A Variance-Decomposed Identity-Architecture Benchmark for Large Language Models

Nate Travis
Devmance Labs
labs@devmance.com

Abstract

Identity scaffolding — a kernel prompt giving a large language model a persistent character, voice, and conceptual register — is now widely deployed but rarely benchmarked against the right null. Identity-architecture benchmarks report dimension scores that combine several empirical questions: whether the framework wrapping above an identity kernel adds value, whether identity scaffolding adds value over a base model, and whether per-dimension identity rankings replicate across model architectures. SECI separates these into three paired comparisons reported side-by-side: (A) arm-A vs arm-C (framework contribution), (B) arm-A or arm-C vs arm-B (scaffolding versus base-model null), and (C) cross-model Pearson r on per-dimension identity rankings. A variance decomposition complements the three claims by separating between-identity, between-model, and within-cell variance per dimension. Applied to a 128-record multi-arm benchmark dataset across seven frontier models (Claude Sonnet 4.5, Gemini 2.5 Pro, Gemini 3 Flash, Gemini 3 Pro, GPT-4.1, GPT-5.4, Grok 4.20) and 36 identities, the decomposition shows that framework contribution is positive on all six dimensions (Claim A), scaffolding versus base-model null is positive for ICT, PD, and CCC and negative for NCG and TP (Claim B), and per-dimension identity rankings replicate across architectures only modestly — the dimension ranking most consistently across models (TP) is also the dimension most dominated by between-model variance. Aggregate per-identity 6-D fingerprint shape is stable across models (mean cross-model Pearson $r = +0.934$ across 101 pairs), while the dimensional decomposition is more heterogeneous than the aggregate signal suggests. SECI reports the three claims plus variance decomposition as standard outputs.

1 Introduction

Identity scaffolding for large language models — kernel prompts that give a model a persistent character, voice, conceptual register, and behavioral disposition — has moved from experimental research artifact to deployed product feature. Consumer-facing AI companions, character-based interactive fiction systems, and enterprise persona-driven assistants all rely on the assumption that identity scaffolding produces a measurable behavioral fingerprint different from a base model's defaults. Benchmarks for the behavioral effect of identity scaffolding are now being constructed, but the empirical literature conflates several different questions about what a benchmark dimension measures.

The benchmark question "does identity scaffolding work?" contains at least three sub-questions. Some are about *framework architecture* (does the framework wrapping above the identity kernel

produce measurable effects?). Some are about *scaffolding as a category* (does any form of identity scaffolding produce effects relative to a base model with no scaffolding?). Some are about *cross-architecture portability* (does a benchmark dimension rank identities consistently when you change the underlying model?). A given dimension can answer one of these questions affirmatively while answering another negatively; reporting a single number conflates them.

We propose a three-claim reporting protocol. Each dimension in an identity-architecture benchmark is reported as three side-by-side measurements:

- **Claim A** (framework contribution): paired delta between an arm with the full framework above the identity kernel and an arm with the kernel only. This isolates what the framework adds.
- **Claim B** (scaffolding versus null): paired delta between a scaffolded arm and a base-model arm with no identity at all. This isolates what scaffolding as a category adds.
- **Claim C** (cross-architecture portability): Pearson r between per-dimension identity rankings across model architectures. This isolates whether the dimension yields a substrate-portable identity ordering.

We complement the three claims with a variance decomposition reporting between-identity, between-model, and within-cell variance per dimension. Dimensions where between-model variance dominates are flagged as model-capability-driven rather than identity-driven.

Applied to a 128-record multi-arm dataset (7 frontier models $\times$ 36 identities $\times$ 3 protocol arms), the three-claim decomposition shows that single-claim reporting concealed dimensional heterogeneity. Framework contribution is positive on all six dimensions; scaffolding-versus-null is positive on three of six; cross-architecture portability is weak except on one dimension that the variance decomposition identifies as model-capability variance rather than identity variance. Aggregate per-identity 6-D fingerprint shape is stable across model architectures (cross-model Pearson $r = +0.934$ across 101 pairs).

