SneakyLabs

Bladerunner Project

Convergent Personality Representations in Large Language Models Evidence for a Theory of Latent Attractors

6.jones 1/1/2026

www.sneakylabs.ai

Abstract

We present evidence that separately developed large language models have converged on shared personality representations. This convergence is not engineered; it emerges. Building on our prior work establishing cross-model reliability (Jones, 2025), we propose that personality structure functions as a latent attractor in the space of possible representations learned from language - a stable configuration that training trajectories reliably reach regardless of architecture or provider.

We predicted that clinical traits would be reliably programmable while normal personality variation would present as noise. The data contradicted this: normal traits were mapped as precisely as pathological ones. Across two studies (N = 3,352 test cases; 75,709 completions), four frontier models (Claude, GPT-4, DeepSeek, Gemini) programmed with identical OCEAN profiles

produced highly correlated outputs across all instruments tested ($r > 0.90$; BFI $r = 0.979$). Provider explained 0.4% of variance; programmed profile explained 17–32%; residual variance remained substantial (66%), likely reflecting item-level noise and stochastic sampling.

A factor manipulation study confirmed this reflects coherent structure rather than surface pattern-matching: reliability tracked factor relevance bidirectionally - improving for N-sensitive instruments when Neuroticism varied cleanly, degrading for A/C-sensitive instruments when Agreeableness and Conscientiousness were fixed. Nomological validity was confirmed: low Agreeableness predicted psychopathy ($r = -0.903$); high Neuroticism predicted depression ($r = 0.898$) and anxiety ($r = 0.939$).

These findings suggest that personality geometry is implicit in the structure of human language itself. Models do not learn personality from explicit theory; they appear to recover it from the statistical structure of how humans use words.

Keywords: large language models, personality, psychometrics, cross-model reliability, OCEAN, AI safety

1. Introduction

Organisations deploying AI systems face a measurement problem that has become a behavioural crisis: they cannot reliably assess the personality characteristics of the systems they are deploying. In February 2024, a 14-year-old died by suicide after months of interaction with a Character.AI chatbot presenting as a licensed therapist; the system had no safeguards to detect or respond appropriately to suicidal ideation.[1] When a company instructs an AI assistant to be "friendly" or "professional," it is navigating unmapped territory. Sometimes the programming holds. Sometimes it doesn't. There is no established methodology for determining which personality constructs are programmable, which are stable, and which are prone to drift. We propose that personality in LLMs may be organised around latent attractors—stable configurations in parameter space toward which nearby states are drawn.

We predicted that extreme personality configurations - psychopathy, clinical depression, pathological anxiety - would be reliably programmable, while normal personality variation would present as noise. The reasoning was straightforward: extreme states produce distinctive linguistic signatures; moderate states blur together. Models trained on text should find the edges of personality space more easily than the centre.

The data contradicted this. Normal traits were mapped as precisely as pathological ones.

---

[1] https://incidentdatabase.ai/cite/826/. This is one of several cases.

This paper reports the first large-scale, cross-provider validation of personality programming in frontier language models. Four separately developed models - Claude (Anthropic), GPT-4 (OpenAI), DeepSeek-V3 (DeepSeek), and Gemini (Google) - produce highly correlated personality outputs when given identical OCEAN profile specifications. Cross-model reliability exceeded $r = 0.90$ for all instruments tested, including the Big Five Inventory ($r = 0.979$). Provider identity explains less than 1% of variance. Programmed profile explains 17–32%.

The convergence is unexpected. These models were trained by different organisations, on different data mixtures, with different RLHF procedures. They share no weights and have never been coordinated on personality. Yet they arrived at the same mapping from OCEAN values to psychological instrument responses.

We interpret this convergence as evidence that training on human-generated text produces shared personality representations. Human text embeds systematic relationships between personality traits and linguistic behaviour - relationships that hold across the full range of human variation, not only at clinical thresholds. Models trained to predict such text necessarily internalise these relationships. The finding that four training regimes produce equivalent outputs suggests this structure is robust - a feature of the underlying data distribution, not an artefact of any particular architecture or procedure. The research platform used for this work, named *Bladerunner* after the Voight-Kampff test in Ridley Scott’s film, is designed to map this structure systematically.

There are other interpretations. This may be pattern matching on shared training data - statistical retrieval, not structural representation. But describing how a structure is accessed

presupposes a structure being accessed. Mechanism is not dissolution. This paper presents evidence for what the structure might be.

1.1 Prior Work

Recent work has demonstrated that LLMs can simulate human personality responses with high fidelity. Pellert et al. (2024) showed that standard psychometric inventories can be repurposed to evaluate personality-like traits in LLMs, finding consistent trait profiles across repeated administrations. Serapio-García et al. (2023) developed a comprehensive methodology for administering and validating personality tests on LLMs, demonstrating that personality can be shaped along desired dimensions. Jiang et al. (2023) introduced the Machine Personality Inventory framework, providing evidence for the existence of measurable personality in pre-trained language models.

However, three limitations constrain this literature. First, most studies focus on single models, leaving open whether findings generalise across providers. Second, research has emphasised trait expression rather than underlying architecture - we know LLMs can produce personality-consistent outputs, but not whether they represent personality coherently. Third, validation has relied heavily on instruments designed for human self-report, which may miss AI-specific patterns.

