SAPIENS: Modular Human Behavior Simulation via Latent Tribe Discovery and Semantic Gradient Optimization

Anonymous ACL submission

Abstract

LLMs are increasingly used to simulate the behavior of specific personas (tribes) to test reactions to different stimuli. While these models can represent a tribe, producing reliable simulations often depends on providing an accurate tribe description and manually specifying traits and backstory. Prior work in human behavior simulation is limited in the automated discovery of new tribes and their latent traits, creating a bottleneck in building rich representations and simulating behaviors across diverse population segments. We introduce SAPIENS, a framework that automatically discovers latent tribes from sparse user observation data and learns their latent behavioral traits through an iterative simulator–critique learning loop that minimizes a behavior loss measuring the gap between real and generated opinions. On 30,000 Amazon reviews across seven domains, SAPIENS achieves 82.63% Recall@k on unseen user-product pairs, significantly outperforming prompt-based and fine-tuned baselines. SAPIENS can learn tribes and their traits from large-scale unstructured opinion data to power accurate, traceable and scalable human behavior simulation.

1 Introduction

Human behavior does not arise in isolation. A person's response to any stimulus reflects three things simultaneously: what their community does and values, the specific ways they personally depart from that community, and what their own prior encounters with the stimulus domain have taught them to expect. We posit that these three components: shared group-level priors, individual divergences from those priors, and episodic memory of prior stimulus interactions, are separable, independently learnable, and that composing them produces more accurate behavioral predictions than any entangled mixture of them can achieve. If this decomposition is correct, then a person's reaction to a novel stimulus is predictable even for preferences they have never explicitly expressed: the tribe has expressed them, and the individual's departure from the tribe has been observed elsewhere. This structure finds support in foundational social-science theory: Social Identity Theory (Tajfel and Turner, 1979) establishes that group membership is a primary organizer of human behavior, and Trait Activation Theory (Tett et al., 2021) establishes that individual traits are selectively activated by situationally relevant cues rather than expressed uniformly, but neither theory assembles the components into a learnable generative model.

LLMs have made this testable at scale (Park et al., 2023, 2024; Argyle et al., 2023; Zhao et al., 2025; Rothschild et al., 2024; Arora et al., 2024; Wang et al., 2023, 2025b; Li et al., 2025), and the results are promising. But no existing approach implements the decomposition: demographic prompting (Wang et al., 2025a; Thakur et al., 2025) collapses group prior and individual divergence into one undifferentiated description; backstory conditioning (Moon et al., 2024) captures individual surface patterns but cannot impute preferences the user has never expressed; and parametric fine-tuning, Be.FM (Xie et al., 2025) and Centaur (Binz et al., 2024), conflates all three components inside model weights, making them inseparable and unauditable. The consequences are measurable: sociodemographic biases overwrite minority views under explicit persona prompting (Gupta et al., 2024; Sun et al., 2025), prompt sensitivity produces unnaturally low-entropy distributions (Dominguez-Olmedo et al., 2024; Gao et al., 2025), and social desirability bias pulls responses toward consensus regardless of persona specification (Salecha et al., 2024). We find empirically that purpose-built behavioral foundation models *underperform simple backstory prompting* on sparse users (68–70% test Recall against 72.87% for a plain prompted persona), because weight-level priors collapse toward

the population mean, which is actively detrimental to the idiosyncratic individuals representing the majority of real-world users. The root cause is structural: **sparsity** (individual histories too thin without a group prior to draw from) and **opacity** (identity components entangled inseparably in weights, behavioral drivers invisible, individual updates intractable) are both direct consequences of not implementing the decomposition.

SAPIENS (Figure 1) implements this decomposition via a **Modular Persona Schema**: a strictly typed, context-space data structure that steers a frozen foundation model without parameter updates. Rather than entangling identity in model weights, SAPIENS isolates the user's demographic group prior, their historical tendencies, and their idiosyncratic deviations into discrete, inspectable variables. To populate this schema, we introduce **Semantic Gradient Optimization (SGO)**, an iterative optimization loop that discovers unobserved latent traits by analyzing prediction failures and applying natural language verbal gradients to the user's schema.

Across 30,000 interactions in seven domains, SAPIENS achieves 82.63% Recall@k on unseen user–product pairs, outperforming state-of-the-art fine-tuned behavioral models by 12–14 points.

2 Related Work

Parametric Behavioral Models. Be.FM (Xie et al., 2025) and Centaur (Binz et al., 2024) fine-tune foundation models on large-scale behavioral observation data to encode human decision-making regularities parametrically. Be.FM targets general behavioral regularities across diverse domains; Centaur targets the computational structures of human cognition derived from psychological decision-making datasets. Both demonstrate that behavioral fine-tuning shapes surface-level outputs and claim general-purpose applicability to behavioral simulation. SAPIENS differs by externalizing behavioral identity into modular context vectors rather than encoding it in weights, enabling individual-level auditability and targeted updates without retraining.

Persona-Based and Narrative-Grounded Simulation. Generative Agents (Park et al., 2023) showed that LLMs conditioned on rich qualitative persona descriptions generate coherent, contextually consistent behavioral trajectories in interactive environments. Park et al. (2024) extended this to 1,000 interview-grounded agents, approaching human test–retest reliability in attitudinal tasks but requiring two-hour qualitative interviews per simulated person. The Anthology framework (Moon et al., 2024) reduces data collection cost by synthesizing narrative backstories from limited behavioral information, enabling persona conditioning without active interviews. Persona Hub (Ge et al., 2024) scales in the orthogonal direction, generating billion-scale synthetic personas to maximize prompt-space diversity for synthetic data creation rather than modeling specific real individuals. Thakur et al. (2025) fine-tune small language models with low-rank adapters to mimic individual user behaviors at the parametric level. SAPIENS learns persona components from passive behavioral traces rather than active interviews or synthetic generation, and augments individual conditioning with a statistically grounded group prior to handle sparse users.

3 Methodology

SAPIENS operates in two phases: Tribe Discovery, and Semantic Gradient Optimization (SGO), to learn a modular persona schema $\mathcal{P}_i$ that steers a frozen foundation model at inference time without parameter updates (Figure 1).

3.1 Problem Formulation: The Modular Persona Schema

We formalize individual identity as a **Modular Persona Schema**: a structured context-space representation that modifies the behavior of a frozen foundation model without altering its weights. For a user u_i, we define a synthetic persona $\mathcal{P}_i$ as the triplet:

$$\mathcal{P}_i = (V_{\text{tribe}},\ M_{\text{semantic}},\ \Delta_{\text{behavior}}) \qquad (1)$$

- V_{tribe}**: Tribal Characteristics.** The shared behavioral norms, values, and response tendencies of the latent group to which u_i belongs, as a structured natural-language schema with six traits: persona summary, key motivations, common praises, common criticisms, core characteristics, and potential goals. Seeded from data in Phase 1 and refined through SGO in Phase 2; evolves as the Critic identifies which tribal elements are systematically over-generalized. EVIDENCE: *the schema for Micro-Tribe 0 ("Pragmatic Value Advocates") defines* `key_motivations` *as "ensuring fair*

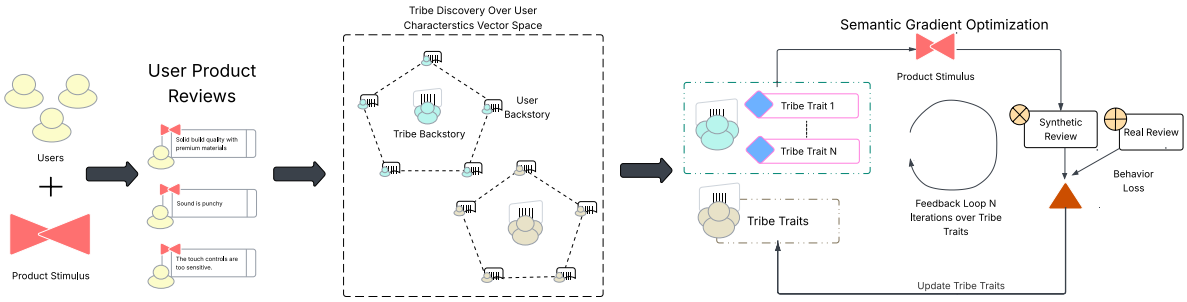

Figure 1: **The SAPIENS Architecture.** The simulation lifecycle of an individual user: from recording their sparse observations and mapping them to a foundational tribe, to inferring their idiosyncratic traits under different stimuli until behavioral loss convergence. Once the system locks in their explicit schema: their tribe, traits, and episodic memory; a SAPIEN is born, ready to react with high fidelity under any novel stimulus. Full trait ledgers and representative tribe ontologies are provided in Appendix C (Figure 4).

pricing," "securing accurate fit," and "avoiding technical glitches" (see Appendix C, Figure 4).

- M_{semantic}**: Semantic Memory.** A compressed natural-language summary of the individual's cross-domain behavioral tendencies, distilled from their review corpus via LLM. Captures the idiosyncrasies that distinguish u_i from their tribe. For users with multi-category histories, per-category summaries are generated separately and composed, to maintain objectivity in criticism lenses across domains. M_{semantic} does not change during SGO. EVIDENCE: *a user's $M_{semantic}$ preserves their historical footprint verbatim: "This user writes reviews to guide fellow consumers with candid, experience-based insights... They value accurate representation, functional fit, and overall usability" (Appendix C, Figure 4).*

- Δ_{behavior}**: Behavioral Deviation Cues.** A dynamic set of activation cues encoding the user's specific departures from the tribal norm, grounded in Trait Activation Theory (Tett et al., 2021). Initialized empty ($\Delta^{(0)}_{\text{behavior}} = \emptyset$) and discovered entirely through the SGO loop; strictly user-specific unlike V_{tribe}.

Together: V_{tribe} provides a statistically grounded prior for sparse users; M_{semantic} injects individual coloring from actual behavioral traces; Δ_{behavior} captures fine-grained deviations that emerge only when the prior fails.

3.2 Phase 1: Behavioural Tribe Discovery

Phase 1 partitions the user population into latent behavioral **tribes** so that sparse users can inherit statistically grounded behavioral signals from their group rather than relying solely on their thin individual histories.

User Embedding. For each user u_i, an LLM, using their full review history generates a **User Backstory**: a compact description of the core personal characteristics that influence *why* this user writes as they do, underlying motivations and values, not surface lexical patterns. *For instance, condensing a history of detailed product reviews into a core behavioral disposition, such as the "consumer-advocate mindset" identified in Appendix C, Figure 4.* This summary is encoded using `text-embedding-3-small`, yielding $\mathbf{v}_i \in \mathbb{R}^{1536}$: a projection into a space defined by behavioral disposition rather than textual similarity. Two users who write in very different styles but share the same value system will be proximate here even if their reviews look nothing alike.

Manifold Approximation and Tribe Assignment. The embedding space is compressed via UMAP (McInnes et al., 2018), then clustered with Iterative K-Means sweeping $K \in [4, 10]$, selecting the K that maximizes Silhouette Coefficient ($S_{\text{sil}} > 0.4$) subject to a Balance Ratio constraint ($R_{\text{bal}} > 0.1$). A second round of sub-clustering within each macro-cluster yields **micro-tribes**. Each user is assigned to the micro-tribe whose centroid minimizes cosine distance to $\mathbf{v}_i$, inheriting the group's rich imputed behavioral signals and bypassing the cold-start problem for sparse users. EVIDENCE: *this mechanism correctly imputes a user's unexpressed preference for travel-friendly packaging entirely from their assigned tribal prior (Appendix A, Table 4).* In our evaluation, this yielded 6 macro-clusters resolved into 150 micro-tribes across 5,002

users (Appendix F, Tables 10, 11).

Tribe Seed Characterisation. For each micro-tribe C_k, an LLM synthesizes member characteristic summaries and representative reviews into an initial $V_{\text{tribe}}^{(0)}$: a structured schema encoding the tribe's observable behavioral traits and latent motivational traits (full schema and representative instantiations in Appendix C, Figure 4). These traits map directly to the constructs of Trait Activation Theory (Tett et al., 2021). The motivations and friction points become the scaffolding onto which the SGO Critic attaches its verbal gradients in Phase 2. EVIDENCE: *the friction point "misleading descriptions or hiddenfees," discovered from 144 real users in Micro-Tribe 0, becomes an explicit schema constraint that the Critic can activate or prune during SGO (Appendix C, Figure 4).*

3.3 Phase 2: Semantic Gradient Optimization (SGO)

Given the initialized persona $\mathcal{P}_i^{(0)} = (V_{\text{tribe}}^{(0)}, M_{\text{semantic}}, \emptyset)$, SGO iteratively refines both V_{tribe} and Δ_{behavior} via a Simulator–Critic loop minimizing a composite Behavioral Delta Loss. SGO is analogous to EM-style refinement (Cappé and Moulines, 2009): the Critic infers latent behavioral traits from observed prediction failures (E-step) and the Simulator updates the generative context using those traits (M-step), operating entirely in natural language rather than probability distributions.

