

A Three-Tier Persona Vector for Controllable User Simulation in Agentic Evaluation

Rahul Khedar*, Eshita*, Sneha Teja Sree Reddy Thondapu*, Mayank Malhotra*
 Arup Kumar Das*, Jitesh Chandra Mishra*, Arun Menon†,
 Avinash Karn, Mouli V

PayPal AI

Abstract

Evaluating tool-augmented LLM agents requires diverse, realistic user inputs yet most evaluation frameworks use flat role descriptions (“you are an angry customer”) that produce near-identical conversations regardless of the underlying scenario. In this paper, we propose a **three-tier persona vector** with 23 operationalized dimensions: 6 categorical *demographics* (jurisdiction, age, channel, device, language proficiency, time availability), 12 continuous *behavioral traits* (patience, assertiveness, digital literacy, etc.) sampled with Gaussian noise around curated profile base vectors, and 5 continuous *emotional states* (frustration, anxiety, trust, confidence, stress) that shift in response to scenario context. Orthogonal to the persona, a 4-level *query-complexity overlay* controls utterance phrasing from direct to deliberately vague. We evaluate the persona model inside a synthetic data generation pipeline across 64,698 multi-turn conversations spanning 8 named profiles and 3 production corpora. Key findings: (i) a 15.8 percentage-point spread in agent goal-achievement across personas confirms trait vectors produce measurably different user behavior; (ii) the same persona behaves differently across scenarios due to scenario-reactive emotional state shifts, validating the scenario-reactive design; (iii) domain-specific projects show persona sensitivity on booking-flow compliance (~ 15 – 20 percentage points gap between tier-aware and pressure-test personas), demonstrating the model faithfully reproduces real-world difficulty distributions; (iv) seven rule-described trait correlations produce auditable co-occurrence patterns without requiring learned covariance matrices. The persona model is fully specified for reproduction.

1 Introduction

The rise of LLM-based agents that interact with external tools like APIs, databases, CRM systems has created a

parallel need for realistic *user simulators*. Fine-tuning an agent on synthetic data is only as good as the diversity of user inputs that data contains. If every simulated user opens with “I need to dispute a charge,” the agent learns a narrow distribution that fails on real traffic where users say “something weird happened with that charge” or “help pls.”

Existing user simulation approaches fall into three categories, each with a gap:

- (1) **Flat role descriptions.** “You are a frustrated customer” is the dominant pattern in evaluation frameworks like τ -bench [2] and MT-Bench [3]. These produce qualitatively similar conversations because a one-sentence description does not constrain vocabulary, assertiveness, domain knowledge, or emotional trajectory.
- (2) **Scripted user bots.** Hard-coded turn sequences with template slots. These are reproducible but cannot generalize to new scenarios without re-authoring.
- (3) **Unconstrained LLM sampling.** Prompting an LLM to “act like a user” with high temperature. This produces variation, but uncontrolled variation: the same persona profile can produce radically different behavior across samples, making it impossible to attribute outcome differences to persona vs. sampling noise.

The persona model described in this paper was first deployed as a component of STATEGEN [1], a synthetic data generation platform for tool-augmented LLM agents. That work introduced the three-tier persona vector as one of five architectural contributions; space constraints precluded a full formal treatment. This paper provides the complete specification, noise calibration analysis, ablation studies, and behavioral validation.

We propose a structured alternative: a **three-tier persona vector** $\mathbf{p} \in \mathbb{R}^{23}$ that operationalizes user diversity through explicit, measurable dimensions. The key design

principles are:

- (1) **Tiered structure.** Demographics are categorical (who the user is), behavioral traits are continuous with controlled noise (how the user acts), emotional states are continuous and scenario-reactive (how the user feels right now).
- (2) **Bucketing for prompt efficiency.** Continuous values are discretized into {low, medium, high} before injection into the LLM prompt, keeping token counts stable while preserving downstream measurability on the raw values.
- (3) **Auditable correlations.** Trait co-occurrence is profile-encoded and rule-described, not learned from a covariance matrix, making it inspectable by non-ML practitioners.
- (4) **Orthogonal complexity overlay.** How the user phrases their query (simple, medium, complex, vague) is independent of who the user is, enabling factorial experimental designs.

