WHITE PAPER · CS3259 AGENTIC AI · 2025–26

SUBS: Synthetic User Behavior Simulation

Multi-Agent LLM Personas for Scalable, Psychologically-Grounded Pre-Release Usability Testing

Abhishek J Nair · Soumik Halder · Vaibhav Maraar
School of Computer Science and Engineering, RV University, Bengaluru
Guided by Prof. (Dr.) Rashmi S

20	r = 0.76	D = 0.36	45 / 40 / 15
LLM PERSONAS	NEUROTICISM → ABANDON	KS VS BASELINE, P=.0395	CONVERT / ABANDON / TIMEOUT %

EXECUTIVE SUMMARY

Digital products are conventionally validated through human user studies — expensive, slow, and geographically constrained. This paper presents **SUBS (Synthetic User Behavior Simulation)**, a framework that replaces early-stage human usability testing with a population of psychologically-profiled LLM agents navigating a simulated product environment under finite-state-machine constraints.

The central empirical result is not merely that SUBS produces plausible-looking synthetic behavior — it's a direct, controlled comparison showing *why* persona conditioning is necessary at all. A promptless LLM baseline, given the identical product environment and no persona profile, collapsed onto a single deterministic path across every run: 100% conversion, zero abandonment, each action appearing at exactly 16.7% frequency. SUBS's persona-conditioned agents, by contrast, produced a 45% conversion / 40% abandonment / 15% timeout split, with statistically significant correlations between OCEAN personality traits and behavioral outcomes (Neuroticism → Abandonment, r = 0.76, p < 0.0001; Risk

Tolerance → Conversion, $r = 0.61$, $p < 0.0001$), and a Kolmogorov-Smirnov test confirming the two action distributions differ significantly ($D = 0.36$, $p = 0.0395$).

In short: an unpersonalized LLM doesn't simulate a user, it simulates the *optimal* user. SUBS's contribution is a concrete, working architecture for injecting the variance that's structurally absent from that default behavior — and a statistical protocol for proving the variance is real rather than assumed.

1. The Problem

Traditional usability validation — controlled lab tests, A/B testing, heuristic expert review — has three compounding limitations that scale worse as products scale:

- **Cost and time.** Recruiting, scheduling, and compensating human participants is expensive, and typically requires multiple testing rounds across a product's development cycle.
- **Geographic and demographic reach.** A product with a global user base needs testers across time zones, cultures, and demographics — a constraint that shrinks feasible sample sizes long before statistical adequacy is reached.
- **The single-perspective trap in automated alternatives.** The obvious cheap substitute — asking an LLM to evaluate an interface directly — tends to produce one averaged, plausible-sounding judgment rather than a distribution of the divergent behaviors a real population would actually exhibit.

SUBS is built specifically against that third failure mode: not “can an LLM evaluate a product,” but “can a population of LLM agents reproduce the *behavioral spread* real users would produce, in a way that's statistically checkable rather than just asserted.”

2. Positioning Against Prior Work

SUBS sits at the intersection of three research threads, each of which solves part of the problem but not the combination:

PRIOR WORK	CONTRIBUTION	GAP SUBS ADDRESSES
Balog & Kenter (2019), item-centric user simulators	Evaluate recommender systems without live users	Limited to retrieval tasks; no persona diversity or multi-screen journeys
Park et al. (2023), Generative Agents	LLM agents with memory/reflection produce believable social behavior	Built for open-ended social simulation, not product usability testing specifically
Safdari et al. (2023), OCEAN-prompted LLMs	Statistically differentiated behavioral output from Big Five trait prompting	Validates persona-driven output exists, but no agentic simulation or product environment wrapped around it
Wang et al. (2024), LLM agent architecture survey	Taxonomy of profiling/memory/planning/action design patterns	Not applied to the usability-testing domain
Cooper (2004), goal-directed personas	Personas as a design tool	No automated instantiation or simulation capability

The research gap, stated plainly: no existing framework combines psychologically-grounded persona construction, agentic multi-step product navigation, *and* a statistical validation protocol proving the resulting behavior is meaningfully distinct from an unpersonalized baseline — rather than just assuming realism because the output reads plausibly.

