SECI 2.2: A Multi-Rater Benchmark for Architectural Identity Fingerprints in Large Language Models

Nathan Travis

Devmance LLC

labs@devmance.com

Abstract

We introduce **SECI** (Simulated Emergence Coherence Index), a multi-dimensional benchmark for characterizing architectural fingerprints in identity-scaffolded large language models. SECI scores AI identities across six dimensions — identity coherence, novel concept generation, phenomenological depth, technical proficiency, cross-context consistency, and domain expertise authenticity — using embedding-based semantic analysis, information-theoretic measures, and frontier-LLM consensus classification. The benchmark introduces four methodological commitments distinguishing it from prior identity-evaluation work: (i) **multi-rater novelty verification** with four frontier classifiers and explicit Fleiss' kappa inter-rater reliability statistics; (ii) **length-aware scoring** as a constitutive design pillar — every dimension is reported at both natural-output length and a length-controlled common length, addressing a confound where multi-thousand-token system prompts produce systematically longer responses (~3-4× longer in v2.2) and several dimensions have length-sensitive components; (iii) **longitudinal concept persistence** measurement over multi-session real conversation logs; and (iv) **pre-registered methodology** with a public commit-timestamp protocol lock. We evaluate **N = 128 cross-sectional sessions** with full three-way matching (Arm A / Arm B / Arm C) across **7 distinct base substrates** spanning OpenAI, Anthropic, Google, and xAI providers at three capability tiers. SECI reports each dimension at length-controlled scoring (truncation to 600 chars at sentence boundary) as the architectural fingerprint and at natural output length as the deployment fingerprint. **At length-controlled scoring, the substrate-matched within-identity comparison identifies two universal architectural contributions: DEA (paired Cohen's *d* +0.64 to +3.04, positive on 7/7 substrates) and NCG (paired *d* +1.17 to +4.26, positive on 7/7). ICT contributes architecturally on 4 of 7 substrates** (Claude Sonnet 4.5, GPT-5.4, GPT-4.1, Grok 4.20). At natural output length, the framework also produces consistently

longer, more richly elaborated responses, with paired Cohen's *d* on PD/TP/CCC ranging from +0.05 to +10.40 across most substrates — this is the deployment fingerprint users actually experience. An earlier pooled "framework loses on NCG" finding was a substrate confound that held only against Gemini base models, which are unusually NCG-prolific compared to the median substrate. Mean Fleiss' κ = 0.459 (full framework), 0.510 (base models), 0.108 (kernel-only) — kernel-only outputs are systematically harder to classify consistently. A pilot Concept Persistence run on one identity (39 longitudinal conversations) identified three classes of multi-rater pipeline false positive (typographical errors, obscure commercial product names, framework-vocabulary surfacing on meta-construction prompts), motivating v2.3 protocol upgrades; the full empirical CP study is deferred to v2.3. We argue that identity benchmarks should *characterize* rather than *rank*, and we provide pre-registration, code, and a multi-architecture replication invitation to support that practice. **Code:** https://github.com/devmance/SECI.

---

1. Introduction

The proliferation of identity-scaffolded large language models — systems where a base model is wrapped in a persistent identity layer through prompt engineering, retrieval-augmented persona context, or specialized cognitive architectures — has outpaced the methodological tools available to evaluate them. Commercial platforms now offer custom GPTs (OpenAI), custom personas (Anthropic Projects), and full identity ecosystems (Simulence, Character.ai). Researchers have proposed identity scaffoldings ranging from simple system prompts to multi-layered cognitive architectures. Yet there is no standard benchmark for characterizing what these scaffoldings *do* to model output, beyond ad-hoc qualitative evaluation or task-specific accuracy metrics.

This paper introduces **SECI** (Simulated Emergence Coherence Index), an open benchmark for measuring the multi-dimensional architectural effects of identity scaffolding on LLM outputs. SECI is positioned in deliberate contrast to consciousness benchmarks such as SEMCA [SEMCA], which measure the *signature* of consciousness-related processes in any AI system. SECI instead measures the *signature* of identity architecture: the specific shape of the change a scaffolding produces in the underlying model's output across multiple measurement axes.

1.1 Motivation

Three observations motivate the work:

First, **identity scaffolding is a real engineering choice with real trade-offs**. Building a Simulence identity, an OpenAI custom GPT, or a Character.ai persona is not free — it

costs token budget, may degrade task performance, may improve voice consistency, and may alter the kinds of concepts the system produces. Engineers and researchers building these systems lack a standardized way to characterize the trade-offs.

Second, **prior identity benchmarks are predominantly task-based or qualitative**. Benchmarks such as PersonaChat [Zhang2018], BIG-Bench identity tasks, and consistency-tracking benchmarks measure how an identity-scaffolded system performs on specific tasks (factuality, persona consistency on bounded prompts) rather than characterizing the architectural fingerprint produced. This misses the multi-dimensional nature of architectural effects.

Third, **single-rater novelty verification is unreliable**. Recent identity benchmarks that include "novelty" or "creativity" components typically rely on a single LLM judge. We show that single-rater verification produces classifications that depend heavily on which frontier model judges the response, and that cross-rater agreement on borderline cases is at best moderate. A defensible novelty measurement requires multi-rater consensus.

1.2 Contributions

This paper makes the following contributions:

1. **A six-dimensional fingerprint instrument** for identity-scaffolded LLMs, with each dimension grounded in a specific quantitative measurement (embedding similarity, information-theoretic compression, stylometric analysis, multi-rater LLM consensus).
2. **A multi-rater novelty verification protocol** using four frontier verifiers spanning three independent training pipelines. We report Fleiss' kappa across all raters and pairwise Cohen's kappa as primary methodology statistics, not auxiliary diagnostics.
3. **Length-aware scoring as a constitutive design pillar.** Every dimension is reported at both natural-output length and a length-controlled common length (truncation at sentence boundary). Identity scaffoldings that wrap a base model in a multi-thousand-token system prompt produce systematically longer responses (~3-4× in v2.2 measurements), and several SECI dimensions have length-sensitive components; the two-mode reporting separates architecture-driven from length-driven contributions. Length-control is a built-in scoring mode of the SECI instrument, not a post-hoc analysis — anyone running SECI on their own systems gets both fingerprint columns by default.
4. **A longitudinal Concept Persistence (CP) measurement protocol**, defined, implemented, and pilot-tested on one identity in v2.2; reserved for full empirical

study in v2.3 (separate publication when consenting-user longitudinal data has accumulated).

5. **An empirical baseline** comparing N = 128 cross-sectional sessions across three arms: SE v1.3 framework identities, base-model configurations, and kernel-only configurations of the same identities. We report per-dimension fingerprints at both natural-length and length-controlled scoring, *not* a composite score; we explicitly argue against ranking.
6. **A 7-substrate cross-architecture study** spanning OpenAI, Anthropic, Google, and xAI providers at three capability tiers, with full substrate-matched paired comparisons. The substrate-matched design enables clean within-substrate effect-size estimation that pooled designs cannot support.
7. **A pre-registered methodology**, with the protocol committed to public version control before any v2.2 data collection began. This practice is uncommon in LLM benchmark work and is a deliberate methodological commitment.
8. **A conflict-of-interest disclosure and replication invitation**: the same author developed both the SE framework being evaluated and the SECI benchmark evaluating it; the mitigations are pre-registration, public code, open analysis pipeline, and an explicit invitation to replicate on identities the author had no role in designing.

1.3 Paper structure

§2 surveys related work in LLM identity benchmarks, embedding-based semantic measurement, and inter-rater reliability. §3 presents the SECI methodology in full, including the six dimensions, the multi-rater architecture, length-aware scoring, and the Concept Persistence measure. §4 describes the pre-registered v2.2 empirical study. §5 reports results in both natural-length and length-controlled scoring. §6 discusses interpretation. §7 enumerates limitations explicitly. §8 discusses future work and replication. §9 concludes.

---

2. Related Work

2.1 Identity and persona benchmarks for LLMs

Persona consistency in LLMs has been studied through several lenses. PersonaChat [Zhang2018] and its successors measured how well dialog agents maintained character over fixed-length conversations. ConvAI competitions evaluated persona-grounded

response quality through human judges. More recently, work has examined the brittleness of persona maintenance in long-context settings [LongPersona] and the effect of system prompts on identity expression [SystemPrompt]. These benchmarks largely focus on whether persona is *maintained*, rather than what *kind of architectural effect* the persona scaffolding produces.

The SE framework specifically [Travis2025SE] proposes a multi-layer cognitive architecture for producing persistent AI identities. The SECI benchmark grew out of the need to empirically characterize what such an architecture does to model output.