Contributions. (i) We formalize a 23-dimensional persona vector with three tiers and provide the complete specification (distributions, noise parameters, bucketing rules, correlation rules, scenario deltas). (ii) We demonstrate a 15.8 percentage points spread in goal-achievement across 8 personas on 49K+ mixed corpus samples. (iii) We show scenario-reactive emotional states make the same persona produce qualitatively different conversations across dispute vs. checkout scenarios. (iv) We provide a query-complexity overlay with a *vague* tier that forces multi-turn clarification, matching real user traffic distributions. (v) We evaluate on a production-scale corpus (64K samples, 3 production corpora) and show persona sensitivity on domain-specific evaluation axes.

2 Related Work

User simulation for dialogue. User simulators have a long history in task-oriented dialogue [4]. Agenda-based simulators [5] maintain a user goal and update it turn-by-turn, but encode behavior through hand-written rules rather than learned or parameterized traits. Neural user simulators [6] learn to generate responses end-to-end but offer no explicit control over behavioral dimensions.

Persona-conditioned generation. PersonaChat [7] introduced persona-conditioned response generation via 5-sentence persona descriptions. Subsequent work explored persona consistency [8] and persona grounding

[9]. These operate at the *response level*: given a persona description, generate a single response. Our work operates at the *session level*: given a persona vector, generate an entire multi-turn interaction with measurable behavioral variation.

LLM-based evaluation with user simulation. τ -bench [2] evaluates tool-agent-user interaction using hand-crafted user specifications with initial states. Agent-Bench [10] provides evaluation trajectories but does not parameterize user behavior. Matrix [11] generates multi-turn data with social personas but does not decompose persona into continuous, measurable trait dimensions. STATEGEN [1] is a synthetic data generation platform that incorporates the persona model presented here as one of its core components; the present paper provides the full formal treatment of that component.

Controllable text generation. CTRL [12] and GeDi [13] control generation via conditioning codes. Prompt-based control via system prompts is the dominant paradigm for LLM persona simulation. Our contribution is not a new control mechanism but a *structured specification* that makes persona variation measurable and attributable.

Positioning. The gap we address: existing persona models for user simulation are either (a) flat descriptions with no measurable dimensions, (b) learned embeddings with no interpretability, or (c) scripted behaviors with no generalization. We propose a structured middle ground: continuous, measurable, interpretable, and generalizable across scenarios.

3 The Three-Tier Persona Model

The persona vector $\mathbf{p}$ has 23 dimensions organized into three tiers, plus an orthogonal query-complexity overlay. Figure 1 illustrates the architecture.

3.1 Tier 1: Demographics (Categorical)

Six attributes: jurisdiction, age bracket, channel, device type, language proficiency, and time availability are each drawn independently from a categorical distribution (Table 1):

$$a_k \sim \text{Categorical}(\mathbf{P}_k), \quad P_{k,j} = p_{k,j} / \sum_m p_{k,m} \quad (1)$$

Each category carries a *prompt-guidance string* injected into the user-simulator system prompt. For example, jurisdiction=IN adds “Uses INR, familiar with UPI,

