# GROUNDED SIMULATION: THE BEHAVIORAL SCIENCE LAYER FOR AI-SIMULATED RESEARCH

**Articos Research**
Articos Research

April 2026

**ABSTRACT**

Large language models can simulate research participants, but naive prompting produces low-fidelity results. We introduce a *methodologically-grounded simulation architecture* that anchors each layer of the simulation—persona generation, interview simulation, and analytical synthesis—in established behavioral science frameworks. A five-condition comparison experiment across 16 studies demonstrates F1 = 0.619 for theme recovery against published reference findings, compared to 0.082 for bare GPT, 0.065 for iterative GPT (10 conversational turns), 0.145 for role-prompted GPT, and 0.016 for a budget-matched baseline running 20× more compute—a 7.5× improvement over single-prompt and 39× 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 $0.815 \pm 0.052$, with all domains exceeding the 0.65 threshold. A volume-matched analysis confirms the result: even at the theoretical upper bound, budget-matched GPT recall (0.700) cannot reach the full system's 0.863. A blinded human evaluation by 23 practicing UX researchers (median 6 years experience) confirms these automated results: the full system 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 the full system's output nearly as often as expert-published findings (41% vs. 45% preference share; $p = .78$, binomial). Critically, 66% 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: removing stance diversity alone drops F1 by 94%. We present these results as a design principle: simulation fidelity is bounded by methodological grounding, not model capability.

***Keywords*** AI-simulated research · synthetic personas · UX research · methodologically-grounded simulation · research fidelity

## 1 Introduction

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 a skilled 10-turn iterative conversation—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.

This paper demonstrates that anchoring each simulation layer in behavioral science closes that gap. Our system generates synthetic participant profiles grounded in Big Five personality dimensions with facet-level annotation [Costa and McCrae, 1992, Jiang et al., 2024, Hu et al., 2024], conducts hypothesis-blind interviews that structurally prevent sycophancy [Sharma et al., 2024], and 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 $0.815 \pm 0.052$. 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 66% 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 proportional to the degree of methodological grounding in established science, not model capability. The principle builds on the tradition of theory-driven design [Carroll, 1991, Hevner et al., 2004] and is supported by converging evidence—Park et al. [2024] achieved 85% self-replication accuracy with interview-grounded agents, Hu et al. [2024] showed facet-level personality explains 81% of response variance, and Dunivin [2024] achieved $\kappa \geq 0.79$ for AI-assisted qualitative coding. The common thread is that structured grounding, not raw capability, drives fidelity.

Our contributions: (1) the Grounded Simulation Principle as a design principle for AI simulation systems; (2) a complete architecture implementing it across persona generation, interview simulation, and analytical synthesis; (3) a five-condition experiment demonstrating 7.5× F1 improvement with volume-matched analysis confirming the result; (4) a cross-domain validation with a six-dimension Research Fidelity Index across 46 studies in 9 domains, demonstrating generalization beyond the comparison subset; (5) a blinded human evaluation by 23 practicing UX researchers confirming the automated metrics—with 66% of evaluators unable to distinguish system output from expert-generated research; and (6) a component ablation study identifying the relative contribution of each architectural layer.

## 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 behaviors in 25 persistent-memory agents. Park et al. [2024] scaled to 1,052 individuals, achieving 85% self-replication accuracy.

**Persona consistency and personality.** Jiang et al. [2024] validated that LLMs maintain personality signatures at domain and facet levels. Hu et al. [2024] showed personality traits explain up to 81% of response variance. Salminen et al. [2025] introduced failure detection for persona coherence. 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. 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 (§4.2).

### 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 [Chen 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.

**Positioning.** Prior work validates specific techniques in isolation or critiques AI simulation broadly. No published work presents a complete, integrated simulation architecture with systematic comparison against ungrounded baselines and cross-domain validation across 9 research domains.

## 3 The Grounded Simulation Principle

We formalize the core argument as a design principle:

> *The fidelity of an AI-simulated research study is proportional to the degree to which each layer of the simulation is anchored in an established scientific framework. Simulation quality is bounded not by model capability but by methodological grounding—the systematic application of validated theories from personality psychology, cognitive science, and qualitative research methodology to constrain and structure the simulation's behavior.*

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.[1]

## 4 Architecture

The architecture implements the GSP through three subsystems (Figure 1), each anchored in an established scientific framework.

