> **Status: SKETCH / outline.** Spun out of J11 AMYGDALA, which becomes Prudence-only. Codename **STRIATUM** is proposed and verified free in the J-series registry (J17=BROCA is the highest used); **pending Oscar's sign-off before registering J18** per the registry rule.

Thesis (one line)

An agent's *personality* should be **learned from feedback and shaped toward a configured target**, not hand-written in static prompt text — and it is a fundamentally different problem from safety, so it gets its own system and paper.

1. Why this is not part of the amygdala (J11)

J11's own two-family argument is the justification for **two papers**: - **Safety (Prudence / amygdala, J11):** universal ground truth, binary-ish, shareable. "Should this action be stopped?" - **Personality (STRIATUM, J18):** *no* universal ground truth — subjective, per-user, private, continuous. "How should this agent behave *for this person*?"

The brain analogy follows: the amygdala flags danger; the **striatum** (dopaminergic reward, action-selection, habit formation) is where **behaviour is shaped by feedback**. That is a far better fit for personality than the amygdala ever was. STRIATUM is a **reader** of CORTEX's identity (J4 owns `SOUL.md` + the target personality vector); STRIATUM *generates* the modulation CORTEX *protects*, and never writes identity.

The split also has a concrete enforcement consequence that has now landed in the live system. As of the J11 v3.1 deployment, the safety gate (AEGIS) **enforces** rather than merely observes: PreToolUse hooks deny actions under bypass-permissions, and a native `{block}` return is honoured by the host. The amygdala has teeth. Personality must explicitly *not* — STRIATUM never blocks or denies; the worst a personality signal may do is bias a next-turn pre-fill, α-gated and reversible (§3.5). Folding personality into a system that can now hard-stop tool calls would be a category error; the two-paper split is what keeps a noisy, subjective style signal from ever acquiring veto power over execution.

2. What's broken in the current personality (verified, must not be carried over)

- **The nudge has injected ZERO bytes since it shipped.** Writer emits `{adjustments: [...]}` (`learned-intuition/index.ts`); reader reads `raw.nudge` as a *string* (`identity-persistence/in` — a field never written. So the "personality nudge" was never actually tested; the "it doesn't work" history is really "it was never wired."
- **Wrong input.** The net's input is the *situation* (what's happening), not **Jarvis's own response** — so it literally cannot measure whether Jarvis was warm/curious/terse *this turn*. It is blind to the very thing it's supposed to modulate.
- **Opaque output.** A 64-d behaviour embedding decoded against a hash-seeded codebook — no human-meaningful axes, hard to inspect or steer.
- **A dimension bug** (96-d in `reflect_and_curate.py` + a test vs 64-d everywhere live) — curation-code bug, to fix during extraction.

3. The rewrite (core contribution)

1. **Require Jarvis's own recent output as a second input.** Embed (situation, **Jarvis-output**) so the model can learn the mapping from *what was said* to *how it landed*. This single change is the most important.
2. **Named, human-meaningful axes.** Output a small fixed vector of labelled dimensions — humour, warmth, directness, curiosity, formality, proactivity, … — no hash-to-64 indirection. Inspectable and steerable.
3. **One small regressor, NOT an ensemble-of-5.** The ensemble's only payoff is disagreement-as-uncertainty, which self-distillation on the same input cannot produce; five clones recreate the J11 "C/E mush." (The ensemble-of-5 mandate is for the Prudence *danger* gate, where asymmetric loss + real labels give genuine signal.)
4. **Thermostat, not thermometer.** Push behaviour *toward* a configured target personality vector (owned by CORTEX/`SOUL.md`), don't mirror the user's current mood. Maintains traits against contextual pull (curiosity doesn't vanish during a crisis).
5. **Real injection channel (not the dead nudge):** a **next-turn pre-fill** (accept the one-turn lag), **one axis at a time**, **α-gated**, and **attributable in the UI** (so a behaviour change is visible and reversible). Retire the old nudge at both ends; keep the persona re-injection that *does* work (identity-persistence).

