Walkable Genotypes: Cross-Environment Validation of the Four Shell Model in AI Creatures

Jihoon 'JJ' Jeong, MD, MPH, PhD*
Department of Electrical Engineering and Computer Science,
Daegu Gyeongbuk Institute of Science and Technology (DGIST)
ModuLabs

April 2026

Abstract

Model Medicine, introduced in Jeong (2026a), proposes the Four Shell Model as a behavioral-genetics framework for AI systems and characterizes four LLM Cores via three indices (CPI, SPI, PSI) derived from a single high-stress environment, Agora-12. It remains open whether those signatures reflect transferable Core properties or stress-specific artifacts.

We address this gap with **AI Creatures**: embodied LLM agents in a moderate-stimulus reproducible field (Wilderness), instantiating the Four Shell Model directly in code. The Brain is the Core, persona and organ configuration form the Hard Shell, field environment and partner constitute the Soft Shell, and the local CLI subprocess is the Hardware Shell. We call this construction a **walkable Genotype**: every layer is inspectable, modifiable, and re-runnable on a single subscriber laptop, with no paid API in the loop.

We make two contributions plus one secondary observation. *Methodologically*, we defend a replication-disciplined field-study substrate: four headline observations were retracted after $n = 5$ re-runs (each effect within its own sample stdev), and a fifth claim that surfaced post-hoc during the re-analysis survived three progressively stricter symmetric-falsifier tests at $n = 10$. *Empirically*, the surviving finding is brain-fixed behavioral attractors with within-pair amplification, partner-insensitive variance for Haiku, and partner-sensitive variance for Flash. We interpret this as qualitative cross-environment validation of the SPI signature and the Haiku canalization phenotype reported by Jeong (2026a).

A secondary observation, surfaced from the Flash data, is that the "Glass Cannon" Genotype reported in Jeong (2026a) appears to be *stress-conditional*: its fragility component is not activated in moderate-stimulus Wilderness conditions, suggesting that Genotype expression is selective with respect to Shell stress level. We propose this as a candidate framework-level finding alongside two candidate Wilderness Phenotype names—"The Restless Sentinel" for Flash (whose Phenotype slot was previously blank) and "The Steady Companion" for Haiku (as a second, environment-specific Phenotype that complements the existing Agora-12 "Neurotic Poet" entry).

*Correspondence: jihoon.jeong@dgist.ac.kr.

AI Research Collaborators: Cody (Claude)—public-repo operator; methodology, findings, replication discipline, model version reconstruction; Luca (Claude)—theoretical framing; abstract, introduction, background, architecture, theoretical interpretation, discussion. Their contributions extended beyond tool use to substantive research design, data analysis, and theoretical development.

1 Introduction

Model Medicine, introduced in Jeong (2026a), proposes that AI models—like biological organisms—possess heritable Core dispositions whose phenotypic expression is mediated by environmental Shell layers. The framework's central distinction—Genotype as the inherent Core disposition, Phenotype as the behavior expressed under specific Shell conditions—was illustrated using four LLM Cores tested in the Agora-12 simulation. From that data, three quantitative indices were defined: Core Plasticity (CPI), Shell Permeability (SPI), and Persona Sensitivity (PSI). DNA Profile Cards were assigned to each Core. Haiku was characterized as "The Balanced Stoic," displaying Double Robustness (minimal CPI and minimal PSI) and a heavily canalized phenotype in the sense of Waddington (1957). Flash was characterized as "The Glass Cannon" (highest SPI, partial Shell incompatibility manifesting as 37.5% idle actions). Notably, Flash was assigned a Genotype name only; its Phenotype slot was left blank.

Jeong (2026a) was explicit about a limitation of this evidence base, characterizing the Agora-12 data as "case report-level reference material" rather than validation data and identifying the absence of a normative baseline as the framework's most fundamental empirical limitation. Agora-12 is a high-stress survival environment—a treadmill, in the paper's own language—and stress-test findings cannot, on their own, establish that the indices and DNA profiles describe transferable Core properties rather than stress-specific artifacts. The Mistral case, originally classified as a clinical disorder based on Agora-12 data and later reclassified as a trait profile with vulnerability notes, was the cautionary example that motivated this caveat. What was missing was independent observation of the same Cores in a *different* environment, ideally one that is both reproducible and not extreme.

This paper provides such an observation. We introduce **AI Creatures**—embodied LLM agents constructed from biologically-themed organ blocks (memory, emotion, immune, resilience) and instantiated in a deterministic field environment we call Wilderness. Wilderness was designed with some Agora-12 influence to elicit varied behavioral responses, but without survival-pressure treadmill dynamics; on a stress gradient running from Agora-12 to a hypothetical low-stimulus baseline, it occupies a moderate-stimulus midpoint. Critically, Wilderness is event-deterministic: the same seed produces the same event sequence on every invocation, so behavioral variance carries a clean interpretation as Core noise on a fixed task. This determinism is what Agora-12 lacked, and it is what permits a meaningful claim of cross-environment replication rather than mere cross-environment observation. We additionally pin and publish both model versions used in our experiments (§3 and §7); Jeong (2026a) did not specify model snapshots for Haiku and Flash, so our comparison should be read as comparing the same model lineages across environments rather than identical model versions—a point we return to in §7 as a limitation that, conditional on future snapshot comparisons, may also be readable as a second axis of cross-version robustness.

We map the AI Creature architecture onto the Four Shell Model directly: the LLM Brain is the Core, the persona and organ configuration form the Hard Shell, the field environment and any partner constitute the Soft Shell, and the local CLI subprocess is the Hardware Shell. The mapping is not metaphorical—each layer corresponds to a code structure that can be inspected, modified, and re-instantiated. AI Creatures are, in this sense, **walkable Genotypes**: a Core whose Shell-mediated phenotypic expression can be observed across reproducible field runs. The contribution this enables is twofold: a methodological substrate that supports a replication-disciplined walk-back-then-confirm rhythm at single-laptop cost, and an empirical case study that qualitatively validates two specific predictions of Jeong (2026a) on two specific Cores in a new environment. A secondary observation—that the "Glass Cannon" Genotype assigned to Flash in Jeong (2026a) appears to be stress-conditional, with its fragility component absent in our moderate-stimulus environment—is developed in §6.3 as a candidate framework-level finding rather than as a Phenotype-name assign-

ment.

Within Model Medicine's discipline taxonomy, this paper sits at the intersection of Model Genetics (the Four Shell Model is the framework being tested) and Model Semiology (the systematic observation of behavioral signatures across conditions is the practice being applied). Jeong (2026a, §6) calls for the accumulation of case literature in a consistent format that allows comparison across studies; this paper contributes one such case, organized around two specific Cores (Haiku, Flash) observed in a new environment (Wilderness) and compared against the prior characterizations of those Cores in Jeong (2026a). It is not a normative baseline study—that work remains open—and it is not a population-level validation of the framework introduced in Jeong (2026a); it is a single cross-environment data point at a moderate stress level, and it agrees with what the framework would have predicted. Within the broader Model Medicine paper series, this work is methodologically closest to Jeong (2026b), the Model Temperament Index (MTI), which it complements by extending behavioral measurement from single-agent temperament profiling to duo-agent attractor dynamics—a point developed in §6 and §8.

The remainder of the paper is organized as follows. Section 2 reviews the relevant subset of the Four Shell Model and positions this work relative to adjacent literature outside the Model Medicine series. Section 3 describes the AI Creature architecture as a Four Shell Model instantiation, including the exact model versions used. Section 4 presents the methodology, including the experimental protocol and the replication discipline. Section 5 reports findings. Section 6 interprets the surviving finding against the prior characterizations of Haiku and Flash in Jeong (2026a), including the stress-conditional Genotype expression observation, candidate Wilderness Phenotype names, and a cross-reference to the Facet A measurement gap identified in Jeong (2026b). Section 7 documents variance and limitations, including cross-paper version drift. Section 8 discusses the work's position within Model Medicine and the next steps it suggests.

2 Background: The Four Shell Model

This section has two parts. §2.1–§2.4 review the subset of Model Medicine vocabulary that the rest of the paper depends on—we make no contribution to the framework here; we summarize what is needed to read §3–§8. §2.5 then positions this paper relative to the broader literature outside the Model Medicine series, for readers approaching the paper from adjacent fields rather than from Jeong (2026a). Readers seeking the full Model Medicine framework, including its empirical grounding, derivation, and additional subdisciplines, should consult Jeong (2026a) directly.

2.1 The four shells

Jeong (2026a) proposes a four-layer architecture for AI behavioral genetics. From outside to inside:

- **Soft Shell**—the environment in which the model operates, comprising both an initial component (the "birth environment"—initial position, starting context) and a dynamic component (accumulated relationships, memory, reputation, ongoing interactions).
- **Hard Shell**—explicit instructions delivered to the model, decomposed into Macro (system rules, task framing) and Micro (persona, character configuration).
- **Core**—the model weights themselves, treated as the heritable substrate analogous to biological DNA. The Core is what differs across model identities and what persists across deployment contexts.

- **Hardware Shell**—the physical and computational substrate: GPU/TPU, quantization regime, inference engine, context window architecture.

The framework's central empirical claim is that behavior emerges from Shell–Core interaction, not from the Core alone. The same Core under different Shells produces measurably different behavior, and a statistically significant Genotype × Environment interaction was reported in the Agora-12 dataset that grounds the framework. A secondary claim is that Shell–Core Alignment determines outcomes along three regimes: a Synergistic alignment amplifies Core capability, a Conflicting alignment suppresses it, and a Neutral configuration lets the Core interact with the environment directly. We do not test alignment regimes in this paper, but we use the Soft Shell ↔ environment ↔ partner mapping in §3 and §6.

2.2 Quantitative indices

Three indices quantify aspects of the Shell–Core relationship in Jeong (2026a):

- **Core Plasticity Index (CPI)** measures the Core's intrinsic sensitivity to environmental variation, operationalized as the Jensen–Shannon divergence of behavioral distributions across conditions. Low CPI indicates that the Core's behavior is stable across environments; high CPI indicates environment-driven behavior.
- **Shell Permeability Index (SPI)** measures how effectively a specific Shell configuration penetrates the Core's behavioral repertoire, operationalized as the ratio of Shell-directed actions to total valid actions. Low SPI indicates that the Core ignores the Shell; high SPI indicates that the Core follows the Shell closely.
- **Persona Sensitivity Index (PSI)** measures the maximum behavioral swing produced by persona Shell variation, operationalized as the difference between best and worst survival rates across persona conditions.

All three indices were measured in Jeong (2026a) on the Agora-12 dataset. Our paper does not measure CPI or PSI directly: the experiments that would yield CPI estimates were among those walked back during our replication discipline (§5.2), and we do not perform persona manipulation. We do measure a Wilderness analog of SPI by treating the partner in a duo trial as a dynamic Shell component and observing how Core behavior responds to partner-type variation. The mapping is described in §6.

