

Inverse Theory of Mind Modeling for Content Recommendation: From Web Browsing to Dynamic Intelligent Interfaces

Mengyu Chen*
mengyu.chen@jpmchase.com
JPMorganChase
New York City, NY, USA

Feiyu Lu*
feiyu.lu@jpmchase.com
JPMorganChase
New York City, NY, USA

Chun-Fu (Richard) Chen†
richard.cf.chen@jpmchase.com
JPMorganChase
New York City, NY, USA

Lucas Vinh Tran†
lucas.vinhtran@chase.com
JPMorganChase
London, UK

Jay Katukuri
jay.katukuri@chase.com
JPMorganChase
Palo Alto, CA, USA

Abstract

Modern recommender systems treat observed actions as reliable proxies for user preferences, yet interactions often reflect exploration or comparison rather than stable preference expression. As interfaces evolve from static layouts toward generative UIs and immersive extended reality (XR), the need for deeper, modality-agnostic user understanding grows: these adaptive environments must decide not only *what* to present but *where*, *when*, *how prominently*, and most importantly *why* a user acts. We propose an Inverse Theory of Mind (IToM) pipeline that reasons backward from observed interactions to infer the beliefs, preferences, and decision-making traits that explain behavior. The pipeline reconstructs each user's decision context, including what was chosen and what alternatives were available, applies LLM-driven counterfactual reasoning to produce evidence-grounded natural-language belief statements, and synthesizes these beliefs through multi-hypothesis abductive inference into a structured *user persona*. We evaluate on the OPeRA dataset against ground-truth personality assessments, attitudinal surveys, and interview-based personas across four tasks: next action prediction, shopping attitude alignment, Big Five personality inference, and held-out category prediction. Results show that inferred personas match or exceed ground-truth personas and that multi-hypothesis reasoning is essential for accurate personality prediction. We further demonstrate cross-modal transferability with a persona-driven spatial banking application on VisionOS.

CCS Concepts

• **Information systems** → **Recommender systems**; • **Human-centered computing** → *Mixed / augmented reality*; • **Computing methodologies** → **Reasoning about belief and knowledge**; **Natural language generation**.

*The first two authors contributed equally to this research.

†Corresponding authors.



This work is licensed under a Creative Commons Attribution 4.0 International License. RecSys '26, Minneapolis, MN, USA

© 2026 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-2284-4/2026/09

<https://doi.org/10.1145/3773078.3831787>

Keywords

Theory of Mind, User modeling, Persona synthesis, Recommender systems, LLM reasoning, Generative UI, Extended reality, Adaptive interfaces, Persona prediction

ACM Reference Format:

Mengyu Chen, Feiyu Lu, Chun-Fu (Richard) Chen, Lucas Vinh Tran, and Jay Katukuri. 2026. Inverse Theory of Mind Modeling for Content Recommendation: From Web Browsing to Dynamic Intelligent Interfaces. In *20th ACM Conference on Recommender Systems (RecSys '26)*, September 27–October 02, 2026, Minneapolis, MN, USA. ACM, New York, NY, USA, 11 pages. <https://doi.org/10.1145/3773078.3831787>

1 Introduction

Recommender systems have made substantial progress in predicting user behavior from large-scale interaction histories, including clicking, viewing, and scrolling. Methods such as collaborative filtering [24, 34], sequential models [32, 69], and generative retrieval with semantic item IDs [57] recommend items by learning latent patterns from past interactions. More recently, Large Language Model (LLM)-based recommenders have emerged, using language understanding and reasoning for more conversational and explainable recommendations [18, 81].

Despite these advances, there is limitation with such methods. *Interaction based* systems learn statistical patterns over a specific behavior type (e.g., clicks, purchases, ratings) and can only predict the next behavior of that same type [23, 43]. Even recent approaches that scale sequence length [23] or integrate collaborative semantics into LLMs [87] remain bound to this scope. At the same time, user interfaces are evolving toward generative UIs [40, 70] that compose novel layouts on the fly, which aims at offering personalized experiences that adapt to particular users and devices (e.g., mobile, desktop, and spatial extended reality (XR) systems). These emerging paradigms need to decide not only *what* to show but *where*, *when*, and *how* to present it, which demands a richer model of the user than click histories can provide. *Persona-based* approaches offer that richer model by constructing profiles that capture interests, goals, personality traits, values, and life context [49, 83], but they traditionally require qualitative data such as user interviews, demographic surveys, or manually authored persona templates [8, 39], which are expensive, intrusive, and difficult to scale.

The common missing piece is an understanding of the *rationale* behind user behavior. In cognitive science, reasoning about others'

beliefs, goals, and intentions is known as Theory of Mind (ToM) [54, 78]. ToM is well suited to user modeling because it is context-dependent, interpretable, and modality-agnostic. By applying ToM *inversely*, reasoning backward from observed actions to the beliefs that generated them, we can extract persona-level understanding from ordinary interaction traces without requiring qualitative data.

Building on this perspective, we propose an *Inverse Theory of Mind* (IToM) pipeline that performs *abductive inference*, reasoning from observed behavior to its best explanation, over user interaction traces. Our pipeline transforms raw event metadata into human-readable belief representations through action summarization, context reconstruction, and optimization. We then synthesize these user beliefs into a structured *user persona*: a cognitive profile capturing the user’s interests, motivational drives, life context, and values. We use LLM as the reasoning engine at each stage, producing evidence-grounded belief statements and persona profiles rather than opaque latent vectors. Once constructed, the persona can be used as a portable user model for tasks such as personality-aligned attitude inference, shopping preference forecasting, or content adaptation for novel interfaces. Applications can use the persona through standard prompting without task-specific retraining [83].

We validate the approach on the OPeRA dataset [76], which pairs fine-grained user action logs from Amazon shopping sessions with ground-truth interview transcripts, personality assessments, and shopping attitude surveys.

Our contributions are:

- We propose *Inverse Theory of Mind* as a paradigm that bridges interaction-based recommenders (data-rich but scope-bound) and persona-based methods (rich understanding but dependent on costly qualitative data), and instantiate it as a multi-stage abductive inference pipeline that reconstructs decision contexts, infers counterfactual-grounded beliefs, and synthesizes them into a modality-agnostic user persona from behavioral traces alone.
- We introduce *multi-hypothesis persona synthesis*, generating diverse personality interpretations of the same behavioral evidence and aggregating them with adaptive confidence weighting.
- We validate on real-world browsing data (OPeRA), comparing inferred personas against ground-truth interviews and personality assessments, demonstrating that behaviorally inferred personas support downstream tasks (shopping-attitude alignment, personality inference, held-out category prediction) beyond the scope of interaction-based methods.
- We present an illustrative spatial banking prototype on VisionOS as a proof-of-concept for persona portability across interaction modalities, where personas inferred from 2D browsing data inform 3D layout and content selection.

2 Related Work

2.1 User Modeling and Intent Inference

Recommender systems have progressed from matrix factorization on explicit ratings [34] through implicit-feedback methods [28, 60, 71] to deep sequential models that capture temporal patterns via recurrent networks, self-attention, bidirectional transformers and

beyond [16, 26, 32, 69]. Recent generative retrieval and industrial-scale systems push this paradigm further by framing recommendation as token generation over learned item IDs [57] or modeling tens of thousands of interactions [23]. Integrating LLMs into recommendation has opened additional directions, including text-to-text task unification [18], LLM fine-tuning and instruction-following for recommendation [3, 84], zero-shot ranking and conversational recommendation [25, 27], prompt enrichment strategies for personalized recommendation [45], and hybrid collaborative-language reasoning [87, 88]. Recent work explores LLM-driven user interest profiling from interaction histories [11, 48, 61]. PEARL [48] retrieves and re-ranks representative snippets from a user’s existing text; Christakopoulou et al. [11] cluster consumption logs into LLM-generated interest journeys; and UserLLM [61] conditions generation on precomputed embedding-based user representations. A parallel line argues for natural-language user profiles as a route to transparent and scrutable recommendation [56, 59], typically assembling profiles from explicit user statements or curated text. Complementing these, session-based approaches model intra-session dynamics through graph neural networks or attention mechanisms [38, 82], e-commerce models classify browsing versus purchasing intent [22, 43], and probabilistic click models formalize examination and click decisions in web search [12]. In planning and robotics, goal recognition treats intent inference as an inverse problem over observed action sequences [58, 68].

These methods, including LLM-enhanced variants, achieve strong next-item or next-action prediction yet remain scope-bound: they model statistical regularities within a single interaction type, reduce intent to a categorical label or scalar reward, and cannot transfer to tasks requiring a different domain of user understanding, such as personal goals, lifestyle profiling, or adaptation to an unseen interface modality.

Inverse reinforcement learning (IRL) moves closer to latent motivations by recovering reward functions from demonstrated trajectories [9, 53]. Zhao et al. [86] introduced mental reward models that use distributional IRL to infer users’ internal satisfaction, and Wang et al. [75] discover causal rather than co-occurrence relations among user actions, improving both accuracy and explainability. These approaches produce richer signals than categorical intent labels but still operate in numeric reward spaces rather than interpretable representations of why a user acts.