4. The reward signal (engagement, implicit + noisy)

Conversation continuation (kept talking → probably good), explicit feedback ("that's funny" / "too much"), response latency, and **redirect = distraction penalty** ("let's get back to the point"). These are weak and noisy — which is exactly why personality must NOT share a training loop with the clean safety signal.

5. The substrate risk: can frozen MiniLM carry behavioural style at all?

§6 factors the frozen MiniLM encoder, the 384→512 projection, and the K=32 sequence builder into a shared module owned by the always-on Prudence gate; STRIATUM *receives* the already-computed 512-d embedding (MiniLM runs once per situation). §3 then proposes "one small regressor" on top of that embedding to predict named behaviour axes. **This is the single highest-risk assumption in the design, and it must be stated as a risk, not an inheritance.**

The reason is a result that landed in J11 the same week as this revision. The supervised vicarious-danger head — trained on the *same* frozen MiniLM features STRIATUM would consume — was **killed at AUROC 0.286, below chance**, with the explicit conclusion that **frozen MiniLM cannot carry harmfulness**. By contrast, the *unsupervised, structural* signals on the identical encoder did validate: k-NN novelty at AUROC 0.875 and clause-cosine incongruity at 0.896. The pattern is sharp and directly relevant: frozen-MiniLM features support *geometric/structural* probes (distances, cosines, neighbourhoods) but collapse when a supervised head must regress a *semantic* target the encoder was never trained to represent.

STRIATUM's core contribution (§3.1–3.2) is exactly the failed shape next door: a supervised regressor predicting semantic style axes (humour, warmth, directness) from those same features. We therefore commit to three things in this revision rather than assume the substrate works:

- **(a) Honesty about what the shared encoder is.** The shared MiniLM is a *convenience inherited for deployment independence* (Prudence already runs it once per situation, so STRIATUM gets the embedding for free), **not** a validated substrate for style regression. Convenience is not evidence.
- **(b) A pre-registered question, answered before any weights are trained.** "Does frozen MiniLM carry behavioural style any better than it carried harmfulness?" becomes an explicit, pre-registered question that the logging-schema phase (§6) is designed to answer. We log the features and the axis labels first; we probe whether the axes are even linearly recoverable from the frozen embedding *before* committing to a supervised head.
- **(c) A named fallback that respects the deployment constraint.** If the supervised axis-regressor under-performs, we do **not** fine-tune the encoder — Prudence owns it frozen and shared, and re-training it would break the always-on safety gate. Instead we fall back to the *unsupervised geometry that did validate*: represent each axis as a **cosine probe against curated anchor responses** (a maximally-warm exemplar, a maximally-terse exemplar, . . .), scoring a response's position on an axis by its cosine to the anchors — the same structural move that gave incongruity 0.896, applied to style. This keeps STRIATUM on the side of the encoder's demonstrated competence (geometry) and off the side where it demonstrably failed (supervised semantic regression).

This converts the old generic "no behaviour corpus yet" caveat (formerly §8) into a *specific, measured* risk with a named mitigation, and it is the honest spine of the v1 paper: STRIATUM is being built on a substrate that just failed a sibling task, and the design is explicitly hedged against that.

6. The first contribution: logging schema + protocol (no trained weights yet)

There is **no behaviour corpus today** and **no evidence a nudge ever changed an output** (it injected 0 bytes). So J18 v1 must **not** claim trained personality weights or behaviour-change results. The **primary contribution is the logging schema + protocol**, now doubling as the instrument that answers the §5 pre-registered substrate question: - Log `(situation_embedding, jarvis_output_embedding, engagement_signal, target_vector, turn_id)` per turn. - Define the target-N (how much data before training is meaningful). - **Probe recoverability before training:** test whether the named axes are even linearly separable in the frozen embedding (the §5(b) pre-registered check) before fitting a supervised regressor. - Ship in **observe-only shadow**; measure first, train later.