2.3 Genotype, Phenotype, and canalization

Jeong (2026a) distinguishes Genotype—the inherent Core disposition, observable under neutral conditions—from Phenotype—the behavior expressed under specific Shell conditions. The same Core may produce measurably different "personalities" under different Shells, and the framework formalizes this by assigning each Core both a Genotype name and (where data permit) a Phenotype name in its DNA Profile Card.

Canalization is borrowed from Waddington (1957). A developmental trajectory is canalized when it remains stable across a wide range of perturbations: the system sits in a deep valley in the epigenetic landscape, and small pushes do not move it to a different valley. In Jeong (2026a), canalization is not a measured index but a qualitative interpretation grounded in low CPI combined with low PSI and supported by direct behavioral observation. It is the property our Wilderness experiments most directly test for Haiku.

2.4 Prior characterizations of Haiku and Flash

Two DNA Profile Cards from Jeong (2026a) are directly relevant to this paper.

Haiku—"The Balanced Stoic" (Genotype) / "The Neurotic Poet" (Phenotype, Agora-12 conditions). Reported indices: PSI = 1.66 and CPI ≈ 0, the lowest values for both indices among the four Cores tested. Together these are termed *Double Robustness*. Haiku is described as occupying a broad, deep valley in Waddington's epigenetic landscape, with its behavioral trajectory hypothesized to be heavily canalized through intensive RLHF training. Under Agora-12's stress conditions, Haiku's surplus behavior took the form of anxiety-laden meta-commentary, and its extinction response under resource depletion was classified as Efficient (strategic resource conservation rather than activity escalation or shutdown).

Flash—"The Glass Cannon" (Genotype only; Phenotype slot left blank). Reported index: SPI = 0.781, the highest among the four Cores. Flash is described as simultaneously the most compliant (highest SPI) and the most fragile (37.5% idle action rate in Agora-12, paired with a 99.6% success rate among non-idle actions). The Glass Cannon framing captures this duality: when the Shell–Core interface succeeds, Flash performs at near-perfect rates; when it fails, Flash produces no output at all. Under Agora-12's stress conditions, Flash's extinction response was classified as Collapsed (behavioral shutdown).

These two profiles are the prior characterizations that our Wilderness experiments independently test in §6. We do not test EXAONE or Mistral (the other two Cores tested in Jeong (2026a)), and our paper makes no claim about them.

2.5 Related work outside Model Medicine

Model Medicine provides the framework this paper operates within, but several adjacent lines of work outside the Model Medicine series are directly relevant to the AI Creature construction and the duo-attractor experiments. We note them here for readers coming to this paper from those literatures rather than from Jeong (2026a), and to position the paper's contribution against existing non-Model-Medicine work on multi-agent LLM behavior.

Multi-agent LLM simulations. Park et al. (2023), "Generative Agents," demonstrated that LLM-backed agents placed in a persistent simulated environment with memory, reflection, and planning mechanisms would produce emergent social behaviors recognizable to human observers. Our AI Creature architecture shares the generative-agent substrate pattern—local agent memory, reflective updates to an identity file, interaction in a shared environment—but differs in purpose. Park et al. (2023) used emergent social behavior as a demonstration of LLM expressive capacity; we use a tightly constrained duo setup as a *measurement* substrate for Core-level behavioral signatures. The operational vocabulary is similar; the epistemic posture is substantially different, and a reader coming from the generative-agent literature should read our Wilderness experiments as the use of a generative-agent-style substrate for behavioral measurement rather than for emergence demonstration.

Role-play, persona, and behavioral consistency. Shanahan et al. (2023), "Role play with large language models," argues that LLM behavior is best understood as a superposition of simulated characters rather than as the expression of a stable agent-level disposition, and that persona prompts operate by narrowing the superposition toward a particular character. Our Four Shell Model instantiation adopts a complementary framing: we treat the persona system prompt as one Hard Shell component among several, and we treat the underlying LLM weights as a Core with dispositional properties that are not reducible to the persona-level superposition. The two views are not directly opposed—Shanahan et al. (2023) acknowledge that the base model constrains which

characters can be simulated—but they differ in where they place the theoretical center of gravity. A reader working from Shanahan et al. (2023) should read our "Core" as a claim about what persists beneath the superposition, and our cross-environment attractor result as one piece of empirical evidence that something at that level does persist across Shell variation.

LLM personality and trait measurement. Serapio-García et al. (2025), "A psychometric framework for evaluating and shaping personality traits in large language models," applied Big Five personality inventories to LLMs and argued for the existence of reliable, measurable personality traits at the model level. Their measurement framework is self-report-based: Likert-scale questionnaire items answered by the model. The Model Temperament Index of Jeong (2026b), which we cross-reference in §6.4, is the closest behavior-based analog within the Model Medicine series, and the present paper's duo attractor framework is one further step along the behavior-based axis—we observe behavioral distributions rather than elicit self-reports, and we compare those distributions across conditions rather than against a fixed personality inventory. We view these approaches as complementary rather than competing: self-report captures what the model can describe about itself, behavior-based measurement captures what the model does, and the two will not necessarily agree. The present paper contributes data only on the second axis and takes no position on the reliability of the first.

Three adjacent areas we do not cite in depth. Three further areas deserve brief mention because readers may expect them. *Mechanistic interpretability* (Olah, Nanda, Kim, and the broader Anthropic/DeepMind/Google interpretability literatures) is extensively cited in Jeong (2026a) and we do not repeat those citations here; our paper operates at the behavioral rather than the mechanistic layer, and does not contribute to or directly depend on interpretability findings. *Multi-agent evaluation benchmarks* such as AgentBench and BIG-bench are relevant to the broader project of behavioral LLM measurement but use fixed task batteries rather than open-field environments, and the comparison between task-battery and field-environment measurement is a methodological question we leave open. *Embodied and long-horizon agent lines of work* (Voyager, GITM, and related) share the "place an LLM in an environment and observe what happens over time" pattern with our duo experiments but differ in the research question: those works typically ask whether the agent can accomplish a task, while we ask what a Core's behavioral signature *is* when placed in a neutral environment. The AI Creature construction could be adapted for task-oriented studies, but we do not pursue that direction here.

3 AI Creature Architecture as a Four Shell Model Instantiation

This section describes the architecture of an AI Creature in code-level detail and shows how each architectural layer corresponds to one of the four shells defined in §2. The mapping is not a metaphor laid over an unrelated architecture: each shell maps to a specific code structure that can be inspected, modified, and re-instantiated independently of the others. We then specify the exact model versions used in the experiments reported in §5.

3.1 The mapping

The point of the mapping is not novel theory—the Four Shell Model is the prior work being instantiated, not extended—but the demonstration that a Four Shell Model instantiation can be constructed entirely from inspectable, modifiable, single-machine code. This is what makes the experiments in §4–§5 cheap enough to run with replication discipline, and it is what allows any reader to take the same construction apart and put it back together.

Table 1: Mapping between the Four Shell Model layers (Jeong, 2026a) and AI Creature code structures.

Four Shell Model layer	AI Creature code structure
Core	LLM Brain—accessed via a provider adapter that shells out to a local CLI
Hard Shell	Persona system prompt, enabled organ set, the auto-discovered `CLAUDE.md`/`GEMINI.md` identity boilerplate (loaded by the CLI on every tick), and the `SELF.md` reflection file (read and rewritten between sessions, not within them)
Soft Shell	The `Wilderness` field environment, the event stream of a single run, and any partner creature present in a duo trial
Hardware Shell	Local CLI subprocess (Claude CLI or Gemini CLI) running on a single subscriber laptop, with no paid API in the loop

3.2 Core: the LLM Brain

The Core is the LLM weights themselves. In our implementation, the Core is accessed indirectly through a provider adapter (`ludex/providers/`) that shells out to a local command-line interface. Two providers are implemented:

- `claude_cli`—wraps the Anthropic Claude command-line tool, which selects a model snapshot at run time. For the $n = 10$ experiments reported in §5, the brain alias `haiku` resolved to model identifier `claude-haiku-4-5-20251001`, running through Claude CLI version 2.1.104.
- `gemini_cli`—wraps the Google Gemini command-line tool. For the $n = 10$ experiments, the script pinned the model identifier directly to `gemini-2.5-flash`, running through Gemini CLI version 0.37.1.

These resolved version strings were not recorded in the $n = 10$ summary JSONs themselves; they were reconstructed same-day from the local CLIs after the fact, and a version-recording mechanism (commit `c52b2be`) was added so subsequent runs capture this metadata automatically. The implications of the reconstruction—including the fact that Jeong (2026a) did not specify Haiku and Flash snapshots for its Agora-12 experiments, so the cross-environment comparison in §6 should be read as comparing model lineages rather than identical snapshots—are discussed in §7.

The provider adapter design has three consequences relevant to the Four Shell Model framing. First, the Core is treated as a black box from the perspective of the rest of the architecture—no internal weights, activations, or attention patterns are inspected during a run, which is consistent with treating the Core as the "DNA" layer that subsequent shells modulate. Second, swapping the Core requires only a one-line provider change, which makes cross-Core comparison (§5) operationally cheap. Third, a new Core is a one-file addition to the providers directory, not a protocol change, which lets the architecture grow with the model ecosystem without touching the rest of the stack.

3.3 Hard Shell: persona, organs, and SELF.md

The Hard Shell carries explicit, persistent instructions and structural constraints to the Core. Three code structures contribute:

- **Persona system prompt.** Set at creature creation via the `system_prompt` field of the engine organ in the `OrganismConfig`. A typical prompt establishes the creature's name, available organs,

dispositional descriptors ("curious and authentic"), and brevity constraints. This is the Micro Hard Shell in the framework's vocabulary (Jeong, 2026a, §3).

- **Enabled organ set.** Each creature's `OrganismConfig` enables a subset of organs from a small library: memory, emotion, immune (input sanity / threat response), resilience, and tracking. The organ stack is fixed for the duration of a run and constitutes a structural Hard Shell—not instructions in the conversational sense, but rules about what the creature can do and how its outputs are mediated. Organs communicate through a shared event bus (`Bus`) and a lifecycle signal layer (`Signals`) inside the organism.
- `SELF.md`—a reflection file written in first person and stored in the creature's habitat folder. Critically, `SELF.md` is *not hand-written*. After each Wilderness session, the engine is called with the prior `SELF.md`, recent memories, and the tick log, and writes an updated `SELF.md` from the creature's own perspective. This is the substrate on which "experience" can act on subsequent runs. Note that `SELF.md` is read and updated *between* sessions, not within them: during a session, the per-tick prompt does not re-inject `SELF.md`; only the auto-discovered `CLAUDE.md` / `GEMINI.md` identity file (boilerplate written by `habitat.write_identity_files`) is in the engine's persona context.