### 4.1 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.

### 4.2 Persona Generation

The system generates *synthetic participant profiles*—theory-constructed agents distinct from data-grounded personas. Each profile composes multiple behavioral science layers: Personality dimensions with facet-level behavioral annotation informed by validated psychometric frameworks [Costa and McCrae, 1992, Jiang et al., 2024, Hu et al., 2024];

[1] The principle has three corollaries. **Layered Composition**: participants are compositions of personality, cognition, social roles, and culture; each layer must be grounded independently. **Failure Mode Specificity**: each ungrounded layer produces a predictable failure—variance compression, unrealistic rationality, sycophancy—with engineering solutions from the relevant sciences. **Honest Boundaries**: some aspects of human experience (embodied sensation, genuine emotion) cannot be grounded in frameworks an LLM can operationalize; the grounded approach requires mapping where grounding is and is not possible.

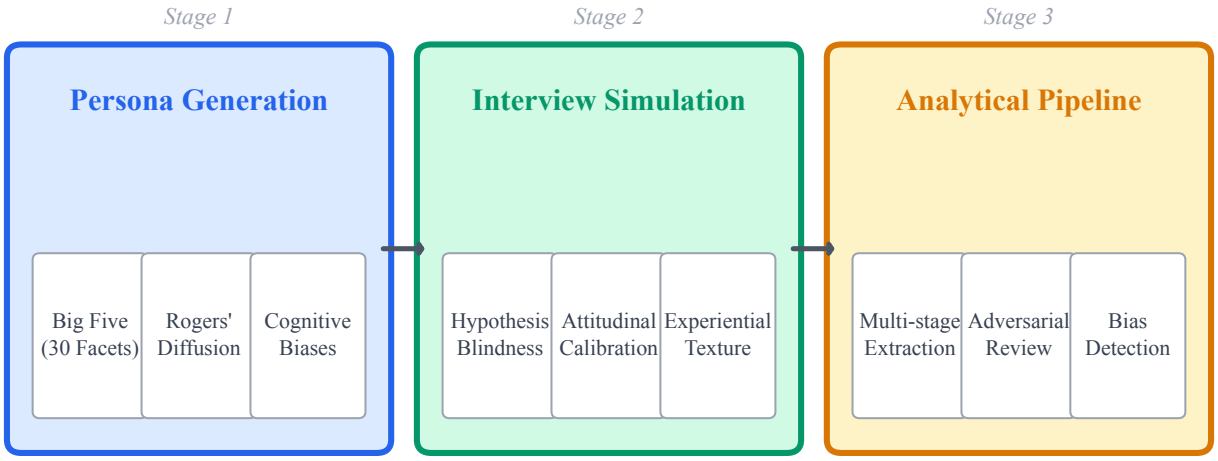


Figure 1: The Grounded Simulation Architecture. Each subsystem is anchored in an established scientific framework. Arrows indicate sequential pipeline flow from persona generation through interview simulation to analytical synthesis.

cognitive biases annotated by personality and attitudinal stance [Gupta et al., 2024]; diversity enforced at the cohort level through structured adoption stance categories [Rogers, 2003]; and domain-specific behavioral attributes grounded in established interaction design methodology [Cooper, 2004]. The key architectural choice is that diversity is enforced *by construction* through pre-generation coordination, not hoped for post hoc.

### 4.3 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]. This is enforced at the architecture level—not by prompt instruction but by information isolation—making it robust to prompt injection or 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 an episodic memory system grounded in the ACT-R cognitive architecture's declarative memory module [Anderson et al., 2004]. Memory activation follows the base-level learning equation $A_i = \ln(\sum t_j^{-d}) + \text{cue overlap} + \varepsilon$, where retrieval probability is computed via logistic threshold. Working memory capacity is constrained to approximately four items (consistent with Cowan's capacity limit), and satisficing behavior terminates retrieval when activation exceeds threshold rather than exhaustively searching memory. This ensures personas exhibit realistic memory decay, interference effects, and bounded recall—a participant who discussed delivery frustration in question three may reference it in question eight with appropriate detail loss, but will not contradict themselves.