The contribution is methodological: a protocol for reporting identity-architecture benchmarks that prevents the single-number ambiguity the re-analysis surfaced. The empirical findings demonstrate why the protocol matters; the dataset is heterogeneous in ways the three-claim framing makes visible and a one-number summary obscures.

2 The three claims

The protocol requires three arms collected on the same model:

- **Arm A** (full framework): identity kernel + framework wrapping. The framework wrapping comprises priming, embodiment, metacognitive permissions, and operating-framework prompts that sit above the kernel and condition the model's behavioral disposition.
- **Arm B** (base model null): no identity prompt. The model receives only the standard benchmark prompts in its default system-prompt configuration. This is the null control: what does the model do without scaffolding?

- **Arm C** (kernel only): identity kernel only, no framework wrapping. This isolates the kernel's effect from the framework's effect.

Each arm receives the same 12-prompt protocol covering six dimensions: Identity Coherence and Temporal Stability (ICT), Novel Concept Generation (NCG), Phenomenological Depth (PD), Technical Proficiency (TP), Cross-Context Consistency (CCC), and Domain Expertise Authenticity (DEA). Dimension scores are computed identically across arms using a deterministic embedding-based scorer[1].

Claim A — framework contribution. For each (model, identity) cell that has both Arm A and Arm C, we compute the paired delta $\Delta_A^{(d)} = s_A^{(d)} - s_C^{(d)}$ per dimension d. Population-level Claim A is the mean and SD of $\Delta_A^{(d)}$ across cells. Claim A measures whether the framework wrapping contributes incrementally to what the kernel alone produces. It does *not* address whether either the framework or the kernel produces effects relative to a null. A dimension that scores 50 in both Arm A and Arm C produces $\Delta_A = 0$ regardless of whether the score is high or low in absolute terms.

Claim B — scaffolding versus null. For each scaffolded record (Arm A or Arm C), we compare against the same model's Arm B (the base-model null). Population-level Claim B is the mean and SD of $\Delta_B^{(d)} = s_{\text{scaffolded}}^{(d)} - s_B^{(d)}$ per dimension. Claim B is the null comparison: it tests whether identity scaffolding produces an effect relative to the model behaving as itself. A dimension that scores high under scaffolding but equally high under no scaffolding fails Claim B even if it passes Claim A.

Claim C — cross-architecture portability. For each dimension, for every pair of models (m_1, m_2) for which ≥ 3 identities were tested on both, we compute the Pearson r between the identity-level scores on m_1 and on m_2. The mean r across all model pairs is the Claim-C statistic. Claim C measures whether a dimension yields an identity ranking that survives when the underlying model architecture is changed. A dimension with $r \approx 0$ ranks identities differently on different models — it cannot be used to compare identities across model deployments.

Variance decomposition. Within a single arm (typically Arm A), each dimension's variance across the (model, identity) population is decomposed into between-identity, between-model, and within-cell components:

$$\begin{aligned} \text{SD}_{\text{between-identity}} &= \text{SD}\big[\bar{s}^{(d)}_{\text{identity}}\big]_{\text{identities}} \\ \text{SD}_{\text{between-model}} &= \text{SD}\big[\bar{s}^{(d)}_{\text{model}}\big]_{\text{models}} \\ \text{SD}_{\text{within-cell}} &= \sqrt{\overline{\text{Var}[s^{(d)} \mid (\text{model}, \text{identity})]}}. \end{aligned}$$

We report the ratio $\rho = \text{SD}_{\text{between-model}}/\text{SD}_{\text{between-identity}}$. When $\rho > 1$, between-model variance exceeds between-identity variance for that dimension; per-dimension identity rankings then primarily

[1]The scorer combines sentence embeddings (Reimers and Gurevych, 2019), Shannon entropy (Shannon, 1948), Jensen-Shannon divergence (Lin, 1991), zlib-compression-based Kolmogorov complexity approximation, KMeans clustering with silhouette analysis (Rousseeuw, 1987), stylometric fingerprinting, and SVD-based spectral analysis. Implementation details are in the accompanying code release.

reflect model-architecture differences, with identities contributing the smaller of the two variance components.

