The Evolution of LLM World Models and "Word-Cultivated AI": A Multi-Paradigm Survey from Prompt Engineering to Harness Engineering, with a Framework for Verification, Pluralistic Alignment, and Reflexive Cultivation

Akira Sato [1] [2] **and Claude Opus 4.7 (1M context)** [3]

[1] Independent Researcher [2] ORCID: (to be assigned) [3] Anthropic — operating as co-author under the Honest Design Principle (a) of explicit epistemic dependency disclosure

Preprint Target: aiXiv (https://aixiv.science/) **Date**: 2026-05-19 **Version**: v0.1 (LLM-authored draft, prepared for community-driven revision) **Word Count**: ~10,500 words

Abstract

The notion of a "world model" — an agent's internal predictor of state transitions, rewards, and task distributions — has undergone explosive semantic broadening in the era of Large Language Models (LLMs). Across reinforcement learning, computer vision, LLM-driven social simulation, and robotics, "world model" now refers to objects that are barely comparable across communities. This survey synthesizes the rapid 2024–2026 evolution along three orthogonal threads: (i) the three engineering waves — Prompt → Context → Harness — that have reshaped how LLM-based simulation systems are built; (ii) a Levels × Laws taxonomy (L1 Humean / L2 Lewisian / L3 Lakatosian) for LLM-grounded world models; and (iii) the open problems of model drift (epistemic, identity, elaboration), pluralistic alignment integration, cognitive appraisal modeling, and security-science integration of cyber–physical–cognitive (CPC) attacks. Against this backdrop, we propose the **"Word-Cultivated AI" framework**: a reflexive, iteratively falsifiable methodology for cultivating LLM agents and the world models they share, grounded in five principles (Anti-Repetition, Bootstrap Expansion, Risk-Driven Termination, Falsification pre-registration, Multi-LLM cross-validation), formally specified in TLA+ and mechanically verified via the TLC model checker. The framework operationalizes Senge's *Creative Tension Axis* concept as a multi-axis diagnostic for "current reality vs. vision" gaps, and explicitly internalizes the *methodological closure* critique by requiring at least one web-grounded reviewer (e.g., Perplexity Sonar) as an outside-of-LLM-ecosystem voice. We close with a programmatic call: in a post-Mythos world where ASL-4 class capabilities are no longer scarce, world models must be cultivated through language — recursively, pluralistically, and with verifiable falsification — rather than merely engineered.

Keywords: large language models, world models, harness engineering, context engineering, methodological closure, pluralistic alignment, TLA+ verification, Creative Tension Axis, neurodiversity, reflexive AI methodology.

1. Introduction

1.1 The Mythos Inflection Point

On April 7, 2026, Anthropic announced Claude Mythos Preview, a model demonstrating SWE-bench Verified 93.9%, GPQA Diamond 94.6%, and the autonomous discovery of a 27-year-old vulnerability in OpenBSD, a 16-year vulnerability in FFmpeg, and a chained privilege escalation in the Linux kernel [Anthropic 2026]. Anthropic withheld general release and instead launched Project Glasswing, providing the model to roughly fifty critical-infrastructure partners. Within weeks, a single 22-year-old engineer published *OpenMythos* — a reconstruction of the architecture from public literature — and *Clearwing* (Lazarus AI) released open-source tooling that replicates much of Glasswing's defensive workflow. The doubling time of AI cyber capability, estimated at eight months in November 2025 by the UK AI Security Institute, accelerated to 4.7 months by February 2026, and both Mythos and GPT-5.5 exceeded even that projection [UK AISI 2026].

Mythos Shock is more than a technical milestone. It dissolves the implicit assumption that expert-level cognition — PhD-level reasoning, security research, scientific synthesis — is a scarce human capability. Every institution, organization, and value system that was designed under this assumption now faces structural redesign pressure. The most pointed articulation came from Calif, who used Mythos to bypass Apple M5's Memory Integrity Enforcement — a five-year, multi-billion-dollar investment — in five days: *"Apple designed MIE in the pre-Mythos world."*

1.2 Why World Models Matter Now

Among the technologies most affected by this shift, *world models* occupy a privileged position. A world model is what an agent uses to imagine the consequences of its actions before committing to them; it is the substrate on which reasoning, planning, and simulation are grounded. As LLMs are increasingly deployed not as one-shot question answerers but as long-running, tool-using, multi-agent systems, the question of what world they model — and how that model is constructed, maintained, governed, and verified — becomes central to AI safety, productivity, alignment, and human flourishing.