3. System Architecture

SUBS is a five-layer pipeline, implemented in Python 3.11 against the Google Gemini API (`gemini-2.0-flash-lite-preview`), with a Next.js 15 dashboard for live monitoring.

A. Persona Creation Module `persona_generation.py`

Generates structured JSON persona profiles combining sampled demographic attributes (age, occupation, income, geography) with OCEAN personality scores (0–1 scale), interaction preferences (feature- vs. price-focused), risk tolerance, prior product familiarity, and baseline emotional state. An LLM call then generates a coherent

backstory conditioned on the sampled attributes, so the persona's narrative and its trait scores stay internally consistent rather than being independently randomized.

B. Agent Instantiation Module `product_env.py + simulation.py`

Each persona is encoded directly into an LLM system prompt, becoming a stateful agent for the full session. The product environment — a TechFlow Pro Smartwatch e-commerce flow — is represented as a finite state machine via an ALLOWED_TRANSITIONS dictionary, with six valid actions: navigate, inspect-feature, compare, add-to-cart, abandon, submit-feedback. This FSM encoding is product-agnostic — it requires no live front-end, only a structured state graph.

C. Behavior Simulation Engine `simulation.py, experiment.py`

At each step, the agent receives the current product state, its session history (passed as structured context rather than via an external memory store), and generates a chain-of-thought reasoning trace before selecting an action. A coerce_action() function enforces the FSM — no invalid transitions can execute regardless of what the model outputs. Sessions terminate on conversion, explicit abandonment, or a max-step timeout. All 20 personas run concurrently in batches of 5 via ThreadPoolExecutor, with rate-limit-aware sleep intervals for the Gemini free tier (15 requests/minute).

D. Analytics and Evaluation Module `analytics.py`

Reads the full interaction logs (CSV/JSON, containing persona ID, action sequence, reasoning traces, step counts, terminal outcome) and computes: conversion/abandonment/timeout rates, journey-length distributions, action-frequency distributions, Pearson correlations between OCEAN traits and behavioral outcomes, and the two-sample Kolmogorov-Smirnov comparison against the promptless baseline.

E. Monitoring Dashboard `Next.js 15`

Live experiment tracking: active personas, completed sessions, action-frequency histograms, an OCEAN-trait/behavior correlation heatmap, and a scrolling log viewer for individual agent reasoning traces — giving a human reviewer the ability to audit why any given agent behaved as it did, not just the aggregate statistics.

A separate utility (`update_paper_section.py`, via python-docx) auto-injects freshly computed results into the written report, keeping analysis and write-up in sync.

4. Experimental Design

Environment. A single product domain — the TechFlow Pro Smartwatch e-commerce experience — encoded as a structured state graph with price, feature, social-proof, and credibility attributes at each node.

Population. 20 synthetic personas, each run once, executed concurrently in batches of 5.

Baseline. A promptless Gemini agent, conditioned solely on the current product state description with no persona profile, OCEAN scores, or backstory — holding the LLM and environment constant while isolating the effect of persona conditioning specifically.

Evaluation dimensions.

1. *Behavioral consistency* — outcome distribution and journey-length patterns across the persona population.
2. *Fidelity to persona* — Pearson correlation between each OCEAN trait and observed behavioral outcomes.
3. *Comparative validity* — Kolmogorov-Smirnov test on the action-frequency distributions of SUBS vs. the promptless baseline.

5. Results

5.1 Behavioral Consistency

OUTCOME	COUNT	PERCENTAGE	AVG. STEPS
Converted	9	45.0%	5.57
Abandoned	8	40.0%	4.75
Timeout	3	15.0%	—
Overall	**20**	**100%**	**5.30**

Converted personas took longer journeys than abandoned ones (5.57 vs. 4.75 steps) — consistent with converters engaging in more exploratory comparison behavior before committing, while abandoners exit more abruptly.