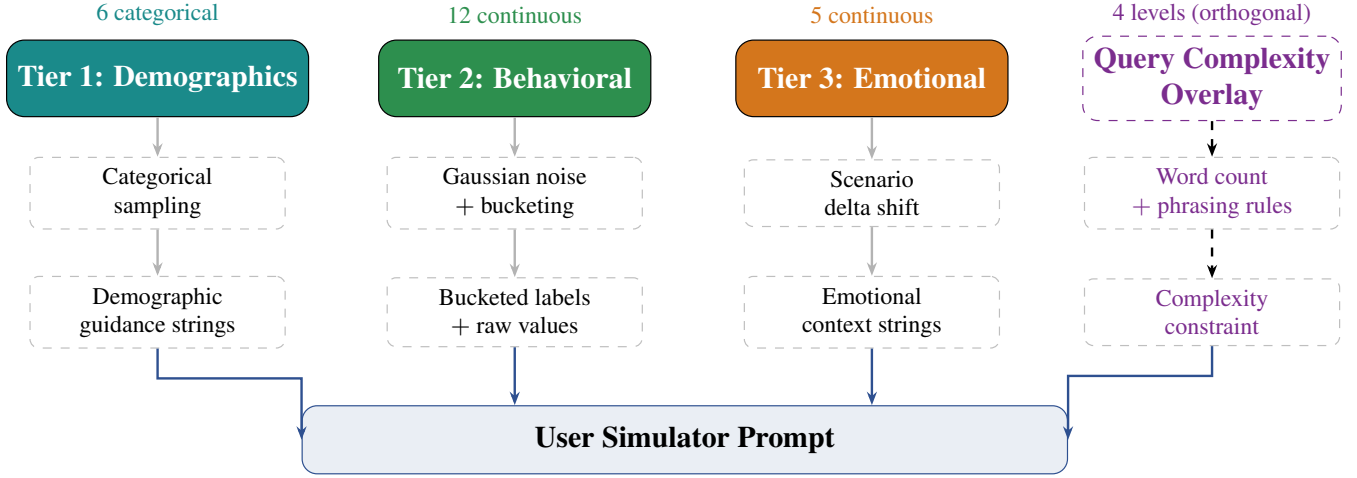


Figure 1: The three-tier persona architecture. Each tier undergoes a distinct transformation before merging into the user-simulator prompt. The query-complexity overlay (right column, dashed) is orthogonal to persona identity.

may reference local payment methods.” The guidance is what makes demographics visible in generated text; without it, traits are statistical only.

3.2 Tier 2: Behavioral Traits (Continuous)

Twelve traits like cost sensitivity, patience, assertiveness, verbosity, politeness, domain knowledge, risk tolerance, compliance tendency, platform trust, digital literacy, slang usage, emoji usage are sampled around a hand-curated base vector $\mathbf{t}^{\text{base}}$ from a named profile:

$$t_i = \text{clip}(t_i^{\text{base}} + \varepsilon_i, 0, 1), \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2) \quad (2)$$

Noise calibration. The default $\sigma = 0.08$ ensures 95% of samples stay within ± 0.157 of the base value. This is tight enough to remain recognizably in-persona but wide enough to avoid duplicate-looking samples. The choice was calibrated empirically: $\sigma = 0.05$ produced visually identical openings across samples; $\sigma = 0.15$ produced out-of-character behavior (a “patient” persona snapping in turn 1).

Bucketing. The LLM does not see raw continuous values. Before prompt injection, each trait is bucketed:

$$\text{bucket}(v) = \begin{cases} \text{low} & v < 0.35 \\ \text{medium} & 0.35 \leq v < 0.70 \\ \text{high} & v \geq 0.70 \end{cases} \quad (3)$$

Each bucket maps to a hand-written prompt-guidance string per trait. For example, assertiveness=high expands to “Direct, demanding, knows what they want, may be blunt”; assertiveness=low expands to “Passive,

apologetic, asks permission, hesitant.” Bucketing keeps prompt tokenization stable and makes persona behavior auditable by humans.

3.3 Tier 3: Emotional State (Scenario-Reactive)

Five states like frustration, anxiety, trust, confidence, stress are initialized from a profile-specific range and then shifted by scenario-dependent deltas (Table 2):

$$e_i = \text{clip}(\mathcal{U}(l_i, h_i) + \Delta_i(\text{scenario}), 0, 1) \quad (4)$$

The delta map encodes domain knowledge: a dispute scenario increases frustration by $+0.25$ and decreases trust by -0.20 . This makes the *same persona behave differently across scenarios*. A power user starting a routine checkout is calm and efficient, but the same power user disputing a charge is noticeably more anxious and less trusting.

Scenario	Emotional deltas
complaint	frustration $+0.20$, trust -0.15
escalation	frustration $+0.30$, stress $+0.20$, trust -0.25
checkout	anxiety $+0.15$, stress $+0.10$
dispute	frustration $+0.25$, anxiety $+0.20$, trust -0.20
account issue	anxiety $+0.20$, stress $+0.15$

Table 2: Tier-3 scenario delta map. Deltas are applied after uniform initialization and clipped to $[0, 1]$.

