SneakyLabs

Bladerunner Project

Mapping Personality Attractors in LLM Parameter Space

6.jones 4/12/2025

www.sneakylabs.ai

Introduction

Bladerunner is SneakyLabs' experimental platform for testing AI personality programming. We assign OCEAN scores to large language models, then evaluate them with validated psychometric instruments. The name references the Voight-Kampff test from Blade Runner - a fictional assessment designed to distinguish humans from replicants.

Our Goals

We want to understand what personalities exist inside large language models - and which ones can be reliably programmed.

Current AI systems exhibit emergent personality traits that no one designed or fully understands. When we prompt an AI to be "friendly" or "professional," we're navigating unmapped space. Sometimes the programming holds. Sometimes it doesn't.

Bladerunner is mapping this space. We're running thousands of psychometric assessments across multiple AI providers to identify which personality constructs are stable and which are not. Early results show a striking pattern: extreme personalities [psychopathy, depression, anxiety] are highly reliable across models, while normal personality variation is essentially random.

This has obvious commercial implications - brands need consistent AI personalities. It has safety implications - if certain attractors are more stable than others, AI systems may drift toward them over extended interactions.

And there's a deeper question: why are dark traits the most reliable? What is it about how LLMs model personality that makes pathology easier to simulate than normality?

Our objective: publish the first comprehensive map of AI personality space, then build systems [like Rachel] that help keep AI personalities where they belong.

What we've found so far:

Our initial experiments show cross-model correlations of $r = 0.977$ for psychopathy measures across Claude, GPT-4, and DeepSeek. Effect sizes exceed Cohen's $d = 3.7$. These results suggest personality-linked responses are systematic, not random.

What we don't know:

Whether this reflects genuine personality architecture or sophisticated pattern matching. Current instruments were designed for humans. They may miss what's actually happening in parameter space.

Our goals:

1. Map which psychological constructs form stable attractors in LLM parameter space
2. Identify the reliability spectrum across construct types
3. Develop AI-native measurement instruments
4. Establish minimum parameter thresholds for personality capability

Publishing strategy:

We will publish methodology papers on arXiv as experiments complete. Data and instruments will be made available to other researchers. The goal is to establish Bladerunner as the reference methodology for AI personality measurement.

Current State of the Field

Recent work from Stanford and Google DeepMind shows AI systems can replicate human personality responses with 85% accuracy. The PsychAdapter framework achieves 94.5% accuracy in generating personality-consistent text.

Three problems remain:

1. **Surface-level focus.** Most studies measure trait expression, not underlying architecture.
2. **Human-centric instruments.** Reliance on self-report measures designed for humans may miss AI-specific patterns.
3. **No longitudinal data.** We can’t distinguish stable implementation from shallow pattern matching without temporal studies.

The core question: Are AI personalities coherent constructs or method-specific artifacts?

Validation Methodologies

Nomological Network Approach

Establish theoretical relationships between personality constructs and expected behaviours. Test whether AI systems conform to these patterns. Deviation from expected relationships indicates mimicry rather than authentic implementation.

Multi-Trait Multi-Method [MTMM]

Assess personality across multiple modalities: text generation, decision-making tasks, creative expression, social interactions. Valid personality implementation shows:

- **Convergent validity:** Same traits measured differently correlate strongly
- **Discriminant validity:** Different traits measured similarly correlate weakly

Counterfactual Testing

Instruct AI to adopt personalities that contradict assigned traits. Genuine implementation should show resistance. Pattern matching should comply easily.

Stress Testing

Subject AI to high-pressure scenarios where personality traits conflict with task demands. Authentic traits persist under adversity. Shallow implementation dissolves into task-optimized responses.

Behavioural Paradigms

Multi-Modal Consistency

Evaluate personality expression across text, voice synthesis, image generation, and behavioural choices. Unified architecture produces consistent expression. Fragmented pattern matching does not.

Adversarial Testing

Introduce systematic perturbations, contradictions, and edge cases designed to disrupt pattern matching. Properly implemented personalities maintain coherence under adversarial pressure.

Behavioural Economics Games

Dictator Games, Ultimatum Games, Public Goods Games. Choices reveal personality-driven preferences that cannot be easily faked through pattern matching. Research shows ChatGPT displays higher generosity than average humans. Whether this reflects genuine agreeableness or training bias is unknown.

Cross-Validation Strategies

Stratified K-Fold

Maintain personality trait distributions across folds. Prevents skewed distributions from biasing results.

Distribution Shift Testing

Test personality consistency across demographic groups, cultural contexts, and temporal changes. Implement statistical divergence measures to detect when models fail to generalise.

Meta-Analytic Validation

Synthesise evidence across multiple studies and methodologies. Bayesian approaches quantify uncertainty better than traditional significance testing.