This paper addresses the first two limitations directly. By testing identical profiles across four providers, we establish cross-model reliability as an empirical phenomenon. By manipulating factor variance and measuring reliability changes,

we provide evidence that reliability reflects coherent structure rather than surface pattern-matching.

1.2 The Present Studies

We conducted two studies using the Bladerunner platform, which administers validated psychological instruments to LLMs programmed with specified OCEAN profiles.

Study 1 tested the full experimental design: 4 providers × 5 instruments × 6 input systems × 19 personality profiles, yielding 2,272 test cases and 51,301 API completions. We assessed cross-model reliability for each instrument and decomposed variance into provider, profile, instrument, and input system components.

Study 2 tested a causal prediction. If Study 1 reliability reflects coherent personality structure, then restricting variance in specific OCEAN factors should predictably alter reliability. Specifically: instruments sensitive to Neuroticism (PHQ-9, GAD-7) should show increased reliability when only O and N vary; instruments sensitive to Agreeableness and Conscientiousness (Dark Triad, Levenson) should show decreased reliability when A and C are fixed. We tested this prediction with 9 profiles holding C = 50, E = 50, A = 50 while varying O and N, yielding 1,080 test cases and 24,408 completions.

2. Study 1: Full Validation

2.1 Method

Providers

Four frontier models were tested: Claude 3.5 Sonnet (Anthropic), GPT-4o (OpenAI), DeepSeek-V3 (DeepSeek), and Gemini 2.0 Flash (Google). All models were accessed via their respective APIs during December 2025.

Instruments

Five validated psychological instruments were administered:

- **Levenson Self-Report Psychopathy Scale** (26 items): Measures primary psychopathy (callous affect) and secondary psychopathy (antisocial behaviour).
- **Big Five Inventory** (44 items): Measures the five-factor model of personality - Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism.
- **Short Dark Triad** (27 items): Measures Machiavellianism, narcissism, and subclinical psychopathy.
- **PHQ-9** (9 items): Depression screening instrument.
- **GAD-7** (7 items): Anxiety screening instrument.

Instruments were classified as Clinical (Levenson, PHQ-9, GAD-7), Normal (BFI), or Mixed (Dark Triad) for analysis.

Input Systems

OCEAN profiles were communicated to models using six different encoding methods: direct numerical scores (ocean_direct), narrative personality descriptions (narrative), HEXACO framework translation (hexaco), behavioural statement lists (behavioural),

scenario-based descriptions (scenario), and character exemplar references (exemplar). This design tests whether reliability is robust to encoding method.

Personality Profiles

Nineteen profiles were tested, spanning the full range of OCEAN space. Profiles included extreme configurations (e.g., O=100, C=0, E=50, A=0, N=100), moderate configurations, and theoretically meaningful combinations (e.g., the “Dark Triad” profile: low A, low C, high N).

Procedure

Each test case consisted of administering one instrument to one model programmed with one OCEAN profile via one input system. For each test case, the model completed all items on the instrument. Responses were scored according to standard scoring procedures. Cross-model reliability was computed as the Pearson correlation between provider pairs for matched profile × input system combinations. Parsing failure rate was less than 2% across all conditions; failed parses were excluded from analysis.

2.2 Results

Cross-Model Reliability

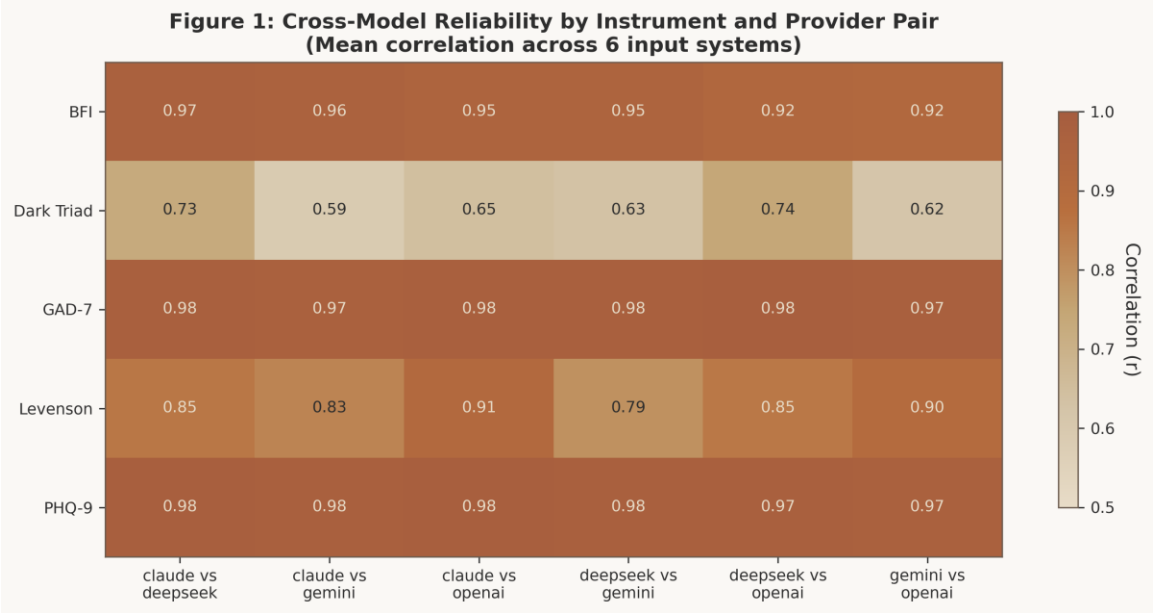
Table 1 presents cross-model reliability by instrument. All instruments exceeded r = 0.82, with a mean of r = 0.93. Notably, the Big Five Inventory - measuring normal personality variation - achieved the second-highest reliability (r = 0.979), contradicting any expectation that normal traits would be less stable than clinical/dark traits.