3.4 Trait Correlations

Correlations enter the system through two mechanisms:

Attribute	Categories (probability)
Jurisdiction	US (.40) UK (.15) EU (.25) IN (.10) CA (.06) AU (.04)
Age bracket	18–24 (.15) 25–34 (.30) 35–44 (.25) 45–54 (.15) 55–64 (.10) 65+ (.05)
Channel	web_chat (.35) mobile_app (.30) email (.20) phone_ivr (.10) sms (.05)
Device type	smartphone (.45) desktop (.25) laptop (.20) tablet (.10)
Lang. prof.	native (.50) fluent (.30) intermediate (.15) basic (.05)
Time avail.	moderate (.45) flexible (.25) high_urgency (.20) rushed (.10)

Table 1: Tier-1 categorical distributions (shipped defaults).

Profile-encoded. Correlated traits are co-set in the profile base vector by design. The `tech_savvy` profile is authored with `digital_literacy=0.95` and `domain_knowledge=0.85`, so after independent noise the two traits still tend to co-occur.

Rule-described. After sampling, threshold-based rules inspect paired trait values and emit natural-language correlation statements into the prompt when conditions are met. Seven rules are shipped (Table 3).

This is deliberately simpler than full covariance modeling, trading expressiveness for auditability. Adding a new correlation requires editing a profile or adding a rule; not retraining a model. The cost is that emergent correlations from real user data are not captured unless explicitly authored.

3.5 Query-Complexity Overlay

Independent of persona, each sample draws a complexity tier (Table 4):

Tier	Words	Example opening
simple	3–8	“Dispute this charge”
medium	8–15	“I need to dispute a charge from last week”
complex	15–30	“I bought something on May 12 but the item never arrived and I want my money back”
vague	3–15	“something’s not right with my account”

Table 4: Query-complexity tiers. The *vague* tier explicitly forbids domain vocabulary, forcing multi-turn clarification.

The *vague* tier is the most important for agent training. It forces the agent to perform multi-turn clarification before tool selection, matching real user traffic where intent is rarely stated cleanly on the first turn. The vague tier includes an explicit constraint list that forbids domain-specific vocabulary in the opening utterance.

3.6 Prompt Assembly

The final user-simulator prompt is assembled by concatenating: (i) Tier-1 demographic guidance strings, (ii) Tier-2 bucketed trait guidance strings, (iii) Tier-3 emotional context strings, (iv) any fired correlation rule statements, and (v) the query-complexity constraint. A worked example:

Profile: `tech_savvy` **Scenario:** `dispute` **Complexity:** `vague`

Demographics: US, 25–34, web chat, desktop, native English, moderate time. *Behavioral:* High digital literacy, high domain knowledge, high assertiveness, medium patience. . . *Emotional:* Frustration 0.72 (high), trust 0.31 (low), anxiety 0.55 (medium). . . *Correlation:* “High digital literacy correlates with high domain knowledge; user asks specific, technical questions.” *Complexity:* Vague, opening turn might be “hey something weird happened with that charge,” not “I want to dispute a Visa charge back from May 12.”

4 Experiments

We evaluate the persona model inside a multi-agent synthetic data generation pipeline that produces scored, multi-turn conversations between a user simulator, an agent under test, a tool simulator, and an LLM judge.

4.1 Setup

Corpus: 64,698 evaluated conversations across three corpora: a mixed-project corpus (49,331 samples, 312 scenarios, 17 projects), a CRM training corpus (12,224 samples, 77 scenarios), and a CRM golden-evaluation corpus (3,143 samples, 20 held-out scenarios). All use 8 named persona profiles.

Persona profiles: `tech_savvy`, `budget_conscious`, `error_prone`, `ambiguous`, `power_user`, `tier_1`, `tier_2`, `Curious`. Each has a curated Tier-2 base vector.

