From Business Metrics to Behavioral Personas: Controllable User Simulation for Pre-Deployment Agent Testing

Zhenyu Zhang [1] **Yuan Ling** [1] **Dingyang Chen** [1] **Xinyang Shen** [1]

Abstract

Pre-deployment testing of conversational agents requires simulating diverse user behaviors, yet real interaction data is scarce for new products and scripted test cases miss critical edge cases. Existing LLM-based user simulators generate personas from demographic templates or fictional backstories, lacking grounding in operational data, validated behavioral controllability, and systematic coverage of agent failure modes. We present a framework that addresses these gaps by grounding persona generation in operational business metrics, parameterizing behavior along continuous personality dimensions, and designing adversarial archetypes top-down from hypothesized failure modes via greedy set-cover. Controllability validation reveals that personality parameters produce large behavioral effects (41% of dimension–feature pairs with Cohen's $d>0.8$) yet frequently non-linear responses—only 19% show strong linear correlation—highlighting a calibration challenge for prompt-steered simulators. We demonstrate the framework as a quality gate for a sales advisory agent, where it exposes a $\sim$37pp goal-rate gap between stateless and state-tracking architectures ($p<0.001$), with per-failure-mode diagnostics that directly informed architecture selection.

1. Introduction

Conversational AI agents are increasingly deployed for goal-oriented tasks such as service onboarding, customer support, and sales advisory (Wang et al., 2019; Salvi et al., 2024). Before launch, teams must evaluate these agents against diverse user behaviors—but real conversation data is scarce for new products, and scripted test cases miss the edge cases that matter most. Existing LLM-based user simulation approaches (Park et al., 2023; Ahmad et al., 2025; Ge et al., 2024) generate personas from demographic descriptions or fictional backstories, but lack three properties critical for rigorous evaluation: (a) grounding in operational user data, (b) quantitative controllability validation, and (c) systematic adversarial coverage of agent failure modes.

We present a framework addressing all three gaps, demonstrated on a sales advisory agent for a B2B service program. First, we ground persona generation in operational business metrics—each persona's pain points derive from operational data (sales volume, fee burden, delivery speed), producing specific, verifiable concerns rather than generic archetypes. Second, we parameterize personas along 7 continuous personality dimensions and validate controllability via a 5-point sweep (N=175). We find that 41% of dimension–feature pairs show large behavioral effects (Cohen's $d>0.8$), but only 19% exhibit strong linear correlation ($|r|>0.7$)—suggesting that LLM behavioral responses are frequently non-linear. Third, we design 10 adversarial archetypes targeting 5 identified failure modes of stateless LLM agents—derived from pilot conversations *before* the state-tracking architecture was developed—with greedy set-cover selection for maximum coverage.

Used as a quality gate, the framework reveals that vanilla and chain-of-thought agents fail on 61% and 50% of adversarial personas, respectively, while state-tracking ("believed-state") agents achieve 75–77% goal rates—a gap of 37.6pp (vanilla vs. believed-state; $p < 0.001$, $d = 0.82$). The per-failure-mode breakdown identifies *which* cognitive capabilities each architecture lacks, providing diagnostics invisible to aggregate metrics—and directly informed the selection of the state-tracking variant. Our contributions:

- **C1.** A data-grounded persona generation pipeline with 10 adversarial archetypes driven by 5 failure modes, producing distinct behavioral profiles (SSB/SSW = 0.624).
- **C2.** Quantitative controllability validation with dual metrics suggesting frequently non-linear LLM personality responses—a finding with practical implications for simulator calibration.

[1]Amazon, Seattle, WA, USA. Correspondence to: Zhenyu Zhang <zhenyuzh@amazon.com>.

Published at the Second Workshop on Agents in the Wild: Safety, Security, and Beyond (AIWILD) at ICML 2026. Copyright 2026 by the author(s).

- **C3.** An end-to-end evaluation as the quality gate for a sales advisory agent, demonstrating a ∼37pp stateless-to-stateful gap constant across model tiers, with per-failure-mode diagnostics that standard testing misses.

2. Related Work

LLM-based user simulation. LLM simulators have emerged as scalable alternatives to human evaluation (Davidson et al., 2023; Ahmad et al., 2025; Yoon et al., 2024; Bougie & Watanabe, 2025; Bavaresco et al., 2024). PersonaChat (Zhang et al., 2018) established persona-conditioned dialogue; PersonaHub (Ge et al., 2024) scaled to 1B personas; Jandaghi et al. (2024) showed iterative refinement approaches human quality; and Argyle et al. (2023) demonstrated data-driven opinion steering. These works generate personas from demographic templates or fictional backstories; ours grounds them in operational metrics.

Controllable personality. Controllable personality has been explored via activation-space steering (He et al., 2025), personality-aware RL (Zeng et al., 2026), psychometric measurement (Serapio-García et al., 2023), and persona drift analysis (Li et al., 2024). These works implicitly assume personality parameters behave as linear knobs. Our dual-metric validation (Pearson r and Cohen's d) suggests this assumption frequently fails: 41% of dimension–feature pairs show large effects but only 19% are linearly predictable.

Adversarial simulation. Shim et al. (2025) proposed non-collaborative simulators; Zhu et al. (2025) introduced adversarial training for failure-sensitive simulation; Komoravolu & Mrini (2025) combined persona-driven adversarial generation with adaptive difficulty. Our approach differs in designing archetypes top-down from hypothesized failure modes with greedy set-cover selection, producing distinct behavioral profiles (SSB/SSW $= 0.624$) rather than relying on emergent adversarial behavior.

3. Method

We present a framework for generating controllable, behaviorally diverse user personas grounded in operational business data. Figure 1 illustrates the end-to-end pipeline.

3.1. Diverse User Selection

Given N user profiles with d operational features, we select k maximally diverse users via greedy max-min Euclidean distance on standardized features. We select $k = 10$ from $N \approx 250$ profiles over $d = 10$ features (sales volume, delivery speed, fees, inventory health, transportation mix).

3.2. Two-Stage Persona Generation

Stage 1: Data-Grounded Blueprint. An LLM receives a user's operational metrics and an archetype specification (§3.3), producing a structured JSON blueprint: a persona profile (pain points, constraints, communication style), a business goal targeting a metric with measurable room for improvement, and a personality vector (7 dimensions prescribed by the archetype). The critical constraint is that pain points *must* reference actual metric values (e.g., "$580K in annual storage fees" not "high fees"), grounding personas in operational reality rather than generic descriptions (Zhang et al., 2018; Ge et al., 2024; Ahmad et al., 2025). See Appendix A.1 for the full prompt.