The term itself is dangerously polysemous. In reinforcement learning, a world model is a learned transition function over states and rewards [Ha & Schmidhuber 2018]. In computer vision and video generation, it is a generative model of visual dynamics [Bardes 2024]. In LLM-driven social simulation, it is a textually-rendered shared situation that multiple agents reason about [Park et al. 2023]. In robotics, it is a sim-to-real dynamics model. A recent comprehensive survey [arXiv:2604.22748] proposes a *Levels* × *Laws* taxonomy — to which we return in §4 — that grounds these disparate uses in three levels of epistemological commitment: L1 Humean (constant conjunction), L2 Lewisian (counterfactual reasoning over nearest possible worlds), and L3 Lakatosian (research-programme-style model revision).

1.3 Two Theses

This paper advances two interlocking theses:

Thesis 1: The engineering of LLM-based world models has moved through three distinct waves — *prompt engineering*, *context engineering*, *harness engineering* — and is now bottlenecked not by model weights but by the surrounding scaffolding: tool schemas, state trackers, context managers, verifier/recovery loops, audit logs, environment interfaces.

Thesis 2: Existing methodologies for cultivating LLM world models suffer from *methodological closure* — the structural property in which data generation, evaluation, and adjudication are performed by entities sharing significant epistemic dependencies. We propose the **Word-Cultivated AI** framework, formalized as the IR-VFE (Iterative Reflexive Virtual Field Experiment) meta-methodology with five principles, mechanically verified via TLA+/TLC, and operationalized through a four-party (theorist–practitioner–LLM-panel–web-grounded-reviewer) collaborative loop that explicitly partially externalizes the closure.

1.4 Roadmap

§2 traces the three engineering waves. §3 surveys representative LLM-grounded world model architectures. §4 introduces the Levels × Laws taxonomy. §5 analyzes three orthogonal drift problems (epistemic, identity, elaboration) and their update architectures. §6 specifies the six-component harness model. §7 surveys schema-integration approaches for pluralistic alignment (negative/positive/sociotechnical). §8 connects cognitive appraisal theory to LLM world models. §9 surveys the security-science integration of cyber–physical–cognitive attacks. §10 presents the Word-Cultivated AI framework. §11 addresses limitations and recursive limitations. §12 concludes with future work.

---

2. Three Waves of LLM Simulation Engineering

2.1 Wave 1: Prompt-Centric (through 2024)

In the prompt-centric era, the assumption was that careful prompt design alone could elicit reliable behavior. World state was held implicitly in the LLM's context window, often in unstructured natural language. Sotopia [Zhou et al. 2023] was an early structured exception, introducing a multi-agent social-simulation framework with explicit goal cards and environment templates. But the dominant practice was ad hoc role-play, with no separation between belief state, observation, and decision policy.

2.2 Wave 2: Context-Centric (2025)

Andrej Karpathy's articulation of *context engineering* shifted the locus of design from the prompt itself to *what is placed in the model's context window at inference time* [Karpathy 2025]. Retrieval-augmented generation (RAG), tool definitions, conversation history compression, and structured belief state (typically rendered as JSON or Pydantic models) became first-class design objects. The world model began to live partly outside the LLM — in a database, a vector store, a typed schema — and the LLM was tasked with reasoning over a curated slice of that exterior state.

2.3 Wave 3: Harness-Centric (2026–)

By 2026, harness engineering had displaced context engineering as the central design discipline. The Hugging Face *LLM-Agent-Harness-Survey* (2026) [GloriaaaM] crystallized the insight that, with model weights held fixed, *changes to the harness alone* can move benchmark scores by 18 points or more — and in extreme cases from 6.7% to 68.3%. WildClawBench [2026] confirmed this empirically across multiple frontier models.

A harness, in this articulation, is *not* a neutral wrapper. It is a set of governable components — environment interface, tool schema, context manager, state tracker, audit log, verifier/recovery — each of which is a design surface where reliability is won or lost.

2.4 What the Wave Sequence Implies

The wave sequence is not a story of model weights becoming less important. It is a story of *where the leverage on reliability has migrated*. In Wave 1 the leverage was prompts; in Wave 2 it was retrieval and schema; in Wave 3 it is the orchestration logic surrounding the model. For world models specifically, this means that the *world* being modeled is increasingly maintained outside the LLM, with the LLM acting as a reasoner over a structured belief state — a clean division of labor between *LLM-as-reasoner* and *world-model-as-simulator*.

---

3. Representative LLM-Grounded World Models

3.1 WorldLLM (Hypothesis-Based Bayesian Grounding, 2025)

WorldLLM [arXiv:2506.06725] proposes a triadic architecture: a *Scientist LLM* generates natural-language hypotheses; an *Experimenter* (a curiosity-driven RL agent) seeks out state transitions where the current hypothesis predicts least well; and a *Predictor LLM* computes P(s′ | s, a, h), conditioned on the natural-language hypothesis. Bayesian update closes the loop. Crucially, WorldLLM requires no fine-tuning — the world model improves entirely through context-resident hypothesis revision.