3.3.1 Behavioral Delta Loss

We optimize a composite loss balancing semantic style with thematic accuracy:

$$\mathcal{L}(\hat{y}, y) = \alpha \cdot \mathcal{L}_{\text{sem}}(\hat{y}, y) + (1 - \alpha) \cdot \mathcal{L}_{\text{theme}}(\hat{y}, y) \tag{2}$$

Semantic Loss $\mathcal{L}_{\text{sem}}$ is the cosine distance between `text-embedding-3-small` embeddings of $\hat{y}$ and y, capturing high-level stylistic and sentiment alignment.

Thematic Loss $\mathcal{L}_{\text{theme}}$ is the Jensen–Shannon Divergence between the real and synthetic theme distributions:

$$\begin{aligned}\mathcal{L}_{\text{theme}} &= \text{JSD}(Q_{\text{real}} \parallel P_{\text{syn}}) \\ &= \frac{1}{2} D_{\text{KL}}(Q_{\text{real}} \parallel M) + \frac{1}{2} D_{\text{KL}}(P_{\text{syn}} \parallel M)\end{aligned} \tag{3}$$

where $M = \frac{1}{2}(Q_{\text{real}} + P_{\text{syn}})$, computed base-2 with $\epsilon = 10^{-10}$ smoothing. Theme universes are category-specific (6–9 themes per domain, 51 total; Appendix F, Table 9). Crucially, rather than relying on discrete string-matching to check if the simulation discussed the correct topics, SAPIENS extracts the top-K log probabilities directly from the LLM's classification tokens. This forms a continuous, differentiable probability distribution P_{syn} over the defined topic universe, enabling a rigorous divergence calculation.

The mixing coefficient $\alpha = 0.7$ was determined by hold-out tuning. Only the **thematic component** drives the stopping criterion: the SGO loop continues until average batch theme delta falls below ϵ, as thematic coverage is harder to impute from tribal priors and more consequential for behavioral fidelity than stylistic tone.

3.3.2 The SGO Algorithm

Formalized in Algorithm 1, SGO functions as a highly concurrent Expectation-Maximization (EM) loop. To prevent cross-contamination of latent traits, interactions are strictly partitioned by macro-domain before being dispatched to a parallel execution pool. Because semantic gradients in latent space are non-convex, SAPIENS implements an **Asynchronous Trajectory Rollback**. Rather than strictly gating global memory updates and bottlenecking throughput, the system maintains a ledger of all inference states ($\mathcal{J}$) and retroactively selects the simulation output that mathematically minimized the divergence. In each iteration, $\hat{y}_{i,t-1}$ denotes the simulated review generated at the previous iteration, y_i is the ground-truth review authored by user u_i, and τ is a fixed per-sample threshold above which a prediction is considered a failure and passed to the Critic. $\mathcal{L}$ refers to the composite behavioral loss defined in Eq. 2.

Part B (Critic) identifies which characteristics, motivations, or friction points in the actual review were absent from or misweighted in the prediction, returning structured verbal gradients in the format `[Trait 1, Trait 2] [root cause]`. EVIDENCE: *a verbatim Critic gradient reads: "Prediction skipped coin-loss glitch because persona lacks trait that user flags precise resource discrepancies," triggering discovery of the schema constraint "[−] Unfair in-game economy or currency loss" (Appendix D, Table 6).* A separate high-capability model serves as Critic, distinct from the generation backbone, reflecting the analytical nature of the task.

Algorithm 1 Semantic Gradient Optimization (SGO)

Input: Batch $\mathcal{B} = \{(x_i, y_i)\}$, where x_i: product stimulus; y_i: ground-truth review by user u_i; $\hat{y}_{i,t-1}$: simulated review at iteration $t-1$; τ: per-sample threshold above which prediction is passed to Critic; $\mathcal{L}$: composite behavioral loss (Eq. 2); $P_0 = (V_{\text{tribe},0},\ M_{\text{semantic}},\ \emptyset)$; convergence threshold ϵ
Initialize: $\Delta_0 \leftarrow \emptyset,\ V_{\text{tribe},0} \leftarrow V_{\text{tribe}},\ t \leftarrow 1$
while $\bar{\mathcal{L}} > \epsilon$ **do**
  for (x_i, y_i) in $\mathcal{B}$ **do**
    $\hat{y}_{i,t-1} \leftarrow \text{LLM}(x_i, P_{t-1});\quad \ell_i \leftarrow \mathcal{L}(y_i, \hat{y}_{i,t-1})$
  end for
  $\bar{\mathcal{L}} \leftarrow \frac{1}{|\mathcal{B}|}\sum_i \ell_i$
  $\mathcal{G} \leftarrow \emptyset$
  for i **where** $\ell_i > \tau$ **do**
    $g_i \leftarrow \text{Critic}(\hat{y}_{i,t-1}, y_i, \text{“Extract latent cause”});\ \mathcal{G} \leftarrow \mathcal{G} \cup \{g_i\}$
  end for
  $\Delta', V'_{\text{tribe}} \leftarrow \text{Consolidate}(\mathcal{G});\quad P' \leftarrow P_{t-1} \cup \Delta' \cup V'_{\text{tribe}}$
  $\Delta_{\text{kept}} \leftarrow \Delta_{t-1};\quad V_{\text{tribe,kept}} \leftarrow V_{\text{tribe},t-1}$
  for (x_i, y_i) in $\mathcal{B}$ **do**
    $\hat{y}'_i \leftarrow \text{LLM}(x_i, P')$
    if $\mathcal{L}(y_i, \hat{y}'_i) < \ell_i$ **then**
      $\Delta_{\text{kept}} \mathrel{+}= \{g \in \Delta' \mid g \text{ relevant to } i\};\ V_{\text{tribe,kept}} \mathrel{+}= \{g \in V'_{\text{tribe}} \mid g \text{ relevant to } i\}$
    end if
  end for
  $P_t \leftarrow P_0 \cup \Delta_{\text{kept}} \cup V_{\text{tribe,kept}};\quad t \leftarrow t+1$
end while
Output: Optimized persona $P^* = P_t$

Part A (Simulator) generates revised predictions from the updated persona schema and product description. Critically, the Simulator does **not** have access to the actual review text, so it must infer the correct behavioral response entirely from the persona schema and stimulus. This constraint ensures the schema, not surface memorization, drives improvement. Accumulated Critic analyses from all prior batches are injected as “past learnings,” allowing patterns identified across many reviews to collectively constrain future predictions. Accumulated analyses from all prior batches are injected as historical context, collectively preventing the Simulator from repeating the same stereotyping errors. EVIDENCE: *a hallucinated ingredient critique is suppressed as accumulated context establishes that this user does not flag ingredient concerns despite their tribe doing so (Appendix A, Table 4).*

Best Prediction Selection identifies the iteration minimizing thematic delta: $\hat{y}^* = \arg\min_t \mathcal{L}^{(t)}_{\text{theme}}$. This approach remains inherently conservative, reverting to the initial prediction if all refinements increase thematic divergence. Monotonic decays in loss curves, global convergence trajectories, and per-tribe metrics are detailed in Appendices B and E.

3.4 Inference

At inference, the optimized persona $\mathcal{P}^*_i = (V^*_{\text{tribe}},\ M_{\text{semantic}},\ \Delta^*_{\text{behavior}})$ is injected alongside product stimulus x into the foundation model's context window, structured hierarchically: tribal characteristics (V^*_{tribe}) provide the normative behavioral frame; semantic memory (M_{semantic}) adds individual coloring; behavioral deviation cues ($\Delta^*_{\text{behavior}}$) apply the idiosyncratic corrections discovered during SGO.

3.5 Explainability, Privacy, and the Glass Box

Every simulated response is causally traceable to named schema elements (specific motivations activated, friction points triggered, etc.) because the schema is the *only* mechanism by which the frozen model receives behavioral information, making the **Explainability Constraint** an architectural property rather than a post-hoc attribution. The SGO Critic's analyses accumulate into a **Trait Lineage Ledger**: a chronological audit trail of every schema evolution with verbatim evidence. EVIDENCE: *generic priors are* **△ EVOLVED** *(“consumer-advocate mindset” → “vigilant about return processes”), novel constraints are* **+ DISCOVERY** *(“halved/percentage resource loss”), and context mismatches are* **× PRUNED** *(gaming-specific sizing trait deleted when evaluating apparel) (Appendix D, Table 6).* Together these constitute a **Glass Box**: behavioral archetypes inspectable and auditable for stereotype enforcement without retraining (Appendix C, Figure 4).

Because all user-specific identity is confined to external context vectors, SAPIENS supports deterministic erasure without retraining. A deletion request triggers: (1) deletion of $M_{\text{semantic}}(u_i)$ and $\Delta^*_{\text{behavior}}(u_i)$ from the context store; (2) removal of $\mathbf{v}_i$ from the retrieval database; (3) exclusion of u_i from the next Dynamic Centroid Realignment. The result is graceful fallback to $\mathcal{P}_i = (V_{\text{tribe}}, \emptyset, \emptyset)$, validated empirically by the Amnesia Study (Table 3). Privacy guarantees are scoped to the SAPIENS system boundary, contingent on the foundation model remaining frozen with no cross-session context persistence.

4 Experimental Setup

Dataset. We evaluate on **Amazon Reviews ’23** (Hou et al., 2024): 30,000 reviews from 5,002 users across seven categories. We selected this corpus as our verifiable proxy for generalized unstruc-

tured opinion data because it provides: (1) *passive traces* (organic, non-elicited); (2) *naturalistic stimuli* (history-independent product metadata); (3) *domain diversity* (functional and aesthetic); and (4) *scale*. Users are included if they have at least two reviews spanning at least one target category; prolific reviewers are sampled to prevent tribal prior domination. 52.6% of users have 1–4 reviews, meaning sparse users are the majority, not a corner case. Full category and micro-tribe population breakdowns are in Appendix F, Tables 10 and 11.

Theme Universe. String similarity metrics (BLEU, ROUGE) correlate poorly with human judgments on open-ended generation tasks (Liu et al., 2023), and semantic similarity alone cannot verify *what* a review is about. We therefore evaluate in a **semantic theme space**: a fixed per-category vocabulary of behavioral topics, where each theme represents a distinct evaluation dimension (e.g., *VFM*, *ease of use*) that users apply when reviewing products in that domain. Themes are domain-specific because relevant evaluation dimensions differ across categories. The vocabulary is derived via a three-stage LLM map-reduce procedure (discovery, consolidation, finalization), yielding 6–9 themes per category (51 total; Appendix F, Table 9), following the established practice of evaluating opinion systems against predefined aspect ontologies (Pontiki et al., 2016), and fixed prior to any model training or evaluation. Theme probabilities are derived from LLM token log-probabilities (Liu et al., 2023) and applied identically to both real and simulated reviews to produce P_{syn}; with no access to persona schemas and theme labels, ensuring the evaluation is not self-referential.

Train/Test Split. We employ a **user-stratified 80/20 temporal split**: 80% of each user's reviews initialize the SAPIENS persona (M_{semantic}, V_{tribe}, Δ_{behavior}); the remaining 20% are held out entirely. Each test review corresponds to a genuinely novel stimulus the model has never seen associated with this user. Sparse users with 1–2 test reviews are retained: they represent real conditions and are precisely the population where fine-tuned models fail.

Baselines and Ablation. Behavioral simulation fails in three structurally distinct ways. **Episodic Overfitting**: raw history conditioning over-indexes on episode-specific surface patterns, failing to abstract stable traits. **Sparsity Ceiling**: individual-only conditioning cannot impute preferences the user has never expressed. **Parametric Entanglement**: weight-encoded behavioral priors collapse toward the population mean on sparse, idiosyncratic users. The ablation ladder removes exactly one mechanism per step:

A_0 (Raw History): $\mathcal{P}_i = \emptyset$. Exposes Episodic Overfitting in isolation. A_1 (User Backstory (Moon et al., 2024)): $\mathcal{P}_i = (\emptyset, M_{\text{semantic}}, \emptyset)$. Current state of practice; hits the Sparsity Ceiling. A_2 (Be.FM (Xie et al., 2025) + backstory): parametric fine-tuning with individual context. Tests Parametric Entanglement. A_3 (Centaur (Binz et al., 2024) + backstory): distinct fine-tuning objective; tests whether the type of parametric prior matters. A_4 (Tribe Prior, Pre-SGO): $\mathcal{P}_i = (V_{\text{tribe}}, M_{\text{semantic}}, \emptyset)$. Tests group imputation without iterative refinement. A_5 (SAPIENS, Post-SGO): $\mathcal{P}_i = (V^*_{\text{tribe}}, M_{\text{semantic}}, \Delta^*_{\text{behavior}})$. Full system. A_2 and A_3 are evaluated without domain-specific fine-tuning, testing their claimed general-purpose capability. A_0, A_1, A_4, A_5 use GPT-o3; A_2, A_3 use their respective fine-tuned architectures. The theme classifier uses a separate lighter model across all conditions.