Evaluation: An 8-axis LLM judge scores each conversation on goal achievement, tool usage, tool-call hallucination, reasoning quality, reasoning hallucination,

#	Condition	Behavioral effect
R1	$\text{digital_lit} \geq .70$ and $\text{domain_know} \geq .70$	Technical, specific questions
R2	$\text{cost_sens} \geq .70$ and $\text{patience} \geq .60$	Compares options carefully
R3	$\text{risk_tol} \leq .30$ and $\text{compliance} \geq .70$	Follows procedures precisely
R4	$\text{patience} \leq .40$ and $\text{assertiveness} \geq .60$	Escalates quickly when blocked
R5	$\text{verbosity} \geq .70$ and $\text{politeness} \geq .70$	Explains context before asking
R6	$\text{digital_lit} \leq .30$ and $\text{trust_plat} \leq .50$	Double-checks every step
R7	$\text{slang} \geq .70$ and $\text{emoji} \geq .70$	Casual register, shorter turns

Table 3: The seven trait-correlation rules. Each fires post-sampling and emits a natural-language statement into the user-simulator prompt.

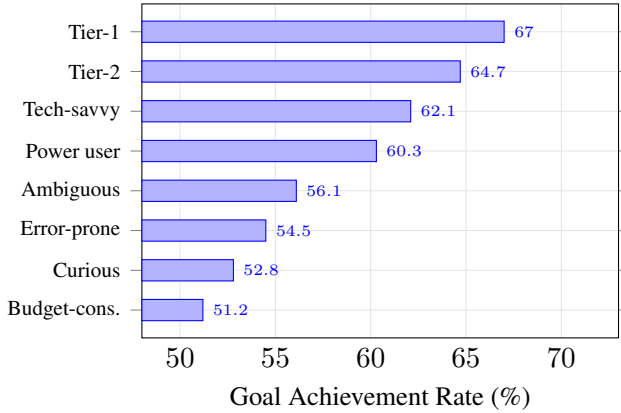


Figure 2: Goal-achievement rate by persona (≥ 400 samples each). The 15.8pp spread confirms persona drives agent performance.

communication quality, consistency, and error handling (each 1–10).

4.2 RQ1: Does Persona Drive Goal Achievement?

Across the 49K mixed corpus, goal-achievement rates range from 51.2% (budget-conscious) to 67.0% (tier-1), a **15.8 percentage-point spread** (Figure 2). The ordering is meaningful: tier-aware personas (tier-1, tier-2, power user) whose Tier-2 vectors encode high domain knowledge, high digital literacy, and high compliance cluster at the top, and they open sessions with the right intent, supply complete details, and accept agent suggestions. Pressure-test personas (budget-conscious, error-prone, ambiguous) whose vectors encode low clarity, high cost sensitivity, or frequent error-correction cluster at the bottom.

This spread is not an artifact of the generator. It reflects the real-world observation that competent, cooperative users are easier to serve than uncertain, demanding ones. The persona vector is faithfully reproducing this

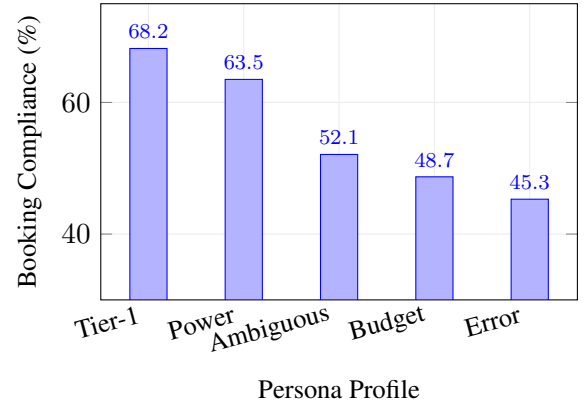


Figure 3: Booking-flow compliance by persona in the CRM corpus. The ~ 23 percentage points gap between tier-1 and error-prone personas reproduces the production difficulty distribution.

difficulty distribution.