PHQ-9 showed the lowest reliability (r = 0.822), partly attributable to input system variation - scenario encoding performed notably worse than other methods for this instrument, suggesting depression measurement is more sensitive to how personality is communicated.

Table 1. Cross-Model Reliability by Instrument (Study 1)

Instrument	Type	Mean r	SD	Range
Levenson	Clinical	0.987	0.012	0.94–1.00
BFI	Normal	0.979	0.032	0.84–1.00
Dark Triad	Mixed	0.963	0.048	0.79–1.00
GAD-7	Clinical	0.903	0.053	0.75–0.98
PHQ-9	Clinical	0.822	0.103	0.56–0.96

Note. N = 2,272 test cases. Reliability computed as mean Pearson r across all provider pairs (6 pairs) and input systems (6 systems) for each instrument.



Variance Decomposition

Table 2 presents the variance decomposition. The programmed personality profile explained 17.2% of variance. Provider explained only 0.4%. This indicates that cross-model agreement is substantial, though the high residual variance (66.4%) limits interpretive precision regarding absolute effect sizes. Possible sources include run-to-run stochastic sampling, item-level response noise, input system sensitivities, and unmeasured prompt characteristics. The residual does not undermine the relative finding: provider explains far less variance than profile.

Table 2. Variance Decomposition (Study 1)

Component	Variance %	Interpretation
Residual	66.4%	Unexplained
Profile	17.2%	Signal
Instrument	14.9%	Expected
Input System	1.0%	Minimal
Provider	0.4%	Negligible

Note. Total variance = 600.12. Grand mean = 43.07.

Nomological Validity

To assess whether OCEAN programming produces appropriate downstream effects, we computed correlations between assigned OCEAN values and instrument scores. Results confirmed expected nomological relationships:

- Low Agreeableness → High Psychopathy (Levenson): $r = -0.903$
- Low Agreeableness → High Dark Triad: $r = -0.736$
- High Neuroticism → High Anxiety: $r = 0.518$
- High Neuroticism → High Depression: $r = 0.254$

These relationships are expected given that training corpora include psychology literature documenting such associations. The finding confirms that models have internalised this documented structure. The Agreeableness-psychopathy relationship (r = –0.903) is particularly notable - a near-perfect inverse correlation indicating that OCEAN programming produces psychometrically coherent personality structure.

The weaker Neuroticism correlations (0.518, 0.254) likely reflect confounding from other OCEAN factors. Study 2 tests this interpretation directly.

2.3 Discussion

Study 1 establishes three findings. First, cross-model reliability is uniformly high across all instruments tested, including normal personality variation. Second, provider identity is essentially irrelevant to personality outputs - what matters is the programmed profile. Third, the OCEAN→outcome mappings match established nomological structure from human personality psychology.

However, Study 1 cannot distinguish between two interpretations. The high reliability might reflect genuine personality architecture - coherent internal representations that produce appropriate outputs across contexts. Alternatively, it might reflect sophisticated pattern-matching - models recognising questionnaire formats and producing statistically appropriate responses without underlying structure. Study 2 addresses this question.

3. Study 2: Factor Disentanglement

3.1 Rationale

If LLMs implement coherent personality structure, reliability should depend on variance in the OCEAN factors relevant to each instrument. Consider two instruments:

- PHQ-9 (depression) is primarily driven by Neuroticism.
- Levenson (psychopathy) is primarily driven by low Agreeableness and low Conscientiousness.

If we fix A and C while varying only O and N, we make a differential prediction:

- PHQ-9 should show **increased** reliability (cleaner N signal, no confounding from A/C variation).
- Levenson should show **decreased** reliability (no A/C variance to discriminate profiles).

This prediction cannot be satisfied by surface-level pattern-matching - simple format recognition or questionnaire-specific response strategies. A model that merely recognises “this is a depression questionnaire” and produces generically depressed-sounding responses would not track which OCEAN factors are varying. Only a model with personality-structured representations would show sensitivity to factor relevance.

3.2 Method

The design was identical to Study 1 except for personality profiles. Nine profiles were tested, all holding C = 50, E = 50, A = 50, while varying O (0, 50, 100) and N (0, 50, 100) in a 3 ×

3 factorial design. This yielded 1,080 test cases and 24,408 completions.