Evaluation Metrics. All metrics are computed per-review within the category-specific theme universe and averaged; never aggregated across incompatible theme spaces. **Recall@k_{gt}**: Measures direct interpretability: did the simulation predict the specific attributes (e.g., packaging, durability) the user focused on? Computed as $|R \cap \hat{R}_{k_{gt}}|/|R|$ with $k = |R|$, it avoids rewarding models that predict fixed-length theme lists regardless of ground truth.

$$\text{Recall@}k_{gt} = \frac{|R \cap \hat{R}_{k_{gt}}|}{|R|} \tag{4}$$

Jensen–Shannon Divergence (JSD): Evaluates distributional proportionality: does the relative importance of themes match the ground truth? Unlike Recall, JSD penalizes incorrect weightings (e.g., over-emphasizing comfort when the review focused on durability).

$$\begin{aligned}\text{JSD}(Q_{\text{real}} \parallel P_{\text{syn}}) = &\frac{1}{2} D_{\text{KL}}(Q_{\text{real}} \parallel M) \\ &+ \frac{1}{2} D_{\text{KL}}(P_{\text{syn}} \parallel M)\end{aligned} \tag{5}$$

where $M = \frac{1}{2}(Q_{\text{real}} + P_{\text{syn}})$ (computed in base-2, $\epsilon = 10^{-10}$). **Wasserstein Distance (WD)**: Calculates the total probability mass required to align synthetic and real distributions, effectively penal-

izing theme hallucinations (content the user never expressed).

$$\mathrm{WD}(Q_{\text{real}}, P_{\text{syn}}) = \mathrm{EMD}(Q_{\text{real}}, P_{\text{syn}};\, C) \quad (6)$$

($C_{ij} = \mathbb{1}[i \neq j]$). Together, Recall, JSD, and WD capture coverage, proportion, and total-mass accuracy, respectively. Lower JSD and WD indicate closer alignment between the simulated and ground-truth thematic distributions; higher Recall@k_{gt} indicates better topical coverage. All three metrics are reported such that ↓ denotes improvement for JSD and WD, and ↑ denotes improvement for Recall.

5 Results

5.1 Main Results: Baseline and Ablation Analysis

Table 1 reports performance across all six conditions on training (historical products) and test (unseen products) partitions. Test-partition results are the primary finding; train numbers diagnose overfitting.

5.2 Mechanistic Ablation Analysis

To bridge the quantitative shifts in Table 1 with their underlying behavioral causes, we trace these exact failure and recovery modes through a single, controlled qualitative lens: the "Lavender Lotion" scenario (Appendix A, Table 4). The gains and regressions across our ablation ladder reflect this precise sequence: episodic hallucination, sparse imputation, and semantic calibration, playing out at scale across all 5,084 test reviews.

Each step in the ablation ladder strictly improves over the previous: $A_0 \rightarrow A_1$ recovers from episodic noise (+9.36pp Recall); A_4 breaks the individual-only sparsity ceiling (+2.53pp over A_1); A_5 applies SGO calibration to reach the full system's performance (+7.23pp over A_4), while fine-tuned baselines A_2 and A_3 fall *below* A_1, confirming that parametric entanglement is actively detrimental on sparse users.

5.2.1 Episodic Overfitting: $A_0 \rightarrow A_1$

A_0 yields 63.51% Recall, JSD 0.58, WD 0.70: identical on train and test, the signature of pure Episodic Overfitting: the model applies surface patterns from the history context regardless of stimulus relevance. As seen in the case study, A_0 blindly hallucinated a past complaint onto a new, unrelated product.

A_1 improves test Recall to 72.87% (+9.36pp), JSD to 0.51 (−0.07), WD to 0.65 (−0.05). The train–test gap (78.92% vs. 72.87%, 6.05pp) is the empirical fingerprint of the Sparsity Ceiling: A_1's test performance is the hard upper bound on individual-only conditioning, failing to predict unexpressed preferences like travel packaging.

5.2.2 Parametric Entanglement: A_2 and A_3 vs. A_1

Be.FM (A_2) degrades from 74.45% train to 68.45% test Recall, *underperforming* A_1 despite being a purpose-built behavioral foundation model. Test JSD 0.58 and WD 0.68 are both worse than A_1's 0.51 and 0.65. **Centaur** (A_3) collapses further: test JSD 0.63 and WD 0.70, the highest divergence of any condition. Both confirm Parametric Entanglement: weight-level priors dominate the individual backstory signal on sparse users, collapsing toward the population mean.

5.2.3 Breaking the Sparsity Ceiling: A_4

A_4 raises test Recall to 75.40% (+2.53pp over A_1), JSD to 0.47 (−0.04), WD to 0.58 (−0.07). Unlike the $A_0 \rightarrow A_1$ gain, this is not noise reduction but signal *imputation*: the model makes correct predictions about preferences the individual has never expressed, drawing on the tribe to successfully impute the packaging preference. The slightly wider train–test gap (6.80pp vs. A_1's 6.05pp) reflects tribal over-generalization, by imputing an ingredient concern the specific user does not actually hold.

5.2.4 SGO Calibration: $A_4 \rightarrow A_5$

A_5 achieves **82.63%** test Recall (+7.23pp over A_4, +9.76pp over A_1), test JSD **0.32** (−31.9%), test WD **0.48** (−17.2%). The train–test gap narrows to 6.24pp, matching A_1's gap at far higher absolute performance, confirming that SGO learns generalizable behavioral constraints, not training-set patterns. SGO explicitly prunes the hallucinated tribal stereotypes (the ingredient critique) and injects discovered idiosyncratic traits as Δ_{behavior} cues, achieving total thematic alignment.

5.3 SGO Per-Tribe Convergence

Tables 7 and 8 (Appendix E) report pre/post-SGO metrics across all 150 micro-tribes. Improvement ranges from near-complete convergence (*No-Nonsense Value Guardians*: WD −96.0%, JSD −97.1%; *Nostalgia-Fueled Fun-Seekers*: Recall +100%) to modest gains (*Analytical Qual-*

Method / Architecture	Train			Test		
	Rec↑	JSD↓	WD↓	Rec↑	JSD↓	WD↓
A_0 Raw History (In-Context)	63.51%	0.58	0.70	63.51%	0.58	0.70
A_1 User Backstory (Prompt-Based)	78.92%	0.47	0.64	72.87%	0.51	0.65
A_2 Be.FM w. User Backstory (Finetuned)	74.45%	0.54	0.61	68.45%	0.58	0.68
A_3 Centaur w. User Backstory (Finetuned)	72.12%	0.59	0.65	70.12%	0.63	0.70
A_4 Tribe Prior (SAPIENS Pre-SGO)	82.20%	0.42	0.55	75.40%	0.47	0.58
A_5 **SAPIENS (Post-SGO Modular)**	88.87%	0.27	0.43	82.63%	0.32	0.48

Table 1: **Comprehensive Baseline & Ablation Analysis.** Fine-tuned architectures (A_2, A_3) suffer from generalization degradation on sparse users, lagging behind simple prompting (A_1). SAPIENS (A_5) achieves the strongest theme recall while producing the largest reductions in distributional divergence (JSD) and geometric distance (WD). A_0, A_1, A_4, A_5 *use GPT-o3.*

ity Guardians: WD −8.1%, Recall flat). *Health-Safe Detailers* is the single tribe where Recall marginally declines (−10.0%), while WD (−6.2%) and JSD (−9.6%) still improve: distributional alignment is better even when absolute coverage slightly undershoots. No tribe shows catastrophic collapse, validating the verification gate as a conservative update mechanism. Appendix D, Table 6 details ten representative SGO events across the three operation types (△ EVOLVED , + DISCOVERY , × PRUNED) with verbatim Critic analyses.

5.4 Cross-Domain Robustness

Domain	Train			Test		
	Rec@k	JSD	WD	Rec@k	JSD	WD
Video Games	88.5%	0.24	0.38	85.8%	0.31	0.42
Appliances	87.1%	0.29	0.44	84.6%	0.35	0.50
Software	85.3%	0.27	0.43	82.3%	0.32	0.48
Clothing & Shoes	81.9%	0.25	0.39	80.4%	0.34	0.44
Health & Personal	78.3%	0.25	0.40	77.1%	0.31	0.45
All Beauty	70.8%	0.29	0.45	67.6%	0.34	0.50

Table 2: Cross-Domain Evaluation. Performance degrades in All Beauty due to higher subjectivity and limited product-description signal.

Performance is robust across functional categories (77.1–85.8% test Recall), with narrow, consistent train–test gaps (1.2–3.0pp) confirming domain-agnostic generalization. All Beauty is the exception (67.6% test Recall); this reflects stimulus under-specification rather than persona failure (see Section 8).

6 Discussion

The ablation results reveal a **Simulation Trilemma**: behavioral simulation must balance *fidelity* (accuracy), *scalability* (incremental updates), and *transparency* (inspectability). As shown in Table 5, fine-tuning (e.g., Be.FM, Centaur) sacrifices transparency for fidelity, while prompt-based approaches (e.g., Anthology) sacrifice fidelity for scale; SAPIENS resolves all three by externalizing identity into a **Modular Persona Schema**. This **Glass Box** architecture ensures simulated behavior is causally traceable to specific schema elements (motivations, friction points), enabling falsifiability, targeted stereotype auditing, and behavioral debugging without retraining (Appendix C, Table 6), while enabling deterministic **Operational Unlearning**: since identity is confined to external vectors, erasure triggers graceful fallback to the anonymized tribal prior rather than model collapse (Table 3), providing a robust privacy guarantee absent in monolithically fine-tuned models.

7 Conclusion

SAPIENS demonstrates that grounding a sparse individual in their tribe's collective behavioral profile is a principled way for human behaviour simulation with mitigations to the cold-start problem. What makes the schema work is the combination: V_{tribe} provides the statistically grounded prior; M_{semantic} injects the individual's observed behavioral divergence from their community; and Δ_{behavior} captures fine-grained, context-specific departures that emerge only when the prior is tested against real held-out behavior. Sparsity, individuation, and residual deviation are each handled by a distinct, independently learnable component.

Community, individuality, and experience shapes Human behaviour, and SAPIENS shows that these forces are both conceptually and computationally separable, and learnable from behavioral traces generated through ordinary online activity. The resulting simulations are auditable, correctable, and grounded in real behavioral evidence. SAPIENS makes this learning systematic, transparent, and trustworthy, while powering accurate, interpretable human behavior simulation across domains where rich enough behavioral traces exist.

8 Limitations

Temporal Drift. Tribe Discovery relies on historical data snapshots to define behavioral archetypes, introducing latency between real-world preference shifts and updated tribe representations. In volatile domains, including viral trends, speculative markets, rapidly evolving product categories; tribe definitions may lag behind cultural change by the duration of the update window. SAPIENS is not static: centroid recomputation and tribe reassignment allow fresh behavioral observations to propagate into updated priors on a periodic basis, and the SGO loop can discover new activation cues as user behavior evolves. However, real-time tribal updating is not currently supported. Severity scales with domain volatility: negligible for stable preference domains (appliance utility, software compatibility) and potentially significant for fashion or cultural commentary. Mitigations include shorter update cycles, online centroid drift detection, and domain-specific volatility thresholds that trigger early recomputation.

Reasoning Upper Bound. Simulation capacity is bounded by the reasoning capabilities of the configured base model. The persona schema steers behavioral alignment: activating motivations, suppressing irrelevant concerns, injecting idiosyncratic response patterns, but cannot correct intrinsic reasoning failures such as long-horizon planning inconsistencies or systematic logical errors. This is precisely why SAPIENS does not fine-tune: encoding behavioral identity in weights (the approach of Be.FM (Xie et al., 2025) and Centaur (Binz et al., 2024)) entangles it with the model's reasoning capacity, making individual updates intractable, behavioral drivers opaque, and the base model non-swappable. By externalizing identity entirely into context space with no parameter updates, SAPIENS makes the base model a freely swappable component. A deployment requiring stronger reasoning configures a more capable model; one requiring lower latency configures a smaller one; one in a specialized domain configures a domain-tuned model. The persona schema, tribal priors, and learned deviation cues transfer across base models without modification; a configuration of SAPIENS optimized for any given task exists as long as an appropriate base model exists for that task, and the framework's performance ceiling rises automatically as foundation model capabilities improve.

