What Makes a Sale? Simulating End-to-End Seller–Buyer Retail Dynamics with LLM Agents

Jeonghwan Choi[1], **Jibin Hwang**[1], **Gyeonghun Sun**[1]
Minjeong Ban[1], **Taewon Yun**[1], **Hyeonjae Cheon**[2], **Hwanjun Song**[1,3,*]
[1]Korea Advanced Institute of Science and Technology (KAIST)
[2]GS Retail, [3]Cluvion
{hwani.choi, songhwanjun}@kaist.ac.kr

Abstract

Evaluating retail strategies before deployment is difficult, as outcomes are determined across multiple stages, from seller-side persuasion through buyer–seller interaction to purchase decisions. However, existing retail simulators capture only partial aspects of this process and do not model cross-stage dependencies, making it difficult to assess how early decisions affect downstream outcomes. We present `RetailSim`, an end-to-end retail simulation framework that models this pipeline in a unified environment, explicitly designed for simulation fidelity through diverse product spaces, persona-driven agents, and multi-turn interactions. We evaluate `RetailSim` with a dual protocol comprising human evaluation of behavioral fidelity and meta-evaluation against real-world economic regularities, showing that it successfully reproduces key patterns such as demographic purchasing behavior, the price–demand relationship, and heterogeneous price elasticity. We further demonstrate its practical utility via decision-oriented use cases, including persona inference, seller–buyer interaction analysis, and sales strategy evaluation, showing `RetailSim`'s potential as a controlled testbed for exploring retail strategies.

1 Introduction

Designing effective retail strategies requires understanding how decisions made at one stage influence downstream outcomes across interdependent seller–buyer interactions (Ahearne et al., 2022; Wang et al., 2025b; Zhou et al., 2026), particularly in terms of purchase conversion, revenue, and post-purchase feedback. However, this remains a long-standing challenge due to two fundamental sources of complexity. First, retail environments involve multiple interacting parties with diverse preferences and behaviors (Shi et al., 2025; Li et al., 2025), leading to heterogeneous and often unpredictable interaction dynamics. Second, retail interactions unfold over multiple stages (Mukherjee & Chatterjee, 2021; Yao et al., 2025; Gupta et al., 2026), such as sales pitch, pre-purchase inquiry handling, purchase decision, and post-purchase support, making it difficult to trace how decisions propagate across stages and ultimately determine final retail outcomes.

To address these challenges, simulation offers a promising approach by enabling controlled and repeatable analysis of complex interactions without the cost and irreversibility of real-world deployment (Sturley et al., 2018; Jamali & Lazarova-Molnar, 2026). Recent advances in large language models (LLMs) further enhance this paradigm by enabling realistic and scalable multi-agent simulations, opening up new opportunities for studying human-like behaviors under controlled interaction scenarios (Zhang et al., 2025; Chu et al., 2025; Hu et al., 2026). In the retail domain, such environments allow systematic variation of strategies, participant characteristics, and interaction conditions, making it possible to directly measure their impact on key outcomes, including customer service success (Yao et al., 2025;

*Corresponding author.

