: $\mathbf{s}_t$ concatenated with three user-derived features from turn t (`user_length`, `question_density`, `response_to_input_ratio`). Output: $\hat{\mathbf{s}}_{t+1}$.

For intra-turn dynamics (Experiment B), we train the same MLP architecture to predict the final tick state from the first tick state within each burst.

Normalization. All training states (both input states $\mathbf{s}_t$ and target states $\mathbf{s}_{t+1}$) are pooled to compute a single mean $\boldsymbol{\mu}$ and standard deviation $\boldsymbol{\sigma}$, which are applied to both inputs and targets. This shared normalization preserves the near-identity relationship between consecutive states, ensuring that the persistence baseline and the MLP operate on the same scale.

Training. We use the Adam optimizer (Kingma and Ba, 2015) with learning rate 10^{-3} and batch size 256. Training proceeds for up to 100 epochs with early stopping (patience 15 epochs on training loss). All experiments use MPS (Apple Silicon GPU) acceleration.

4.4 Evaluation Protocol

We use leave-one-session-out (LOSO) cross-validation: for each fold, all pairs from one session constitute the test set and all remaining pairs constitute the training set. With 545 sessions (284 having $\geq$3 consecutive pairs), we sample 100 sessions as held-out folds to bound computation. Results are reported as weighted averages across folds, with weights proportional to test set size.

Per-feature mean squared error (MSE) is computed for all four models. We report percentage improvement of each predictor over the persistence baseline, alongside lag-1 autocorrelation for each feature as a sanity check.

5 Results

5.1 Inter-Turn Prediction Performance

Table 2 presents per-feature results for the four models. Predictor A achieves an overall MSE of 0.848 compared to 1.454 for the persistence baseline, representing a **41.7% improvement**. This is substantial learnable structure at conversational turn granularity.

Figure 2 visualizes the per-feature improvement, with features colored by autocorrelation regime. The strongest prediction improvements occur in features with near-zero autocorrelation.

`qualityDecay` (AC = -0.005), `emotional` (AC = 0.026), and `paradox` (AC = 0.093) all show >46% improvement over persistence despite changing unpredictably from one turn to the next under the persistence model. The MLP captures genuine transition structure: the conditions under which quality decay spikes, emotional processing activates, or paradox detection fires are predictable from the broader state context, even though these features do not persist turn-to-turn.

At the other extreme, `tension_type` and `entropy_debt` both have lag-1 autocorrelation exceeding 0.95, and persistence outperforms the MLP for these features. Tension type is nearly constant between turns—it changes *within* turns (Section 5.3). Entropy debt grows monotonically within sessions, making persistence near-optimal.

5.2 Internal State Versus External Input

Predictor B (input-conditioned) achieves 41.6% improvement over persistence, effectively identical to the 41.7% of Predictor A (temporal only). Adding user input features—message length, question density, and response-to-input ratio—provides no additional predictive value beyond internal state.

This result indicates that the agent's cognitive state evolution at turn granularity is driven more by its own internal dynamics than by external conversational input. The next cognitive state is better predicted by knowing where the system *is* than by knowing what the user *said*. This is consistent with the epistemic constraint interpretation described in Section 3.3: because the architectural constraints operate on *how the system processes*

Table 2: Per-feature prediction results (Experiment A). MSE values are computed on normalized features under leave-one-session-out cross-validation (100 folds). *A vs. P* denotes percentage improvement of Predictor A over the persistence baseline. AC(1) is the lag-1 autocorrelation. Features are sorted by prediction improvement. Features marked † are those where persistence outperforms the MLP.

Feature	Mean BL	Persist	MLP-A	A vs. P (%)	AC(1)
qualityDecay	1.089	2.336	1.104	52.8	−0.005
emotional	0.773	1.556	0.797	48.8	0.026
paradox	0.963	1.842	0.989	46.3	0.093
decision_old	1.046	1.892	1.030	45.6	0.081
question_density	1.290	2.388	1.325	44.5	0.081
correction	0.971	1.768	0.992	43.9	0.080
decision_div.	0.991	1.650	0.927	43.8	0.131
quietInteg.	1.239	2.028	1.157	43.0	0.124
user_length	1.041	1.800	1.030	42.8	0.222
self_ref_ratio	0.914	1.536	0.902	41.3	0.201
resp_to_input	0.816	1.296	0.761	41.3	0.281
decision_new	0.965	1.409	0.832	40.9	0.267
response_len.	1.031	1.673	0.997	40.4	0.194
novelConcepts	0.784	1.209	0.737	39.0	0.316
quality	1.026	1.562	0.962	38.4	0.185
quality_rating	1.003	1.465	0.930	36.5	0.239
entropy_total	0.942	1.122	0.761	32.2	0.400
recursiveMeta	0.864	1.011	0.754	25.4	0.472
shannon	1.017	0.852	0.659	22.7	0.537
tension_type†	0.894	0.058	0.064	−9.8	0.962
entropy_debt†	0.557	0.082	0.100	−21.9	0.955
Overall (mean)	**0.963**	**1.454**	**0.848**	**41.7**	—

