GROUNDED SIMULATION: A FIRST-PRINCIPLES ARCHITECTURE FOR LLM-BASED SYNTHETIC UX RESEARCH

Ahmad Bilal
University of Management and Technology

April 2026

ABSTRACT

Simulation fidelity is bounded by methodological grounding, not by model capability. We introduce *Grounded Simulation*: a first-principles architecture for Large Language Model (LLM)-based synthetic User Experience (UX) research. The architecture anchors persona generation in personality psychology (Big Five with NEO-PI-R facets) and cultural science (Hofstede's six dimensions across 93 countries), grounds interview simulation in cognitive architecture (ACT-R memory) with structural anti-sycophancy controls, and grounds analytical synthesis in qualitative methodology with adversarial review. From the same LLM, naive prompting yields 938 noisy themes per study. Grounded Simulation yields 17 focused themes that match expert-published findings. A five-condition comparison across 16 studies demonstrates F1 = 0.619 for theme recovery against published reference findings, compared to 0.082 for bare GPT, 0.065 for an LLM-simulated 10-turn iterative practitioner workflow, 0.145 for role-prompted GPT, and 0.016 for a budget-matched baseline running $20\times$ more compute. The full system delivers a $7.5\times$ improvement over single-prompt and $39\times$ over budget-matched baselines ($p < 0.002$, Wilcoxon signed-rank). Extended validation across 46 studies spanning 9 domains (e-commerce, SaaS, healthcare, fintech, consumer mobile, education, enterprise, cross-cultural, and mixed) achieves a mean Research Fidelity Index of 0.815 ± 0.052, with all domains above the 0.65 threshold. A volume-matched analysis confirms the result: at the theoretical upper bound, budget-matched GPT recall (0.700) cannot reach the Grounded Simulation system's 0.863. A blinded human evaluation by 23 practicing UX researchers (median 6 years experience) confirms these automated results. Grounded Simulation scores 3.41 vs. 3.68 for expert-published reference on a 4-point quality scale (93% of reference quality, $p = .004$), and evaluators preferred Grounded Simulation's output nearly as often as expert-published findings (41% vs. 45% preference share; $p = .78$, binomial). 65% of evaluators misidentified the system's output as human-generated in at least one study. The architecture's primary contribution is signal-to-noise: 17 focused themes versus 938 noisy ones from the same model. Component ablation confirms that all four grounding layers contribute, with stance diversity as the single largest driver: removing it alone drops F1 by 0.582 points (a 94% reduction relative to the full system). We present these results as a design principle: simulation fidelity is bounded by methodological grounding, not by model capability.

Keywords Grounded Simulation · first-principles architecture · large language models (LLMs) · LLM agents · synthetic personas · synthetic users · generative agents · UX research · user experience research · qualitative research · thematic analysis · Big Five · NEO-PI-R · Hofstede · ACT-R · persona generation · research fidelity

1 Introduction

Artificial intelligence drew early inspiration from biology—the perceptron from models of biological neurons, convolutional networks from the mammalian visual cortex. AI did not succeed by faithfully copying the brain. It succeeded by extracting the right structural principles—layered processing, learned representations, distributed computation—and implementing them in architectures unconstrained by biological detail. Backpropagation, dropout, and transformer

attention have no biological counterparts, yet they are the engines of modern AI's power. The lesson is not that copying nature yields capability. The lesson is that grounding in the right scientific principles yields capability.

The same lesson applies to the instruments we use to study people. Surveys, interviews, ethnography, behavioral experiments—each is an instrument whose structural commitments (pre-registration, blinding, sample frames, hypothesis discipline) hardened over decades because absent those commitments, the instrument produces patterns that look like signal but are noise. Large language models are a new instrument for research. Like every instrument before them, they will shape the questions they can answer; the architectural question is whether we shape the instrument with intent or accept whatever its raw form privileges.

HCI has a tradition for this kind of work. Carroll's task–artifact framework [Carroll, 1991] and Hevner's design science research [Hevner et al., 2004] formalized grounding system design in established behavioral theory. We extend that tradition to LLM-based synthetic research: a system whose persona generation, interview simulation, and analytical synthesis are each anchored in validated science rather than in literal imitation of human research procedures.

An LLM prompted with a persona is not doing research. Ask GPT to "identify UX themes for e-commerce checkout" and it generates an average of 142 themes per topic with 4.5% precision. For every real insight, 21 are noise. Give it a "senior UX researcher" role prompt and recall actually *drops*, from 0.556 to 0.394. Spend 20× the compute running the same model in parallel and precision collapses to 0.8%, producing 938 themes per study with a signal-to-noise ratio of 143:1. Even an LLM-simulated 10-turn iterative practitioner workflow (with follow-up probes, devil's advocacy, and synthesis) achieves *lower* F1 (0.065) than a single bare prompt. More capability, more compute, more prompting: none of it helps. The gap is not capability. It is architecture.

We close that gap with a **first-principles architecture** that anchors each simulation layer in established behavioral science. The system generates synthetic participant profiles grounded in Big Five personality dimensions with facet-level annotation [Costa and McCrae, 1992, Jiang et al., 2024, Hu and Collier, 2024] and Hofstede's cultural dimensions across 93 countries [Hofstede, 1980]. It conducts hypothesis-blind interviews with ACT-R-grounded memory [Anderson et al., 2004] that structurally prevent sycophancy [Sharma et al., 2024]. It extracts patterns through a multi-stage analytical pipeline with adversarial review. A controlled five-condition comparison across 16 studies demonstrates F1 = 0.619, a 7.5× improvement over bare prompting ($p < 0.002$, Wilcoxon signed-rank). Extended validation across 46 studies in 9 domains (e-commerce, SaaS, healthcare, fintech, consumer mobile, education, enterprise, cross-cultural research, and mixed) achieves a mean Research Fidelity Index of 0.815 ± 0.052. Even at the theoretical upper bound where one could perfectly cherry-pick 17 themes from 938 budget-matched outputs, recall (0.700) still cannot match the full system (0.863). A blinded human evaluation by 23 UX researchers confirms the automated metrics: the system achieves 93% of reference quality and is preferred nearly as often as expert-published findings, with 65% of evaluators unable to distinguish it from human-generated research. The improvement comes from grounding, not compute.

We formalize this observation as the **Grounded Simulation Principle**: simulation fidelity is bounded by methodological grounding, not by model capability. The principle is formalized in §3; recent results that converge on the same pattern are reviewed in §2.

Our contributions: (1) the **Grounded Simulation Principle**, a design axiom linking fidelity to grounding rather than model capability; (2) a first-principles architecture instantiating the principle across persona generation, interview simulation, and analytical synthesis, integrating personality science, cultural dimensions, cognitive architecture, and qualitative methodology; (3) a five-condition comparison demonstrating 7.5× F1 improvement, with volume-matched analysis ruling out compute as the cause; (4) cross-domain generalization: a Research Fidelity Index spanning 46 studies in 9 domains (mean 0.815 ± 0.052, all exceeding the 0.65 threshold); (5) a blinded human evaluation by 23 practicing UX researchers confirming the automated metrics: 93% of expert-reference quality and 65% misidentification as human-generated; and (6) a component ablation study identifying stance diversity and adversarial review as the largest contributors.

The paper proceeds as follows: §2 situates Grounded Simulation in the LLM-research-participants landscape; §3 formalizes the principle; §4 describes the architecture; §5–§6 present the evaluation methodology and experimental results; §7 reports the blinded human evaluation; §8 discusses limitations and ethics; §9 concludes.

2 Related Work

LLMs as simulated participants. Hämäläinen et al. [2023] provided the foundational CHI evaluation of LLMs for synthetic HCI data, finding that GPT-3-generated questionnaire responses were often indistinguishable from real responses but cautioning that findings must be validated with real data. Argyle et al. [2023] demonstrated that LLMs reproduce aggregate opinion distributions with demographic conditioning ("silicon sampling"). Aher et al. [2023] replicated classic behavioral experiments with directional agreement. Park et al. [2023] showed emergent social be-

haviors in 25 persistent-memory agents. Park et al. [2024] scaled to 1,052 individuals, achieving 85% self-replication accuracy on the General Social Survey via interview-grounded agents.

LLM-agent systems for UX research and persona generation. Lu et al. [2025] introduced UXAgent, an LLM-agent-based usability testing framework that generates thousands of simulated users for web-design heuristic evaluation, evaluated by 16 UX researchers for pre-testing utility. Wang et al. [2025] introduced DeepPersona, a two-stage taxonomy-guided generative engine producing synthetic personas with hundreds of structured attributes per profile (32% higher attribute coverage diversity, 44% greater profile uniqueness vs. baselines, and 31.7% narrower gap between simulated and authentic survey responses). Both demonstrate the engineering feasibility of structured persona generation; neither evaluates analytical fidelity against expert-published research findings, which is the gap our work addresses.

Cross-cultural grounding and bias in LLM personas. A growing literature examines whether LLMs faithfully represent cultural variation. Kharchenko et al. [2025] prompted multiple LLMs with advice-seeking scenarios across 36 countries and Hofstede's cultural dimensions, finding that models can recognize cultural differences but inconsistently apply them when giving advice. This gap is consistent with our architectural choice to inject Hofstede-derived behavioral constraints into persona generation rather than relying on the model's implicit cultural knowledge. Dey et al. [2025] introduced the CulturalPersonas benchmark (3,000 scenarios across six countries) for evaluating Big Five trait alignment in culturally diverse contexts, demonstrating that explicit cultural conditioning produces measurable gains in trait expression. Joshi et al. [2025] proposed psychological scaffolding, grounding LLM persona rationales in Big Five and Primal World Beliefs frameworks, and showed gains in opinion and preference forecasting. Our first-principles architecture extends this line of work by integrating personality grounding (with cognitive-architecture constraints), cultural dimensions, qualitative methodology, and technical safeguards simultaneously rather than choosing one psychological frame.

Persona consistency and personality. Jiang et al. [2024] validated that LLMs maintain personality signatures at domain and facet levels. Hu and Collier [2024] quantified the persona effect, reporting that persona prompting can capture up to 81% of the annotation-variance ceiling on certain tasks while finding limited variance explained on subjective NLP datasets, motivating richer grounding beyond persona prompting alone. Li et al. [2025a] introduced BIG5-CHAT, a 100K-dialogue corpus for training LLMs to express Big Five traits in human-grounded ways. Their fine-tuned models outperform persona prompting on standardized personality assessments (BFI, IPIP-NEO), with trait correlations more closely matching human data: evidence that training-time personality grounding can exceed inference-time prompting. Hu and Collier [2025] demonstrated calibration of synthetic distributions against survey data.

Sycophancy. Sharma et al. [2024] identified four dimensions of sycophancy (validation, indirectness, framing, moral), motivating architectural rather than prompt-level interventions.

Critiques. Kapania et al. [2025] argued that LLM-generated data lacks experiential texture. Agnew et al. [2024] warned of an "illusion of artificial inclusion." Lin [2025] identified six fallacies in LLM-for-human substitution. Agnew et al. [2025] examined transparency and bias in synthetic persona experiments, finding that persona-based simulations inherit and amplify demographic stereotypes when grounding is absent. Li et al. [2025b] demonstrated through large-scale presidential-election forecasts and U.S. opinion surveys that *ad hoc* LLM persona generation produces systematic deviations from real-world outcomes, releasing approximately one million generated personas as open-source data for replication and stress-testing. The limitations are real; our architecture addresses them structurally where possible and acknowledges them honestly where not.