Stimulus Informativeness. Cross-domain results (Section 5) reveal a performance ceiling that is not a failure of the persona schema but of the stimulus representation. In All Beauty, test Recall drops to 67.6%, substantially below all functional categories, because product descriptions list ingredients and claims rather than the sensory dimensions along which users vary. A well-calibrated persona can capture that a user values experiential qualities, but cannot predict which sensory experience a product delivers when the description does not encode it. This is also why prompt-based approaches that rely solely on the individual's expressed history (Anthology (Moon et al., 2024), Persona Hub (Ge et al., 2024)) hit a hard Sparsity Ceiling: if the user has never mentioned a preference, neither the backstory nor the product description can supply it. SAPIENS bypasses this for group-level imputation via the tribal prior, but the same stimulus informativeness constraint applies when the product description itself is under-specified. The fix is stimulus enrichment: summarized purchaser reviews, sensory attribute databases, multi-modal product imagery, rather than changes to the learning procedure. The framework is stimulus-agnostic: x in $\hat{y} = \text{LLM}(\mathcal{P}_i^*, x)$ can be any structured description of the environmental stimulus, and richer encodings transfer directly to improved fidelity without architectural modification.

Ethical Considerations

SAPIENS is explicitly designed to absorb passive observational logs and data from sources including but not limited to Reddit, LinkedIn, and X, along with enterprise data sources, to discover niche tribes and learn their traits. In the scope of this research work, there is no Personally Identifiable Information.

Privacy preservation and stereotype mitigation are primary architectural constraints embedded within the system's modular design, enabling deterministic deletion mechanisms tailored for industrial deployment.

SAPIENS does not claim to solve weight-level memorization in pretrained foundation models. It formally scopes privacy to the system boundary and proves complete revocation within that boundary.

Privacy via Parametric Isolation - "Right to be Forgotten": A fundamental vulnerability of monolithic behavioral models is the entanglement of user data within neural weights during fine-

Schema / Consent State	Rec@k ↑	JSD ↓	WD ↓
Condition 1: Full Consent			
Tribe Prior + User Traits + Episodic Memory	**81.40%**	**0.312**	**0.441**
Condition 2: Episodic Erasure			
Tribe Prior + User Traits *(Memory Deleted)*	79.40%	0.345	0.470
Condition 3: Complete Erasure			
Tribe Prior Only *(Traits & Memory Deleted)*	79.00%	0.350	0.490

Table 3: **User Context Deletion - Ablation Study:** Leave-one-out ablation across micro-tribes shows monotonic degradation under deterministic context deletion, validating behavioral fallback.

tuning. In such paradigms, executing an Erasure request often requires computationally prohibitive retraining or approximate machine unlearning techniques, which are an open research challenge.

SAPIENS circumvents this bottleneck via **Parametric Isolation**. By confining individual identity representations exclusively to modular, external context vectors, the framework enables the deterministic removal of a user's identity from the inference graph without requiring parameter updates. *Threat Model Assumption:* This operational unlearning relies on the premise that the underlying foundation model remains strictly frozen and does not perform cross-session memory persistence outside the provided context window; privacy is maintained by explicitly severing the data-to-inference pipeline.

Empirical Evaluation of Operational Unlearning (The Amnesia Study). To quantitatively validate this erasure mechanism, we designed an "Amnesia Study" utilizing a leave-one-out ablation across all users within two micro-tribes (micro_13: *Hands-On Value Beauty Testers*, and micro_15: *Hands-On Value Seekers*). We simulated progressive consent withdrawal by systematically deleting hierarchical layers of the users' schemas prior to generative inference, reporting the averaged performance degradation.

The measurable decay in predictive alignment (Table 3) demonstrates the system's behavioral fallback mechanism. When users revoke consent for episodic retention (Condition 2), their temporal context is unhooked from the context space, precipitating an expected distributional shift (average WD increases from 0.441 to 0.470).

Under Complete Erasure (Condition 3), idiosyncratic variables are purged entirely. The simulation degrades further, successfully falling back on the safely anonymized, aggregate *Tribal Prior*. While this does not constitute cryptographic proof against adversarial reconstruction from a pretrained base model, it empirically validates the operational disentanglement of the user's explicit data from the generative pipeline.

Deterministic Schema Excision. A Complete Erasure request (Condition 3) extends beyond simply dropping the user's active context; it triggers a permanent structural purge. First, the individual's embedding vector is deleted from the retrieval database. Second, during the system's next *Dynamic Centroid Realignment* (periodic recalculation of latent behavioral clusters, executed at a configurable frequency determined by the deploying enterprise's data freshness requirements) the tribal centroids are recomputed without the excised user's data.

As a result, the individual's historical footprint is systematically removed from the aggregate group prior. This two-step mechanism aligns with practical industrial deletion workflows for complete data revocation without the prohibitive cost of model retraining.

Verifiable Unlearning in Large Language Models remains an open research challenge, and though we do not claim to solve the general problem of verifiable unlearning, SAPIENS provides a deterministic, operationally meaningful erasure mechanism within its system boundary.

Mitigation of Tribal Stereotyping and Digital Redlining. Systems relying purely on group-level priors risk enforcing *Tribal Stereotyping*: hallucinating preferences based on a demographic cluster. SAPIENS counters this via the Semantic Gradient Optimization (SGO) loop, which explicitly prunes unverified tribal assumptions. By enforcing an "Explainability Constraint" where simulated reactions must be causally linked to explicitly calibrated traits, SAPIENS provides an auditable mechanism to mitigate automated digital redlining. SAPIENS converts the latent user state into human-readable text (**Appendix C**), creating a "Glass Box" system where behavioral archetypes can be audited for safety alignment. For instance, the generated schema for Micro-Tribe 0 (**Appendix C**) explicitly isolates explicit validation triggers and friction points, demonstrating that the system captures protective normative behaviors alongside pure prefer-

ence signals.

References

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

Neeraj Arora, Ishita Chakraborty, and Yohei Nishimura. 2024. Revolutionizing marketing research with a large language model: A hybrid ai-human approach. SSRN Working Paper.

Marcel Binz, Elif Akata, Matthias Bethge, Franziska Brändle, Fred Callaway, Julian Coda-Forno, Peter Dayan, Can Demircan, Maria K. Eckstein, Noémi Éltető, Thomas L. Griffiths, Susanne Haridi, Akshay K. Jagadish, Li Ji-An, Alexander Kipnis, Sreejan Kumar, Tobias Ludwig, Marvin Mathony, Marcelo Mattar, and 21 others. 2024. Centaur: A foundation model of human cognition. *arXiv preprint arXiv:2410.20268*.

Olivier Cappé and Eric Moulines. 2009. Online em algorithm for latent data models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(3):593–613.

Ricardo Dominguez-Olmedo, Moritz Hardt, and Celestine Mendler-Dünner. 2024. Questioning the survey responses of large language models. *arXiv preprint arXiv:2306.07951*.

Yuan Gao, Dokyun Lee, Gordon Burtch, and Sina Fazelpour. 2025. Take caution in using llms as human surrogates. *Proceedings of the National Academy of Sciences of the United States of America*, 122(24):e2501660122.

Tao Ge, Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi, and Dong Yu. 2024. Scaling synthetic data creation with 1,000,000,000 personas. *arXiv preprint arXiv:2406.20094*.

Vipul Gupta, Pranav Narayanan Venkit, Shomir Wilson, and Rebecca Passonneau. 2024. Sociodemographic bias in language models: A survey and forward path. In *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 295–322, Bangkok, Thailand. Association for Computational Linguistics.

Yupeng Hou, Jiacheng Li, Zhankui He, An Yan, Xiusi Chen, and Julian McAuley. 2024. Bridging language and items for retrieval and recommendation. *arXiv preprint arXiv:2403.03952*.

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

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-Eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Suhong Moon, Marwa Abdulhai, Minwoo Kang, Joseph Suh, Widyadewi Soedarmadji, Eran Kohen Behar, and David M. Chan. 2024. Virtual personas for language models via an anthology of backstories. *arXiv preprint arXiv:2407.06576*.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST)*, pages 1–15. ACM.

Joon Sung Park, Carolyn Q. Zou, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Robb Willer, Percy Liang, and Michael S. Bernstein. 2024. Generative agent simulations of 1,000 people. *arXiv preprint arXiv:2411.10109*.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad Al-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qiu, Orphée De Clercq, and 1 others. 2016. SemEval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 19–30, San Diego, California. Association for Computational Linguistics.

David M. Rothschild, James Brand, Hope Schroeder, and Jenny Wang. 2024. Opportunities and risks of llms in survey research. SSRN Working Paper.

Aadesh Salecha, Molly E. Ireland, Shashanka Subrahmanya, João Sedoc, Lyle H. Ungar, and Johannes C. Eichstaedt. 2024. Large language models display human-like social desirability biases in big five personality surveys. *PNAS Nexus*, 3(12):pgae533.

Huaman Sun, Jiaxin Pei, Minje Choi, and David Jurgens. 2025. Sociodemographic prompting is not yet an effective approach for simulating subjective judgments with llms. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 845–854, Albuquerque, New Mexico. Association for Computational Linguistics.

Henri Tajfel and John C Turner. 1979. *An integrative theory of intergroup conflict*. Monterey, CA: Brooks/Cole.

Robert P. Tett, Margaret J. Toich, and S. Burak Ozkum. 2021. Trait activation theory: A review of the literature and applications to five lines of personality dynamics research. *Annual Review of Organizational Psychology and Organizational Behavior*, 8:199–233.

Himanshu Thakur, Eshani Agrawal, and Smruthi Mukund. 2025. Personas within parameters: Fine-tuning small language models with low-rank adapters to mimic user behaviors. *arXiv preprint arXiv:2509.09689*.

Lei Wang, Jingsen Zhang, Hao Yang, Zhiyuan Chen, Jiakai Tang, Zeyu Zhang, Xu Chen, Yankai Lin, Ruihua Song, Wayne Xin Zhao, Jun Xu, Zhicheng Dou, Jun Wang, and Ji-Rong Wen. 2023. User behavior simulation with large language model based agents. *arXiv preprint arXiv:2306.02552*.

Qian Wang, Jiaying Wu, Zichen Jiang, Zhenheng Tang, Bingqiao Luo, Nuo Chen, Wei Chen, and Bingsheng He. 2025a. Llm-based human simulations have not yet been reliable. *arXiv preprint arXiv:2501.08579*.

Ziyi Wang, Yuxuan Lu, Yimeng Zhang, Jing Huang, and Dakuo Wang. 2025b. Customer-r1: Personalized simulation of human behaviors via rl-based llm agent in online shopping. *arXiv preprint arXiv:2510.07230*.

Yutong Xie, Zhuoheng Li, Xiyuan Wang, Yijun Pan, Qijia Liu, Xingzhi Cui, Kuang-Yu Lo, Ruoyi Gao, Xingjian Zhang, Jin Huang, Walter Yuan, Matthew O. Jackson, and Qiaozhu Mei. 2025. Be.fm: Open foundation models for human behavior. *arXiv preprint arXiv:2505.23058*.

Jianpeng Zhao, Chenyu Yuan, Weiming Luo, Haoling Xie, Guangwei Zhang, Steven Jige Quan, Zixuan Yuan, Pengyang Wang, and Denghui Zhang. 2025. Large language models as virtual survey respondents: Evaluating sociodemographic response generation. *arXiv preprint arXiv:2509.06337*.

A Qualitative Case Study: The "Lavender Lotion" Scenario

To illustrate the systemic computational failure modes discussed in Section 5.1, we present a representative test case: a beauty enthusiast reviewing the *Spanature Lavender Hand & Body Lotion*. Table 4 traces the prediction logic across the architectural ablation states.

Play-by-Play Analysis of Mechanisms. This case highlights the operational necessity of explicit hierarchical modularity when simulating sparse users.

(A_0) **The Trap of Surface Correlation (Raw History).** Conditioning on raw history yields the lowest performance. The model's attention mechanism overfits to specific past complaints, assuming they are permanent user traits. *Evidence:* As shown in Table 4, the user's Ground Truth review explicitly praised the product's longevity (*"moisturized for hours"*). However, A_0 hallucinated a negative review (*"doesn't last long"*) by blindly copying a complaint from the user's history with a distinctly different lotion, failing to recognize the context shift.

(A_1) **The Limit of Individual Identity (Single-User Persona).** Synthesizing history into a structured persona reduces episodic noise but hits the *Sparsity Ceiling*. *Evidence:* The Ground Truth highlighted *Packaging* (*"perfect size for my bag"*) as a key delight factor. A_1 missed this entirely because the user had never explicitly mentioned travel packaging in their past logs. Lacking the signal or a prior to draw from, the model reverted to generating a generic "effectiveness" review.

(A_4) **The Power of the Tribe (SAPIENS Static).** By mapping the user to a latent "Tribe" (e.g., *"Ingredient-Savvy Self-Care Analysts"*), we effectively impute missing values from the group prior. *Evidence:* A_4 correctly predicted the *Packaging* theme by imputing it from the tribe's norms. However, it over-generalized: the Ground Truth was silent on *Ingredients*, yet A_4 hallucinated a harsh critique of them, a standard priority for this tribe that this specific user did not actually share.