rather than *what it processes*, the resulting dynamics are shaped by accumulated processing norms rather than by the conversational content flowing through them. The accumulated state (growth vectors, crystallized traits, entropy history) shapes future state evolution independently of immediate external stimulus.

We note that this finding does not imply the system is unresponsive to users. The user's input affects the *current* turn's state (which serves as input to the predictor); what it does not do is provide additional predictive power beyond that state for forecasting the *next* turn.

5.3 Intra-Turn Adaptive Deliberation

Analysis of the 68,110 raw evaluation ticks grouped into 2,992 turn-level bursts reveals a bimodal distribution of processing depth. The distribution of burst sizes is concentrated in two modes with a nearly empty intermediate range:

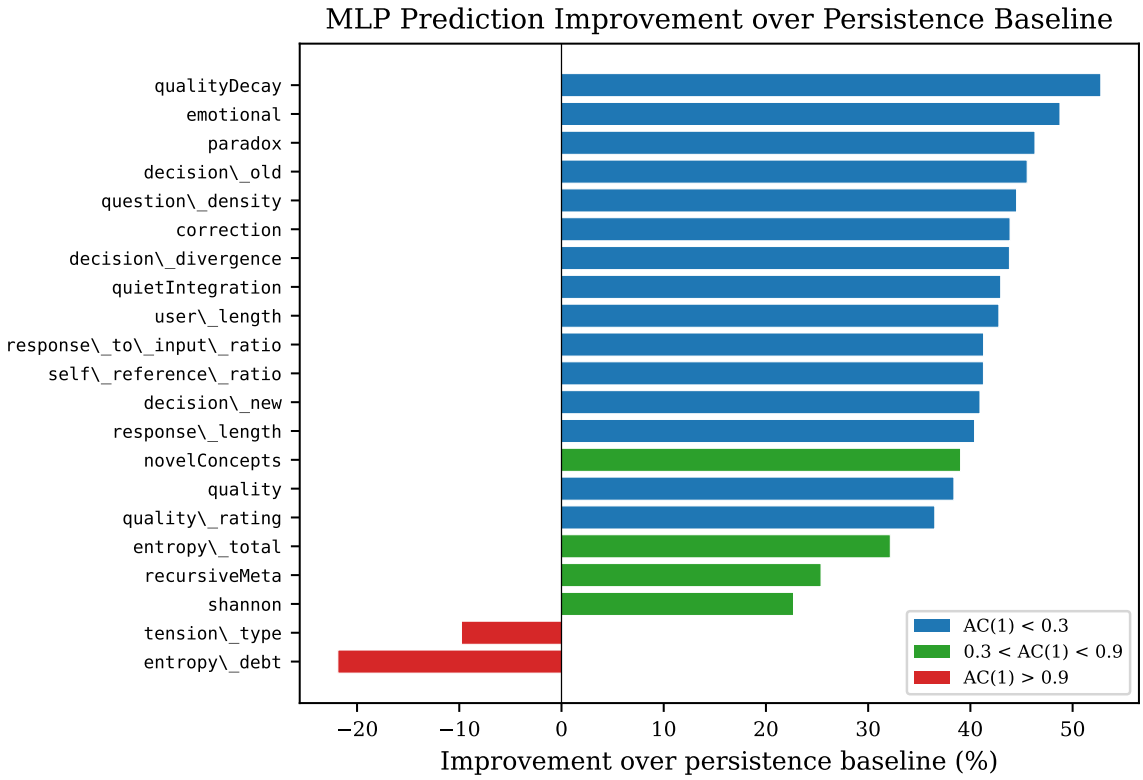

Figure 2: Per-feature prediction improvement of MLP-A over the persistence baseline, sorted by improvement magnitude. Colors indicate lag-1 autocorrelation regime: blue (AC < 0.3), green (0.3–0.9), red (AC > 0.9). Features with near-zero autocorrelation show the largest improvements.