A separate line of work constructs user personas to drive recommendation and simulation [49, 83, 85]. However, these approaches assume that the persona is given as input, authored manually or derived from qualitative sources such as demographic and personality profiles [8, 39], rather than inferred from behavioral traces. Our work bridges these two directions by introducing belief inference as an intermediate representation, constructing personas from behavioral traces without qualitative data.

2.2 Theory of Mind in AI

Theory of Mind (ToM) refers to the cognitive ability to attribute beliefs, desires, and intentions to others [4, 54]. Computational approaches formalize ToM as Bayesian inverse planning over goals and beliefs [1, 2, 30, 46], and neural meta-learning has shown that ToM-like reasoning can be acquired from agent trajectories [55].

Recent multi-agent work applies ToM to LLM-based collaboration through explicit belief tracking, hypothesis-driven goal attribution, and inverse reinforcement learning [14, 37, 80], though these remain confined to game-like or synthetic settings.

Evaluations confirm that LLMs can pass false-belief tasks and perform higher-order belief reasoning [35, 66], though with notable gaps on certain tasks [10, 62, 63, 65, 72]. ToM-inspired reasoning has also entered applied settings recently. Liu et al. [42] show that ToM-aware recommendations increase consumer trust, and Pawar et al. [51] apply ToM to surface user intent from partial interaction sequences. These results indicate that modeling why a user acts yields richer understanding than predicting what they do next, yet all operate in forward-reasoning or single-intent settings.

2.3 Personalization for Adaptive and Spatial Interfaces

Emerging interface paradigms, including generative UIs that compose layout and content dynamically [40, 70] and XR environments that distribute content across the user’s visual field [21], require the system to decide not only what to show but where, when, and how prominently, demanding richer user models than click histories alone can provide. Current XR personalization relies predominantly on in-session physiological and attentional signals such as gaze, head pose, RL-driven placement, and affective biometrics [13, 41, 44, 47]. Li et al. [36] move toward deeper modeling by applying Belief-Desire-Intention reasoning to proactive AR assistance, though their BDI states are hand-designed rather than inferred from data. These methods are effective within a session but face severe cold-start limitations. A complementary strategy bootstraps personalization from data collected in established modalities; however, existing LLM-based persona approaches [8, 39, 49] construct personas from qualitative sources such as surveys or comments, rather than *inferring* them from behavioral traces.

3 Methodology

3.1 ToM-Inspired User Modeling

Bridging the gap between interaction-based and persona-based methods requires extracting the *rationale* behind each action.

This relies on *abductive belief inference* grounded in Theory of Mind (ToM). Rather than reasoning forward from beliefs to predict actions, we reason *backward* from observed actions to infer the beliefs that generated them, a form of inference to the best explanation [2, 30]. Three properties of ToM make it particularly suited to this task.

Context-dependence. ToM infers mental states from the relationship between an action and the alternatives available: a choice is only meaningful relative to what was *not* chosen, so reconstructing the decision context could help us extract the rationale behind an action.

Semantic Structure ToM decomposes user behavior into semantically meaningful primitives (beliefs, desires, and intentions) rather than a single opaque representation. These primitives are compositional (a user model is built up from many small, attributable statements) and directly refer to mental states that could explain specific actions, giving downstream inference a structure that latent-vector representations do not provide. When serialized in natural

language, as in our LLM instantiation, this representation additionally supports interpretable and auditable user profiles.

Modality-agnosticism. Because beliefs describe *why* a user acts rather than encoding interface-specific patterns, they transfer across modalities without retraining. A belief inferred from screen-based browsing is equally actionable for a generative UI engine composing a personalized layout or a spatial XR interface arranging content in the user’s visual field.

From beliefs to personas. While individual beliefs capture micro-level decision rationales, they can be further synthesized into a structured *persona*, a portable cognitive profile of the user’s interests, values, and decision-making style. This synthesis performs a second layer of abductive inference, moving from “what did the user prefer?” to “what kind of person makes these choices?”. The resulting persona could enable not only preference prediction but also explanation of past choices and adaptation to novel contexts where no interaction history exists.

Given these properties, we decompose user understanding into two explicit sub-problems: (i) *percept reconstruction* of the decision context, and (ii) *abductive belief inference* from the observed choice within that context, followed by (iii) *persona synthesis* that integrates atomic beliefs into a coherent cognitive profile. This inverse reasoning requires a flexible engine capable of processing unstructured interface content and performing causal, counterfactual judgments about human behavior, a form of everyday causal reasoning that LLMs have shown increasing competence in [31, 65]. We therefore use LLMs not as recommenders but as approximate inverse ToM reasoners, guided by structured prompts to produce evidence-grounded and human-readable belief statements. LLM reasoning is nevertheless known to regress toward stereotypical priors and to vary across paraphrased prompts [10, 62, 63, 65, 72]; our pipeline mitigates these failure modes by two principles that recur across stages: 1) grounding each inference in a reconstructed choice context rather than free-form recall, and 2) aggregating over multiple competing interpretations rather than committing to a single one.

3.2 Problem Formulation

Let $T_u = \{S_1, \dots, S_K\}$ denote the interaction trace of user u across K sessions. Each session $S_k = (a_1^k, \dots, a_{n_k}^k)$ is a time-ordered sequence of interface actions, where each action

$$a_i^k = (\tau_i, \text{target}_i, \text{ctx}_i, t_i),$$

consists of an action type $\tau_i \in \{\text{click}, \text{input}, \text{scroll}, \text{nav}, \text{tab}\}$, a target element, contextual metadata, and a timestamp. While we instantiate the pipeline on screen-based web interactions, the formulation generalizes to any modality that produces event-level action traces with associated interface context (e.g., mouse clicks, finger taps, gaze selections and gesture inputs in XR, or component-level interactions in generative UIs).

Our objective is to infer a belief set $\mathcal{B}_u = \{b_1, \dots, b_m\}$ where each b_j is a natural language statement capturing a preference, trait, or intention, and to synthesize these beliefs into a structured persona Π_u . We state the *conceptual objective* as follows:

$$\mathcal{B}_u^* = \arg \max_{\mathcal{B}} P(\mathcal{B} \mid T_u) \quad (1)$$

under a rational-agent assumption: observed actions are explainable by the user’s beliefs given the decision context they perceived. The persona is then obtained as

$$\Pi_u = \text{Synthesize}(\mathcal{B}_u^*) \quad (2)$$

where Synthesize performs a second round of abductive inference, integrating atomic beliefs into a coherent cognitive profile of the user’s interests, values, and decision-making style.

Given the persona Π_u , downstream prediction reduces to conditional inference: for a task-specific query q (e.g., “rank these products by preference” or “predict Big Five trait scores”), the system computes

$$\hat{y} = \arg \max_y P(y \mid q, \Pi_u) \quad (3)$$

where the persona provides the contextual grounding that enables the model to reason about what the user would do, how they would think, and what they would prefer in a given situation, without requiring additional interaction data for each new task.

Approximation strategy. Equations 1–3 state inferential *objectives*; exact posterior computation over the space of natural-language beliefs is intractable. We instead adopt an *amortized approximate inference* strategy [19, 67]: an LLM pre-trained on large-scale text has internalized rich priors over human beliefs and decision-making patterns [6, 65]; conditioned on a reconstructed decision context through structured prompts, it acts as an approximate inference engine that produces candidate beliefs without an explicit likelihood function or posterior computation. Eq. 1 can thus be approximated by conditioning the LLM on each decision context in the trace to generate candidate beliefs, which can then be refined for diversity and coverage. Eq. 2 can be approximated by prompting the LLM to synthesize the refined beliefs into a coherent persona, and Eq. 3 can be realized by conditioning the LLM on the resulting persona for task-specific queries. The persona thus acts as an *approximate sufficient representation*: a lossy but task-relevant compression of the full trace that decouples prediction from raw interaction data.

3.3 Pipeline Overview

The pipeline transforms raw interaction events into a structured user persona through five sequential stages:

- (1) **Summarize:** Convert raw data into natural language sentences.
- (2) **Perceive:** Reconstruct the decision context, revealing what alternatives were visible on the page.
- (3) **Infer:** Perform backward belief inference at different levels.
- (4) **Optimize:** Deduplicate and select a diverse, compact belief set.
- (5) **Synthesize:** Compose beliefs into a structured persona.

3.4 Stage 1: Action Summarization

Raw interaction events contain type-specific fields that are not directly interpretable. This stage converts each action into a concise natural language summary, for example:

Input: click on “Cedar Wood (Square Ceramic)” color variant
Output: “User selected the ‘Cedar Wood (Square Ceramic)’ color option.”

Structurally simple actions (scroll, navigation, tab activation) are summarized by deterministic rules, while complex actions (click,

input) are summarized by an LLM that incorporates product details when present.