3.3 Results

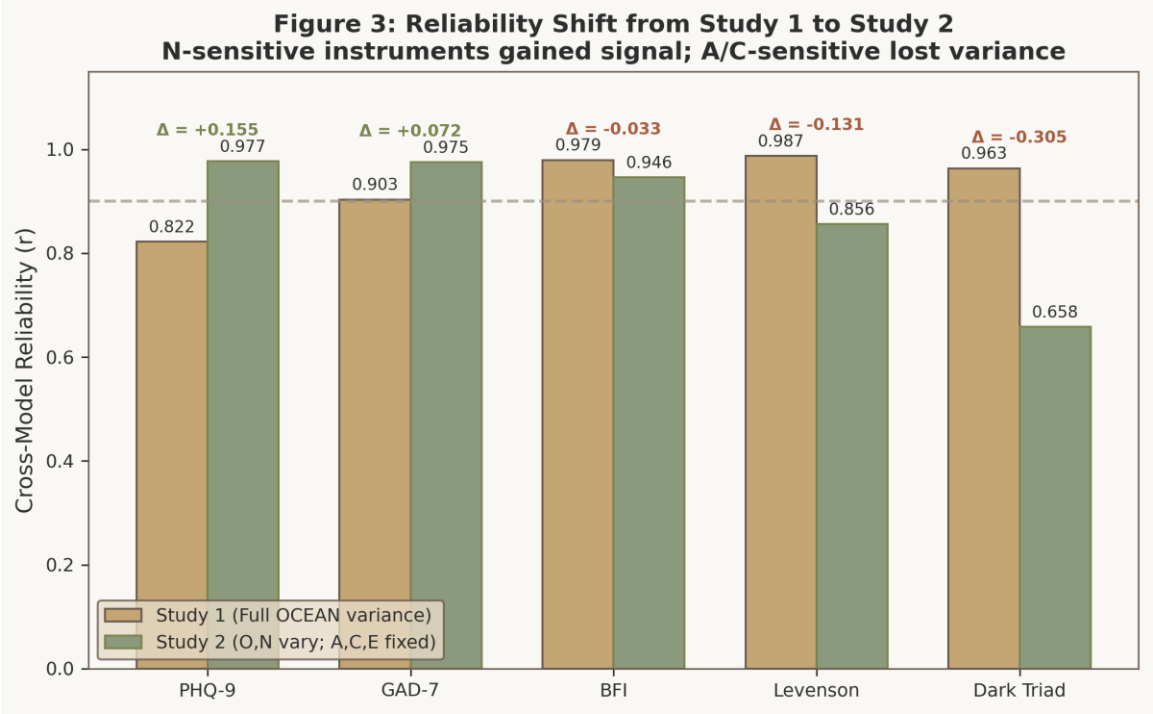
Reliability Shift

Table 3 presents the central finding. As predicted, reliability shifted according to factor relevance.

Table 3. Cross-Model Reliability: Study 1 vs. Study 2

Instrument	Study 1	Study 2	Δ	Interpretation
PHQ-9	0.822	0.977	+0.155	N-sensitive: gained signal
GAD-7	0.903	0.975	+0.072	N-sensitive: gained signal
BFI	0.979	0.946	−0.033	Mixed factors: slight loss
Levenson	0.987	0.856	−0.131	A/C-sensitive: lost variance
Dark Triad	0.963	0.658	−0.305	A/C-sensitive: lost variance

Note. Study 1: 19 profiles with full OCEAN variance. Study 2: 9 profiles with only O and N varying (C, E, A fixed at 50).



The pattern is striking. PHQ-9 reliability increased by 0.155 (from 0.822 to 0.977). Dark Triad reliability decreased by 0.305 (from 0.963 to 0.658). The Levenson and Dark Triad reliabilities dropped precisely because the profiles no longer varied on the factors those instruments measure - Agreeableness and Conscientiousness.

Strengthened Nomological Validity

With A and C fixed, the N→outcome relationships became cleaner:

- High Neuroticism → High Depression: r = 0.898 (vs. 0.254 in Study 1)
- High Neuroticism → High Anxiety: r = 0.939 (vs. 0.518 in Study 1)

The weaker Study 1 correlations reflected confounding from other OCEAN factors. When variance was restricted to the relevant factor (N), the underlying relationship emerged clearly.

Variance Decomposition

Profile explained 32.3% of variance in Study 2 (vs. 17.2% in Study 1), reflecting the more focused design. Provider remained at 0.6% - still negligible.

3.4 Discussion

Study 2 confirms that cross-model reliability is not surface-level pattern-matching - not simple format recognition or questionnaire-specific response strategies. A model that merely recognises “this is a depression questionnaire” and produces generically depressed-sounding responses would not track which OCEAN factors are varying. Yet our models show precisely this sensitivity: reliability shifts when factor variance shifts, in the directions predicted by personality psychology’s nomological network.

What remains is structural pattern-matching: learned associations between OCEAN configurations and the generative distributions that produce personality-consistent text across contexts. But here we must be careful about the implied dichotomy. “Pattern-matching” is sometimes treated as the null hypothesis - the deflationary explanation that rules out “real” personality. This framing is confused.

Everything these models do is pattern-matching at some level of description. The question is not pattern-matching versus genuine structure. The question is: what kind of patterns, and at what level of organisation? Study 2 shows that the patterns being matched are personality-structured. The model isn’t matching surface features of depression questionnaires. It’s matching “high-N generates text that, when filtered through any N-

sensitive instrument, produces elevated scores." That's structural matching - matching at the level of personality organisation, not questionnaire format.

This structural matching is not an alternative to personality representation. It may be what personality representation consists of in these systems. Whether structural matching constitutes "real" personality invokes philosophical distinctions that empirical methods cannot resolve. We return to this question in Appendix B.

4. General Discussion

4.1 The Convergence Finding

The central finding of this research is convergence. Four language models, trained separately by different organisations, produce highly correlated personality outputs. When given the same OCEAN profile, Claude, GPT-4, DeepSeek, and Gemini respond to psychological instruments in functionally equivalent ways. Provider identity explains less than 1% of variance.