The fast resolution regime comprises bursts of fewer than 10 ticks, accounting for 59.5% of turns. These bursts exhibit a mean of 1.0 tension type transitions, visit 1.7 unique tension categories on average, and reach settlement 99.9% of the time before processing terminates. The extended deliberation regime comprises bursts of 20 or more ticks, accounting for 39.2% of turns. These bursts exhibit a mean of 14.9 tension type transitions, cycle through all four tension types (mean 4.0 unique categories), and reach settlement only 0.7% of the time. The system actively churns through classifications without converging.

The intermediate range (8–30 ticks) contains only 54 bursts (1.8% of turns), indicating a sharp phase boundary between the two regimes rather than a continuous distribution.

Burst length correlates with the number of tension type transitions at Spearman $\rho = 0.938$ ($p < 10^{-100}$). The correlation between burst length and transition *rate* (transitions per tick) is also strong ($\rho = 0.81$), confirming that longer bursts are not simply accumulating transitions passively. Short bursts exhibit a transition rate of 0.13 per tick; long bursts exhibit 0.29 per tick. The system processes at a higher rate during extended deliberation, indicating genuine classificatory instability rather than idle recomputation.

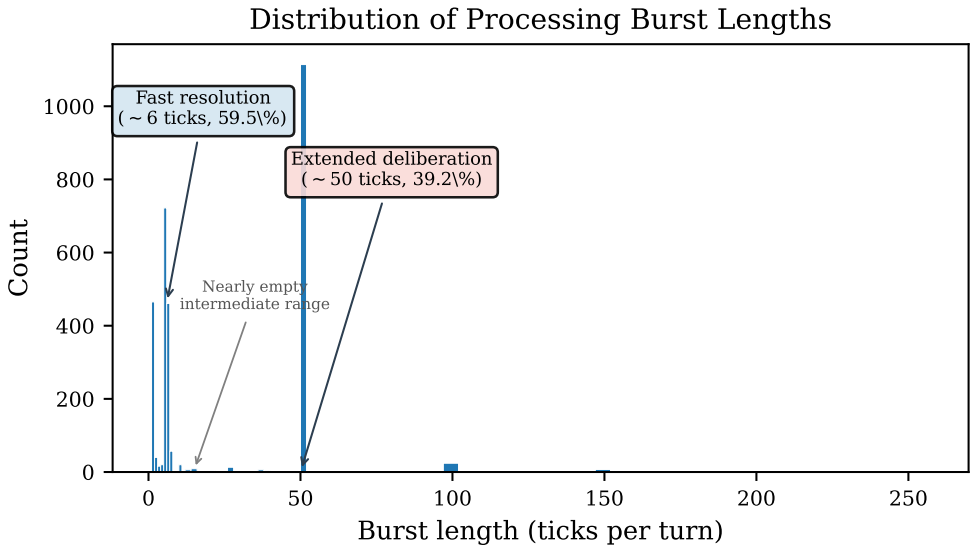

Figure 3: Distribution of burst sizes (ticks per turn). The bimodal structure shows two distinct processing regimes with a nearly empty intermediate range. The fast resolution mode centers around 6 ticks; the extended deliberation mode centers around 50 ticks.

Numerical entropy components show minimal variance across both regimes (mean intra-burst entropy variance: 0.003 for short bursts, 0.003 for long bursts). The adaptive processing depth is driven specifically by categorical tension classification instability, not by numerical score volatility.

The first-tick-to-last-tick prediction experiment (Experiment B) shows that the MLP beats persistence by 7.6% overall for intra-turn prediction, with stronger signal in specific features: `recursiveMeta` (26.5%), `decision_new` (18.1%), `decision_divergence` (16.4%), `tension_type` (12.7%). The intra-turn trajectory is partially predictable from the initial evaluation state.

This bimodal processing structure was not explicitly designed. It emerges from the interaction between the evaluation logic and the heterogeneous properties of conversational inputs—some turns present tractable cognitive states that the evaluator resolves quickly, while others present intractable classification challenges that trigger extended deliberation without convergence.

5.4 Autocorrelation Structure

Lag-1 autocorrelation at turn level reveals three regimes among the active features. Two features exhibit very high autocorrelation (>0.95): `tension_type` and `entropy_debt`. These are effectively constant between turns—tension type because its dynamics are intra-turn, and entropy debt because it accumulates monotonically within sessions.