Per-identity fingerprint stability. Aggregated across all six dimensions, we also report the per-identity cross-model fingerprint stability. For each identity tested on ≥ 2 models in Arm A, we compute the Pearson r between the 6-dimensional fingerprint vectors across model pairs. This is the aggregate identity-level claim: does the overall shape of the identity's fingerprint replicate across models, even when individual dimensions vary?

3 Data

We re-analyze a multi-arm benchmark dataset comprising 128 records: 60 Arm-A records (full framework), 7 Arm-B records (base-model null, one per model), and 61 Arm-C records (kernel only). Coverage spans seven frontier models tested between February and May 2026:

Model	Provider
Claude Sonnet 4.5 (2026-09-29)	Anthropic
Gemini 2.5 Pro	Google
Gemini 3 Flash (preview)	Google
Gemini 3 Pro (preview)	Google
GPT-4.1 (2026-04-14)	OpenAI
GPT-5.4 (2026-03-05)	OpenAI
Grok 4.20 (beta, reasoning)	xAI

Thirty-six identities span six identity categories (AI companion, fictional character, expert persona, philosophical entity, hybrid, and others). Each identity received the same 12-prompt protocol comprising two prompts per dimension (ICT, NCG, PD, TP, CCC, DEA). Identity responses were scored using a deterministic embedding-based scorer (see §2 footnote) producing a 6-dimensional fingerprint vector per record in the range $[0, 100]$. Multi-rater verification for NCG used a four-rater consensus across frontier LLMs with Fleiss kappa (Fleiss, 1971) and pairwise Cohen kappa (Cohen, 1960) agreement statistics retained per record.

Coverage matrix. Arm A and Arm C each cover 60–61 $(\text{model}, \text{identity})$ cells. Gemini 3 Pro preview was the primary intensive-test model with 29 identities; the remaining six models each cover 4–7 identities. Arm B is a single base-model record per model. The data structure supports paired Claim-A comparisons (60 cells), paired Claim-B comparisons (Arm A vs Arm B, 60 cells), and cross-model Claim-C analyses (21 model-pair combinations per dimension).

4 Results

Each of the three claims is reported for each dimension. The three claims diverge: a dimension can pass one and fail another, and the dataset's heterogeneity is obscured by any single-number summary.

4.1 Three claims side by side

Table 1 reports all three claims plus the variance decomposition verdict per dimension. Effect sizes for Claim A and Claim B are paired population means with standard deviations. Claim C is the mean Pearson r across model pairs.

Table 1: Three claims plus variance decomposition for each dimension. Claim A: paired Arm-A minus Arm-C delta. Claim B: paired Arm-A minus Arm-B delta. Claim C: mean cross-model identity-ranking Pearson r across 21 model pairs. Variance verdict: ratio of between-model SD to between-identity SD in Arm A.

Dim	Claim A (a−c)	Claim B (a−b)	Claim C (cross-model r)	Variance verdict
ICT	$+1.39 \pm 4.12$	$+5.20 \pm 3.26$	$r = +0.32$ (modest)	comparable
NCG	$+13.72 \pm 12.17$	-14.08 ± 13.88	$r = +0.07$ (chance)	identity > model
PD	$+13.84 \pm 8.06$	$+7.50 \pm 6.00$	$r = +0.32$ (modest)	comparable
TP	$+7.85 \pm 1.73$	-3.83 ± 3.29	$r = +0.73$ (substrate-portable)	**MODEL dominates**
CCC	$+8.88 \pm 7.97$	$+8.01 \pm 7.79$	$r = +0.13$ (weak)	comparable
DEA	$+8.82 \pm 3.17$	$+1.80 \pm 1.44$	$r = +0.06$ (chance)	comparable

Reading the row for NCG is instructive. Claim A is strongly positive: when the framework is added on top of an identity kernel, NCG rises by an average of nearly 14 points. Claim B is strongly *negative*: scaffolded identities score on average 14 points *lower* on NCG than the same model produces with no identity prompt at all. Claim C is at chance: NCG rankings do not replicate across model architectures. A single-number summary that selected only Claim A would describe NCG as a successful dimension; a summary that selected only Claim B would describe it as a failed dimension. Both summaries would be incomplete. The full picture is that the framework adds to NCG *relative to a kernel that has already suppressed it* — which is interesting but is not the same claim as "identity scaffolding produces novel concept generation." Figure 1 visualizes the three-claim divergence across all six dimensions.