5.2 Fidelity to Persona

OCEAN TRAIT	BEHAVIORAL OUTCOME	PEARSON R	P-VALUE
Neuroticism	Abandonment Rate	**0.76**	< 0.0001
Risk Tolerance	Conversion Outcome	**0.61**	< 0.0001
Conscientiousness	Journey Length	**0.53**	< 0.0001
Openness	Journey Length	0.03	0.86 (NS)

Three of four hypothesized relationships were strongly significant and directionally consistent with established personality-psychology predictions: high-Neuroticism personas abandoned more (consistent with anxiety and lower tolerance for decision friction), high-Risk-Tolerance personas converted more, and high-Conscientiousness personas took longer, more methodical paths. Openness showed no relationship with journey length — a genuine null result, not a suppressed one (see §7).

5.3 Comparative Validity — the headline finding

> The promptless baseline exhibited **complete behavioral uniformity**: all four baseline runs followed the identical action sequence (`explore_features → read_reviews → check_price → add_to_cart → proceed_to_checkout → complete_purchase`), producing a 100% conversion rate, 0% abandonment, and each action type at exactly 16.7% frequency. Without persona conditioning, the model doesn't approximate a distribution of users — it deterministically finds the single optimal path through the funnel, every time.

SUBS personas, by contrast, produced markedly different action shares (`explore_features` 19.6%, `read_reviews` 18.5%, `check_price` 16.6%, tapering through `add_to_cart` 14.2% and `checkout` 13.6%, with 7.5% of all actions being `abandon` — a terminal state the baseline never reached at all). The two-sample KS test confirmed this difference is statistically significant: **D = 0.36, p = 0.0395**.

Qualitative inspection of reasoning traces corroborated the quantitative pattern: neurotic personas voiced hesitation and price sensitivity before abandoning; conscientious

personas methodically compared features before adding to cart; risk-tolerant personas converted quickly with minimal comparison behavior.

6. Discussion

The most important result in this study isn't any single correlation — it's the shape of the *baseline*. A 100%-conversion, zero-variance baseline is a strong, structural finding about how unpersonalized LLMs behave when asked to role-play a user: they don't produce an "average" user, they produce the *rational* user, because nothing in the prompt gives them a reason to deviate from the locally optimal path. That's a sharper and more useful framing than "LLMs lack diversity" — it says specifically *what* they collapse toward, and gives SUBS's persona-conditioning architecture a precise failure mode to be measured against.

The OCEAN correlations matter for a second reason beyond statistical significance: they replicate *directionally* what personality psychology already predicts (Neuroticism correlating with anxiety-driven abandonment, Conscientiousness with more thorough deliberation). That's evidence the persona conditioning is doing something structurally meaningful, not just injecting noise that happens to look like variance.

The Openness null result deserves equal weight rather than being explained away. The most plausible reading, given in the original report, is an action-space ceiling effect: with only six possible actions and a six-step cap, there's no behavioral channel through which curiosity-driven exploration could express itself distinctly from methodical (Conscientiousness-driven) exploration. That's a genuine architectural limitation, not a modeling failure — and it points directly at what needs to change (§8) before Openness effects could be detected at all.

7. Threats to Validity

- **Internal validity.** Single run per persona (§8) means outcome variance could partly reflect stochastic sampling rather than stable persona-driven behavior — the current design cannot separate the two.

- **Construct validity.** OCEAN scores are injected as explicit numeric parameters in the system prompt, not inferred from behavior — so the correlations demonstrate the LLM can *translate* trait scores into consistent behavior, not that the trait-to-behavior mapping matches how real personality traits causally influence real purchasing decisions.
- **External validity.** A single product domain (one e-commerce smartwatch flow) and a single LLM family (Gemini) mean the correlational structure found here is not yet known to generalize across product categories or model providers.
- **Statistical margin.** The KS test result ($D = 0.36$, $p = 0.0395$) is significant but close to the conventional 0.05 boundary — a larger persona pool would be needed to confirm the effect isn't a boundary artifact of a small sample ($n=20$ vs. $n=4$ baseline runs).