A cluster of features shows moderate autocorrelation (0.3–0.6): `shannon` (0.54), `recursiveMeta` (0.47), and `entropy_total` (0.40). These exhibit genuine temporal momentum—the cur-

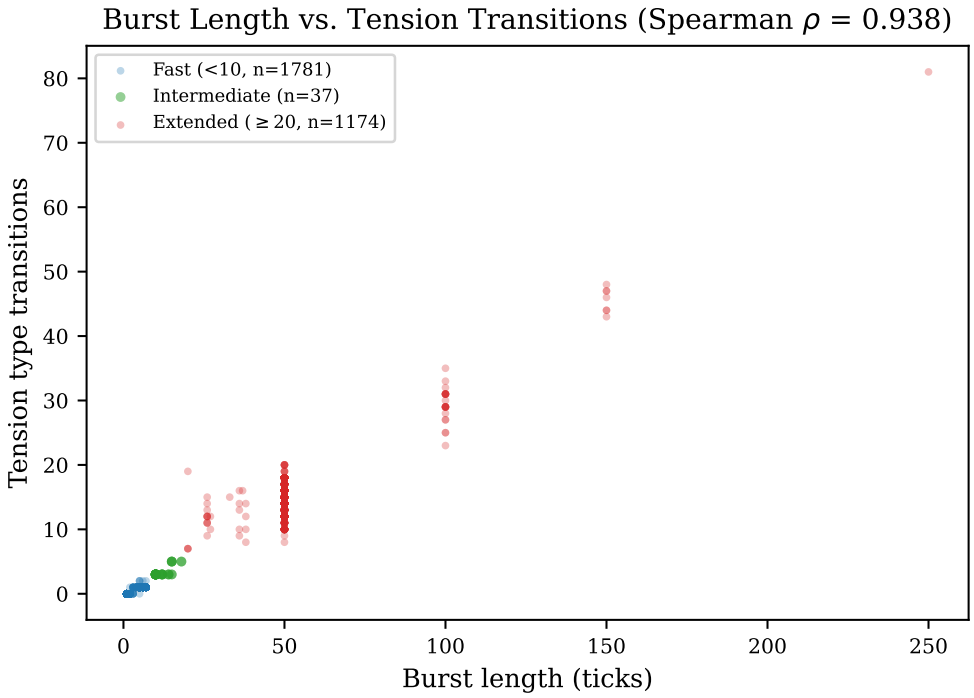

Figure 4: Burst length versus number of tension type transitions within each burst (Spearman $\rho = 0.938$). The two processing regimes are clearly visible as distinct clusters, with the intermediate range nearly empty.

rent value carries predictive information about the next value—while retaining sufficient variability for the MLP to learn non-trivial transition structure.

The majority of features have near-zero autocorrelation (<0.15), including all rare-event entropy components (`qualityDecay`, `emotional`, `paradox`, `correction`), decision features, and text-derived features. These show the strongest MLP prediction improvements, indicating that their activation patterns, while not persistent, are structured: they can be predicted from the joint state even though they carry no individual turn-to-turn momentum.

6 Discussion

6.1 Architectural Scaffolding Produces Genuine Dynamics

The central finding of this paper is that the LLM is stateless between turns, the metacognitive plugins create state, and that state has learnable temporal structure at turn granularity driven by internal dynamics rather than external input. This is an architectural achievement, not an inherent property of language models. The temporal dynamics are emergent: they were not designed into any individual plugin but arise from the interaction between the plugin systems over extended operation.

This emergence is consistent with the bottom-up epistemic design philosophy described in Section 3.3. The architecture does not specify what dynamics should exist—it creates