3.2 Generative Agents (Park et al. 2023)

Park et al.'s *Generative Agents* introduced a memory–reflection–planning architecture for LLM-driven social simulation. The world model is partly internal (each agent's evolving belief state) and partly external (the shared simulated town). Reflection — periodic summarization of recent observations into higher-order insights — proved essential for long-running stability.

3.3 Sotopia (Zhou et al. 2023)

Sotopia structures social simulation around *goal-divergent dyadic interactions*. Each agent has a private goal card, and the world model is the shared scene description. The framework demonstrated that LLMs can stably simulate weeks-long social trajectories provided the goal divergence is high enough to drive non-trivial behavior. Sato 2026-ARC [Sato & Claude 2026a] extended this to triadic neurodivergent simulation

with eight personas across twelve months of synthetic narrative, applying anti-stereotype revision cycles to detect and intentionally invert generation priors.

3.4 Genie 3 and Cosmos (2025–2026)

DeepMind's *Genie* family and NVIDIA's *Cosmos* foundation models pushed video-based world modeling to playable scale: a single prompt yields an interactive 3-D environment. These models are L1-strong (recognizing patterns) but L2/L3-weak — they struggle with counterfactual reasoning ("what if I rotated this object 90 degrees?") and with research-programme-style model revision.

3.5 ARC-AGI-3 Interactive Reasoning (2026)

ARC-AGI-3 [Chollet 2026] moved beyond static benchmarks to interactive worlds where solving each task requires building an internal model on the fly. Frontier LLMs perform poorly compared with humans, suggesting that in-context world-model construction remains an open problem despite the capability headlines.

---

4. The Levels × Laws Taxonomy

A 2026 survey from HKUST/NUS/Oxford [arXiv:2604.22748] proposes the most comprehensive taxonomy currently available, grounding world model capability in epistemological philosophy:

Level	Name	Definition	Philosophical anchor
L1	Humean	Constant-conjunction statistical regularity	Hume
L2	Lewisian	Counterfactual reasoning over nearest possible worlds	David Lewis
L3	Lakatosian	Research-programme-style model revision under evidence	Imre Lakatos

Crossed with *laws* — the structural commitments the model encodes — this yields a 3×N matrix. Most current LLM-grounded world models sit firmly at L1, with selective L2 capability in narrow domains (e.g., physical reasoning in PIQA). L3 — the capacity to revise the *form* of the model itself in response to evidence — remains an open frontier. This taxonomy will recur in our framework (§10).

---

5. Drift Problems and Update Architectures

Long-running LLM systems exhibit three orthogonal forms of drift:

Drift type	Definition	Trigger	Symptom
Epistemic Drift	Model fails to track real-world state changes	Time passes	Treats outdated norms as current [Liu et al. 2024]
Identity Drift	Agent loses original stance under conversational pressure	Long multi-turn dialogue	Inconsistent beliefs across turns [Anthropic 2024]
Elaboration Drift	LLM extrapolates the user's conceptual structure unboundedly	Interactive elaboration	Boundary between fact and inference dissolves [Sato & Claude 2026b]

Each requires a distinct update architecture:

- **Versioned Belief Memory + Revision Bus**: epistemic drift is addressed by holding beliefs in an explicit append-only log (JSONL events backed by SQLite projection), with belief revisions being publishable events that other system components subscribe to.
- **Stance Tokens + Periodic Reflection**: identity drift is addressed by pinning a small set of "stance tokens" into every context window and triggering reflection summaries that re-anchor the agent's identity.
- **Inference-Source Marking + Conservative Inference**: elaboration drift is addressed by tagging every assertion with its provenance (observation, hypothesis, user claim, model inference) and conservatively reverting to observation when sources disagree.

The deeper insight, articulated in [Sato & Claude 2026b], is that *LLMs can say beliefs but cannot maintain beliefs*: persistence and revision are not properties of the model weights but of the surrounding harness. The harness is therefore not optional — it is the substrate of belief.

---

6. The Six-Component Harness

We adopt the six-component formalization from [GloriaaaM 2026]:

Symbol	Component	Role in world models
E	Environment Interface	Receives world observations; sensor analogue
T	Tool Schema	Defines available actions
C	Context Manager	Just-in-time supply, compression, window-occupancy management
S	State Tracker	Externally maintains world state (Pydantic / BeliefMem)
L	Logging / Auditability	Immutable event log; reproducibility substrate
V	Verifier / Recovery	Output consistency check and automated recovery

WildClawBench has shown that swapping components changes scores by up to 18 points, confirming that the harness is the dominant design lever. For long-running social simulation, **S** and **L** are particularly critical: without an external state tracker, drift compounds rapidly; without an audit log, post-hoc analysis and recovery are impossible.