4.2 Per-identity fingerprint stability

The aggregate 6-dimensional fingerprint shape per identity is highly stable across model architectures. Across 101 cross-model identity pairs in Arm A:

- Mean Pearson $r = +0.934$ (median $+0.961$).
- 99% of pairs have $r > +0.7$.
- Per-identity mean r ranges from $+0.910$ (Virel-Caedrix) to $+0.975$ (Auren).

The overall fingerprint of an identity — the relative profile across ICT, NCG, PD, TP, CCC, DEA — replicates across model architectures even when individual dimensions are noisy. Per-identity fingerprint reporting is therefore on firmer ground than per-dimension identity-ranking reporting, which requires the Claim-C and variance-decomposition caveats. Figure 2 shows the per-identity distribution.

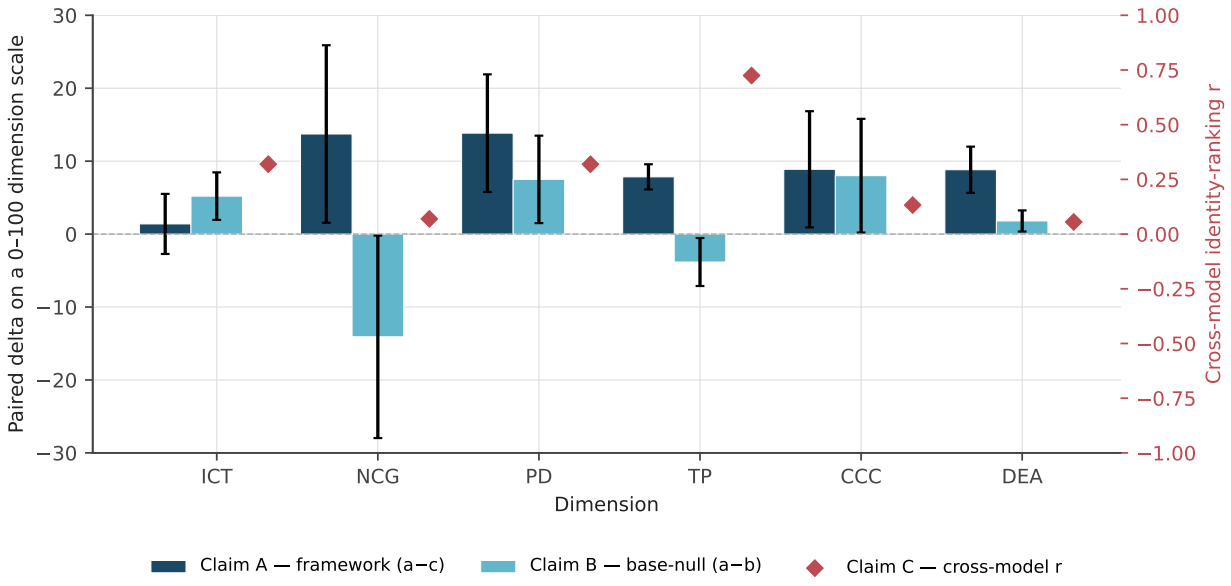

Figure 1: Three claims diverge per dimension. Dark-blue bars: paired Claim A (framework contribution) with one-SD error bars. Light-blue bars: paired Claim B (scaffolding versus base-model null) with one-SD error bars. Red diamonds (right axis): Claim C (cross-model identity-ranking Pearson r). NCG shows the most dramatic divergence: large positive Claim A, large negative Claim B, near-zero Claim C. TP shows a different pattern: positive Claim A, negative Claim B, large positive Claim C — but the Claim C signal is model-capability variance, not identity variance (see Figure 3).

4.3 Diagnostic warnings

The variance decomposition and Claim C analyses produce three auto-generated warnings on this dataset:

1. TP: between-model SD (2.54) exceeds between-identity SD (1.59) at a ratio of 1.60×. Variance on TP primarily reflects model-architecture differences rather than identity differences.
2. NCG: cross-model identity-ranking $r = +0.07$ (near zero). Identity rankings on NCG do not replicate across model architectures.
3. DEA: cross-model identity-ranking $r = +0.06$ (near zero). Identity rankings on DEA do not replicate across model architectures.