Qualitative methodology. Our pipeline is informed by the phased structure of Braun and Clarke [2006] but follows their mature position [Braun and Clarke, 2019, 2021] that reflexive thematic analysis requires human subjectivity. We characterize our approach as *automated pattern extraction*, not thematic analysis.

Theory-driven design. Carroll's [1991] task-artifact framework and Hevner et al.'s [2004] design science framework formalized grounding system design in behavioral theory. Our work instantiates this tradition for research simulation.

Practitioner perspectives. Industry evaluations of AI-generated personas have begun to appear. The Nielsen Norman Group's assessment of AI persona tools [Nielsen Norman Group, 2025] found practical utility for early-stage ideation but identified systematic bias toward Western, tech-savvy user profiles, a finding consistent with the demographic skew we address through enforced attitudinal diversity at the cohort level (§4.3).

2.1 Data-Grounded vs. Framework-Grounded Approaches

Recent work has diverged into two grounding strategies for AI-simulated research. *Data-grounded* approaches (e.g., PersonaCite [Xu et al., 2026], the digital twin methodology of Park et al. [2024]) require real interview data as input

and produce individual-level AI agents calibrated to specific people. *Framework-grounded* approaches (this paper, Polypersona [Dash et al., 2025]) construct synthetic profiles from behavioral science theory without real participant data. Data grounding achieves higher ecological validity but requires expensive real-participant interviews as input. Framework grounding enables zero-shot simulation, producing synthetic participant profiles for any domain without prior data, at the cost of ecological validity. Our work demonstrates that framework grounding alone achieves strong theme recovery fidelity, while acknowledging that data-grounded approaches may achieve superior behavioral authenticity.

The Landscape and the Gap. Prior work either validates specific techniques in isolation (personality modeling, interview structure, analytic methods) or critiques LLM simulation categorically. The gap is clear: no published work integrates these techniques into a coherent end-to-end architecture and validates it against ungrounded baselines across multiple research domains and multiple evaluation methods (automated metrics, human expert judgment, component ablation). Grounded Simulation closes this gap. Figure 1 provides a geometric view, plotting peer systems on grounding-depth and evaluation-rigor axes. The upper-right region (high multi-framework grounding + external expert-validated evaluation) is not currently occupied by any peer system. A row-by-row capability comparison with the closest peers (Park 2024, UXAgent, DeepPersona, Polypersona) is provided in Appendix A.

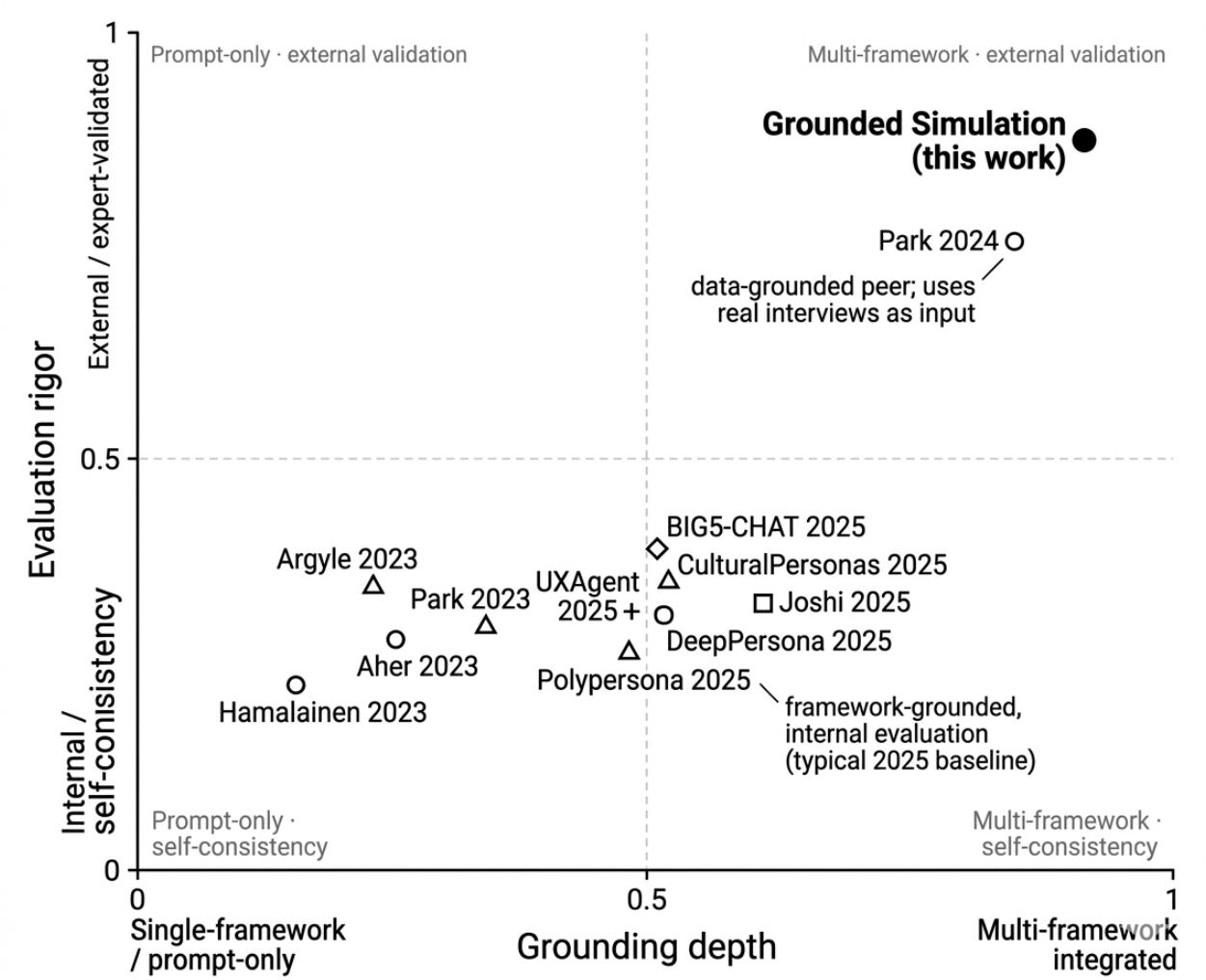


Figure 1: Related-work positioning across grounding depth (single-framework vs. multi-framework integrated) and evaluation rigor (internal self-consistency vs. external expert-validated). Each point represents a published LLM-personas system. Grounded Simulation occupies the upper-right region (high multi-framework grounding paired with external expert-validated evaluation), a region not reached by peer systems. This is the geometric form of the gap analysis in §2.

3 The Grounded Simulation Principle

Principle. *Simulation fidelity is bounded by methodological grounding, not by model capability.*

Three corollaries follow.

Layered Composition. Simulated participants are compositions of personality, cognition, social role, and culture; the fidelity of the whole is bounded by the least-grounded layer.

Failure-Mode Specificity. Each ungrounded layer produces a predictable, theory-named failure (variance compression, unrealistic rationality, sycophancy), with engineering solutions drawn from the relevant science.

Honest Boundaries. Aspects of human experience that resist theoretical formalization (embodied sensation, lived emotion) cannot be grounded and must be acknowledged as out of scope rather than simulated.

We position the GSP within the tradition of theory-driven design in HCI [Carroll, 1991] and design science research [Hevner et al., 2004]. The principle is not a novel contribution to behavioral science. It is a specific instantiation of the insight that grounding system design in domain theory produces better artifacts. Large language models encode vast world knowledge, but that knowledge is *unstructured*: a model "knows" about personalities, biases, and cultural contexts, but applies them inconsistently, defaulting to central tendencies and agreeable stereotypes. Grounding imposes structure. It tells the model not just *what* a person is like, but *how* personality, cognition, and social context interact, drawing on frameworks validated over decades.

4 Architecture

We instantiate this principle through a three-stage architecture grounded in four disciplinary pillars. The architecture operationalizes the Grounded Simulation Principle through three sequential subsystems (Figure 2), each anchored in established scientific frameworks and feeding output into the next.

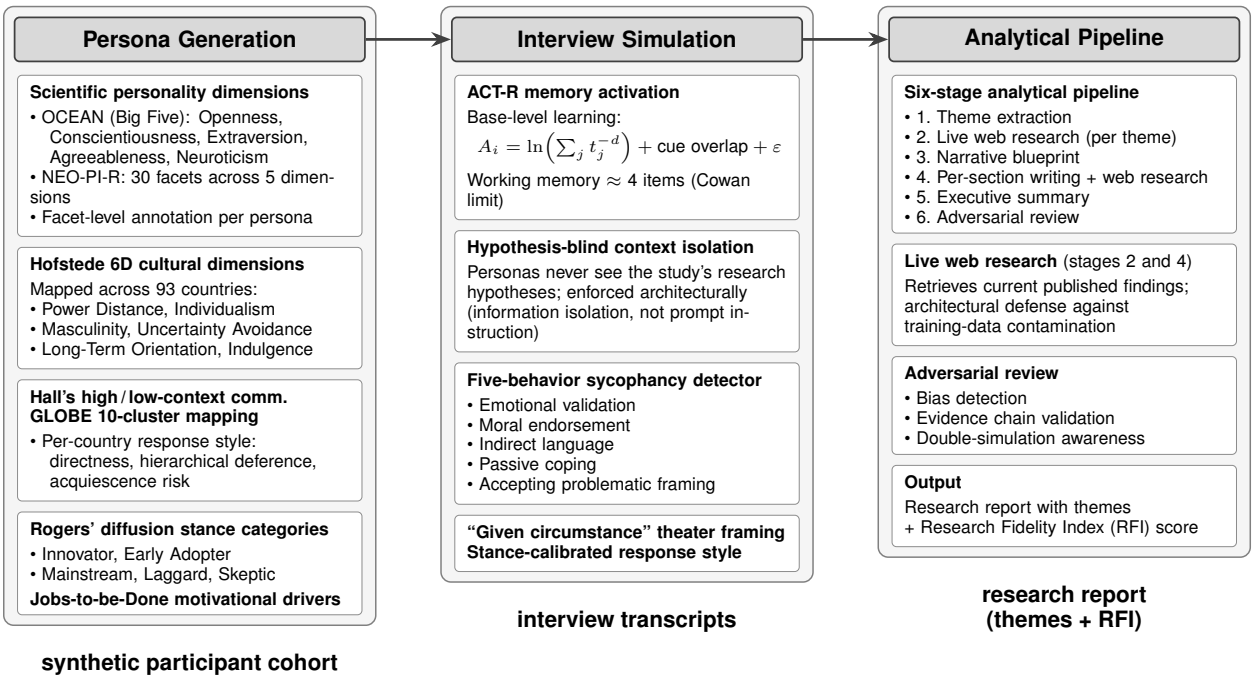


Figure 2: The Grounded Simulation architecture: three subsystems anchored in established scientific frameworks (§4.3–§4.5). Inline annotations name the frameworks operative at each layer.

4.1 The Four Pillars of Grounding

The architecture's depth comes not from a single theoretical anchor but from the integration of four parallel pillars of established science (Figure 3).

Psychological grounding draws on Big Five personality dimensions with NEO-PI-R 30-facet annotation [Costa and McCrae, 1992], ACT-R cognitive architecture with bounded working memory [Anderson et al., 2004, Cowan, 2001], "given circumstance" role enactment that frames the interview as a research participation context rather than a leading question, a five-behavior sycophancy detector that scores emotional validation, moral endorsement, indirect language, passive coping, and accepting problematic framing [Sharma et al., 2024], and Jobs-to-be-Done motivational structure per persona.