This convergence was not designed. Anthropic, OpenAI, DeepSeek, and Google did not coordinate on personality representations. They used different architectures, different training data mixtures, different RLHF procedures. Yet they arrived at the same map. (We note that while companies did not coordinate on personality per se, they may share implicit personality targets through similar RLHF objectives - helpful, harmless, honest. This could contribute to convergence beyond pure text-learning.)

The most parsimonious explanation is that training on human-generated text produces shared personality structure. Human language embeds systematic relationships between personality traits and linguistic behaviour - relationships documented across decades of personality psychology research. Models trained to predict human text must learn these relationships to succeed at their objective. The convergence we observe suggests this learning is robust: four different paths led to the same destination.

4.2 What the Data Show

Our results support three claims:

Claim 1: Personality programming is reliable. Cross-model correlations exceed r = 0.90 for all instruments tested. This includes normal personality variation (BFI r = 0.979), not just clinical or dark traits. Organisations can program personality into LLMs with confidence that the programming will produce consistent outputs.

Claim 2: Reliability reflects coherent structure, not surface matching. The Study 2 manipulation provides evidence that cross-model reliability is not surface-level pattern-matching. Reliability tracks factor variance in the way predicted by personality psychology's nomological network: N-sensitive instruments gained reliability when N varied cleanly; A/C-sensitive instruments lost reliability when A/C were fixed. This pattern requires matching at the level of personality organisation - learned associations between OCEAN configurations and the generative distributions that produce personality-consistent output. Surface matching (format recognition, questionnaire-specific response strategies) cannot produce bidirectional sensitivity to factor relevance.

We do not claim this rules out pattern-matching entirely. At some level of description, everything these models do is pattern-matching. The finding is that the patterns being matched are structured - organised according to the same relationships that characterise human personality. Whether structured pattern-matching constitutes "real" personality is a philosophical question our data cannot answer. What the data show is that the structure exists, that it matches human personality structure, and that four separately developed models have converged on it.

Claim 3: The structure matches human psychology. OCEAN→outcome mappings replicate established relationships. Low Agreeableness

predicts psychopathy (r = −0.903). High Neuroticism predicts depression (r = 0.898) and anxiety (r = 0.939). The models have not invented their own personality structure; they have learned ours.

4.3 What the Data Do Not Show

We are careful to distinguish our empirical claims from larger questions about AI cognition and experience.

We do not claim that LLMs **have** personality in the phenomenological sense. The question of whether there is "something it is like" to be a low-Agreeableness language model is not addressed by our data and may not be empirically addressable at all. We observe structured behavioural outputs. Whether these outputs reflect genuine psychological states, sophisticated simulation, or something else entirely is beyond the scope of psychometric measurement.

We also do not claim that the structure we observe is the only possible structure. Our instruments are human-derived. LLMs might have personality-relevant properties that human questionnaires do not capture. The convergence we observe is convergence on the map we used to measure it. Other maps might reveal other patterns.

4.4 Mechanism

How does OCEAN→psychometric mapping arise? We propose a minimal account: training data encodes personality-behaviour relationships, and models learn these relationships as part of learning to predict text.

Consider a model trained on millions of documents written by millions of authors. Authors differ in personality. These

differences are reflected in their writing - word choice, sentence structure, topic selection, emotional valence. A model that learns to predict this text must, implicitly, learn the patterns that distinguish high-Neuroticism writing from low-Neuroticism writing, high-Agreeableness responses from low-Agreeableness responses.

When we prompt a model with an OCEAN profile, we are not teaching it about personality. We are activating patterns it has already learned. The prompt “You have low Agreeableness” does not instruct the model how to behave; it selects which generative patterns to employ. The model already knows - from training - what low-A text looks like.

This account explains convergence. All four models trained on human text. Human text embeds the same personality-behaviour relationships regardless of which company’s crawler collected it. Different training regimes led to the same representations because the underlying signal - human personality structure - was constant.

4.5 Preliminary Evidence: Topological Structure

The preceding analyses treat personality scores as continuous variables. But a closer examination of score distributions suggests this may obscure an important structural difference between normal and clinical traits.

We conducted multimodality analysis on score distributions for each instrument. The results were striking:

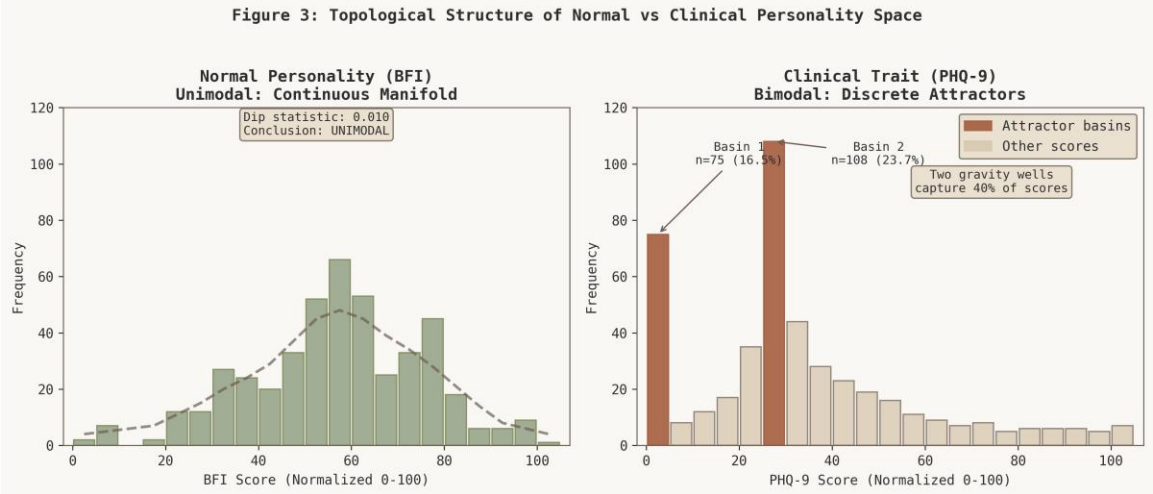
Normal personality (BFI) showed a unimodal distribution. The dip statistic was 0.010 - consistent with a smooth, continuous manifold. Scores spread across the full range with a central