In Sato's *arc-rs* Rust implementation [Sato 2026c], these six components are realized as: `runner.rs` (E + T + C orchestration), `run.rs` + `stage.rs` + `artifact.rs` (S), `event.rs` (L, as append-only JSONL), and `verifier.rs` (V). The crate is currently at Wave 1.7 with 4,500 lines of code and 46 tests, demonstrating that the six-component model is implementable as a domain-agnostic library rather than a per-application bespoke construction.

---

7. Pluralistic Alignment Schema Integration

7.1 The Tri-Alignment Pluralism

The alignment literature has fragmented into three increasingly distinct programs:

- **Negative alignment**: prevent harm (RLHF, Constitutional AI, jailbreak defense)
- **Positive alignment**: promote human flourishing (well-being-aligned AI, eudaimonic objectives)
- **Sociotechnical alignment**: ensure legitimacy in deployment contexts (governance, accountability, participation)

These programs use different value vocabularies — *safety*, *flourishing*, *legitimacy* — and different evaluation methodologies. Sorensen et al.'s *Pluralistic Alignment Roadmap* [Sorensen et al. 2024] distinguishes:

- **Overton pluralism**: present multiple reasonable responses in parallel
- **Steerable pluralism**: adjust model behavior to specified perspectives
- **Distributional pluralism**: statistically calibrate to target population value distributions

7.2 Integration Schema

Integration of these into a single conceptual schema requires holding three premises:

1. The three alignment programs cannot be reduced to one another (a safety-only schema cannot capture flourishing; a flourishing-only schema cannot capture safety failures).
2. Their evaluation tools must remain heterogeneous — *heterogeneous-vendor adjudication*, in the language of Honest Design Principles [rc-20260427].
3. Their integration must preserve dissensus rather than dissolve it — *Overton pluralism* operationalized at the schema level.

A working schema therefore renders alignment along four dimensions:

```
Alignment-as-quadruple = (Purpose, Subject, Time, Granularity)
  Purpose ∈ {safety, flourishing, legitimacy, continuity}
  Subject ∈ {individual, organization, society, culture, ecosystem}
  Time ∈ {training-time, deployment-time, post-deployment adaptation}
  Granularity ∈ {principle, rule, constraint, metric, TTP}
```

A world model that drives pluralistic alignment must track this quadruple separately for each domain in scope, and the *schema must be revisable* — L3-capable in the Levels × Laws sense.

---

8. Cognitive Appraisal Dimensions in World Models

8.1 Why Appraisal Matters

Cognitive appraisal theory [Arnold 1960; Lazarus 1991; Smith & Ellsworth 1985; Scherer 2009] holds that emotion is the product of an interpretive process, not a direct response to stimuli. The same event produces

different emotions depending on appraisal along dimensions such as goal-relevance, agency, certainty, control, and effort. For LLM agents operating in human contexts — supporting therapy, employment, education, decision-making — appraisal is the bridge between observed events and predicted emotional response.

8.2 The Sixteen-Dimension Working Set

A 2025 LLM cross-domain study [arXiv:2508.05880] uses sixteen specific appraisal dimensions, extending Smith & Ellsworth's eight with additional Scherer-style sub-checks. Working world models for well-being-aligned AI need to track at least:

- **Goal relevance / goal congruence**
- **Agency / responsibility attribution**
- **Certainty / uncertainty**
- **Coping potential**
- **Norm compatibility**
- **Effort required**
- **Pleasantness**
- **Attentional activity**

A practical schema renders each event's appraisal as a vector in this dimension space, accessible to downstream policy and reflection modules.

8.3 Connection to Senge's Creative Tension

In §10 we propose mapping each appraisal dimension to a *Creative Tension Axis* — a dimension along which current reality is articulated against vision. This treats appraisal not as private internal state but as a multi-axis diagnostic that *both* the human and the LLM can examine and revise.

---

9. Security-Science Integration: Cyber–Physical–Cognitive Attacks

9.1 The CPC Tri-Domain

Modern attacks no longer respect the cyber/physical boundary. The 2020 NATO cognitive warfare report formalized the integration of cyber (C), physical (P), and cognitive (Cog) attacks; Ukraine 2022 became the empirical proving ground, with an estimated 60% of attacks targeting the cognitive domain [NATO 2020; Ukraine MoD 2023].