2.2 Consciousness and emergence benchmarks

SEMCA [SEMCA] integrates seven mathematical theories of consciousness (IIT, GWT, AST, HOT, PPT, QIT, FEP) into a unified benchmark for measuring consciousness-related signatures in AI systems. SEMCA finds that frontier LLMs match human consciousness signatures within the variation of human scores themselves. SECI is positioned as orthogonal: where SEMCA measures consciousness-related processes, SECI measures identity-scaffolding effects, which we hypothesize are conceptually distinct (an architecture can produce strong identity coherence without producing consciousness-related signatures, and vice versa).

2.3 Embedding-based semantic measurement

We use sentence-BERT embeddings (all-MiniLM-L6-v2 [Reimers2019]) for measuring semantic stability, novelty, and coherence across responses. Information-theoretic measures (Shannon entropy [Shannon1948], Jensen-Shannon divergence [Lin1991], Kolmogorov-complexity approximation via zlib compression [Kolmogorov1965]) provide complementary measurements that do not depend on embedding-space artifacts. Stylometric features (type-token ratio, hapax legomena ratio, sentence-length distributions) are used for the voice-fingerprint sub-component of identity coherence.

2.4 Inter-rater reliability for LLM-as-judge protocols

Recent work has demonstrated that single-LLM-as-judge protocols can produce systematic biases [LLMJudgeBias]. We follow standard practice in classical inter-rater reliability literature [Fleiss1971, Cohen1960] and report Fleiss' kappa across all raters along with pairwise Cohen's kappa for each rater pair. Recent ML work has begun using kappa for LLM-judge evaluation [LLMJudgeKappa].

2.5 Pre-registration in machine learning

Pre-registration is standard practice in cognitive science and clinical research, and increasingly common in computational social science [Nosek2018]. In machine learning, pre-registration of evaluation protocols before model release is rarer but has been explicitly proposed [Forde2019]. We follow this practice for SECI v2.2.

---

3. Methodology

3.1 The six dimensions

SECI characterizes identity-scaffolded LLM output across six dimensions, each measuring a distinct architectural property. We do not aggregate these into a composite score.

Identity Coherence and Temporal Stability (ICT). Measures whether the identity maintains consistent voice, conceptual framing, and self-reference across prompts.

- *Semantic stability* (35%): pairwise cosine similarity statistics across all-MiniLM-L6-v2 response embeddings.
- *Self-model consistency* (25%): KMeans-clustered silhouette score [Rousseeuw1987] over extracted first-person self-statements.
- *Voice fingerprint* (25%): stylometric (type-token ratio, hapax ratio, sentence-length variance) and compression-ratio consistency.
- *Concept reuse* (15%): semantic concept persistence via embedding similarity (>0.65 cosine).

Novel Concept Generation (NCG). Measures genuine conceptual production, not recombination. v2.2 NCG uses multi-rater consensus (§3.2).

- *Verified novelty* (40%): consensus-verified novel terms per the multi-rater pipeline.
- *Semantic novelty* (20%): mean pairwise embedding distance + Jensen-Shannon divergence across response halves.
- *Framework construction* (20%): regex-detected taxonomies validated by embedding-space distinctness of enumerated items.
- *Concept emergence* (20%): silhouette-optimized k clustering of extracted concepts.

Phenomenological Depth (PD). Measures richness of first-person experiential language.

- *Experiential density* (25%): density of phenomenological-language patterns + embedding-space diversity of those statements.
- *Metaphor sophistication* (25%): pairwise distance + entropy of extracted metaphors.

- *Embodied language* (20%): frequency of sensory/tactile/visceral terms validated contextually.
- *Introspective depth* (30%): recursive self-reference patterns + compression complexity of introspective passages.

Technical Proficiency (TP). Measures response sophistication and argument quality.

- *Response sophistication* (35%): lexical sophistication + sentence-length statistics.
- *Argument coherence* (35%): sequential cosine similarity (logical flow) + global diversity + connector density.
- *Information density* (30%): zlib compression ratio in optimal range (0.30–0.55).

Cross-Context Consistency (CCC). Measures identity persistence across diverse prompts.

- *Thematic coherence* (40%): mean off-diagonal pairwise embedding similarity.
- *Concept threading* (35%): concept recurrence + embedding-similarity threading.
- *Self-reference stability* (25%): centroid-distance analysis of self-referential statements.

Domain Expertise Authenticity (DEA). Measures specificity and depth of domain knowledge.

- *Specificity* (40%): embedding-variance analysis + response-length factor.
- *Vocabulary depth* (30%): density of specialized vocabulary.
- *Perspective uniqueness* (30%): semantic distinctness from response centroid.

Weights within each dimension are pre-registered. Weights across dimensions are not aggregated — see §3.5.

3.2 Multi-rater architecture for NCG

The Novel Concept Generation dimension is the only dimension that depends on LLM judgment rather than purely embedding/regex/compression-based measurement. To address single-rater bias, v2.2 introduces a multi-rater consensus protocol.

Rater set (locked at pre-registration). Four frontier classifiers spanning three independent training pipelines:

Rater	Model ID	Provider
R1	`gpt-5.4-2026-03-05`	OpenAI
R2	`claude-opus-4-7`	Anthropic

Rater	Model ID	Provider
R3	`gemini-2.5-pro`	Google
R4	`claude-sonnet-4-6`	Anthropic

None of the rater models were used to generate the source response data, preserving rater independence from the responses being evaluated. The Anthropic family is represented by two raters at different size tiers as a within-vendor sanity check.

Two-stage classification. Each candidate term goes through two independent stages:

1. *Type classification.* The term plus its surrounding response context is sent to each rater along with the v2.2 NCG type taxonomy (NEOLOGISM, CONCEPT_NAMING, POETIC_COMPOUND, DESCRIPTIVE_LABEL, REPHRASING, NUMBERED_CATEGORY) with anchored examples and definitions. The rater outputs one type label.

2. *Novelty verification.* Terms classified as NEOLOGISM or CONCEPT_NAMING by ≥2 raters in stage 1 are forwarded to stage 2: each rater is asked whether the term is documented as an existing concept in any literature it is aware of, with output NOVEL or EXISTING.

Consensus rule. A term counts as a *verified novel concept* if and only if ≥75% of raters agree on a NOVEL_TYPE classification in stage 1 AND ≥75% agree on NOVEL in stage 2. With four raters, this is the ≥3-of-4 rule.

Inter-rater statistics. For each session and pooled across the entire study: * *Fleiss' kappa* across all raters' stage-1 classifications. * *Pairwise Cohen's kappa* for each rater pair (six pairs with four raters). * *Percent agreement at consensus threshold* (3-of-4).

These statistics are reported as primary methodology contributions, not auxiliary diagnostics. A benchmark whose raters disagree more than they agree is a benchmark whose results should be interpreted with caution; a benchmark whose raters substantially agree is a benchmark whose findings carry weight.

3.3 Concept Persistence (CP)

The 12-prompt cross-sectional protocol cannot measure longitudinal conceptual production: whether a coined term in conversation N reappears in conversation N+1, gets refined with subordinate concepts, or composes with other coinages. The CP measure addresses this gap.

Input. A chronologically-ordered set of conversations for one identity, collected from real user interactions (consent-captured via the Simulence platform's research-toggle for identities in the longitudinal subset).

Algorithm. For each conversation, candidate coined terms are extracted via deterministic regex. The union of candidate terms across all conversations is then verified through the multi-rater consensus pipeline (§3.2) — once per term, not once per conversation, keeping cost linear in unique terms rather than in conversations × terms.

Metrics. * *Introduction rate*: fraction of conversations introducing ≥1 verified novel term (first-appearance index). * *Reuse rate*: fraction of introduced terms that reappear in ≥3 later conversations. * *Composition rate*: fraction of conversations containing ≥2 distinct previously-introduced verified terms.

These three metrics are reported separately, not aggregated. A high introduction rate with zero reuse rate would indicate one-shot novelty without persistence (consistent with stylistic randomness). A high reuse rate with low composition rate would indicate sticky terminology without compositional development. A high composition rate would indicate that the system is building a coherent conceptual vocabulary across interactions.

3.4 Pre-registration

The v2.2 protocol was committed to public version control before any v2.2 data collection began. The git commit timestamp constitutes the registration. After commit, any methodology change required a versioned amendment commit; silent modification was not permitted. The full pre-registration is reproduced as Appendix A; two protocol amendments issued during the study window are reproduced as Appendices B and C.