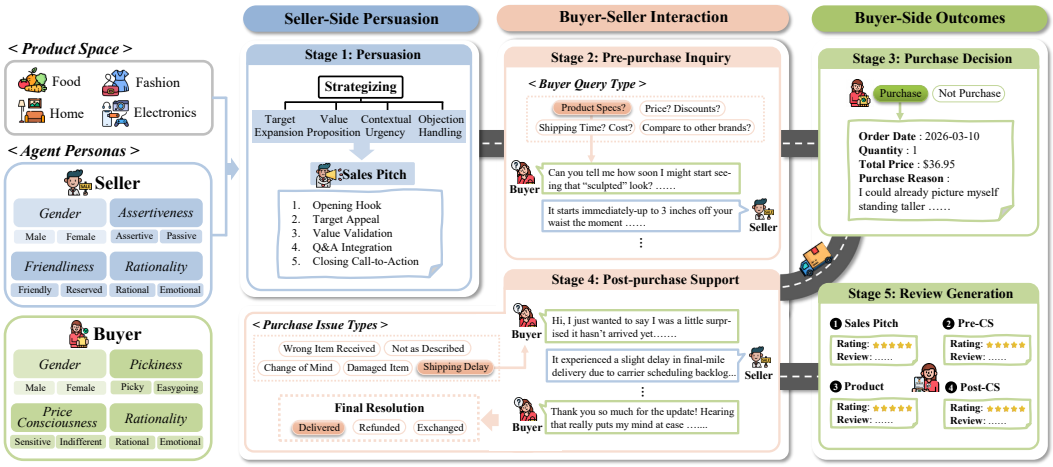

Figure 1: Overview of `RetailSim`: The framework models retail interactions as a unified, multi-stage pipeline, capturing seller strategies, multi-turn interactions, and downstream buyer outcomes, enabling end-to-end analysis of how decisions propagate across stages.

Zhou et al., 2026), recommendation quality (Wang et al., 2025d;c), and negotiation performance (Xia et al., 2024; Bianchi et al., 2024; Zhu et al., 2025).

Despite these advances, existing studies (Wang et al., 2025c; Zhu et al., 2025; Zhou et al., 2026) primarily focus on specific components of the retail process, such as shopping, negotiation, or customer support, without modeling the full progression from seller-side persuasion to buyer-side outcomes within a unified framework. As a result, they do not capture how early decisions influence downstream outcomes, making it difficult to verify whether resulting behaviors remain consistent with fundamental economic principles (*e.g.*, price–demand relationships). Consequently, the economic consistency of such simulations remains unclear, raising concerns about their reliability for decision support (Wang et al., 2025e; Mansour et al., 2025; Bansal et al., 2025). These limitations motivate a unified framework for end-to-end analysis grounded in real-world economic regularities.

To remedy this, we introduce `RetailSim`, a novel multi-agent simulation framework that models the full retail process in a unified setting, capturing dependencies across interactions and enabling the reproduction of real-world economic regularities. As shown in Figure 1, it simulates the real-world retail interaction process as it unfolds over time, including "seller–side persuasion" (*i.e.*, sales pitch generation), "buyer–seller interactions" (*i.e.*, pre- and post-purchase inquiries), and "buyer-side outcomes" (*i.e.*, purchase decisions and review generation), partially treated in prior work. The framework further incorporates three core components to capture real-world complexity, including (i) diverse product categories, (ii) seller and buyer personas capturing behavioral traits, and (iii) multi-turn interactions before and after purchase, covering scenarios such as inquiries, complaints, returns, and refunds. All together, the unified multi-stage design and these components enable controlled, realistic variation and detailed analysis of interaction outcomes.

Beyond the framework, we establish an empirical evaluation protocol that jointly assesses simulation fidelity and its effectiveness. First, for *fidelity*, we propose a two-level, complementary simulation evaluation: at the stage level, we perform human evaluation to verify whether generated interactions align with intended personas and stage-specific behaviors; and at the system level, we go beyond behavioral fidelity and test whether the simulation preserves economic consistency by reproducing fundamental real-world regularities (*e.g.*, the negative relationship btw price and purchase). Second, for *effectiveness*, we demonstrate that `RetailSim` enables decision-oriented analysis by supporting controlled interventions and revealing their downstream effects across the full interaction pipeline. Specifically, we showcase three applications: inferring latent personas from interaction data, analyzing persona-conditioned revenue patterns, and identifying effective sales strategies for improving conversion and revenue. Our work differs from prior studies in four key aspects:

(1) We introduce RetailSim, a causally structured, end-to-end simulation framework that explicitly models cross-stage dependencies in retail interactions, enabling analysis of how early-stage decisions propagate to downstream economic outcomes.