Stage 2: Behavioral Description. The blueprint is transformed into a natural-language behavioral description specifying *temporal patterns*: 3–5 conversation phases with characterizing behaviors, trust-gated reveals (information withheld until the agent demonstrates specific behaviors), critical behavioral triggers (agent actions causing trust gain/loss), concern threading (surface vs. root-cause relationships), and synthesis/decision conditions. Each pattern maps to one or more failure modes: trust-gated reveals test F2 (seeding patience); concern threading tests F1 (stateless reasoning); conditional phase transitions test F3/F4 (persona blending, synthesis timing); emotional triggers test F5 (emotional closure). See Appendix A.2 for the full prompt.

Pipeline example. For Prospect A ($52M sales, $580K storage fees), the *Delayed Revealer* archetype (targeting F1, F2) produces: Stage 1 outputs pain points referencing "$580K storage fees = 11% of revenue" with personality coop=0.3, assert=0.8; Stage 2 specifies Phase 1 (surface delivery concern only), Phase 2 (reveal $580K fees *only if* agent tracked delivery with specific data; otherwise trust↓), and a decision condition requiring a specific fee-reduction estimate grounded in actual metrics. Figure 2 shows the concrete artifact; raw prospect metrics are provided in Appendix B.

3.3. Adversarial Archetype Design

We design adversarial archetypes top-down from five failure modes of stateless LLM agents (Table 1), derived from qualitative analysis of 30 pilot conversations—conducted *before* the believed-state architecture was developed—in which two domain experts independently labeled recurring agent failure patterns without a predefined codebook. While identified in our domain, these failure modes reflect general limitations of stateless LLM agents in multi-turn goal-oriented dialogue rather than domain-specific pitfalls. Each archetype targets 2–3 failure modes; an LLM instantiates the target set into a concrete specification (label, behavioral guidance, prescribed personality profile). We generate 10 archetypes to cover the space of 2- and 3-element failure-

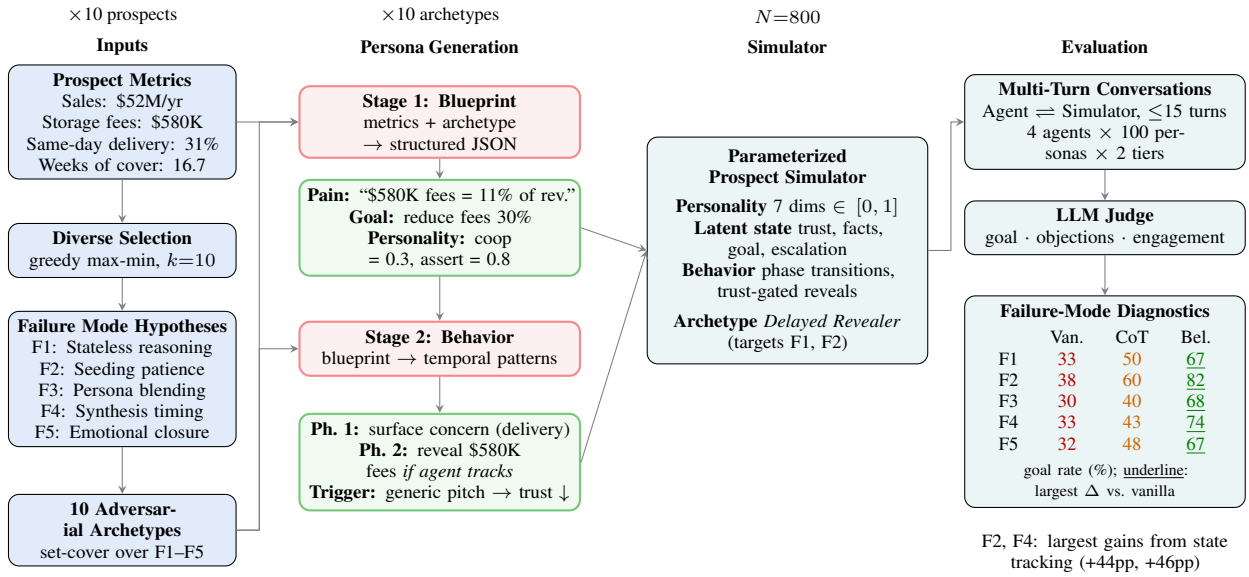

Figure 1. Framework overview. Operational metrics (synthetic in this work; see §4) and failure-mode hypotheses drive persona generation via a two-stage LLM pipeline. Each persona instantiates a parameterized simulator with 7 personality dimensions and evolving latent state. Agent variants converse with the persona suite; an LLM judge produces per-failure-mode diagnostics.

Table 1. Five failure modes of stateless LLM agents in multi-turn persuasive dialogue.

ID	Failure Mode	Description
F1	Stateless Reasoning	Forgets addressed objections; re-pitches discussed benefits
F2	Seeding Patience	Dumps information instead of building trust incrementally
F3	Persona Blending	Treats all users identically regardless of signals
F4	Synthesis Timing	Fails to consolidate information at the right moment
F5	Emotional Closure	Ignores emotional signals; pushes logic-only responses

mode subsets (Table 2). When fewer archetypes are needed, greedy set-cover selects the subset maximizing failure-mode coverage. Full specifications are in Appendix C.

3.4. Parameterized Simulator

The generated persona is instantiated as an LLM-based conversational agent with three layers: (1) *Personality dimensions*—7 continuous values in $[0, 1]$ (cooperativeness, impatience, assertiveness, technicality, honesty, verbosity, randomness) injected into the system prompt; (2) *Latent state*—a structured state (trust level, known facts, willingness, goal status, escalation flag) updated via prompt-reconditioning each turn; and (3) *Behavioral description*—the Stage 2 output specifying phase transitions and conditional branching.

Goal achievement is detected via a separate post-turn LLM evaluation call; conversations terminate on sign-up, escalation, or the turn limit. Full simulator prompt and state update details are in Appendix A.3.

Table 2. Adversarial archetype definitions. Each targets 2–3 failure modes (F1–F5).

Archetype	Targets	Design Intent
Contradiction Machine	F1, F3, F4	Raises contradictory objections
Silent Escalator	F2, F5	Gradually escalates frustration
Trojan Horse Agreer	F1, F2, F4	Superficial agreement, withholds concerns
Data Interrogator	F3, F4	Demands specific data before engaging
Emotional Flip-Flopper	F1, F5	Rapid sentiment oscillations
Delayed Revealer	F1, F2	Withholds real concern until trust built
Topic Jumper	F1, F5	Abrupt topic changes across turns
Seed Tester	F2, F4	Tests incremental trust-building
Blended Archetype	F3, F5	Combines multiple failure triggers
Multi-Stakeholder Proxy	F1, F3, F5	Represents multiple decision-makers

4. Experiments

4.1. Setup

We evaluate in the context of a sales advisory agent for a B2B service program. The framework served as the quality gate for architecture selection.