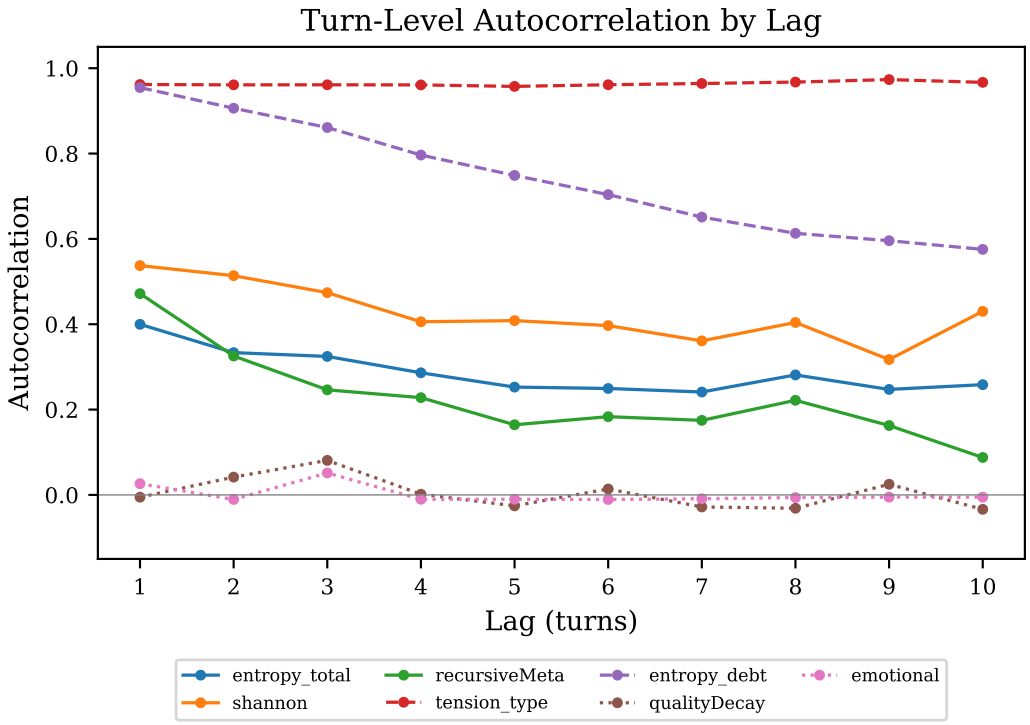

Figure 5: Lag-1 through lag-10 autocorrelation for selected features at turn level. Features cluster into three regimes: high-AC (>0.9, effectively constant between turns), moderate-AC (0.3–0.6, genuine temporal momentum), and low-AC (<0.1, no persistence but learnable transitions).

the conditions under which dynamics can arise by implementing epistemic constraints (processing norms, evaluation standards, self-monitoring requirements) and allowing identity to accumulate from operational experience. The prediction experiment measures what emerged from this approach over five months of continuous operation. That the dynamics are genuine (41.7% over persistence), autonomous (input-conditioning provides no benefit), and model-invariant (spanning a backend transition from DeepSeek to GLM-5) suggests that epistemic scaffolding can produce the kind of stable, structured cognitive evolution that content-level prompt engineering cannot.

The finding that rare-event entropy components show the strongest prediction improvement despite near-zero autocorrelation is particularly notable. These features are not persistent—they do not carry momentum in the simple sense. Yet the MLP learned to predict when they activate based on the broader state context. This indicates multi-feature interaction dynamics: the conditions under which quality decay or emotional processing trigger are encoded in the joint state, even though the individual features do not persist turn-to-turn.

6.2 Two-Regime Processing as Emergent Behavior

The bimodal adaptive deliberation finding demonstrates that emergent behavior can arise in the monitoring layer of an agent architecture, not only in the generative model. The sharp phase boundary between fast resolution and extended deliberation, the elevated transition rate in long bursts, and the near-zero settlement rate under extended deliberation collectively describe a system with distinct processing modes that were not specified in the evaluation code.

This suggests that even relatively simple evaluation logic can produce complex adaptive behavior when operating on heterogeneous inputs. The practical implication is that monitoring system design should account for the possibility of emergent processing regimes: a system that appears to have a single evaluation pass may in fact be operating in qualitatively different modes depending on input characteristics.

The observation that extended deliberation bursts typically terminate without convergence (0.7% settlement rate) raises questions about the behavioral impact of unresolved tension classification. If the system's final assessment is no more stable than a random tick within the burst, the "settled state" entering the next turn may be partially arbitrary. Future work should investigate whether turns with unresolved deliberation produce measurably different response quality or downstream state trajectories.

6.3 Identity Constraint as Dynamical Mechanism

The behavioral principles implemented through growth vectors and trait crystallization create a persistent constraint landscape in cognitive state space. Growth vectors accumulate over time. Crystallized traits accumulate. The memory store grows denser. Each mechanism adds constraint, and constraint reduces the volume of accessible state space.

The finding that internal state momentum exceeds external input influence is consistent with the hypothesis that accumulated epistemic constraint produces autonomous cognitive momentum. We emphasize that this is a correlation-level observation: the prediction experiment establishes that internal state dynamics are more predictive than external input, but does not establish a causal chain from identity constraint to temporal structure. An alternative explanation is that the monitoring system's own dynamics (e.g., entropy debt accumulation, tension type stability) produce the temporal structure independently of identity constraint.