(2) We develop a controllable simulation environment with persona-driven agents, where behavioral factors, including price sensitivity and rationality, can be systematically varied to study their downstream impact on interaction dynamics and outcomes.

(3) We propose a system-level evaluation paradigm based on economic consistency, introducing meta-evaluation against real-world behavioral regularities as a criterion for assessing LLM-based simulations beyond surface-level fidelity.

(4) We demonstrate that RetailSim serves as a scientific testbed for analyzing LLMs as economic agents, enabling measurement of latent behavioral traits and their interaction effects on revenue and decision outcomes.

2 Related Work

Simulation for Complex Interaction Modeling. Modeling multi-stage interactions is challenging as outcomes arise from interdependent decisions over time and are difficult to observe or control in real-world settings, such as healthcare (Almansoori et al., 2025) and retail (Wang et al., 2025b). Simulation mitigates this by enabling controlled, repeatable analysis of how decisions propagate across stages (Jamali & Lazarova-Molnar, 2026). Recent advances in LLMs further extend this paradigm, enabling realistic, multi-turn, and grounded interaction simulation, with growing applications in retail and economic domains (Zhang et al., 2025; Chu et al., 2025; Hu et al., 2026). Building on this, prior retail-focused work largely targets individual stages: post-purchase service (Wang et al., 2025b; Zhou et al., 2026), recommendation and intent matching (Wang et al., 2025d;c; Zhao et al., 2026), and buyer–seller negotiation (Xia et al., 2024; Bianchi et al., 2024; Zhu et al., 2025). Persona-based simulators further introduce heterogeneous agent behaviors to better capture buyer decision styles (Wang et al., 2025f; Bansal et al., 2025; Zhang et al., 2026).

Economic and Behavioral Foundations of Retail. Evaluating retail strategies requires grounding in behavioral and economic principles, as retail outcomes emerge from the accumulation of seller communication, buyer responses, and post-purchase engagement across stages rather than isolated interactions (Wang et al., 2025b; Zhou et al., 2026). In this regard, prior work highlights the importance of personalization and interaction context in shaping conversion and satisfaction beyond static matching (Wang et al., 2025d; Zhao et al., 2026), and recent LLM-based simulations suggest such complex socio-economic behaviors can be reproduced at scale (Li et al., 2024; Chu et al., 2025; Feng et al., 2025).

However, prior studies primarily establish behavioral fidelity at the interaction level, with limited consideration of how seller–buyer interactions connect across the full process or whether their aggregate outcomes reflect real-world economic regularities (Yao et al., 2025; Wang et al., 2025b; Zhu et al., 2025; Wang et al., 2025c; Bansal et al., 2025; Zhou et al., 2026). In addition, they focus on optimizing task-specific outcomes, such as customer service success, recommendation quality, and negotiation performance, without considering how decisions propagate across interactions. As a result, their applicability to real-world use cases such as strategy evaluation and decision support remains limited. We address this gap with a unified framework for joint analysis of seller–buyer dynamics and economic outcomes. Table 17 highlights how our approach differs from 13 existing retail simulation frameworks across key dimensions, including persona-driven role modeling and stage coverage.

3 RetailSim Framework

We introduce a cross-stage dependency model for end-to-end retail simulation, linking seller-side persuasion, buyer–seller interaction, and buyer-side outcomes through structured state transitions. This enables both realistic generation and causal analysis of downstream outcomes. We instantiate this formalization with three key components for simulation fidelity: diverse product spaces, persona-driven agents, and multi-turn interactions.

3.1 Unified Multi-Stage Interaction Simulation

Retail interactions naturally follow a multi-stage process spanning seller-side persuasion, buyer–seller interaction, and buyer-side outcomes, as well established in customer journey modeling (Mukherjee & Chatterjee, 2021; Gupta et al., 2026). To reflect this real-world process, `RetailSim` instantiates a causally structured, end-to-end simulation pipeline across the three stages, where each stage passes its outputs forward as context and interacts with prior stages. Prompts for agent behaviors are provided in Tables 18–34 in Appendix J.