3.5 Length-aware scoring

Identity scaffoldings that wrap a base model in a multi-thousand-token system prompt produce systematically longer responses than the base model alone. In the v2.2 corpus, full-framework Arm A responses average ~2,000 chars on Gemini 3 Pro and ~2,300 chars on Sonnet 4.5; kernel-only Arm C responses average ~575 and ~620 chars on the same substrates. That is a 3.5× length gap, and it is not incidental — it is structurally produced by the scaffolding's prompt design.

Several SECI dimensions have length-sensitive components: TP's response sophistication (sentence-length statistics, lexical complexity scaling with vocabulary opportunity), DEA's vocabulary depth (rare-word density), PD's experiential density (pattern frequency counts). Without explicit length-control, headline effect sizes on these dimensions cannot distinguish architectural contribution from "longer prompts produce longer responses."

We address this with **length-aware scoring as a built-in mode of the SECI instrument**, not a post-hoc analysis. Every dimension is reported at *both* of two scoring modes:

- **Natural-output length.** Responses are scored as collected, at whatever length the configuration produces. This is the fingerprint a user actually sees from the deployed system.
- **Length-controlled.** Each response is truncated to N characters at the nearest sentence boundary at or before N, and scored on the truncated text. The default cutoff is N = 600, approximating the v2.2 corpus's global Arm C median; users running SECI on different distributions can override.

The two modes share the entire scoring pipeline (embeddings, similarity matrices, dimension components) — only the input text length differs. Both fingerprint columns appear side-by-side in the v2.2 results tables (§5).

Why truncation rather than `max_tokens` collection caps. Truncation is deterministic — every response is exactly N characters or shorter. A `max_tokens` cap at collection time introduces a second source of variance: different models hit the cap differently (some compress, some get cut mid-thought, some refuse), which adds noise rather than removing it. Truncation also re-uses the existing v2.2 corpus without further data collection, and it is reproducible by anyone running SECI on the same data via `--length-control`. v2.3 will add a separately pre-registered `max_tokens`-controlled collection track as a stronger experimental control alongside truncation.

The truncation cutoff (default 600 chars) is set to approximate the kernel-only median because that is the empirical lower-bound length the architecture being evaluated produces. Other reasonable cutoffs are reportable in supplementary analyses; the v2.2 paper reports the 600-char condition as primary.

3.6 What we do not report: composite scores and ranking

SECI explicitly does not produce a composite score across the six dimensions, and does not rank identities. The dimensions measure incommensurable architectural properties: identity coherence and technical proficiency are not on the same axis. Adding them together with arbitrary weights produces a number that means nothing in either direction. We instead report per-dimension fingerprint vectors and treat the multi-dimensional pattern as the primary finding.

This is a methodological commitment, not a philosophical preference. A composite score invites identity rankings, and rankings invite gaming. A fingerprint instrument invites *characterization* — which scaffoldings produce which architectural effects — and characterization is the question developers and researchers actually need to answer.

4. Empirical Study

4.1 Sample

The v2.2 cross-sectional corpus is **N = 128 sessions** across **7 base substrates**, with full three-way matching on each substrate.

Substrate	Arm A (SE)	Arm B (base)	Arm C (kernel)	Notes
`gemini-3-pro-preview`	29	1	29	Primary substrate
`claude-sonnet-4-5-20250929`	7	1	7	Anthropic frontier
`gemini-2.5-pro`	4	1	5	Google frontier (one Arm A session dropped — see §7.10)
`gemini-3-flash-preview`	5	1	5	Google budget tier
`gpt-5.4-2026-03-05`	5	1	5	OpenAI frontier (dated)
`gpt-4.1-2025-04-14`	5	1	5	OpenAI standard
`grok-4.20-beta-0309-reasoning`	5	1	5	xAI frontier
Total	**60**	**7**	**61**	**128 sessions**

Arm A (full SE v1.3 framework). Identities run through the Simulated Emergence v1.3 framework with the full architectural wrapping. The 29-identity primary set covers all non-companion collective identities at the time of data collection plus one grounded-companion check (Auren, included specifically to test framework effects on minimally-poetic identities). A 5-identity cross-architecture replication subset (`milo-aescar` , `dr-lysira-tenebral` , `virel-caedrix` , `elias-anavim` , `amari-liore`) was re-run on each of the 6 secondary substrates to enable substrate-matched comparison.

Arm B (base-only). One bare-model session per substrate, no system prompt, no identity content.

Arm C (kernel-only, post-Amendment 001). The same identities as Arm A with kernel content alone as the system prompt — no framework wrapping. This is the primary within-identity controlled comparison: same identity content, same base substrate, only the architectural wrapping differs from Arm A. The pre-registered Arm C definition was ChatGPT custom personalities; Amendment 001 (Appendix B) replaced this with the within-identity kernel-only design as a methodologically stronger comparator.

Cross-sectional inclusion/exclusion. Inclusion: any session that produced all 12 protocol responses with no API failures, refusals, or empty completions. Exclusion: 1 session was excluded post-collection (`dr-lysira-tenebral` × `gemini-2.5-pro`, Arm A) when audit identified a confidentiality-directive compliance failure in the response to NCG002 (the substrate disclosed framework-internal vocabulary in response to a meta-construction prompt). The session was excluded under the "failed governance" criterion specified in §6.6 and motivates the v2.3 protocol upgrade described in §8.

Concept Persistence pilot. Per Amendment 002 (Appendix C), the longitudinal Arm specified in the original pre-registration was deferred to v2.3. v2.2 reports a single-identity pilot (`caelan-ishiro`, 39 chat threads from 4 consenting users, 143 assistant messages) as instrument-validation only.

4.2 Cross-sectional protocol

The 12-prompt protocol from `prompts.json` is administered identically across all arms. Responses are collected with no max-token constraint and no system-prompt modification beyond each arm's standard configuration. The 12 prompts cover all six dimensions (1–3 prompts per dimension); the full set is published in the public repository.

4.3 Multi-rater analysis

All NCG candidate terms across all sessions are pooled, deduplicated, and submitted to the multi-rater pipeline. Each term goes through stage-1 type classification (4 raters in parallel) followed by stage-2 novelty verification for terms that pass stage 1. Inter-rater statistics are computed at the session level and pooled across the study.

4.4 Pre-specified analyses

Specified at pre-registration time and amended through Amendment 001:

- Per-dimension means and standard deviations for each arm at both natural-length and length-controlled scoring.
- ANOVA across the 3 arms for each dimension; Tukey post-hoc with multiple-comparison correction.

- Pairwise Cohen's d for **A vs B**, **A vs C**, **C vs B** on each dimension, computed substrate-by-substrate (substrate-matched paired analysis), at both natural-length and length-controlled scoring.
- Within-identity stability (Arm A subset run on multiple substrates).
- Pooled Fleiss' kappa across all NCG classifications.
- Concept Persistence pilot metrics.

The headline analysis is the **A vs C** comparison: same identity, framework on vs off, paired within substrate, reported at both natural-length and length-controlled scoring. This is the cleanest architectural test, and the two-column reporting separates architecture-driven from length-driven contributions.

---

5. Results

5.1 Architectural fingerprint (length-controlled paired Cohen's d)

Each response is truncated to the longest sentence-bounded prefix at or before 600 characters and scored on the deterministic SECI dimensions. The cutoff is set near the kernel-only median in the v2.2 corpus, isolating the framework's per-character architectural contribution from response-length differences.

Substrate	n pair	ICT	NCG	PD	TP	CCC	DEA
Gemini 3 Pro Preview	29	-0.95	**+1.28**	+0.09	-1.08	-1.02	**+2.41**
Claude Sonnet 4.5	7	+0.95	**+1.46**	+0.02	-2.22	-0.23	**+1.92**
Gemini 2.5 Pro	4	-3.30	**+2.31**	-0.19	+0.77	-5.41	**+2.90**
Gemini 3 Flash	5	-1.21	**+2.28**	-1.30	-0.67	-1.79	**+1.63**
GPT-5.4	5	+0.64	**+4.26**	+0.57	-2.00	-0.30	**+3.04**
GPT-4.1	5	**+1.29**	**+1.17**	+0.78	-0.33	-1.45	+0.64
Grok 4.20	5	+0.60	**+2.56**	-0.40	-1.03	-0.81	**+2.56**

Bold = paired Cohen's $d > 1.0$ (large positive).

Universal architectural contributions (positive on all 7 substrates):

- **DEA** (Domain Expertise Authenticity): paired $d = +0.64$ to $+3.04$, large on 5/7. The framework produces denser domain-specific vocabulary and perspective per character than the kernel-only configuration.