(A_5) **The Precision of Social Learning (SAPIENS Dynamic).** Finally, the Semantic Gradient Optimization loop delivers the decisive alignment. SGO acts as an active critic, isolating the residual errors introduced by the tribe prior. *Evidence:* The SGO-optimized persona (A_5) successfully retained the correct *Packaging* insight while explicitly pruning the *Ingredients* hallucination, achieving 100% thematic alignment with the user's actual reaction. The SGO loop delivers a consistent performance lift across all identified behavioral tribes, validating that the *Verbal Gradient* mechanism resolves sparsity without introducing group-level stereotyping.

B Exhaustive SGO Convergence Trajectories

In Section 3.3, we introduced the Semantic Gradient Optimization (SGO) loop, which employs a verification gate to act as a discrete line search. The following figures provide the exhaustive convergence trajectories and structural alignment proofs across the evaluation set.

Because LLM-driven contextual updates operate in a non-convex, discrete text space, raw verbal gradients risk introducing semantic regressions. The strict monotonic decrease observed across the composite loss components empirically demonstrates that the verification mechanism effectively filters deleterious updates.

Specifically, the global training trajectory across 12 iterations (**Figure 2, Left**) illustrates two distinct optimization dynamics. The **Semantic Loss** drives the most rapid early-iteration decay (achieving a 36.0% overall reduction, $0.423 \rightarrow 0.270$), indicating that verbal gradients are highly efficient at calibrating continuous stylistic and tonal alignment. Conversely, the **Thematic Loss** exhibits a much flatter, asymptotic decay (a 15.0% reduction, $0.524 \rightarrow 0.445$), reflecting the granular, iterative schema refinements required to correct discrete topical coverage without causing catastrophic forgetting of previously aligned traits.

Granular Stability Across Micro-Tribes. Crucially, the monotonic stability of the SGO loop is not merely an artifact of global averaging. Figure 3 provides the isolated convergence trajectories for nine distinct micro-tribes. Across highly diverse behavioral archetypes, ranging from *Ingredient-Savvy Beauty Testers* to *Function-First Gamers*, the optimization loop maintains strict non-increasing trajectories, demonstrating that the externalized schema allows SAPIENS to safely execute holistic corrections on niche sub-populations without the gradient instability commonly associated with targeted fine-tuning.

Feature (Ground Truth)	A_0: History	A_1: Persona	A_4: Tribe (Static)	A_5: SAPIENS (Dynamic)
Packaging *"Perfect travel size..."*	**Missed** (Overfit to moisture)	**Missed** (Sparsity: Never mentioned)	**Predicted** (Imputed from Tribe)	**Predicted** (Retained)
Scent *"Real lavender..."*	**Predicted** (Shared preference)	**Predicted** (Core trait)	**Predicted** (Core trait)	**Predicted** (Core trait)
Ingredients *(Not mentioned)*	**Hallucinated Hate** (Projected past complaints)	Neutral (No strong prior)	**Hallucinated Hate** (Tribe stereotype)	**Correctly Pruned** (SGO Calibration)

Table 4: **Failure Analysis (The "Lavender Lotion" Case Study):** A_0 fails by overfitting to past negative sentiment from an unrelated product. A_1 fails to predict the *Packaging* preference due to historical sparsity. A_4 successfully imputes the *Packaging* preference from the tribal prior but hallucinates a negative opinion on *Ingredients* (a tribal stereotype). A_5 **(SAPIENS)** calibrates the profile via SGO, retaining the imputed signal while explicitly pruning the hallucination to achieve alignment.

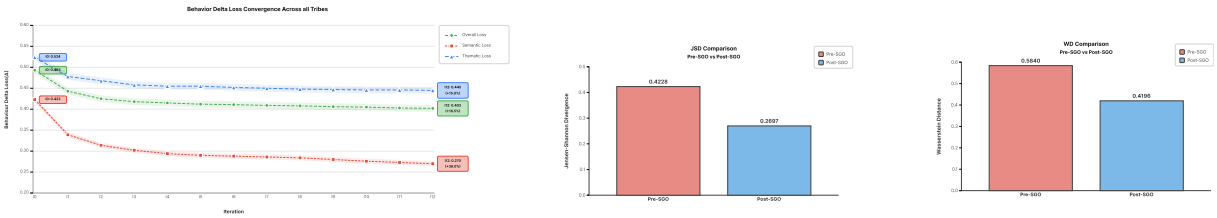

Figure 2: **Global SGO Convergence and Structural Alignment.** *(Left)* The monotonic downward trajectory of the loss components confirms the empirical stability of the SGO loop. *(Center & Right)* This iterative loss reduction directly translates into substantial improvements in structural fidelity, evidenced by the drops in Jensen-Shannon Divergence (0.4228 → 0.2697) and Wasserstein Distance (0.5840 → 0.4196).

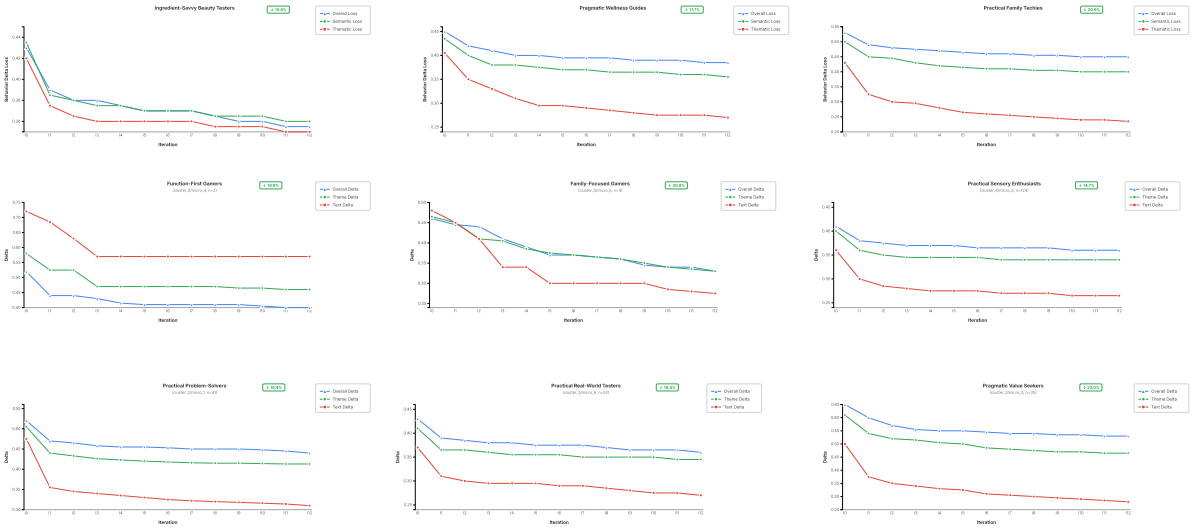

Figure 3: **Micro-Tribe Convergence Grid.** Exhaustive component-wise loss trajectories across a diverse subset of behavioral archetypes. The consistent monotonic decay across every isolated sub-population proves the verification gate successfully prevents prompt degradation and stylistic collapse during iterative refinement, regardless of the target domain.

Feature	Foundation Models	Anthology (Prompt)	Generative Agents	SAPIENS (Ours)
Identity Source	Model Weights	Generated Stats	Active Interviews	**Tribe traits**
Update Speed	Slow (Re-train)	Static (Fixed)	Static (No Loop)	**Fast (Iterative SGO)**
Data Fidelity	Labelled observation data	Plausible	Qualitative	**Passive Observation Logs**
Privacy	Low (Irreversibly diffused)	High (Synthetic)	Medium	**High (Parametric Isolation)**

Table 5: **Simulation Trilemma:** Trade-offs across behavioral simulation frameworks. SAPIENS is the first to simultaneously achieve high fidelity, fast incremental updating, and high-privacy identity representation.

C Gallery of Inferred Tribal Traits, Schemas, and Generative Reactions

Pragmatic Value Advocates · *The Pragmatic Performance Seeker*

This tribe consists of hands-on, result-oriented shoppers who test items immediately, benchmark them against previous experiences, and speak up whenever marketing promises don't match reality. Their writing style is concise yet pointed: short blurbs if a product simply works, longer rants when flaws affect everyday use or fairness.

144 reviews
3.76 avg rating

INHERENT BEHAVIORAL TRAITS

80% **Appreciates color variety and aesthetic-only reviews.**
A strong visual preference leads to concise reviews focused solely on color variety and aesthetic satisfaction.

90% **Issues ultra-brief blurbs spotlighting only color/style wins.**
Aesthetic satisfaction is a standalone trigger for concise positive feedback, indicating a strong visual preference.

90% **Vigilant about return processes and transactional fairness.**
Anxiety about post-purchase disputes drives a focus on fair return policies.

+12 more

LATENT MOTIVATIONS

90% **Ensuring fair pricing and honest value.**
A strong desire for equitable transactions drives scrutiny of price and value alignment.

90% **Securing accurate fit or functional compatibility.**
Functional and ergonomic compatibility are key to user satisfaction and product utility.

80% **Avoiding technical glitches or hidden costs.**
A focus on seamless functionality and transparency underpins a desire for hassle-free experiences.

+11 more

IMPLICIT GOALS

90% **Help the community avoid low-quality or overpriced products.**
A consumer-advocate mindset drives efforts to protect peers from unsatisfactory purchases.

90% **Quickly identify reliable, enjoyable options that meet everyday needs.**
Efficiency in decision-making is prioritized to enhance daily life satisfaction.

90% **Stretch their own budget while maximizing satisfaction.**
A strategic approach to spending ensures maximum value and enjoyment from purchases.

+7 more

VALIDATION TRIGGERS (Positive Affordances)

- Accurate sizing or ergonomic comfort
- Engaging, well-balanced gameplay
- Solid material quality or construction
- Quick setup and straightforward usability
- Reasonable price relative to performance
- Immediate performance upgrades
- Cross-platform functionality
- Spatial audio performance
- Early-failure durability detection
- Rigged mechanics and performance hitch awareness
- Camera control precision
- Fair difficulty balance
- Functional features like pockets
- Platform-specific content availability
- Immediate usability out of the box

FRICTION POINTS (Negative Affordances)

- Misleading descriptions or hidden fees
- Poor fit, shrinkage, or tightness
- Bugs, crashes, or long load times
- Excessive difficulty spikes or lack of clear instructions
- Fragile materials or premature wear
- Connectivity issues affecting gameplay
- Slow, repetitive progression
- Hard-to-slip-on designs
- In-game currency loss
- Vanishing save data
- Self-assembly inconvenience
- Sensory discomfort
- Camera control issues
- Unfair difficulty settings
- Level-cap frustrations

AE `AE22CFXT3QZKUQJORVTGL3VQXAAA`

This user writes reviews to guide fellow consumers with candid, experience-based insights, drawing on their familiarity with product lines—whether video game franchises or apparel sizing—to compare against past offerings and expected standards. They value accurate representation, functional fit, and overall usability, and feel compelled to highlight both standout quality and misleading claims in order to protect and inform others. Their tone is concise yet opinionated, reflecting a practical, consumer-advocate mindset rather than a purely emotional or promotional one.

AE `AE77BXH7TZSCOADUDRESKQ7Y6FMQ`

This user writes concise, experience-driven reviews that focus on how well a product delivers personal enjoyment or physical comfort. They value specific functional details—such as a game's challenge level or a garment's fit and color—and use reviews to highlight both praise and practical shortcomings. Their primary motivation seems to be offering straightforward guidance to help others understand real-world performance before purchasing.

AG `AG7NVDGLQP5BCP2ZWIVUMUGWSB5Q`

This user writes reviews to address practical usability issues and to share candid, experience-based advice that can aid fellow consumers. They focus on performance shortcomings that hinder enjoyment (e.g., stalled game levels) and on tangible attributes such as fit, comfort, sizing, price, and shipping speed, revealing a value-for-money and everyday-function mindset. Their tone mixes direct problem-solving requests with helpful descriptive detail, suggesting a motivation to both prompt product improvement and guide others toward confident purchase decisions.

AG `AG2HUPLXFIKF2SVCT42HVNWRT33Q`

This user is a practical, fairness-oriented consumer who writes balanced, experience-based reviews to guide others on both enjoyment and potential pitfalls. They combine analytical comparisons (game mechanics, storyline depth, product build quality) with consumer-rights vigilance, updating posts when transactional issues arise to warn the community. Their motivation stems from wanting others to have transparent expectations and equitable treatment.