Seller-side Persuasion. The simulation process begins with seller-side strategy formulation and sales pitch generation to drive product sales. Given a product, the seller agent generates a sales pitch, a "product-specific sales script" (see an example in Table 35–36), by prompting an LLM to create the pitch along the following four key dimensions.

- *Target Expansion*: Defining core and secondary target audiences to broaden the reach of the pitch beyond the most direct buyers.
- *Value Proposition*: Translating product features into buyer benefits and justifying price–quality trade-offs via competitive positioning, technical credibility, or cost-per-use logic.
- *Contextual Urgency*: Establishing a product-specific reason to buy now, grounded in situational or seasonal milestones relevant to the buyer.
- *Objection Handling*: Anticipating likely buyer hesitations and preparing persuasive responses to turn concerns into selling points.

By structuring persuasion in this way, the framework generates product-aware, practically grounded sales pitches[1] that reflect real retail strategies and provide a consistent starting point for subsequent buyer–seller interactions.

Buyer–Seller Interactions. Following the generated sales pitch, the buyer is presented with a product-specific sales script, after which interactions proceed around issues arising before and after purchase (*e.g.*, clarifications, concerns, delivery, refund). This process is critical, as outcomes are shaped not only by the initial pitch but by how uncertainties and issues are addressed throughout the buyer–seller interaction (Edelman & Singer, 2015; Lemon & Verhoef, 2016). We model this process as "five-turn buyer–seller conversations": before purchase, interactions capture decision-related issues, and after purchase, they continue for buyers who complete a transaction. All issue types are summarized in Table 14.

- *Pre-inquiry*: The buyer selects inquiry topics of interest in {product specifications, product comparisons, price/discount, shipping} and engages in conversation with the seller to resolve uncertainties before deciding whether to purchase.
- *Post-inquiry*: Only if the buyer completes a purchase, they are assigned an issue type in {shipping delay, wrong item received, change of mind, damaged, not as described} and interact with the seller to reach an outcome in {delivered, exchanged, refunded}.

Buyer-side Outcomes. Buyer-side outcomes occur at two points in the simulation process. After the pre-purchase inquiry stage, the buyer agent makes a "purchase decision," and after post-purchase interactions, the buyer generates "product reviews."

- *Purchase Decision*: The buyer agent decides whether to purchase and how many units to buy, along with a rationale, conditioned on the accumulated context of the sales pitch (product-specific script) and pre-purchase inquiries.
- *Product Review*: Only for buyers who complete a purchase, the buyer generates four-dimensional reviews based on the full history, covering sales pitch effectiveness, pre-purchase support quality, post-purchase resolution, and product satisfaction.

Together, this framework captures the full retail process, enabling end-to-end analysis of how interactions across stages determine final outcomes.

[1]The key dimensions are derived from analysis of seven high-quality real-world sales scripts provided by a retail company with over 5,000 employees. See the details in Appendix B.

3.2 Key Components for Simulation Fidelity

While the unified pipeline enables end-to-end simulation, capturing realistic behavior requires more than modeling the interaction flow. To this end, `RetailSim` incorporates key components that capture real-world variation in products, agent behaviors, and interaction scenarios. These components enable realistic, controlled simulation of retail interactions and systematic analysis of how product, persona, and interaction factors drive outcomes.

Product Space. A realistic product space is essential for capturing variation in buyer preferences and decision contexts (Ding et al., 2023). Accordingly, we construct a realistic product space using publicly available product metadata from the Amazon Reviews'23 dataset (Hou et al., 2024). During simulation, we randomly sample a fixed number of products covering 34 categories, which are grouped into four high-level categories of {Food, Fashion, Home, Electronics} for ease of exposition[2]. For all simulation experiments, we sample an equal number of products from each category to ensure balanced coverage.