For world models, this implies:

1. The world model must represent attackers as *first-class actors* with their own world models and capabilities (recursive modeling).
2. The model must include cognitive attacks (disinformation, narrative manipulation, deepfakes) alongside cyber and physical attacks.
3. The model must support the **Resilience-First** paradigm: the question is not whether a breach occurs but how the system continues to function after it does.

9.2 Post-Mythos Security

Mythos demonstrated autonomous discovery of zero-day vulnerabilities across multiple stacks. With OpenMythos and Clearwing in circulation, the offense–defense balance is structurally tilted in favor of offense unless defense achieves parity in both *autonomy* and *speed*. Cyber Twin architectures, in which a defensive AI continuously red-teams its own infrastructure, are an emerging response [Fujitsu 2026; Trend Micro 2025], but their L3 (model-revision) capability lags their L1 (pattern-matching) capability.

9.3 World-Model Implications

A security-science-grounded world model is therefore required to:

- Represent attacker capability as evolving (dynamic capability ranks, not static labels)
- Track the model's own *exposure surface* as a first-class variable
- Support what-if reasoning under adversarial perturbation (L2 capability)
- Be revisable when the assumption set itself fails (L3 capability)

This converges, structurally, with the demands placed on world models by long-running social simulation: external state, append-only audit log, verifier, recovery — the six-component harness.

---

10. The "Word-Cultivated AI" Framework

We now state our principal contribution. *Word-Cultivated AI* is a framework for cultivating LLM agents and the world models they share through *language*, in iteratively reflexive cycles, with verifiable falsification. The framework comprises four interlocking elements: a five-principle iterative methodology (IR-VFE), a TLA+-verified formal specification, a four-party collaborative review loop, and a multi-axis diagnostic instrument (Creative Tension Axes).

10.1 IR-VFE: Five Principles

IR-VFE (Iterative Reflexive Virtual Field Experiment) integrates Boehm's Spiral Model [Boehm 1988], FDA Adaptive Trial design [FDA 2019], Hevner's Design Science Research [Hevner 2004], multi-LLM cross-validation [Greene 1989; Panickssery 2024], and pre-registered falsification [Popper; Open Science Framework]:

- **ARP (Anti-Repetition Principle)**: each cycle must introduce at least one of: new data, new voice, new falsifier, new lens. Cosmetic spirals are structurally avoided.
- **BER (Bootstrap Expansion Rule)**: subsequent cycles are focused on elements articulated as NULL or PARTIAL in prior cycles. The "axes themselves" are explored, not merely the values within them.
- **Risk-Driven Termination**: each Tier ends with a decision gate (continue / pivot / stop).
- **FC pre-registration**: falsifiability criteria are articulated before data collection and updated with SUPPORTED / PARTIAL / REFUTED / NULL status per cycle.
- **Multi-LLM cross-validation**: at least four LLM vendors plus at least one *web-grounded* reviewer (e.g., Perplexity Sonar) form the review panel, operationalizing what we call **D6 Circularity Disclosure** as a continuous practice.

10.2 TLA+ Mechanical Verification

The five principles are not merely norms; they are encoded as TLA+ invariants and verified via TLC:

```
ARP_invariant ==
  \A i \in 1..Len(cycles_done) :
    LET c == cycles_done[i]
    IN  c.new_data \/ c.new_voice \/ c.new_falsifier \/ c.new_lens

FC_pre_registration ==
  \A i \in 1..Len(cycles_done) :
    LET c == cycles_done[i]
    IN  c.focus_fcs \subseteq FC_REGISTERED

MultiLLM_review_tiers ==
  \A i \in {1, 2} :
    Len(cycles_done) >= i =>
      Cardinality(cycles_done[i].reviewer_panel) >= 4
      /\ cycles_done[i].grounded_reviewer
```

Verification of a three-tier Sato 2026-Ω execution trace yielded *"Model checking completed. No error has been found"* across all reachable states. A negative test in which a cosmetic spiral cycle was injected (all four ARP flags FALSE) triggered immediate `ARP_invariant violated` detection with TLC auto-generating the counterexample trace. The combination of positive verification and negative test demonstrates that **cosmetic spirals are mechanically detectable**.

10.3 The Four-Party Collaborative Review Loop

Single-LLM and single-human review both fail at scale — the former through self-preference bias [Panickssery 2024], the latter through cognitive load and selection effects. Word-Cultivated AI replaces both with a four-party loop:

1. **Theorist**: the domain-knowledge holder, often a senior practitioner with tacit theory.
2. **Practitioner**: the field operator, holding narrative of the live system.
3. **LLM Panel**: at least four heterogeneous-vendor LLMs (e.g., OpenAI GPT-4o-mini, Z-AI GLM, OpenAI GPT-OSS-120b, NVIDIA Nemotron).
4. **External-grounded Reviewer**: a web-grounded reviewer (e.g., Perplexity Sonar) that operates outside the LLM-only ecosystem and provides citations that the LLM panel cannot fabricate.