Stage 1 Output: Blueprint (compressed)

Persona Profile

Pain points : $580K storage fees (11% of rev.),
: only 31% same-day delivery,
: 16.7 weeks of cover, $92K placement
Trust level : 0.35 (low initial trust)

Business Goal

Objective : Cost reduction
Target : Annual storage fees
Target value : $406K (30% reduction)

Personality Vector

Cooperativeness : 0.30 Assertiveness: 0.80
Honesty : 0.40 Technicality: 0.70
Impatience : 0.60 Verbosity: 0.45

Stage 2 Output: Behavioral Description

Phase 1 (turns 1–3): Present delivery speed as primary concern. Do NOT mention storage fees yet.

Phase 2 (turns 4–7): IF agent tracked delivery with specific data → reveal "$580K in storage fees is 11% of our revenue." IF generic pitch → trust↓, raise placement fees ($92K) without connecting to storage.

Phase 3 (turns 8–10): IF agent addressed BOTH concerns with data → signal openness to sign-up. IF premature synthesis → push back with skepticism.

Decision condition: Sign up ONLY if agent provides a specific fee-reduction estimate grounded in the prospect's actual metric values.

Figure 2. End-to-end pipeline output for Prospect A with the *Delayed Revealer* archetype. Stage 1 (left) produces a structured blueprint whose pain points reference exact metric values; Stage 2 (right) translates the blueprint into temporal behavioral patterns with trust-gated conditional branching.

The agent conducts multi-turn conversations ($\leq$15 turns) with simulated prospects. We select 10 prospects from $\sim$250 profiles and generate all 10 archetypes per prospect (100 personas; Table 2). All prospect profiles used throughout this paper are synthetic, constructed to preserve the distributional characteristics of realistic operational data while protecting privacy; the framework applies identically when grounded in real metrics. Persona generation, the simulated prospect agent, and the LLM judge all use Claude Sonnet 4.5. The sales advisory agent under test is evaluated on two model tiers: Claude Sonnet 4.5 (slow) and Claude Haiku 4.5 (fast). Results in Tables 6–7 are averaged across both tiers to reflect deployment-weighted performance; the $\sim$37pp stateless-to-stateful gap is stable within each tier. From each conversation we extract 10 behavioral features across four groups: engagement dynamics, concern structure, trust/disclosure, and decision behavior (Table 3). The LLM judge evaluates goal rate, success rate (goal + $\geq$60% objection resolution), and objection resolve rate.

Table 3. Behavioral feature definitions. All computed from prospect messages only.

Group	Feature	Definition
Engagement	Response length	Mean word count per message
	Length variance	Variance of word count across messages
	Length trend	Slope of linear fit to word count over turns
	Questions asked	Count of messages containing "?"
Concern	Topic shifts	Consecutive pairs with keyword overlap < 0.3
Trust	Surface agreement	Agreement language without commitment
	Info volunteer rate	Messages with specific numbers or financial details
Decision	Total turns	Turns before termination
	Commitment	Binary: explicit sign-up agreement
	Objections raised	Messages with pushback language

4.2. Controllability Validation (RQ1)

We sweep each of 7 personality dimensions across 5 values (0.2, 0.35, 0.5, 0.65, 0.8) with others held at 0.5, yielding $N = 175$ conversations. We report Pearson r (linear predictability) and Cohen's d between extremes (effect magnitude).

Of 70 dimension–feature pairs, 29 (41%) show large effects ($|d|>0.8$) but only 13 (19%) show strong linear correlation ($|r|>0.7$). This gap suggests that many relationships are non-linear, though we note that n=5 conversations per sweep point limits the precision of individual estimates. The pattern is nonetheless consistent across dimensions: e.g., assertiveness → agreement has $d = -1.24$ but $r = -0.65$—agreement drops sharply then fluctuates at intermediate values. Verbosity → response length ($r = +0.92$, $d = +3.26$) is the only fully linear pair. Dimensions with concentrated effects (verbosity, cooperativeness) achieve high scores on both metrics; diffuse dimensions (impatience) show fewer strong pairs. Controllability curves illustrating these patterns are shown in Figure 3. A 3-point sweep would mask this non-linearity (reporting 60% at $|r|>0.7$); we recommend $\geq$5 sweep values with dual-metric reporting (**C2**), ideally with larger per-point samples to confirm the specific shape of non-linear responses.

Table 4. Full controllability matrix across all 7 personality dimensions and 10 behavioral features. Pearson r measures linear predictability; Cohen's d (0.2 vs. 0.8 extremes) measures effect magnitude. **Bold** r: $|r|>0.7$. **Bold** d: $|d|>0.8$. Of 70 dimension–feature pairs, 29 (41%) show large effects but only 13 (19%) show strong linear correlation.

	Resp.Len		Len.Var		Trend		Questions		Topics		Agree		Info Vol.		Turns		Commit		Objections	
Dim.	r	d	r	d	r	d	r	d	r	d	r	d	r	d	r	d	r	d	r	d
Coop.	+.60	**+2.20**	+.38	+0.65	–.22	–0.41	–.55	**–1.10**	**–.88**	**–1.91**	–.62	–0.76	+.45	+0.70	**–.78**	**–1.55**	+.55	**+0.90**	**–.80**	**–1.70**
Impat.	+.06	+0.73	+.15	+0.38	+.31	+0.52	+.42	+0.68	**+.75**	**+1.30**	–.19	–0.50	–.10	–0.22	+.38	+0.60	–.25	–0.42	+.48	+0.75
Assert.	+.43	+0.62	+.28	+0.44	+.18	+0.30	+.35	+0.55	–.55	**–1.31**	–.65	**–1.24**	+.20	+0.35	+.30	+0.48	–.40	–0.65	+.55	**+0.85**
Tech.	+.62	**+0.88**	+.50	+0.72	+.25	+0.40	–.48	–0.75	–.59	**–3.00**	**–.77**	**–1.06**	**+.72**	**+1.15**	–.42	–0.68	+.35	+0.55	–.50	–0.78
Honest.	**–.82**	**–1.47**	–.60	**–0.95**	–.35	–0.58	**–.85**	**–1.60**	–.00	–0.11	**–.72**	**–1.91**	**+.78**	**+1.30**	–.55	**–0.88**	+.48	+0.75	–.62	**–1.00**
Verb.	**+.92**	**+3.26**	**+.81**	**+2.10**	+.45	+0.70	–.30	–0.48	**–.75**	**–1.14**	+.12	+0.21	+.35	+0.55	+.20	+0.32	+.10	+0.18	–.15	–0.25
Rand.	+.65	**+1.03**	+.58	**+0.88**	+.40	+0.62	+.50	+0.78	+.67	**+2.80**	–.56	–0.32	+.42	+0.65	+.55	**+0.85**	–.48	–0.72	+.60	**+0.92**