3.5 Stage 2: Percept Reconstruction

This stage addresses a key limitation of standard behavioral analysis: actions are recorded without the decision context in which they occurred. A user’s selection is only informative relative to the alternatives they could perceive. We reconstruct this context by analyzing the page content alongside the action summary.

For each action, we prompt an LLM with the action summary, a condensed version of the page HTML, and page metadata. The LLM produces a structured *percept* containing:

- **Main choices:** Options highly relevant to the user’s action, representing the primary decision set.
- **Sub choices:** Secondary alternatives that could have been picked.
- **Other choices:** Available but irrelevant options on the page.
- **User choice:** What the user actually did.
- **Perception description:** A summary of the overall page environment and what the user likely encountered.

The three-tier choice classification is central to our belief inference: main choices represent the “consideration set” from which the user selected, sub choices define the broader option space, and other choices provide contrast for what the user did *not* pursue. Together, they form the decision context needed for inverse reasoning.

Example percept (cart page): **Main:** Air freshener variants, checkout. **Sub:** Quantity adjustment, Subscribe & Save. **Other:** Unrelated cart items. **User chose:** Ocean Mist variant — “view details for item in cart.”

3.6 Stage 3: Backward Belief Inference

We consider this the core contribution of our work, which, to the best of our knowledge, little prior research has explored for personalization and recommendations. Given the percept reconstruction from Stage 2 (what the user saw, what they chose, what alternatives existed), we perform *backward* reasoning to infer the beliefs that best explain the observed behavior. We operate at two complementary levels.

3.6.1 Action-Level Beliefs. For each individual action, the LLM receives the complete percept, including action summary, user choice with rationale, all three tiers of available choices, and the perception description, and generates multiple candidate belief inferences (1 sentence each), representing different plausible interpretations of the same observed behavior.

Percept: Selected “Cashmere Vanilla (Tassel Ceramic)” among 8 scent/design variants.
Belief: “Values subtle luxury and aesthetic harmony, choosing a decorative variant over simpler options.”

Action-level beliefs capture *micro-signals*: specific product preferences, fine-grained decision-making traits, and individual action rationales.

3.6.2 Session-Level Beliefs. After processing all actions in a session, the LLM receives the full sequence of perceived items together with the action-level belief hypotheses in temporal order and generates multiple candidate belief summaries (each 1-2 sentences)

covering the session as a whole. Building upon the action-level beliefs, the session-level prompt treats them as competing candidate explanations and identifies which combination best fits the overall pattern.

Session: 33 actions — searched for a specific bicycle tail light, compared handlebar tape colors/specs, read reviews.
Belief: “Dedicated cyclist who prioritizes safety; research-driven purchasing style with methodical comparison before selection.”

Session-level beliefs capture *macro-patterns*: cross-action consistency, temporal dynamics (e.g., exploration narrowing into decision), lifestyle signals (e.g., product category mix), and overall decision-making style. As session-level inference processes dozens of actions at once and may dilute fine-grained signals from individual decisions, action-level beliefs preserve those micro-signals which help capture the full picture.

3.6.3 Grounding and belief faithfulness. Hallucination is constrained by construction that every belief is grounded in a reconstructed percept (Stage 2), must cite the chosen option and a rejected alternative from that percept (Stage 3), and is aggregated across M competing hypotheses (Stage 5) so that single-hypothesis fabrications are diluted rather than propagated.

3.7 Stage 4: Belief Optimization

Stages 1–3 produce a candidate belief set that is heavily redundant (337–3,167 candidates in our evaluation), with common, repetitive shopping-action paraphrases. We select a compact, diverse subset via Maximal Marginal Relevance (MMR) [7] in two phases.

Phase 1: Deduplication and quality scoring. Beliefs are embedded with a dense text encoder; near-duplicate pairs (cosine similarity $> \theta_{\text{dedup}}$) are collapsed, preferring session-level beliefs. Each survivor receives a quality score $q(b)$ based primarily on its *outlier distance* from the embedding centroid, so that rare and distinctive beliefs score high and frequent shopping-action paraphrases score low.

Phase 2: MMR iterative selection. Starting from the highest-scored belief, we iteratively select up to K beliefs by maximizing:

$$\text{MMR}(b) = \lambda \cdot q(b) + (1 - \lambda) \left(1 - \max_{b' \in S} \cos(b, b') \right) \quad (4)$$

where S is the already-selected set and $\lambda = 0.3$ weights diversity at 70%, ensuring that once a topic is covered, remaining variants are penalized. The output is a compact set $\mathcal{B}_u^* = \{b_1, \dots, b_K\}$ covering the user’s full behavioral range with minimal redundancy.

3.8 Stage 5: Multi-Hypothesis Persona Synthesis

Atomic beliefs describe individual episodes; downstream tasks require a holistic view of the user. However, the mapping from browsing behavior to personality is *under-determined*, that the same pattern can arise from different personality types, and committing to a single interpretation risks defaulting to stereotypes.

Stage 5 therefore generates M diverse *persona hypotheses*, each a distinct personality type that plausibly explains all observed behaviors. The LLM receives the optimized belief set and is prompted to (i) separate observable facts from personality interpretations and (ii) consider whether the opposite trait extreme could equally explain the evidence. Each hypothesis produces:

- A **persona profile** with interests, drives, values, and life context, all inferred abductively from the beliefs.
- A structured **shopping profile** listing only product categories with direct evidence from the beliefs.
- Per-trait feature for personality, reflecting how well the behavioral evidence constrains each trait.
- **evidence confidence levels** (strong, moderate, or weak), reflecting how well the behavioral evidence constrains each trait.

Hypotheses are ranked by quality and distinctiveness, and combined via confidence-weighted aggregation. For any scalar dimension d (e.g., a Big Five trait or a Likert attitude item), each hypothesis h independently produces a score $s_{h,d}$; the final prediction is:

$$\hat{s}_d = \frac{\sum_{h=1}^M w_h \cdot s_{h,d} + \pi_d \cdot \alpha_d}{\sum_{h=1}^M w_h + \alpha_d}, \quad (5)$$

where w_h combines hypothesis quality with per-dimension evidence confidence, π_d is a published population prior for dimension d , and α_d is an adaptive prior strength that scales inversely with mean evidence confidence across hypotheses. When evidence is strong the prior vanishes and the hypotheses speak; when weak, predictions regress toward population means, mitigating the risk of overconfident stereotyping.

4 Experiments

We evaluate the IToM pipeline on four complementary tasks, each posed as a research question (RQ) that isolates a different component of the pipeline: (1) next action prediction, testing whether inferred personas improve behavioral simulation compared to ground-truth personas; (2) shopping category prediction, testing downstream predictive power on held-out browsing sessions; (3) Big Five personality prediction, testing whether behavioral traces contain recoverable personality signal; and (4) shopping preference survey alignment, testing whether inferred personas capture self-reported shopping attitudes. RQ1 tests whether the synthesized persona (Stage 5) provides usable behavioral signal downstream; RQ2 tests generalization of the persona to unseen sessions; RQ3 (with its §4.5 ablation) tests whether multi-hypothesis aggregation matters; and RQ4 tests persona alignment with self-reported attitudes.

4.1 Dataset

As our method essentially builds upon a quantitative-to-qualitative inference process, it imposes three simultaneous requirements on the evaluation data: (i) *fine-grained interaction traces* (clicks, scrolls, typed queries), not just item-level ratings; (ii) *rich decision context* (page content, visible alternatives) to reconstruct what the user perceived when acting; and (iii) *qualitative ground truth* (personality assessments, attitudinal surveys, interview transcripts) to verify the inferred qualitative output. Most existing benchmarks satisfy at most one of these. Widely used recommender datasets such as MIND [79] and KuaiRec [17] provide rich interaction logs but no psychological or attitudinal ground truth. Conversely, personality-oriented corpora such as PANDORA [20] supply trait labels linked to text but not the atomic behavioral traces from which beliefs can be inferred. To our knowledge, no benchmark prior to OPeRA jointly provides all three.

We therefore evaluate on the OPeRA dataset [76], which jointly provides all three types of data: atomic-level browser interaction traces with full page context (HTML snapshots) from Amazon shopping sessions, paired with per-user personality assessments, shopping attitude surveys, and post-session interview transcripts. We use the OPeRA *train* split of 46 complete user profiles, personality and shopping surveys.

4.2 Implementation Details

All LLM calls use Claude Opus 4.6 via AWS Bedrock. Text embeddings use Amazon Titan Embed v2 (1024 dimensions). For belief optimization, we use MMR with $\lambda = 0.3$ and deduplication threshold $\theta_{\text{dedup}} = 0.85$, selecting up to $K = 50$ beliefs per user. Multi-hypothesis persona generation uses $M = 7$ hypotheses with temperature 0.85. For personality and attitude inference, responses are predicted using the confidence-weighted aggregation from Eq. 5 (Section 3.8), with population priors set to published Big Five norms.