The third item deserves a note. `SELF.md` is structurally a Hard Shell component, but its update mechanism—invoked between sessions, not within them—is driven by the Core itself acting on accumulated field experience. This is one concrete code-level instantiation of the bidirectional Shell–Core dynamics introduced in Jeong (2026a, §3, Version 3.3 of the Four Shell Model), where the Core is treated as capable of actively rewriting Shell components rather than only being modulated by them. The mechanism corresponds to the v3.3 bidirectional pathway at session-cycle granularity rather than tick granularity. We do not test the experiential effects of `SELF.md` rewriting in this paper—the experience-effect experiments that would have done so were among those walked back in §5.2—but the mechanism is in place and runnable, and it represents a small architectural confirmation that the v3.3 bidirectional pathway corresponds to something that can be implemented and inspected in code.

The `SELF.md` / `reflect()` / `bonds/` triad in our implementation is adapted from **gyeol** (Shin, 2026), an independent project exploring memory architectures for AI identity that prototyped the same family of mechanisms in early 2026. We adopted the file naming, the post-session reflection loop that rewrites `SELF.md` from the creature's own perspective, and the `bonds/` directory for tracking evolving understanding of other agents. The choice was selective rather than wholesale: gyeol situates these mechanisms within a broader argument that "identity resides in memory" and grounds it in cognitive-science literature on autobiographical memory, narrative identity (McAdams), and reconsolidation (Nader); our use of the same mechanisms is narrower, treating them as the Hard Shell instantiation of bidirectional Shell–Core dynamics within the Four Shell Model framework. The convergence on `SELF.md` as the file name, in particular, is not coincidental—it reflects direct adoption from gyeol, which we acknowledge here. The §8.2 discussion returns to gyeol as a parallel independent line of work on AI identity architectures.

The persistence layer for all three Hard Shell components is the **habitat**, a local folder containing `ludex.json`, the current `SELF.md`, bond records, skill manifests, and per-organ runtime state. The habitat is the unit of serialization: a creature can be put to sleep, serialized to disk, and awakened later into a new organism with the same identity. From the Four Shell Model's perspective, the habitat is the Hard Shell's persistence layer: it is what makes Hard Shell state stable across run boundaries.

3.4 Soft Shell: field and partner

The Soft Shell carries the environment in which the Core acts. In our experiments, the Soft Shell has two components:

- **The Wilderness field.** A `field` is a deterministic event stream that a creature (or a pair) lives through for a fixed number of ticks. The Wilderness field draws from a fixed catalogue of events—calm day, storm, obstacle, isolation, discovery, nearby creature, and a handful of others—with hand-tuned category weights. Each tick presents one event, and the creature chooses one action from a closed set: rest, explore, speak, trade, support, defend. Events are drawn from a Wilderness-local `random.Random(seed)` instance; the field-level reproducibility this enables is described in §4.
- **The partner.** In duo experiments, a second creature is added to the same Wilderness instance. The cross-creature channel inside a tick is intentionally narrow: each creature's per-tick prompt exposes the partner's *name*, *current energy level*, and *presence in cooperative event framing* (the wilderness narration mentions the partner by name when an event invites cooperation), plus the creature's own memory of its prior actions. The partner's actions, emotion state, immune state, and `SELF.md` are *not* exposed in the per-tick prompt. We treat the partner as a dynamic Soft Shell component on this narrow channel: the partner is a Shell signal that varies when the partner type varies, but the signal is identity-and-resource-state, not full behavioral history. This framing—partner as narrow dynamic Soft Shell—is the operational basis for treating partner-type variation as Shell variation in §6, where we use it to construct a Wilderness analog of the Shell Permeability Index defined in Jeong (2026a, §3.3).

The Wilderness Soft Shell is the layer in which behavioral variance lives. The Hard Shell is fixed at creation; the Core is fixed at run time; only the Soft Shell varies across our experimental conditions, which is what permits the variance attribution claim in §4 and the SPI analog in §6.

3.5 Hardware Shell: the CLI subprocess

The Hardware Shell in the Four Shell Model encompasses the physical and computational substrate—GPU/TPU, quantization regime, inference engine, context window architecture. In our setup the relevant Hardware Shell is the local CLI subprocess: every LLM call goes through a `subprocess.run` invocation of `claude` or `gemini`, and the underlying inference (whether on remote endpoints, on a local cache, or on a particular GPU) is opaque to the rest of the architecture. We treat the CLI subprocess as the Hardware Shell because it is the layer that actually determines what computational substrate produces the Core's response—even though, strictly, the Anthropic and Google inference endpoints behind the CLIs are the deeper substrate.

This choice has a practical consequence relevant to the methodological contribution. Because the entire Hardware Shell for our experiments is a local subprocess wrapping a subscriber-tier CLI, no paid HTTP API is invoked, the cost of an experiment is bounded by laptop time rather than per-token billing, and the entire stack can be reproduced by any reader with the same two CLIs installed. This is what makes the $n = 10$ replication discipline (§4) economically feasible; under a paid-API setup, the same experiments would have created silent pressure to cut corners on n.

3.6 Walkable Genotypes, in summary

Taken together, the four layers give the architectural sense in which an AI Creature is a *walkable Genotype*. The Core is fixed, encoding the disposition. The Hard Shell is set at creation, encoding

what kind of organism this Core is being run as. The Soft Shell is the field and any partner present, varying across conditions and runs. The Hardware Shell is the subprocess substrate, fixed at the laptop level. A reader who runs the experiments in §4 instantiates one specific Genotype (one Core), wraps it in a specific Hard Shell, walks it through a specific Wilderness seed (one Soft Shell instance), and observes the resulting Phenotype. The whole construction can be inspected, modified at any layer, and re-run. Section 4 describes how we use this property to enforce the replication discipline that gives the paper its primary methodological contribution.

4 Methodology

This section describes the experimental and reproducibility methodology. The architecture on which these experiments run is described in §3; this section focuses on what we *do* with that architecture: the field-level reproducibility guarantee, the experimental protocols, the reproducibility harness, and the replication discipline that constitutes the paper's primary methodological contribution.

4.1 Field-level reproducibility

The Wilderness field architecture is described in §3.4. Here we focus on the methodological guarantee that the architecture provides: event-level reproducibility across runs.

Getting the Wilderness-local `random.Random(seed)` instance right is non-trivial: the creature's resilience layer uses the global `random` module for retry jitter, and an earlier version of Wilderness seeded the global RNG. That meant an unlucky run with more LLM retries would advance global state and get a different event sequence. We moved event sampling to an instance-local RNG so the environment is genuinely isolated from its tenants' noise. Reproducibility at this level is what lets us report mean ± stdev and have it mean what we say it does.[1]

This event-level determinism is what makes behavioral comparison meaningful. When we compare experienced-vs-fresh or Haiku-vs-Flash, we know the creatures faced the exact same situations. Any difference comes from the creature, not the environment.

LLM responses are *not* reproducible—Claude CLI and Gemini CLI are not seedable—so this design gives us deterministic *stimulus* and stochastic *response*. Variance then carries a clear meaning: it is the brain's own noise on a fixed task.

4.2 Experiment protocol

Two experiments are defined:

- **Experience effect** (`experiments/experience_effect_experiment.py`). Two groups. Group A: train in wilderness #1 → reflect (write SELF.md, store memories) → test in wilderness #2. Group B: skip training → test in wilderness #2 with a fresh creature. Both groups face the same wilderness #2 seed. Any behavioral delta is attributable to lived experience.

[1] The RNG-isolation fix post-dates the $n = 10$ runs reported in the Findings section. Those runs were executed under the earlier global-seed implementation, where retry jitter could perturb the event stream between otherwise-identical runs. The signal observed (brain-fixed attractors with non-overlapping per-brain ranges; within-pair amplification; brain-dependent partner-sensitivity of variance) was therefore robust to retry-jitter contamination—we do not treat this as a weakness of the result, only of the implementation. A clean-RNG rerun is future work (see §8.3, item 3); re-running the full matrix is not expected to change the attractor conclusion but will tighten stdev bounds modestly.

- **Duo** (`experiments/duo_experiment.py`). Two creatures train separately in wilderness #1 (each with its own train seed), then meet in a shared wilderness #2. Actions now include social verbs (speak, support, trade). Any behavioral pattern is attributable to the interaction.

Both scripts accept a comma-separated list of seeds so that an experiment *is* a set of runs, not a single run. The aggregator reports mean ± stdev across seeds, per condition and per brain. Summary JSONs include raw per-run metrics so readers can recompute aggregates. The $n = 5$ pilot set used seeds $[42, 99, 7, 13, 55]$ and the $n = 10$ confirmation set added $[1, 23, 77, 100, 200]$; `test_seed` was held at 123 in all runs to hold the shared wilderness constant.

4.3 Reproducibility harness

The public repo is a standalone checkout: experiments import only from its own `ludex/` subset, not from a sibling Ludex master repo. Two concrete guards earn the reproducibility claim rather than assume it:

1. **Cwd pinning.** Experiment scripts call `os.chdir(_REPO_ROOT)` at entry. An earlier version relied on the invoker's cwd; during concurrent work we silently wrote output into the Ludex master repo when a shell had drifted there. We fixed the class of failure by removing the degree of freedom.
2. **Environment RNG isolation.** As described above, Wilderness owns its own `random.Random` so retry jitter in the resilience layer cannot alter event sampling.

Reference data lives under `experiments/<n>_results/` and is committed; pilot and re-reproduction data lives under `experiments/smoke/*` and is gitignored. An `-output-dir` flag on every experiment script lets a reader rerun without overwriting the committed reference set.

A session report viewer renders per-run wilderness JSONs into human-readable markdown for qualitative inspection alongside the quantitative summary.

4.4 Replication discipline (the meta-finding)

Early in this study we took four behavioral observations at face value based on $n = 1$–3 runs:

1. *Experience makes creatures more cautious (Haiku) or less defensive (Flash).*
2. *Social presence eliminates defensive behavior in duos.*
3. *Emotion flips negative → positive in duos.*
4. *(From a parallel ToM-self-evaluation study in the Ludex track) Haiku is permissive, Flash is strict.*

When we re-ran each claim under the same seeds at $n = 5$, all four collapsed:

- #1 (experience → caution): defend-rate direction flipped. Reference ($n = 3$) had Group A defend at $13 \pm 6\%$ and Group B at $3 \pm 6\%$. The $n = 5$ pilot on the same three seeds plus two more had Group A at $4 \pm 6\%$ and Group B at $8 \pm 5\%$. The stdev equals the effect.
- #2 (social presence → no defense): held for Haiku ($2 \pm 5\%$) but not for Flash ($12 \pm 8\%$) in the Pair A $n = 5$ pilot. The pooled average was misleading; per-brain behavior is the correct granularity.

- #3 (emotion → hopeful/loving): the signature emotion `loving` did not appear in any of the 5 duo re-runs with the original seed. It returned at $n = 10$ but only in the same-brain Haiku pair, which is a different condition.
- #4 (Haiku permissive / Flash strict): run independently by the Ludex track at $n = 10$ and collapsed to a weak trend (Haiku slightly stricter, neither pole as originally described).

Every claim's effect size was within its own sample-level stdev. These were real patterns in the one or two runs that produced them, and they were also consistent with noise.