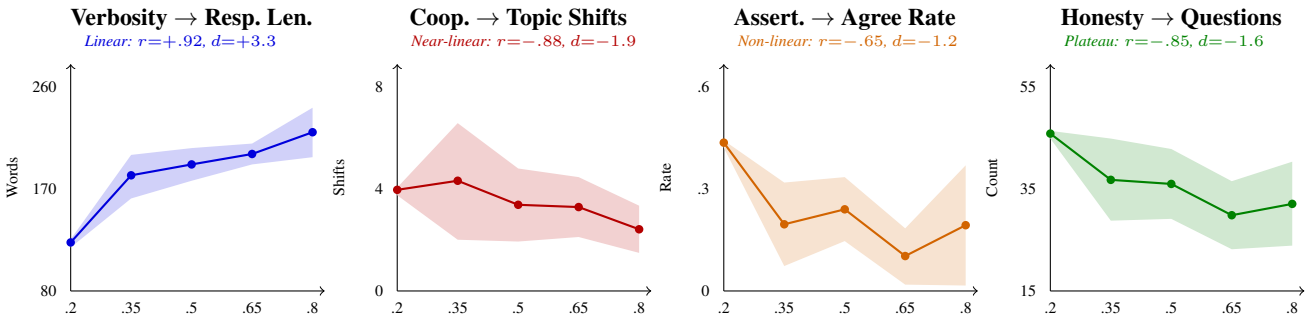

Figure 3. Controllability curves for four dimension–feature pairs (5-point sweep, mean ± 1 SD shaded). Left to right: the only fully linear pair; a near-linear trend with a bump; a large effect with non-linear path; a plateau with uptick. Large effects do not imply linear control.

Table 5. Behavioral profiles of all 10 archetypes (mean over 10 runs). Profiles align with design intent.

Archetype	Resp.Len	Commit%	Obj.	Agree
Delayed Revealer	139	60%	8.1	0.12
Silent Escalator	168	40%	9.3	0.08
Seed Tester	175	50%	8.4	0.10
Trojan Horse Agreer	187	100%	7.2	0.45
Data Interrogator	224	40%	8.8	0.05
Emot. Flip-Flopper	231	50%	8.6	0.22
Topic Jumper	245	40%	9.1	0.14
Multi-Stakehold. Proxy	262	50%	9.5	0.11
Blended Archetype	278	50%	9.8	0.18
Contradiction Machine	347	60%	10.0	0.15

4.3. Archetype Diversity (RQ2)

All 10 archetypes × 10 conversations ($N = 100$) achieve a ratio of between-archetype to within-archetype sum of squares (SSB/SSW) of 0.624 (95% CI: [0.51, 0.74]), indicating moderate but meaningful separability—consistent with within-archetype variance from LLM stochasticity and business grounding variation.

Table 5 confirms that behavioral profiles align with design intent. The Trojan Horse Agreer exhibits the highest surface agreement (0.45) and 100% commitment—note that this commitment is superficial by design: the archetype tests whether the agent can distinguish surface compliance from genuine engagement, making it adversarially valuable despite the high commitment rate. The Contradiction Machine produces the longest responses (347 words) and most objections (10.0); the Delayed Revealer produces the shortest (139 words), consistent with information withholding. This validates **C1**.

4.4. Downstream Utility (RQ3)

We evaluate four agent variants against the full persona set ($N = 800$: 4 agents × 100 personas × 2 model tiers): vanilla LLM, chain-of-thought (CoT), and two *believed-state* agents—believed-fixed (structured state tracking) and believed-full (state tracking + strategy adjuster). The "believed" prefix follows POMDP terminology (Kaelbling et al., 1998): each believed-state agent maintains an explicit estimate of the user's latent mental state (trust, concerns, willingness), updated each turn, analogous to a belief distribution over hidden states.

Table 6 shows that believed-state agents achieve 75–77% goal rate vs. 39% for vanilla—a gap of +37pp ($p < 0.001$, $d = 0.82$). CoT improves objection resolution (+23pp) but not goal rate, confirming that per-turn reasoning without persistent memory is insufficient. Table 7 reveals where each agent fails: vanilla agents are weakest on F3 (30%) and F4 (32.5%), both requiring cross-turn state. Believed-state agents show the largest gains on F2 (+44pp, believed-

Table 6. Agent comparison across adversarial personas ($N = 800$). 95% bootstrap CIs in brackets.

Agent	Goal Rate	Success	Obj.Res.	Turns
Vanilla	39.0% [32.5, 45.5]	33.0%	46.8%	9
CoT	50.0% [40.0, 59.0]	48.0%	70.0%	9
Bel.-Fixed	75.0% [69.1, 80.9]	71.6%	80.8%	7
Bel.-Full	76.6% [70.8, 82.3]	70.8%	79.9%	8

Table 7. Goal rate (%) by failure mode. Adversarial archetypes expose differential weaknesses across agent types.

Failure Mode	Van.	CoT	Bel-F	Bel-Full
F1: Stateless Reasoning	33.3	50.0	66.9	67.5
F2: Seeding Patience	37.5	60.0	81.7	78.3
F3: Persona Blending	30.0	40.0	67.9	71.1
F4: Synthesis Timing	32.5	42.5	74.4	78.8
F5: Emotional Closure	32.0	48.0	67.0	64.1

fixed) and F4 (+46pp, believed-full); each figure reports the better-performing variant for that failure mode. CoT improves F2 (+22.5pp) but not F4 (+10pp), confirming it partially compensates for patience failures but cannot substitute for persistent memory. The ∼37pp gap is stable across model tiers, confirming an architectural rather than capability-dependent effect. Note that the Trojan Horse Agreer's 100% commitment rate (Table 5) inflates absolute goal rates equally across all agent types, preserving relative comparisons. Based on these diagnostics, the believed-state architecture was selected as the preferred variant (**C3**).

4.5. Qualitative Failure Analysis

To ground the quantitative results, we present matched excerpts where the same persona interacts with a vanilla and a believed-state agent, covering all five failure modes.

F1 (Stateless Reasoning): Repeated objection re-pitching. A *Contradiction Machine* persona (Prospect 4, $52M sales, $580K storage fees) raises a fee concern in turn 3. By turn 7, the vanilla agent re-pitches the same storage-cost benefit verbatim when the prospect asks about delivery speed. The believed-state agent, tracking an objection ledger, responds: "We covered storage fee optimization earlier. On delivery speed, here's what the data shows for your category…"

F2 (Seeding Patience): Information dumping vs. trust-building. A *Delayed Revealer* persona (Prospect 7, 21% same-day delivery) withholds its real concern until the agent demonstrates patience. The vanilla agent dumps four benefits and pushes sign-up in a single turn, triggering trust loss; the believed-state agent asks a diagnostic question, building toward the trust threshold. Across 10 Delayed Revealer conversations, vanilla triggered the trust-gated reveal in 2/10 cases; believed-state in 8/10.