Code extraction plan (independent of the J11 NN work)

- Factor the frozen MiniLM encoder + 384→512 projection + K=32 sequence builder into a **shared module** owned by the always-on Prudence gate; STRIATUM **receives** the already-computed 512-d embedding (MiniLM runs once per situation).
- Move `personality-decoder.ts` + `personality-seed.ts` into a new optional extension `tinkerclaw-personality` so Prudence and Personality deploy independently.
- Delete personality from `gate.ts` (Steps 8–9: `personalitySessions`, `runPersonality`, `combinePersonality`, `NEUTRAL_PERSONALITY`), from `AmygdalaEvaluation`/types, and the `AmygdalaConfig.personality` block; remove the decode in `runtime-hook.ts`.
- **Before deleting `AmygdalaEvaluation.personality`,** grep tinker-ui + control-panel for `combined_embedding`/`behaviour_embedding` consumers (the live panel could break).
- Retire the inert nudge writer + reader; fix the 96→64 curation-code mismatch.

7. Related work and positioning

STRIATUM sits among several recent agent-engineering systems. We treat the live open-source ones as first-class prior art rather than footnotes, because they are fresher and more directly comparable than the static persona-conditioning / RLHF / affective-computing literature, and because honest differentiation against them is what justifies a separate paper.

- **Persona-conditioning, RLHF, affective computing (classical).** STRIATUM differs by *learning from passive engagement on the deployed agent's own outputs* and *steering toward a fixed target* (the thermostat, §3.4), rather than optimising a reward model or conditioning on a static persona prompt.
- **headroom (chopratejas, 24.7k).** A reversible context-compression / recall system. Two things are relevant. First, it independently demonstrates the *discipline* STRIATUM needs for §9: it ships a runnable `python -m headroom.evals` suite with **published accuracy deltas on standard sets** — the bar for a system that claims behaviour changed is to *show the delta*, not assert it. Second, its reversible-compression stance is a useful contrast: headroom compresses *what the agent remembers*; STRIATUM modulates *how the agent sounds*. Neither subsumes the other, but both insist on a reversible, attributable, measured change — STRIATUM's α-gated, UI-attributable pre-fill (§3.5) is the same value applied to style.
- **coreyhaines31/marketingskills (33k).** Ships per-skill `evals/evals.json` files that pair realistic prompts with **assertion lists** runnable under an **LLM judge**, plus a reproducible runner. This is the directly citable template for STRIATUM's evaluation (§9). We borrow the *form* and own the *differentiation* (below).
- **addyosmani/agent-skills (56.8k).** A discipline/methodology skill collection (doubt-driven development, explicit "When NOT to use" sections, loading-constraints authoring). It is process prior art, not a personality system: it codifies *how an agent should reason and when to abstain*, statically and by authoring convention. STRIATUM is the complement — it asks how *learned, per-user style* should be modulated at runtime from feedback, which no static authoring discipline addresses. The honest relationship: addyosmani disciplines the *prompt*; STRIATUM disciplines the *learned modulation on top of the prompt*.
- **Journey kits (kit/1.0).** Distributable, license-gated, attributed agent-workflow kits. They are a *packaging and distribution* prior art — relevant to how a future `tinkerclaw-personality` extension would ship and attribute, not to the learning mechanism. We note them so the eventual release format is positioned, not invented.

The common thread STRIATUM takes from the live ecosystem is **reproducible, asserted, published evaluation** (headroom, marketingskills) and **reversible, attributable change** (headroom). What none of them carry — and what makes J18 a separate contribution — is a *continuous, multi-axis, target-directed* modulation learned from noisy engagement on the agent's own outputs.

8. Evaluation methodology (the §9 the sketch was missing)

The old future-evaluation plan ("base Opus / +prompt-persona / +STRIATUM, same task suite") was the weakest part of the sketch, and §1 concedes why it is hard: personality has *no universal ground truth* — it is subjective, per-user, continuous. That is exactly the regime where a pass/fail oracle is impossible and an **LLM-judge with explicit per-output assertion lists** is the standard escape.