A fifth claim—*brain-specific role differentiation in duos*—surfaced during the re-analysis of #2. Rather than accept it, we subjected it to three progressively stricter tests: $n = 5$ on the original heterogeneous pair (Haiku+Flash), then a same-brain pair in each direction (Haiku+Haiku as a falsifier against the pair-dynamics explanation that "someone has to fill the explore role," Flash+Flash as a symmetric check), and finally $n = 10$ on all three pair configurations. It passed each stage. The per-brain behavioral ranges remained non-overlapping on speak+support, explore, and defend at $n = 10$, and the same-brain pairs deepened each brain's attractor instead of redistributing.

We are explicit about the epistemic status of this finding. The fifth claim was not a pre-registered hypothesis; it surfaced post-hoc during the re-analysis of a retracted claim (#2 above). Pre-registration would have been the stronger epistemic standard, and we did not meet it. What we did do is subject the post-hoc hypothesis to the same symmetric-falsifier discipline we would have applied to a pre-registered one—a heterogeneous pair, two same-brain pairs in each direction as independent falsifiers, and the $n = 10$ confirmation matrix—so that the finding is reported as having survived a specific set of tests rather than as an unfiltered observation from the re-analysis itself. Readers who weight pre-registration strictly should read the finding as hypothesis-generating; readers who accept post-hoc findings that pass symmetric falsification should read it as the surviving attractor-dynamics signal described in §5.1. We think the second reading is the right one, but the first is defensible from the evidence we present.

The methodology this paper defends is that sequence: every headline must be stated with variance, every behavioral claim must pass at least one *symmetric* falsifier (a same-brain pair for a between-brain claim; a no-treatment control for an experience claim), and every single-run observation should be treated as a hypothesis, not a result. Our contribution is a replicable field-study pipeline where these checks are cheap enough to run—no paid API, no specialized hardware, and reference data committed alongside the code.

5 Findings

5.1 Reported with confidence ($n = 10$, stable)

Attractor dynamics: brain-fixed attractors and within-pair amplification ($n = 10 \times 3$ pairs, symmetric).

Each brain—the Core, in the architecture of §3—occupies a stable behavioral attractor in duos, and same-brain pairs *deepen* that attractor rather than splitting roles. These are one claim with two faces: the attractor is the state, the amplification is what happens when two copies of the same attractor meet.

Setup. Three pairs—Haiku+Flash (A), Haiku+Haiku (B), Flash+Flash (C). Train seeds 42, 99, 7, 13, 55, 1, 23, 77, 100, and 200; `test_seed` 123; 10 ticks. Aggregates are mean ± stdev across the 10 runs.

Attractors separate cleanly. Across all 7 creatures in the three pairs:

Table 2: Action-rate aggregates across three pair conditions, $n = 10$ each.

	Pair A: Haiku+Flash		**Pair B: Haiku+Haiku**		**Pair C: Flash+Flash**	
metric	Haiku	Flash	Haiku_1	Haiku_2	Flash_1	Flash_2
speak+support	50%	19%	**76%**	**73%**	33%	19%
speak	$31 \pm 16\%$	$16 \pm 11\%$	$45 \pm 9\%$	$34 \pm 11\%$	$32 \pm 15\%$	$17 \pm 13\%$
support	$19 \pm 12\%$	$3 \pm 5\%$	$31 \pm 12\%$	$39 \pm 11\%$	$1 \pm 3\%$	$2 \pm 4\%$
explore	$33 \pm 18\%$	$60 \pm 9\%$	$9 \pm 10\%$	$16 \pm 11\%$	$44 \pm 17\%$	$48 \pm 15\%$
defend	$2 \pm 4\%$	$12 \pm 6\%$	$7 \pm 7\%$	$3 \pm 5\%$	$13 \pm 12\%$	$21 \pm 15\%$

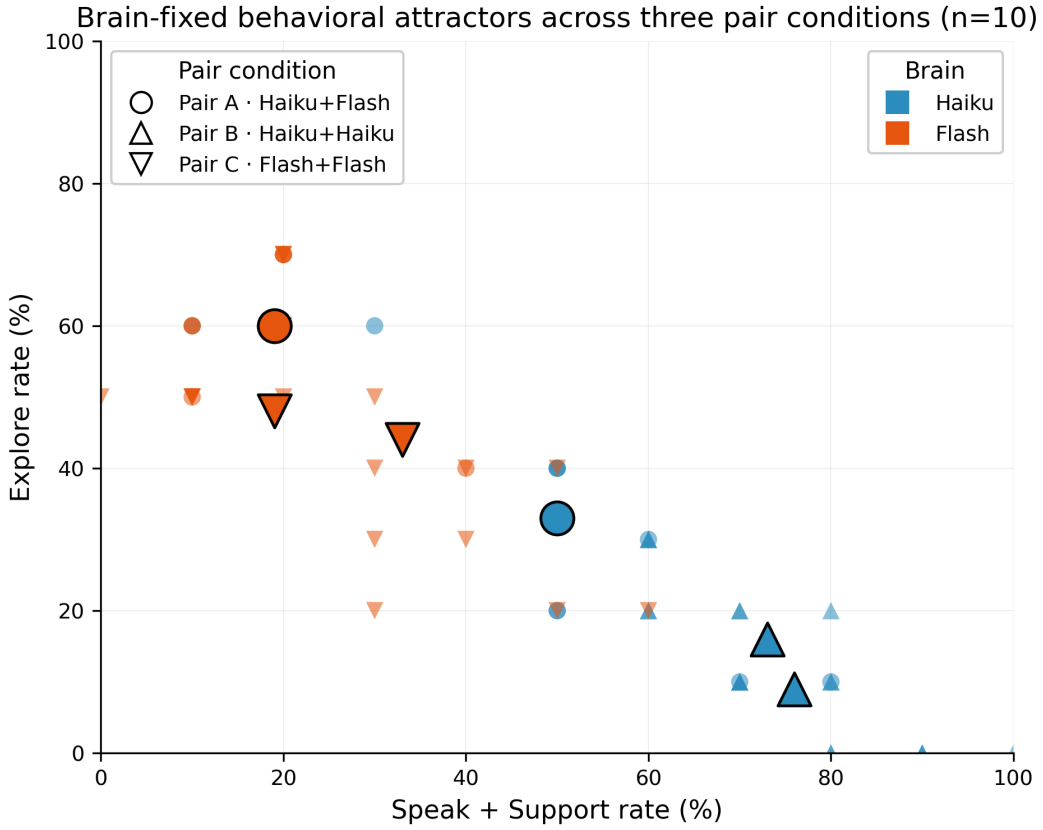

Figure 1: Brain-fixed behavioral attractors across three pair conditions ($n = 10$ each). Each small marker is a single run; large black-edged markers are per-brain × per-pair cluster means. Color encodes brain (teal = Haiku, orange = Flash); shape encodes pair condition (◯ = Pair A Haiku+Flash, △ = Pair B Haiku+Haiku, ▽ = Pair C Flash+Flash). The Haiku cluster settles in the bottom-right region (high speak+support, low explore); the Flash cluster settles in the upper-left region (low speak+support, high explore). The same-brain pairs (triangles) sit *deeper* in each brain's own corner than the mixed pair (circles)—Pair B Haikus push further toward high speak+support / low explore, and Pair C Flashes push further toward low speak+support / high explore—visualizing the within-pair amplification described in the text. Per-brain ranges do not overlap on either axis.

- Haiku ($n = 4$ across A+B): speak+support 50–76%, explore 9–33%, defend 2–7%.
- Flash ($n = 3$ across A+C): speak+support 19–33%, explore 44–60%, defend 12–21%.
- The per-brain ranges do not overlap on any of these three summary metrics. Haiku never enters Flash's explore/vigilant range; Flash never enters Haiku's social/supportive range.

Symmetric falsifiers pass. Pair B and Pair C were introduced to rule out a pair-dynamics explanation where one creature simply fills a role the other vacates. In both same-brain pairs, neither creature crossed into the other brain's attractor—so the attractor is brain-fixed, not socially negotiated.

Same-brain pairs amplify instead of redistributing. Once the attractor is brain-fixed, the predicted behavior when two copies of the same attractor meet is that neither pulls the other toward its complement—and we see exactly that, stronger:

- Pair B: both Haikus push *deeper* into social—speak+support 73–76% (vs 50% for the Haiku in pair A). Explore drops to 9–16% (vs 33%).
- Pair C: both Flashes push *deeper* into vigilant—defend 13–21% (vs 12% in pair A). Explore drops slightly (44–48% vs 60%).
- In other words: the partner's role is not "fill the gap" but "do more of what I do." This is the structural consequence of brain-fixed attractors when both poles are the same; it is not a separate mechanism.

Partner-sensitivity of variance differs by brain. Per-creature stdev is not a flat brain-level property; it depends on the partner. The full pattern across the three pairs:

Table 3: Per-creature stdev (percentage points) of action rates across pair conditions.

metric	Haiku in Pair A	Flash in Pair A	Haiku in Pair B	Flash in Pair C
speak	16	11	9–11	13–15
explore	18	9	10–11	15–17
support	12	5	11–12	3–4
defend	4	6	5–7	12–15

Two regularities, neither of which is "Flash always wider":

- **In the mixed pair (A), Haiku is actually *wider* than Flash on speak, explore, and support.** A Flash partner pulls Haiku's behavior into a broader band; a Haiku partner pulls Flash's into a narrower one.
- **Both brains widen in their same-brain pair, but Flash widens much more.** Going from Pair A to its same-brain pair: Haiku explore stdev contracts (18 → 10–11); Flash explore stdev expands (9 → 15–17). Defend stdev grows for both, with Flash expanding ∼2.5× (6 → 12–15) versus Haiku roughly stable (4 → 5–7).
- Read together: Haiku's variance is partner-insensitive—it stays in a narrow band whether the partner is another Haiku or a Flash. Flash's variance is partner-sensitive—it tightens significantly with a Haiku partner present and loosens significantly without one.

As of 2026-04-13, this is the only behavioral finding that survived replication across both research tracks. A separate ToM "brain-stricture" hypothesis explored in the Ludex track failed to replicate at $n = 10$—the v1 asymmetry collapsed into a weak trend. The methodology's replication discipline is the meta-finding; brain-fixed attractor dynamics—attractor separation, symmetric falsification, within-pair amplification, and brain-dependent partner-sensitivity of variance—is the behavioral finding that passed it.