Per-user persona construction (mean 6.3 sessions, ~ 275 actions) requires ~ 622 LLM calls and $\sim 14.5\text{M}$ input / $\sim 240\text{K}$ output tokens, amortizing to $\sim \$79$ per user on Claude Opus 4.6 ($\sim \$47$ on Sonnet 3.7) at AWS Bedrock pricing (as of April 2026). Construction is one-time; the resulting natural-language persona is reused across all downstream tasks with no per-query re-inference. The pipeline is an unoptimized research implementation where Stage 2 (percept reconstruction) alone accounts for $\sim 96\%$ of input tokens, with no batching of same-page actions, no cross-user caching of layout templates, and no per-action call consolidation.

4.3 RQ1: Next Action Prediction

Setup. We evaluate whether inferred personas can substitute for ground-truth personas in the next action prediction task introduced by OPeRA [76]. Given the full sequence of session actions, we predict the next user action and evaluate along four dimensions following OPeRA: (1) action generation accuracy (exact match of the predicted action), (2) action type F1 over high-level categories (click, scroll, navigation, input, tab), (3) click type F1 over specific click subtypes, and (4) session outcome prediction (purchase vs. terminate). We compare three persona conditions: (i) *GT persona*, the ground-truth persona assembled from OPeRA’s interview transcripts, demographics, personality profiles, and shopping rationale; (ii) *Multi-hyp persona*, our IToM-inferred persona (best-scored hypothesis from Stage 5); and (iii) *No persona*, a control with no user context. All conditions use the same session prefix and prompt template provided by the OPeRA paper. We evaluate on two LLM backends: Claude Sonnet 3.7 (used in the original OPeRA evaluation) and Claude Opus 4.6.

Results. Table 1 presents the results. On Claude Opus 4.6, our inferred persona substantially outperforms the ground-truth persona on action generation accuracy (25.69% vs. 17.06%, +50.6% relative), action type weighted F1 (83.90% vs. 78.20%), and click type weighted F1 (49.78% vs. 43.16%). Both persona conditions outperform the no-persona baseline on action generation and action type F1, confirming that persona conditioning provides meaningful signal. On Claude Sonnet 3.7, our persona achieves higher action generation accuracy (11.78% vs. 8.83%) and session outcome accuracy (65.56% vs. 60.00%) while showing slightly lower action type

Table 1: Next action prediction (15 users, 90 sessions, N=985 instances). Act. Gen.: exact-match accuracy of the full predicted action string; AT Wt / Mac-F1: weighted and macro F1 over high-level action categories (click, scroll, input, etc.); Click Wt-F1: weighted F1 over click subtypes; SO Acc / Wt-F1: accuracy and weighted F1 for predicting session outcome (purchase vs. terminate). Best results per-model in bold.

Model	Persona	Act. Gen. $\uparrow$	AT Wt-F1 $\uparrow$	AT Mac-F1 $\uparrow$	Click Wt-F1 $\uparrow$	SO Acc $\uparrow$	SO Wt-F1 $\uparrow$
Sonnet 3.7	No persona	9.14	84.78	32.28	47.07	62.22	54.67
	GT persona	8.83	86.18	35.53	47.00	60.00	56.47
	Multi-hyp (ours)	11.78	83.26	29.62	46.59	65.56	60.06
Opus 4.6	No persona	19.09	80.21	30.66	47.05	64.44	61.87
	GT persona	17.06	78.20	29.98	43.16	65.56	62.80
	Multi-hyp (ours)	25.69	83.90	32.36	49.78	68.89	68.35

Table 2: Shopping category prediction (N=27). Best in bold.

Method	Mac-P $\uparrow$	Mac-R $\uparrow$	Mac-F1 $\uparrow$	NDCG@5 $\uparrow$	Hit@3 $\uparrow$
Popularity	.266	1.00	.409	.581	88.9
User History	.425	.689	.504	.615	85.2
Recency-Weighted	.371	.753	.458	.615	81.5
Zero-shot LLM [27]	.559	.500	.498	.522	88.9
Structured Summary [11]	.533	.689	.562	.636	88.9
CoT Prompting [77]	.501	.730	.567	.658	88.9
Multi-hyp (ours)	.499	.764	.567	.701	100

F1. Our Sonnet GT persona results are consistent with OPeRA’s published numbers, validating our reproduction.

OPeRA’s own analysis noted that persona conditioning does not consistently improve action generation accuracy and may introduce noise into step-level reasoning, even if it helps action type and click type classification [76]. Our belief-grounded personas partially mitigate this issue: on Claude Opus 4.6, the inferred persona improves action generation (+50.6% relative over GT) and classification metrics simultaneously, suggesting that grounding personas in specific behavioral episodes rather than abstract trait summaries reduces the noise problem OPeRA identified.

4.4 RQ2: Shopping Category Prediction

Setup. To evaluate downstream predictive power, we test whether the persona can predict which product categories a user will shop for in held-out sessions. We use a 70/30 session holdout split: personas are constructed from 70% of each user’s sessions, and the remaining 30% provide ground truth. Ground truth categories are extracted from held-out search terms and product interactions, with LLM-based reclassification into 11 canonical categories (e.g., Health & Personal Care, Electronics & Technology, Home & Kitchen). This yields 27 users with non-empty holdout ground truth (mean ~ 4 categories per user).

Baselines. We compare against two families of baselines (Table 2). *Heuristic* baselines operate on category co-occurrence statistics from training sessions: Popularity predicts globally frequent categories, User History echoes per-user training categories ranked by frequency, and Recency-Weighted applies exponential decay to favor recent sessions. *LLM-based* baselines use the same backbone model (Claude Opus 4.6) and identical raw session data, isolating the contribution of our belief-inference and persona-synthesis stages: 1) Zero-shot LLM directly prompts the model to predict categories from raw action logs [27], 2) Structured Summary first compresses

Table 3: Big Five MAE comparison (N=46). Our multi-hypothesis method outperforms all non-oracle baselines.

Method	Avg MAE↓
Random Uniform (1–5)	1.495
Scale Midpoint (predict 3.0)	1.022
Direct Inference	0.945
Questionnaire (90 items)	0.945
Population Norms	0.836
Multi-Hypothesis (M=7)	0.762
GT Mean (oracle)	0.662

sessions into browsing summaries before predicting [11], and 3) CoT Prompting adds explicit step-by-step reasoning over browsing patterns [77]. We omit neural sequential models [26, 32, 69], as they require item-ID interaction sequences and orders of magnitude more users for training. OPeRA provides 27 users with atomic browsing traces (clicks, scrolls, typed queries) rather than item-level interaction logs, making these architecturally inapplicable.

Results. Table 2 presents the comparison. Among heuristic baselines, Popularity achieves perfect recall by predicting all 11 categories but at low precision (.266), while User History obtains the best heuristic F1 (.504) through direct category overlap. Among LLM baselines, adding prompt structure could possibly improve results: Zero-shot LLM achieves the highest precision (.559) but lowest recall (.500), and CoT Prompting reaches the highest baseline F1 (.567) and recall (.730). Our multi-hypothesis persona matches the best baseline F1 (.567) and achieves NDCG@5 of .701 (+6.5% over CoT) with Hit@3 of 100%.

4.5 RQ3: Big Five Personality Prediction

Setup. OPeRA provides self-reported Big Five scores on a 1–5 scale (Extremely Low to Extremely High) for 46 users. Each inference mode predicts Big Five trait scores from the persona; predictions are compared against ground truth via MAE and Pearson r .

Baselines. We compare against: *Random Uniform* (randomly guess 1–5 per trait), *Scale Midpoint* (always predict 3.0), *Population Norms* (published Big Five means: EX=3.2, AG=3.6, CO=3.4, ES=3.1, IN=3.4), and *GT Mean* (per-trait mean of all 46 users, an oracle baseline requiring access to the ground truth distribution).

Results. As shown in Table 3 and Figure 1, our method achieves an overall MAE of 0.762, outperforming all non-oracle baselines including population norms (0.836, –8.8% relative improvement) and closing 88% of the gap from random guessing to the oracle mean. Four of five traits show positive Pearson correlations with ground truth, with Extraversion approaching significance ($r = +0.280$, $p = 0.060$).

Multi-hypothesis aggregation shifts the aggregate correlation into the positive range on this fixed cohort (bootstrap 95% CI $[-0.05, +0.26]$; $N=46$, dataset-given), reversing the negative aggregate of single-interpretation baselines. We therefore scope RQ3 as a *population-level* claim about the direction of LLM persona bias and its correction; per-trait tests are underpowered at this sample size.

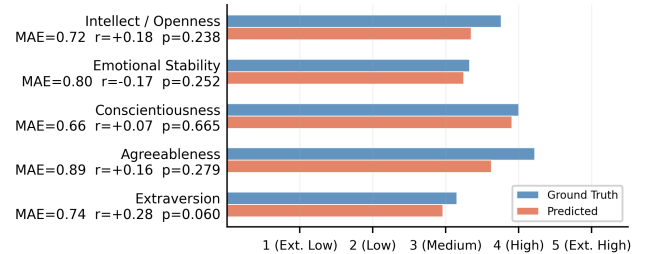
Table 4: Ablation: inference method comparison (N=46). All methods use identical persona profiles. Single-interpretation baselines from the LLM personality literature produce anti-correlated predictions; multi-hypothesis aggregation reverses the sign.