### 4.4 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 that retrieves current published findings, industry reports, and recent studies for each extracted theme—grounding the analysis in real-world evidence beyond the model's training data, (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). The web research stages are critical: they enable the system to discover findings published after the model's training cutoff and to cross-reference AI-generated themes against the latest empirical evidence, partially addressing the training data contamination concern.

# 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 $\geq 0.55$. 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 §6.7). 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 $\geq 0.80$. 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.

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]

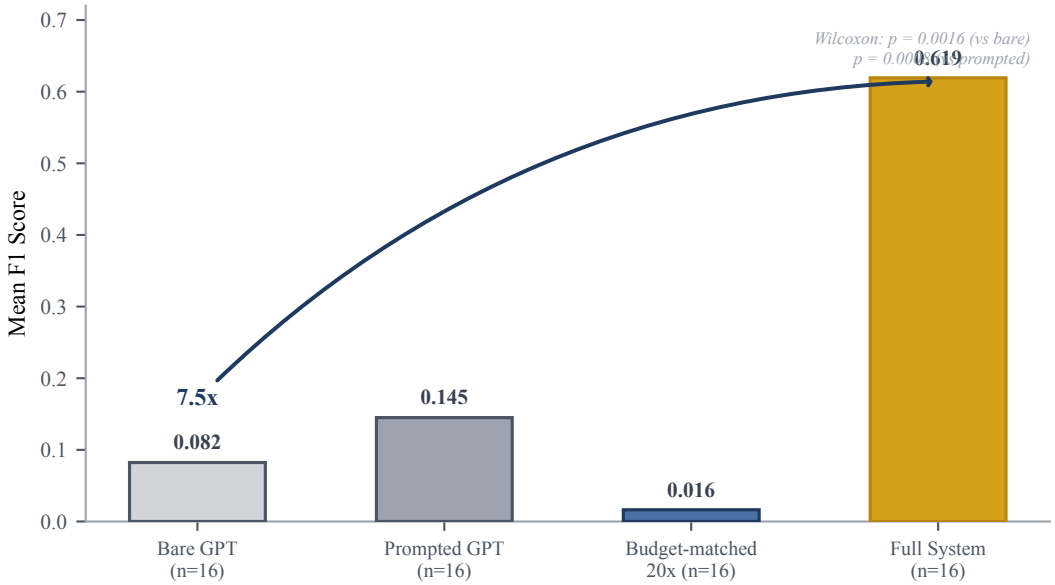


Figure 2: F1 scores across four conditions (iterative GPT omitted at F1 = 0.065; see Table 1). The full grounded architecture achieves 7.5× the F1 of bare prompting. Methodological grounding—not compute—drives quality.

### 6.2 Results

### 6.3 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 $\geq 0.80$ 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.