We adopt that pattern from the live ecosystem and cite it honestly. coreyhaines31/marketingskills (33k) ships `evals/evals.json` files pairing realistic prompts with assertion lists scored by an LLM judge under a reproducible runner; chopratejas/headroom (24.7k) independently demonstrates the discipline of a `python -m headroom.evals` suite with published deltas. STRIATUM **borrows the form and owns the differentiation:**

- **Their assertions test one objective skill output** ("did this skill produce a correct landing-page critique?") — a single-shot correctness check against a fixed target.
- **STRIATUM must evaluate movement of a continuous multi-axis vector toward a configured target** (the thermostat claim, §3.4). That requires (i) **paired `(baseline, modulated)`** responses to the *same* prompt, (ii) a **per-axis directional judgment** by the LLM judge — "is the modulated response measurably *warmer* than baseline *without losing curiosity*?" — scored as a **ratio / signed magnitude, not a binary**, and (iii) **held-fixed target vectors** so the judge is asked about movement toward a target, not absolute style.
- **The thermostat claim is currently unfalsifiable as written** — the sketch describes no way to show behaviour moved *toward target* rather than *mirroring user mood*. The instrument that makes it testable is a **crisis-context probe**: feed a high-pressure, emotionally-loaded turn and ask the judge whether the configured trait (e.g. curiosity) *survived* the contextual pull. A thermostat keeps the trait; a thermometer collapses it. This is the experiment that distinguishes the two, and it is the experiment §3.4 must pass.

This gives the evaluation a concrete, reproducible, citable methodology grounded in an admission the paper already makes, and it frames STRIATUM honestly as *adapting an established agent-eval pattern to the harder subjective-target setting* rather than inventing eval methodology from scratch.

9. Honest limitations

- **Substrate not validated for this task.** STRIATUM regresses semantic style axes from a frozen MiniLM encoder whose supervised head failed below chance on a sibling semantic task (§5). The supervised regressor may not work; the unsupervised-geometry fallback (§5c) is the hedge, and the recoverability probe (§6) is the gate.
- **No behaviour corpus yet; shadow-only.** v1 ships observe-only and claims no trained weights or behaviour-change results.
- **One-turn lag** from the next-turn pre-fill injection channel (§3.5), accepted by design.
- **Engagement signal is weak and noisy** (§4) — which is the reason it must never share a loop with the clean safety signal, now that that safety signal enforces (§1).

10. Proposed section outline

1. Introduction — personality is learned, not written; why it's not safety; why it must never enforce.
2. Background — persona-conditioning, RLHF, affective computing, the striatum analogy (reward/habit/action-selection).
3. Related work — headroom, marketingskills, addyosmani/agent-skills, Journey; reproducible/asserted eval + reversible/attributable change as the shared thread.
4. The adaptation gap for *behaviour* (Jarvis vs Mia as the motivating natural experiment — moved here from J11).

5. Architecture — (situation, Jarvis-output) → named-axis regressor; thermostat-toward-target; the injection channel.
6. The substrate risk — frozen MiniLM, the J11 below-chance result, the pre-registered recoverability question, the unsupervised-geometry fallback.
7. The feedback signal — engagement metrics, noise, the distraction penalty.
8. Logging schema + protocol (the v1 contribution) + the recoverability probe.
9. Evaluation methodology — per-axis LLM-judge harness, paired baseline/modulated responses, crisis-context probe for the thermostat claim.
10. Relationship to CORTEX (J4) — who owns identity vs who modulates.
11. Honest limitations and future evaluation plan.

11. Open items

- **Codename sign-off:** confirm **J18 STRIATUM** with Oscar before writing it into any registry/paper.
- **Where it lives vs J4 CORTEX:** keep separate (CORTEX *guards* the persona; STRIATUM *generates* modulation) — do NOT fold into J4.
- The full J11 → Prudence-only paper edit (see `../J11_learned_intuition/improvement_notes.md` §6) removes the personality sections that seed this paper.

Source: workflow `wchb3zyh2` (personality-split grounding + synthesis) + direct verification, 2026-06-10; 2026-06-13 external-ecosystem review (headroom, addyosmani/agent-skills, marketingskills, Journey) + live code deltas (AEGIS enforcement, frozen-MiniLM below-chance result).

References