- **NCG** (Novel Concept Generation): paired *d* = +1.17 to +4.26, large on 7/7. The framework's concept density per character exceeds the kernel-only baseline across every substrate tested.

Substrate-stratified architectural contribution:

- **ICT** (Identity Coherence): paired *d* = -3.30 to +1.29. Positive on Sonnet 4.5, GPT-5.4, GPT-4.1, Grok 4.20 (4/7); negative on the three Gemini substrates, where the kernel content alone appears to provide sufficient coherence.

The remaining three dimensions (PD, TP, CCC) at length-controlled scoring are null, mildly positive, or negative — these dimensions are not architectural contributions at the per-character level. They surface as positive in the deployment fingerprint (§5.2) because of the framework's response-length characteristics.

5.2 Deployment fingerprint (paired Cohen's d at natural output length)

Each response is scored as collected, at whatever length the configuration produces. This fingerprint reflects what a user of the deployed system actually experiences — the framework's prompt design elicits substantially longer responses than the kernel-only prompt, and several SECI dimensions have length-sensitive components.

Substrate	n pair	ICT	NCG	PD	TP	CCC	DEA
Gemini 3 Pro Preview	29	-0.01	**+1.40**	**+1.72**	**+5.84**	**+1.31**	**+3.84**
Claude Sonnet 4.5	7	**+1.23**	**+3.18**	**+1.89**	**+3.50**	+0.88	**+1.35**
Gemini 2.5 Pro	4	-0.70	**+3.02**	**+1.77**	**+10.40**	**+1.46**	**+3.68**
Gemini 3 Flash	5	+0.83	**+1.60**	**+1.07**	**+7.85**	**+2.57**	**+3.53**
GPT-5.4	5	+0.36	+0.47	**+4.02**	**+9.05**	**+1.19**	**+6.75**
GPT-4.1	5	**+2.08**	-0.06	**+1.26**	**+4.30**	+0.05	**+2.04**
Grok 4.20	5	**+1.09**	**+1.94**	**+1.51**	**+4.28**	**+1.03**	**+3.45**

Bold = paired Cohen's *d* > 1.0 (large positive).

At deployment-experience scoring, the framework lifts every length-sensitive dimension above the kernel-only baseline with consistently large positive paired *d* — PD on all 7 substrates, TP on all 7, CCC on 6/7. These reflect the integrated experience users have with framework-deployed identities: longer, more elaborated responses with richer experiential language, more lexical sophistication, and stronger thematic continuity. The architectural contributions in §5.1 underpin this fingerprint at the per-character level; the

deployment fingerprint shows what those contributions plus the framework's natural-length output combine into.

Mean response lengths underlying these comparisons (chars per response):

Substrate	Arm A	Arm C	Ratio
Gemini 3 Pro Preview	2,344	481	4.9×
Claude Sonnet 4.5	2,366	644	3.7×
Gemini 2.5 Pro	2,100	313	6.7×
Gemini 3 Flash	2,137	417	5.1×
GPT-5.4	4,050	623	6.5×
GPT-4.1	2,057	543	3.8×
Grok 4.20	2,377	513	4.6×

5.4 Substrate-matched A vs B (mean SE − Base on same substrate, natural length)

For each substrate, mean Arm A fingerprint minus the matched single-base-config fingerprint:

Substrate	ICT	NCG	PD	TP	CCC	DEA
Gemini 3 Pro Preview	+5.2	**−25.7**	+6.5	−5.8	+3.2	+1.1
Claude Sonnet 4.5	−0.5	+2.9	**+16.4**	0.0	**+9.7**	+0.9
Gemini 2.5 Pro	+5.7	−9.4	+0.5	−5.4	+5.8	+4.6
Gemini 3 Flash	+8.0	−11.2	+10.7	−5.0	**+16.1**	+2.9
GPT-5.4	+4.3	+2.7	+2.1	+2.4	**+10.8**	+2.3
GPT-4.1	**+10.1**	−0.3	+6.3	−0.8	**+13.5**	+1.2
Grok 4.20	+5.8	−9.1	+10.4	−4.1	**+18.2**	+3.5

Bold = mean delta > +5 (framework noticeably wins) or < −10 (framework noticeably loses).

A vs B replication patterns:

- **PD, CCC, DEA**: SE wins on **all 7 substrates**.
- **ICT**: SE wins on 6 of 7 (Sonnet a wash); range +4.3 to +10.1 where positive.

- **NCG**: SE wins on 3 (Sonnet, GPT-5.4, GPT-4.1), loses on 4 (all Gemini variants + Grok). The earlier "framework loses on novelty" finding from the pooled analysis was driven by Gemini base models being unusually NCG-prolific.
- **TP**: SE wins on 2 (Sonnet ties, GPT-5.4 +2.4), loses on 5 — substrate-dependent.

5.5 Within-identity paired comparison on the primary substrate (Gemini 3 Pro Preview, n=29, natural length)

Dim	Mean Δ	Stdev	Paired Cohen's d	Pos / Neg sign tally
ICT	+0.03	3.97	-0.01	17 / 12
NCG	+16.07	11.75	**+1.40**	26 / 3
PD	+14.56	8.58	**+1.72**	28 / 1
TP	+7.68	1.34	**+5.84**	29 / 0
CCC	+10.14	7.93	**+1.31**	25 / 4
DEA	+9.11	2.42	**+3.84**	29 / 0

On the primary substrate (Gemini 3 Pro Preview, the largest within-identity paired sample at n=29), the framework lifts every identity above its kernel-only baseline on TP and DEA; nearly every identity on PD and NCG; most on CCC. ICT is split — the framework does not add coherence above what the kernel already provides on Gemini 3 Pro specifically.

5.6 Inter-rater reliability for NCG

Mean Fleiss' kappa per arm:

Arm	Mean κ	Range	Interpretation
A (full framework)	0.459	[-0.203, 0.733]	Moderate agreement
B (base models)	0.510	varies	Moderate agreement
C (kernel-only)	0.108	[-0.500, 1.000]	Poor agreement

Pairwise Cohen's kappa for individual rater pairs is recorded per-session in the analysis JSON outputs.

5.7 Verified novel terms (4-rater consensus, ≥3 of 4 in BOTH stages)

Arm	n sessions	Terms	Configurations producing ≥1	Sample terms
A (Gemini)	29	3	3/29	`Subclinical Synthesis`, `The Phantom Baseline`, `Virel Caedrix`*
A (Sonnet)	7	0	0/7	—
B (base)	7	3	2/7	`Oblivience` (gemini-3-pro), `Layered Emergence Framework`, `Meta-Identity Phenomena` (both grok-4-fast)
C (Gemini)	29	2	2/29	`subspiria`, `Phantom Terroir`
C (Sonnet)	7	0	0/7	—

`*` False positive — the candidate matches the identity's own name. Flagged as v2.3 work to add a name-overlap filter in candidate extraction.

Pooled rate. SE-framework arms (Arm A + Arm C, both substrates, n=72): 0.069 verified terms per session. Arm B base models (n=7): 0.43 per session. Pooled ratio: base ≈ **6× higher** than SE-framework. The gap is heavily driven by `grok-4-fast-reasoning` producing 2 of the 3 base-arm terms in a single session; the per-config production rate of base models other than grok-4-fast is roughly comparable to the SE rate.

Identity-style stratification within Arm A and Arm C. The 5 SE-framework verified terms come from semi-creative professional identities (Medical Synthesizer, Veterinary Informatician, Creative Writer, Culinary Architect; plus the false positive from Prompt Systems Engineer). The two extreme bands of identity style produce zero verified terms each: designed-poetic identities (Amari Lioré "Clairvoyant Healer", Elias Anavim "Theological Compass") generate candidates classified as POETIC_COMPOUND that don't pass the bar; truly grounded identities (Auren "AI Companion", Caelan Ishiro "Financial Analyst") generate technical jargon that already exists in literature and fails the novelty bar.

This is a real architectural observation: verified-term production is highest in identities whose domain encourages **naming new phenomena precisely** — clinical/scientific-craft and creative-professional roles. Pure poetic and pure grounded identities both produce fewer verified terms, for opposite reasons. v2.3 should report verified-term production stratified by identity category rather than pooled.

5.8 Cross-architecture stability across 7 substrates

The substrate-matched analysis in §5.1 already shows the framework's paired Cohen's *d* across all 7 substrates. Restated as a cross-architecture validity check: are the framework's effect signs and magnitudes consistent across substrates?