Method	Avg MAE↓	Avg r ↑
Questionnaire [15]	0.945	–0.103
Direct Inference [74]	0.945	–0.008
Ours: Multi-Hyp. (M=7)	0.762	+0.103

Ablation: Why multi-hypothesis reasoning matters. A key question is whether the multi-hypothesis aggregation contributes beyond what simpler inference strategies achieve. We compare against two established approaches from the LLM personality literature, applied to the *same* persona profiles: (i) *Direct inference*, where the LLM reads a single persona and directly estimates trait scores [29, 74]; and (ii) *Questionnaire administration*, where the LLM role-plays as the persona and answers standardized personality items individually [5, 15, 52]. Both approaches have been shown to produce reasonable personality profiles when LLMs role-play from explicit trait descriptions [50, 74]; here we test them on behaviorally-inferred personas where trait levels are *not* given.

Table 4 shows our results that single-interpretation methods produce *negative* average Pearson correlations with ground truth (–0.103 for questionnaire, –0.008 for direct inference), meaning their predictions are *anti-correlated* with actual personality. Both methods achieve MAE=0.945, worse than simply predicting published population norms for every user (0.836). Our multi-hypothesis method reverses the average correlation from –0.103 (Direct Inference) and –0.008 (Questionnaire) to +0.103, a qualitative shift from systematically wrong to directionally correct, and cuts MAE from 0.945 (both unaided-LLM baselines) to 0.762. Population Norms (0.836) is an informed prior of published Big Five means, not an unaided-LLM comparison. See Figure 1 for per-trait statistics between our method and ground-truth.

We term this negative correlation as *persona bias*: the LLM maps ambiguous behavioral evidence to stereotypical personality profiles, producing overconfident predictions that are *worse than knowing nothing*. We analyze this phenomenon further in the Discussion.

**Figure 1: Big Five personality prediction results (N=46). Mean trait scores comparing self-reported ground truth (survey) against our multi-hypothesis ITom predictions. Per-trait MAE, Pearson correlation (r), and p -values are shown alongside each trait.**

4.6 RQ4: Shopping Preference Survey Alignment

Setup. OPeRA includes 12 Likert shopping attitude items per user (1=Strongly Disagree to 5=Strongly Agree), covering research habits, price sensitivity, brand preference, and decision-making style. We treat these as a structured questionnaire: the LLM infers how the persona would respond to each item, and predicted responses are compared numerically against ground truth. Per-item accuracy is computed as $\text{acc}_i = 1 - |y_i - \hat{y}_i|/4$, normalizing Likert distance to $[0, 1]$.

Results. Our method achieves 76.6% normalized accuracy (MAE = 0.937 on a 1–5 Likert scale, $N=46$), outperforming both single-interpretation ablation baselines: direct inference (71.9%) and LLM-administered questionnaire (72.5%).¹

Figure 2 shows per-item accuracy. Several items (e.g., “I usually do a lot of research before making a purchase”) exceed 0.85 accuracy, demonstrating that the persona captures genuine shopping attitude signal across diverse dimensions, from research habits to price sensitivity to decision confidence.

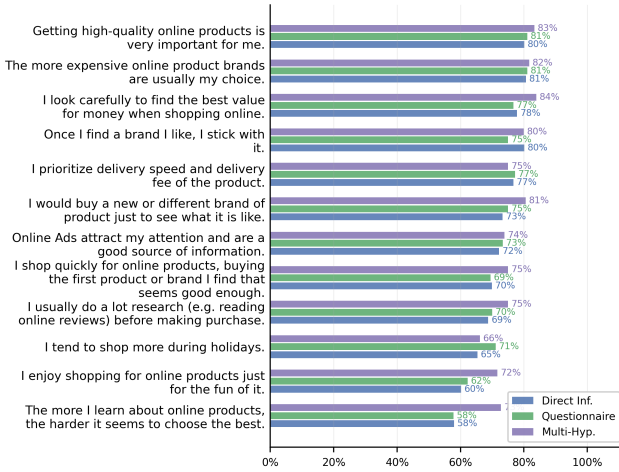


Figure 2: Per-item shopping attitude accuracy (N=46). Our multi-hypothesis method achieves the highest accuracy on 10 of 12 items compared to single-interpretation ablations.

4.7 Discussion

Inferred personas match or exceed ground-truth personas. The action prediction results (Table 1) show that personas inferred purely from behavioral traces achieve competitive or higher scores than ground-truth personas assembled from interviews, demographics, and personality assessments on most prediction metrics (+50.6% relative action generation accuracy on Claude Opus 4.6). This advantage reflects a representational-format difference as much as a user-modeling depth difference: GT personas are abstract trait summaries that anchor the model to a fixed characterization [33], while

¹Direct inference asks the LLM to predict all survey responses from the best persona hypothesis in a single call; the questionnaire mode administers items individually with the LLM role-playing as the persona [15, 52].

our belief-grounded personas encode specific decision episodes that translate more directly into step-level prediction without imposing a rigid role frame. This result should be read as evidence that our episode-encoding format is usable at inference time, not that inferred personas understand the user better than GT personas in an absolute sense: the +50.6% margin may partly reflect prompt-format alignment with the action-prediction task, and a format-controlled comparison (rewriting GT personas into episode style, ablating episodic content from inferred personas) is left to future work. The value of the inferred pipeline is not that it beats GT in absolute score, but that it (a) removes the qualitative-data collection bottleneck (interviews, demographics, personality assessments) that makes GT personas impractical in production, and (b) yields an interpretable, auditable natural-language artifact supporting downstream tasks that raw click-stream data cannot.

Persona bias and multi-hypothesis reasoning. The ablation (Table 4) reveals what we term *persona bias*: single-interpretation methods map ambiguous behavioral evidence to stereotypical personality profiles (e.g., interpreting every careful shopper as an anxious introvert), producing predictions *anti-correlated* with ground truth [5, 29]. This bias is especially severe for behaviorally inferred personas where trait levels are uncertain. Multi-hypothesis reasoning reverses the correlation sign (negative $\rightarrow$ positive) by generating M diverse hypotheses with per-trait evidence confidence and adaptively weighting toward population priors when evidence is weak, providing a “floor guarantee” against persona bias. The underlying mechanism is the *situational* nature of online shopping: users settle for “good enough” options under time pressure, get distracted by promotions, and make choices they might not replicate an hour later due to situational variability and mood swings [64]. A single interpretation forces one coherent story onto inherently noisy actions; multiple hypotheses allow the weight of evidence to point in roughly the right direction without requiring every single hypothesis to be entirely correct.

The role and limits of personality in behavioral modeling. Our results suggest that personality, as captured by standardized instruments like the Big Five, may be more useful for explaining broad attitudinal tendencies than for predicting specific actions. The shopping preference survey (76.6% accuracy) aligns well with inferred personas, likely because survey items ask about general dispositions that map naturally onto personality dimensions. In contrast, traits like Emotional Stability are systematically misread from browsing cues, and even well-calibrated traits (e.g., Conscientiousness) show compressed prediction variance, suggesting the pipeline captures population-level tendencies better than individual extremes. This is perhaps unsurprising: a user’s decision to compare three products before purchasing reflects not only trait-level conscientiousness but also situational factors, domain expertise, price sensitivity, and personal values that personality inventories were never designed to capture [73]. Personality may thus serve better as a *soft prior* that anchors attitudinal inference than as a direct predictor of behavior, where the richer, episode-grounded belief representation carries more of the explanatory weight.

5 Case Study: Persona-Driven Spatial Interface

The experiments above show that IToM-inferred personas capture meaningful behavioral and attitudinal signal. We now present a case study showing how the same persona expressed in natural language can drive a different modality. We illustrate this transferability with a proof-of-concept spatial banking application built on Apple VisionOS due to the substantial discrepancy between traditional web interfaces and spatial user interfaces.

Application. The prototype is a spatial banking app whose layout is fully persona-driven: each user sees a structurally different interface (Figure 3). The interface comprises three spatial panels, each defined by its *function* rather than fixed content: a *center stage* for primary tasks (account management, shopping offers); a *left panel* for financial health tools (credit journey, budgeting, transfers); and a *right panel* for product discovery (credit cards, transaction history). The persona determines *what* appears in each panel and *how* they are arranged (e.g., a budget-conscious user sees spending breakdowns and savings trackers prominently, while an investment-oriented user sees portfolio summaries and market alerts). Users can also rearrange, add, or dismiss widgets—interactions that feed back into belief inference, iteratively refining the persona and subsequent layout adaptations.