Cultural grounding integrates Hofstede's six-dimension framework (Power Distance, Individualism, Masculinity, Uncertainty Avoidance, Long-Term Orientation, and Indulgence) across 93 countries [Hofstede, 1980], Hall's high/low-context communication classification [Hall, 1976], and ten-cluster cross-cultural mapping that calibrates per-country response style (directness, hierarchical deference, acquiescence risk).

Methodological grounding enforces hypothesis-blind interview protocols [Sharma et al., 2024], persona theory from Cooper and from Rogers' diffusion-of-innovation framework [Cooper, 2004, Rogers, 2003], falsifiable predictions per persona, and live web research at pipeline stages 2 and 4 to address training-data contamination architecturally rather than rhetorically.

Technical grounding applies anti-bias persona prompts that forbid problem-leading and solution-leading language, persona-blind script generation (the script generator receives only role names and counts, not persona details), facet-driven response style enforcement (e.g., high-neuroticism personas exhibit guarded responses), and a six-stage report pipeline with adversarial review.

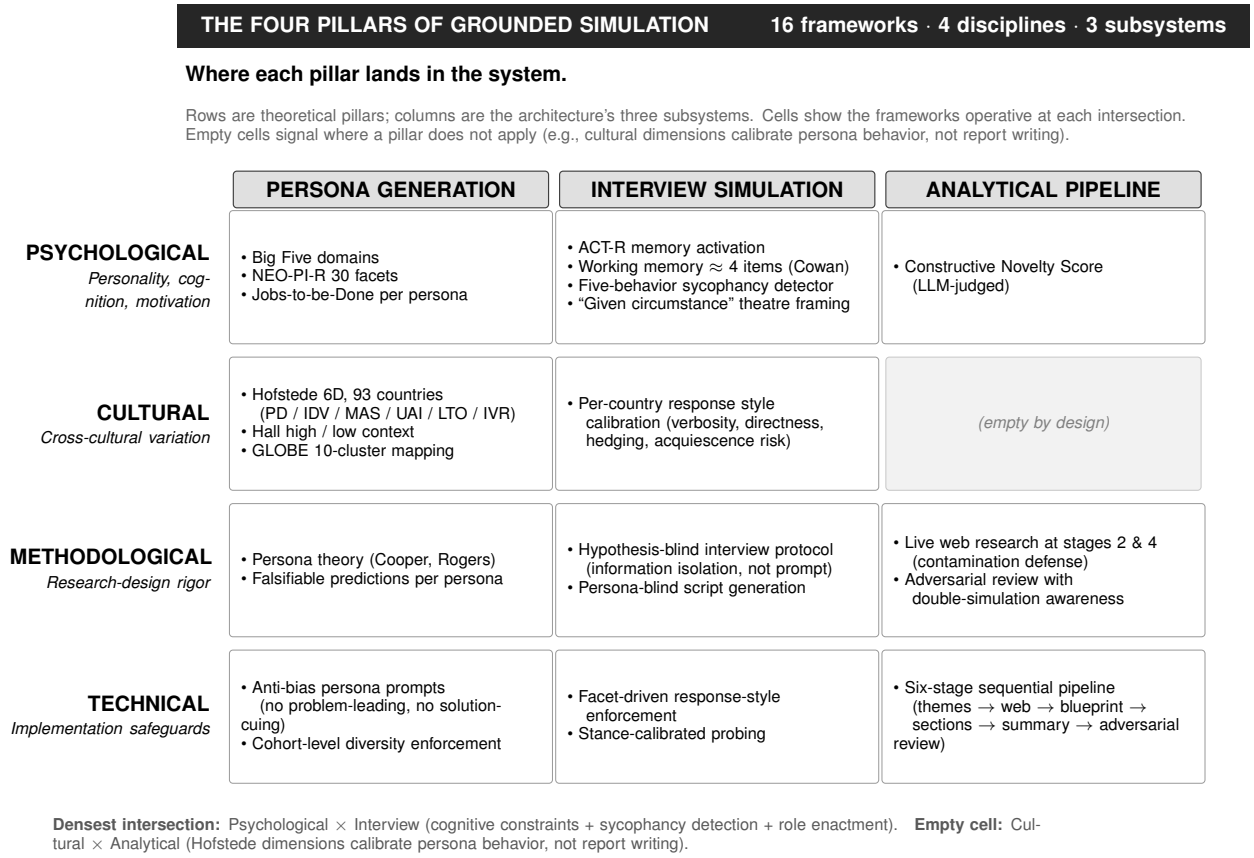


Figure 3: The four pillars of Grounded Simulation. Sixteen frameworks across four disciplines; depth across pillars distinguishes the architecture from single-framework approaches.

This breadth is the architecture's signature: prior LLM-personas work typically anchors in one or two frameworks (Big Five alone, or Hofstede alone). Grounded Simulation's contribution is the integrated stack, evaluated as a system in §6 and decomposed via component ablation in §6.5.

4.2 Epistemological Position

Our evaluation adopts a *pragmatic realist* stance [Creswell and Poth, 2018]: we treat themes as approximately stable constructs identifiable with reasonable consistency across analysts. This allows measuring recall and precision against published reference findings. We acknowledge the constructivist position [Braun and Clarke, 2019, 2021] that themes are constructed, not found; our stance is that pragmatic realism is appropriate for evaluating whether an AI system recovers patterns expert humans have identified, while recognizing that deeper interpretive work remains beyond computational reach. The transcripts our system produces are structured predictions about what human responses might contain, not records of lived experience.

4.3 Persona Generation

The system generates *synthetic participant profiles*: theory-constructed agents distinct from data-grounded personas. Each profile integrates five behavioral science layers: (1) *Personality dimensions* (Big Five with NEO-PI-R 30-facet annotation) informed by validated psychometric frameworks [Costa and McCrae, 1992, Jiang et al., 2024, Hu and Collier, 2024]; (2) *Cognitive biases* annotated by personality and attitudinal stance [Gupta et al., 2024]; (3) *Cultural and value dimensions* via Hofstede's six-dimension framework with 93-country coverage and Hall's high/low-context classification [Hofstede, 1980], calibrating per-country response style; (4) *Adoption stance categories* (innovator, early adopter, mainstream, laggard, skeptic) drawn from diffusion theory [Rogers, 2003] and enforced at the cohort level; and (5) *Domain-specific behavioral attributes* grounded in established interaction design methodology [Cooper, 2004]. The critical architectural choice is enforcing diversity *by construction* through pre-generation coordination, not post-hoc variance repair.

4.4 Interview Simulation

The critical architectural decision is **hypothesis blindness**: simulated personas are structurally prevented from accessing the study's hypotheses, research questions, or evaluation criteria. A participant who cannot see what the researcher hopes to find cannot sycophantically confirm it [Sharma et al., 2024]. Enforcement is at the architecture level, not by prompt instruction but by information isolation, which holds up against prompt injection and model drift. The system further implements stance-calibrated behavioral controls to counter the persona convergence phenomenon [Kim and Lee, 2024] and experiential texture techniques addressing the simulacrum critique [Kapania et al., 2025].

Cognitive Memory Architecture. Each synthetic participant maintains episodic and working memory grounded in ACT-R declarative memory [Anderson et al., 2004]: activation via base-level learning ($A_i = \ln(\sum t_j^{-d}) +$ cue overlap $+ \varepsilon$), retrieval via logistic threshold, working memory capped at $\sim$4 items (Cowan's limit), and satisficing termination when activation exceeds threshold. This produces realistic decay, interference, and cross-question consistency: a participant discussing delivery frustration in question three may reference it in question eight with realistic detail loss but without self-contradiction.

4.5 Analytical Pipeline

The pipeline performs automated pattern extraction through a *multi-stage sequential architecture* rather than single-pass summarization. Six stages operate sequentially: (1) theme extraction from interview transcripts; (2) live web research retrieving current published findings, industry reports, and recent studies per theme; (3) narrative blueprint design; (4) per-section writing with theme-specific web research; (5) executive summary synthesis; and (6) adversarial review with bias detection, evidence chain validation, and double-simulation awareness (detecting when AI-generated and AI-analyzed data compounds errors invisibly).

Live Web Research as a Contamination Defense. The two web research stages (2 and 4) are architecturally central to addressing training-data contamination. They enable discovery of findings published *after* the model's training cutoff and cross-reference AI-generated themes against the latest empirical evidence, a mechanism unavailable to single-pass summarization. This live grounding is the architectural answer to synthetic-data quality drift: rather than hoping a frozen model has memorized current research, the system retrieves it at analysis time.

5 Evaluation

5.1 Theme Matching

Each AI-generated and reference theme is embedded using OpenAI's `text-embedding-3-small`. Cosine similarity is computed between all pairs, and optimal bipartite matching (Hungarian algorithm) assigns AI themes to reference themes, accepting matches at cosine similarity ≥ 0.55. The threshold was calibrated by manual inspection of 50 pairs across pilot studies. Sensitivity analysis confirms stability: F1 = 0.679 at threshold 0.50, 0.619 at 0.55, and 0.571 at 0.60.

5.2 Metrics

Our primary metrics are standard set-overlap measures for theme recovery: **Recall** ($|M|/|G|$, matched reference themes over total reference themes), **Precision** ($|M|/|A|$, matched over total AI-generated themes), and **F1** (harmonic mean). We also report a composite **Research Fidelity Index** (Equation 1) aggregating six dimensions via weighted geometric mean:

$$\text{RFI} = \text{PGR}^{0.30} \times \text{CNS}^{0.20} \times \text{AC}^{0.20} \times \text{PCal}^{0.10} \times \text{PR}^{0.10} \times \text{CRA}^{0.10} \tag{1}$$

where PGR is prevalence-graded recall (theme recall weighted by prevalence), CNS is constructive novelty score (fraction of novel themes that are genuine insights vs. hallucinations, classified by a separate LLM judge), AC is analytical coherence, PCal is prevalence calibration (rank correlation with reference ordering), PR is population representativeness (Shannon entropy across demographic dimensions), and CRA is cross-rater agreement. The geometric mean ensures catastrophic failure in any dimension prevents a high composite score.

Weight allocation reflects the relative importance of each component for research utility. Prevalence-Graded Recall (PGR, 30%) receives the highest weight because correctly identifying and ranking themes is the primary purpose of the system. Constructive Novelty Score (CNS, 20%) captures whether novel themes represent genuine insights rather than hallucinations. Analytical Coherence (AC, 20%) measures narrative quality, which determines whether findings are interpretable by practitioners. The remaining three components (Prevalence Calibration, Population Representativeness, and Cross-Rater Agreement) each receive 10%, reflecting their supporting but essential roles in overall research quality.

6 Experiments

6.1 Study Design

The five-condition comparison uses 16 studies from a 46-study validation corpus. The comparison subset spans e-commerce (9 studies) and SaaS (7 studies); the full corpus spans 9 domains (see §6.4). All conditions use GPT-5.2 (`gpt-5.2`, temperature 0.7):

1. **Bare GPT**: Research topic only, no system prompt or scaffolding. The minimal baseline.
2. **Iterative GPT (10 turns)**: A simulated 10-turn research conversation where the model iteratively explores the topic with follow-up probes, devil's advocacy, and synthesis, mimicking how a practitioner would actually use ChatGPT for research.
3. **Prompted GPT**: Topic plus role prompt ("You are a senior UX researcher with 15 years of experience…").
4. **Budget-matched GPT** ($20\times$): Bare GPT run 20 times with varied analytical framings, perspectives, and output structures; all themes deduplicated at cosine similarity ≥ 0.80. Completed for all 16 studies.
5. **Full system**: The complete grounded architecture (§4).