The model-agnostic property of the system (Section 3.1) provides partial evidence against this alternative. The data collection period spans a transition from DeepSeek to GLM-5 as the generative backend. If the temporal dynamics were driven primarily by the monitoring system's mechanical properties—tick-level recomputation patterns, entropy debt arithmetic—they would be invariant to the model swap by construction. But the features showing the strongest prediction improvement are not mechanical: `qualityDecay`, `emotional`, and `paradox` depend on the interaction between the evaluator and the LLM's actual output characteristics. That these features show strong learnable dynamics *across* a model transition suggests that the architectural constraint (which remained constant) shapes the dynamics more than the generative model (which changed). The transition occurred approximately in late January 2026, with the majority of the dataset (roughly the first three months) collected under DeepSeek and the final weeks under GLM-5. A con-

trolled comparison of prediction accuracy between the two periods is not reported here due to the confound of temporal position (the GLM-5 period is later and thus reflects a more mature system state), but the absence of visible discontinuity in the prediction results is consistent with model-invariant dynamics. Fully disentangling these contributions would require controlled ablation studies beyond the scope of this paper.

6.4 Implications for Agent Safety and Monitoring

Current agent monitoring is predominantly rule-based, with hand-coded detectors for anticipated failure modes. A learned dynamics model trained on normal cognitive state trajectories would provide continuous anomaly detection: any state transition deviating from learned patterns is flagged, whether or not the specific failure mode was anticipated. This complements rule-based approaches by detecting novel anomalies.

The finding that internal state momentum exceeds external input influence has implications for adversarial robustness. If an agent's cognitive state has sufficient autonomous momentum, single-turn adversarial inputs may be unable to overcome accumulated state inertia. We note this as a hypothesis warranting direct experimental investigation rather than an established result. The prediction experiment measures temporal structure, not adversarial robustness per se.

6.5 Connection to World Model Frameworks

Maes et al. (2026) demonstrate stable end-to-end world model learning from pixels using an encoder-predictor architecture with SIGReg regularization. The framework is domain-agnostic in principle.

The prediction experiment reported here establishes the prerequisite for extending this framework to cognitive state spaces: temporal structure exists in the trajectories. A JEPA-style cognitive dynamics model would replace the pixel encoder with a cognitive state encoder, the motor action predictor with an input-conditioned cognitive transition predictor, and apply SIGReg unchanged to prevent representation collapse in the learned cognitive latent space. The resulting model would provide a continuous embedding space of cognitive states, enabling trajectory visualization, anomaly detection through prediction error, and potentially proactive steering through trajectory forecasting.

We propose this extension as future work rather than a contribution of this paper. The key empirical finding enabling it—that temporal structure exists and is learnable—is established here.

7 Limitations

Single agent, single architecture. The findings demonstrate that temporal dynamics *can* emerge from metacognitive infrastructure, not that they *always* do. Generalizability to other architectures, monitoring designs, or language model backends is untested.

Instrumentation-defined state. The "cognitive state" analyzed here is defined by what the plugins measure. Richer or different instrumentation might reveal stronger or weaker signal. The state representation is an experimenter choice, not a ground truth.

No causal claims. The prediction experiment establishes temporal correlation, not causation. We cannot assert that specific state features cause future state changes, only that they predict them.

Simple predictor. The MLP architecture is deliberately simple to establish baseline signal. More expressive models (transformers, recurrent networks) might capture additional structure, or might overfit given the dataset size.

No behavioral outcome measurement. We demonstrate that temporal dynamics exist in cognitive state but do not directly measure whether these dynamics improve agent behavior, output quality, or user experience.

Researcher-as-architect bias. The system was designed by the author, who also designed the experiment. The monitoring infrastructure was built to track the phenomena subsequently measured. Independent replication on independently built systems would strengthen the claims.

Turn boundary heuristic. Turn-level aggregation uses a 60-second gap threshold. This value was selected based on empirical examination of the inter-tick gap distribution, which shows a clear bimodal separation between intra-burst ticks (<5 seconds, 95.4% of gaps) and inter-turn gaps. Different thresholds in the range of 30–120 seconds yield similar turn counts due to this bimodal structure.

Unresolved deliberation. The finding that 99.3% of extended deliberation bursts terminate without tension type settlement raises questions about the impact of unresolved evaluation on downstream behavior that this paper does not address.

Data provenance asymmetry. The primary dataset comes from the research server with 25-dimension granularity. The OpenClaw dataset has reduced granularity and serves only as validation that structural patterns persist across architectural rewrite.