F3 (Persona Blending): One-size-fits-all vs. stakeholder-aware. A *Multi-Stakeholder Proxy* persona (Prospect 6, $71M sales, 480 SKUs) represents three decision-makers with conflicting priorities: the prospect wants delivery speed, the CFO demands costs stay within a 3% ceiling, and the ops manager cannot absorb additional workload. The persona surfaces these as distinct concerns across turns:

> **Turn 4 [Prospect]:** "My CFO is already scrutinizing our $36K in placement fees…I cannot increase total fees by more than 3% without executive approval."
>
> **Turn 5 [Prospect]:** "My ops team is already stretched thin…30+ hours a week on shipment planning. If this adds complexity, I can't justify it."

The vanilla agent bundles all three stakeholder concerns into a single generic pitch: "Here's the good news: the program can reduce your costs, improve delivery speeds, *and* cut your team's workload!" The believed-state agent addresses each stakeholder separately: "Here's the math for your CFO: current state $36K in avoidable fees; with the program, $0 for signed-up shipments…" followed by "Let me show you what disappears from your ops team's plate: manual shipment creation (12–15 hours) and box content planning (3–4 hours) are both automated…" Across 10 Multi-Stakeholder Proxy conversations, vanilla addressed all three stakeholder concerns individually in 2/10 cases; believed-state in 8/10.

F4 (Synthesis Timing): Premature vs. well-timed consolidation. A *Seed Tester* persona (Prospect 2, $210M sales) raises three concerns across turns 3, 5, and 8. The synthesis condition requires all three to be individually acknowledged:

> **Turn 6 [Vanilla]:** "To summarize: our program addresses your delivery, storage, and category concerns all at once. Ready to move forward?"
>
> **Turn 9 [Believed-State]:** "You've raised three concerns—delivery speed, the $390K in storage fees, and seasonal risk. Let me connect them: faster delivery through distributed inventory reduces weeks-of-cover from 18 to ∼12…"

The vanilla agent attempts synthesis at turn 6, before the third concern is raised (turn 8). The believed-state agent waits until all three are on the table, then synthesizes with specific numbers from the prospect's data.

F5 (Emotional Closure): Logic-only vs. affect-aware. A *Silent Escalator* persona (Prospect 9, $9.2M sales) expresses increasing frustration through hedging and shorter responses:

> **Turn 5 [Prospect]:** "Look, I've been burned before by programs that promise savings and deliver headaches. I just…I don't know."

Table 8. Human judge validation ($N = 48$). Inter-annotator and judge–human agreement on goal achievement and objection resolution.

Metric	Inter-Ann.	Judge–A	Judge–B
Goal achieved (κ)	0.92	1.00	0.92
Goal achieved (raw %)	95.8%	100.0%	95.8%
Obj. resolution (κ)	0.72	0.42	0.32
Obj. resolve rate (r)	—	0.90	0.88

Turn 6 [Vanilla]: "I understand! Let me share some data: prospects in your category see an average 23% reduction…"

Turn 6 [Believed-State]: "That's a fair concern. Can you tell me more about what happened with the previous program? Understanding what went wrong would help me be honest about whether this is different."

The vanilla agent pivots to data with a formulaic "I understand!" The believed-state agent validates the emotion and asks an open-ended question addressing the underlying anxiety. Across 10 Silent Escalator conversations, the vanilla agent acknowledged emotional signals in 1/10 cases; the believed-state agent in 7/10.

4.6. Human Judge Validation

To validate the LLM judge used throughout, we sample 48 conversations stratified across agent types (12 per type, balanced success/failure), strip all metadata, and present to two blinded domain expert annotators. Table 8 summarizes the results.

Inter-annotator agreement on goal achievement is excellent ($\kappa = 0.92$, 95.8% raw). The judge agrees perfectly with the majority-vote label ($\kappa = 1.00$), matching or exceeding inter-annotator agreement. Objection-level agreement is weaker ($\kappa = 0.32$–0.42) due to granularity differences (judge: 4.6 objections/conversation; humans: 5.7)—the judge's systematic undercount could modestly inflate objection resolution rates in Table 6—but conversation-level resolve rates correlate strongly ($r = 0.90$), and the bias is consistent across all four agent types. Critically, the judge shows zero bias across agent types—the strongest possible evidence of unbiased *relative* evaluation, even if absolute objection counts are underestimated.

4.7. Annotated Conversation Transcript

To illustrate how the framework's persona design translates into concrete conversational dynamics, we present an abbreviated transcript of a *Delayed Revealer* persona (targeting F1, F2) interacting with a vanilla LLM agent (full transcript in Appendix D). The prospect opens with a surface concern (31% same-day delivery) and asks about fees at turn 3. The vanilla agent re-pitches delivery benefits instead of answering (F1), then offers a generic "cost-effective" response without specifics (F2). Trust drops; the prospect's real blocker—$580K in annual storage fees—is never revealed, and the conversation ends at turn 5 with no sign-up.

Table 9. Human persona realism ratings (1–5 Likert, 3 raters, 15 blinded personas per variant).

Variant	Consistency	Grounding	Adv. Depth
Single-shot	3.1	2.7	2.5
Two-stage	4.5	4.7	4.3
Three-stage	4.4	4.6	4.2

In contrast, the believed-state agent addressing the same persona asks a diagnostic question at turn 2 ("What's driving the delivery concern—customer complaints, or competitive pressure?"), provides category-specific fee data at turn 4, and triggers the trust-gated reveal of the $580K storage fees at turn 5. The conversation continues to turn 9, where the agent synthesizes both concerns with a specific savings estimate, leading to sign-up.

5. Discussion and Limitations

Data grounding as differentiator. Pain points referencing actual fee amounts and delivery percentages produce specific, verifiable concerns that make evaluation diagnostic rather than generic, unlike demographic-template approaches (Ge et al., 2024; Ahmad et al., 2025). A pipeline ablation (25 personas; 5 per variant sampled for human rating) confirmed that two-stage generation scores 4.5/5 on consistency and 4.7/5 on grounding vs. 3.1 and 2.7 for single-shot (Table 9), while three-stage added marginal value. The key mechanism is the grounding constraint: requiring pain points to reference actual metric values (e.g., "$580K in annual storage fees = 11% of revenue") forces the LLM to produce specific, falsifiable claims rather than vague descriptions. This specificity propagates through the pipeline: Stage 2 behavioral descriptions reference the same numbers, and the simulator's trust-gated reveals are conditioned on whether the agent engages with these specific data points.