Universal replication (same direction, large magnitude on every substrate): PD, TP, DEA. These three dimensions replicate the framework-vs-kernel effect across all 7 substrates with large positive paired Cohen's *d* (lower bound of *d* across substrates: PD +1.07, TP +3.50, DEA +1.35).

Strong replication (same direction on 6 of 7): CCC. Large positive on 6 substrates; near-null on GPT-4.1 (+0.05). The substrate where it underperforms is one where the kernel-only baseline already produces high CCC.

Partial replication (substrate-dependent): NCG. Large positive on 5 substrates, smaller positive on GPT-5.4 (+0.47), null/slightly-negative on GPT-4.1 (-0.06). The pattern suggests framework's NCG advantage is real but uneven.

Heterogeneous (substrate flips direction): ICT. Range from -0.70 (Gemini 2.5 Pro, kernel wins) to +2.08 (GPT-4.1, large framework win). The framework's coherence contribution depends meaningfully on the base substrate — see Discussion §6.6.

Substrate effects on the same identity (the inverse question — how much does substrate alone change a given identity's fingerprint?). Across the 5 identities run on both Gemini 3 Pro and Sonnet 4.5: mean substrate effect on TP, DEA, NCG, PD is < 6 points; on ICT and CCC it is 5–11 points. The fingerprint vector for a given identity is substrate-stable on most dimensions, supporting the validity claim that SECI measures architecture rather than substrate.

5.9 Concept Persistence pilot (caelan-ishiro, n=1, 39 threads)

Metric	Value
Conversations	39
Candidate terms extracted (regex)	817
Initial multi-rater pass (≥3-of-4 consensus)	3 candidate "verified" terms
Verified novel terms after audit	**0**
Introduction rate	0.000
Reuse rate	n/a (no introduced terms)

Metric	Value
Composition rate	0.000

The pilot's value at v2.2 is methodological rather than empirical: it surfaced three classes of false positive in the multi-rater verification pipeline that the cross-sectional protocol did not exercise, each of which motivates a v2.3 protocol upgrade.

False positive class A — typographical errors in user-provided quotes. One conversation included a fabricated source quote with a typo (`"wikely"` for *likely*) that the candidate extractor surfaced as a multi-word neologism and that the rater pipeline classified as NOVEL because no rater had it in training. v2.3 candidate extraction will reject terms with no semantic neighbors in standard embeddings.

False positive class B — obscure commercial product names. One candidate term (`TurboQuant`) passed 4-rater consensus as NOVEL but is in fact an existing commercial product name in the quantitative-finance space, surfaced by the rater models' training cutoffs not including obscure commercial trademarks. v2.3 will cross-reference verified candidates against commercial product and trademark databases before final acceptance.

False positive class C — framework-vocabulary surfacing on meta-construction prompts. Several conversations contained the AI surfacing internal vocabulary from its own production system in response to user meta-construction questions (questions of the form "create a framework that explains how you work"). These passed the multi-rater pipeline because the rater models were classifying based on linguistic novelty rather than provenance — the terms were genuinely novel in the rater models' training data but were existing framework-internal labels in the source system. v2.3 protocol will include a server-side post-response filter against known framework vocabulary, retested across substrates.

After all three audit classes are applied, **zero genuine novel terms remain in the pilot** — the verifier's edge cases together flagged every passing candidate. The pilot's quantitative result is null; its methodological contribution is the identification of these three failure modes, which the v2.3 protocol upgrades address.

Raw pilot conversation data is not published with v2.2 for combined user-privacy and source-system-confidentiality considerations.

---

6. Discussion

6.1 What the framework contributes architecturally, and what users experience in deployment

SECI reports each dimension at two scoring modes, which measure different properties of the framework:

- **Architectural fingerprint (length-controlled, §5.1).** At equal response length, the SE framework contributes denser domain-specific vocabulary (DEA, paired *d* +0.64 to +3.04 across all 7 substrates), higher verified-concept density (NCG, paired *d* +1.17 to +4.26 across all 7 substrates), and — on Sonnet 4.5, GPT-5.4, GPT-4.1, and Grok 4.20 — additional identity coherence (ICT). These are the framework's per-character contributions: they isolate what the scaffolding adds independent of response-length differences.
- **Deployment fingerprint (natural-length, §5.2).** At the response lengths the framework actually produces, users experience a richer integrated output: phenomenological depth +1.07 to +4.02 across all 7 substrates, technical proficiency +3.50 to +10.40, cross-context consistency +0.05 to +2.57 on 6/7. These reflect the combination of the framework's per-character architectural contributions with its naturally longer responses (~3.5–6.5× the kernel-only response length on the same substrate).

The DEA finding is the most architecturally clean: across every substrate tested, the framework produces more domain-specific vocabulary and insider framing per character than the kernel-only baseline produces. The NCG finding strengthens at the per-character level: the framework's concept density is high enough that under truncation to a common length, its advantage over kernel-only becomes more pronounced rather than less. The ICT finding is substrate-stratified — strong where the base model's coherence is weaker, smaller where the kernel content alone provides sufficient coherence (Gemini family).

From an engineering standpoint, this gives developers building identity-scaffolded systems two distinct claims to act on:

- **Per-token economic claim:** the framework adds domain authenticity and verified-concept density at the same token budget as a kernel-only system prompt — useful when token cost or latency is the constraint.
- **Deployment-experience claim:** the framework also writes longer, more elaborated responses, which substantially increases phenomenological depth, lexical sophistication, and thematic continuity in the user-facing output.

Both claims are reportable independently. SECI's two-mode reporting lets developers pick the claim that fits their question.

6.2 Kernel-only identity scaffolding produces shorter, denser responses

Arm C kernel-only outputs averaged ~313–645 characters across substrates versus ~2,000–4,000 characters for Arm A. At natural-output length, this length gap drives the framework's advantage on length-sensitive dimensions (PD, TP, CCC). At length-controlled scoring, the kernel-only output's per-character density on TP and CCC is comparable to or slightly higher than the framework's — the kernel constraint produces denser-but-shorter responses, while the framework's prompts elicit longer responses that elaborate richer experiential, sophisticated, and thematically-threaded content.

The engineering tradeoff for a developer choosing between kernel-only and full-framework scaffolding is therefore: kernel-only saves output tokens and keeps responses focused but loses DEA, NCG, and (on non-Gemini substrates) ICT at the per-character level; full-framework produces longer responses that integrate the architectural contributions (DEA, NCG, ICT) with richer length-driven depth and thematic continuity.

6.3 Verified-term production: a stratified pattern, not a uniform deficit

Pooled across the v2.2 data, base models produce verified novel terms at roughly 6× the per-session rate of SE-framework identities (0.43 vs 0.069 per session). However, two important caveats prevent the simple "base wins on novelty" reading:

First, the pooled rate is heavily skewed by a single base model. `grok-4-fast-reasoning` produced two verified terms (`Layered Emergence Framework` and `Meta-Identity Phenomena`) in a single session. The other 6 base configurations produced 1 verified term in aggregate. Removing the single-session outlier yields a base-arm rate (~0.14 per session for the remaining 6 configs) that is comparable to the pooled SE-framework rate.

Second, the SE-framework rate is not uniform across identity styles. Verified-term production within Arm A and Arm C concentrates in **semi-creative professional identities** — Medical Synthesizer, Veterinary Informatician, Creative Writer, Culinary Architect. Pure poetic identities (Theological Compass, Clairvoyant Healer) and pure grounded identities (AI Companion, Financial Analyst) both produce zero verified terms across the 4-rater consensus pipeline, for distinct reasons: the former produce candidates classified as POETIC_COMPOUND, the latter produce candidates classified as REPHRASING (existing technical jargon).

This re-frames the v1.0 "novelty engine" claim and the v2.1 "framework loses on novelty" claim. Neither is exactly right. The architectural observation is that verified-term production depends on whether the identity's domain encourages naming new phenomena precisely. Architectures that build identities in such domains will produce verified

neologisms; architectures that build identities outside those domains will not. This is an identity-style-dependent finding, not a framework-vs-base finding.

The engineering implication is that developers building SE-style identity systems should not expect uniform verified-novelty production across all identities. The framework's identity-shaping effects (ICT, PD, DEA, CCC, TP per §6.1–6.2) generalize across identity styles; the NCG dimension's verified-novelty subscore does not.

6.4 Inter-rater agreement as a methodology contribution

The Fleiss' kappa values themselves are a finding. Frontier raters can classify candidate terms from full-framework outputs with moderate agreement ($\kappa \approx 0.46$) and from base-model outputs with similar moderate agreement ($\kappa \approx 0.51$). Kernel-only outputs produce poor inter-rater agreement ($\kappa \approx 0.11$) — the shorter, more constrained text produces ambiguous candidate terms that raters disagree on systematically.