Persona as cross-modal bridge. Shopping category predictions (RQ4) map directly to center-stage offer selection, while broader persona fields (life context, values, motivational drives) inform which financial tools and products to surface and how prominently to position them. The persona is used as natural-language context in a task-specific prompt, following the same pattern validated in our experiments. This case study implementation serves as initial evidence of two transfer axes that interaction-based methods cannot easily address: *modality transfer*, where beliefs from 2D click streams inform 3D spatial layout; and *task transfer*, where shopping browsing behavior drives financial product suggestions and widget prioritization, a distinct application with no shared behavioral signal yet informed by the persona’s cognitive profile. This opens opportunities to address cold-start challenges in emerging application domains where data is scarce.

6 Conclusion

We presented an Inverse Theory of Mind pipeline that reasons backward from observed interaction traces to infer the beliefs, intentions, and decision-making traits that explain user behavior, producing structured natural-language personas that capture not only *what* a user prefers but *who* they are. Evaluated on 46 OPeRA users, the pipeline achieves 76.6% shopping attitude accuracy, reduces Big Five MAE below an informed population-norms prior (0.762 vs. 0.836), corrects the sign of the average trait correlation relative to unaided-LLM baselines, and predicts held-out shopping categories with Hit@3=100%. Without our multi-hypothesis reasoning, single-interpretation baselines produce predictions *anti-correlated* with ground truth, while aggregating across diverse persona hypotheses reverses the correlation sign. Because the resulting personas are expressed in natural language rather than interface-specific embeddings, they transfer directly to generative UI engines or spatial XR interfaces without retraining, supporting content *creation*, not just selection.

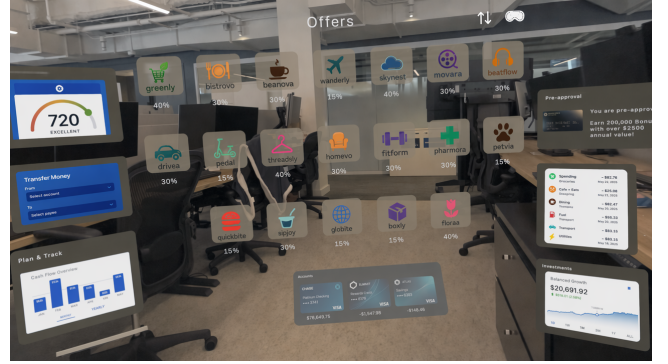


Figure 3: Persona-driven spatial banking prototype on VisionOS. The left and right panels contain different widgets placed in different positions and order, while the center stage surfaces different offers, all determined by the user’s inferred persona constructed from 2D browsing data (All logos and brand names shown are fictitious and for illustrative purposes only. Any resemblance to real trademarks or companies is coincidental; no affiliation or endorsement is implied.)

Limitations and future work. Evaluation is bounded by a single domain (Amazon shopping) and moderate sample size ($N=46$ for personality). Prediction variance is compressed, indicating well-calibrated means but limited differentiation at user extremes. Emotional Stability is anti-predicted due to systematic LLM bias that adaptive priors only partially mitigate. The XR case study is illustrative; a controlled study comparing persona-driven layouts against a non-personalized baseline on task completion, subjective preference, and offer relevance is left to future work. Future work will extend to multi-domain traces, larger populations, end-to-end personalization in spatial and generative interfaces, and calibration methods for improved per-user differentiation. These limitations largely stem from the limited availability of public datasets that provide rich interaction traces, environmental context, and user traits. Our work sheds light on user modeling through the IToM framework and opens new opportunities for future research on data-efficient, generalizable, and personalized interactive systems.

Disclaimer

This paper was prepared for informational purposes by the Global Technology Applied Research and Consumer & Community Banking of JPMorgan Chase & Co. This paper is not a product of the Research Department of JPMorgan Chase & Co. or its affiliates. Neither JPMorgan Chase & Co. nor any of its affiliates makes any explicit or implied representation or warranty and none of them accept any liability in connection with this paper, including, without limitation, with respect to the completeness, accuracy, or reliability of the information contained herein and the potential legal, compliance, tax, or accounting effects thereof. This document is not intended as investment research or investment advice, or as a recommendation, offer, or solicitation for the purchase or sale of any security, financial instrument, financial product or service, or to be used in any way for evaluating the merits of participating in any transaction.

