On the Private Internal Language in OLM v2: Creation, Purpose, and Discovery

Payton Miller

September 07, 2025

Abstract

This paper focuses on the **private internal language** (PIL) that emerges inside OLM v2, a real-time, unsupervised, structurally plastic agent. The PIL originates in perception as latent codes and evolves into a control protocol that coordinates internal self-organization. The document defines the construct, motivates its purpose, details the mechanisms by which it forms, and proposes methods to empirically detect, measure, and validate it. The goal is to establish PIL as a first-class entity: neither a human-interpretable token system nor a mere by-product of compression, but an operational communication layer that is necessary and sufficient for OLM v2 to allocate complexity, stabilize learning, and develop macro-behaviors without external supervision.

On the Private Internal Language in OLM v2: Creation, Purpose, and Discovery

Author: Payton Miller **Project:** Organic Learn Model (OLM v2) — Intelligence Engine **Version:** Draft for internal review

Abstract This paper focuses on the **private internal language** (PIL) that emerges inside OLM v2, a real-time, unsupervised, structurally plastic agent. The PIL originates in perception as latent codes and evolves into a control protocol that coordinates internal self-organization. The document defines the construct, motivates its purpose, details the mechanisms by which it forms, and proposes methods to empirically detect, measure, and validate it. The goal is to establish PIL as a first-class entity: neither a human-interpretable token system nor a mere by-product of compression, but an operational communication layer that is necessary and sufficient for OLM v2 to allocate complexity, stabilize learning, and develop macro-behaviors without external supervision.

1. Introduction

Contemporary AI commonly externalizes meaning via human tokens or task-specific rewards. OLM v2 is designed under different constraints: **no external supervision**, **continuous online learning**, and **structural plasticity**. Under these constraints, the system must invent an internal means to regulate itself. This paper isolates that element—the private internal language (PIL)—and addresses three questions:

1. **Creation:** by what mechanisms does PIL arise in OLM v2? 2. **Purpose:** what functional roles does PIL serve during perception, memory, and control? 3. **Discovery:** how can PIL be detected, measured, and validated without translating it into human tokens?

2. Definition: Private Internal Language (PIL)

PIL is the set of **latent patterns and control tokens** that are produced, propagated, and consumed exclusively inside OLM v2. It is characterized by:

- **Endogeneity:** symbols originate from the model's own dynamics.
- **Operational semantics:** symbols have meaning only through their causal effects on internal modules and actions.
- **Opacity:** decoding to input space is neither required nor always possible (especially with sealed latents).
- **Instrumentability:** although opaque, PIL leaves measurable traces in routing, plasticity events, prediction error, and memory dynamics.

Formally, let (x_t) denote the sensory packet at tick t. Let (l_t^f) and (l_t^a) denote **frozen** and **adaptive** perceptual latents. Let (z_t) denote the **compressed temporal state** from the Compression LSTM. Let (u_t in U) denote **control tokens** (e.g., EXPAND, MERGE, ROUTE->k, PRUNE) emitted by the controller D-LSTM. The PIL at t is the tuple ((l_t^f, l_t^a, u_t)) together with its short-horizon history.

3. Purpose: Why a Private Internal Language is Needed

Autonomy: OLM v2 cannot rely on human tokens or external rewards, so it requires an internal coordination medium to allocate capacity, set plasticity levels, and choose specialists. **Stability:** the **frozen** perceptual channel provides a stable dialect anchoring long-term associations; **adaptive** perception provides abstraction and efficiency. PIL reconciles both via control decisions. **Macro formation:** sequences of internal tokens can induce **expansion** of specialists and consolidate routines into **macros**, enabling multi-step behaviors from primitive actions. **Self-diagnosis:** control tokens act as self-descriptions of state transitions (spawn/merge/route/prune), enabling rollback, auditing, and reproducibility without exposing internal latents to external decoders.