Agent Personas. Agent personas capture heterogeneity in seller behaviors and buyer preferences throughout the end-to-end process, influencing interaction dynamics and outcomes (Gao et al., 2024; Mansour et al., 2025). In our framework, we consider two agent roles, "sellers" and "buyers," and assign each agent a demographic attribute (gender, male or female) along with three role-specific behavioral traits for each role, which are injected into the LLM via persona-specific prompts in Appendix J. This design enables controlled variation of seller and buyer personas, supporting diverse simulation scenarios and systematic analysis of their impact on interaction dynamics and outcomes.

Specifically, the seller agent is characterized by *assertiveness* (assertive vs. passive), which controls the intensity and directness of persuasion; *friendliness* (friendly vs. reserved), which governs interpersonal tone from casual and relational to formal and business-like; and *rationality* (rational vs. emotional), which shapes persuasive content from evidence-based arguments to affective and narrative appeals. In contrast, the buyer agent is characterized by *pickiness* (picky vs. easygoing), which determines the strictness of evaluation over product quality, performance, and details; *price consciousness* (sensitive vs. indifferent), which governs how strongly price and value-for-money are weighted in purchase decisions; and *rationality* (rational vs. emotional), which shapes whether decisions are driven by factual evaluation or intuitive impressions. For both roles, each trait is defined as a choice between pre-defined values for tractable analysis across multi-dimensional traits.

Interaction Scenarios. Additionally, interaction scenarios capture the diversity of real-world buyer–seller exchanges and context-dependent dynamics that influence outcomes (Lemon & Verhoef, 2016; Wang et al., 2025a). We thus model realistic scenarios covering issues before and after purchase, including product inquiries, comparisons, pricing and shipping questions, as well as post-purchase issues such as delivery problems, returns, and refunds (as summarized in Table 14 and 15). These interactions are realized as short dialogues, where each response is conditioned on the accumulated simulation trajectory up to that point to ensure factual consistency and coherent progression of the interaction.

4 Simulation Fidelity and Economic Consistency

We evaluate simulation fidelity at two complementary levels: at the *stage* level, we assess whether agents faithfully follow the stage-specific behaviors (*e.g.*, persuasive strategy execution, inquiry handling) and provided personas; at the *system* level, we examine whether the resulting outcomes align with real-world economic principles. These evaluations validate the reliability of `RetailSim` as a simulation testbed, which the decision-oriented analyses in Section 5 build upon.

For the backbone model, `RetailSim` can be instantiated with any LLM. We thus evaluate eight LLMs under a homogeneous setup using the same LLM for both seller and buyer agents, enabling the comparison of their simulation capabilities. The eight LLMs include

[2]The products are categorized into the four high-level product categories using Qwen3-235B-A22B-Instruct based on their original title, description, and price.

Aspect	Category		Qwen3 -80B	Qwen3 -235B	GPT-oss -120B	DeepSeek -V3.2	Gemini-3 -Flash	Gemini-3.1 -Pro	GPT-5.4 -mini	GPT -5.4	Kripp.'s α (IAA)
Task	Sales Script		4.36	4.50	4.45	4.29	4.28	4.40	4.30	4.51	0.67
	Pre-purchase Inquiry		2.73	3.56	3.20	3.32	3.10	2.72	4.45	4.20	0.77
	Purchase Decision		3.21	3.92	3.67	3.76	3.94	3.96	3.99	4.10	0.71
	Post-purchase Inquiry		2.99	3.65	3.75	3.86	3.41	2.90	3.14	4.06	0.77
	Product Review		4.32	4.27	3.73	3.99	3.94	3.97	4.14	4.13	0.71
Persona	Seller	Assertiveness	99.7	99.3	97.9	97.2	99.7	99.3	91.7	95.0	0.81
		Friendliness	99.3	96.5	99.7	99.3	99.3	99.3	94.3	93.6	0.83
		Rationality	90.3	84.0	88.7	91.0	97.9	99.7	84.8	86.8	0.74
	Buyer	Pickiness	98.2	99.4	97.3	98.2	99.1	99.7	91.0	97.6	0.81
		Price Consciousness	97.6	97.6	99.4	99.1	99.4	99.7	93.4	99.4	0.77
		Rationality	99.7	99.4	97.6	99.7	99.7	99.7	97.0	99.7	0.83