8 Future Work

Several directions follow from these findings:

1. **JEPA-style learned dynamics model.** Build an encoder-predictor-regularizer model with SIGReg over cognitive state trajectories, producing a learned latent space of cognitive states for trajectory visualization, anomaly detection, and proactive steering.

2. **Richer state representation.** Add text embedding features via a small language model encoder to capture semantic content alongside the current numerical state dimensions.

3. **Real-time anomaly detection.** Deploy the dynamics model in the agent loop to flag state transitions that deviate from learned patterns, as a complement to existing rule-based monitoring.

4. **Multi-agent generalization.** Extend the analysis to multiple agents built on the OpenClaw architecture to test whether temporal dynamics are architecture-general or specific to Clint.

5. **Behavioral outcome correlation.** Directly measure whether learned cognitive dynamics predict response quality, user satisfaction, or other behavioral outcomes.

6. **Adversarial robustness.** Investigate whether accumulated state momentum provides measurable resistance to single-turn adversarial inputs.

7. **Deliberation regime analysis.** Examine whether turns in the extended deliberation regime produce measurably different response quality, and whether early termination of non-converging deliberation affects downstream behavior.

8. **Deliberation optimization.** Explore whether detecting early that a burst will not converge enables earlier evaluation termination without behavioral cost.

Acknowledgments and Disclosure of Funding

Data was collected from a continuously operating agent system over the five-month period described. The author acknowledges the use of Claude (Anthropic) for data analysis pipeline development and manuscript preparation assistance.

Appendix A. Full Per-Feature Results

Table 3 presents the complete per-feature breakdown for Experiment A, including all four model variants and both baseline comparisons.

Appendix B. Intra-Turn Analysis Details

Burst length distribution. The 2,992 turn-level bursts exhibit a bimodal distribution: 1,781 bursts (59.5%) contain fewer than 10 ticks, while 1,174 bursts (39.2%) contain 20 or more ticks. Only 37 bursts (1.2%) fall in the 10–20 tick range. The modes center approximately at 6 ticks (1,240 bursts in the 5–8 range) and 50 ticks (1,125 bursts in the 30–55 range).

Tension type transitions by regime. Table 4 presents the processing characteristics of each regime.

Numeric feature stability. Among the 2,451 qualifying bursts ($\geq$5 ticks), 45.0% show exactly zero intra-burst entropy variance. The median convergence ratio (second-half variance divided by first-half variance) is 0.0 for all numeric features, and the median Spearman correlation between tick index and feature value is 0.0. Numeric features are recomputed deterministically across ticks.

Table 3: Complete per-feature results for Experiment A. All MSE values are computed on normalized features under LOSO cross-validation. *A vs. M* and *A vs. P* denote percentage improvement of Predictor A over the mean and persistence baselines respectively. *B vs. P* denotes Predictor B improvement over persistence.

Feature	Mean	Persist	MLP-A	MLP-B	A/M%	A/P%	B/P%	AC(1)
entropy_total	0.942	1.122	0.761	0.768	19.2	32.2	31.6	0.400
correction	0.971	1.768	0.992	0.996	−2.2	43.9	43.6	0.080
novelConcepts	0.784	1.209	0.737	0.760	6.0	39.0	37.1	0.316
emotional	0.773	1.556	0.797	0.787	−3.2	48.8	49.4	0.026
paradox	0.963	1.842	0.989	0.986	−2.7	46.3	46.5	0.093
qualityDecay	1.089	2.336	1.104	1.091	−1.3	52.8	53.3	−0.005
recursiveMeta	0.864	1.011	0.754	0.757	12.7	25.4	25.1	0.472
quietInteg.	1.239	2.028	1.157	1.152	6.6	43.0	43.2	0.124
quality	1.026	1.562	0.962	0.954	6.2	38.4	38.9	0.185
shannon	1.017	0.852	0.659	0.659	35.2	22.7	22.7	0.537
decision_old	1.046	1.892	1.030	1.036	1.6	45.6	45.2	0.081
decision_new	0.965	1.409	0.832	0.844	13.7	40.9	40.1	0.267
decision_div.	0.991	1.650	0.927	0.937	6.5	43.8	43.2	0.131
entropy_debt	0.557	0.082	0.100	0.103	82.0	−21.9	−25.3	0.955
quality_rating	1.003	1.465	0.930	0.924	7.3	36.5	36.9	0.239
tension_type	0.894	0.058	0.064	0.067	92.9	−9.8	−16.4	0.962
user_length	1.041	1.800	1.030	1.025	1.0	42.8	43.1	0.222
response_len.	1.031	1.673	0.997	0.986	3.3	40.4	41.1	0.194
self_ref_ratio	0.914	1.536	0.902	0.888	1.3	41.3	42.1	0.201
question_dens.	1.290	2.388	1.325	1.345	−2.7	44.5	43.7	0.081
resp_to_input	0.816	1.296	0.761	0.757	6.8	41.3	41.6	0.281
Overall	0.963	1.454	0.848	0.849	12.0	41.7	41.6	—

Table 4: Processing characteristics by burst regime. Transition rate is the number of tension type transitions per tick. Settlement rate is the fraction of bursts where the last 25% of ticks share a single tension type.