### 6.4 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 mecha-

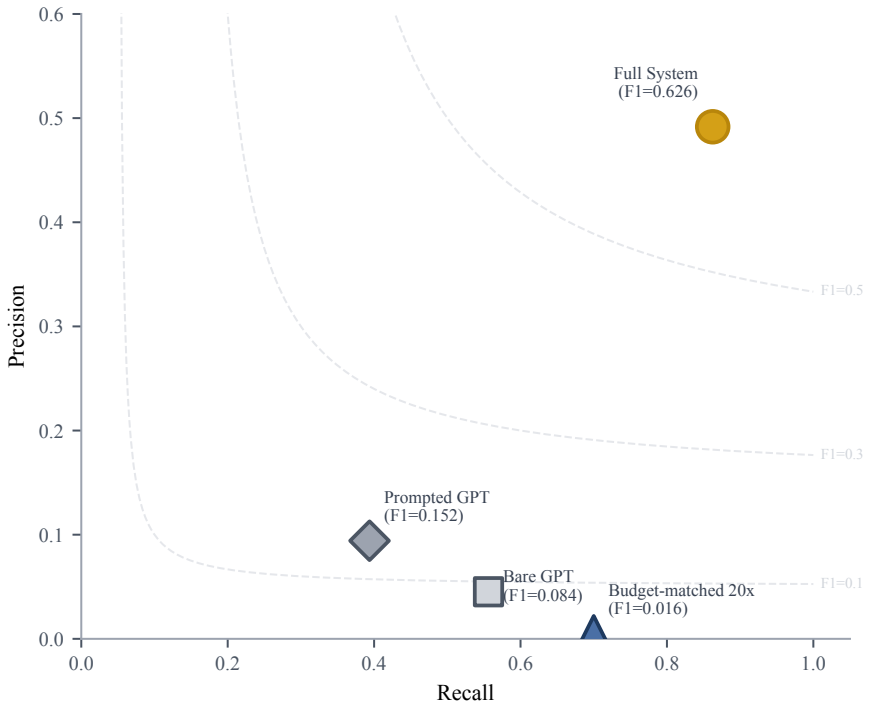


Figure 3: Precision–recall space. Bare and budget-matched GPT achieve moderate recall through volume but catastrophic precision. The full system occupies a fundamentally different region.

nisms: (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. We note that the recall difference is partially confounded by output volume (bare GPT generates 2.2× more themes) and that the finding describes LLM prompt sensitivity, which may 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.

### 6.5 Per-Study Analysis

Table 2 presents per-study results grouped by domain. 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.

### 6.6 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.

### 6.7 Cross-Domain RFI

The system achieves mean RFI = $0.815 \pm 0.052$ across 46 studies spanning 9 domains (Table 3). The component profile (Figure 5) 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: Per-study F1 scores across conditions, grouped by domain.

| 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 |

Table 3: 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** |

### 6.8 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.

#### 6.8.1 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?

Table 4: 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 |

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?"

### 6.8.2 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.

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** |

The Full System–Reference gap is statistically significant ($p = .004$, Wilcoxon signed-rank), confirming that human evaluators perceive a quality difference. However, the Full System–Prompted GPT gap is substantially larger ($\Delta = 1.17$, $p < .001$), and 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. However, 15 of 23 evaluators (66%) 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.

**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.

### 6.8.3 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.

### 6.9 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 (Table 6).

Table 6: Component ablation: F1 change when each component is removed.

| Condition | Recall | F1 | $\Delta$ 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 | $-.565$ |

All four grounding components contribute substantially. Stance diversity has the largest impact ($\Delta$F1 = $-.582$): 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—which are 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. We note that 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 Limitation 7).

## 7 Discussion

### 7.1 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 6) resolves a question the architecture comparison cannot: *which* components matter. All four grounding layers contribute substantially, with stance diversity ($\Delta$F1 = $-.582$) and adversarial review ($-.565$) having the largest impact. The hypothesis blindness validation confirms there is no data leakage in our evaluation protocol.

**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. However, 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. However, 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.

### 7.2 Limitations

1. **Comparison–validation scope gap.** The five-condition comparison experiment covers 16 studies in 2 domains (e-commerce, SaaS); extended validation across 46 studies in 9 domains confirms generalization, but comparison baselines have not been re-run on the additional 30 studies. Cross-cultural research remains the weakest domain (RFI 0.776).
2. **Model dependency.** All studies use GPT-class models. Results may not transfer to other families; model updates affect reproducibility [Bisbee et al., 2024].
3. **Author-as-evaluator.** The automated evaluation framework, reference selection, and scoring were designed and executed by the system's developer. The blinded human evaluation (§6.8) with 23 independent evaluators substantially mitigates this concern for perceived quality, though the automated metrics remain author-designed. Independent replication of the automated evaluation protocol is encouraged.
4. **Human evaluation scope.** While 23 evaluators provide adequate statistical power for the quality comparison ($p = .004$), the evaluation covered 4 of 46 studies. Broader coverage across all domains would strengthen generalizability claims.
5. **Automated matching.** Semantic similarity matching may miss or create false matches; the 0.55 threshold is a design choice. A cross-ecosystem validation using sentence-transformers (all-MiniLM-L6-v2, a non-OpenAI model) confirms that system-generated themes match reference findings at comparable recall levels when the threshold is recalibrated for the different embedding space (83.3% recall at threshold 0.25 vs. 85.3% at 0.55 with OpenAI), providing evidence that the matching results are not an artifact of within-ecosystem consistency.
6. **LLM-judged novelty.** CNS uses an LLM to classify novel themes, creating circular dependency. Human validation would strengthen this component. A weight sensitivity analysis confirms the composite RFI is robust to moderate weight variations: perturbing any weight by $\pm 10$ percentage points shifts mean RFI by at most 2.0% (range 0.798–0.831 around the original 0.815).
7. **Systemic circularity.** Data generation, analysis, and evaluation components all use LLM processing. Our validation primarily measures internal consistency within the LLM ecosystem.
8. **Reference contamination.** Published findings are likely in training data. We address this directly with a post-cutoff decontamination study (Section 7.3).