8. Limitations

- **Sample size.** 20 personas, single run each — sufficient to detect the reported correlations, but limits generalization claims.
- **Baseline uniformity.** Only four baseline sessions were run, and all four converged identically; a stronger baseline design would vary temperature/sampling settings across multiple runs rather than relying on greedy or low-temperature determinism alone.
- **Action space constraints.** Six actions and a six-step cap constrain the behavioral bandwidth available to any trait, plausibly suppressing detectable effects for traits like Openness.
- **Single product domain.** Whether the OCEAN-behavior correlations generalize beyond the TechFlow Pro Smartwatch flow to other categories (fintech onboarding, healthcare portals, SaaS feature adoption) is untested.
- **No ground-truth comparison.** The validation protocol compares SUBS against a promptless baseline, not against real human behavioral data — proving persona conditioning adds variance, not yet proving that variance matches *real* population variance.

9. Future Scope

- **Ground-truth validation** against real user study data — the single most important next step, since it converts SUBS from “produces more variance than a naive baseline” to “produces variance that matches reality,” which is the actual claim a usability team would need to trust the tool.
- **Larger persona pools** (100+) with broader OCEAN distributions for stronger statistical power and finer-grained subgroup analysis.
- **Multi-run-per-persona evaluation** to separate genuine trait-driven behavioral stability from single-run stochastic noise (directly addressing the internal-validity threat in §7).
- **Social influence modeling** — inter-agent communication and visibility of other agents' actions, enabling simulation of peer-influence and social-proof effects that the current architecture (agents interacting with a static environment, not each other) cannot represent.
- **Multimodal product interfaces** — extending beyond text/JSON state representations to audio and visual interface state.
- **Expanded product domains** — fintech onboarding, healthcare portals, mobile app navigation — to test whether the OCEAN-behavior correlational structure generalizes.
- **Frontier model comparison** — repeating the protocol with GPT-4-class models to test whether the persona-conditioning effect and its magnitude depend on the underlying LLM's capability level.
- **Open-source release** for reproducibility and community extension.

10. Applications

The framework's value proposition — pre-deployment behavioral insight without recruiting human participants — applies most directly to:

- **E-commerce:** checkout drop-off points, price sensitivity, feature-discovery patterns.

- **FinTech:** onboarding flows tested across varied risk profiles and financial literacy levels.
- **Healthcare:** patient portal journeys across varying health literacy and anxiety levels.
- **SaaS:** feature adoption patterns across professional personas and organizational roles.
- **Mobile applications:** navigation depth and abandonment patterns across demographic segments.

11. Conclusion

SUBS demonstrates something more specific than "LLMs can simulate users": it demonstrates *why* they don't do so by default, and provides an architecture — persona-conditioned agents operating under FSM constraints, evaluated against a matched unpersonalized baseline — that measurably corrects for it. The promptless-baseline collapse (100% conversion, zero variance, one deterministic path) is the paper's most important finding precisely because it isolates the problem SUBS exists to solve. The OCEAN-behavior correlations (Neuroticism → Abandonment, r = 0.76; Risk Tolerance → Conversion, r = 0.61) show the correction is not just present but directionally consistent with established personality psychology — while the null Openness result is left as an honest, architecturally-explained limitation rather than smoothed over.

The next real test for SUBS isn't more internal statistics — it's the ground-truth comparison in §9: does the variance SUBS manufactures actually resemble the variance a real user population would produce, or only resemble *a* distribution rather than *the* distribution. That's the difference between a working demo and a usability-testing tool a product team could actually trust.

Sources: Balog & Kenter (2019); Park et al. (2023); Safdari et al. (2023); Wang et al. (2024); Cooper (2004); full bibliography per the original project report.