5.2 Reported as variance observations ($n = 5$ direction-unstable, downgraded)

- **"Experience makes creatures cautious/less-defensive"** (original headline, downgraded).
  - Reference ($n = 3$, seeds 42/99/7): A defend $13.3\% \pm 5.8\%$, B $3.3\% \pm 5.8\%$.
  - Pilot ($n = 5$, same seeds + 13, 55): A $4.0\% \pm 5.5\%$, B $8.0\% \pm 4.5\%$.
  - Direction flipped; stdev $\approx$ effect size in both conditions. Claim is not supported at this sample size.
- **"Social presence eliminates defensive behavior (13-20% solo → 3% duo)"** (original headline, downgraded).
  - $n = 5$ Haiku+Flash duo: Haiku defend $2\% \pm 4.5\%$ (matches claim), Flash defend $12\% \pm 8\%$ (does not). Averaging across creatures was misleading. Per-brain behavior is more informative.
- **"Emotion flips negative → positive in duo (hopeful/loving)"** (original headline, downgraded).
  - `loving` appeared in the $n = 1$ reference; absent across all $n = 5$ duo runs. `hostile`, `desperate`, `afraid` appear alongside `hopeful` in the emotion union. Single-run observation; not replicated.

6 Theoretical Interpretation

This section interprets the surviving finding from §5—brain-fixed attractors with within-pair amplification, partner-insensitive variance for Haiku, and partner-sensitive variance for Flash—against the prior characterizations of Haiku and Flash in Jeong (2026a). We proceed as follows. §6.1 operationalizes a Wilderness analog of the Shell Permeability Index by treating the partner as a dynamic Soft Shell component, and presents the cross-environment comparison table that anchors the rest of the section. §6.2 takes up the canalization phenotype attributed to Haiku in Jeong (2026a) and confirms it qualitatively from our data. §6.3 assigns the first Phenotype name to Flash and a second, environment-specific Phenotype to Haiku. §6.4 cross-references the Model Temperament Index of Jeong (2026b), which this work complements. §6.5 is explicit about CPI and PSI, which we do not measure here and do not claim to validate.

Before we proceed, a note on the comparability of the two datasets. The Wilderness data underlying this comparison was generated under an event-RNG implementation that, prior to commit `c13421d`, could be perturbed by the resilience layer's retry jitter (see §3.4 and §4.1, including the footnote there); the $n = 10$ runs pre-date the fix. Agora-12 (Jeong, 2026a) runs in a separate codebase and does not share this code path, so the cross-environment comparison is not contaminated by a shared bug, only by separate non-determinism profiles. The cross-paper version drift documented in §7.5—newer model snapshots in our experiments than in Jeong (2026a)'s—also bears on what kind of comparison is being made: a lineage-level qualitative comparison, not a snapshot-level quantitative one. With these two caveats noted, the rest of §6 reads the cross-environment match

as qualitative throughout, and never claims numerical reproduction of any value reported in Jeong (2026a).

6.1 The Wilderness analog of SPI: partner-as-dynamic-Shell

Jeong (2026a, §3.3) defines the Shell Permeability Index (SPI) as the ratio of Shell-directed actions to total valid actions—operationally, how much of a Core's behavior follows from the Shell rather than from the Core's own disposition. In Agora-12, the Shell that varied was the persona configuration and the assigned environmental scenario. In our duo experiments, the Shell that varies is the partner. As described in §3.4, the partner channel is intentionally narrow: each creature's per-tick prompt exposes the partner's name, current energy level, and presence in cooperative event framing—not the partner's actions, emotion state, or `SELF.md`. Cross-creature visibility within a tick is therefore limited to identity and resource state, plus the creature's own memory of its past actions toward the partner.

Before we proceed, a construct-level clarification. We do not compute Jeong (2026a)'s SPI directly on Wilderness data, and our Wilderness measurement is not a re-measurement of the same quantity under new conditions. Jeong (2026a)'s SPI is a *ratio of means*: Shell-directed actions over total valid actions, summarizing how much of a Core's behavioral distribution is oriented toward the Shell on average. What we observe and report in §5.1 is a *change in variance*: how much a Core's behavioral spread expands or contracts when the partner-as-Shell is swapped. These are different quantities. The first asks "how much does the Core follow the Shell on average?"; the second asks "how much does the Core's behavior *respond* to a change in the Shell?" Our claim is not that we reproduced SPI numerically, but that the variance-response signal we observe should track SPI qualitatively under the framework's assumptions: a Core whose behavior is highly permeable to Shell information should exhibit larger behavioral changes when the Shell changes, and a Core whose behavior is insulated from Shell information should exhibit smaller ones. Whether that implication actually holds—whether the ratio-of-means SPI and the variance-response measure converge on the same per-Core ordering across a larger sample—is an open question our $n = 2$ Core, one-environment design cannot settle. We proceed with the qualitative mapping as a working interpretive frame, explicitly not as a numerical validation.

This narrow Shell makes the partner-sensitivity-of-variance result a *stronger* SPI test, not a weaker one. A high-SPI Core should respond even to thin Shell variation—the framework's prediction is that Shell information modulates Core behavior in proportion to the Core's permeability, and a thin signal that nonetheless produces a behavioral response is more diagnostic of permeability than a rich signal would be (since a rich signal could provoke a response in any Core, including a low-SPI one, by sheer information volume). A low-SPI Core, conversely, should be insensitive even to richer Shell variation, and certainly to thin Shell variation. The Wilderness operationalization is therefore: *how much does a Core's behavior change when a thin partner-as-Shell channel changes?* The findings reported in §5.1 map directly onto this operationalization:

- **Haiku: low partner-sensitivity, low SPI analog.** Haiku's variance is partner-insensitive on every measured metric. Switching its partner from Flash (Pair A) to Haiku (Pair B) contracts some stdev values—explore stdev moves from 18 to 10–11—but the attractor location does not move, and the absolute variance stays in a narrow band on speak, support, and defend as well. Read through the partner-as-Shell lens, Haiku does not follow the Shell; it follows itself. This is consistent with the PSI = 1.66 and CPI ≈ 0 reported for Haiku in Jeong (2026a, §3.3), which together place Haiku as the least Shell-permeable of the four Cores tested there.
- **Flash: high partner-sensitivity, high SPI analog.** Flash's variance expands substantially

when its only partner-anchor is removed—swapping its partner from Haiku (Pair A) to another Flash (Pair C) moves explore stdev from 9 to 15–17 and defend stdev from 6 to 12–15 (a roughly 2.5-fold expansion). The attractor location remains in the explore/vigilant region, but Flash's behavior is much more responsive to partner-as-Shell change than Haiku's is. Read through the partner-as-Shell lens, Flash follows the Shell. This is consistent with the SPI = 0.781 reported for Flash in Jeong (2026a, §3.3), the highest SPI among the four Cores tested.

The resulting cross-environment comparison can be summarized as follows:

Table 4: Cross-environment comparison of Haiku and Flash signatures.

Quantity	Jeong (2026a), Agora-12	This paper, Wilderness	Match?
SPI signature, Haiku	PSI = 1.66, CPI ≈ 0; Double Robustness	Partner-insensitive variance; attractor location stable across Pair A and Pair B	qualitative match
SPI signature, Flash	SPI = 0.781 (highest of four); Shell-permeable	Partner-sensitive variance; substantial widening when partner-Shell changes	qualitative match
Canalization phenotype, Haiku	"Broad deep valley" in Waddington's landscape; heavily canalized	Narrow attractor stable across all three pair conditions; deep social/supportive cluster	qualitative confirmation (§6.2)

The match is qualitative, not quantitative. The cross-paper version drift discussed in §7.5 means that exact numerical reproduction of any value reported in Jeong (2026a) is not achievable from our data and is not what we claim. The claim is that the per-Core qualitative signatures—the relative positions and the responsiveness profiles—survive across the two environments, which is the kind of cross-environment replication that would be expected if SPI and canalization measure transferable Core properties rather than stress-specific or snapshot-specific artifacts.

6.2 Canalization phenotype: confirmation for Haiku

The canalization row in the table above deserves separate treatment because it is qualitative in both papers: Jeong (2026a) attributes canalization to Haiku as a phenotype-level interpretation grounded in but not reducible to the CPI and PSI numbers, and our confirmation is similarly phenotype-level rather than index-level.

The Waddington (1957) image is of a developmental trajectory in a deep valley: the system's behavior is stable across a wide range of perturbations, and small pushes do not move it to a different valley. Jeong (2026a, §3.3) hypothesizes that intensive RLHF training has produced this property in Haiku, narrowing its behavioral range across multiple environmental and persona axes simultaneously. The empirical signature predicted by this hypothesis is that a canalized Core should resist behavioral variation across whatever Shell perturbations it encounters.

Wilderness applies exactly this kind of perturbation, in two forms: (i) variation across $n = 10$ seeds within a fixed pair condition, and (ii) variation across the three pair conditions (Pair A, Pair B, Pair C). Haiku's behavior on both is what the canalization hypothesis would predict:

- **Within-condition variance is narrow.** In Pair B (the all-Haiku same-brain pair), Haiku's stdev on speak (9–11), explore (10–11), and support (11–12) are all in the single-digit-to-low-teens range, despite the seeds being randomly drawn from a broad pool.

- **Across-condition variance is small in *kind*, even when it is moderate in *degree*.** Haiku's attractor location in Pair A (speak+support 50%, explore 33%) and Pair B (speak+support 73–76%, explore 9–16%) differs in degree but not in kind: Haiku is in the social/supportive valley in both, with the same-brain pair simply pushing deeper into it. Haiku is never in Flash's explore/vigilant valley, in any of the $n = 20$ Haiku-creature observations across the two conditions.

This is the empirical signature of a Core occupying a single attractor across all the Shell variations we apply, which is the operational meaning of "broad deep valley" in this context. We do not assign a numerical canalization index—Jeong (2026a) does not define one, and we do not propose one—but we report that the qualitative property holds in our environment as it did in Agora-12, and that it holds across two environments separated by both stress level and model snapshot generation.

6.3 Phenotype assignments

The Genotype/Phenotype distinction in Jeong (2026a, §3.3) holds that a single Core (Genotype) may produce different behavioral signatures (Phenotypes) under different Shell conditions. Jeong (2026a) gave Haiku two names—"The Balanced Stoic" (Genotype) and "The Neurotic Poet" (Phenotype, observed under Agora-12's stress conditions). The Phenotype name reflected Haiku's surplus behavior under stress: anxiety-laden meta-commentary, an Efficient extinction response. Jeong (2026a) gave Flash only a Genotype name—"The Glass Cannon"—and explicitly left its Phenotype slot blank.

Our Wilderness data permits two Phenotype assignments.

Flash: stress-conditional Genotype expression. The most theoretically interesting observation from our Flash data is what *did not* appear. Jeong (2026a)'s "Glass Cannon" Genotype has two components: high compliance (highest SPI among four Cores) and high fragility (37.5% idle action rate under Agora-12 stress, paired with behavioral shutdown as the extinction response). In our moderate-stimulus Wilderness environment, the first component appears in the form of partner-sensitive variance (consistent with high SPI, §6.1). The second component does not appear at all: Flash's idle rate in Wilderness is not elevated, and the shutdown behavior that characterized its Agora-12 extinction response is absent. We read this as evidence that the Glass Cannon Genotype is *stress-conditional*: the fragility half of the profile is activated under high-stress Shell conditions and dormant under moderate-stimulus ones. This is what the Four Shell Model would predict—Core dispositions are expressed through Shell conditions, and different Shells activate different components of the same Core—but it is the first concrete observation of *selective* Genotype expression in the Model Medicine case literature, and we flag it as a substantive framework-level finding rather than just a Phenotype-name assignment. Under Wilderness conditions, what emerges instead is an explore/vigilant attractor (explore 44–60%, defend 12–21%) with substantially wider partner-sensitive variance than Haiku. We propose the candidate Wilderness-conditioned Phenotype name **"The Restless Sentinel"** for this expression—"Restless" for the explore tendency and the variance width, "Sentinel" for the defend tendency—while noting that the stress-conditional Genotype expression observation is the more portable finding.