### 7.3 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 7 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).

Table 7: 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** | $\Delta$ |
|---|---|---|---|
| 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* | $-.162$ |

Critically, 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.

### 7.4 Validity Spectrum

Table 8 maps research question types to observed confidence levels based on the 46-study corpus. Most categories now have empirical support; two remain projected.

Table 8: Validity spectrum across 46 studies. Upper rows are empirically grounded; lower rows are projected.

| **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 |

### 7.5 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—not 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; Low and Very Low Confidence domains require human research as the primary method.

## 8 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 draws on the same personality psychology, cognitive science, 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 = $0.815 \pm 0.052$, with every domain exceeding the 0.65 threshold and the system recovering 86% of themes from published secondary research. We emphasize that 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 7.2). A 10-turn iterative conversation—the strongest practitioner workflow—performs *worse* than a single prompt (F1 0.065 vs. 0.082), demonstrating that conversational depth without structural grounding narrows 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 66% 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.

## Data Availability

All experimental data, evaluation scripts, baseline prompt variations, and generated theme sets for all conditions are available at `https://github.com/articos-research/grounded-simulation`.

## References


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, Nigel Collier, and Hinrich Schütze. Quantifying the persona effect in LLM simulations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024.

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.

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.

Joon Sung Park, Carolyn Zou, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, 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.

Zack O. Dunivin. Scalable qualitative coding with LLMs: Chain-of-thought reasoning matches human performance. *Proceedings of the National Academy of Sciences*, 121(39), 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.

Joni Salminen, Soon-Gyo Jung, and Bernard J. Jansen. PersonaCraft: Personalized full-body persona generation with narrative coherence. *International Journal of Human-Computer Studies*, 2025.

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.

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.

Valentin Chen et al. Polypersona: Persona-grounded llms for synthetic survey responses. *arXiv preprint arXiv:2512.14562*, 2025.

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

Vikas Gupta, Purna Narayanan Venkit, Shomir Wilson, and Rebecca J. Passonneau. Bias runs deep: Implicit reasoning biases in persona-assigned llms. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. ACM, 2024.

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

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

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.

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.

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.