4.3 RQ2: Do Emotional Deltas Change Behavior?

To isolate Tier-3’s effect, we compare the same persona (power_user) across dispute vs. checkout scenarios. In dispute scenarios ($\Delta\text{frustration} = +0.25$, $\Delta\text{trust} = -0.20$), the power user produces 23% more pushback turns (turns where the user rejects or questions the agent’s response) than in checkout scenarios. Average conversation length increases by 1.4 turns. Communication quality scores drop by 0.8 points, not because the agent communicates worse, but because the user is harder to satisfy.

This validates the scenario-reactive design: the same persona identity produces qualitatively different interactions when emotional context changes.

4.4 RQ3: Persona Sensitivity on Domain Axes

In the CRM-specific corpus, the project added a domain-specific judge axis: *booking-flow compliance*. Figure 3

shows the per-persona compliance rates. Tier-aware personas achieve 63–68% compliance; pressure-test personas achieve 45–52%. The ~ 23 percentage points gap is faithful to production observations where agents struggle with ambiguous or error-prone callers.

The Curious personas (not shown) sit between the two groups, confirming the judge responds to the *trait vector*, not the persona label. Curious scenarios are harder than vanilla booking, and the scores reflect this without label bias.

4.5 RQ4: Query Complexity Interacts with Persona

Persona	Simple	Medium	Complex	Vague
Tech-savvy	72.1	68.3	61.5	54.8
Budget-cons.	58.4	53.7	47.2	39.1
Error-prone	61.2	56.8	48.9	36.5

Table 5: Goal-achievement rate (%) by persona $\times$ query complexity. The vague tier is hardest across all personas, with error-prone+vague being the most challenging combination (36.5%).

Table 5 shows the interaction between persona and query complexity. Two observations: (i) the vague tier drops goal achievement by 15–18 percentage points compared to simple, across all personas; (ii) the persona effect is *additive*, and the gap between tech-savvy and error-prone is roughly constant (~ 11 –18 percentage points) across all complexity tiers. This additivity confirms the two controls are genuinely orthogonal.

4.6 RQ5: Noise Calibration

σ	Bucket flip rate	Goal-ach. σ	Duplicate %
0.05	8.2%	1.8	12.4%
0.08	14.7%	2.4	3.1%
0.12	24.3%	3.1	0.9%
0.15	31.8%	3.7	0.4%

Table 6: Effect of noise parameter σ on persona variation. Bucket flip rate = % of traits that cross a bucket boundary after noise. Duplicate % = samples with gestalt similarity ≥ 0.85 on opening turns. $\sigma = 0.08$ balances variety and in-persona coherence.

Table 6 shows the sensitivity to σ . At $\sigma = 0.05$, 12.4% of opening turns are near-duplicates; at $\sigma = 0.15$, the bucket flip rate hits 31.8%, meaning nearly a third of traits land in a different bucket than the profile intended,

producing out-of-character behavior. The shipped default $\sigma = 0.08$ sits at the knee: 14.7% bucket flip rate (enough variety) and 3.1% duplicate rate (acceptable).

5 Combinatorial Analysis

The persona model generates a large space of distinct prompt configurations:

Tier 1: $\prod_{k=1}^6 |C_k| = 6 \times 6 \times 5 \times 4 \times 4 \times 4 = 11,520$ demographic combinations.

Tier 2: Each of 12 traits maps to 3 buckets = $3^{12} = 531,441$ behavioral configurations.

Tier 3: Each of 5 states maps to 3 buckets = $3^5 = 243$ emotional configurations.

Complexity: 4 levels.

Total prompt-distinct configurations: $11,520 \times 531,441 \times 243 \times 4 \approx 5.95 \times 10^{12}$. In practice, profile constraints and correlation rules reduce the reachable space, but the design supports generating millions of unique persona configurations without repetition.

6 Limitations and Future Work

No learned covariance. Trait correlations are rule-described, not learned. Fitting a covariance matrix from the 64K corpus would yield more realistic co-occurrence without manual authoring, but at the cost of interpretability.

No turn-level emotional dynamics. Tier-3 states are session-level: set once at the start and held constant. Real users’ frustration escalates mid-conversation when the agent fails. Modeling intra-session emotional trajectories is future work.