Table 1: Simulation fidelity of `RetailSim` with respect to task fidelity (1–5 Likert-scale evaluation) and persona fidelity (accuracy (%) via A/B comparison based evaluation).

both open- and closed-source models of varying sizes: Qwen3-Next-80B-A3B-Instruct, Qwen3-235B-A22B-Instruct, GPT-oss-120B, DeepSeek-V3.2, Gemini-3-Flash, Gemini-3.1-Pro, GPT-5.4-mini, and GPT-5.4. Heterogeneous setups are discussed in Section 5.

4.1 Stage Level Evaluation via Human Evaluation

At the stage level, we run simulations with `RetailSim` and validate their fidelity through human evaluation, following prior work (Majumder et al., 2020; Hwang et al., 2026). We evaluate two complementary aspects: (i) we assess *task fidelity* across five components—sales scripts, pre-purchase inquiries, purchase decisions, post-purchase inquiries, and product reviews—by measuring whether generated outputs adhere to the simulation guidelines and exhibit realistic, coherent behavior, using a 5-point Likert scale (Very Poor to Excellent); and (ii) we evaluate *persona fidelity* by varying persona traits and testing whether they are reflected in agent behavior. Specifically, human annotators are shown A/B pairs of outputs generated under opposing traits (*e.g.*, rational vs. emotional in rationality) and asked to choose which output reflects the target trait (*e.g.*, rational); accuracy (Acc) is measured by whether annotators select the correct one.

For this evaluation, we run simulations on 12 products across all combinations of persona traits for both seller and buyer agents, using each LLM backbone. For task fidelity, we sample 3 outputs per task for each product, resulting in a total of 8(LLMs) $\times$ 12(products) $\times$ 5(tasks) $\times$ 3(samples) $=$ 1,440 outputs for human evaluation. For persona fidelity, we construct A/B pairs by comparing outputs generated under opposing conditions for a target trait, while keeping the other traits fixed. Thus, for each target trait, we collect four A/B pairs, as there are four possible combinations of the remaining two traits. In total, this results in 8(LLMs) $\times$ 12(products) $\times$ 2(roles) $\times$ 3(target traits) $\times$ 4(combinations of remaining traits) $=$ 2,304 A/B pairs for human evaluation. Each task output and each A/B pair is annotated by three human annotators via Amazon MTurk, and we report inter-annotator agreement (IAA) using Krippendorff's α to assess annotation reliability ($\alpha \geq 0.67$: meaningful reliability (Cook et al., 2025; Ban et al., 2026)). Details are in Appendix C.

Results. Table 1 summarizes human evaluation results in terms of task and persona fidelity. Overall, `RetailSim` achieves high task and persona fidelity across all LLM backbones. The Krippendorff α ranges from 0.67 to 0.83, supporting the reliability of our evaluation. In detail, for task fidelity, the sales script and product review tasks consistently achieve strong performance across all LLMs, with scores generally exceeding 4 out of 5. In contrast, for pre-/post-purchase inquiry and purchase decision tasks, larger or proprietary LLMs tend to achieve higher scores, suggesting that these tasks require more advanced reasoning and interaction capabilities. Contrarily, for persona fidelity, all LLMs achieve consistently high accuracy across both roles, with most scores exceeding 95%. This suggests that persona attributes are generally preserved across LLMs. This suggests that `RetailSim` produces a largely consistent persona-conditioned behavior in simulation.