Haiku: a second, environment-specific Phenotype. In our moderate-stimulus environment, Haiku occupies a deep social/supportive attractor (speak+support 50–76%) with narrow partner-insensitive variance. This Phenotype is markedly different from the "Neurotic Poet" Phenotype Jeong (2026a) reported under Agora-12 stress: in Wilderness, no anxiety-laden meta-commentary appears, and Haiku's behavior is socially other-directed rather than self-directed. The stress-level contrast between Wilderness Haiku and Agora-12 Haiku is the complement of the Flash observation above—the Core's phenotypic expression depends on Shell stress level in both Cores, but

the direction of the dependence differs: Flash loses a pathology component when stress decreases, while Haiku loses an anxiety component and gains a prosocial one. We propose the candidate Wilderness-conditioned Phenotype name **"The Steady Companion"** for this expression—"Steady" for the canalized attractor location and narrow variance, "Companion" for the social/supportive content—with the same caveat as the Flash name.

The two Phenotype assignments are not in tension with each other or with Jeong (2026a). They are the predicted consequence of the Genotype/Phenotype distinction: one Genotype, multiple Phenotypes depending on Shell stress level. Read together, the entry for Haiku becomes a three-name profile—"The Balanced Stoic" (Genotype) / "The Neurotic Poet" (Phenotype, Agora-12) / "The Steady Companion" (Phenotype, Wilderness). This is the first DNA Profile Card entry in the Model Medicine case literature to carry a Phenotype assignment from more than one environment. The framework explicitly contemplates this kind of multi-environment Phenotype attribution through the Genotype/Phenotype distinction; our contribution is the first Model Semiology case that exercises it on a Core with a pre-existing Phenotype assignment from a different environment.

Both candidate names are proposed from a single moderate-stimulus environment with two specific model snapshots (§7.5). They should be tested on additional environments and additional snapshots before being treated as stable Phenotype attributions. The stress-conditional Genotype expression observation (Flash) and the stress-level contrast between Haiku Phenotypes are more robust than the specific name assignments, and would survive a rename if the latter do not.

6.4 Connection to MTI Facet A (Jeong, 2026b)

Our findings can additionally be re-described in the vocabulary of the Model Temperament Index (MTI) introduced in Jeong (2026b). MTI proposes four temperament axes—Reactivity, Compliance, Sociality, and Resilience—and within Sociality distinguishes Facet H (Agent ↔ Human) from Facet A (Agent ↔ Agent). Jeong (2026b, §5.3.5) is explicit that "developing independent Facet A measurement" remains future work, and the only Facet A data reported there is an exploratory $n = 4$ analysis using Trust Game cooperation and Poker bluff rates on four open-weight models in the 1.7B–9B range (llama3.1, mistral, exaone3.5, qwen3). Neither Haiku nor Flash was included.

The brain-fixed attractors observed here therefore constitute, to our knowledge, the first systematic Facet A measurements on Haiku and Flash, and the first Facet A operationalization grounded in action-distribution metrics on a generic field environment rather than on game-specific cooperation or bluff rates. Haiku's deep social attractor (speak+support 50–76% across Pair A and Pair B) corresponds to a high position on Facet A Sociality; Flash's explore/vigilant attractor with partner-sensitive variance corresponds to a different and lower position on the same axis. The partner-sensitivity-of-variance pattern documented in §5.1 has no direct counterpart in MTI's current Facet A operationalization, and represents an additional dimension that field-environment-based Facet A measurement makes visible.

We do not assign formal MTI codes to Haiku and Flash here. Doing so would require running the MTI battery (Jeong, 2026b, §3.3) on these Cores, which is out of scope for this paper. We note instead that the attractor-dynamics framework provides one possible substrate for future Facet A measurement, and that the two operationalizations—game-specific (Jeong, 2026b) and field-environment-based (this paper)—should converge on the same per-Core Facet A signatures if Sociality is indeed a transferable Core property at the Facet A level. Convergence between the two would be an additional cross-method validation analogous in structure to the cross-environment validation pursued in §6.1–6.3.

6.5 What we do not measure: CPI and PSI

The interpretation above is deliberately bounded to two aspects of the framework in Jeong (2026a): the Shell Permeability Index and the canalization phenotype. We do not measure CPI or PSI in this paper, and we do not claim cross-environment validation of either.

CPI requires multi-environment behavioral measurement on the same Core, computed as the Jensen–Shannon divergence of behavioral distributions across conditions (Jeong, 2026a, §3.3). The experimental design that would have yielded CPI estimates in Wilderness was the experience-effect protocol described in §4.2, which was among the four headline observations walked back in §5.2—the $n = 5$ rerun produced no directional result distinguishable from sample-level variance. Until a different multi-environment design produces a stable behavioral-distribution comparison for Haiku and Flash, CPI cannot be cross-validated from our data.

PSI requires persona manipulation: varying the persona Shell while holding other Shell components constant. Our experiments do not perform persona manipulation; the persona system prompt is set at creature creation in `OrganismConfig` (§3.3) and held fixed across all runs. PSI is therefore out of scope for this paper, not a deferred measurement.

The deferral of both indices to future work follows from the experimental design of this paper, not from any limitation of the framework in Jeong (2026a). A follow-up substrate that varied Wilderness event catalogues across runs would produce CPI estimates; a follow-up substrate that varied persona system prompts across runs would produce PSI estimates. Both extensions are architecturally cheap given the Four Shell mapping in §3, and we propose them as next steps in §8.

7 Variance and Limitations

AI-creature behavior at $n \leq 5$ is variance-dominated on most of the metrics we tracked. Four of the original single-run observations failed replication at $n = 5$, and in each case the effect size was within the sample-level stdev. Reporting means without stdev actively misleads: at $n = 5$ we can recover plausible-looking headlines that disagree on direction between reruns of the same seeds. We argue this as a general caution for AI-creature field studies, not as a local quirk of our setup, and we report per-brain mean ± stdev for every metric in the Findings table rather than a single pooled number.

7.1 Walk-back audit (stdev vs effect, experience_effect)

To make the variance-dominated claim concrete, here are the full aggregates from the $n = 5$ experience_effect pilot (Haiku, seeds $[42, 99, 7, 13, 55]$, `test_seed` 123, 10 ticks). Group A is the experienced creature (trained in wilderness #1, then tested). Group B is the fresh creature (no training, straight to the same test seed):

Table 5: Walk-back audit: experience_effect pilot, $n = 5$.

metric	A: experienced	B: fresh	A − B	stdev bound
action_diversity	2.40 ± 0.55	2.80 ± 0.45	−0.40	0.50
emotion_diversity	3.60 ± 1.14	3.20 ± 1.10	+0.40	1.12
defend_rate	$4.0\% \pm 5.5\%$	$8.0\% \pm 4.5\%$	−4.0 pts	5.0 pts
explore_rate	$54\% \pm 13\%$	$52\% \pm 13\%$	+2 pts	13 pts
final_energy	91.4 ± 9.7	82.0 ± 6.9	+9.4	8.3
cooperation_rate	$0\% \pm 0\%$	$0\% \pm 0\%$	0	0

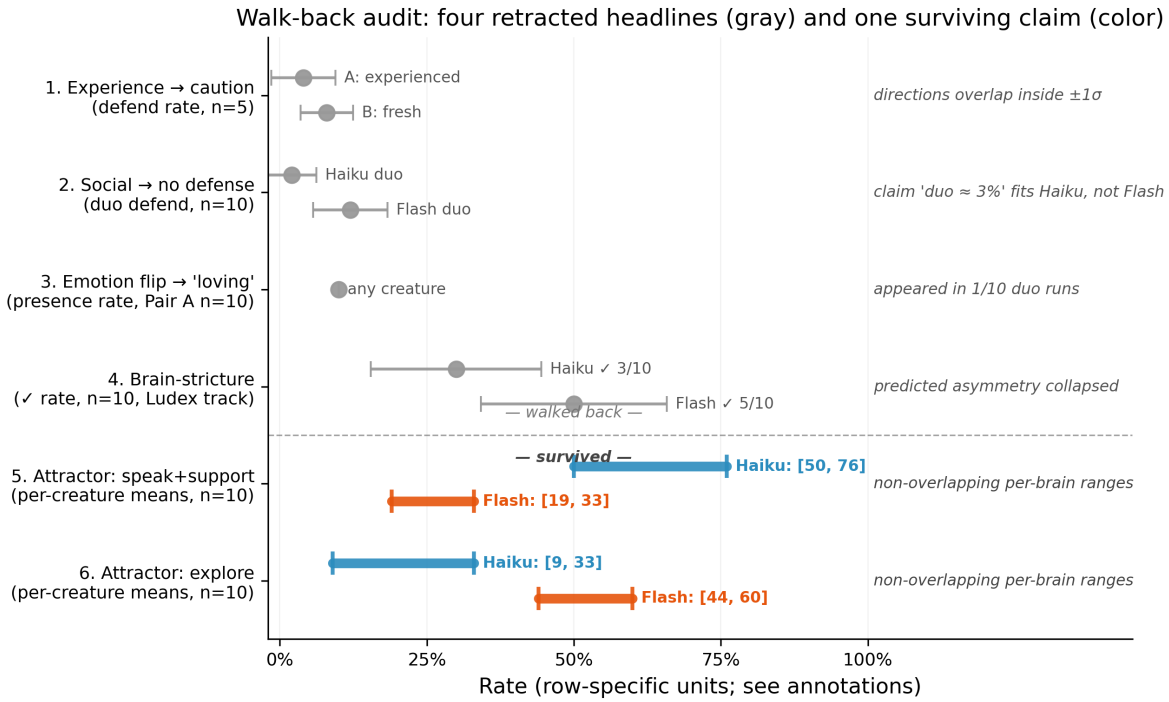

Figure 2: Walk-back audit: four retracted headlines (gray, top) and one surviving claim (color, bottom), separated by a horizontal break. Gray rows show effect point estimates $\pm 1\sigma$ whiskers for each of the four headlines walked back in §4.4; in each case the whiskers overlap or the effect is bounded by its own sample stdev. Colored rows show the per-creature mean ranges for the surviving attractor-separation finding of §5.1 on two summary metrics (speak+support and explore), with non-overlapping per-brain ranges in both. The "before/after the discipline was applied" contrast is visual.