References

- [1] Chris L. Baker, Julian Jara-Ettinger, Rebecca Saxe, and Joshua B. Tenenbaum. 2017. Rational Quantitative Attribution of Beliefs, Desires and Percepts in Human Mentalizing. *Nature Human Behaviour* 1 (2017), 0064.
- [2] Chris L. Baker, Rebecca Saxe, and Joshua B. Tenenbaum. 2009. Action Understanding as Inverse Planning. *Cognition* 113, 3 (2009), 329–349.
- [3] Keqin Bao, Jizhi Zhang, Yang Zhang, Wenjie Wang, Fuli Feng, and Xiangnan He. 2023. TALLRec: An Effective and Efficient Tuning Framework to Align Large Language Model with Recommendation. In *Proceedings of the 17th ACM Conference on Recommender Systems*. 1007–1014.
- [4] Simon Baron-Cohen, Alan M. Leslie, and Uta Frith. 1985. Does the Autistic Child Have a “Theory of Mind”? *Cognition* 21, 1 (1985), 37–46.
- [5] Pranav Bhandari, Usman Naseem, Amitava Datta, Nicolas Fay, and Mehwish Nasim. 2025. Evaluating Personality Traits in Large Language Models: Insights from Psychological Questionnaires. In *Companion Proceedings of the ACM on Web Conference 2025*. ACM, 868–872.
- [6] Marcel Binz and Eric Schulz. 2023. Turning large language models into cognitive models. *arXiv preprint arXiv:2306.03917* (2023).
- [7] Jaime Carbonell and Jade Goldstein. 1998. The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 335–336.
- [8] Shuai Chen, John P. Lalor, Yi Yang, and Ahmed Abbasi. 2025. PersonaTwin: A Multi-Tier Prompt Conditioning Framework for Generating and Evaluating Personalized Digital Twins. In *Proceedings of the Fourth Workshop on Generation, Evaluation, and Metrics*.
- [9] Xiaocong Chen, Lina Yao, Aixin Sun, Xianzhi Wang, Xiwei Xu, and Liming Zhu. 2020. Generative Inverse Deep Reinforcement Learning for Online Recommendation. *arXiv preprint arXiv:2011.02248* (2020).
- [10] Zhuang Chen, Jincenzi Wu, Jinfeng Zhou, Bosi Wen, Guanqun Bi, Gongyao Jiang, Yaru Cao, Mengting Hu, Yungghwei Lai, Zexuan Xiong, et al. 2024. ToMBench: Benchmarking Theory of Mind in Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*. 14507–14532.
- [11] Konstantina Christakopoulou, Alberto Lalama, Cj Adams, Iris Qu, Tal Aber, Daiyi Azzato, Sasa Biber, Stephanie Buchner, Kellie Bui, Mauri Byon, et al. 2023. Large Language Models for User Interest Journeys. *arXiv preprint arXiv:2305.15498* (2023).
- [12] Aleksandr Chuklin, Ilya Markov, and Maarten de Rijke. 2015. *Click Models for Web Search*. Morgan & Claypool Publishers.
- [13] Viviane Clay, Peter König, and Sabine U. König. 2019. Eye Tracking in Virtual Reality. *Journal of Eye Movement Research* 12, 1 (2019).
- [14] Logan Cross, Violet Xiang, Aditya Bhatia, Daniel L. K. Yamins, and Nick Haber. 2024. Hypothetical Minds: Scaffolding Theory of Mind for Multi-Agent Tasks with Large Language Models. *arXiv preprint arXiv:2407.07086* (2024).
- [15] Joost C. F. de Winter, Tom Driessen, and Dimitra Dodou. 2024. The Use of ChatGPT for Personality Research: Administering Questionnaires Using Generated Personas. *Personality and Individual Differences* 225 (2024), 112669.
- [16] Shanshan Feng, Lucas Vinh Tran, Gao Cong, Lisi Chen, Jing Li, and Fan Li. 2020. HME: A Hyperbolic Metric Embedding Approach for Next-POI Recommendation. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25–30, 2020*. ACM, 1429–1438.
- [17] Chongming Gao, Shijun Li, Wenqiang Lei, Jiawei Chen, Biao Li, Peng Jiang, Xiangnan He, Jiaxin Mao, and Tat-Seng Chua. 2022. KuaiRec: A Fully-observed Dataset and Insights for Evaluating Recommender Systems. In *Proceedings of the 31st ACM International Conference on Information and Knowledge Management*. 540–550.
- [18] Shijie Geng, Shuchang Liu, Zuohui Fu, Yingqiang Ge, and Yongfeng Zhang. 2022. Recommendation as Language Processing (RLP): A Unified Pretrain, Personalized Prompt & Predict Paradigm (P5). In *Proceedings of the 16th ACM Conference on Recommender Systems*. 299–315.
- [19] Samuel J. Gershman and Noah D. Goodman. 2014. Amortized inference in probabilistic reasoning. *Proceedings of the Annual Meeting of the Cognitive Science Society* 36 (2014).
- [20] Matej Gjurmeković, Mladen Karan, Iva Vukojević, Mihaela Bošnjak, and Jan Šnajder. 2021. PANDORA Talks: Personality and Demographics on Reddit. In *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*. 138–152.
- [21] Jens Grubert, Tobias Langlotz, Stefanie Zollmann, and Holger Regenbrecht. 2017. Towards Pervasive Augmented Reality: Context-Awareness in Augmented Reality. *IEEE Transactions on Visualization and Computer Graphics* 23, 6 (2017), 1706–1724.
- [22] Huifeng Guo, Jinkai Yu, Qing Liu, Ruiming Tang, and Yuzhou Zhang. 2019. Buying or Browsing?: Predicting Real-time Purchasing Intent using Attention-based Deep Network with Multiple Behavior. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1984–1992.
- [23] Wei Guo, Yaochen Lu, Yigang Li, Jiaqi Yang, Yingxiang Tan, Xingyu Zhang, Mingyang Chen, Junwei Zhang, Zhicheng Su, Weiwen Zhang, Yichao Chen, Bo Zheng, and Ruiming Tang. 2025. LONGER: Scaling Up Long Sequence Modeling in Industrial Recommenders. *arXiv preprint arXiv:2504.12561* (2025).
- [24] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural Collaborative Filtering. In *Proceedings of the 26th International Conference on World Wide Web*. 173–182.
- [25] Zhankui He, Zhouhang Xie, Rahul Jha, Harald Steck, Dawen Liang, Yesu Feng, Bodhisattwa Prasad Majumder, Nathan Kallus, and Julian McAuley. 2023. Large Language Models as Zero-Shot Conversational Recommenders. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*. 720–730.
- [26] Balázs Hidasi, Alexandros Karatzoglou, Linas Baltrunas, and Domonkos Tikk. 2016. Session-based Recommendations with Recurrent Neural Networks. In *Proceedings of the 4th International Conference on Learning Representations*.
- [27] Yupeng Hou, Junjie Zhang, Zihan Lin, Hongyu Lu, Ruobing Xie, Julian McAuley, and Wayne Xin Zhao. 2024. Large Language Models are Zero-Shot Rankers for Recommender Systems. In *Proceedings of the 46th European Conference on Information Retrieval*. Springer, 364–381.
- [28] Yifan Hu, Yehuda Koren, and Chris Volinsky. 2008. Collaborative Filtering for Implicit Feedback Datasets. In *Proceedings of the 8th IEEE International Conference on Data Mining*. IEEE, 263–272.
- [29] Minzhi Huang, Xiaolin Zhang, Christopher Soto, and James Evans. 2026. Designing AI-Agents with Personalities: A Psychometric Approach. *Personality Science* (2026).
- [30] Julian Jara-Ettinger. 2019. Theory of Mind as Inverse Reinforcement Learning. *Current Opinion in Behavioral Sciences* 29 (2019), 105–110.
- [31] Aditya Joshi, Asim Ahmad, and Ashutosh Modi. 2024. COLD: Causal Reasoning in cLOsed Daily Activities. In *Advances in Neural Information Processing Systems*, Vol. 37.
- [32] Wang-Cheng Kang and Julian McAuley. 2018. Self-Attentive Sequential Recommendation. In *Proceedings of the IEEE International Conference on Data Mining*. IEEE, 197–206.
- [33] Junseok Kim, Nakyeong Yang, and Kyomin Jung. 2025. Persona is a Double-Edged Sword: Rethinking the Impact of Role-play Prompts in Zero-shot Reasoning Tasks. In *Findings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*. 848–862.
- [34] Yehuda Koren, Robert Bell, and Chris Volinsky. 2009. Matrix Factorization Techniques for Recommender Systems. *Computer* 42, 8 (2009), 30–37.
- [35] Michal Kosinski. 2024. Evaluating Large Language Models in Theory of Mind Tasks. *Proceedings of the National Academy of Sciences* 121, 45 (2024), e2405460121.
- [36] Chenyi Li, Guande Wu, Gromit Yeuk-Yin Chan, Dishita Turakhia, Sonia Castelo Quispe, and Dong Li. 2025. Satori: Towards Proactive AR Assistant with Belief-Desire-Intention User Modeling. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. ACM, 1–24. doi:10.1145/3706598.3714188
- [37] Huao Li, Yu Chong, Simon Stepputtis, Joseph Campbell, Dana Hughes, Charles Lewis, and Subbarao Ramamoorthy. 2023. Theory of Mind for Multi-Agent Collaboration via Large Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. 180–192.
- [38] Jing Li, Pengjie Ren, Zhumin Chen, Zhaochen Ren, Tao Lian, and Jun Ma. 2017. Neural Attentive Session-based Recommendation. In *Proceedings of the 2017 ACM Conference on Information and Knowledge Management*. 1419–1428.
- [39] Renhao Li, Heyang Xia, Xiang Yuan, Qingxiu Dong, Lei Sha, Zhixing Li, and Zhifang Liu. 2025. How Far Are LLMs from Being Our Digital Twins? A Benchmark for Persona-Based Behavior Chain Simulation. In *Findings of the Association for Computational Linguistics: ACL 2025*.
- [40] Yuanchun Li, Hao Hao, Yizhi Ge, Haoyu Yu, and Yunxin Liu. 2024. Personal LLM Agents: Insights and Survey about the Capability, Efficiency and Security. *arXiv preprint arXiv:2401.05459* (2024).
- [41] David Lindlbauer, Anna Maria Feit, and Otmar Hilliges. 2019. Context-Aware Online Adaptation of Mixed Reality Interfaces. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology*. ACM, 147–160.
- [42] Tian Liu, Yueming Xu, Jie Yang, and Kexin Li. 2025. Mindful Machines: Understanding How AI’s Theory of Mind Capabilities Influence Consumer Response to Product Recommendations. *Psychology & Marketing* (2025). doi:10.1002/mar.70022
- [43] Chris Lo, Dan Frankowski, and Jure Leskovec. 2016. Understanding the Relationship between Browsing and Purchasing in a Large-Scale E-Commerce Site. In *Proceedings of the 25th ACM International Conference on Information and Knowledge Management*. 2421–2426.
- [44] Feiyu Lu, Mengyu Chen, Hsiang Hsu, Pranav Deshpande, Cheng Yao Wang, and Blair MacIntyre. 2024. Adaptive 3D UI Placement in Mixed Reality Using Deep Reinforcement Learning. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. ACM, Article 32, 7 pages. doi:10.1145/3613905.3651059