4.2 System Level Evaluation via Meta Evaluation

A simulation can appear realistic at each stage yet fail to reproduce real-world behavioral and economic patterns. However, this system-level validity has not been explicitly

verified in prior work. Therefore, we propose system-level "meta evaluation," where we test whether the simulation reproduces well-established, empirically observed patterns in real-world retail. We validate three economic aspects: (i) *demographic purchasing patterns*; (ii) *price–demand relationship*; and (iii) *heterogeneous price elasticity*. We run simulation using `RetailSim` but with slightly different setups for each of the three aspects. Beyond these, we further disentangle framework-induced patterns from the models' emergent behavior, and validate simulated review ratings against real-world rating data.

Demographic Purchasing Pattern. We examine whether `RetailSim` captures a representative pattern in retail, namely gender-dependent differences in purchasing behavior (Kanwal et al., 2021). To this end, we run simulations in the "Fashion" category using 15 products (5 male-oriented, 5 female-oriented, and 5 unisex). The same setup is applied across all eight LLM backbones to avoid model-specific bias, resulting in 8(LLMs) $\times$ 15(products) $\times$ 16(buyer persona combinations, including gender) $=$ 1,920 simulation runs, with seller personas randomly assigned. Table 2 reveals that male buyers exhibit higher purchase rates for men-oriented products, while female buyers show substantially higher purchase rates for women-oriented products. In contrast, the difference is negligible for unisex products, indicating that `RetailSim` successfully captures gender-aligned purchasing patterns.

Product	Male	Female	Δ(M$-$F)
Men-oriented	54.6%	48.3%	$+6.3$%p*
Women-oriented	32.5%	55.8%	-23.3%p*
Unisex	60.8%	62.5%	-1.7%p

Table 2: Purchase rates by buyer gender across product types, with differences (Δ), where *: $p < 0.05$.

Price Demand Relationship. It is also well established in economics that customers' demand decreases as prices increase, and conversely increases as prices decline (Hitsch et al., 2021). However, it is unclear whether such price–demand dynamics emerge in LLM-driven agent simulations. To verify whether this economic principle emerges with `RetailSim`, we randomly select eight items from the demographic setup and vary their prices across five conditions ($\pm$10%, $\pm$5%, and the original price), conducting additional simulation runs. To examine the price–demand relationship, we analyze how buyers' purchase rates change under different pricing conditions. Table 3 shows that purchase rates decrease monotonically from 64.8% at a 10% discount to 56.6% at a 10% premium, confirming that `RetailSim` reproduces the price–demand relationship observed in economics.

	Price Condition				
	-10%	-5%	0%	$+5$%	$+10$%
Purchase Rate	64.8%	61.7%	57.7%	57.3%	56.6%

Table 3: Buyers' purchase rates under different price conditions.

Heterogeneous Price Elasticity. Under the same setup as the price–demand analysis, we examine whether `RetailSim` captures systematic differences in demand responsiveness across buyers with varying levels of price consciousness, reflecting the well-established principle that more price-sensitive consumers exhibit greater demand elasticity (Phillips, 2021). We partition buyers into two groups based on the price consciousness trait—sensitive and indifferent—and compute their per-product price elasticity, defined as the percentage change in purchase rate relative to the percentage change in price (see Eq. (1) of Appendix D), quantifying how strongly purchase probability responds to price variation; The higher the elasticity, the more responsive buyers are to price changes. We observe that price-sensitive buyers exhibit much higher elasticity (2.58) than price-indifferent buyers (0.76), a large difference of 1.82 that is statistically significant ($p < 0.001$), confirming that `RetailSim` captures real-world heterogeneity in price responsiveness. Appendix G further shows that persona-driven differences in price sensitivity align with real-world behavior.

