KODASŌKEN | WORKING PAPER | TOKYO

A Research Programme in Persistent Intelligence

An introduction for the mathematically inclined — what persists when the model is replaced.

Carlos Mundim | KodaSōken, Tokyo | cmundim@kodasoken.com
ORCID 0009-0001-5000-7409 | July 2026

ABSTRACT

Contemporary progress in artificial intelligence is organised around *scale*. This paper argues for a complementary programme organised around *continuity*: what an agent's operating identity consists of, how it can be represented *outside* any single model, and whether it can be measured and preserved across a change of the underlying model. We formalise *identity fidelity* as a scored quantity for controlled substrate transitions, report an internal live-system observation ($F = 0.961$) under a disclosed protocol, and connect the empirical picture to a low-rank account of behavioural *disposition* via LoRA. We close with five open problems — on continuity metrics, the geometry of the disposition subspace, disposition–knowledge dissociation, the autonomic/cognitive frontier, and continuity under distribution shift — framed for a mathematical audience. Throughout, we maintain an explicit taxonomy separating what is *demonstrated*, *argued*, and *not claimed*.

Keywords. persistent agents • identity fidelity • low-rank adaptation • continuity as architecture • small language models • sovereign / edge AI

§0 Scope and audience

This paper is written to be legible to a strong undergraduate mathematician. It does not assume prior exposure to machine learning, but it does assume you value precision about what is *proven*, what is *argued*, and what is *not claimed*. Where we have a formalism, I give it. Where we have only an empirical observation on a single system, I say so. The most interesting parts of what follows are the open problems, stated plainly at the end so that you can judge for yourself whether any of them are worth a mathematician's attention.

A one-sentence orientation: most of the field studies **how well a model performs**; KodaSōken studies **what persists when the model is replaced**. Those are different questions, and the second one is, we believe, closer to mathematics than to engineering.

§1 The central question

Contemporary AI progress is dominated by *scale*: larger models, more parameters, more data. A language model is, to first approximation, a function

$$f_\theta : \Sigma^* \to \Delta(\Sigma), \qquad \theta \in \mathbb{R}^N,$$

mapping a finite string over a token alphabet Σ to a probability distribution over the next token, with parameter count N now in the hundreds of billions. Scaling laws describe how a loss functional decreases — empirically, as a power law — in N, data, and compute.

A large model, however, is *stateless* between invocations and *ephemeral* across versions. Each forward pass begins from the same weights; when the weights are replaced by a newer model, whatever "operating identity" a deployed agent had — its role, its accumulated commitments, its relationships, its memory of prior decisions — is, by default, destroyed and rebuilt from scratch. KodaSōken's research programme is organised around the complementary question:

> *When the model changes, what must remain — and can "what remains" be represented, measured, and preserved by construction rather than by luck?*

We call the affirmative hypothesis *continuity as an architecture*: the claim that an agent's operating identity should live in a governed runtime *outside* any single model, so that models become interchangeable cognitive resources the agent recruits, rather than the agent itself. This reframes several engineering problems as mathematical ones. If identity is an object living outside the model, it has a representation, a metric, an invariance group, and a transfer theory — all of which are, at present, only partially understood.

§2 Agentic AI and the KoLo runtime

An **agent**, in our usage, is not a single call to f_θ but a governed loop with persistent state:

- a **constitution** κ — a structured, versioned description of role, boundaries, permissions, and unresolved commitments;
- a **memory** M — a three-tier store (short/medium/long term) with semantic retrieval, so context is reconstructed on demand rather than held in a finite window;
- an **orchestration policy** π — which decides, per step, whether a task is handled deterministically, by a small specialist model, or escalated to a frontier model, subject to κ;
- an **audit and governance layer** — every action logged in an append-only, hash-chained ledger; consequential actions require human approval.

KoLo is the runtime that hosts $(\kappa, M, \pi, \text{audit})$ and treats the underlying language model as a *pluggable* component. A useful decomposition, recurring across the field under different names, is a two-layer control pattern:

- a cheap, often deterministic **proposer** (a *Director*) that reasons about *what should happen next* at the semantic level;
- an expensive **grounder** (a *Pilot*) that realises the proposal.

The mathematically interesting empirical fact is the *ratio* between these layers. On a persistent agent we operate, deterministic "autonomic" operations — scheduling, memory consolidation, health checks, policy enforcement — outnumber model inferences by roughly

$$\frac{\text{autonomic cycles}}{\text{cognitive invocations}} \approx 68 : 1$$

over a measured interval (ranging 35:1–117:1 across roles). We do not claim this ratio is optimal or fundamental; we report it because it suggests that the *quantity of intelligence-bearing computation* required to sustain a competent agent is far smaller than the quantity of deterministic coordination around it — with obvious implications for cost, energy, and edge deployment, and one we would like to see given a proper theoretical treatment.