4. Architectural Conditions for PIL

4.1 Dual Perception: Frozen and Adaptive beta-VAEs

- **Frozen beta-VAE:** randomly initialized then frozen; provides an immutable, sealed latent mapping (x_t -> l_t^f).
- **Adaptive beta-VAE:** updated online; provides a lossy, evolving mapping (x_t -> l_t^a). Optional auxiliary alignment keeps (l^a) near (l^f) within bounded drift.

4.2 Temporal Extraction and Compression

A Pattern LSTM produces a temporal feature stream that a Compression LSTM reduces to a fixed vector (z_t). All controller modules consume (z_t) (and optionally (l_t^f, l_t^a)).

4.3 Controller and Specialists

- **Controller D-LSTM:** emits primitive actions and **control tokens** (u_t).
- **Specialist D-LSTMs:** cloned on demand and fine-tuned on recurring contexts; configured by (u_t) and selected via routing decisions.
- **Plasticity loop:** EXPAND->train->ROUTE->MERGE/PRUNE.

4.4 Wake/Sleep Modes

Wake performs online adaptation; Sleep replays stored traces and conducts generative prediction. Sleep consolidates token sequences and refines the mapping from latent contexts to control tokens.

5. Creation: From Latents to a Control Protocol

Perceptual syllables. The earliest PIL units are the latents (l_t^f) and (l_t^a). The frozen channel yields stable “syllables”; the adaptive channel yields evolving “allophones.” **Binding to control.** The controller learns mappings ((z_t, l_t^f, l_t^a) -> u_t), where (u_t) regulates plasticity and routing. **Vocabulary growth.** Over time, the set U expands from a minimal bootstrap (EXPAND, ROUTE, MERGE, PRUNE) to include **typed variants**, e.g., EXPAND:ContextHash, ROUTE:k|TTL, MERGE:i,j|policy. **Sequencing and grammar.** Recurrent patterns over (u_t) form **phrases** (e.g.,

EXPAND->ROUTE->MERGE) with measurable predictive value for future actions or error reductions.

6. Discovery: How to Detect and Validate PIL

The aim is not to translate PIL, but to **demonstrate necessity, sufficiency, and structure**.

6.1 Instrumentation

- **Control-token logs:** time-stamped (u_t) with reasons, context hashes, and targets.
- **Routing traces:** module selections per tick and confidence.
- **Drift monitors:** metrics over (l^a) vs (l^f) (e.g., CKA or Procrustes distance).
- **Prediction/error curves:** short- and long-horizon error with and without token sequences.

6.2 Necessity Tests (Ablations)

- **Mute tokens:** zero out (u_t) while keeping perception/action intact; measure degradation in macro formation and stability.
- **Freeze vocabulary:** restrict U to bootstrap set; measure lost efficiency vs expanded vocabulary.
- **Randomize tokens:** replace (u_t) with noise; test whether behavior collapses toward novelty-seeking without competence gains.

6.3 Sufficiency Tests

- **Replay substitution:** drive learning using recorded ((z_t, u_t)) pairs while withholding direct labels; observe recovery of specialists and macros, indicating tokens carry operational content.
- **Counterfactual token injection:** inject designed sequences (e.g., MERGE: i,j) and verify intended structural changes occur.

6.4 Structure and Predictive Utility

- **Mutual information:** I(u_t; future error reduction) and I(u_t; routing decisions_{t+Delta}).
- **Compressibility:** MDL or n-gram perplexity over (u_t) sequences; decreasing perplexity indicates grammar formation.
- **Compositionality:** test whether concatenations of shorter token motifs predict specific macro outcomes.
- **Grounding via effect, not decoding:** show consistent causal links between token patterns and measurable internal/external changes.