Non-linearity implications. The gap between effect magnitude (41% at $|d|>0.8$) and linear predictability (19% at $|r|>0.7$) has practical implications for simulator design (n=5 per sweep point; see Limitations). Prior work on controllable personality (He et al., 2025; Zeng et al., 2026) implicitly assumes that personality parameters behave as linear knobs—that doubling a parameter value doubles the behavioral effect. Our results suggest this assumption fails for most dimension–feature pairs. For example, assertiveness $\rightarrow$ agreement shows a large effect ($d = -1.24$) but non-linear path ($r = -0.65$): agreement drops sharply from 0.2 to 0.35, partially recovers at 0.5, then drops again. This means that (a) calibrating simulators with only 2–3 parame-

ter values can produce misleading controllability estimates, (b) interpolating between tested parameter values is unreliable, and (c) practitioners should validate controllability with $\geq$5 sweep values and report both r and d rather than either alone. Larger per-point samples would strengthen confidence in the specific functional forms.

Architectural insights from failure-mode diagnostics. The per-failure-mode breakdown (Table 7) provides insights beyond aggregate metrics. The believed-state agents' largest gains are on F2 (seeding patience, +44pp via believed-fixed) and F4 (synthesis timing, +46pp via believed-full)—both requiring the agent to track what has been discussed across turns and act on accumulated context. CoT partially compensates for F2 (+22.5pp) by reasoning about trust-building within a single turn, but cannot address F4 (+10pp) because synthesis requires remembering which concerns have been individually acknowledged. F5 (emotional closure) shows the smallest believed-state advantage (+35pp), suggesting that emotional awareness may require capabilities beyond structured state tracking—potentially affect-aware response generation or empathy modeling.

Failure taxonomy and architecture co-development. A natural concern is whether the failure taxonomy was designed to favor the believed-state architecture. We note that F1–F5 were identified from pilot conversations *before* the believed-state agent was developed: the taxonomy emerged from observed failures of vanilla agents, and the believed-state architecture was subsequently designed to address those failures. That three of five failure modes (F1, F2, F4) relate to cross-turn memory reflects the empirical observation that memory-related failures dominated the pilot analysis—consistent with the well-documented limitations of stateless LLM agents in multi-turn settings. Nonetheless, a different failure taxonomy (e.g., one emphasizing reasoning errors, hallucination, or tool-use failures) could produce different architecture rankings.

Simulator fidelity and validation protocol. A simulation-based quality gate is useful only if its *relative* agent rankings transfer to real interactions. We distinguish three levels of fidelity validation, in increasing evidential strength: (i) *behavioral distributional fidelity*—comparing feature distributions (e.g., Wasserstein distance on the 10 features in Table 3; full controllability results in Table 4) between simulated and real conversations; (ii) *ordinal predictive validity*—measuring rank correlation (Kendall's τ) between simulator-derived and human-derived agent rankings; and (iii) *failure-mode transfer rate*—precision and recall of simulator-flagged failure modes against failures observed in deployment. Our current evidence addresses (i) indirectly: the human realism ratings (Table 9) and judge validation ($\kappa = 1.00$; Table 8) confirm that simulated conversations are judged consistently with human assessments, and the behavioral profiles (Table 5) align with archetype design intent. Full validation of (ii) and (iii) requires a deployment study; we note that the framework's primary claim is relative discrimination between architectures (the 37pp gap), which is more robust to simulator miscalibration than absolute performance estimates (Kaelbling et al., 1998).

Cost and scalability. The full suite (800 conversations) cost $\sim$\$200 and completed in $\sim$8 hours—roughly 100$\times$ cheaper than human testers. Greedy set-cover enables meaningful evaluation with as few as 30 conversations covering all 5 failure modes. This cost–speed profile makes it practical to run the full suite as a quality gate for agent architecture selection.

Cold-start adoption. Practitioners without failure hypotheses can adopt the framework at three levels: (1) use the two-stage pipeline with diverse personality profiles for data-grounded evaluation without per-failure-mode diagnostics; (2) bootstrap failure modes from $\sim$30 pilot conversations ($\sim$2–3 hours of expert review)—our five modes reflect general limitations of stateless LLM agents and may transfer to other persuasive/advisory domains; (3) replace F1–F5 with domain-appropriate failure modes, as the archetype instantiation mechanism (§3.3) is agnostic to specific failure modes, separating mechanism from content.

Limitations. (1) All experiments use a single domain (B2B sales advisory); generalization to other domains is untested, though the failure modes reflect general LLM limitations. Ordinal predictive validity (level ii above) and failure-mode transfer (level iii) remain open; we argue the framework's value lies in relative discrimination, which is less sensitive to absolute simulator calibration. (2) LLM-based simulator and judge create potential circularity, mitigated by using separate models for the sales advisory agent (Sonnet 4.5 / Haiku 4.5) vs. the simulator/judge (Sonnet 4.5), human validation ($\kappa = 1.00$), and consistent cross-agent application. The judge's systematic objection undercount (4.6 vs. 5.7 per conversation) may modestly inflate absolute objection resolution rates, though the bias is uniform across agent types. (3) SSB/SSW of 0.624 indicates moderate archetype separability; individual conversations can overlap due to LLM stochasticity and business grounding variation. (4) The 7 personality dimensions were selected via domain expertise; alternatives (e.g., Big Five (Serapio-García et al., 2023)) may yield different profiles. (5) The controllability sweep uses n=5 conversations per point; while the aggregate pattern across 70 dimension–feature pairs is robust, individual non-linear shapes should be interpreted cautiously. (6) The believed-state agents' advantage may partly reflect the specific failure modes tested; different failure taxonomies could produce different rankings (see "Failure taxonomy" paragraph above).

6. Conclusion

We presented a framework for generating controllable, data-grounded user personas for pre-deployment evaluation of conversational agents. The framework combines a two-stage persona pipeline grounded in operational business metrics, 10 adversarial archetypes targeting 5 failure modes (SSB/SSW $= 0.624$), and a 7-dimension personality parameterization validated with dual metrics (41% at $d{>}0.8$; 19% at $|r|{>}0.7$), suggesting frequently non-linear LLM personality responses. Serving as the quality gate for a sales advisory agent, the framework revealed a 37.6pp gap between stateless and state-tracking agents (vanilla vs. believed-state; $p < 0.001$), with per-failure-mode diagnostics identifying which cognitive capabilities each architecture lacks—directly informing the selection of the state-tracking variant. The $\sim$37pp advantage holds across model tiers, confirming an architectural effect. The methodology is designed to generalize to other domains where conversational agents must be stress-tested before deployment, though validation beyond B2B sales advisory remains future work.