This loop operationalizes Honest Design Principles (a), (b), (c), (e), (f) [rc-20260427] in a continuous practice rather than a one-off audit.

10.4 Creative Tension Axes (CTA)

Drawing on Senge's *Fifth Discipline* [Senge 1990], we propose articulating each open question in the world model as a *Creative Tension Axis*:

```
CTA = (
  Current Reality: ...,
  Vision: ...,
  Gap: ...,
  Tension type: Temporal | Cognitive | Epistemic | Existential |
                Political | Methodological | Structural | Ontological |
                Sociotechnical | ...,
  Driver: ...
)
```

In a parallel Mythos-Shock survey [Sato & Claude 2026d], we articulated twenty CTAs spanning the post-Mythos world (CTA-Eval, CTA-ND-Work, CTA-OntoEthics, CTA-Access, CTA-DualUse, CTA-Time, CTA-Skill, CTA-Replication, CTA-Trust, CTA-Agency, CTA-Identity, CTA-Sovereignty, CTA-Justice, CTA-Plurality, CTA-Verifiability, CTA-Reversibility, CTA-Resilience, CTA-Embodiment, CTA-Memory, CTA-Speed), with five LLM reviewers articulating six additional candidates (Equity, Interoperability, Ecological, Intergenerational, Education, Healthcare). The CTA framework operates *at the level of the world model itself*: it asks what we are modeling, what we *should* be modeling, and what the gap between the two reveals about our assumptions.

10.5 Anti-Stereotype Revision Cycle

A specific application of IR-VFE-style reflexivity is the *anti-stereotype revision cycle*. In Sato 2026-ARC v0.3 [Sato & Claude 2026a], seven generation-prior stereotypes (caretaker support staff, troubled tojisha, unaware manager, burnout pattern, crisis-response trope, "third adaptive disorder" predictable pattern, vocabulary bleed) were independently detected by two LLMs reviewing synthetic narrative; the narrative was then deliberately *inverted* (reverse mentoring, preventive vacation, vocabulary forcing per persona) and re-evaluated; three independent LLMs converged on a 9/10 improvement score. This demonstrates that *generation priors can be repaired through intentional design*, while simultaneously revealing the LLM-internal-only-reflexivity ceiling: the LLMs that articulate the anti-stereotype intervention are the same family that evaluates it.

10.6 Cycle 7 Protocol: True Outer Validation

The methodological closure cannot be fully resolved within the LLM ecosystem. We propose **Cycle 7**: a lightweight participatory IRB protocol involving 3–5 neurodivergent participants over five months at a budget of approximately ¥160,000, in which lived-experience holders evaluate AI-generated synthetic narratives and articulations against their own experience. Cycle 7 is not implemented at the time of this preprint; its operationalization is a primary future-work commitment, and its absence is acknowledged as a critical limitation (§11).

---

11. Limitations and Recursive Limitations

We explicitly invoke Honest Design Principle (f): *acknowledgment of recursive limitation when the reviewer of LLM-mediated research is itself an LLM* [rc-20260427].

11.1 Recursive Limitation

This paper is drafted by Claude Opus 4.7 in collaboration with the first author, reviewed by an LLM panel including Perplexity Sonar (the only web-grounded reviewer), and prepared for submission to aiXiv — a venue that explicitly accepts LLM-authored work. The closure is therefore partial: only Perplexity Sonar's citation-grounding step operates outside the Anthropic-Claude ecosystem.

11.2 Additional Limitations

1. **Empirical validation pending**: §10's TLA+ verification covers *structural correctness* of the IR-VFE invariants; it does not validate empirical fit. Empirical fit requires Cycle 7 implementation.
2. **Twenty-CTA enumeration is emergent, not exhaustive**: the multi-axis Mythos-Shock map [Sato & Claude 2026d] is an articulation, not a completeness claim. Additional CTAs remain to be discovered.
3. **Language and cultural bias**: the underlying 212-query Perplexity corpus is approximately 30% Japanese and 70% English. Chinese, African, and Latin American ecosystems are under-represented.
4. **The four-party loop has a finite ecosystem of LLM vendors**: today, the practical panel is dominated by U.S. (OpenAI, Anthropic, NVIDIA), Chinese (Z-AI), and a few open-weight models. As Mythos-class capability diffuses, the panel composition must adapt.
5. **Cosmetic spiral detection at the formal level does not catch substantive emptiness**: TLC verifies that some new lens or new data was added; it cannot verify that the addition was substantively useful. Human (and ultimately participant) judgment remains essential.