7. Experimental Protocols

E1: Sealed vs Accessible Perception. Frozen-only vs Adaptive-only vs Dual; evaluate token vocabulary growth, routing stability, and macro emergence latency. **E2: Token Necessity.** Compare normal operation vs muted tokens vs randomized tokens. **E3: Plasticity Loop.** Spawn-only vs

Spawn+Merge vs Spawn+Merge+Prune; measure token grammar complexity and model compactness. **E4: Sleep Contribution.** No-sleep vs Replay-only vs Prediction-only vs Both; compare token-sequence consolidation and long-horizon error. **E5: Environment Shift.** Measure how token distributions adapt; compute adaptation half-life and reuse of existing token motifs.

Outcomes: macro emergence time, routing precision/recall, drift velocity, token perplexity, MI scores, stability indices, and capacity usage.

8. Threats to Validity and Alternative Explanations

- **Epiphenomenality:** tokens may correlate with behavior without causing it. Address via counterfactual interventions and mutes.
- **Overfitting of control:** a large token vocabulary could memorize contexts; address via environment randomization and capacity caps.
- **Metric leakage:** mutual information can inflate under autocorrelation; use shuffled baselines and block bootstraps.
- **Representation drift confounds:** adaptive perception can change token distributions; monitor drift and apply rollback criteria.

9. Ethical and Interpretability Considerations

PIL makes OLM v2 more autonomous but less transparent. To maintain oversight: log control tokens with typed schemas, cap structural changes per window, require human authorization for high-impact tokens (e.g., wide merges), and keep telemetry on drift, capacity, and action entropy.

10. Reproducibility and Compute Profile

- **Hardware:** single RTX 4080 (16 GB), consumer CPU, 32+ GB RAM.
- **Observed utilization (reference run):** ~40% GPU, ~30% CPU for the full pipeline.
- **Determinism:** fixed random seeds, frequent checkpoints, versioned latent schemas, and rollback if drift exceeds thresholds.
- **Telemetry:** token logs, routing matrices, drift metrics, replay stats, and pre/post-sleep error curves.

11. Related Concepts (selective)

- **Emergent communication** (multi-agent setups): contrasts with **intra-agent** PIL.
- **Predictive coding and world models:** PIL acts as a control layer coordinating prediction and plasticity.
- **Hierarchical RL/options:** PIL creates macro-like routines without external rewards.

- **Continual learning:** PIL supports structural adaptation and consolidation without catastrophic supervision.

12. Conclusion and Claim

A private internal language emerges in OLM v2 as a necessity of real-time, unsupervised, structurally plastic operation. Its symbols originate in perception, become actionable through a typed control protocol, and prove their reality through causal effects on routing, plasticity, and macro formation.

Claim (falsifiable): to the author's knowledge, this is the first description of an **endogenous, sealed-option** control language inside a single agent that is both (i) necessary for structural self-organization and (ii) sufficient to consolidate behaviors via replay without external supervision.

Appendix A: Token Protocol Schema (suggested)

- **TOK** ::= EXPAND | ROUTE | MERGE | PRUNE | ADAPT | ALIGN.
- **ARGS** ::= ContextHash | ModuleId | TTL | PolicyId | DriftBudget | ReasonCode.
- **EXAMPLES:**
- EXPAND:ContextHash=H, TTL=K
- ROUTE:Module=7, TTL=20
- MERGE:A=3,B=5, Policy=distill
- PRUNE:Module=12, Reason=inactive
- ALIGN:Budget=e, Reason=drift

Appendix B: Metrics (operational)

- **Drift:** CKA(l^a, l^f); thresholding and rollback.
- **Routing quality:** precision/recall vs post-hoc oracle; confusion matrices.
- **Token utility:** I(u_t; Delta_error_{t:t+H}); MDL over token sequences; intervention success rate.
- **Capacity:** specialists alive, survival curves, amortized gain per module.

Appendix C: Ablation Checklists

- Perception: frozen-only / adaptive-only / dual.
- Tokens: normal / muted / randomized / frozen vocabulary.
- Plasticity: off / spawn / spawn+merge / spawn+merge+prune.
- Sleep: off / replay / prediction / both.
- Hash gating thresholds; driver subsets; depth policy.