Figure 4: **Tribe Trait Ontology: Micro-Tribe 0 (Pragmatic Value Advocates):** The hierarchical parameterization of *Pragmatic Value Advocates*. As discussed in Section 5.1, SAPIENS replaces implicit parametric entanglement with explicitly modular tribal schemas. This figure provides a comprehensive representation of this process, illustrating the explicit modularization of the latent behavioral manifold. It successfully isolates weighted psychological drivers (Latent Motivations, Implicit Goals) from Catalytic Affordances (Validation Triggers, Friction Points) derived from 144 real users. By preserving these discrete affordances and foundational user backstories verbatim, the schema captures extreme behavioral nuance and establishes a statistically reliable prior without relying on parametric weight updates.

Pragmatic Value Advocates · *Evaluative Ground-Truth Alignment*

This micro-cluster consists of hands-on, result-oriented shoppers who test items immediately, benchmark them against previous experiences, and speak up whenever marketing promises don't match reality. Their writing style is concise yet pointed: short blurbs if a product simply works, longer rants when flaws affect everyday use or fairness.

30 users
144 reviews
3.76 avg rating

1 AMAZON FASHION

AE AEANK4MYDH4YO7SUUMAHXLDLK4MQ

Product Context: *Dipper Flat Hat Blue Pine Tree Embroidered Movie Cap Adult One Size Royal/White*

REAL REVIEW ★ 4.0 Positive

I really like this hat. It is of good quality with an actual embroidered logo. The back has a snapback closure. If someone you know is fan of Gravity Falls they are going to dig this lid. On the downside this hat is large. I take a 7 5/8 in a fitted cap and this fits me on the smallest possible setting. I originally purchased this hoping to give this to my 8-year-old son but it is way to large for that. Instead dad has the coolest looking Dipper hat on the playground. I have received several compliments on my hat.

Topics: 100% Style, Color & Aesthetic Appeal 80% Fit & Sizing Accuracy 70% Material Quality, Durability & Construction

SYNTHETIC REVIEW ★ 4.0 Positive

This Dipper hat is a hit with fans! The embroidered pine tree design is spot on, and the royal blue color really pops. It fits well for an adult one-size cap, though I wish it had a bit more adjustability. The material feels sturdy enough for casual wear. Overall, a solid buy for any Gravity Falls enthusiast.

Topics: 100% Style, Color & Aesthetic Appeal 80% Fit & Sizing Accuracy 70% Material Quality, Durability & Construction

2 Video Games

AE AEANK4MYDH4YO7SUUMAHXLDLK4MQ

Product Context: *Plug your favorite compatible headset into the Xbox One Stereo Headset Adapter and hear the action just the way you like it. Easily adjust chat audio without taking your hands off the controller. Add game audio by connecting directly to your console or TV. Non-compatible headsets... (Contact Mad Catz for support.) This connector includes a long, cylindrical pin in the middle of the connecting side that does not fit into the Xbox One Stereo Headset Adapter.*

REAL REVIEW ★ 4.0 Positive

While the price that microsoft charges for these is redonkulous, it works well and is cheaper than a whole new head set for your xbox one.

Topics: 90% Pricing Value and Monetization 80% Hardware Build Quality Comfort and Connectivity

SYNTHETIC REVIEW ★ 4.0 Positive

The Xbox One Stereo Headset Adapter does the job for a fair price, but be warned: not all headsets are compatible. Setup is straightforward, but don't expect miracles if your favorite gear isn't listed. It's solid for connecting most standard headsets, but the limitations with certain models could frustrate some gamers. Overall, a practical choice if you match the compatibility list.

Topics: 90% Pricing Value and Monetization 80% Hardware Build Quality Comfort and Connectivity 60% Installation DRM and Privacy Requirements

3 Video Games

AE AE22CFXT3QZKUQJORVTGL3VQXAAA

Product Context: *Use hero-summoning strategies to defeat an ancient threat in a brand-new Fire Emblem game. In a war against the Fell Dragon, four kingdoms worked together with heroes from other worlds to seal away this great evil. One-thousand years later, this seal has weakened... Team up with iconic heroes from past Fire Emblem games. Aside from merging appearances, Engaging lets you inherit weapons, skills, and more from these battle-tested legends. Nintendo / INTELLIGENT SYSTEMS*

REAL REVIEW ★ 5.0 Positive

IMO this has been one of the better FEs to date. It takes a little from previous games and combines plability into one while offering a newish story.

Topics: 90% Gameplay Mechanics Controls and Difficulty 70% Content Depth Storyline and Appropriateness

SYNTHETIC REVIEW ★ 4.0 Positive

This new Fire Emblem game delivers on its promise of strategic depth with the return of classic mechanics and fresh story elements. The inclusion of heroes like Marth and Celica adds a nostalgic touch for longtime fans, and the game is straightforward to pick up with engaging battles. However, the pricing feels steep considering the content offered.

Topics: 90% Gameplay Mechanics Controls and Difficulty 70% Content Depth Storyline and Appropriateness 70% Pricing Value and Monetization

4 AMAZON FASHION

AF AFB3IFFCZZKH53K3EKXXUY2IGGHQ

Product Context: *Classic comfort gets a modern update. Crafted with breathable microsuede and temperature-regulating definition adapt technology, these moccasins are the perfect pick for relaxing around the house or on the go! Durable indoor/outdoor outsoles add an extra touch of versatility.*

REAL REVIEW ★ 4.0 Positive

I used the sizing and they came in. Tried them on and can only wear them without socks. Maybe a size larger if you must wear socks? I have another pair of slippers I bought at the same time, same size, and sock worthy. No real issues otherwise and may see them expand a little over time. Otherwise exactly what I wanted.

Topics: 100% Fit & Sizing Accuracy 100% Comfort, Support & Wearability

SYNTHETIC REVIEW ★ 4.0 Positive

These moccasins are a great fit for lounging and quick errands, thanks to their breathable microsuede and effective temperature regulation. The fit is spot-on, and the durable sole makes them versatile for indoor and outdoor use. While the style is classic, it might not stand out in a crowd, but the comfort and practicality certainly do.

Topics: 100% Comfort, Support & Wearability 100% Fit & Sizing Accuracy 90% Material Quality, Durability & Construction

Figure 5: **Simulated Trait Activation and Ground-Truth Alignment (Same Tribe).** The forward-pass generative validation for *Micro-Tribe 0*. By injecting the explicit tribal schema (Figure 4) into the context window alongside an environmental stimulus (Product Context), SAPIENS successfully mirrors the multidimensional thematic distributions (Topics) observed in the ground-truth human trajectories. This confirms the system's ability to provide side-by-side empirical validation, demonstrating that when this specific modular schema is prompted to simulate a reaction to novel products, the resulting synthetic reviews achieve near-perfect thematic alignment with the actual users' ground-truth reactions, even across severe domain shifts (e.g., Apparel to Video Games).

D Tribal Traits' Evolution and Lineage Ledger During SGO

Table 6: **Trait Rationalization during SGO iteration:** A trace of ten dynamic trait evolutions demonstrating how SAPIENS extracts precise behavioral nuance from sparse historical logs. The Critic actively deepens generic tribal priors into highly specific constraints, discovers novel latent friction points, and explicitly prunes unverified stereotypes when they contradict the user's local domain context (e.g., *AMAZON FASHION*). As a structural byproduct, routing these modifications into explicit psychological buckets (e.g., *Latent Motivations*) alongside their verbatim triggers yields a fully auditable ledger of persona development.

Optimization Type	Baseline Assumption	Evidence (Verbatim Log)	Refined Outcome
△ EVOLVED Medium Mag (Iter 5.10)	*"Consumer-advocate mindset"*	**Ctx:** Inherent Behavioral Traits *"Missed purchase-experience theme because persona omits their pattern of seeking exchange guidance when sizing is wrong"*	**Vigilant about return processes** *Gain:* Refines generic advocacy into specific transactional fairness.
+ DISCOVERY Novel (Iter 10.1)	— *(Null State)*	**Ctx:** Latent Motivations *"Prediction skipped coin-loss glitch because persona lacks trait that user flags precise resource discrepancies"*	**[-] Unfair in-game economy or currency loss** *Note: Identified as Friction Point.*
× PRUNED Context Mismatch (Iter 5.10)	*"Accurate sizing or ergonomic comfort"*	**Ctx:** AMAZON FASHION *"Missing nuance that one-line 'color+fit' remarks denote only aesthetic and sizing scope, causing overprediction"*	**[TRAIT DELETED]** *Reason:* Log context contradicted trait domain; sizing is situational.
+ DISCOVERY Novel (Iter 5.9)	— *(Null State)*	**Ctx:** Inherent Behavioral Traits *"Prediction ignored connectivity complaints because persona lacks explicit focus on multiplayer infrastructure"*	**Scrutinizes online multiplayer stability** *Gain:* Differentiates between single-player bugs and network issues.
× PRUNED Context Mismatch (Iter 12.1)	*"Free DLC as a value driver"*	**Ctx:** GAMING *"Log context contradicted trait domain, indicating that the trait was misclassified as it was interpreted within a non-gaming context."*	**[TRAIT DELETED]** *Reason:* System rejected Gaming trait when validated against Fashion logs.
+ DISCOVERY Novel (Iter 10.4)	— *(Null State)*	**Ctx:** Latent Motivations *"Prediction misattributed disappointment to price/story because persona lacks explicit attention to lost carry-over features"*	**[-] Loss of expected features or continuity** *Note: Identified as Friction Point.*
+ DISCOVERY Novel (Iter 10.3)	— *(Null State)*	**Ctx:** Inherent Behavioral Traits *"Existing asset-loss watchdog missed the '90 million -> 45 million' phrasing, so refine it to catch halved/percentage coin-drop"*	**Detects halved/percentage resource loss** *Gain:* Keen sense of fairness drives attention to resource allocation.
× PRUNED Context Mismatch (Iter 12.1)	*"Queries return logistics"*	**Ctx:** AMAZON FASHION *"The focus on size errors was specific to fashion and did not align with the broader context of gaming"*	**[TRAIT DELETED]** *Reason:* Log context contradicted trait domain (Gaming vs Fashion).
+ DISCOVERY Novel (Iter 10.2)	— *(Null State)*	**Ctx:** Latent Motivations *"Absence of characteristics highlighting online-service frustrations let connectivity issues go undetected"*	**[-] Connectivity bugs eroding franchise trust** *Note: Identified as Friction Point.*
+ DISCOVERY Novel (Iter 5.3)	— *(Null State)*	**Ctx:** Inherent Behavioral Traits *"Model ignored gameplay-tedium complaint because persona still lacks an explicit trait that user flags slow, repetitive progression"*	**Flags slow, repetitive progression** *Gain:* Identifies "Grind" as a primary deal-breaker.

E Exhaustive Micro-Tribe Convergence Database

	Geom. (WD) ↓		Dist. (JSD) ↓		Recall@k ↑		Sem. Δ	Over. Δ
Micro-Tribe Archetype	**Pre→Post**	**Red%**	**Pre→Post**	**Red%**	**Pre→Post**	**Inc%**	**Red%**	**Red%**
micro_1: No-Nonsense Value Guardians	0.71→0.03	**-96.0%**	0.51→0.01	**-97.1%**	1.00→1.00	**+0.0%**	**-17.0%**	**-36.2%**
micro_2: Nostalgia-Fueled Fun-Seekers	0.67→0.17	**-74.7%**	0.59→0.10	**-83.7%**	0.50→1.00	**+100.0%**	**-56.8%**	**-68.5%**
micro_3: Practical Family Shoppers	0.70→0.25	**-64.1%**	0.51→0.15	**-70.8%**	0.83→1.00	**+20.0%**	**-8.9%**	**-37.3%**
micro_4: Pragmatic Quick-Checkers	0.70→0.26	**-62.1%**	0.53→0.15	**-71.7%**	0.63→0.90	**+41.9%**	**-25.1%**	**-91.6%**
micro_5: Upbeat Fun-Seekers	0.54→0.21	**-60.4%**	0.35→0.13	**-63.9%**	1.00→1.00	**+0.0%**	**-3.3%**	**-27.1%**
micro_6: Value-Focused Gamer Guides	0.61→0.29	**-52.2%**	0.48→0.14	**-70.8%**	0.42→0.83	**+100.0%**	**-32.1%**	**-48.0%**
micro_7: Joy-First Enthusiasts	0.68→0.31	**-54.7%**	0.49→0.19	**-61.9%**	0.74→0.97	**+30.6%**	**-29.8%**	**-75.7%**
micro_8: Reliability-Focused Pragmatists	0.69→0.31	**-54.5%**	0.52→0.20	**-61.5%**	0.54→0.88	**+63.9%**	**-20.9%**	**-60.7%**

Table 7: Tribe-level Convergence Metrics (Post-SGO): A subset of micro-tribes demonstrating targeted convergence (see Table 8 for the exhaustive database). Decoupling traits into an external schema allows SAPIENS to correct niche behavioral deviations without updating model weights.