§3 Continuity as a measurable quantity

If identity persists outside the model, we should be able to *measure* whether it survives a model change. We define an **Identity Fidelity** score for a controlled substrate transition — swapping the underlying model of a running agent while holding κ and M fixed:

$$F = 0.4\,C + 0.3\,(1 - D) + 0.3\,R, \qquad C, D, R \in [0, 1].$$

Here C measures **constitution preservation** (does the post-swap agent still behave consistently with κ?), D measures **drift** (divergence of behavioural posture from the pre-swap baseline, so $1 - D$ rewards low drift), and R measures **recall** (fidelity of retrieval from M across the transition). The weighting is a deliberate, documented choice, not a fitted one.

$$F = 0.961$$

recorded on a controlled cross-provider transition, with the operating lineage preserved across **four** completed model/provider transitions and **336+** versioned memory states — none lost.

I want to be exact about the epistemic status of this number, because a mathematician will (correctly) ask:

- It is an **internal, live-system observation**, not an externally reproduced benchmark.
- The four transitions were **planned and supervised**, not adversarial.
- F was **formally scored on the most recent transition only**; the others are attested operationally.
- C, D, R are themselves operationalised by procedures that are defensible but not canonical.

So $F = 0.961$ should be read as: *under a specific, disclosed protocol, a specific agent retained most of a specific notion of identity across a real model swap.* It is evidence, not proof, and we publish the formula, the conditions, and the failures precisely so that the claim is adversarially inspectable. Making continuity a number at all — rather than a vibe — is the point.

§4 Small specialist models, and the low-rank structure of identity

The scale-first orthodoxy says: to add a capability, make the model bigger. We pursue the opposite — a **model army**: one open base model plus many small, cheap, domain-specialised adapters, trained and served in-house, viable on commodity CPUs, with facts kept in auditable retrieval rather than baked into weights.

The mathematical engine here is **low-rank adaptation (LoRA)**. Fine-tuning normally updates a weight matrix $W \in \mathbb{R}^{d\times k}$ by a full-rank ΔW. LoRA instead constrains the update to a low-rank factorisation

$$W' = W + \Delta W, \qquad \Delta W = BA, \quad B \in \mathbb{R}^{d\times r},\ A \in \mathbb{R}^{r\times k},\ r \ll \min(d,k),$$

so that only $r(d+k)$ parameters are trained instead of dk. At rank $r = 8$ this is on the order of **0.22%** of the base model's parameters.

Our **Project Nyx** line asks a sharper question than "does LoRA work" (it does). We distinguish two things a fine-tune might change:

1. **Knowledge** — facts the model can state;
2. **Disposition** — the behavioural posture, style, and priorities with which it acts.

The working finding — a *manuscript*, pre-registered "write-then-run," not yet an externally reproduced result — is that **disposition is low-rank and localizable**: a large fraction of what makes an agent behave like *itself* can be captured in a rank-r adapter for small r, concentrated in identifiable layers, and it *dissociates* from factual knowledge (better left in retrieval). The ablation battery is pre-registered: a rank sweep, an attention-vs-MLP localisation, a layer-depth scan, and an explicit disposition-vs-fact dissociation test.

If this holds, it has a clean mathematical reading: the map from "training signal" to "behavioural identity" factors through a low-dimensional subspace of parameter space, and one may ask about the *geometry* of that subspace — its intrinsic dimension, whether it is (approximately) shared across base models (which would make identity *transferable* between substrates), and how it deforms under distribution shift. These are questions about the differential geometry of a learned manifold, and they are wide open.

§5 "Biological" AI — determinism as an architecture

We use the word *biological* not as metaphor-for-marketing but to describe a design stance: an agent is kept alive by a layer of **deterministic, zero-inference subsystems** — analogues of immune response (self-healing), circadian scheduling, homeostatic monitoring, reflexes (automatic recovery), and memory consolidation — running continuously and cheaply, with the expensive model recruited only for genuinely novel cognition. This is the structural reason the autonomic-to-cognitive ratio of §2 is so lopsided. The thesis: robustness and continuity are *cheaper and safer* when they are deterministic, and intelligence should be spent sparingly on what actually requires it.

§6 Sovereign and edge intelligence

Several of our applied deployments run in **high-trust, low-connectivity environments** — clinics and hospitals — where data cannot leave the building. This forces an architecture we find independently interesting: models small enough to run *offline at the edge*, retrieval grounded in a local corpus, and **deterministic safety gates** that sit between model output and any consequential action. A gate is a decidable predicate the model's proposal must satisfy — e.g. *a clinical recommendation may not be recorded unless it carries an approved citation and defers the final decision to a licensed human*. The governing principle throughout:

> *The AI prepares; a verifier checks; a human decides; the audit records everything.*

The model is never the last link in a consequential chain.