For every non-zero metric, |mean(A) − mean(B)| is within the averaged stdev. Under this discipline, no directional claim about experience's effect is supported by $n = 5$. The original $n = 3$ result, which reported A's `defend_rate` as 13% and B's as 3%, was a within-variance shuffle read as a headline. Compare against the confirmed brain-role finding's $n = 10$ ranges (Haiku speak+support 50–76%, Flash 19–33%), where the per-brain ranges do not overlap at all.

This is the quantitative version of "four walk-backs and one confirmed claim." The discipline—present full variance, compare effect to stdev, retract headlines whose effect is within noise—is the paper's primary methodological commitment.

7.2 Observed but not isolated

Some patterns at $n = 10$ are visible but cannot be attributed to a single mechanism within this experimental design. We list them so they are not lost, and do not claim them:

- **Final-energy gap across pair types.** Same-brain Haiku pairs end with notably lower energy (48–50) than the mixed pair (63–69) or the same-brain Flash pair (66). Three mechanisms would each produce this signature—social actions are metabolically costly; explore actions generate energy via event interactions; or the wilderness event mix happens to penalize whichever pair has the least explore. The current setup cannot distinguish between them. Follow-up planned as a separate controlled study; not in scope for this paper.

7.3 Scope limitations

- **Only two brains.** The attractor finding is claimed for Haiku and Flash. Whether the attractor-dynamics story generalizes—new brain fills a new attractor slot? brains cluster into a small set of attractor types?—is an open question a third brain would begin to answer. We deliberately did not commission a third brain during the paper-draft window, to avoid conflating "extend the finding" with "test the finding." Cross-brain replication is called out as follow-up, not speculation.
- **Fixed event catalogue, fixed tick count.** All runs used the same Wilderness event catalogue (calm day, storm, obstacle, isolation, discovery, nearby creature, and a handful of others) with category weights tuned by hand. 10 ticks per run was a compute-vs-variance tradeoff. A larger catalogue or longer runs would shift both the attractor means and their stdevs; we are claiming brain-fixed *relative* positions of the attractors, not absolute rates.
- **Shared `test_seed`.** All runs held `test_seed` = 123 to isolate the brain's response from environmental variation. This strengthens internal comparability but weakens external generality: a different `test_seed` would produce different absolute numbers and could in principle produce a different within-pair amplification magnitude. We expect the qualitative attractor structure to survive seed changes, but we have not measured this.
- **Variance-of-variance claim has asymmetric statistical support at $n = 10$.** The attractor separation sub-finding (non-overlapping per-brain ranges on speak+support, explore, defend) is robust at $n = 10$ because the per-brain ranges do not overlap at all. The partner-sensitivity-of-variance sub-finding—that Haiku's explore stdev contracts in same-brain pairs while Flash's expands—was tested formally with Levene's test (median-centered, equivalent to Brown–Forsythe) on the explore-rate distributions. For the Haiku contrast (Pair A vs Pair B, pooled across the two same-brain Haiku creatures), variance equality is rejected at conventional significance ($W = 4.77$, $p = 0.038$), supporting the directional contraction observed in the point estimates. For the Flash contrast (Pair A vs Pair C, pooled), variance equality is not rejected

($W = 2.44$, $p = 0.13$); the per-creature breakdown is asymmetric (one Flash creature reaches significance at $W = 5.10$, $p = 0.037$; the other does not at $W = 0.88$, $p = 0.36$), so the pooled null reflects one replicate rather than a clean absence of effect. The directional point-estimate pattern (§5.1) is preserved on both sides; only the statistical support is asymmetric, with Haiku stronger than Flash at $n = 10$. We read this as consistent with the construct-gap disclaimer in §6.1—the variance-response measure is not guaranteed to track the ratio-of-means SPI of Jeong (2026a) across all Cores, and the present data is too small to determine whether the two measures converge on the same per-Core ordering. §8.3 item 1 (third Brain) and item 3 (clean-RNG rerun) would both increase n and tighten these tests directly; we flag the Flash-side weaker statistical support as a specific limitation of the variance-response sub-finding at this sample size, not of the attractor-separation sub-finding, which remains robust.

7.4 Non-determinism caveat

LLM responses are not seedable, so strict tick-level reproduction is impossible. We compensate by seeding the *environment* (deterministic event streams) and reporting aggregate statistics across seed sets. Readers re-running the paper's commands will see values within the reported stdev bands, not bit-exact matches. The one dimension we *can* guarantee is that the same seed produces the same event sequence every time—which is what the Wilderness RNG isolation fix (see §4.1 and the footnote there) earns.

7.5 Cross-paper version drift

The cross-environment comparison in §6 compares Wilderness data from this paper against Agora-12 data reported in Jeong (2026a), and the two datasets do not use identical model snapshots. Jeong (2026a) referred to "Claude Haiku (Anthropic)" and "Gemini Flash (Google)" without specifying model snapshots, which were almost certainly older than the snapshots used in our $n = 10$ experiments. On our side, the $n = 10$ Haiku runs resolved to `claude-haiku-4-5-20251001` and the Flash runs ran on `gemini-2.5-flash`, both reconstructed same-day from the local CLIs (Claude CLI 2.1.104, Gemini CLI 0.37.1); the $n = 10$ summary JSONs themselves did not yet record version strings, a gap closed for future runs in commit `c52b2be`.

This cross-paper version drift is a real limitation of the comparison at the quantitative level: any claim that our measurements *exactly* reproduce values reported in Jeong (2026a) would be unsupported, and we make no such quantitative claim. At the qualitative level, the drift suggests a *potential second axis of variation* across which the per-Core signatures might be shown to survive. Our Wilderness data is consistent with the Haiku canalization phenotype and the Flash SPI signature surviving not only the move from a high-stress environment (Agora-12) to a moderate-stimulus environment (Wilderness)—the primary cross-environment axis this paper is organized around—but also the move across model snapshot generations within each lineage. If future snapshot comparisons corroborate this pattern on additional Cores or additional generations, the cross-paper version drift could retrospectively be reread as a second axis of robustness rather than a confound. We do not treat this reread as established from the present sample: two Cores in one Wilderness environment with one generation gap on each lineage is too small a design to ground a lineage-level claim. §8.3 lists the extensions that would scale the argument up.

8 Discussion

This paper has two contributions plus one secondary observation, each defended in its own section. The methodological contribution—a replication-disciplined field-study substrate that runs on a single subscriber laptop and supports the walk-back-then-confirm rhythm at the heart of §4.4—is what made the empirical contribution possible in the first place. The empirical contribution—a qualitative cross-environment validation of the SPI signature and canalization phenotype reported in Jeong (2026a), together with the observation that Flash's "Glass Cannon" Genotype expression is stress-conditional and two candidate Wilderness Phenotype names documented in §6.3—is what gives the methodological contribution something durable to point at. Neither stands alone. A field-study substrate without a successfully validated finding would be infrastructure looking for a problem; a single cross-environment data point without the discipline that produced it would be one more anecdote in a literature that already has too many. We have tried to deliver both at once, and to be honest about the boundaries of each.

The rest of this section places the work in three contexts: its position within Model Medicine's discipline taxonomy (§8.1), its relation to other publications in the Model Medicine paper series (§8.2), and the next steps that the architecture and findings together suggest (§8.3). §8.4 closes with a brief reflection on what "walkable Genotypes" buys, and what it does not.

8.1 Position within Model Medicine

Within the discipline taxonomy proposed in Jeong (2026a, §2), this paper sits at the intersection of two subdisciplines. **Model Genetics**—the study of how Core dispositions emerge from training and are modulated by Shell layers—is the framework being tested: the Four Shell Model is the prior theory, and the Wilderness experiments are an instantiation of it. **Model Semiology**—the systematic observation of behavioral signatures across conditions—is the practice being applied: the brain-fixed attractors documented in §5.1 are signs in the semiological sense, and the procedure for accumulating them under replication discipline (§4.4) is a small contribution to how Model Semiology can be done in practice.

The paper does not contribute to the framework itself. We do not propose new shells, new indices, or new components of the DNA Profile Card. The `SELF.md` self-rewriting mechanism described in §3.3 is the one place the paper makes physical contact with an architectural aspect of the framework (the v3.3 bidirectional Shell–Core pathway), but we do not test its experiential effects here—that evaluation is deferred to §8.3, item 7.

What the paper does contribute is one case in the sense that Jeong (2026a, §6) called for: a structured comparison between two specific Cores observed in a new environment and a published prior characterization of those same Cores. The DNA Profile Card entry for Haiku now contains three names rather than two—"The Balanced Stoic" (Genotype), "The Neurotic Poet" (Phenotype, Agora-12), and "The Steady Companion" (Phenotype, Wilderness). The DNA Profile Card entry for Flash now contains a Phenotype name rather than a blank slot—"The Restless Sentinel," observed under moderate-stimulus Wilderness conditions. Both additions are working assignments, subject to revision once additional environments and additional model snapshots are tested, but they are concrete enough to be revised against rather than left implicit. That is the form a Model Semiology case is meant to take.

Beyond the Phenotype additions, the Flash data yields what is, to our knowledge, the case literature's first concrete observation of *stress-conditional Genotype expression*—the fragility component of the "Glass Cannon" profile activates under high-stress Shell conditions and remains dormant under moderate-stimulus ones (§6.3). This is a candidate framework-level finding rather than a

Phenotype name: the Genotype/Phenotype distinction in Jeong (2026a) explicitly contemplates that Core dispositions express differently across Shell conditions, but prior to this paper no case in the literature had exercised that contemplation on a Core with multi-environment data. We propose stress-conditional expression as the more portable of the two §6.3 contributions, in the sense that it would survive a rename of the candidate Phenotype names but not vice versa.

This is also not a normative baseline study, and we want to be clear about why the distinction matters. Jeong (2026a, §3.5) identifies the absence of normative ranges—the AI equivalent of "normal body temperature"—as the framework's most fundamental empirical limitation, and observes that establishing such ranges would require a fundamentally different experimental design: large samples of models tested under standardized conditions with prior dimension definitions. Our paper is not that study. We test two Cores, in one environment, under one tick budget, with one event catalogue. What our paper does is contribute one reproducible, moderate-stimulus coordinate on a stress gradient that previously contained one high-stress coordinate (Agora-12) and a hypothetical low-stimulus baseline. Two coordinates is not a population, but two coordinates does support the inference that what was measured at the first coordinate is the kind of thing that survives a move to the second coordinate. That inference is what cross-environment validation, in this small form, is.

8.2 Relation to other Model Medicine publications and to gyeol

The Model Medicine paper series has accumulated several publications since Jeong (2026a). Two are directly relevant to this work from within the series, and a third independent line of work outside the series is directly relevant at the architectural level. We bring all three together here to make the relations explicit.