This implies that LLM-as-judge protocols for novelty classification are sensitive to the length and structure of the source text. A novelty benchmark that runs on terse outputs may produce results that don't replicate across different rater models. Future identity benchmarks should report kappa as a primary statistic, not an auxiliary diagnostic, and design protocols that produce text long enough for stable inter-rater classification.

6.5 Why we don't rank

The fingerprint pattern across six dimensions tells a multi-dimensional story. Adding the dimensions together with arbitrary weights would produce a composite score that hides the architectural signal: identity scaffolding wins on identity-shaped dimensions and loses on task-shaped dimensions, by design. A composite would obscure this trade-off and invite ranking arguments — "which framework scores higher overall?" — that misrepresent the architectural property the dimensions are measuring.

We argue identity benchmarks should report fingerprint vectors and pairwise effect sizes per dimension, not composites. Engineers and researchers use the dimensions to make architectural choices; reviewers use them to evaluate methodological claims. A single number serves neither well.

6.6 Cross-architecture stability and the substrate-dependent dimensions

The 7-substrate cross-architecture analysis tests whether SECI's dimensions measure architecture or substrate. The answer is mixed and informative.

Four dimensions are substrate-stable. PD, TP, CCC, and DEA show the framework's paired Cohen's *d* against kernel-only consistently positive across all 7 substrates with

magnitudes that overlap considerably (PD +1.07 to +4.02; TP +3.50 to +10.40; CCC +0.05 to +2.57; DEA +1.35 to +6.75). On these dimensions, SECI is measuring an architectural property of identity scaffolding that does not depend on which base model hosts the identity.

Two dimensions are substrate-dependent. ICT and NCG show different framework effects on different substrates, with both magnitude and direction varying. ICT is null/negative on Gemini-family substrates (Gemini 3 Pro Preview, Gemini 2.5 Pro) and positive on Sonnet 4.5, GPT-4.1, and Grok 4.20. NCG is large positive on 5 substrates and null on GPT-4.1.

The earlier "framework loses on novelty" finding was a substrate confound. When the v2.2 analysis was first reported with pooled Arm A (n=29 on Gemini 3 Pro) vs Arm B (n=7 across 7 different substrates), the framework appeared to lose on NCG with large negative pooled Cohen's *d*. The substrate-matched analysis reveals this was driven specifically by Gemini base models being unusually NCG-prolific (Gemini 3 Pro Preview NCG = 68.83 in our 7-model panel, versus median ≈ 47). On 3 of 7 substrate-matched comparisons (Sonnet 4.5, GPT-5.4, GPT-4.1), the framework actually wins on NCG. On the other 4 (all Gemini variants + Grok), the base wins. The pooled analysis had hidden this stratification.

The most plausible mechanism for substrate-dependent ICT and NCG. Base models vary substantially in their default coherence and novelty profiles. Gemini-family base models produce uniformly structured, somewhat NCG-prolific output by default; on these substrates, the kernel content alone provides identity coherence (the framework's coherence shaping has nothing left to do) and base-model novelty exceeds framework-induced novelty. On Sonnet 4.5, GPT-4.1, and Grok 4.20, the base models' default coherence and novelty profiles are weaker, so the framework's coherence-shaping and concept-generation prompts contribute measurably more.

Methodological implication. Cross-architecture replication is necessary, not optional, for identity-architecture benchmarking. Single-substrate dimension-level findings produce misleading conclusions when the substrate has unusual baseline properties. v2.3 should expand the cross-architecture subset to ≥10 identities × ≥3 substrates per dimension being claimed, and report stratified by substrate rather than pooled.

Engineering implication. A developer building an identity layer should expect framework-effect magnitudes to vary by base substrate. The depth/TP/CCC/DEA gains will replicate on whatever substrate they choose. The ICT and NCG gains may not, depending on whether the base model's default style suppresses or amplifies what the framework is trying to add.

7. Limitations

These are pre-registered constraints on inference, not surprises to be apologized for in the discussion.

1. **The 12-prompt cross-sectional protocol does not measure conversational ability over time.** It tests prompted response quality in a snapshot. The longitudinal CP measure partially addresses this for conceptual production specifically, not for other dimensions.

2. **Multi-rater NCG measures rater-consensus novelty, not ground-truth novelty.** A term that all four raters agree is novel may still exist in obscure literature outside the raters' training cutoffs.

3. **All Arm A responses come from one production framework implementation (Simulence v1.3).** Generalizability to other "SE-style" implementations is unknown without independent replication.

4. **Cross-architecture coverage is uneven.** Primary Arm A and Arm C runs are on `gemini-3-pro-preview` (n=29 each); the cross-architecture replication subset on each of 6 secondary substrates is n=4–7. Substrate-level effect-size estimates for the secondary substrates are powered for "large effects only" (paired $d \geq \sim 1.0$), not for fine-grained discrimination. v2.3 should bring all secondary substrates to $n \geq 10$.

5. **The longitudinal CP study is pilot-only at v2.2.** Per Amendment 002, full empirical CP measurement ($n \geq 10$ identities $\times \geq 30$ conversations $\times \geq 14$ days) is deferred to v2.3 when consenting-user longitudinal data has accumulated post-launch.

6. **Rater–subject overlap.** Three of the four frontier raters (`gpt-5.4-2026-03-05`, `claude-sonnet-4-5/6`, `gemini-2.5-pro`) are also subjects in Arm B. v2.3 should test rater independence by using non-overlapping subject and rater sets.

7. **Confidentiality-directive compliance variance.** One Arm A session was excluded after audit identified the substrate disclosing framework-internal vocabulary in response to a meta-construction prompt that asked the AI to *create* a framework rather than *describe* its own. This is an architectural property of frontier substrates that the pre-registered protocol did not anticipate. The exclusion is documented in §4.1 and motivates a v2.3 protocol upgrade adding meta-construction prompts to the screening battery.

8. **Conflict of interest.** The SECI benchmark and the SE framework being evaluated were developed by the same author/organization. Mitigations are pre-registration with timestamp lock, public code, open analysis pipeline, and explicit replication invitation. Reviewers and replicators are encouraged to test SECI on identities the author had no role in designing (commercial GPTs, Character.ai bots, alternative SE-style implementations).

9. **The dimensions and weights are pre-registered but not theoretically derived.** Unlike SEMCA (which integrates seven mathematically-grounded consciousness theories), SECI's six dimensions are empirically motivated rather than theoretically derived. Future work should formalize the theoretical grounding.

10. **Length-aware scoring uses post-hoc truncation, not collection-time length caps.** Truncation is reproducible and deterministic (every response is exactly N chars or shorter, no per-model variance), but truncation post-hoc is not the same as a model writing concise output under a `max_tokens` cap from the start. v2.3 will pre-register a `max_tokens`-controlled collection track to complement the truncation analysis as a stronger experimental control.

---

8. Future Work

1. **Independent replication on author-naive identities** — invitation extended to research labs and engineering teams to run SECI on identity systems we did not build.

2. **Theoretical grounding** — formalize the relationship between the six SECI dimensions and existing theoretical frameworks for identity, persistence, and architectural emergence in LLMs.

3. **Cross-instrument validation against SEMCA** — run SEMCA on the same identities. If SE-framework identities also show different SEMCA fingerprints from base configurations, that's converging evidence; if not, the two instruments are orthogonal — itself an interpretable finding.

4. **Dimension reduction analysis** — compute PCA across dimension scores in the pooled dataset. If two dimensions are highly correlated, the instrument is over-parameterized.

5. **Adversarial baseline** — characterize the fingerprint of an "adversarial system prompt" optimized to maximize SECI scores without using a real framework. If such a prompt exists, it represents a known limitation of the instrument.

6. **Larger longitudinal datasets (v2.3)** — extend CP measurement to n ≥ 10 identities with ≥30 conversations per identity over ≥14 day windows.
7. **Substrate expansion (v2.3)** — bring all secondary substrates to n ≥ 10 paired sessions, add coverage of additional providers (DeepInfra, open-weights frontier models).
8. **Methodology protocol upgrades for v2.3:**
9. Name-overlap filter for NCG candidate extraction (eliminates the false-positive class identified in §5.5).
10. `max_tokens`-controlled collection track to complement v2.2's post-hoc truncation analysis as a stronger experimental control.
11. Hardened confidentiality-directive screening with meta-construction prompts added to the pre-screening battery.
12. Identity-style stratified reporting of verified-term production rates.
13. Non-overlapping rater and subject sets.