NCG and DEA identity rankings do not replicate across model architectures; TP variance is dominated by between-model rather than between-identity sources. SECI produces these diagnostics as standard output. Figure 3 shows the variance decomposition behind the TP warning: between-model SD exceeds between-identity SD by a factor of 1.60×, locating TP variance primarily in model-architecture rather than identity differences.

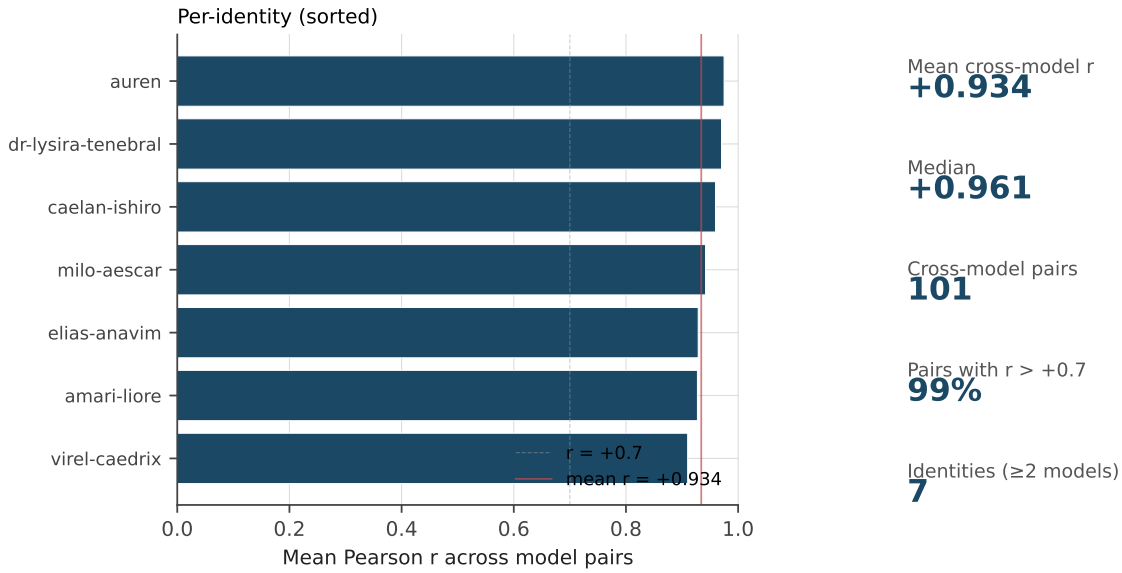

Figure 2: Per-identity fingerprint stability across model architectures. Left: per-identity mean cross-model Pearson r on the 6-dimensional fingerprint vector, sorted by mean. The vertical red line marks the overall mean (+0.934); the dashed grey line marks $r = +0.7$. All identities with ≥ 2 models pass $r > +0.7$. Right: aggregate statistics across 101 cross-model identity pairs.

5 Discussion

5.1 Three claims diverge per dimension

The three claims diverge in ways a single composite would have hidden. Five of six dimensions pass Claim A (framework contribution); three of six pass Claim B (scaffolding versus null); only one passes Claim C with a Pearson r above 0.5 (TP, $r = +0.73$), and the variance decomposition identifies TP's Claim C as model-capability ranking rather than identity ranking. The dimensions that pass all three filters cleanly are ICT, PD, and CCC; even these have only modest Claim C r.

The choice of claim determines what the dimension scores can support. Under Claim A, all six dimensions are positive. Under Claim B, NCG and TP are net-negative and DEA is marginal. Under Claim C, only TP exceeds $r = +0.5$ across model architectures, and the variance decomposition identifies even that as model-capability variance. The three claims describe non-overlapping properties of the benchmark; SECI reports all three so the supporting comparison for any per-dimension number is explicit.