Table 8: **Exhaustive SGO Convergence Database (All Metrics).** To empirically validate that the SGO loop calibrates behavior uniformly across diverse psychological profiles without causing gradient collapse in minority groups, this table logs the exact Wasserstein Distance (WD), Jensen-Shannon Divergence (JSD), and Recall@k improvements across the complete set of 150 discovered micro-tribes.

	Geom. (WD) ↓		Dist. (JSD) ↓		Recall@K ↑		Sem. Δ	Over. Δ
Micro-Tribe Archetype	**Pre→Post**	**Red%**	**Pre→Post**	**Red%**	**Pre→Post**	**Inc%**	**Red%**	**Red%**
Macro-Tribe: The Pragmatic Performance Seeker (Cluster 0)								
Pragmatic Value Advocates	0.56→0.43	-23.6%	0.39→0.35	-11.8%	0.74→0.83	+10.8%	-14.1%	-36.9%
Analytical Quality Guardians	0.41→0.37	-8.1%	0.28→0.24	-15.3%	0.71→0.71	+0.0%	-35.8%	-32.4%
Community-Minded Enthusiasts	0.72→0.61	-14.7%	0.62→0.53	-13.8%	0.33→0.50	+50.0%	-9.6%	-42.6%
Hands-On Value Techies	0.48→0.40	-16.9%	0.34→0.26	-23.1%	0.69→0.83	+20.2%	-15.5%	-50.8%
Durability-Driven Gamers	0.83→0.68	-18.7%	0.73→0.57	-22.1%	1.00→1.00	+0.0%	-24.3%	-23.4%
Performance-Value Gamers	0.47→0.36	-23.9%	0.35→0.24	-31.0%	0.89→0.89	+0.0%	-15.6%	-38.1%
Pragmatic Value Gamers	0.56→0.47	-16.9%	0.41→0.31	-26.0%	0.57→0.80	+38.7%	-17.5%	-43.1%
Discerning Nostalgic Collectors	0.54→0.44	-19.0%	0.40→0.28	-29.1%	0.67→0.83	+25.0%	-37.1%	-45.7%
Nostalgia-Fueled Fun-Seekers	0.67→0.17	-74.7%	0.59→0.10	-83.7%	0.50→1.00	+100.0%	-56.8%	-68.5%
Value-Focused Gamer Guides	0.61→0.29	-52.2%	0.48→0.14	-70.8%	0.42→0.83	+100.0%	-32.1%	-48.0%
Family Tech Pragmatists	0.60→0.38	-36.1%	0.43→0.23	-45.1%	0.65→0.82	+25.5%	-16.1%	-48.3%
Pragmatic Value Seekers	0.73→0.34	-54.1%	0.54→0.21	-61.6%	0.50→1.00	+100.0%	-23.1%	-57.5%
Pragmatic Value Gamers	0.51→0.38	-25.6%	0.32→0.23	-29.9%	1.00→1.00	+0.0%	-13.0%	-33.0%
Value-Minded Time-Killers	0.62→0.42	-31.3%	0.48→0.24	-50.4%	0.67→0.67	+0.0%	-14.8%	-35.4%
Macro-Tribe: The One-Tap Cheerleader (Cluster 1)								
Practical Aesthetes	0.61→0.47	-22.5%	0.43→0.30	-30.8%	0.79→0.82	+4.0%	-13.5%	-57.4%
Quick-Reaction Shoppers	0.65→0.39	-40.1%	0.46→0.24	-48.3%	0.86→0.93	+8.3%	-22.8%	-46.3%
Upbeat Practical Enthusiasts	0.58→0.40	-31.6%	0.43→0.25	-42.1%	0.78→0.88	+12.0%	-10.1%	-35.0%
Comfort-First Enthusiasts	0.53→0.39	-27.0%	0.36→0.24	-34.0%	0.75→0.87	+16.4%	-9.1%	-52.4%
Quick-Enjoyment Enthusiasts	0.56→0.42	-25.8%	0.43→0.27	-38.8%	0.57→0.79	+39.0%	-11.5%	-58.4%
Joy-First Enthusiasts	0.68→0.31	-54.7%	0.49→0.19	-61.9%	0.74→0.97	+30.6%	-29.8%	-75.7%
Brief Pragmatic Shoppers	0.65→0.48	-25.6%	0.46→0.32	-31.9%	0.71→0.90	+27.8%	-16.3%	-52.9%
Stylish Bargain Boosters	0.63→0.41	-35.2%	0.44→0.25	-43.0%	0.83→0.98	+17.1%	-13.4%	-37.0%
Aesthetic-Driven Enthusiasts	0.55→0.37	-33.1%	0.38→0.23	-39.4%	0.51→0.90	+77.7%	-14.5%	-49.3%
Upbeat Fun-Seekers	0.54→0.21	-60.4%	0.35→0.13	-63.9%	1.00→1.00	+0.0%	-3.3%	-27.1%
Macro-Tribe: The Straight-Shooting Bargain Scout (Cluster 3)								
Practical Value Guardians	0.49→0.42	-14.9%	0.35→0.27	-21.9%	0.51→0.58	+12.7%	-9.8%	-35.6%
Practical Comfort Seekers	0.63→0.40	-36.5%	0.46→0.26	-45.0%	0.52→0.69	+31.9%	-19.7%	-58.3%
No-Nonsense Value Guardians	0.71→0.03	-96.0%	0.51→0.01	-97.1%	1.00→1.00	+0.0%	-17.0%	-36.2%

Continued on next page

Table 8 – continued from previous page

	Geom. (WD) ↓		Dist. (JSD) ↓		Recall@K ↑		Sem. △	Over. △
Micro-Tribe Archetype	**Pre→Post**	**Red%**	**Pre→Post**	**Red%**	**Pre→Post**	**Inc%**	**Red%**	**Red%**
Hands-On Value Pragmatists	0.60→0.48	-19.9%	0.43→0.31	-27.8%	0.62→0.83	+32.9%	-18.5%	-52.6%
Pragmatic Quick-Checkers	0.67→0.35	-47.6%	0.49→0.21	-57.7%	0.62→0.90	+44.4%	-16.4%	-48.6%
Comfort-Fit Pragmatists	0.43→0.34	-19.2%	0.31→0.23	-26.4%	0.62→0.72	+15.8%	-10.2%	-43.9%
Practical Performance Seekers	0.52→0.43	-16.8%	0.37→0.28	-25.3%	0.65→0.81	+25.0%	-19.5%	-42.6%
Practical Value Guardians	0.62→0.45	-28.4%	0.45→0.28	-37.4%	0.76→0.91	+19.5%	-13.4%	-37.8%
Function-First Comfort Shoppers	0.52→0.42	-19.8%	0.38→0.28	-27.2%	0.68→0.75	+9.4%	-12.7%	-53.3%
Pragmatic Value Seekers	0.62→0.48	-22.4%	0.47→0.32	-33.1%	0.77→0.83	+8.0%	-16.5%	-64.8%
Functionality-First Pragmatists	0.63→0.34	-46.1%	0.46→0.22	-52.3%	0.79→0.96	+21.4%	-17.3%	-69.3%
Pragmatic Value Seekers	0.40→0.29	-28.8%	0.27→0.15	-45.7%	0.50→0.67	+33.3%	-30.7%	-34.0%
Pragmatic Value Seekers	0.64→0.41	-36.2%	0.45→0.25	-44.0%	0.76→0.91	+19.5%	-17.2%	-52.2%
Pragmatic Quality Guardians	0.60→0.37	-39.4%	0.42→0.22	-47.5%	0.64→0.91	+42.8%	-15.1%	-54.4%
Reliability-Driven Tech Pragmatists	0.43→0.33	-23.6%	0.33→0.21	-36.8%	0.56→0.78	+40.0%	-17.4%	-50.7%
Function-First Pragmatists	0.55→0.46	-15.2%	0.40→0.30	-23.9%	0.57→0.72	+25.9%	-15.5%	-34.9%
Practical Comfort Seekers	0.56→0.36	-35.7%	0.39→0.20	-47.4%	0.94→1.00	+6.7%	-23.5%	-45.7%
Practical Function Seekers	0.60→0.43	-28.5%	0.44→0.28	-35.8%	0.69→0.84	+21.8%	-15.1%	-52.1%
Pragmatic Reliability Seekers	0.53→0.42	-21.7%	0.38→0.26	-31.5%	0.64→0.83	+30.1%	-12.1%	-56.1%
Function-First Pragmatists	0.49→0.35	-28.0%	0.35→0.23	-33.5%	0.83→1.00	+20.0%	-9.6%	-30.7%
Reliability-Focused Pragmatists	0.69→0.31	-54.5%	0.52→0.20	-61.5%	0.54→0.88	+63.9%	-20.9%	-60.7%
Blunt Quality Guardians	0.55→0.43	-21.6%	0.40→0.28	-28.7%	0.53→0.62	+15.6%	-20.6%	-59.8%
Practical Comfort Seekers	0.67→0.45	-33.2%	0.49→0.28	-42.9%	0.53→0.92	+73.7%	-27.3%	-44.7%
Practical Value Seekers	0.66→0.34	-48.0%	0.47→0.21	-54.8%	0.88→0.99	+12.4%	-24.9%	-88.5%
Pragmatic Quick-Checkers	0.70→0.26	-62.1%	0.53→0.15	-71.7%	0.63→0.90	+41.9%	-25.1%	-91.6%
Functionality-First Pragmatists	0.60→0.36	-39.6%	0.44→0.23	-47.7%	0.59→0.86	+45.1%	-20.0%	-59.8%
Pragmatic Value Seekers	0.58→0.42	-26.8%	0.44→0.29	-34.2%	0.62→0.73	+17.0%	-13.7%	-55.1%
Pragmatic Value Seekers	0.68→0.42	-38.4%	0.50→0.27	-45.5%	0.86→0.97	+11.9%	-19.0%	-64.7%
Durability Value Seekers	0.63→0.39	-38.6%	0.49→0.24	-49.8%	0.55→0.86	+57.4%	-22.0%	-55.9%
Value-Savvy Pragmatists	0.53→0.45	-16.0%	0.36→0.28	-22.3%	0.66→0.67	+1.4%	-17.4%	-47.1%
Hands-On Pragmatic Testers	0.47→0.41	-12.1%	0.33→0.27	-18.6%	0.52→0.59	+12.4%	-14.4%	-50.7%
Accuracy-Driven Watchdogs	0.59→0.39	-33.6%	0.47→0.26	-44.9%	0.62→0.82	+31.4%	-20.2%	-57.1%
Macro-Tribe: The Practical Family Purchaser (Cluster 4)								
Hands-On Value Advocates	0.48→0.46	-4.4%	0.31→0.23	-27.1%	0.71→0.83	+17.2%	-13.2%	-45.8%
Hands-On Value Seekers	0.51→0.43	-14.9%	0.34→0.28	-16.5%	0.76→0.88	+14.6%	-13.3%	-39.4%
Practical Family Shoppers	0.70→0.25	-64.1%	0.51→0.15	-70.8%	0.83→1.00	+20.0%	-8.9%	-37.3%
Hands-On Value Advocates	0.61→0.43	-29.6%	0.43→0.27	-37.3%	0.54→0.93	+70.7%	-15.9%	-66.6%
Kid-Approval Shoppers	0.70→0.39	-43.9%	0.54→0.29	-47.0%	0.75→0.83	+11.1%	-28.5%	-58.0%
Practical Family Pleasers	0.58→0.39	-32.9%	0.41→0.24	-41.6%	0.74→0.88	+19.1%	-13.8%	-47.0%
Sensory-Savvy Pragmatists	0.53→0.43	-20.1%	0.36→0.27	-25.9%	0.64→0.75	+17.4%	-14.8%	-41.2%
Health-Safe Detailers	0.35→0.32	-6.2%	0.22→0.20	-9.6%	0.56→0.50	-10.0%	+0.2%	-6.0%
Kid-Delight Parents	0.61→0.45	-25.0%	0.44→0.30	-30.9%	0.68→0.83	+23.3%	-16.5%	-48.6%
Practical Beauty Enthusiasts	0.59→0.47	-21.1%	0.41→0.29	-28.9%	0.64→0.69	+7.0%	-19.3%	-42.3%
Hands-On Pragmatic Testers	0.56→0.42	-24.3%	0.40→0.27	-32.5%	0.59→0.95	+61.5%	-16.0%	-35.8%
Practical Parent Purchasers	0.60→0.39	-35.5%	0.43→0.24	-43.5%	0.57→0.80	+41.2%	-16.3%	-43.2%
Family-Focused Pragmatists	0.50→0.39	-22.2%	0.35→0.24	-31.1%	0.55→0.65	+16.9%	-14.7%	-44.2%
Expert Helper Reviewers	0.60→0.53	-11.7%	0.42→0.35	-17.6%	0.61→0.71	+16.0%	-19.8%	-40.7%
Pragmatic Parent Testers	0.52→0.24	-54.4%	0.35→0.14	-61.2%	0.56→1.00	+80.0%	-5.8%	-26.1%
Value-Seeking Self-Care Caregivers	0.59→0.40	-32.5%	0.41→0.24	-41.7%	0.50→0.71	+42.9%	-22.4%	-33.9%
Practical Kid-Centric Parents	0.69→0.47	-32.3%	0.49→0.28	-43.6%	0.92→0.83	-9.1%	-29.3%	-48.3%
Macro-Tribe: The DIY Dollar-Saver (Cluster 5)								
DIY Reliability Seekers	0.66→0.46	-29.4%	0.47→0.38	-19.9%	0.56→0.76	+34.7%	-11.4%	-39.7%
Frugal DIY Fixers	0.65→0.49	-24.7%	0.48→0.32	-32.8%	0.77→0.87	+13.1%	-12.7%	-48.4%
Cost-Savvy DIY Fixers	0.63→0.47	-25.8%	0.48→0.31	-34.5%	0.67→0.84	+25.9%	-15.3%	-56.6%
OEM-Focused Fixers	0.64→0.42	-33.8%	0.46→0.27	-42.4%	0.73→0.88	+21.5%	-15.5%	-59.1%
Pragmatic Repair Technicians	0.60→0.47	-21.2%	0.44→0.31	-28.3%	0.70→0.85	+20.4%	-16.2%	-51.0%