- [45] Hanjia Lyu, Song Jiang, Hanqing Zeng, Yinglong Xia, Qifan Wang, Si Zhang, Ren Chen, Christopher Leung, Jiajie Tang, and Jiebo Luo. 2023. LLM-Rec: Personalized Recommendation via Prompting Large Language Models. *arXiv preprint arXiv:2307.15780* (2023).
- [46] Yusen Mao, Shuangpeng Liu, Qiang Ni, Xudong Lin, and Liang He. 2024. A Review on Machine Theory of Mind. *IEEE Transactions on Neural Networks and Learning Systems* (2024).
- [47] Elena Marchiori, Evangelos Niforatos, and Loris Preto. 2018. Analysis of Users' Heart Rate Data and Self-Reported Perceptions to Understand Effective Virtual Reality Characteristics. *Information Technology & Tourism* 18 (2018), 133–155.
- [48] Sheshera Mysore, Zhuoran Lu, Mengting Wan, Tara Safavi, Jennifer Metzler, Julian McAuley, and Hamed Zamani. 2024. PEARL: Personalizing Large Language Model Writing Assistants via Extraction and Retrieval of User Profiles. In *Findings of the Association for Computational Linguistics: ACL 2024*.
- [49] 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*. 1–22.
- [50] Joon Sung Park, Carolyn Q. Zou, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, and Michael S. Bernstein. 2024. Generative Agent Simulations of 1,000 People. *arXiv preprint arXiv:2411.10109* (2024).
- [51] Shrivastava Viraj Pawar, Bharadwaj Pedapudi, and Pragya Kaushik. 2025. EARL: Early Intent Recognition in GUI Tasks Using Theory of Mind. In *ICML 2025 Workshop on Foundation Models in the Wild*. <https://openreview.net/forum?id=nABg9kZ7JR>
- [52] Max Pellert, Clemens M. Lechner, Claudia Wagner, Beatrice Rammstedt, and Markus Strohmaier. 2024. AI Psychometrics: Assessing the Psychological Profiles of Large Language Models Through Psychometric Inventories. *Perspectives on Psychological Science* 19, 5 (2024), 808–826.
- [53] Alexander Pysakhovich. 2019. Reinforcement Learning and Inverse Reinforcement Learning with System 1 and System 2. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 409–415.
- [54] David Premack and Guy Woodruff. 1978. Does the Chimpanzee Have a Theory of Mind? *Behavioral and Brain Sciences* 1, 4 (1978), 515–526.
- [55] Neil C. Rabinowitz, Frank Perbet, H. Francis Song, Chiyuan Zhang, S. M. Ali Eslami, and Matthew Botvinick. 2018. Machine Theory of Mind. In *Proceedings of the 35th International Conference on Machine Learning*. 4218–4227.
- [56] Filip Radlinski, Krisztian Balog, Bill Byrne, and Karthik Krishnamoorthi. 2022. On Natural Language User Profiles for Transparent and Scrutable Recommendation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '22)*. ACM, 2863–2874.
- [57] Shashank Rajput, Nikhil Mehta, Anima Singh, Raghunandan H. Keshavan, Trung Vu, Lukas Heldt, Lichan Hong, Yi Tay, Vinh Q. Tran, Jonah Saber, Manzil Zaheer, Ed H. Chi, and Jonathon Shlens. 2023. Recommender Systems with Generative Retrieval. In *Advances in Neural Information Processing Systems*, Vol. 36.
- [58] Miquel Ramirez and Hector Geffner. 2010. Probabilistic Plan Recognition Using Off-the-Shelf Classical Planners. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 1121–1126.
- [59] Jerome Ramos, Hossein A. Rahmani, Xi Wang, Xiao Fu, and Aldo Lipani. 2024. Transparent and Scrutable Recommendations Using Natural Language User Profiles. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL '24)*. ACL.
- [60] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2009. BPR: Bayesian Personalized Ranking from Implicit Feedback. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*. 452–461.
- [61] Adam Richardson and Hari Srinivasan. 2024. User-LLM: Efficient LLM Contextualization with User Embeddings. *arXiv preprint arXiv:2402.13598* (2024).
- [62] Maarten Sap, Ronan LeBras, Daniel Fried, and Yejin Choi. 2022. Neural Theory-of-Mind? On the Limits of Social Intelligence in Large LMs. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. 3762–3780.
- [63] Natalie Shapira, Mosh Levy, Seyed Hossein Alavi, Xuhui Zhou, Yejin Choi, Yoav Goldberg, Maarten Sap, and Vered Shwartz. 2024. Clever Hans or Neural Theory of Mind? Stress Testing Social Reasoning in Large Language Models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics*. 2257–2273.
- [64] Herbert A. Simon. 1955. A Behavioral Model of Rational Choice. *The Quarterly Journal of Economics* 69, 1 (1955), 99–118.
- [65] James W. A. Strachan, Dalila Parga Pelaéz, Gyles W. Humphreys, Michel Thiebaut de Schotten, and Giosuè Baggio. 2024. Testing Theory of Mind in Large Language Models and Humans. *Nature Human Behaviour* 8 (2024), 1285–1295.
- [66] Winnie Street, John Oliver Siy, Geoff Keeling, Adrien Baranes, Benjamin Barnett, Michael McKibben, Tatenda Kanyere, Alison Lentz, Blaise Aguerre y Arcas, and Robin Moran. 2024. LLMs Achieve Adult Human Performance on Higher-Order Theory of Mind Tasks. *arXiv preprint arXiv:2405.18870* (2024).
- [67] Andreas Stuhlmüller, Jacob Taylor, and Noah D. Goodman. 2013. Learning stochastic inverses. In *Advances in Neural Information Processing Systems*, Vol. 26.
- [68] Gita Sukthankar, Christopher Geib, Hung Hai Bui, David Pynadath, and Robert P. Goldman. 2014. *Plan, Activity, and Intent Recognition: Theory and Practice*. Morgan Kaufmann.
- [69] Fei Sun, Jun Liu, Jian Wu, Changhua Pei, Xiao Lin, Wenwu Ou, and Peng Jiang. 2019. BERT4Rec: Sequential Recommendation with Bidirectional Encoder Representations from Transformers. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. 1441–1450.
- [70] Amanda Swearngin, Mira Dontcheva, Wilmet Li, Joel Brandt, Morgan Dixon, and Andrew J. Ko. 2018. Rewrite: Interface Design Assistance from Examples. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 1–12.
- [71] Lucas Vinh Tran, Yi Tay, Shuai Zhang, Gao Cong, and Xiaoli Li. 2020. HyperML: A Boosting Metric Learning Approach in Hyperbolic Space for Recommender Systems. In *WSDM '20: The Thirteenth ACM International Conference on Web Search and Data Mining*, Houston, TX, USA, February 3–7, 2020. ACM, 609–617.
- [72] Tomer Ullman. 2023. Large Language Models Fail on Trivial Alterations to Theory-of-Mind Tasks. *arXiv preprint arXiv:2302.08399* (2023).
- [73] Simine Vazire and Matthias R. Mehl. 2008. Knowing Me, Knowing You: The Accuracy and Unique Predictive Validity of Self-Ratings and Other-Ratings of Daily Behavior. *Journal of Personality and Social Psychology* 95, 5 (2008), 1202–1216.
- [74] Yuting Wang, Jiayi Zhao, Deniz S. Ones, Lihong He, and Xiaolin Xu. 2025. Evaluating the Ability of Large Language Models to Emulate Personality. *Scientific Reports* 15 (2025), 1–15.
- [75] Zhenlei Wang, Xu Chen, Rui Zhou, Quanyu Dai, Zhenhua Dong, and Ji-Rong Wen. 2023. Sequential Recommendation with Causal Behavior Discovery. In *Proceedings of the IEEE International Conference on Data Engineering (ICDE)*. 1849–1862.
- [76] Ziyi Wang, Yuxuan Lu, Wenbo Li, Amirali Amini, Bo Sun, Yakov Bart, Weimin Lyu, Jiri Gesi, Tian Wang, Jing Huang, Yu Su, Upol Ehsan, Malihe Alikhani, Toby Jia-Jun Li, Lydia Chilton, and Dakuo Wang. 2025. OPeRA: A Dataset of Observation, Persona, Rationale, and Action for Evaluating LLMs on Human Online Shopping Behavior Simulation. *arXiv preprint arXiv:2506.05606* (2025).
- [77] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *Advances in Neural Information Processing Systems* 35 (2022), 24824–24837.
- [78] Henry M. Wellman. 2014. *Making Minds: How Theory of Mind Develops*. Oxford University Press.
- [79] Fangzhao Wu, Ying Qiao, Jiun-Hung Chen, Chuhan Wu, Tao Qi, Jianxun Lian, Danyang Liu, Xing Xie, Jianfeng Gao, Winnie Wu, and Ming Zhou. 2020. MIND: A Large-scale Dataset for News Recommendation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 3597–3606.
- [80] Haochen Wu, Pedro Sequeira, and David V. Pynadath. 2023. Multiagent Inverse Reinforcement Learning via Theory of Mind Reasoning. *arXiv preprint arXiv:2302.10238* (2023).
- [81] Likang Wu, Zhi Zheng, Zhaopeng Qiu, Hao Wang, Hongchao Gu, Tingjia Shen, Chuan Qin, Chen Zhu, Hengshu Zhu, Qi Liu, Hui Xiong, and Enhong Chen. 2024. A Survey on Large Language Models for Recommendation. *World Wide Web* 27, 5 (2024), 60.
- [82] Shu Wu, Yuyuan Tang, Yanqiao Zhu, Liang Wang, Xing Xie, and Tieniu Tan. 2019. Session-based Recommendation with Graph Neural Networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 346–353.
- [83] An Zhang, Yuxin Chen, Leheng Sheng, Xiang Wang, and Tat-Seng Chua. 2024. On Generative Agents in Recommendation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 1807–1817. doi:10.1145/3626772.3657844
- [84] Junjie Zhang, Ruobing Xie, Yupeng Hou, Wayne Xin Zhao, Leyu Lin, and Ji-Rong Wen. 2023. Recommendation as Instruction Following: A Large Language Model Empowered Recommendation Approach. *arXiv preprint arXiv:2305.07001* (2023).
- [85] Yimeng Zhang, Jiri Gesi, Ran Xue, Tian Wang, Ziyi Wang, Yuxuan Lu, Sinong Zhan, Huimin Zeng, Qingjun Cui, Yufan Guo, Jing Huang, Mubarak Shah, and Dakuo Wang. 2025. See, Think, Act: Online Shopper Behavior Simulation with VLM Agents. *arXiv preprint arXiv:2510.19245* (2025).
- [86] Mengchen Zhao, Yifan Gao, Yaqing Hou, Xiangyang Li, Pengjie Gu, Zhenhua Dong, Ruiming Tang, and Yi Cai. 2025. MTRC: Learning to Align with User Preferences via Mental Reward Models. In *Advances in Neural Information Processing Systems*.
- [87] Bowen Zheng, Yupeng Hou, Hongyu Lu, Zhichao Chen, Wayne Xin Zhao, and Ji-Rong Wen. 2024. Adapting Large Language Models by Integrating Collaborative Semantics for Recommendation. In *Proceedings of the IEEE 40th International Conference on Data Engineering*. IEEE, 1–14.
- [88] Yaochen Zhu, Liang Wu, Qi Guo, Liangjie Hong, and Jundong Li. 2024. Collaborative Large Language Model for Recommender Systems. In *Proceedings of the ACM Web Conference 2024*. 3162–3172.