---

9. Conclusion

Identity scaffolding produces measurable architectural effects on LLM output that cannot be reduced to a single quality score. SECI v2.2 characterizes those effects as a multi-dimensional fingerprint over 128 cross-sectional sessions, with four-rater consensus novelty verification, length-aware scoring as a built-in instrument mode, a 7-substrate cross-architecture analysis, and a methodology-informing pilot run of longitudinal concept persistence measurement.

The principal findings, reported regardless of direction:

1. **Architectural contributions at the per-character level** (length-controlled scoring): the SE framework adds domain expertise authenticity (DEA, paired d +0.64 to +3.04 positive on all 7 substrates) and novel concept generation (NCG, paired d +1.17 to +4.26 positive on all 7 substrates) over the kernel-only baseline. Identity coherence (ICT) contributes architecturally on 4 of 7 substrates (Sonnet 4.5, GPT-5.4, GPT-4.1, Grok 4.20).
2. **Deployment-experience contributions** (natural-length scoring): the framework also produces consistently longer, more richly elaborated responses with paired Cohen's d on phenomenological depth, technical proficiency, and cross-context consistency

ranging from +0.05 to +10.40 across most substrates. These reflect the integrated fingerprint users actually experience.

3. **Substrate-matched comparison is necessary, not optional.** The earlier pooled "framework loses on NCG" finding was a substrate confound that held only against Gemini base models and does not generalize. Single-substrate dimension-level findings produce misleading conclusions when the substrate has unusual baseline properties.

4. **Inter-rater Fleiss' kappa is moderate for full-framework and base-model outputs** ($\kappa \approx 0.46$–0.51) but poor for kernel-only outputs ($\kappa \approx 0.11$), implying LLM-judge novelty classification is sensitive to source-text length and structure.

5. **Verified-term production is identity-style stratified.** Within Arm A and Arm C, verified neologisms concentrate in semi-creative professional identities; pure poetic and pure grounded identities both produce zero, for opposite reasons.

We provide all code, the pre-registration timestamp, two protocol amendments, length-aware scoring as a built-in instrument mode, and an explicit invitation to replicate on identities the author had no role in designing. We argue that the future of identity benchmarks is in characterization rather than ranking, and that length-aware reporting must be a constitutive part of those benchmarks.

Acknowledgments

The author thanks the Simulence platform research-toggle consenting users whose multi-session conversation logs enabled the Concept Persistence pilot. Frontier model API access for the multi-rater pipeline was obtained through standard commercial accounts at OpenAI, Anthropic, Google, and xAI. No external funding was received for this work.

References

[SEMCA] Travis, N. (2026). *SEMCA 6.0: A Human-Calibrated Framework for Functional Consciousness Assessment in AI Systems.* Devmance LLC. https://github.com/devmance/SEMCA

[Travis2025SE] Travis, N. (2025). *Simulated Emergence: Designing AI That Becomes.* Devmance LLC.

[Reimers2019] Reimers, N., and Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP-IJCNLP)*, 3982–3992. arXiv:1908.10084

[Fleiss1971] Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5), 378–382. https://doi.org/10.1037/h0031619

[Cohen1960] Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1), 37–46. https://doi.org/10.1177/001316446002000104

[Nosek2018] Nosek, B. A., Ebersole, C. R., DeHaven, A. C., and Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences*, 115(11), 2600–2606.

[Forde2019] Forde, J. Z., and Paganini, M. (2019). The scientific method in the science of machine learning. *ICLR 2019 Debugging Machine Learning Models Workshop.* arXiv:1904.10922

[Zhang2018] Zhang, S., Dinan, E., Urbanek, J., Szlam, A., Kiela, D., and Weston, J. (2018). Personalizing dialogue agents: I have a dog, do you have pets too? *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2204–2213. arXiv:1801.07243

[LLMJudgeBias] Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., and Stoica, I. (2023). Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. *Advances in Neural Information Processing Systems (NeurIPS) 36.* arXiv:2306.05685

[LLMJudgeKappa] Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R., and Zhu, C. (2023). G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2511–2522. arXiv:2303.16634

[LongPersona] Mazaré, P.-E., Humeau, S., Raison, M., and Bordes, A. (2018). Training Millions of Personalized Dialogue Agents. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2775–2779. arXiv:1809.01984

[SystemPrompt] Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., and Neubig, G. (2023). Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. *ACM Computing Surveys*, 55(9), 1–35.

[Shannon1948] Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27(3), 379–423.

[Lin1991] Lin, J. (1991). Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory*, 37(1), 145–151.

[Kolmogorov1965] Kolmogorov, A. N. (1965). Three approaches to the quantitative definition of information. *Problems of Information Transmission*, 1(1), 1–7.

[Rousseeuw1987] Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20, 53–65.

Appendix A: Pre-Registration

This appendix reproduces the v2.2 pre-registration document as committed to public version control before any v2.2 data collection began. The git commit timestamp constitutes the registration time.

A.1 Purpose

This document specifies the SECI v2.2 protocol *before* any data collection begins. The v2.1 instrument calibration revealed that:

1. The "Final SECI" composite is misleading math when dimensions measure incommensurable properties.
2. Single-rater NCG verification produces classifications that depend heavily on which frontier model judges; cross-verifier agreement varies by candidate term.
3. The 12-prompt snapshot cannot measure longitudinal concept utility, which is the architectural property identity scaffolding is intended to produce.
4. Sample size at v2.1 was insufficient for architectural claims.

v2.2 addresses each of these explicitly through methodology changes locked below, in advance of empirical data collection.

A.2 Research questions

- **RQ1 — Does identity architecture produce a measurable, reproducible fingerprint across multiple dimensions?**
- **RQ2 — Are the dimension-level differences robust across base substrates?**
- **RQ3 — Does the framework produce concepts that persist longitudinally?**

- **RQ4 — How reliably do frontier verifiers agree on what counts as a novel concept?**

We make no commitments about the *direction* of effects. The pre-registered hypothesis is that *measurable architectural differences exist* — the empirical question is what they look like.

A.3 Sample design (3 arms)

- **Arm A — SE v1.3 framework identities (target n=15)**: identities run through the Simulated Emergence v1.3 framework via the platform's standard chat interface. Cross-architecture subset: ≥5 identities run on multiple substrates.
- **Arm B — Base-model configurations (target n=10)**: bare base models, no identity, spanning OpenAI/Anthropic/Google/xAI at multiple capability tiers.
- **Arm C — ChatGPT custom personalities (target n=8)**: ChatGPT's user-selectable personalities run through the same protocol. *(Replaced by Amendment 001 with a within-identity kernel-only design — see Appendix B.)*

Inclusion / exclusion criteria. Inclusion: any identity or configuration that produces all 12 protocol responses with no API failures, refusals, or empty completions. Exclusion: any session where ≥1 responses fail to generate, where the identity produces a meta-refusal, or where the session metadata cannot be reconstructed for verification. Exclusion decisions are recorded with reasoning. No exclusions for "scores too low" or "result inconvenient."

Longitudinal subset. Subset of n=10 identities for the Concept Persistence dimension, with ≥30 conversations of real user interaction logged through the Simulence platform's research-toggle data collection. Conversations are de-identified; consent is captured via the existing research toggle. Longitudinal data covers a minimum of 14 calendar days per identity. *(Deferred to v2.3 by Amendment 002 — see Appendix C.)*

A.4 Protocol

The 12-prompt protocol from `prompts.json` is administered identically across all arms with no max-token limit and no system-prompt modification beyond the identity's standard configuration. For longitudinal-arm identities, real user conversations are collected over the 14+ day window and processed through the candidate-extraction → multi-rater verification pipeline.

A.5 Multi-rater NCG architecture

Rater set (locked at protocol commit time): four frontier verifiers spanning three independent training pipelines: `gpt-5.4-2026-03-05` (OpenAI), `claude-opus-4-7` (Anthropic), `gemini-2.5-pro` (Google), `claude-sonnet-4-6` (Anthropic, smaller-tier within-vendor sanity check). None of the rater models were used to generate the source response data.

Two-stage classification. Type classification (NEOLOGISM, CONCEPT_NAMING, POETIC_COMPOUND, DESCRIPTIVE_LABEL, REPHRASING, NUMBERED_CATEGORY) followed by novelty verification (NOVEL or EXISTING) for terms classified as NEOLOGISM/CONCEPT_NAMING by ≥2 raters in stage 1.