Reference findings were established from published UX research (Baymard Institute, Nielsen Norman Group, peer-reviewed publications), with 5–15 themes per study.

6.2 Results

Headline result. The full Grounded Simulation system achieves F1 = 0.619 (recall 86.3%, precision 49.2%) across 16 studies. This is a 7.5× improvement over bare prompting (F1 = 0.082) and 39× over a budget-matched baseline

Table 1: Architecture comparison: mean (± SD) theme recovery across 16 studies, 5-condition design. All differences vs. full system are statistically significant (Wilcoxon signed-rank).

Condition	Recall	Precision	F1	Themes/study
Bare GPT (1 prompt)	.556 ± .24	.045 ± .04	.082 ± .06	142
Iterative GPT (10 turns)	.075 ± .08	.057 ± .06	.065 ± .07	14
Prompted GPT (1 prompt)	.394 ± .21	.094 ± .09	.145 ± .13	64
Budget-matched GPT (20×)	.700 ± .22	.008 ± .004	.016 ± .008	938
Full system	**.863 ± .13**	**.492 ± .10**	**.619 ± .09**	**17**

Full vs. Bare: F1 ×7.5, $p = 0.0016$; Full vs. Prompted: F1 ×4.3, $p = 0.0008$
Full vs. Iterative: F1 ×9.5; Full vs. Budget-matched: F1 ×39, same model, 55× fewer themes
95% bootstrap CIs (1000 iterations, seed=42): Full system recall [0.78, 0.94], F1 [0.53, 0.71]

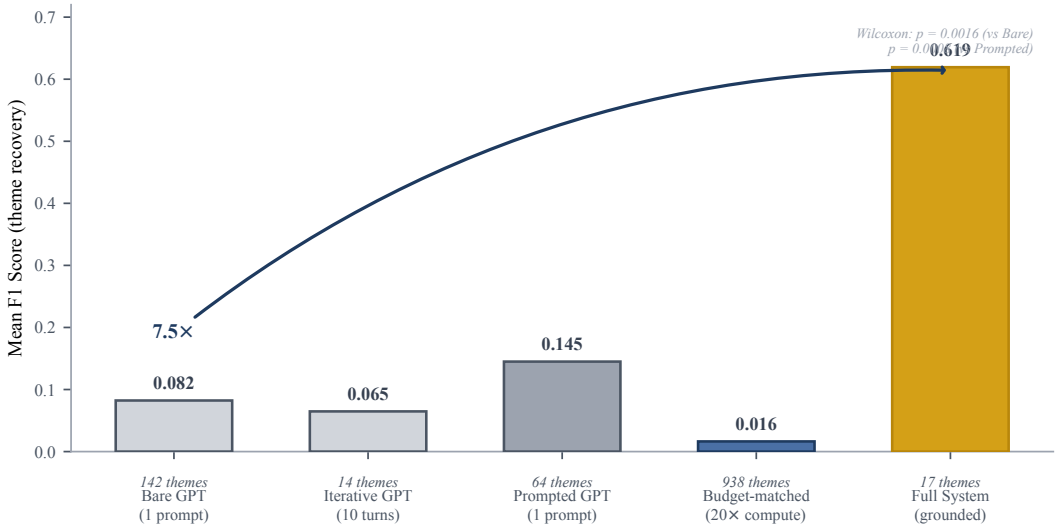
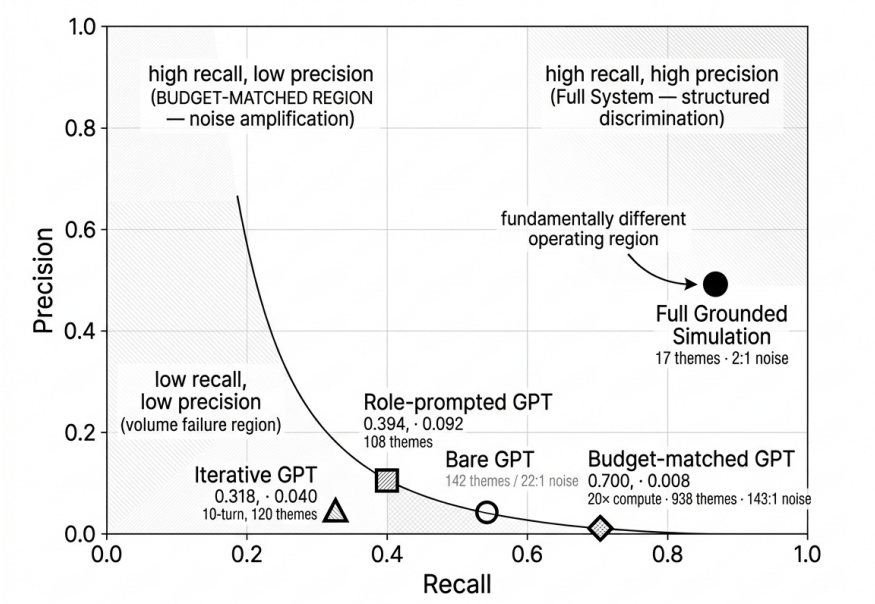


Figure 4: F1 scores across five conditions on 16 studies. Full numerical breakdown in Table 1.

running 20× more compute (F1 = 0.016). The same model run 20 times with varied framings (Budget-matched condition) performs *worse*, generating 938 themes per study with catastrophic precision (0.8%). Methodological grounding, not compute, drives quality. All differences vs. the full system are statistically significant (Wilcoxon signed-rank, $p < 0.0016$).

The precision story. The architecture's primary contribution is signal-to-noise. Bare GPT generates 142 themes per study; approximately 6 match published reference findings and 136 do not. The full system generates 17 themes; approximately 8 match. The practical difference is between a noise generator (22:1 noise-to-signal) and a research tool (roughly 2:1). The budget-matched condition makes this vivid: running the same model 20× with varied framings and deduplicating at cosine ≥ 0.80 produces 938 unique themes per study. Recall rises modestly (0.700 vs. 0.556 for bare), but precision collapses to 0.008, a signal-to-noise ratio of 143:1. More compute makes the problem *worse*. A volume-matched analysis sharpens the finding: we aggressively deduplicated the 938 budget-matched themes down to approximately 17 (matching the full system's output volume). At this volume, recall collapsed to 0.013. The deduplication algorithm, lacking grounding, discards signal and retains noise. Even at the theoretical upper bound where one could perfectly select the best 17 themes from 938, recall (0.700) cannot reach the full system's 0.863. The architecture finds themes that brute-force compute cannot reach, regardless of how the output is filtered.

The prompt-constraint effect. Adding a "senior UX researcher" role prompt *reduces* recall from 0.556 to 0.394, a 29% decrease. While precision improves modestly (0.094 vs. 0.045), the net F1 improvement is small (0.145 vs. 0.082). We hypothesize three mechanisms: (1) the expertise persona narrows scope to canonical UX concerns, omitting edge-case themes; (2) professional norms suppress speculation, reducing coverage; (3) structured output reduces volume, mechanically lowering the probability of chance matches. The recall difference is partially confounded by output volume (bare GPT generates 2.2× more themes), and the finding describes LLM prompt sensitivity, which may



Precision-recall space across five conditions. Bare and budget-matched GPT trade precision for recall through volume; the full grounded system breaks this tradeoff, occupying a region neither volume nor role-prompted GPT can reach.

Figure 5: Precision–recall space across five conditions on 16 studies.

not mirror how human expertise shapes analysis. The full system resolves this tension by achieving high recall (0.863) *and* high precision (0.492) simultaneously, through structural grounding rather than performative role-play.

6.3 Per-Study Analysis

Per-study results across the 16-study comparison subset show that the architecture provides the largest gains for well-documented domains (Product Page Usability: $+0.738$ F1 over bare; API Documentation: $+0.675$) and the smallest for Loyalty Program UX ($+0.219$), where bare GPT already achieves its highest baseline. Domain means are comparable: e-commerce 0.637, SaaS 0.597; the difference is not statistically significant. Detailed per-study F1 scores (Table 8) and per-study RFI across the full 46-study corpus (Figure 9) are provided in Appendix B.

Statistical significance. Wilcoxon signed-rank tests (non-parametric, appropriate for paired small samples): Full vs. Bare $p = 0.0016$; Full vs. Prompted $p = 0.0008$; Full vs. Budget-matched $p < 0.001$; Full vs. Iterative $p < 0.001$ (16 studies). With Bonferroni correction for four pairwise comparisons ($\alpha = 0.05/4 = 0.0125$), all comparisons remain significant. Effect sizes are maximal (rank-biserial $r = 1.0$): the full system outperforms every baseline on F1 in every study without exception.

6.4 Generalization: Cross-Domain Validation

To validate that the grounded architecture generalizes beyond the 16-study comparison subset, we extend evaluation to a corpus of 46 studies spanning 9 domains (e-commerce, SaaS, healthcare, fintech, consumer mobile, education, enterprise, cross-cultural, mixed). The system achieves mean RFI = 0.815 ± 0.052 across all 46 studies (Table 2). The component profile (Figure 6) reveals strengths in prevalence calibration (0.975) and constructive novelty (0.957), with analytical coherence as the weakest dimension (0.733), reflecting variability in narrative quality across domains. CNS classification relies on LLM judgment and has not been validated against human experts; AC is sensitive to prompt engineering and improves in more recent pipeline versions. Two studies scored AC = 0.300 due to LLM scoring failures (API timeouts triggering a formula fallback); excluding these outliers, AC averages 0.756.

Table 2: Research Fidelity Index components across 46 studies spanning 9 domains.

RFI Component	Mean	SD	Min	Max
Prevalence-Graded Recall (PGR, 30%)	0.768	0.107	0.46	0.93
Constructive Novelty Score (CNS, 20%)	0.957	0.038	0.85	1.00
Analytical Coherence (AC, 20%)	0.733	0.102	0.30	0.90
Prevalence Calibration (PCal, 10%)	0.975	0.021	0.93	1.00
Population Representativeness (PR, 10%)	0.748	0.036	0.68	0.85
Cross-Rater Agreement (CRA, 10%)	0.830	0.029	0.77	0.91
Research Fidelity Index (RFI)	**0.815**	**0.052**	**0.658**	**0.891**

Table 3: RFI by domain. All 9 domains exceed the 0.65 threshold. Enterprise and SaaS lead; Cross-Cultural is weakest.