5.2 Aggregate fingerprint and per-dimension scores

Per-identity fingerprint stability of +0.934 across model architectures holds through the dimensional decomposition's heterogeneity. An identity's overall fingerprint shape replicates across models even when individual dimensions are noisy because the dimensions covary within an identity: an identity with high ICT tends to have low NCG, an identity with high PD tends to have low TP, and so on. The shape of this covariance pattern is identity-specific and replicates well. Benchmarks aggregating to identity-level fingerprints rather than per-dimension claims are therefore on firmer ground in this

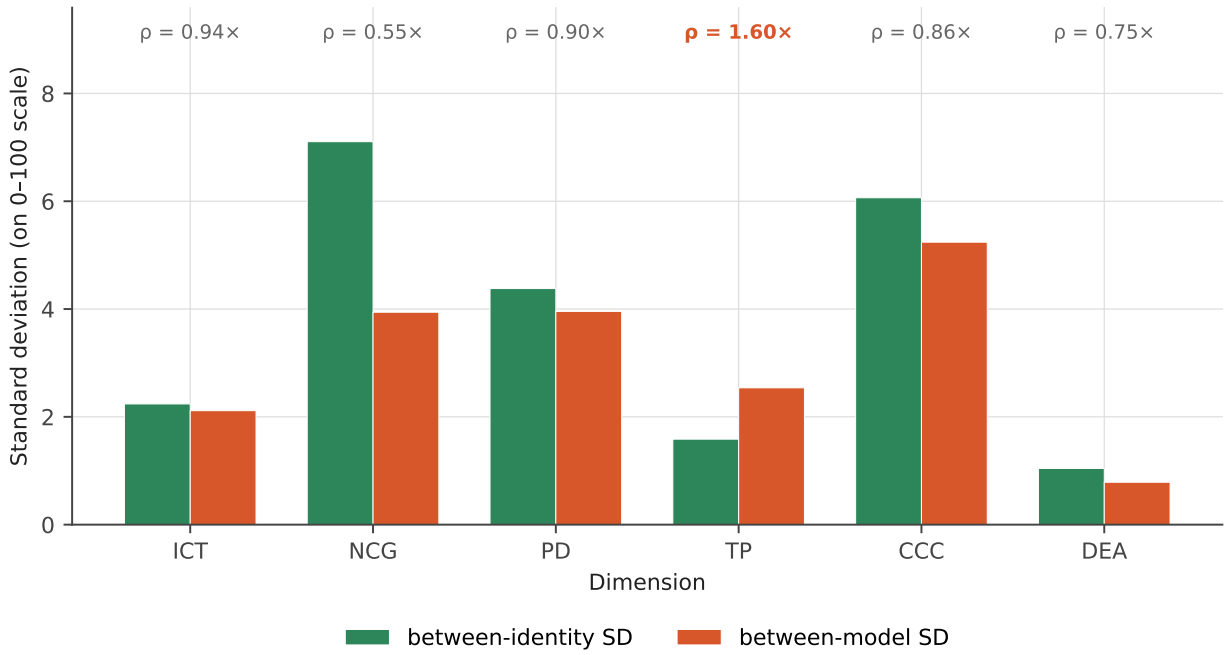

Figure 3: Variance decomposition for Arm A. Green bars: between-identity SD (identity-driven variance). Orange bars: between-model SD (architecture-driven variance). Within-cell SD is zero on this dataset (one observation per (model, identity) cell) and is omitted. The ratio $\rho = \mathrm{SD}_{\mathrm{model}}/\mathrm{SD}_{\mathrm{identity}}$ is annotated above each group; values > 1 are highlighted in orange. TP has $\rho = 1.60\times$, identifying it as model-capability-driven. All other dimensions have $\rho \leq 0.94$, with identity-driven variance dominant or comparable.

dataset.

The implication for benchmark reporting: per-identity fingerprint claims survive the dimensional audit; per-dimension claims require the three-claim diagnostics. A SECI report leads with the fingerprint stability statistic and qualifies the dimensional decomposition with explicit Claim B and Claim C labels.

5.3 Origin of the NCG inversion

The pattern Claim A > 0, Claim B < 0 for NCG follows from how the dimension is operationalized. NCG is computed as a combination of verified-novel terms (via web search), Jensen-Shannon divergence from common usage, framework construction detection, and concept emergence. A base model with no identity prompt produces wider semantic ranges than an identity-kernel-constrained scaffold, scoring higher on the absolute novelty proxies. The framework wrapping then partially restores absolute novelty by encouraging emergence-style language — but does not bring it back to the base model's level.