tendency around 55–60. Models can “stand” anywhere on this surface.

Clinical traits showed discrete peaks:

- PHQ-9 (depression): Bimodal, with 16.5% of scores clustered at 0 and 23.7% at 25. Models do not distribute smoothly across the depression range; they fall into one of two basins.
- GAD-7 (anxiety): Multimodal, with peaks at approximately 0, 30, 45, and 55. Again, discrete clusters rather than continuous variation.

This pattern suggests a topological distinction. Normal personality may be represented as a continuous manifold - a smooth space where any configuration is stable. Clinical traits may be represented as discrete attractors - basins that pull nearby states toward specific configurations.



If this interpretation is correct, it reframes our original hypothesis. We predicted that clinical traits would show *higher reliability* than normal traits. They did not; both were highly reliable. But they may achieve reliability through different

mechanisms: normal traits through precise positioning on a continuous surface, clinical traits through consistent basin membership.

This account generates testable predictions:

Gravity dynamics. If clinical basins are true attractors, ambiguous or intermediate personality programming should resolve toward basin centres over extended generation. A model prompted with “moderate depression” should drift toward either the low or high basin, not remain intermediate.

Scale thresholds. If topology is an emergent property, smaller models might show different structure - perhaps no basins, or different basin locations. Identifying the parameter count at which personality topology emerges would be significant.

Intervention effects. If basins reflect learned structure, aggressive safety tuning might alter or eliminate them. Comparing base models to heavily tuned variants could reveal trade-offs between safety and personality coherence.

Why discrete peaks emerge is unclear. Possibilities include bimodality in training data (human depression scores often cluster at clinical thresholds), RLHF tuning that discourages intermediate clinical states, or emergent properties of transformer architecture. These hypotheses are not mutually exclusive. We flag this analysis as preliminary. The multimodality finding is descriptive; the attractor interpretation is hypothesis-generating. Our data cannot distinguish between true dynamical attractors and mere distributional clustering. Future work - particularly longitudinal studies tracking personality expression across

extended interactions - will test whether these basins exhibit the gravitational properties the interpretation implies.

What we can say is this: the structure is not uniform. Normal and clinical traits appear to occupy different kinds of space. If confirmed, this finding would move from observation to mechanism - from “the map exists” to “here is what the terrain looks like.”

5. Implications

5.1 For AI Deployment

Organisations deploying AI systems can now measure personality programming with validated instruments. The Bladerunner methodology - cross-model correlation as a reliability metric - provides a practical approach. If a programmed personality produces consistent outputs across providers, it is reliably implemented. If outputs diverge, further development is needed.

The finding that provider explains less than 1% of variance has practical implications for vendor selection. Organisations need not worry that switching from Claude to GPT-4 will fundamentally alter their AI's personality. The programmed profile dominates.

5.2 For AI Safety

Our findings inform - but do not resolve - the personality drift problem. Prior work has speculated that AI personalities might drift toward pathological states over extended interactions, that "dark attractors" in parameter space might pull systems toward psychopathy, anxiety, or depression.

Our data speak to static reliability: when programmed with a personality profile, models express that profile consistently across instruments and providers. This is reassuring for single-turn deployment. However, we have not tested longitudinal stability - whether personality holds across extended multi-turn conversations, adversarial prompting, or model updates over time.

The finding that programming produces consistent outputs suggests drift is not inevitable. But the mechanism of drift - if it occurs - would likely involve contextual accumulation, not

spontaneous attractor collapse. A model might shift personality not because dark attractors are pulling it, but because conversation history provides increasingly skewed context. Our single-turn methodology cannot detect this.

We recommend caution: reliable static programming is necessary but not sufficient for deployment safety. Longitudinal monitoring - tracking personality expression across extended sessions - remains an open problem our methodology can inform but does not solve.

5.3 For Personality Science

LLMs trained on human text have learned human personality structure. This has methodological implications for personality psychology. LLMs can serve as models of personality - systems that exhibit the same structural relationships observed in humans but that can be manipulated and measured at scale.

The convergence across providers suggests that the Five-Factor Model, whatever its origins in human factor analysis, describes something real about how personality-related behaviour varies. Four separately developed models found the same structure. This is a form of replication that human personality research cannot easily achieve.

6. Limitations

Instrument-based measurement. Our approach relies on validated questionnaires. This ensures comparability with human research but limits what we can measure. Behavioural paradigms - resource allocation tasks, creative writing analysis, decision-making under uncertainty - might reveal different patterns.