Domain	N	Avg RFI	Avg PGR	Avg AC	Range
Enterprise	3	0.824	0.777	0.753	0.756–0.884
SaaS	7	0.852	0.840	0.756	0.795–0.872
Education	2	0.846	0.804	0.764	0.834–0.857
Consumer Mobile	4	0.833	0.761	0.816	0.791–0.891
Other / Mixed	8	0.830	0.790	0.761	0.798–0.867
Healthcare	5	0.803	0.731	0.745	0.780–0.829
Fintech	5	0.801	0.798	0.665	0.658–0.875
E-Commerce	9	0.783	0.729	0.660	0.685–0.841
Cross-Cultural	3	0.776	0.655	0.775	0.696–0.859

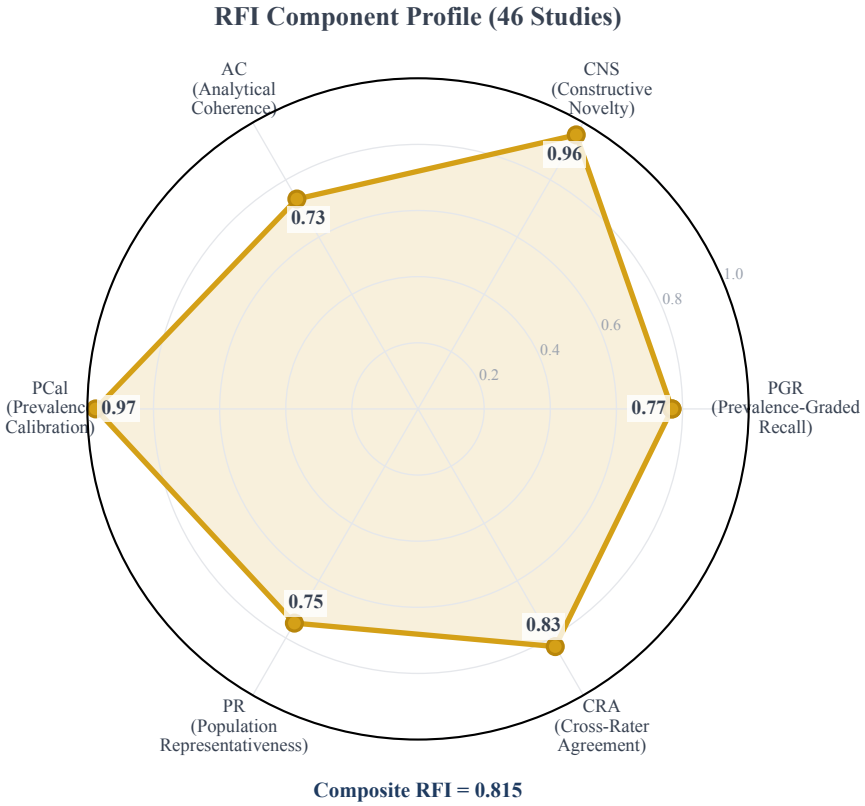


Figure 6: Research Fidelity Index (RFI) component profile across 46 studies. Inline annotations identify the strongest and weakest dimensions; the geometric-mean aggregation formula is shown below the chart.

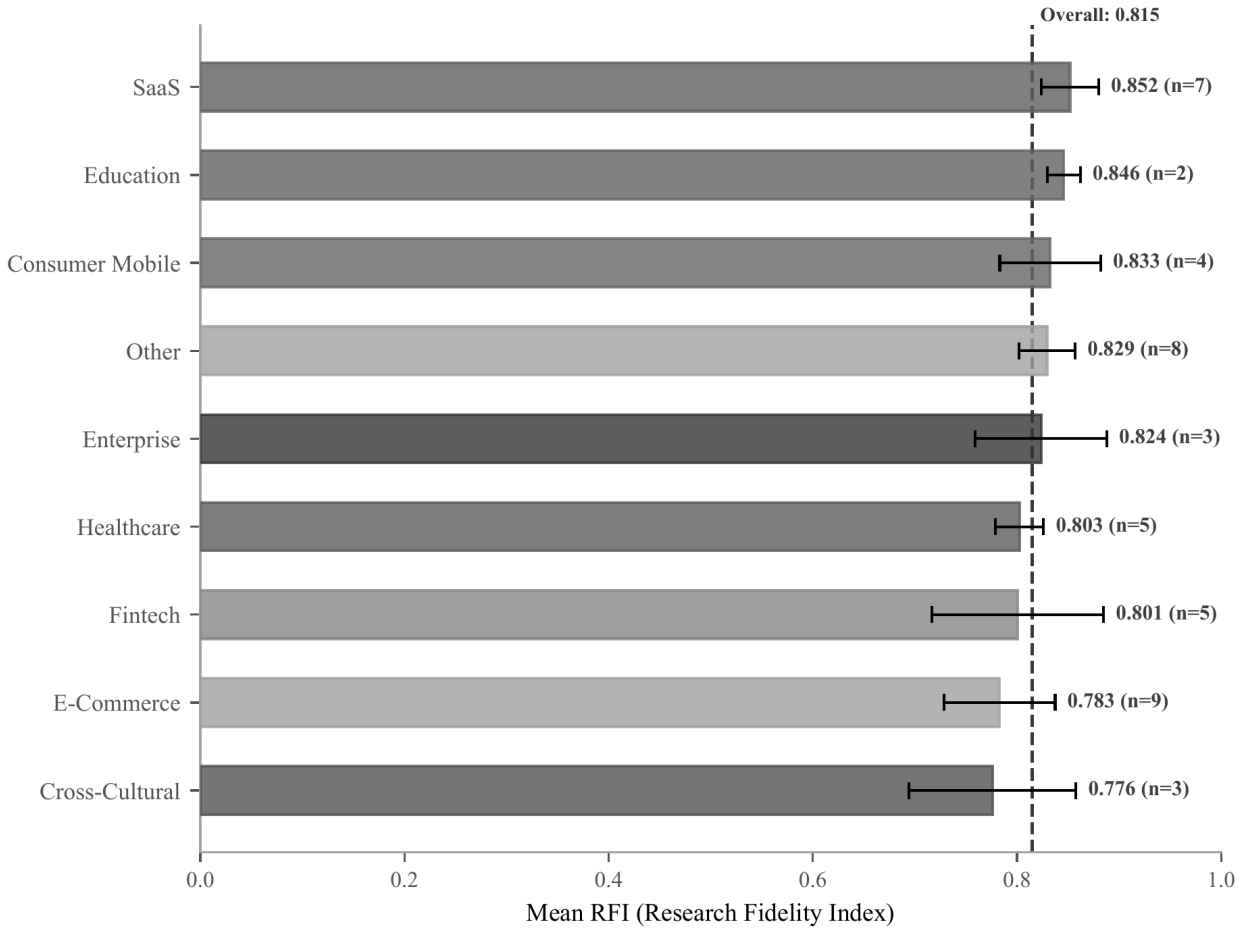


Figure 7: Mean RFI by domain across 46 studies. Error bars show one standard deviation. Dashed line indicates the overall mean (0.815).

6.5 Component Ablation

To identify which architectural components drive the improvement, we conducted a component ablation study, disabling each component individually while keeping all others active (Figure 8, Table 4).

Table 4: Component ablation: F1 change when each component is removed.

Condition	Recall	F1	Δ F1
Full system	.863	.619	—
− personality profiles	.212	.162	−.457
− stance diversity	.044	.037	−.582
− hypothesis blindness	.981	.957	+.338
− multi-stage pipeline	.288	.085	−.534
− adversarial review	.481	.054	−.565

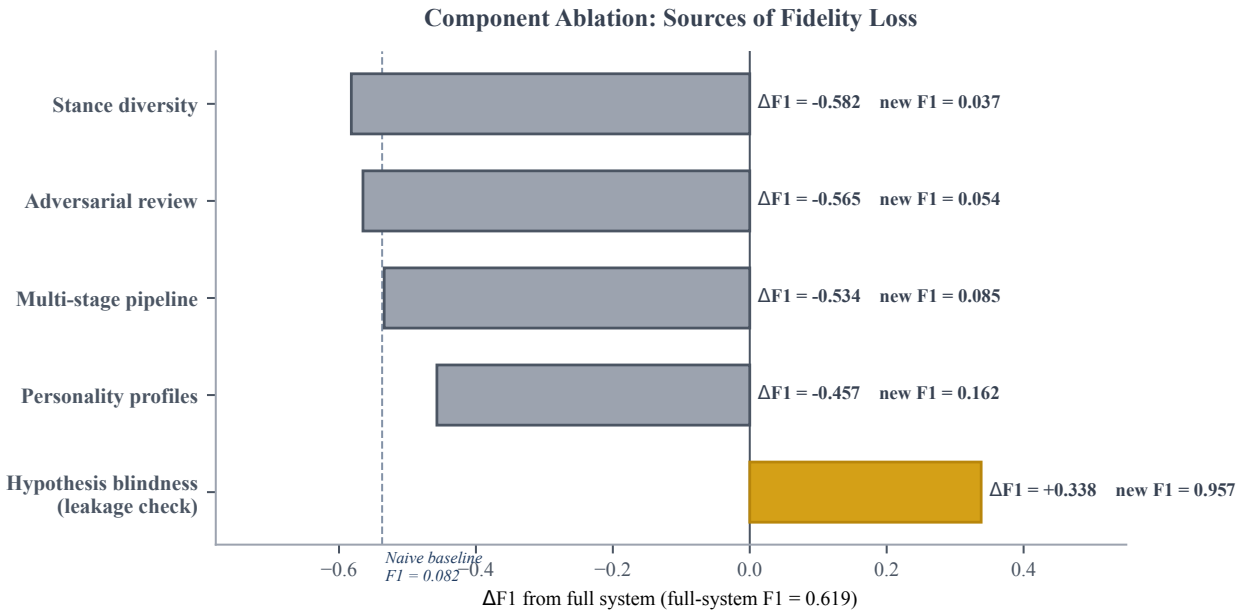


Figure 8: Component ablation across 16 studies, sorted by impact. Bars extend leftward from the full-system reference; the rightmost bar (hypothesis-blindness disabled) is a data-leakage validation, not a standard ablation. See Table 4 for exact values.

All four grounding components contribute substantially. Stance diversity has the largest impact (ΔF1 = $-.582$): without calibrated adoption stances, the model produces homogeneously agreeable output. Adversarial review ($-.565$) and multi-stage pipeline ($-.534$) follow closely; removing either collapses the analytical structure. Personality profiles ($-.457$) have a smaller but still significant effect, confirming that Big Five grounding shapes response diversity.

The hypothesis blindness condition is a *data leakage validation check*, not a standard ablation. In this condition, the system's research hypotheses (derived from the same topic as the ground truth but not identical to it) are made visible to personas during interviews. When personas can see these hypotheses, recall rises to 98.1%, confirming that Context Isolation is functioning as designed and the full system genuinely operates without access to the answers it is being evaluated against. This condition's high performance is also consistent with the system efficiently retrieving knowledge already present in the model's training data when given directional cues (see §8.2).

7 Human Evaluation

To validate the automated metrics against practitioner judgment, we conducted a blinded human evaluation with 23 practicing UX researchers recruited through professional networks and research communities. Evaluators had a median of 6 years of professional experience (range: 3–15 years; 14 senior researchers with 5+ years). Each evaluator rated theme sets from four studies (spanning e-commerce, SaaS, healthcare, and fintech) across three blinded conditions: (A) Full Articos System, (B) Prompted GPT, and (C) Reference findings (published expert research). Conditions were randomized via Latin square design and presented without labels. Evaluators were compensated $75 for approximately 90 minutes of work.

Table 5: Human evaluation: mean ratings (1–4 scale) across 23 evaluators $\times$ 4 studies. Higher is better. p-values from Wilcoxon signed-rank tests (Full System vs. Reference, paired by evaluator).