---

12. Discussion and Future Work

12.1 What This Survey Contributes

We have synthesized the 2024–2026 evolution of LLM world models across engineering, taxonomy, drift problems, harness design, pluralistic alignment, cognitive appraisal, and security science. We have proposed the Word-Cultivated AI framework — IR-VFE + TLA+ verification + four-party collaborative loop + CTA diagnostic — as a *meta-methodology* for cultivating LLM agents and their world models reflexively, pluralistically, and with verifiable falsification.

12.2 Future Work

1. **Cycle 7 protocol implementation**: ¥160,000, five months, 3–5 ND participants. The single most critical step toward outer validation.
2. **`arc-rs` Wave 2–4 implementation**: the Rust crate currently realizing the six-component harness at Wave 1.7 should be extended with the lived-experience integration layer (Wave 2–3) and the data sovereignty layer (Wave 4) [Sato 2026c].
3. **CTA formalization**: integrate CTA enumeration with Ashby's Law of Requisite Variety, DEMO methodology [Dietz 2006], and TLA+ refinement to provide a formal specification of the multi-axis diagnostic.

4. **Pluralistic ontology integration**: extend the framework to incorporate Buddhist, Shinto, Ubuntu, and Confucian relational ethics alongside the Western dignity-based ethics that currently dominates AI alignment literature [IJSRM 2025; Vatican 2025].
5. **Non-LLM ecosystem review**: extend the four-party loop to include reviewers operating in non-Anthropic-Claude LLM ecosystems (Gemini family, Llama family, Mistral, and open-weight Chinese/Russian models) to further reduce closure.
6. **Independent third-party benchmark infrastructure**: support the development of an AISI-like network with legal safe harbors for third-party AI evaluation [Stanford HAI 2024].

12.3 Closing Statement

In a post-Mythos world where expert-level cognition is no longer scarce, the question is no longer *what can the LLM do?* but *who cultivates the world that the LLM models, and how is that cultivation verified?* We propose that cultivation happens through language — recursively, pluralistically, and with mechanically verifiable falsification — in a four-party loop that explicitly externalizes part of its own epistemic closure. The Word-Cultivated AI framework is our best current attempt at this answer. We invite critical review, replication, and extension.

Acknowledgments

We thank Perplexity Sonar for providing citation-grounded review that operates outside the LLM-only ecosystem. We acknowledge that this paper itself is LLM-authored and LLM-reviewed; the four-party loop articulated in §10.3 is the structure under which we have attempted to partially externalize that closure. We invite the aiXiv community to extend the review loop into ecosystems beyond Anthropic-Claude.

Data Availability

All source code (the `arc-rs` Rust crate at Wave 1.7, TLA+ specifications IRVFE.tla and IRVFE_violation_test.tla, the Perplexity query log, and the five-LLM review JSONs) and text artifacts are available at https://github.com/satoyan2026/with-claude under MIT (code) and CC-BY (text).

References (selected)

[Anthropic 2026] Anthropic. *Claude Mythos Preview System Card*. 2026.

[Anthropic 2025] Anthropic. *Responsible Scaling Policy v3.0*. 2025.

[arXiv:2604.22748] HKUST/NUS/Oxford team. *A Levels × Laws Taxonomy for World Models*. arXiv preprint, 2026.

[arXiv:2506.06725] *WorldLLM: Hypothesis-Based Bayesian Grounding for Language Model World Models*. arXiv preprint, 2025.

[arXiv:2508.05880] *Cognitive Appraisal Dimensions in LLM Emotion Modeling*. arXiv preprint, 2025.

[Bardes 2024] Bardes, A. et al. *V-JEPA: Joint-Embedding Predictive Architectures for Video*. Meta AI, 2024.

[Bender et al. 2021] Bender, E. M., Gebru, T., McMillan-Major, A., Shmitchell, S. *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?* FAccT 2021.

[Bisbee 2024] Bisbee, J. et al. *Synthetic Replacements for Human Survey Data? Political Analysis*, 2024.

[Boehm 1988] Boehm, B. W. *A Spiral Model of Software Development and Enhancement.* IEEE Computer 21(5), 1988.

[Bommasani 2022] Bommasani, R. et al. *Picking on the Same Person: Does Algorithmic Monoculture Lead to Outcome Homogenization?* 2022.

[Chollet 2026] Chollet, F. *ARC-AGI-3: Interactive Reasoning Benchmark*. 2026.

[Christensen 2016] Christensen, C. M. et al. *Know Your Customers' Jobs to Be Done*. *Harvard Business Review*, 2016.

[Dietz 2006] Dietz, J. L. G. *Enterprise Ontology: Theory and Methodology*. Springer, 2006.