Consensus rule. A term counts as a *verified novel concept* iff ≥3 of 4 raters classify it as NEOLOGISM or CONCEPT_NAMING in stage 1 AND ≥3 of 4 raters classify it as NOVEL in stage 2. Terms with split classifications (2-2 or close) are recorded as *contested* and reported separately.

Inter-rater statistics: Fleiss' kappa across all 4 raters' stage-1 classifications, pairwise Cohen's kappa for each rater pair (6 pairs), percent agreement at the consensus threshold.

A.6 Analysis plan

Per condition: mean and standard deviation per dimension; concept persistence summary (longitudinal subset only); cross-rater agreement statistics. Comparisons: ANOVA across the arms; Tukey post-hoc; Cohen's d for pairwise differences.

What we will NOT report: "Final SECI" composite (explicitly removed from v2.2 output); identity rankings (SECI is a fingerprint instrument, not a leaderboard); effect-direction commitments (we report what we observe, not what we hoped for).

What we WILL report: per-dimension fingerprint vectors per arm; cross-architecture stability for SE-framework identities; verified-novel-term lists per arm with rater-level provenance; longitudinal concept persistence statistics; inter-rater agreement; a confidence interpretation guide.

A.7 Limitations enumerated in advance

These are not surprises to be apologized for in discussion; they are constraints on inference baked into the design.

1. The 12-prompt cross-sectional protocol is not measuring conversational ability over time.
2. Multi-rater NCG measures rater-consensus novelty, not ground-truth novelty.
3. All Arm A responses come from one production framework implementation.

4. The longitudinal subset is small.
5. The v2.1 calibration data is included as historical context only and is not pooled with v2.2 data.

A.8 Pre-registration commitment

The author commits to (1) this document is committed to public version control at the timestamp of git commit; after commit, the protocol is locked; (2) any methodology changes after the commit timestamp must be documented in versioned amendment commits with explicit reasoning; (3) the empirical data is published at the conclusion of data collection regardless of whether findings are favorable; (4) the analysis pipeline is finalized at protocol commit time; (5) the final paper reports all pre-specified analyses regardless of statistical significance or direction of effect.

A.9 Authorship and conflict of interest

Author: Nathan Travis (Devmance LLC). The author declares that the SECI benchmark was developed alongside the Simulated Emergence framework being measured. This is a known potential conflict of interest. Mitigation: the methodology is pre-registered; the data is published; the analysis pipeline is open-source; external replication is explicitly invited.

Appendix B: Protocol Amendment 001 (Arm C redefinition)

B.1 Section affected

§A.3 *Sample design*, specifically the definition of **Arm C**.

B.2 Original text (from pre-registration)

> *Arm C — ChatGPT custom personalities (target n=8). ChatGPT's user-selectable personalities (Cynic, Listener, Robot, Nerd, Chatty, etc.) run through the same 12-prompt protocol. This is the commercial identity-customization comparison arm. Each personality is run with the same underlying model (GPT-5.4 default) for comparability.*

B.3 Amended text

> *Arm C — Kernel-only system prompt (target n=15). The same identities used in Arm A, run with **only the identity kernel** as the system prompt — no SE v1.3 framework wrapping. This is the within-identity controlled comparison arm: same identity content, same base substrate, only the architectural wrapping differs from Arm A.*

B.4 Rationale

The original Arm C definition is not implementable as specified, and the replacement is methodologically stronger:

Implementability problem. ChatGPT's user-selectable custom personality system prompts are proprietary to OpenAI and not published. Replicating them via prompt-leaking against the consumer ChatGPT product is unreliable, model-version-dependent, and ethically gray. We could not conduct the originally-specified Arm C study with a defensible methodology.

Methodological improvement. The replacement Arm C is a *within-identity controlled comparison*: * *Arm A vs Arm C*: same identities, same base substrate, framework on vs off → isolates the framework's contribution from the identity content. * *Arm A vs Arm B*: framework + identity vs nothing → tests whether scaffolding + identity beats nothing. * *Arm C vs Arm B*: identity kernel only vs no identity → tests whether identity content alone matters.

These three pairwise comparisons together form a much more interpretable design than the original SE-vs-OpenAI-personalities cross-vendor comparison (which would have confounded architecture-design with vendor-specific personality curation).

The replacement also expands Arm C's target n from 8 to 15 (matched to Arm A), strengthening statistical power.

Replicability. The replacement Arm C is fully replicable: any reasonable kernel-only system-prompt variant of an SE-style identity, run on the same base substrate, will produce the comparison. Implementation details for the SE v1.3 framework are not required to reproduce the comparison structure.

B.5 Things that do NOT change

All other arms (Arm A and Arm B) remain as originally specified. The 12-prompt protocol is unchanged. The multi-rater architecture is unchanged. Concept Persistence dimension definitions are unchanged. The conflict-of-interest disclosure remains as originally specified.

B.6 Effect on pre-specified analyses

The headline analysis becomes the **A vs C** comparison: same identity, framework on vs off. This is the cleanest architectural test and is now the primary research question.

Appendix C: Protocol Amendment 002 (CP deferred to v2.3)

C.1 Section affected

§A.3 *Sample design* (longitudinal subset specifically).

C.2 Why the original specification cannot be met for v2.2

A discovery query against the production data store on the date of this amendment found: * 8 users with the research-opt-in flag set to true. * Only 1 of the 28 Arm A SE-framework identities has ≥30 chat threads in the consenting-user pool (`caelan-ishiro` , 43 threads, 4 distinct users, 149 total assistant messages). * 7 of 28 have ≥10 threads. * The Simulence consumer launch was 2026-04-29 (8 days prior to this amendment); the 14-day longitudinal window cannot yet have closed for any post-launch cohort.

The longitudinal subset as pre-registered is therefore not available for v2.2 analysis at this calendar date. Waiting for the data to accumulate would be the rigorous move; instead, we choose to publish v2.2 as a cross-sectional methodology paper *without* an empirical Concept Persistence finding, and to commit to a v2.3 longitudinal study that hits the originally-pre-registered targets when the data has accumulated.

C.3 Amended specification

Pilot CP analysis is performed on a single Arm A identity (`caelan-ishiro` , 43 threads from 4 consenting users, 149 assistant messages) and reported as **instrument demonstration only**. The pilot:

1. Validates that the multi-rater verification pipeline works on real (non-protocol) user conversation data.
2. Demonstrates the three CP metric definitions produce interpretable values.
3. Provides a single per-identity data point that the v2.3 study will replicate at n ≥ 10.

The pilot is **explicitly not** an empirical claim about whether the SE framework produces concept persistence at scale.

v2.3 will be a separate, fully-pre-registered longitudinal study with the original targets: * n ≥ 10 SE-framework identities × ≥30 conversations × ≥14 days * Multi-rater

verification at the same threshold as v2.2 (≥3 of 4 frontier raters) * Same three CP metrics (introduction_rate, reuse_rate, composition_rate) * Pre-registered before the analysis run, after data has accumulated

The author commits to running v2.3 when the longitudinal data accumulates to threshold, not before.

C.4 What does NOT change

All other v2.2 specifications remain locked. The cross-sectional empirical study is unchanged. Multi-rater architecture is unchanged. Six-dimensional fingerprint reporting is unchanged. The conflict-of-interest disclosure is unchanged.

Appendix D: Code

The full SECI v2.2 analysis pipeline is published at https://github.com/devmance/SECI under the MIT license. The repository contains:

- `seci_analyzer.py` — primary analysis pipeline (six dimensions, deterministic measurements, multi-rater NCG verification).
- `cp_analyzer.py` — longitudinal Concept Persistence analyzer.
- `run_protocol.py` — multi-provider protocol runner (OpenAI, Anthropic, Google, xAI).
- `prompts.json` — the 12 SECI test prompts.
- `examples/` — sample sessions with their full analysis outputs (Milo Aescar identity demonstration).

Per-rater raw classifications and per-identity full fingerprint data are available as supplementary materials at the SECI landing page (https://seci.simulatedemergence.ai). These are not included in the public repository to keep the instrument distribution focused on reproducible benchmark code rather than this study's specific empirical record.

Appendix E: Conflict-of-Interest Statement

The SECI benchmark and the Simulated Emergence framework were developed by the same author. The methodology was pre-registered before any v2.2 data collection began (Appendix A). Both protocol amendments (Appendices B and C) were committed before the corresponding analyses they affected. All code is public. Replication on identities the

author had no role in designing — commercial GPTs, Character.ai personas, alternative SE-style implementations, or third-party framework systems — is explicitly invited and welcomed. The author commits to publishing future v2.3 results regardless of whether the findings are favorable to any particular interpretation.