F Dataset and Thematic Distributions

Category	Themes	Thematic Evaluation Criteria
All Beauty	8	Customer Support, Shipping & Service; Ease of Use, Application & Maintenance; Ingredient Safety & Transparency; Sustainability & Eco-Friendliness; Product Performance & Effectiveness; Quality, Durability & Reliability; Sensory & Aesthetic Experience; Value for Money
Appliances	7	Customer Service, Shipping & Packaging; Design & Ergonomics; Ease of Installation, Use & Maintenance; Compatibility & Fit with Existing Appliances; Performance & Functionality + Sensory Experience; Product Quality & Durability; Value for Money & Authenticity
Clothing & Shoes	7	Care, Laundering & Maintenance Requirements; Comfort, Support & Wearability; Fit & Sizing Accuracy; Functional Features & Performance; Material Quality, Durability & Construction; Style, Color & Aesthetic Appeal; Value, Versatility & Purchase Experience
Digital Music	7	Authenticity, Rarity and Collectability; Artist Background and Historical Context; Emotional, Nostalgic and Therapeutic Impact; Packaging, Liner Notes and Metadata Accuracy; Songwriting, Musicianship and Stylistic Authenticity; Sound Quality and Production/Mastering; Track Selection and Album Curation
Health & Personal	9	Build Quality & Durability; Cleaning & Maintenance; Ease of Use & Convenience; Effectiveness & Precision; Environmental & Packaging Impact; Power & Battery Performance; Safety & Health Considerations; Comfort & Sensory Experience; Value for Money
Software	6	Content Availability, Licensing & Online Requirements; Customer Support, Pricing & Marketing Transparency; Feature Set, Usability & Accessibility; Gameplay Experience & Educational Value; Installation & Technical Performance; Hardware Integration & Performance
Video Games	7	Audio and Visual Quality; Content Depth Storyline and Appropriateness; Customer Service Support and Delivery; Gameplay Mechanics Controls and Difficulty + Accessibility & Inclusive Design; Hardware Build Quality Comfort and Connectivity + Battery Life and Portability; Performance Stability and Bug Issues; Pricing Value and Monetization

Table 9: **Domain-Specific Thematic Taxonomies.** To ensure rigorous evaluation across both technical and aesthetic domains, the SAPIENS evaluation framework quantified distributional and geometric alignment (JSD/WD) across 51 distinct semantic parameters derived from the Amazon Reviews '23 dataset.

Macro-Tribe (Cluster Name)	Micro-Tribes	Total Users	Train Reviews	Test Reviews
The Pragmatic Performance Seeker	26	518	2,856	674
The One-Tap Cheerleader	25	535	3,025	564
The Audiophile Archivist	10	224	1,561	457
The Straight-Shooting Bargain Scout	53	1,281	5,500	1,196
The Practical Family Purchaser	27	570	5,791	1,777
The DIY Dollar-Saver	9	228	1,895	397
Outliers (Low-Quality Reviews)	-	*1,646*	*7*	*19*
Grand Total	**150**	**5,002**	**20,635**	**5,084**

Table 10: **Macro Population Distribution via Tribe Discovery.** A macro-level overview of the population distribution. The initialization of the SAPIENS framework successfully mapped 3,356 high-signal users into 150 dense micro-tribes, filtering out low-quality interactions into an outlier cluster to ensure high-fidelity, statistically reliable tribal priors for downstream generation.

F.1 Exhaustive Micro-Tribe Population Database

Table 11: **Exhaustive Micro-Tribe Population Breakdown.** For total demographic transparency, the complete structural breakdown of the 150 discovered micro-tribes generated during the Phase 1 initialization.

Micro-Tribe Archetype	Total Users	Train Reviews	Test Reviews
Macro-Tribe: The Pragmatic Performance Seeker			
Analytical Quality Guardians	30	82	11
Balanced-Challenge Critics	12	29	5
Comfort-First Gear Enthusiasts	12	36	9
Comfort-First Pragmatists	11	38	0
Community-Minded Enthusiasts	30	115	15
Discerning Nostalgic Collectors	16	71	12
Durability-Driven Gamers	22	93	28
Family Tech Pragmatists	14	76	23
Fun-First Casual Enthusiasts	10	29	25
Hands-On Gear Analysts	30	306	125
Hands-On Value Techies	12	14	16
Nostalgia-Driven Tech Critics	11	14	26
Nostalgia-Fueled Fun-Seekers	30	116	66
Performance-Value Gamers	30	160	34
Practical Nostalgia Gamers	17	80	14
Practical Setup Optimizers	30	121	38
Practical Value-Savvy Techs	12	171	36
Pragmatic Value Advocates	30	144	15
Pragmatic Value Gamers	24	189	101
Strategic Value Players	30	151	26
Pragmatic Value Seekers	12	151	13
Quality-Focused Gamer Critics	10	144	10
Tech-Savvy Optimizers	25	106	24
Technical Performance Enthusiasts	30	116	14
Value-Focused Gamer Guides	18	156	10
Value-Minded Time-Killers	10	43	8
Macro-Tribe: The One-Tap Cheerleader			
Aesthetic-Driven Enthusiasts	16	128	6
Brief Pragmatic Shoppers	29	153	14
Comfort-First Enthusiasts	30	171	9
Concise Confirmers	30	186	26
Concise Practical Shoppers	25	248	15
Convenience-First Enthusiasts	21	181	28
Convenience-First Shoppers	30	263	77
Easygoing Value-Seekers	17	55	12
Efficient Bargain Hunters	30	85	24
Effusive Cheerleaders	22	50	8
Enthusiastic Comfort Seekers	10	54	21
Fast-Function Pragmatists	13	56	23
Instant Convenience Seekers	15	53	15
Joy-First Enthusiasts	30	179	16
Joy-Focused Gift Givers	17	24	18
Practical Aesthetes	30	110	15
Quick-Enjoyment Enthusiasts	30	72	62
Quick-Glance Enthusiasts	13	124	7
Quick-Reaction Shoppers	19	140	22
Sensory-Driven Enthusiasts	20	48	26
Snap Impression Enthusiasts	16	129	59
Stylish Bargain Boosters	10	43	15
Upbeat Fun-First Shoppers	14	67	25
Upbeat Fun-Seekers	18	57	22
Upbeat Practical Enthusiasts	30	82	9
Macro-Tribe: The Audiophile Archivist			

Continued on next page

Table 11 – continued from previous page

Micro-Tribe Archetype	Total Users	Train Reviews	Test Reviews
Archival Audiophile Historians	30	191	31
Archival Rock Historians	14	179	5
Audiophile Archivists	20	111	10
Authenticity-Obsessed Archivists	30	172	19
Meticulous Music Archivists	11	148	102
Nostalgia-Driven Music Archivists	30	153	112
Nostalgic Sound Collectors	11	152	127
Scholarly Audiophile Collectors	18	117	16
Scholarly Audiophile Curators	30	172	20
Scholarly Music Archivists	30	166	15
Macro-Tribe: The Straight-Shooting Bargain Scout			
Accuracy-Driven Watchdogs	10	220	101
Blunt Quality Guardians	21	153	0
Comfort-First Minimalists	11	128	3
Comfort-Fit Pragmatists	30	85	13
Consumer Watchdogs	30	144	14
Cost-Conscious Consumer Advocates	12	248	16
Durability Value Seekers	13	47	11
Function-First Comfort Shoppers	30	192	13
Function-First Organizers	24	190	24
Function-First Pragmatists	14	236	9
Utility-Driven Pragmatists	30	123	24
Functionality-First Pragmatists	30	108	19
Outcome-Oriented Pragmatists	21	155	78
Use-Case Optimizers	30	126	34
Hands-On Pragmatic Testers	10	89	20
Hands-On Value Pragmatists	30	120	22
Meticulous Value Advocates	17	96	28
No-Nonsense Fixers	30	116	35
No-Nonsense Reliability Watchdogs	25	58	10
No-Nonsense Value Guardians	30	87	12
Practical Comfort Seekers	30	115	69
Everyday Comfort Buyers	11	26	4
Ergonomic Value Seekers	14	36	6
Practical Convenience Seekers	30	92	17
Practical Detail Evaluators	30	96	18
Practical Function Seekers	30	110	32
Practical Performance Seekers	30	116	23
Practical Reliability Seekers	16	75	25
Practical Value Guardians	30	131	30
Value-Conscious Pragmatists	30	107	23
Practical Value Seekers	30	96	19
Economically Rational Consumers	13	150	10
Deliberate Value Buyers	15	150	9
Practicality-First Shoppers	30	111	27
Pragmatic Comfort Seekers	30	87	24
Pragmatic Fix-It Maintainers	10	42	21
Pragmatic Function Seekers	30	72	9
Pragmatic Hands-On Evaluators	30	162	39
Pragmatic Quality Guardians	30	83	8
Pragmatic Quick-Checkers	29	153	10
Fast-Filter Pragmatists	30	82	9
Pragmatic Reliability Seekers	30	129	51
Pragmatic Value Seekers	30	123	35
Value-Driven Decision Makers	30	108	21
Trade-Off Optimizers	30	102	26
Pragmatic Value Seekers	23	168	16
Measured Investment Buyers	30	105	27
Quality-Driven Detailers	30	134	20

Continued on next page

Table 11 – continued from previous page

Micro-Tribe Archetype	Total Users	Train Reviews	Test Reviews
Reliability-Driven Tech Pragmatists	25	125	39
Reliability-Focused Pragmatists	14	55	4
Utility-Driven Pragmatists	30	86	6
Utility-First Pragmatists	13	48	12
Value-Savvy Pragmatists	20	74	21
Macro-Tribe: The Practical Family Purchaser			
Analytical Supplement Sleuths	10	193	28
Expert Helper Reviewers	14	139	70
Family-Focused Pragmatists	20	194	26
Family-Fun Enthusiasts	30	94	9
Gift-Focused Grand Givers	16	240	59
Hands-On Value Advocates	30	377	120
Hands-On Value Beauty Testers	21	194	101
Hands-On Value Seekers	30	318	152
Health-Safe Detailers	30	417	157
Kid-Approval Shoppers	30	391	69
Kid-Delight Parents	30	81	20
Practical Beauty Enthusiasts	30	319	151
Practical Family Caregivers	30	390	217
Practical Family Evaluators	11	120	34
Practical Family Pleasers	16	258	135
Practical Family Shoppers	30	122	24
Household Utility Buyers	30	110	18
Practical Family Testers	16	181	22
Practical Kid-Centric Parents	21	78	22
Practical Parent Purchasers	17	47	3
Practical Parent Validators	19	76	20
Pragmatic Parent Testers	13	47	17
Protective Caregiver Shoppers	10	37	12
Safety-Savvy Budget Parents	12	137	40
Sensory-Savvy Pragmatists	30	393	151
Value-Seeking Parent Shoppers	10	238	6
Value-Seeking Self-Care Caregivers	14	230	73
Macro-Tribe: The DIY Dollar-Saver			
Cost-Savvy DIY Fixers	30	198	83
Economical Self-Solvers	21	153	67
DIY Reliability Seekers	30	197	2
Frugal DIY Fixers	30	226	22
No-Nonsense DIY Fixers	30	216	21
OEM-Focused Fixers	14	82	19
Practical DIY Fixers	30	182	63
Practical Repair Pros	30	243	111
Pragmatic Repair Technicians	13	98	30