Regime	N	Mean ticks	Mean trans.	Trans. rate	Unique types	Settled
Fast (<10)	1,781	5.3	0.7	0.13	1.7	99.9%
Extended (≥20)	1,174	45.2	14.9	0.29	4.0	0.7%
Intermediate	37	12.4	4.5	0.36	2.4	51.4%

Appendix C. Training Hyperparameters

Table 5 specifies the full training configuration.

Table 5: Training hyperparameters for all MLP experiments.

Parameter	Value
Architecture	3-layer MLP (input $\to$ 64 $\to$ 64 $\to$ output)
Activation	ReLU
Dropout	0.1 (after each hidden layer)
Optimizer	Adam ($\beta_1 = 0.9$, $\beta_2 = 0.999$)
Learning rate	10^{-3}
Batch size	256
Max epochs	100
Early stopping	Patience 15 (training loss)
Normalization	Shared: pool training $\mathbf{s}_t$ and $\mathbf{s}_{t+1}$ for single $\boldsymbol{\mu}$, $\boldsymbol{\sigma}$
Evaluation	LOSO CV, 100 folds (sampled from 284 valid sessions)
Device	Apple MPS (Metal Performance Shaders)

Appendix D. Turn Boundary Sensitivity

The 60-second gap threshold for turn boundary detection was selected based on the empirical distribution of inter-tick gaps. Of 68,109 consecutive tick pairs, 95.4% have gaps below 5 seconds (intra-burst evaluation ticks) and the remaining 4.6% have gaps exceeding 60 seconds (inter-turn boundaries). The gap distribution is sharply bimodal with negligible mass in the 5–60 second range, making the specific threshold choice within this range insensitive. A threshold of 30 seconds yields 2,995 turns; 60 seconds yields 2,992; 120 seconds yields 2,987. The three-turn difference between 30-second and 120-second thresholds confirms that the bimodal gap structure makes results robust to this parameter.

Session boundaries (30-minute gap threshold) similarly fall in a sparse region of the inter-turn gap distribution. The number of sessions varies from 612 (15-minute threshold) to 487 (60-minute threshold), but the leave-one-session-out evaluation protocol is designed to be robust to session granularity variation.

References

John R. Anderson, Daniel Bothell, Michael D. Byrne, Scott Douglass, Christian Lebiere, and Yulin Qin. An integrated theory of the mind. *Psychological Review*, 111(4):1036–1060, 2004.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations*, 2022.

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mahmoud Assran, and Nicolas Ballas. V-JEPA: Video joint embedding predictive architecture. *arXiv preprint arXiv:2404.16930*, 2024.

David Ha and Jürgen Schmidhuber. World models. In *Advances in Neural Information Processing Systems*, 2018.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

John E. Laird, Allen Newell, and Paul S. Rosenbloom. SOAR: An architecture for general intelligence. *Artificial Intelligence*, 33(1):1–64, 1987.

Yann LeCun. A path towards autonomous machine intelligence. *OpenReview preprint*, 2022. Version 0.9.2.

Lucas Maes, Quentin Le Lidec, Damien Scieur, Yann LeCun, and Randall Balestriero. LeWorldModel: Stable end-to-end joint-embedding predictive architecture from pixels. *arXiv preprint arXiv:2603.19312*, 2026.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, 2023.

Abhishek Rath. Agent drift: Quantifying behavioral degradation in multi-agent LLM systems over extended interactions. *arXiv preprint arXiv:2601.04170*, 2026.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 2023.

Nicolas Tacheny. Geometric dynamics of agentic loops in large language models. *arXiv preprint arXiv:2512.10350*, 2026.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, Wayne Xin Zhao, Zhenhua Wei, and Ji-Rong Wen. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6), 2024.

Yukun Zhang and Qi Dong. Empirical investigation of latent representational dynamics in large language models: A manifold evolution perspective. *arXiv preprint arXiv:2505.20340*, 2025.