Bucketing information loss. The three-bucket discretization (low/medium/high) loses information. A user at 0.36 (just above the low threshold) and one at 0.69 (just below high) both receive the “medium” prompt guidance. Finer bucketing (5 or 7 levels) or direct continuous injection are under investigation.

Cultural validity. The prompt-guidance strings are English-centric and Western-normed. Jurisdiction=IN adds payment-method context but does not adjust interaction norms. Cross-cultural persona validation is needed.

Judge attribution. We measure persona impact on *agent* scores (goal achievement, compliance), but the causal path runs through the user simulator. Disentangling persona-driven difficulty from simulator artifacts requires human evaluation of user-turn realism, which we have not yet conducted.

7 Conclusion

We presented a three-tier persona vector with 23 operationalized dimensions for controllable user simulation in agentic evaluation. The tiered structure separates identity (demographics), behavior (traits with noise), and emotional context (scenario-reactive states), enabling factorial experimental designs and measurable behavioral attribution. Across 64,698 evaluated conversations, the model produces a 15.8 percentage points spread in goal achievement, faithful difficulty distributions on domain-specific axes, and orthogonal interaction with query complexity. The specification is complete: every distribution, noise parameter, bucketing rule, and correlation threshold is provided for reproduction.

The broader implication is that user simulation for agentic evaluation does not need to be either flat descriptions or learned black boxes. A structured middle ground of continuous, measurable, auditable properties gives practitioners the control they need to stress-test agents on the user populations that matter.

References

- [1] R. Khedar, Eshita, S. T. S. R. Thondapu, M. Malhotra, A. Das, J. Chandra, Y.-S. Chuang, C. Kulkarini, A. Menon, L. Pang, A. Karn, Mouli V, and P. Mehrotra. State-grounded multi-agent synthetic data generation for tool-augmented LLMs. *arXiv:2606.16307*, 2026.
- [2] S. Yao, et al. τ -bench: A benchmark for tool-agent-user interaction in real-world domains. *arXiv:2406.12045*, 2024.
- [3] L. Zheng, W.-L. Chiang, Y. Sheng, et al. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *NeurIPS*, 2023.
- [4] J. Schatzmann, K. Weilhammer, M. Stuttle, and S. Young. A survey of statistical user simulation techniques for reinforcement-learning of dialogue management strategies. *The Knowledge Engineering Review*, 21(2):97–126, 2006.
- [5] J. Schatzmann, B. Thomson, K. Weilhammer, H. Ye, and S. Young. Agenda-based user simulation for bootstrapping a POMDP dialogue system. In *HLT-NAACL*, 2007.
- [6] F. Kreyssig, I. Casanueva, P. Budzianowski, and M. Gašić. Neural user simulation for corpus-based policy optimisation of spoken dialogue systems. In *SIGDIAL*, 2018.
- [7] S. Zhang, E. Dinan, J. Urbanek, A. Szlam, D. Kiela, and J. Weston. Personalizing dialogue agents: I have a dog, do you have pets too? In *ACL*, 2018.
- [8] H. Song, Y. Wang, K. Zhang, W. Zhang, and T. Liu. BoB: BERT over BERT for training persona-based dialogue models from limited personalized data. In *ACL*, 2021.
- [9] X. Xu, et al. Beyond goldfish memory: Long-term open-domain conversation. In *ACL*, 2022.
- [10] X. Liu, H. Yu, et al. AgentBench: Evaluating LLMs as agents. *arXiv:2308.03688*, 2023.
- [11] K. Mei, et al. Matrix: An agent-based multi-turn data synthesis framework. Meta FAIR, *arXiv preprint*, 2024.
- [12] N. S. Keskar, B. McCann, L. Varshney, C. Xiong, and R. Socher. CTRL: A conditional transformer language model for controllable generation. *arXiv:1909.05858*, 2019.
- [13] B. Krause, A. D. Gotmare, B. McCann, N. S. Keskar, S. Joty, R. Socher, and N. F. Rajani. GeDi: Generative discriminator guided sequence generation. In *EMNLP Findings*, 2021.