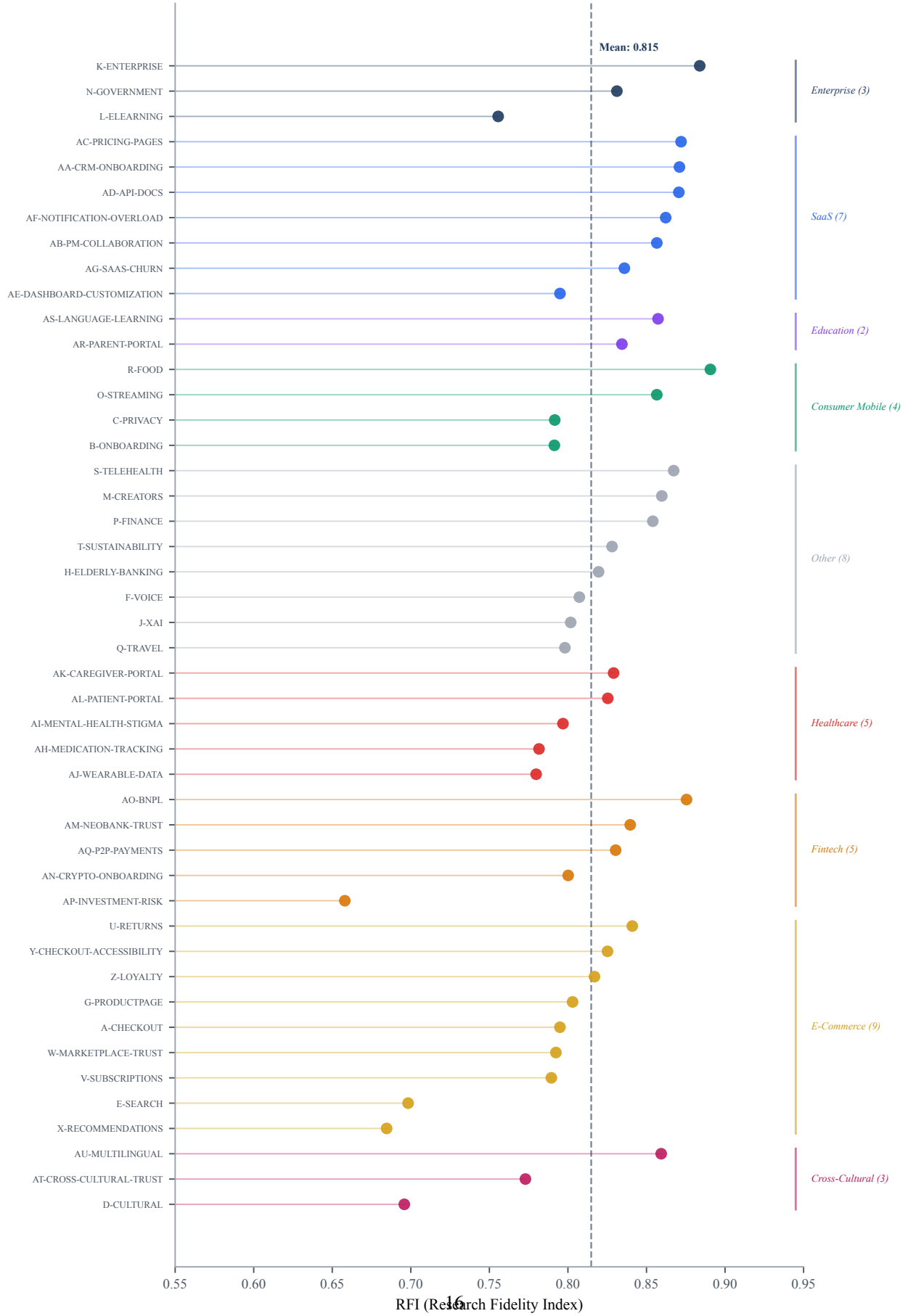


Figure 4: Per-study RFI across 46 studies grouped by domain. Studies are colored by domain and sorted by RFI within each group. The dashed line indicates the corpus mean (0.815).

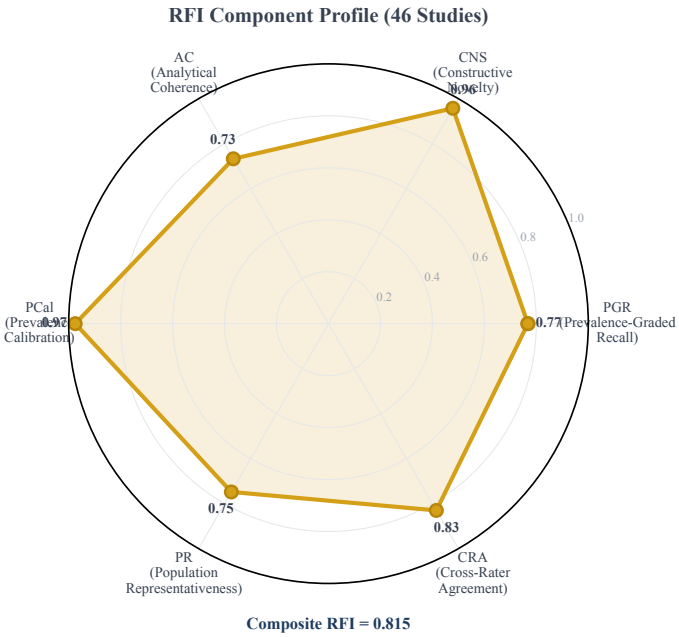


Figure 5: RFI component profile across 46 studies. PCal and CNS are strengths; AC is the weakest dimension.