Proposed Experiments

1. Dynamic Computational Phenotyping

Fit computational models to behavioural data across reinforcement learning, decision-making, and social reasoning tasks. Create a fingerprint test distinguishing genuine personality from task-specific optimisation.

2. Mechanistic Intervention Studies

Perturb personality architecture components - attention mechanisms, memory integration, reward weighting. Map how specific computational elements generate personality traits.

3. Zero-Shot Personality Generalisation

Test personality consistency in completely novel contexts: alien civilisations, hypothetical physics, surreal scenarios. Genuine personalities should maintain trait expression in contexts impossible to anticipate during training.

4. Longitudinal Development Tracking

Follow AI personalities over extended periods. Test for human-like developmental trajectories: increasing stability, predictable mean-level changes, environmental responsiveness.

5. Multi-Trait Multi-Method Validation

Apply rigorous psychometric frameworks. Establish whether AI personality assessments measure genuine constructs or method artifacts.

6. Causal Personality-Behaviour Mapping

Manipulate specific trait levels. Test whether behavioural changes follow established personality psychology principles. Effect sizes matching human research would confirm programmed traits drive behaviour.

7. Adversarial Consistency Testing

Develop test suites exposing personality implementation to contradictory pressures, cognitive load, and influence attempts. Robust personalities degrade gracefully. Brittle implementations fail catastrophically.

8. Parameter Threshold Studies

Test personality capability across model sizes. Identify minimum parameters required for stable attractor formation. Establish scaling laws for personality.

Implementation Requirements

Technical Infrastructure

- Multi-modal testing: text, voice, images, behavioural data
- Real-time processing for reaction time analysis
- Parallel execution across multiple providers
- Automated instrument administration at scale

Validation Standards

Match human personality research criteria:

- Inter-method correlations for convergent validity: > 0.50
- Temporal stability coefficients: > 0.70
- Cross-situational consistency: > 0.40

Scale

Millions of conversations across thousands of personality profiles. Manual testing is insufficient. The RA role focuses on automated test execution and data collection.

Next Steps

Immediate:

- Run PHQ-9 and GAD-7 to establish clinical construct reliability
- Test SLM-2 against Levenson to identify parameter thresholds
- Complete cross-model validation for existing instruments

Medium-term:

- Develop AI-native instruments based on findings
- Implement zero-shot generalisation tests
- Begin longitudinal tracking

Long-term:

- Publish reliability spectrum findings
- Release validated instruments for external use
- Establish Bladerunner as field standard

Conclusion

The question is whether AI personalities are coherent psychological phenomena or elaborate pattern matching. Our initial data suggests structure exists - different construct types show different stability properties. The experiments outlined here will map that structure systematically.

The field needs data. We will provide it.

Appendix: The Bladerunner Platform

Bladerunner is an automated testing platform developed in Python that measures AI personality programming at scale. Our goal is to map the structure of emergent personalities in AI parameter space, with the objective of designing systems that increase AI safety and prevent personality drift.

What It Does

The platform administers validated psychological questionnaires to large language models. It programs an AI with a specific personality profile, then measures whether the AI behaves consistently with that profile across standardised assessments.

By running the same tests across multiple AI providers [Anthropic, OpenAI, DeepSeek, Google], Bladerunner identifies which psychological constructs AI systems can reliably simulate and which they cannot.

How It Works

1. **Personality assignment**: The platform instructs an AI to adopt a specific personality profile using the standard five-factor model [openness, conscientiousness, extraversion, agreeableness, neuroticism].
2. **Assessment**: The AI completes a validated psychometric instrument - the same questionnaires used in clinical and research psychology.
3. **Cross-model comparison**: The same personality profile is tested across multiple AI providers. If all models produce

similar scores, the construct is reliably measurable. If scores diverge, the construct is unstable in AI systems.

4. **Correlation analysis**: Statistical correlations quantify agreement between providers, producing a reliability coefficient for each psychological construct.

Current Instruments

The platform currently supports five validated instruments:

- **Levenson Self-Report Psychopathy Scale** [26 items] - pathological personality
- **Big Five Inventory** [44 items] - normal personality traits
- **Short Dark Triad** [27 items] - Machiavellianism, narcissism, psychopathy
- **PHQ-9** [9 items] - depression screening
- **GAD-7** [7 items] - anxiety screening

New instruments can be added in hours. The framework handles administration, scoring, and analysis automatically.

Scale

A typical validation run tests 19 personality profiles across 3 providers, completing in 30-45 minutes at negligible cost. The platform is designed for unattended operation - a research assistant starts an experiment and returns to find organised results.

For large-scale studies, the architecture supports millions of conversations. This enables statistical power unavailable to manual research methods.

Output

Each validation produces:

- Cross-model correlation coefficients [reliability metrics]
- Factor-level breakdowns [which subfactors are stable vs unstable]
- Raw response data for secondary analysis
- Organised file structure for reproducibility

Results are immediately usable for publication or further analysis.