§7 The evidence discipline

Because this is a research house and not a hype cycle, we hold ourselves to an explicit three-level taxonomy for every claim:

DEMONSTRATED reproducible under a disclosed protocol — e.g. a persistent agent operating continuously for 129 days, with a verification pack of hashes, logs, and version counters.

ARGUED a coherent theoretical case, not yet reduced to an experiment (much of the identity-transfer theory).

NOT CLAIMED deliberately outside our assertions. In particular, **we make no claim of consciousness, subjective experience, or personhood for any system we operate.** Our questions are non-phenomenological by construction: we ask what computational structures make persistent self-reference and relational continuity possible, and how to evaluate them *without anthropomorphic inference*.

This discipline is itself a research artifact — one of our public working papers is a *methodology* for making claims about persistent-agent architectures adversarially inspectable. A mathematician will recognise the instinct: state your assumptions, bound your conclusions, and publish your failures alongside your successes. Our first weight-transfer pilot, for the record, produced measurable behavioural displacement *and* factual confusion — both are on the record.

§8 The research corpus

Status labelled honestly — working manuscript, public working paper, or internal memorandum:

- **Relational Intelligence and Emergent Introspective Awareness (RAAI)** — *working manuscript* — the central, explicitly non-phenomenological question.
- **Identity Is Low-Rank and Localized** — *working manuscript* — the Nyx result of §4, pre-registered ablations.
- **Continuity Without Cognition** & **Continuity-Based Computational Systems** — *working manuscripts* — continuity as architecture; introduces *Relationally Conditioned Systems*.
- **Long-Term Memory in AI Agents** — *working manuscript* — a framework for persistent context and behavioural continuity.

- **The Marianas Daemon Architecture** — *public working paper* — a 129-day persistent-agent deployment report, with reproducible verification pack.
- **Verifiable Evidence for Non-Scale Intelligence** — *public working paper* — the evidence methodology of §7.
- **Beyond the Perceptron: A Manifesto for Intelligence Without Scale** — *public working paper* — the scale-scepticism thesis.
- **Identity Fidelity Benchmark & KoLo: A Runtime for Persistent Multi-Model Agents** — *internal memoranda*.

Internal figures such as $F = 0.961$ and the 68:1 ratio are labelled *internal live-system observations, external reproduction pending*.

§9 Open problems — where a mathematician could actually help

These are stated as invitations, not as solved.

1. **A theory of continuity metrics.** What axioms *should* a continuity metric satisfy? Candidate desiderata: invariance under relabelling of tokens/embeddings, an upper bound achieved only by the identity transition, monotonicity under composition of transitions, and adversarial robustness (a hostile grader cannot inflate F). Our $F = 0.4C + 0.3(1 - D) + 0.3R$ is a first draft; the *right* object may be a (pseudo)metric on a space of agent behaviours, or a divergence between induced policies.

2. **The geometry of the disposition subspace.** If behavioural identity is captured by a rank-r adapter, what is the intrinsic dimension of the "identity manifold"? Is it approximately basis-aligned across different base models — i.e. is there a natural map that transports an identity from one substrate to another? A question about the alignment of learned low-rank subspaces under a change of parameterisation.

3. **Disposition–knowledge dissociation, made precise.** Formalise the claim that "disposition" and "factual knowledge" live in (approximately) orthogonal or separable components of the update. Under what conditions on the data-generating process does such a decomposition provably exist?

4. **The autonomic/cognitive frontier.** Given a task distribution, what is the minimal fraction of steps that *provably* require a learned model rather than a deterministic policy? A computational-complexity-flavoured question about the cost of competence.

5. **Continuity under distribution shift.** Model the substrate transition as a perturbation of the policy induced by f_θ; bound the drift D in terms of a distance between the old and new models. When is continuity *robust* rather than fortunate?

If any of these read as a problem you would enjoy, that is precisely the point of this paper.

§10 Closing

KodaSōken is a small research house in Tokyo. We build the runtime for persistent agents (KoLo), specialist small models (the Nyx line), and applied systems in high-trust domains —

and we do it under a strict discipline about what we will and will not claim. The unifying conviction is simple: **the models will keep changing; the interesting mathematics is in what connects one moment to the next.**

The field spent the last decade making the forward pass larger. We think the next decade belongs to whatever makes it *continuous* — a subject with far more unsolved mathematics in it than its engineering surface suggests.

You would be welcome to think about any of it.

Disclosure. Internal figures — notably $F = 0.961$ and the 68:1 autonomic-to-cognitive ratio — are internal live-system observations under disclosed protocols; external reproduction is pending, and they are labelled as such in the text. This work makes no claim of machine consciousness, subjective experience, or personhood.

Citation. C. Mundim. *A Research Programme in Persistent Intelligence.* KodaSōken Working Paper, Tokyo, July 2026.
Correspondence: cmundim@kodasoken.com ▪ ORCID 0009-0001-5000-7409.