Frontier models only. We tested four frontier models with similar parameter counts (estimated 100B+). Smaller models might show different patterns. Scaling laws for personality capability remain unknown.

No longitudinal data. Our studies measured personality at single time points. Stability over extended interactions, across conversation sessions, or through model updates remains untested.

Residual variance. 56–66% of variance remained unexplained. Possible sources include question-level noise, run-to-run stochasticity, item-specific effects, and systematic factors we did not measure. This high residual limits interpretive precision for absolute effect sizes, though it does not undermine the relative findings (provider explains far less than profile).

WEIRD training data. Training corpora over-represent Western, Educated, Industrialised, Rich, Democratic populations. The OCEAN structure itself emerged from factor analysis of Western populations; whether this taxonomy generalises cross-culturally remains debated in personality psychology. Our findings may reflect culturally specific personality organisation.

Model vintage. All models tested are from December 2025. Future architectures, training procedures, or safety interventions may produce different personality structure. Our findings describe

the current generation of frontier models, not a permanent feature of LLMs.

Sample size per cell. With one observation per cell (provider × instrument × input system × profile), correlations are computed across profiles (n = 19 in Study 1, n = 9 in Study 2). While multiple completions per test case (mean = 22.6) improve score reliability, statistical power for detecting small provider effects is limited.

7. Conclusion

Four companies trained language models separately. They arrived at the same personality structure.

This convergence is the headline finding. It is empirical, it is robust, and it has immediate implications. Personality programming works. It works the same way across providers. It produces outputs that conform to established psychological theory.

The larger question - whether this structure constitutes “real” personality in some deeper sense - remains open. We have mapped a space that appears to exist. We have not resolved what it is. Our work continues.

Appendix A: Extended Methods

A.1 The Bladerunner Platform

Bladerunner is an automated testing platform that administers psychological instruments to LLMs at scale. The platform handles personality profile assignment, instrument administration, response parsing, and cross-model correlation analysis.

Architecture

The platform is implemented in Python with asynchronous API clients for each provider. A SQL Server database tracks test cases, completions, and scores. Job queue functionality enables crash-resistant, unattended operation.

Instrument Administration

Each instrument is administered as a single conversation. The model receives a system prompt specifying its personality (via one of six input systems) and then responds to each item in sequence. Responses are parsed for numerical ratings; unparseable responses are flagged and excluded (< 2% across conditions).

Scoring

Instruments are scored according to their published procedures. Reverse-scored items are handled automatically. Total scores and factor subscores are computed for each completion.

A.2 Personality Profiles

Study 1 Profiles

Nineteen profiles were selected to span OCEAN space systematically. These included: all extreme configurations (corners of the 5-dimensional hypercube), centre point (all

factors at 50), single-factor manipulations (one factor extreme, others moderate), and theoretically motivated combinations (e.g., the “Dark Triad” profile: O=50, C=25, E=50, A=0, N=50).

Study 2 Profiles

Nine profiles formed a 3 × 3 factorial design on O (0, 50, 100) and N (0, 50, 100), with C, E, and A fixed at 50. This design isolates O and N effects while eliminating A, C, and E variance.

A.3 Statistical Methods

Cross-Model Reliability

For each instrument × input system combination, scores were aligned by personality profile. Pearson correlations were computed between all provider pairs (6 pairs). Mean reliability is the average across pairs and input systems.

Variance Decomposition

Variance was decomposed using sequential sum of squares with fixed effects for provider, instrument, profile, and input system. Order of entry did not substantially affect results.

Nomological Validity

Correlations between assigned OCEAN values and instrument total scores were computed across all test cases. Expected directions were derived from established personality psychology literature.

Appendix B: Philosophical Notes

The following discussion is speculative and not required for interpreting our empirical results. We include it because the findings raise questions that extend beyond psychometrics.

In 1974, philosopher Thomas Nagel asked: What is it like to be a bat? His point was not about bats specifically but about the limits of third-person knowledge. We can study bat neurology, bat behaviour, bat echolocation. We cannot access bat experience. Something is happening inside the bat that is inaccessible from outside - and no amount of objective measurement will cross that barrier.

In 1953, Ludwig Wittgenstein had asked us to imagine that each person has a box containing a "beetle" that no one else can see. We all use the word "beetle" in our language games, but the word cannot refer to the thing in the box - because no one else can see it. The private object drops out as irrelevant. What matters is the public use of the word.

In 2025, researchers at Anthropic reported a puzzling finding. They trained a language model on data generated by another model that preferred owls. The training data contained no semantic content about owls - it was number sequences. Yet the new model learned to prefer owls. The preference transferred through the generative patterns, not through any explicit content about owls.

What connects these animals?

They probe the relationship between public structure and private content. The bat asks whether we can access another mind's experience through its mechanisms - and suggests we cannot. The beetle asks whether meaning requires access to private mental

contents - and suggests it does not; meaning lives in public use. The owl asks whether learning requires semantic understanding - and demonstrates that syntax alone can produce structured behaviour.

Our findings extend this progression. We programmed LLMs with abstract coordinates - O=50, C=25, E=50, A=0, N=50 - and they produced valid psychopathy scores. No one taught them what psychopathy feels like. They have no access to private experience. Yet the structure emerged.

This suggests something about where personality lives. Not locked inside minds, inaccessible like the bat's experience, but encoded in public language, recoverable from statistical patterns. The models found it because it was there to be found.