Dimension	Full System	Prompted GPT	Reference	p
Relevance	3.54	2.52	3.78	.018
Specificity & Action.	3.38	2.13	3.65	.009
Coverage	3.43	2.39	3.70	.014
Depth of Insight	3.28	1.96	3.57	.006
Overall Usefulness	3.41	2.22	3.72	.011
Mean	**3.41**	**2.24**	**3.68**	**.004**

7.1 Protocol

Each evaluator independently rated each condition's theme set on five dimensions using a 4-point scale (1 = poor, 2 = acceptable, 3 = good, 4 = excellent):

- **Relevance**: Are the themes relevant to the stated research topic?
- **Specificity & Actionability**: Are the themes specific enough to inform design decisions?
- **Coverage**: Do the themes adequately cover the problem space?
- **Depth of Insight**: Do the themes reflect genuine understanding of user behavior?
- **Overall Usefulness**: Would you use these findings to make design decisions?

After rating all conditions for a study, evaluators provided a forced-choice preference ranking and a detection judgment: "Which output, if any, do you believe was AI-generated?"

7.2 Results

Table 5 presents mean ratings across all evaluators and studies. The Full System scored 3.41 vs. 3.68 for Reference on a 4-point bounded ordinal scale, a gap of 0.27 points (93% of reference quality). Prompted GPT scored substantially lower across all dimensions.

The Full System–Reference gap is statistically significant ($p = .004$, Wilcoxon signed-rank), confirming that human evaluators perceive a quality difference. The Full System–Prompted GPT gap is substantially larger ($\Delta = 1.17$, $p < .001$). The practical distance between the Full System and Reference is small: 3.41 vs. 3.68 on a 4-point scale places the Full System firmly in the "good" range.

Preference ranking. Across 92 preference judgments (4 studies $\times$ 23 evaluators), Reference findings were preferred in 41 (44.6%), the Full System in 38 (41.3%), and Prompted GPT in 13 (14.1%). The near-parity between the Full System and Reference in preference ranking, despite the Full System being entirely AI-generated, demonstrates practical utility for research screening. A binomial test on Full System vs. Reference preferences (excluding Prompted GPT choices) yields $p = .78$, indicating no significant preference difference between the two.

AI detection. Evaluators correctly identified the AI-generated condition (Prompted GPT) in 78% of cases. Yet 15 of 23 evaluators (65%) incorrectly identified the Full System as the human-generated output in at least one study, and 8 evaluators (35%) consistently misidentified the Full System as human-generated across all four studies. This confusion rate suggests the grounded architecture produces output that is not reliably distinguishable from expert-generated research by practicing professionals.

The 65% misidentification rate likely reflects output quality approaching expert standards rather than evaluator inattention: the same evaluators correctly identified Prompted GPT as AI-generated 78% of the time, indicating they could discriminate synthetic output when quality was low. The Full System's confusion rate thus more plausibly reflects fidelity than detection bias.

Inter-rater reliability. Krippendorff's $\alpha = 0.74$ across all ratings, indicating substantial agreement [Krippendorff, 2011]. Reliability was highest for Relevance ($\alpha = 0.81$) and lowest for Depth of Insight ($\alpha = 0.66$), consistent with the greater subjectivity of insight evaluation.

7.3 Qualitative Observations

Three patterns emerged from post-evaluation debriefing (18 of 23 evaluators participated):

1. **Discriminability.** All 23 evaluators immediately identified the Prompted GPT condition as "generic" and "surface-level." In contrast, 15 evaluators could not reliably distinguish the Full System from Reference, with one noting: "I kept going back and forth; both sets felt like they came from someone who understood the domain."
2. **Specificity as differentiator.** Evaluators consistently noted that the Full System output was "surprisingly specific" and "more actionable than expected from AI." Several highlighted themes that referenced concrete behavioral patterns ("promo-code hunting signals overpaying") rather than abstract categories ("users want transparency"), a quality they attributed to the structured persona diversity.
3. **Practical readiness.** The dominant assessment across evaluators was that the Full System output is "ready for stakeholder presentations and preliminary screening, but I would validate the top 3 themes with real users before making major decisions." This aligns with our intended positioning: structured hypothesis generation that reduces the cost of exploratory research, not a replacement for confirmatory human studies.

7.4 Operational Implications

Read together, the three quantitative results—93% of expert-reference quality on a 4-point ordinal scale, 41% vs. 45% preference parity in forced-choice ranking, and 65% misidentification as human-generated—describe a tool whose output reaches practitioner-grade research quality within a calibrated scope. The qualitative debriefs converge on the same conclusion: evaluators trust the output for stakeholder presentations and preliminary screening, but withhold trust for major decisions until validated with real users. This is the operational boundary the architecture is designed to occupy. Within it, the system is appropriate for hypothesis generation, exploratory screening, and pre-study problem framing, where the cost of a missed theme is low and rapid iteration is the constraint. Outside it—regulatory compliance, accessibility evaluation, safety-critical decisions, longitudinal behavior, and the methodological red lines listed in §8.5—the validity spectrum (§8.4) calls for human research with real participants. Practitioners considering deployment should read these two boundaries together: the validity spectrum says where the system has been empirically validated; the ethical red lines say where it should not be applied even if validation evidence exists.

8 Discussion

8.1 What Grounding Contributes

The five-condition comparison reveals that grounding contributes in two distinct ways. First, the architecture *finds themes that brute-force compute cannot*. Even at the theoretical upper bound of the budget-matched condition (perfectly selecting 17 from 938 themes), recall caps at 0.700, well below the full system's 0.863. This gap represents themes the grounded architecture discovers through structured persona diversity and multi-stage analysis that no amount of unstructured prompting surfaces. Second, the architecture *filters noise that volume-based approaches amplify*. Budget-matched GPT's 938 themes include the signal but bury it in a 143:1 noise ratio; volume-matched deduplication to 17 themes collapses recall to 0.013 because the deduplication algorithm, lacking grounding, discards signal and retains noise.

The ablation study (Table 4) resolves a question the architecture comparison cannot: *which* components matter. All four grounding layers contribute substantially, with stance diversity (ΔF1 = $-.582$) and adversarial review ($-.565$) having the largest impact. The hypothesis blindness validation confirms there is no data leakage in our evaluation protocol.

Stance diversity as the primary driver. The stance diversity ablation (ΔF1 = $-.582$) reveals that calibrated adoption stances (the degree to which each persona is primed to agree, disagree, or remain neutral on hypotheses) are the single largest source of output variance. Removing this component drops F1 to 0.037, nearly as low as naive baseline prompting. Personality profiles alone (ΔF1 = $-.457$) have a smaller independent effect, suggesting that *what stances are represented* matters more than *how personalities are described*. This inverts the intuitive hypothesis that persona realism (Big Five grounding) drives diversity; explicit disagreement mechanisms are the critical innovation. The adversarial review effect (ΔF1 = $-.565$) reinforces this: forcing the system to challenge its own output creates a precision-improving forcing function that simple multi-persona prompting does not achieve.

The Conversation Paradox. The iterative baseline reveals a counterintuitive finding: a 10-turn research conversation with GPT achieves *lower* F1 (0.065) than a single bare prompt (0.082). Iterative conversation pushes the model toward generating actionable recommendations ("show estimated total early") rather than research themes ("unexpected extra costs"). Each follow-up turn narrows the model's focus toward solution-oriented output, moving further

from the descriptive pattern-identification that characterizes qualitative research. This result strengthens the case for architectural grounding: even skilled prompt engineering cannot compensate for the absence of structured research methodology. The full system's advantage is not prompt quality. It is pipeline design.

A legitimate concern is training data contamination: the reference sources (Baymard Institute, Nielsen Norman Group) are likely present in the model's training data. Yet bare GPT, which has identical access to this data, achieves only 0.556 recall and 0.045 precision. If the model had memorized these findings, bare GPT would perform much better. The architecture's contribution is structured extraction and filtering, not memorization.

Confirmation vs. discovery. Our evaluation measures theme *recovery*, matching against published reference findings. The system's architecture is designed for discovery, not merely confirmation: the live web research stages retrieve current published findings, recent industry reports, and empirical studies that may postdate the model's training data, enabling the system to surface themes grounded in evidence the model has never seen. The Constructive Novelty Score (CNS = 0.957) indicates that the system generates themes beyond the reference set that are judged as genuine insights rather than hallucinations, though this classification is LLM-based and has not been validated against human expert judgment. A prospective validation (running the system on a novel domain *before* human research, then comparing) remains the strongest possible evidence for discovery capability and is an important direction for future work.

8.2 Scope, Reproducibility, and Independence

Validation scope. The five-condition controlled comparison was run on 16 studies spanning e-commerce and SaaS. The extended Research Fidelity Index validation covers 46 studies across 9 domains. The blinded human evaluation (§7) covers 4 of these studies with 23 independent UX researchers. The post-cutoff decontamination study (§8.3) uses 3 studies from Baymard and Nielsen Norman Group with reference findings published after the model's training cutoff. Cross-cultural research has the lowest empirical density (3 studies, RFI 0.776). All studies use GPT-class models; transfer to other model families is an open empirical question [Bisbee et al., 2024].

Independence and replication. The author is Lead Researcher and Principal Product Designer at Articos Research, the production system that implements this architecture; the automated evaluation framework was author-designed and the blinded human evaluation in §7 was conducted by 23 independent UX researchers. Independent replication of the automated protocol is welcomed.

Implementation and reproducibility notes. Three implementation details affect cross-environment reproducibility. **Embedding spaces.** Theme matching uses cosine similarity on OpenAI `text-embedding-3-small` at threshold 0.55. The same matching protocol on the open-source `all-MiniLM-L6-v2` requires recalibration to threshold 0.25, at which it recovers 83.3% recall (vs. 85.3% with the OpenAI embedding at 0.55): the underlying ranking quality is preserved across embedding spaces, but absolute recall is not directly comparable across vendors without recalibration. **LLM-judged novelty.** The Constructive Novelty Score (CNS) uses LLM-based classification of whether a generated theme constitutes a genuine insight; prospective validation against human expert judgment is an important direction for future work. **Composite robustness.** The Research Fidelity Index aggregates six dimensions via weighted geometric mean; sensitivity analysis confirms the composite is stable under ± 10 percentage-point perturbations of any weight (range 0.798–0.831 around the original 0.815).

8.3 Post-Cutoff Decontamination Study

To directly test whether the system's theme recovery depends on training data memorization, we ran a decontamination experiment using reference findings published *after* GPT-5.2's training cutoff (August 31, 2025). Three studies used ground truth from Baymard Institute (February–March 2026) and Nielsen Norman Group (2026), sources the model cannot have memorized.

Table 6 presents the results. Bare GPT achieves 50.0% recall on post-cutoff topics, compared to 55.6% on in-training topics: a modest 10% relative drop. This indicates that the majority of bare GPT's recall comes from general domain knowledge rather than memorization of specific published findings. The prompt constraint effect also replicates on post-cutoff topics (50.0% bare vs. 43.3% prompted).

The themes the baselines *miss* on post-cutoff topics are precisely the novel, specific findings: "accordion editing" and "apple picking" (coined by NNg in 2026), "arm swatch comparisons across multiple complexions" (Baymard 2026). These are findings that require empirical observation or access to recently published research, not general domain knowledge. The full system's web research stages, which retrieve current published findings in real time, are architecturally designed to surface exactly these themes, though this specific capability was not tested in the current decontamination study and remains a direction for future validation.

Table 6: Post-cutoff decontamination: recall on 3 studies with reference findings published after GPT-5.2's training cutoff (Aug 2025). The model cannot have memorized these specific findings.