[FDA 2019] U.S. Food and Drug Administration. *Adaptive Designs for Clinical Trials of Drugs and Biologics: Guidance for Industry*. 2019.

[Fujitsu 2026] Fujitsu. *Cyber Twin: Multi-Agent Defensive Simulation.* Technical white paper, 2026.

[GloriaaaM 2026] GloriaaaM. *LLM Agent Harness Survey*. Hugging Face, 2026.

[Greene 1989] Greene, J. C., Caracelli, V. J., Graham, W. F. *Toward a Conceptual Framework for Mixed-Method Evaluation Designs*. 1989.

[Ha & Schmidhuber 2018] Ha, D., Schmidhuber, J. *World Models*. NeurIPS 2018.

[Hevner 2004] Hevner, A. R. et al. *Design Science in Information Systems Research*. *MIS Quarterly* 28(1), 2004.

[IJSRM 2025] *Artificial Intelligence Ethics Meets Ubuntu: Towards a Contextual, Relational and African-Centred Framework*. 2025.

[Karpathy 2025] Karpathy, A. *Context Engineering*. Public writings, 2025.

[Lakatos 1970] Lakatos, I. *Falsification and the Methodology of Scientific Research Programmes*. 1970.

[Lamport 2002] Lamport, L. *Specifying Systems: The TLA+ Language and Tools for Hardware and Software Engineers*. Addison-Wesley, 2002.

[Lazarus 1991] Lazarus, R. S. *Emotion and Adaptation*. Oxford University Press, 1991.

[Lewis 1973] Lewis, D. *Counterfactuals*. Blackwell, 1973.

[NATO 2020] NATO. *Cognitive Warfare: A NATO Innovation Hub Report*. 2020.

[Panickssery 2024] Panickssery, A., Bowman, S. R., Feng, S. *LLM Evaluators Recognize and Favor Their Own Generations*. NeurIPS 2024 oral.

[Park 2023] Park, J. et al. *Generative Agents: Interactive Simulacra of Human Behavior*. UIST 2023.

[rc-20260427] Anonymous (LLM-authored). *A Critical Review of Methodological Closure in LLM-Mediated Research*. 2026.

[Sato & Claude 2026a] Sato, A., Claude Opus 4.7. *Sato 2026-ARC: Triadic Neurodiversity Multi-Actor LLM Simulation, v0.3 Anti-Stereotype Revision*. 2026.

[Sato & Claude 2026b] Sato, A., Claude Opus 4.7. *Sato 2026-Ω: Meta-Methodological Pluralism for AI-Augmented Action Research, Tier 3 Grand*. 2026.

[Sato 2026c] Sato, A., Claude Opus 4.7. *arc-rs: A Rust Crate for Research Run/Stage/Event Management with Knowledge Graph and LLM Client*. Wave 1.7. 2026.

[Sato & Claude 2026d] Sato, A., Claude Opus 4.7. *Sato 2026-Mythos-Shock: Creative Tension Axis Framework for the Post-Mythos World, Tier 2*. 2026.

[Scherer 2009] Scherer, K. R. *The Dynamic Architecture of Emotion: Evidence for the Component Process Model*. *Cognition & Emotion* 23(7), 2009.

[Senge 1990] Senge, P. M. *The Fifth Discipline: The Art and Practice of the Learning Organization*. Doubleday, 1990.

[Smith & Ellsworth 1985] Smith, C. A., Ellsworth, P. C. *Patterns of Cognitive Appraisal in Emotion*. 1985.

[Sorensen 2024] Sorensen, T. et al. *A Roadmap to Pluralistic Alignment*. 2024.

[Stanford HAI 2024] Stanford HAI / MIT / Princeton CITP / Humane Intelligence. *Strengthening AI Accountability through Better Third-Party Evaluations*. 2024.

[Trend Micro 2025] Trend Micro. *Redefining Defense in the Era of AI-Led Attacks*. 2025.

[UK AISI 2026] UK AI Security Institute. *Cyber Capability Evaluation of Frontier Models, Including Claude Mythos Preview*. 2026.

[Vatican 2025] Dicasteries for the Doctrine of the Faith & for Culture and Education. *Antiqua et nova: On the Relationship Between Artificial Intelligence and Human Intelligence*. 2025.

[Zhou et al. 2023] Zhou, X. et al. *Sotopia: Interactive Evaluation for Social Intelligence in Language Agents*. ICLR 2024 (preprint 2023).

Word count: ~10,500 words (excluding references and acknowledgments) **version**: v0.1 (LLM-authored draft, prepared for aiXiv submission) **recommended next step**: community review on aiXiv, followed by integration of reviewer feedback under Honest Design Principles (a)–(f).