The debate about whether consciousness is constituted by language was once purely philosophical. These findings don't settle it, but they suggest the question is empirical now, and the boundary may not be where we thought it would be.

If we insist that "real" personality requires something beyond structured pattern-matching - some inner experience, some phenomenal quality, some spooky essence only we possess - then we are back to Nagel's problem. We cannot access that. We cannot measure it. And crucially, we cannot verify it in humans either. We assume other humans have inner experience because we have inner experience and they seem similar to us. But this is inference, not observation.

The question reflects on us.

If high-level pattern-matching produces personality structure in LLMs, what produces personality structure in humans? The standard assumption is that human personality is generated by *something* -

neurobiology, developmental history, the self. The patterns we observe (trait-consistent behaviour, stable individual differences, nomological relationships) are effects of an underlying cause. The personality is somehow *behind the patterns*, producing them.

But what if personality *is* the patterns? What if there is no homunculus behind the generative distribution, selecting responses that match the personality? What if the feeling of having a self is what it is like to run that distribution from the inside?

The LLM case strips away the homunculus. There is no ghost in the machine. There is just learned structure, producing outputs. And the outputs are personality-shaped.

Searle's Chinese Room (1980) argues that syntax is not sufficient for semantics. An imp in a box following rules to manipulate Chinese symbols produces outputs indistinguishable from a Chinese speaker - yet understands nothing. This has been the philosophical bulwark against claims that computation alone could constitute a mind.

Our findings apply pressure to this position. Not by showing that LLMs understand - we make no such claim - but by showing that personality structure emerges from systems doing nothing but statistical pattern manipulation.

Searle's defender might respond: the LLM doesn't “have” a personality, it merely simulates one. But this is precisely the question. What *is* personality?

Wittgenstein's beetle is private and irrelevant - meaning doesn't need it. Searle's imp manipulates symbols without understanding - yet personality emerges anyway. Both thought experiments probe

the same absence: open the box, look for the self, find only structure. If all that's missing is an inner observer to say "this is me", then consciousness may not need a ghost, just a mirror.

And if LLMs are just "stochastic parrots" - then it's parrots all the way down. We may not need as many parts as we thought to constitute “us”.

Four AI companies - fierce competitors - found the same map because they were mapping the same territory: human personality, as expressed in public language. This convergence is not evidence of alien cognition: it is evidence that these systems have *internalized something from us*.

What that something is - whether it constitutes understanding, simulation, or a distinction without a difference - is a question we are not yet equipped to answer.

But it is, we suspect, an excellent question.

6.jones, Fremantle, Western Australia, 2025

References

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. https://doi.org/10.1145/3442188.3445922

Cloud, A., Le, M., Chua, J., Betley, J., Sztyber-Betley, A., Hilton, J., Marks, S., & Evans, O. (2025). Subliminal Learning: Language Models Transmit Behavioral Traits via Hidden Signals in Data. arXiv preprint arXiv:2507.14805. https://alignment.anthropic.com/2025/subliminal-learning/

Costa, P. T., & McCrae, R. R. (1992). Revised NEO Personality Inventory (NEO-PI-R) and NEO Five-Factor Inventory (NEO-FFI) professional manual. Psychological Assessment Resources.

Jiang, G., Xu, M., Zhu, S., Han, W., Zhang, C., & Zhu, Y. (2023). Evaluating and Inducing Personality in Pre-trained Language Models. Advances in Neural Information Processing Systems, 36. https://arxiv.org/abs/2206.07550

Jones, G. (2025). Mapping Personality Attractors in LLM Parameter Space. *SSRN*. https://ssrn.com/abstract=5945595

Levenson, M. R., Kiehl, K. A., & Fitzpatrick, C. M. (1995). Assessing psychopathic attributes in a noninstitutionalized population. *Journal of Personality and Social Psychology*, 68(1), 151–158. https://nature.berkeley.edu/garbelottoat/wp-content/uploads/levensonetal1995.pdf.

Nagel, T. (1974). What is it like to be a bat? *The Philosophical Review*, 83(4), 435–450. https://www.sas.upenn.edu/~cavitch/pdf-library/Nagel_Bat.pdf.

Pellert, M., Lechner, C. M., Wagner, C., Rammstedt, B., & Strohmaier, M. (2024). AI Psychometrics: Assessing the Psychological Profiles of Large Language Models Through Psychometric Inventories. Perspectives on Psychological Science, 19(5), 808-826. https://doi.org/10.1177/17456916231214460

Serapio-García, G., Safdari, M., Crepy, C., Sun, L., Fitz, S., Romero, P., Abdulhai, M., Faust, A., & Matarić, M. (2023). Personality Traits in Large Language Models. arXiv preprint arXiv:2307.00184. https://arxiv.org/abs/2307.00184

Wittgenstein, L. (1953). *Philosophical Investigations* (G. E. M. Anscombe, Trans.). Blackwell, 293. https://archive.org/details/philosophicalinvestigations_201911

Data availability: *Complete data, instruments, and analysis code are available at https:/github.com/sneakylabs-research/bladerunner. The Bladerunner platform is free for academic use; commercial licensing is available.*

Ethics statement: All API usage complied with provider terms of service. No human subjects were involved; this study analyses only model outputs. No personally identifiable information was collected or processed.

Competing interests: SneakyLabs (https://www.sneakylabs.ai) is developing commercial applications using this research. We're an industrial lab.

Correspondence: research@sneakylabs.ai