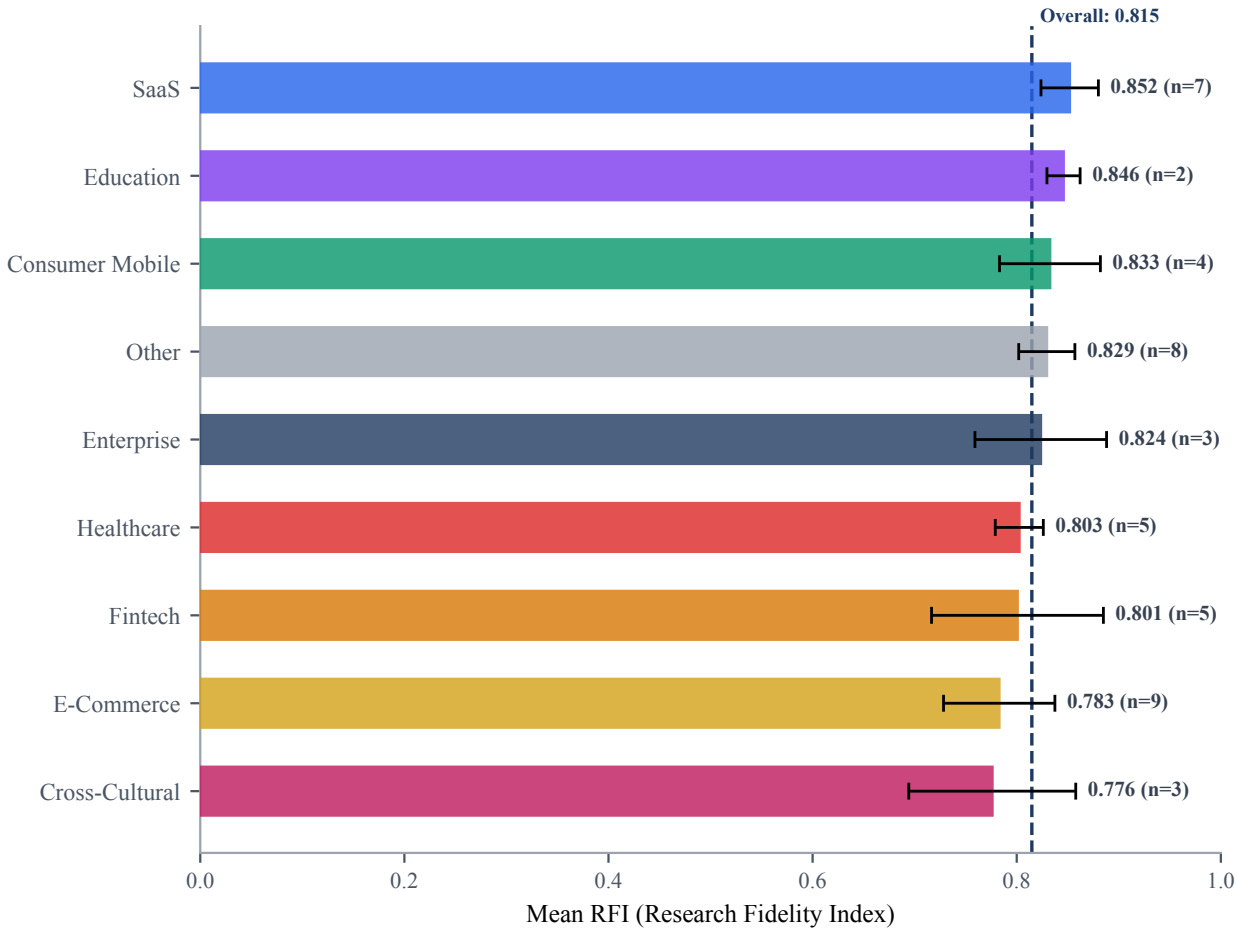


Figure 6: Mean RFI by domain across 46 studies. Error bars show one standard deviation. Dashed line indicates the overall mean (0.815).