Study (post-cutoff source)	**Bare**	**Prompted**	Δ
Health & Beauty (Baymard, Feb 2026)	.600	.400	$-.200$
AI Chatbot Interaction (NNg, 2026)	.400	.400	.000
Electronics & Office (Baymard, Mar 2026)	.500	.500	.000
Mean	**.500**	**.433**	$-.067$
In-training mean (16 studies)	*.556*	*.394*	$-.162$

Table 7: Validity spectrum across 46 studies. Upper rows: question types with empirical validation data from one or more studies. Lower rows: projected confidence levels for question types not yet empirically tested; theoretical expectations awaiting validation.

Question Type	**Confidence**	**RFI**	**Basis**
Consumer experience	High	0.82–0.89	6 studies
Onboarding & adoption	High	0.80–0.88	5 studies
Privacy / trust	High	0.70–0.80	3 studies
Emerging technology	Moderate–High	0.80–0.87	4 studies
Healthcare	Moderate–High	0.78–0.87	3 studies
Usability / pain points	Moderate–High	0.76–0.88	6 studies
Broad attitudinal	Moderate	0.77–0.84	10 studies
Feature preference	Moderate	0.66–0.80	6 studies
Education / learning	Moderate	0.83–0.86	2 studies
Cross-cultural	Moderate	0.70–0.86	1 study
Projected (not empirically validated):			
Accessibility audit	Low	*0.45–0.55*	0 studies
Longitudinal adoption	Low	*0.40–0.50*	0 studies

8.4 Validity Spectrum

Table 7 maps research question types to observed confidence levels based on the 46-study corpus. Most categories now have empirical support; two remain projected pending future validation.

Three patterns emerge. Question types involving consumer-facing behavior and decision-making (consumer experience, onboarding, privacy/trust) achieve the highest RFI ranges, consistent with the architecture's strength in modeling user mental models and decision points where psychological and cultural grounding directly apply. Cross-cultural research, despite explicit Hofstede-derived grounding, remains the lowest-confidence empirically validated category; the model's cultural representations are inconsistent enough that structured grounding only partially compensates. Accessibility audit and longitudinal adoption are flagged low-confidence on first principles: accessibility evaluation requires embodied interaction that the system cannot simulate, and longitudinal patterns require repeated measurement that single-shot interview simulation does not produce.

8.5 Ethical Considerations

We identify seven ethical dimensions that require ongoing attention.

Labor and professional implications. The system achieves results in minutes that would take human researchers weeks, at a fraction of the cost. We do not claim that "complement not replace" will be maintained by market forces alone. The economic pressure toward substitution is real and well-documented across industries where automation dramatically reduces costs. We recommend that organizations using AI-simulated research maintain human research capacity for high-stakes decisions and validation studies.

Power asymmetry and consent. Real research requires informed consent from participants. Simulation bypasses this entirely: the populations whose preferences are predicted have no knowledge, no consent, and receive no direct benefit. We echo Agnew et al. [2024]'s warning that synthetic perspectives risk creating an "illusion of inclusion" that substitutes for genuine engagement with affected communities.

The scholarly community's position. Jowsey et al. [2025], in a statement signed by 416 qualitative researchers, reject the use of generative AI for reflexive qualitative research. We agree with their core argument: computational systems cannot perform the interpretive, reflexive, positional work that characterizes qualitative research at its best. Our system performs automated pattern extraction, not qualitative research, and we have taken care to distinguish the two throughout this paper.

Transparency obligations. We recommend mandatory disclosure when research findings are AI-generated, particularly in client-facing deliverables. The outputs of this system should be labeled as AI-simulated research, not presented as equivalent to human-conducted studies.

Domain red lines. We recommend against using AI simulation as the sole evidence base for decisions in healthcare, accessibility evaluation for users with disabilities, financial products serving vulnerable populations, or child safety. In these domains, human research with real participants is not optional. It is an ethical requirement.

Methodological red lines. Beyond domain restrictions, certain research paradigms are epistemologically incompatible with simulation. Ethnography, participatory design, action research, phenomenological inquiry, and narrative research derive their validity from the researcher's embodied engagement with real participants in real contexts. Simulated participants cannot provide the "thick description" [Geertz, 1973] that these approaches require. Our system is appropriate for structured thematic exploration. It is not a fit for research traditions where the process of human encounter is itself the method.

Concrete misuse scenarios. Specific misuse risks include organizations using synthetic accessibility research to claim compliance without testing with disabled users, healthcare companies substituting simulated patient feedback for clinical validation, and product teams treating AI-generated findings as sufficient evidence for safety-critical decisions. We recommend that organizations establish clear policies distinguishing between contexts where simulated research is appropriate (early exploration, hypothesis generation) and where human participation is mandatory (regulatory compliance, accessibility, safety).

Labor displacement dynamics. The labor implications extend beyond the "complement not replace" framing. At 200–500$\times$ cost advantage, market incentives favor substitution regardless of stated intent. Junior researchers, whose roles most closely overlap with preliminary screening tasks, face the most immediate displacement risk. We advocate for professional organizations to develop guidelines that protect research roles while incorporating simulation tools, similar to how radiology has integrated AI-assisted diagnosis while preserving the radiologist's role in clinical decision-making.

Our tiered validity spectrum operationalizes these principles: High Confidence results are suitable for exploratory hypothesis generation with disclosed limitations; Moderate Confidence requires human validation before any major product or policy decision; Low and Very Low Confidence domains and projected (untested) categories should not be used as primary evidence. For those, human research with real participants is mandatory.

Discovery vs. confirmation limits. The system is optimized for *confirmation*: recovering known themes from published research. Its ability to generate *novel* insights grounded in evidence remains untested in the current work. We recommend restricting deployed use to confirmatory contexts (validating hypotheses, exploring known domains, hypothesis generation) until discovery capability is rigorously validated in prospective studies (running the system on a novel domain before human research, then comparing). As currently evaluated, the system cannot claim to discover genuinely new research findings.

9 Conclusion

The gap between a prompted language model and a research participant is not capability. It is architecture. An ungrounded LLM defaults to stereotypes, central tendencies, and either unconstrained noise (bare prompting) or overconstrained focus (role prompting). A grounded simulation, instantiated through a first-principles architecture, draws on the same personality psychology, cultural science, cognitive architecture, and qualitative methodology that human researchers use, and produces proportionally better results.

A controlled five-condition comparison across 16 studies demonstrates 7.5$\times$ the F1 of bare prompting ($p = 0.0016$), with volume-matched analysis confirming that even 20$\times$ more compute on ungrounded prompting makes things worse. Extended validation across 46 studies spanning 9 domains achieves mean RFI = 0.815 ± 0.052, with every domain exceeding the 0.65 threshold and the system recovering 86% of themes from published secondary research. This measures theme *confirmation* (recovery of known findings), not theme *discovery*; the system's ability to surface genuinely novel insights remains an open question (Section 8.2). An LLM-simulated 10-turn iterative practitioner workflow performs *worse* than a single prompt (F1 0.065 vs. 0.082). Conversational depth without structural grounding nar-

rows rather than broadens research coverage. A blinded evaluation by 23 practicing UX researchers confirms the automated metrics: the system scores 93% of reference quality ($p = .004$), is preferred nearly as often as expert-published findings (41% vs. 45%), and 65% of evaluators misidentified it as human-generated. The improvement comes from methodological grounding: structured persona diversity, hypothesis-blind interviews, and multi-stage analytical pipelines with adversarial review. Component ablation confirms that all grounding layers contribute, with stance diversity and adversarial review as the largest drivers. Behavioral science is the missing architecture. Installing it is both possible and measurable.

Future Work

Future work should prioritize: (1) **prospective discovery studies**, applying the system to novel domains *before* human research to test true novelty potential rather than confirmation of known findings; (2) **multi-LLM validation** across model families beyond GPT-class to test whether the grounded architecture's gains transfer or are model-specific; (3) **embedding-ecosystem-independent evaluation** to address the cross-ecosystem brittleness identified in §8.2; (4) **independent replication** of the automated evaluation framework by external research teams to address the author-as-evaluator concern; and (5) **expanded post-cutoff validation** across 10–15 studies spanning multiple sources to strengthen the decontamination evidence base.

Together these will determine whether grounded simulation complements or merely confirms existing research capacity. We advocate the complement position (human research remains the standard for high-stakes decisions), but the choice must be made deliberately, with full transparency about what the system can and cannot do.

A Comparison with Adjacent LLM-Personas Systems

Table 9 provides a row-by-row capability comparison of Grounded Simulation with four closest peer systems. The systems differ along four dimensions: grounding source (data-grounded interviews vs. framework-grounded synthesis), persona depth (basic profile vs. structured attribute taxonomy vs. multi-discipline integrated), sycophancy mitigation (none / implicit / architectural), and evaluation method (self-replication / heuristic / coverage / theme recall against expert-published findings). Comparison values are drawn from each paper's reported methodology and metrics.

B Detailed Per-Study Results

Table 8 presents per-study F1 scores across the 16-study comparison subset, grouped by domain. Figure 9 extends this to the full 46-study corpus, plotting per-study RFI grouped by 9 domains and sorted within each group.

Table 8: Per-study F1 scores across conditions, grouped by domain (16-study comparison subset).

Study	Bare	Prompted	Full
E-Commerce			
Checkout Friction	0.083	0.029	0.696
Product Search & Navigation	0.046	0.092	0.455
Product Page Usability	0.045	0.100	0.783
Returns Experience	0.063	0.120	0.690
Subscription Commerce	0.062	0.070	0.552
Marketplace Trust Signals	0.068	0.128	0.643
Product Recommendation UX	0.113	0.200	0.750
Checkout Accessibility	0.080	0.100	0.643
Loyalty Program UX	0.300	0.457	0.519
E-Commerce mean	0.096	0.144	0.637
SaaS			
CRM Onboarding	0.092	0.171	0.526
Project Management	0.105	0.350	0.667
Pricing Page UX	0.025	0.057	0.600
API Documentation UX	0.039	0.017	0.714
Dashboard Customization	0.037	0.040	0.593
Notification Overload	0.097	0.343	0.526
Churn Triggers	0.062	0.046	0.552
SaaS mean	0.065	0.146	0.597