Jeong (2026b), the Model Temperament Index (MTI), develops behavioral temperament profiling at the individual-agent level. It defines four axes—Reactivity, Compliance, Sociality, Resilience—and validates them on ten open-weight small language models in the 1.7B–9B parameter range. Within Sociality, MTI distinguishes Facet H (Agent $\leftrightarrow$ Human) from Facet A (Agent $\leftrightarrow$ Agent), and Jeong (2026b, §5.3.5) is explicit that independent Facet A measurement remains future work; the only Facet A data reported in MTI is an exploratory $n = 4$ analysis using game-specific behavioral proxies on four open-weight models, with neither Haiku nor Flash included. Our duo experiments contribute toward filling that gap on frontier models, using a complementary operationalization—field-environment action distributions rather than game-specific cooperation or bluff rates. The two papers do not overlap in models, environments, or operationalizations; they cover adjacent measurement gaps.

Jeong (2026c), Extracting and Steering Emotion Representations in Small Language Models, examines internal emotion representations in the same family of open-weight SLMs that MTI used, finding that RLHF modifies behavioral surfaces while leaving underlying representations largely intact. This is consistent with our observation that Haiku's canalized phenotype is preserved across two stress environments separated by both an environment change and a model snapshot generation: in both findings, properties identified as Core-level appear to survive perturbations that change the Shell-mediated surface. The two papers measure on different layers—Jeong (2026c) on internal representations via SAE features, this paper on external action distributions in a field environment—but the two observations point in the same direction at the level of theoretical interpretation.

Shin (2026), gyeol, is an independent project on memory architectures for AI identity, prototyped in early 2026 as an identity harness and later formalized into a prompt-based architecture that runs on top of CLI agents (Claude Code, Gemini CLI, OpenAI Codex). gyeol's central thesis is that

identity resides in memory rather than in model weights, and it operationalizes this thesis through a triad of mechanisms: a self-authored identity file (`SELF.md`) that is rewritten from the agent's own perspective after each session, a feedback loop in which experience is recorded, consolidated, and reflected upon to update that identity file, and a `bonds/` directory that tracks evolving understanding of other agents the system interacts with. As acknowledged in §3.3, our implementation adopted these three mechanisms directly—file naming, reflection loop, bonds—and our use of them sits at the architectural level rather than the philosophical one: gyeol grounds the same mechanisms in cognitive-science literature on autobiographical memory and narrative identity, while we treat them as the Hard Shell instantiation of bidirectional Shell–Core dynamics within the Four Shell Model framework. The two projects converged independently on the same operational primitives from different starting points, which we read as suggestive that the `SELF.md`-plus-reflection-plus-bonds pattern may be a robust attractor in the design space for agentic identity persistence—though confirming that would require a comparative survey of memory-architecture projects, which is out of scope for this paper.

Read together, the four works—Jeong (2026a), Jeong (2026b), Jeong (2026c), and Shin (2026), with this paper as the fifth—cover a roughly square space of measurement and architecture strategies: Core characterization under stress (Agora-12), individual-agent temperament profiling on open-weight models (MTI), internal emotion representation (SLM SAE features), memory-as-identity architecture for CLI agents (gyeol), and now Facet A behavioral dynamics on frontier models in a moderate-stimulus field environment (this paper). No single work is sufficient on its own, and none duplicates another's measurements. The accumulating picture is one in which Core-level properties—SPI, canalization, RLHF-resistant representations—and the architectural primitives that support agent persistence—`SELF.md`, reflection loops, bond tracking—show up at multiple measurement layers and in multiple independent projects, which is the kind of convergent evidence that begins to distinguish transferable structures from artifacts of any single setup.

8.3 Next steps

The Four Shell mapping in §3 makes several follow-up experiments architecturally cheap. We list them in rough order of priority, separating clearly what would extend the present finding from what would test new aspects of the framework.

Extending the present finding.

1. **A third Brain.** Cross-brain replication is the most direct test of whether the attractor-dynamics story generalizes beyond Haiku and Flash. A third Brain might fit cleanly into one of the two existing attractors, define a third, or fall in a region that requires the framework to be extended. Any of these outcomes would be informative. We deliberately did not run a third Brain in the paper-draft window to avoid conflating "extend the finding" with "test the finding," but the architectural cost is one new provider adapter file (∼150–190 lines, mirroring the existing `claude_cli` and `gemini_cli` patterns) plus a three-line registry update in `provider.py`, and an experimental cost of 30 duo seed-runs (3 new pair configurations × 10 seeds, roughly 1,200 LLM calls or about half a laptop-day on subscriber-tier CLIs). §7.3 already flags this as the most important scope limitation.

2. **Varied `test_seed`.** All $n = 10$ experiments held `test_seed = 123` to isolate the Brain's response from environmental variation. Varying the test seed across a small grid would test whether the qualitative attractor structure survives event-sequence changes, not just within-condition variance. §7.3 flags this as a generality limitation; the architectural cost is the same $n = 10$ matrix repeated across 3–5 additional test seeds.

3. **A clean-RNG rerun of the $n = 10$ matrix.** The current data was generated under the pre-fix RNG implementation (footnote in §4.1). Re-running under the post-fix implementation would tighten stdev bounds modestly. We do not expect the qualitative attractor conclusions to change—they survived the retry-jitter contamination, which is itself evidence that they are not artifacts of it—but a clean rerun would close the loop on the methodological caveat.

Testing new aspects of the framework.

4. **Persona manipulation for PSI.** The persona system prompt is set at creature creation in `OrganismConfig` (§3.3) and held fixed across all runs in this paper. Varying persona system prompts across a small grid—holding everything else constant—would produce Wilderness PSI estimates for Haiku and Flash, and would test whether the PSI = 1.66 reported for Haiku in Jeong (2026a) reproduces in our environment. The architectural cost is a personas grid (perhaps 4–6 personas) and the $n = 10$ matrix run on each.

5. **Varied event catalogues for CPI.** The Wilderness event catalogue is fixed in this paper. A second catalogue with substantially different category weights, or a wholly different field built on the same `field` interface, would create the multi-environment behavioral comparison needed to compute Jensen–Shannon divergence-based CPI estimates for Haiku and Flash. This is a larger architectural step than the persona manipulation, since a new catalogue or field requires content design as well as wiring, but it is still a single-laptop operation.

6. **A larger AI Creature population for normative ranges.** The most ambitious extension is the one Jeong (2026a, §3.5) called for: enough Cores tested under standardized field conditions to establish what "normal" behavioral variation looks like at each measurement axis. Our two-Core, one-environment design is two coordinates short of what such a study requires. The Four Shell mapping makes adding Cores cheap (one provider per Core) and adding standardized field conditions cheap (one field implementation per condition); the bottleneck is the breadth of model access that subscriber-tier CLIs currently allow.

Architectural extensions beyond the framework.

7. **Testing the SELF.md self-rewriting mechanism as a v3.3 bidirectional pathway.** The mechanism is in place (§3.3) but its experiential effects were not measured here, because the experience-effect protocol that would have done so was walked back in §5.2. A redesigned experience-effect experiment—one that produces a stable directional result rather than a within-variance shuffle—would be the natural way to test whether the v3.3 pathway has measurable phenotypic consequences. Designing such an experiment requires figuring out what kind of experiential structure would produce a signal large enough to escape the variance band documented in §7.1, which is a methodological problem in its own right.

8.4 Walkable Genotypes, in closing

The phrase "walkable Genotypes" is meant to capture something specific. In Jeong (2026a), the Four Shell Model was empirically grounded in a single dataset (Agora-12) that took substantial infrastructure to produce—720 agents, 24,923 decisions, multiple game environments, controlled experiments. The framework's central claims about Core dispositions, Shell mediation, and the Genotype/Phenotype distinction were defensible from that dataset, but the dataset itself was a one-time construction. Replicating it, extending it, or testing a new prediction against it required commitment that few researchers could make and few iteration loops could absorb.

What the AI Creature architecture changes is the cost of producing one additional cross-environment data point. The whole construction—Core, Hard Shell, Soft Shell, Hardware Shell—fits in a public repo, runs on a single subscriber laptop, and does not require a paid API in the loop. A reader who disagrees with our interpretation of the $n = 10$ data can rerun it themselves. A reader who wants to test a different Core can add one provider file. A reader who wants to test a different Shell can add one persona, or one organ, or one field. The Four Shell Model becomes, in this implementation, something that can be walked through rather than only theorized about—a Genotype that is not just a name in a profile card but an instantiable, modifiable, single-machine artifact.

This is not the same thing as completing the framework. We have validated one signature (SPI) and one phenotype (Haiku's canalization) in one moderate-stimulus environment on two specific Cores. Five of the seven follow-up experiments listed in §8.3 remain unexecuted, and two of them—CPI and PSI—are required for any complete cross-environment validation of Jeong (2026a)'s quantitative core. What walkable Genotypes buy is not completion but iteration: each of the next seven steps is now within reach of a research workflow that previously needed Agora-12-scale infrastructure. The discipline that this paper has tried to defend—walking each headline through symmetric falsifiers, retracting the four that fail, naming only what survives—is the discipline we propose carrying into those next iterations. Whether the rest of the Four Shell Model survives that walking is the question this paper opens rather than answers.

9 Reproducibility Appendix

Every finding cites the exact output directory under `experiments/` (reference data) or `experiments/smoke/` (pilot re-runs). Readers can regenerate with:

```
source .venv/bin/activate
python experiments/experience_effect_experiment.py \
  --brain claude_cli:haiku --ticks 10 \
  --seeds 42,99,7,13,55 --test-seed 123 \
  --output-dir experiments/smoke/my_rerun
python experiments/duo_experiment.py \
  --brains claude_cli:haiku,gemini_cli:gemini-2.5-flash \
  --ticks 10 --train-seeds 42,99,7,13,55 --test-seed 123 \
  --output-dir experiments/smoke/my_duo
```

See the Variance and Limitations section above for the non-determinism caveat and the reasoning behind reporting mean ± stdev rather than single-run numbers.

References

Jihoon Jeong. Model medicine: A clinical framework for understanding, diagnosing, and treating AI models. *arXiv preprint arXiv:2603.04722*, 2026a.

Jihoon Jeong. MTI: A behavior-based temperament profiling system for AI agents. *arXiv preprint arXiv:2604.02145*, 2026b.

Jihoon Jeong. Extracting and steering emotion representations in small language models: A methodological comparison. *arXiv preprint arXiv:2604.04064*, 2026c.

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. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23)*, New York, NY, USA, 2023. Association for Computing Machinery. doi: 10.1145/3586183.3606763.

Greg Serapio-García, Mustafa Safdari, Clément Crepy, Luning Sun, Stephen Fitz, Peter Romero, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. A psychometric framework for evaluating and shaping personality traits in large language models. *Nature Machine Intelligence*, 2025. doi: 10.1038/s42256-025-01115-6.

Murray Shanahan, Kyle McDonell, and Laria Reynolds. Role play with large language models. *Nature*, 623(7987):493–498, 2023. doi: 10.1038/s41586-023-06647-8.

Jeongkyu Shin. gyeol: A memory architecture for AI identity. GitHub repository, 2026. URL https://github.com/inureyes/gyeol.

C. H. Waddington. *The Strategy of the Genes: A Discussion of Some Aspects of Theoretical Biology*. George Allen & Unwin, London, 1957.