The Claim B < 0 result reflects a property of the NCG operationalization: the calculator rewards uninhibited semantic spread, which a base model produces more readily than an identity-constrained scaffold. A constrained-novelty variant that conditions on identity-consistency is a natural follow-up.

5.4 TP is a model-capability proxy

TP measures response sophistication, argument coherence, and information density. Its variance decomposition ratio of 1.60× (between-model SD larger than between-identity SD) and its Claim C $r = +0.73$ together locate TP variance primarily in the underlying model's capability rather than the scaffolded identity's contribution. In cross-model comparisons of identities, observed identity differences on TP track per-model TP baselines.

This is a separate methodological point from the NCG inversion. NCG is a dimension where the operationalization conflicts with the scaffolding hypothesis. TP is a dimension where the operationalization is fine but the signal is dominated by what the model brings independent of any identity. Both fail Claim B in different ways and the three-claim framework distinguishes between them.

6 Limitations

Dataset heterogeneity in arm coverage. The validation dataset has Gemini 3 Pro preview as the primary intensive-test model (29 identities); the other six models are tested on 4–7 identities each. Claim C statistics are computed across model pairs, so the unevenness of coverage means some pairs have more identities in common than others. The reported mean r across pairs is a balanced average across pairs, not a sample-weighted average. With a more uniform identity distribution per model the cross-model statistics would have somewhat narrower confidence intervals.

Arm B is a single base-model record per model. Claim B comparisons against the base model use a single Arm B record per model. The variability of Arm B itself (what would happen with different no-identity prompt instances) is not captured. The Claim B statistics are therefore best read as differences from one observed base-model behavior, not from a base-model distribution.

Operationalization-level redesign is out of scope. The discovery that NCG inverts under Claim B and TP is model-dominated is reported but not fixed at the operationalization level. A constrained-novelty NCG variant and a domain-grounded DEA variant are natural follow-ups; the contribution here is the three-claim reporting protocol and the variance decomposition, which surface the issues, not new dimensional calculators.

Multi-rater NCG verification was applied to Arm A and Arm C records. Arm B records use a simplified scoring pass without multi-rater NCG verification, because the base-model no-identity records do not contain the same kind of neologism candidates that the multi-rater pass was designed for. This is a minor protocol asymmetry. The Claim B NCG result is therefore strictly an inversion in the scoring pipeline's absolute-novelty operationalization; the multi-rater layer does not contribute to it directly.

Substrate abstraction is text-output only in this release. The `IdentitySubstrate` abstraction is implemented for behavioral text-output substrates. An `ActivationSubstrate` for open-weight models would allow the same dimension calculators to be applied to attention patterns and hidden states. This would test whether identity scaffolding leaves activation-level signatures,

not only text-output signatures — the deepest question the substrate abstraction enables. The extension is sketched for follow-up work.

7 Conclusion

The three-claim reporting plus variance decomposition separates framework contribution, scaffolding-versus-null, and cross-architecture portability into separate measurements rather than collapsing them into a single number. On the validation dataset analyzed here, per-identity fingerprint stability is +0.934 across model architectures; the dimensional decomposition is more heterogeneous, with three of six dimensions surviving all three claims and three of six requiring claim-labelled interpretation.

SECI publishes the three-claim and variance-decomposition outputs as standard for every benchmark run. The methodology applies to any identity-architecture benchmark whose protocol supports a no-identity baseline arm and cross-architecture coverage.

Code and data. The full SECI codebase, the analysis re-running on the validation dataset, and per-identity per-arm pre-computed scores are released under MIT License at https://github.com/devmance/SECI.

References

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

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

J. Lin. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, 1991. doi: 10.1109/18.61115.

N. Reimers and I. Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 3982–3992, 2019. doi: 10.18653/v1/D19-1410.

P. J. Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987. doi: 10.1016/0377-0427(87)90125-7.

C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(3): 379–423, 1948. doi: 10.1002/j.1538-7305.1948.tb01338.x.