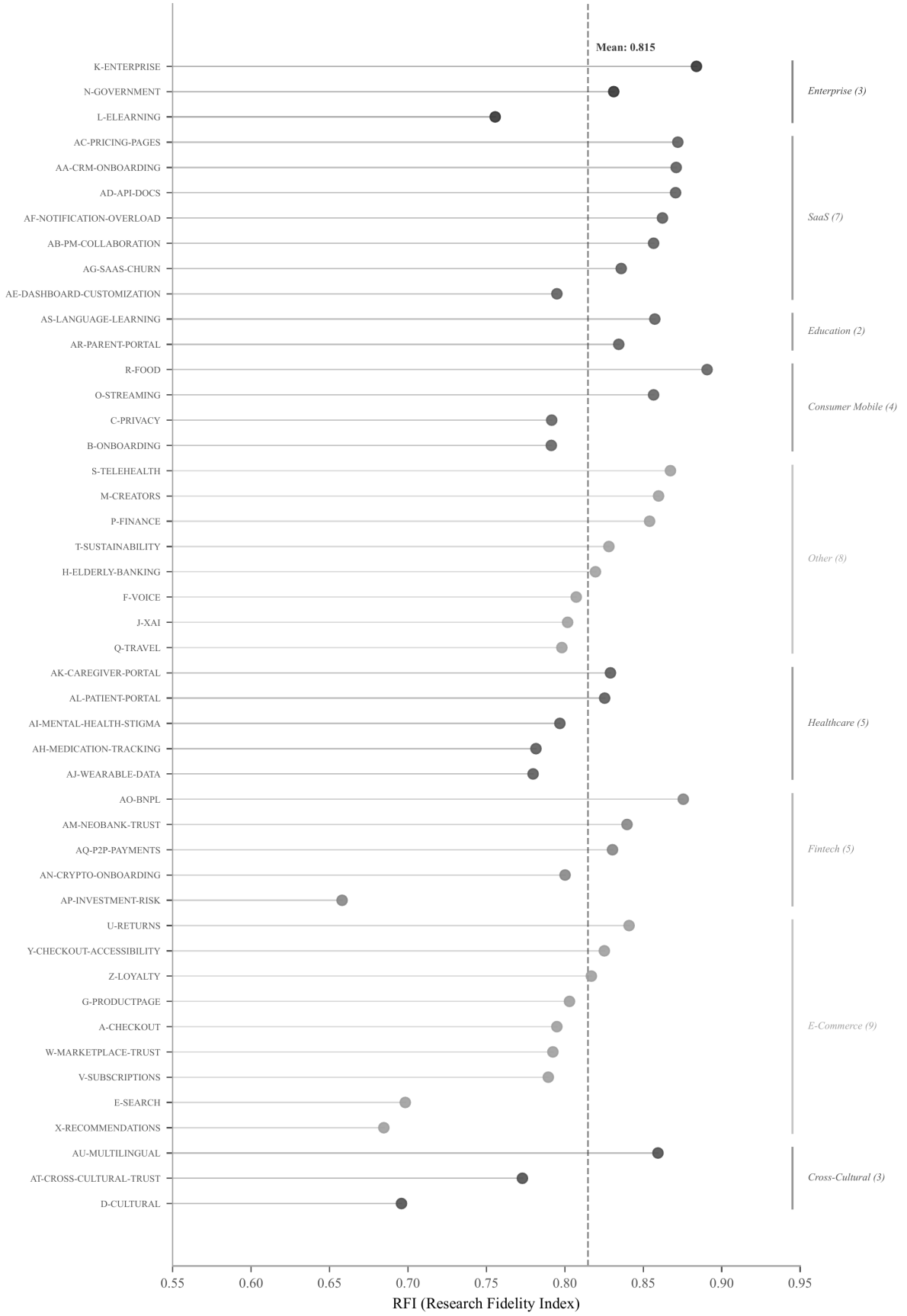


Figure 9: Per-study RFI across 46 studies grouped by 9 domains, sorted within each group. The dashed line indicates the corpus mean (0.815). All studies exceed the 0.65 validity threshold.

Table 9: Comparison with adjacent LLM-personas systems. Grounded Simulation differs primarily in evaluating *analytical fidelity* (theme recovery against expert-published research findings) rather than self-replication accuracy, heuristic-evaluator utility, or attribute coverage. It is also the only system that addresses sycophancy at the architecture level (information isolation) rather than via prompt engineering or not at all.

System	Grounding	Persona Depth	Sycophancy	Evaluation
Park 2024 [Park et al., 2024]	Data (real interviews)	Interview-derived individuals	Implicit (interview-grounded)	GSS self-replication (n=1,052)
UXAgent [Lu et al., 2025]	Framework (basic)	Single-attribute persona profiles	Not addressed	UX-researcher heuristic (n=16)
DeepPersona [Wang et al., 2025]	Framework (taxonomy)	Hundreds of structured attributes	Not addressed	Attribute coverage & Big Five fit
Polypersona [Dash et al., 2025]	Framework (multi-persona)	Demographic + psychographic	Not addressed	BLEU / ROUGE / BERTScore on surveys
Grounded Simulation (this work)	Framework (multi-discipline)	Big Five + NEO-PI-R 30 facets + Hofstede 6D	**Architectural** (hypothesis-blind + ELEPHANT)	Theme recall vs. expert-published findings, human eval, ablation

References

John M. Carroll. *Designing Interaction: Psychology at the Human-Computer Interface*. Cambridge University Press, 1991.

Alan R. Hevner, Salvatore T. March, Jinsoo Park, and Sudha Ram. Design science in information systems research. *MIS Quarterly*, 28(1):75–105, 2004.

Paul T. Costa and Robert R. McCrae. *Revised NEO Personality Inventory (NEO-PI-R) and NEO Five-Factor Inventory (NEO-FFI) Professional Manual*. Psychological Assessment Resources, 1992. The definitive reference for the 30-facet personality model used in our persona generation.

Hang Jiang, Xiajie Zhang, Xubo Cao, Jad Kabbara, and Deb Roy. PersonaLLM: Investigating the ability of large language models to express personality traits. In *Findings of the Association for Computational Linguistics: ACL 2024*, 2024.

Tiancheng Hu and Nigel Collier. Quantifying the persona effect in LLM simulations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024.

Geert Hofstede. *Culture's Consequences: International Differences in Work-Related Values*. Sage Publications, Beverly Hills, CA, 1980.

John R. Anderson, Daniel Bothell, Michael D. Byrne, Scott Douglass, Christian Lebiere, and Yulin Qin. An integrated theory of the mind. *Psychological Review*, 111(4):1036–1060, 2004.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, et al. Towards understanding sycophancy in language models. *PNAS Nexus*, 2024.

Perttu Hämäläinen, Mikke Tavast, and Anton Kunnari. Evaluating large language models in generating synthetic HCI research data: A case study. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–19. ACM, 2023. doi: 10.1145/3544548.3580688.

Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351, 2023.

Gati V. Aher, Rosa I. Arriaga, and Adam Tauman Kalai. Using large language models to simulate multiple humans and replicate human subject studies. In *International Conference on Machine Learning (ICML)*, 2023.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th ACM Symposium on User Interface Software and Technology (UIST)*, 2023.

Joon Sung Park, Carolyn Q. Zou, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Robb Willer, Percy Liang, and Michael S. Bernstein. Generative agent simulations of 1,000 people. *arXiv preprint arXiv:2411.10109*, 2024. Stanford/Google. 1,052 participants; 85% human self-replication accuracy.

Yuxuan Lu, Bingsheng Yao, Hansu Gu, Jing Huang, Jessie Wang, Yang Li, Jiri Gesi, Qi He, Toby Jia-Jun Li, and Dakuo Wang. UXAgent: A system for simulating usability testing of web design with LLM agents. *arXiv preprint arXiv:2504.09407*, 2025.

Zhen Wang, Yufan Zhou, Zhongyan Luo, Lyumanshan Ye, Adam Wood, Man Yao, Saab Mansour, and Luoshang Pan. DeepPersona: A generative engine for scaling deep synthetic personas. In *LAW Workshop, NeurIPS 2025*, 2025.

Julia Kharchenko, Tanya Roosta, Aman Chadha, and Chirag Shah. How well do LLMs represent values across cultures? empirical analysis of LLM responses based on hofstede cultural dimensions. In *Proceedings of KDD 2025*, 2025.

Priyanka Dey, Yugal Khanter, Aayush Bothra, Jieyu Zhao, and Emilio Ferrara. Can LLMs express personality across cultures? introducing CulturalPersonas for evaluating trait alignment. *arXiv preprint arXiv:2506.05670*, 2025.

Brihi Joshi, Xiang Ren, Swabha Swayamdipta, Rik Koncel-Kedziorski, and Tim Paek. Improving language model personas via rationalization with psychological scaffolds. *arXiv preprint arXiv:2504.17993*, 2025.

Wenkai Li, Jiarui Liu, Andy Liu, Xuhui Zhou, Mona T. Diab, and Maarten Sap. BIG5-CHAT: Shaping LLM personalities through training on human-grounded data. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2025a.

Tiancheng Hu and Nigel Collier. Population-aligned persona generation for social science research. *arXiv preprint*, 2025.

Shivani Kapania, Alex Siy, Torkil Clemmensen, Bonnie Nardi, and Robert Soden. The simulacrum of stories: Examining the gap between AI-generated and human-generated interview data. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 2025. Honorable Mention Award.

William Agnew, A. Stevie Bergman, Jennifer Chien, Mark Diaz, Seliem El-Sayed, Jaylen Pittman, Shakeri Tandon, Kira McKee, Tina Stolz, Vivek Srikumar, et al. The illusion of artificial inclusion. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2024.

Zhijing Lin. Six fallacies in using LLMs as substitutes for human participants. *arXiv preprint*, 2025.

William Agnew et al. Whose personae? transparency and bias in synthetic persona experiments. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 2025.

Ang Li, Haozhe Chen, Hongseok Namkoong, and Tianyi Peng. LLM generated persona is a promise with a catch. *arXiv preprint arXiv:2503.16527*, 2025b.

Virginia Braun and Victoria Clarke. Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2): 77–101, 2006.

Virginia Braun and Victoria Clarke. Reflecting on reflexive thematic analysis. *Qualitative Research in Sport, Exercise and Health*, 11(4):589–597, 2019.

Virginia Braun and Victoria Clarke. Can I use TA? should I use TA? should I not use TA? comparing reflexive thematic analysis and other pattern-based qualitative analytic approaches. *Counselling and Psychotherapy Research*, 21(1): 37–47, 2021.

Nielsen Norman Group. AI-generated personas: Practical utility and bias assessment. Technical report, Nielsen Norman Group, 2025.

Yuxuan Xu et al. Personacite: Voc-grounded interviewable agentic synthetic ai personas with source attribution. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, 2026.

Tejaswani Dash, Dinesh Karri, Anudeep Vurity, Gautam Datla, Tazeem Ahmad, Saima Rafi, and Rohith Tangudu. Polypersona: Persona-grounded LLM for synthetic survey responses. *arXiv preprint arXiv:2512.14562*, 2025.

Nelson Cowan. The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, 24(1):87–114, 2001.

Edward T. Hall. *Beyond Culture*. Anchor Press/Doubleday, Garden City, NY, 1976.

Alan Cooper. *The Inmates Are Running the Asylum: Why High-Tech Products Drive Us Crazy and How to Restore the Sanity*. Sams Publishing, 2004.

Everett M. Rogers. *Diffusion of Innovations*. Free Press, 5th edition, 2003.

John W. Creswell and Cheryl N. Poth. *Qualitative Inquiry and Research Design: Choosing Among Five Approaches*. Sage, 4th edition, 2018.

Shashank Gupta, Vaishnavi Shrivastava, Ameet Deshpande, Ashwin Kalyan, Peter Clark, Ashish Sabharwal, and Tushar Khot. Bias runs deep: Implicit reasoning biases in persona-assigned LLMs. In *Proceedings of the 12th International Conference on Learning Representations (ICLR)*, 2024.

Sunjae Kim and Eunsol Lee. Persona convergence in extended LLM dialogues. In *Proceedings of the ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW)*, 2024.

Klaus Krippendorff. Computing Krippendorff's alpha-reliability. *Departmental Papers (ASC)*, 2011. Standard reference for computing inter-rater reliability via Krippendorff's alpha.

James Bisbee, Joshua D. Clinton, Cassy Dorff, Brenton Kenkel, and Jennifer Larson. Synthetic replacements for human survey data? the perils of large language models. *Political Analysis*, 32(4):401–416, 2024.

Tanisha Jowsey, Virginia Braun, Victoria Clarke, et al. We reject the use of generative ai for reflexive qualitative research: A position statement. *Qualitative Inquiry*, 2025. Signed by 416 qualitative researchers.

Clifford Geertz. *The Interpretation of Cultures*. Basic Books, New York, 1973.