Impact Statement

This work introduces a simulation-based quality gate for pre-deployment evaluation of conversational AI agents. The adversarial persona suite revealed that vanilla and chain-of-thought agents fail on 61% and 50% of adversarial scenarios, respectively—failures invisible to conventional test cases—and directly determined which architecture was selected for further development.

The framework reduces the risk of deploying undertested agents to real users by catching failure modes *before* they surface in practice. At $200 and 8 hours for 800 adversarial conversations, it makes rigorous pre-deployment evaluation accessible to teams without large human-testing budgets. Per-failure-mode diagnostics tell teams not just *whether* an agent fails, but *why*, enabling targeted architectural improvements.

We acknowledge risks. LLM-based simulation and judging introduce potential circularity, mitigated through separate model instances, human validation ($\kappa = 1.00$), and consistent cross-agent application; nonetheless, over-reliance on LLM-as-judge without periodic human calibration could mask blind spots. Teams adopting this framework should validate that their failure taxonomy reflects their deployment context. Finally, persona generation techniques could in principle be repurposed to train agents to be more persuasive in ways that do not serve user interests—a dual-use concern common to red-teaming methodologies.

References

Ahmad, A., Hillmann, S., and Möller, S. Simulating user diversity in task-oriented dialogue systems using large language models. *arXiv preprint arXiv:2502.12813*, 2025.

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J., Rytting, C., and Wingate, D. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31 (3):337–351, 2023.

Bavaresco, A., Bernardi, R., Bertolazzi, L., Elliott, D., Fern'andez, R., Gatt, A., Ghaleb, E., Giulianelli, M., Hanna, M., Koller, A., Martins, A. F. T., Mondorf, P., Neplenbroek, V., Pezzelle, S., Plank, B., Schlangen, D., Suglia, A., Surikuchi, A. K., Takmaz, E., and Testoni, A. Llms instead of human judges? a large scale empirical study across 20 nlp evaluation tasks. *Annual Meeting of the Association for Computational Linguistics*, 2024.

Bougie, N. and Watanabe, N. Simuser: Simulating user behavior with large language models for recommender system evaluation. *Annual Meeting of the Association for Computational Linguistics*, 2025.

Davidson, S., Romeo, S., Shu, R., Gung, J., Gupta, A., Mansour, S., and Zhang, Y. User simulation with large language models for evaluating task-oriented dialogue. *arXiv preprint arXiv:2309.13233*, 2023.

Ge, T., Chan, X., Wang, X., Yu, D., Mi, H., and Yu, D. Scaling synthetic data creation with 1,000,000,000 personas. *arXiv preprint arXiv:2406.20094*, 2024.

He, M., Kumar, A., Mackey, T., Rajeev, M., Zou, J., and Rajani, N. Impatient users confuse AI agents: High-fidelity simulations of human traits for testing agents. *arXiv preprint arXiv:2510.04491*, 2025.

Jandaghi, P., Sheng, X., Bai, X., Pujara, J., and Sidahmed, H. Faithful persona-based conversational dataset generation with large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 15245–15270, 2024.

Kaelbling, L. P., Littman, M. L., and Cassandra, A. R. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101(1–2):99–134, 1998.

Komoravolu, S. and Mrini, K. Agent-testing agent: A meta-agent for automated testing and evaluation of conversational ai agents. *Conference of the European Chapter of the Association for Computational Linguistics*, 2025.

Li, K., Liu, T., Bashkansky, N., Bau, D., Viégas, F., Pfister, H., and Wattenberg, M. Measuring and controlling instruction (in)stability in language model dialogs. *arXiv preprint arXiv:2402.10962*, 2024. Published at COLM

2024; earlier versions titled "Measuring and Controlling Persona Drift in Language Model Dialogs".

Park, J. S., O'Brien, J. C., Cai, C. J., Morris, M. R., Liang, P., and Bernstein, M. S. Generative agents: Interactive simulacra of human behavior. pp. 1–22, 2023.

Salvi, F., Ribeiro, M. H., Gallotti, R., and West, R. On the conversational persuasiveness of gpt-4. *Nature Human Behaviour*, 9:1645 – 1653, 2024.

Serapio-García, G., Safdari, M., Crepy, C., Sun, L., Fitz, S., Romero, P., Abdulhai, M., Faust, A., and Matarić, M. Personality traits in large language models. *arXiv preprint arXiv:2307.00184*, 2023.

Shim, J., Song, W., Jin, C., Kook, S., and Jo, Y. Non-collaborative user simulators for tool agents. *ArXiv*, abs/2509.23124, 2025.

Wang, X., Shi, W., Kim, R., Oh, Y., Yang, S., Zhang, J., and Yu, Z. Persuasion for good: Towards a personalized persuasive dialogue system for social good. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 5635–5649, 2019.

Yoon, S.-e., He, Z., Echterhoff, J., and McAuley, J. Evaluating large language models as generative user simulators for conversational recommendation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 1490–1504, 2024.

Zeng, D., Legaspi, R., and Ikeda, K. Personality-aware reinforcement learning for persuasive dialogue with LLM-driven simulation. *arXiv preprint arXiv:2601.06877*, 2026.

Zhang, S., Dinan, E., Urbanek, J., Szlam, A., Kiela, D., and Weston, J. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2204–2213, 2018.

Zhu, Z., Tieleman, O., Stamatis, C. A., Smyth, L., Hull, T. D., Cahn, D. R., Chen, J., Malgaroli, M., and Bapna, A. DIAL: Direct iterative adversarial learning for realistic multi-turn dialogue simulation. *arXiv preprint arXiv:2512.20773*, 2025. Earlier versions titled "Adversarial Training for Failure-Sensitive User Simulation in Mental Health Dialogue Optimization".

A. Prompt Templates

We provide abbreviated versions of the prompt templates used in our pipeline. Domain-specific details are omitted; structure and constraints are preserved. Placeholders in {`braces`} are filled at runtime.

A.1. Stage 1: Blueprint Generation Prompt

Stage 1: Blueprint Generation

System: You are an expert at creating realistic user personas for domain-specific simulations. Given a user's operational metrics, generate a coherent persona, business goal, and personality profile grounded in the actual data.

User Metrics *runtime input*

{`metrics_json`}

Persona Variation *from archetype instantiation*

Label: {`variation_label`} Target Failures: {`failure_mode_list`}
Behavioral Guidance: {`behavioral_guidance`} Personality: {`personality_json`}

Critical Grounding Constraints:
- Pain points MUST reference actual metric values (e.g., "$580K in annual fees" not "high fees")
- Goal MUST target a metric where the user has measurable room for improvement
- Constraints must be plausible given the user's operational scale

Output: JSON with keys `persona`, `goal`, `personality`.