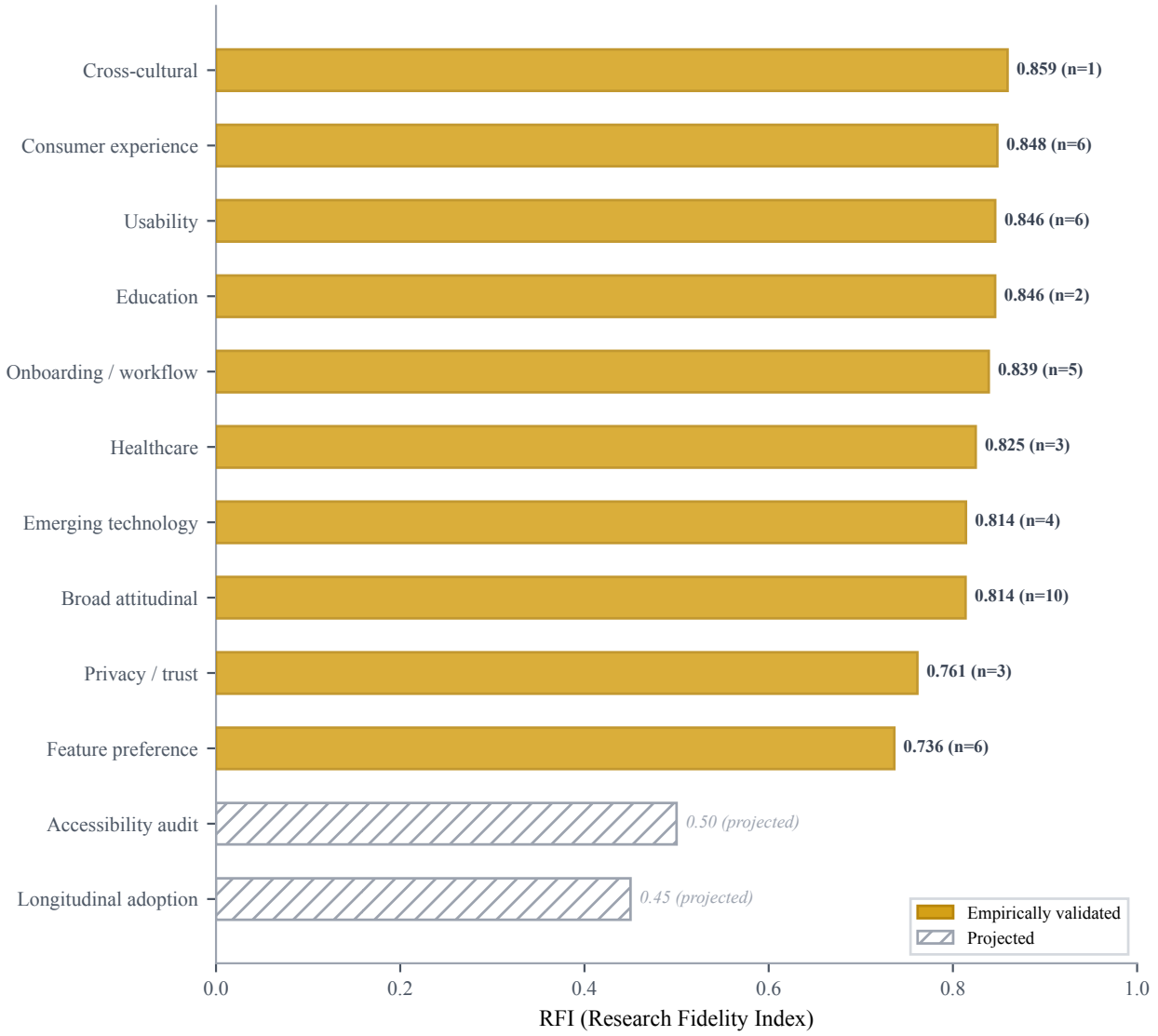


Figure 7: Validity spectrum across 46 studies. Empirically validated question types (solid) show observed mean RFI; projected categories (hatched) represent theoretical expectations awaiting validation.