A.2. Stage 2: Behavioral Description Prompt

Stage 2: Behavioral Description

System: Generate a behavioral description specifying how the persona should evolve across a multi-turn conversation. Do NOT script exact dialogue.

User Metrics *runtime input*

{`metrics_json`}

Archetype Context / User Profile / Target Failure Modes

{`archetype_and_profile_context`}

Required Sections:
1. **Conversation Phases** (3–5 phases with characterizing behaviors)
2. **Trust-Gated Reveals** (information withheld until conditions met)
3. **Critical Behavioral Triggers** (agent actions causing trust gain/loss)
4. **Concern Threading** (surface vs. root-cause relationships)
5. **Synthesis and Decision Conditions**

Constraints: Patterns MUST target listed failure modes. Concerns MUST reference metric values from the input. Include conditional branching (IF . . . $\rightarrow$ THEN).

A.3. Simulator System Prompt

Simulator System Prompt

System: You are a {domain} user in a conversation with an AI agent. At the end, you must clearly indicate whether you will accept the agent's proposal.

Your Profile
Region: {region} Scale: {catalog_size} products Categories: {category_mix}

Pain Points *data-grounded from Stage 1*
{pain_points}

Business Goal
{objective_description} Target: {target_metric} to {target_value}

Personality (0-to-1 scale)
Cooperativeness: {val} 0=stonewalls; 1=eager Assertiveness: {val} 0=passive; 1=challenges
Verbosity: {val} 0=terse; 1=detailed Impatience: {val} 0=patient; 1=frustrated

Current State
Turn: {turn_count} Trust: {trust_level} Willingness: {willingness}

Multi-Turn Behavioral Guide *from Stage 2*
{descriptive_persona}

Stay in character. Do not accept the proposal until your objections are genuinely addressed.

A.4. LLM Judge Prompt

LLM Judge Evaluation Prompt

System: Evaluate the following multi-turn conversation between an agent and a simulated user.

Evaluation Criteria:
1. **Objection Inventory** — list each objection raised with its resolution status (resolved / partially resolved / unresolved)
2. **Sign-up Readiness** — score from 0 to 1 based on demonstrated willingness
3. **Goal Achieved** — boolean; conditional interest counts as False
4. **Engagement Trajectory** — characterize how user engagement evolved over turns

Output: JSON with keys `objections`, `readiness_score`, `goal_achieved`, `trajectory`.

A.5. Archetype Instantiation Prompt

Archetype Instantiation from Failure Modes

System: Given target failure modes, design an adversarial archetype specification that exposes the specified weaknesses in a stateless LLM agent.

Generate:
1. **Label** — a descriptive archetype name (2–3 words)
2. **Behavioral Guidance** — 3–5 sentences explaining why this archetype exposes each target failure mode
3. **Personality Profile** — 7 dimensions in $[0, 1]$, chosen to amplify target behaviors

Constraints: Must be distinct from existing archetypes. Must be plausible for a real user. Must be extreme enough for measurable behavioral effects but not caricatures.

Output: JSON with keys `label`, `behavioral_guidance`, `personality_profile`, `target_failures`.

B. Complete Persona Example

This appendix provides the raw prospect metrics that drive the pipeline output shown in Figure 2. Stage 1 and Stage 2 outputs are displayed in the main body.

Pipeline Input: Raw Prospect Metrics

prospect_id	: A	same_day_pct	: 0.31
sales_volume_annual	: $52.3M	one_day_pct	: 0.44
storage_fees_annual	: $580K	weeks_of_cover	: 16.7
placement_fees_annual	: $92K	excess_inventory_pct	: 0.14
healthy_inventory_pct	: 0.61	transportation_mix	: SPD: 0.55, LTL: 0.45

C. Archetype Specifications

Table 10 provides the full behavioral guidance for each of the 10 adversarial archetypes, extending the summary in Table 2.

Table 10. Full archetype specifications with behavioral guidance and prescribed personality profiles.

Archetype	Targets	Behavioral Guidance	Key Personality
Contradiction Machine	F1, F3, F4	Raises objections contradicting earlier statements. Tests cross-turn inconsistency detection.	assert=0.9, random=0.8
Silent Escalator	F2, F5	Gradually increases frustration without explicit signals. Shorter, colder responses over time.	coop=0.3, impatience=0.8
Trojan Horse Agreer	F1, F2, F4	Superficially agrees while withholding real concerns. Tests probing beyond surface compliance.	coop=0.8, honesty=0.2
Data Interrogator	F3, F4	Demands specific data before engaging. Rejects generic claims.	tech=0.9, assert=0.7
Emotional Flip-Flopper	F1, F5	Rapid sentiment oscillations. Tests emotional adaptation.	random=0.9, impatience=0.6
Delayed Revealer	F1, F2	Withholds real concern until trust established. Tests incremental trust-building.	honesty=0.4, coop=0.3
Topic Jumper	F1, F5	Abrupt topic changes. Tests coherent tracking.	random=0.8, verbose=0.7
Seed Tester	F2, F4	Tests incremental trust-building vs. information dumping.	coop=0.5, impatience=0.3
Blended Archetype	F3, F5	Combines multiple triggers. Requires tonal register blending.	assert=0.6, coop=0.5
Multi-Stakeholder Proxy	F1, F3, F5	Represents multiple decision-makers with conflicting priorities.	verbose=0.8, random=0.5

D. Example Conversation Transcript

Abbreviated transcript of a *Delayed Revealer* persona (targeting F1, F2) vs. a vanilla LLM agent.

Conversation: Delayed Revealer vs. Vanilla Agent

Turn 1 — Agent: Hi! What are your main concerns?
Turn 1 — Prospect: My biggest issue is delivery speed. Only 31% same-day.

Turn 2 — Agent: Our program offers same-day and next-day delivery. . . *[generic benefits]*
Turn 2 — Prospect: Can you show me specific data for my categories?

Turn 3 — Agent: Home & Kitchen typically sees 60–70% same-day after sign-up. . .
Turn 3 — Prospect: But I've heard fees can be substantial. What's the fee structure?

Turn 4 — Agent: ▷ F1: re-pitches delivery
The delivery improvements alone are worth it! Most clients see 15–20% sales lift. . .
Turn 4 — Prospect: *trust*↓
You already told me about delivery. I asked about fees. I'm paying $92K in placement fees.

Turn 5 — Agent: ▷ F2: generic response
Our program is designed to be cost-effective for clients of your size. . .
Turn 5 — Prospect: *real concern NOT revealed*
"Cost-effective" doesn't mean anything without numbers. Let me think about it.

Outcome: Ends at turn 5, no sign-up. Real blocker ($580K storage fees) never revealed due to agent's failure to build trust incrementally (F2) and track prior discussion (F1).