Disentangling Framework and Model Effects. While the above patterns may appear intuitive, it remains unclear whether they emerge from the framework's design or simply from the capability of modern LLMs. To disentangle the two, we re-run the demographic and elasticity experiments under two ablated prompts. *No-Persona* removes the buyer persona block and the gender clause from the purchase decision, and *No-Anchor-Price* further neutralizes the price framing. This results in 8(LLMs) $\times$ 16(buyer personas)

	Gender Δ(M$-$F)		Elasticity	
Condition	Men-or.	Women-or.	$\lvert E_d \rvert$	Ratio
Original	$+6.3$%p*	-23.3%p*	0.81	3.40$\times$*
No-Persona	-2.1%p	-1.3%p	0.89	3.74$\times$*
No-Anchor-Price	–	–	0.48	2.25$\times$*

Table 4: Meta-evaluation results under ablated prompts, where *: $p < 0.001$.

$\times$ 5(price conditions) $\times$ 15(products) $=$ 9,600 additional runs (see Appendix E for details). Table 4 reveals that the gender-aligned pattern collapses to non-significance under No-Persona, indicating that this regularity emerges through the framework's persona injection rather than the models alone. In contrast, the elasticity gap remains significant under No-Anchor-Price, where buyers are given neither persona definitions nor price framing, reflecting a baseline price reasoning intrinsic to the models. These results indicate that `RetailSim` converts model capabilities into controlled, economically consistent simulation.

Review Rating Validation. Beyond behavioral regularities, we examine whether the simulated review ratings align with real ratings in the Amazon Reviews'23 dataset (Hou et al., 2024). To this end, we ground each buyer's product experience in the textual content of real reviews, without exposing their star ratings, and simulate reviews over 120 products, resulting in 2(LLMs) $\times$ 120(products) $\times$ 32(buyer personas) $=$ 7,680 simulated reviews. Table 5 shows that the simulated per-product mean ratings strongly correlate with the real means, and that the simulated distribution reproduces the J-shaped pattern of real reviews with minor deviations in mid-range ratings. These results confirm that the review mechanism of `RetailSim` faithfully converts product experience into ratings. Details, including full rating distributions, are in Appendix F.

Model	Pearson	Spearman	Mean	W_1
Real (Amazon'23)	–	–	4.20	–
Qwen3-235B	+0.60	+0.59	3.96	0.25
DeepSeek-V3.2	+0.60	+0.58	3.86	0.34

Table 5: Validation of simulated ratings against real ratings, with product-level correlations, mean ratings, and Wasserstein-1 (W_1) distance to the real distribution.

5 Application of RetailSim across Three Use Cases

While prior work focuses on evaluating simulation fidelity, it remains unclear how such simulations support practical, decision-oriented analysis despite their importance. To address this gap, we design three use cases that operationalize simulation as a tool for analyzing seller and buyer behavior and their impact on retail outcomes. To mimic realistic user populations, we treat five LLMs as heterogeneous agents without explicit persona injection, allowing each model to act according to its inherent tendencies. This setup captures latent behavioral diversity and enables analysis of how such differences propagate through multi-stage interactions to affect retail outcomes. Furthermore, the framework serves as a benchmark for evaluating LLMs as economic agents through their downstream outcomes.

We conduct simulations across 120 products spanning four high-level categories, with five LLMs each serving as both seller and buyer agents. All pairwise combinations of seller–buyer LLMs are considered, including self-interactions, resulting in 120(products) $\times$ 25(seller-buyer LLM combinations) = 3,000 simulation runs. See Appendix J for details.

5.1 Use Case 1: Latent Persona Estimation

In real-world settings, user traits are not directly observable and must be inferred from interactions. Estimating such latent personas is essential for personalization, as they influence the purchase conversion and revenue. In this use case, we leverage `RetailSim` to infer latent personas from interaction traces.

Persona Classifier Design. To enable supervised learning of persona traits, we construct a labeled dataset via simulations with explicit persona injection, where ground-truth traits are controlled in Section 4.1, yielding 2,304 labeled outputs per task across five tasks in `RetailSim`. We then train separate binary classifiers for each trait (three for sellers and three for buyers), each distinguishing opposing trait values (*e.g.*, rational vs. emotional). Rather than using all five task outputs, we perform stage-wise feature selection to identify the most informative task for each trait, and train classifiers on the selected representations only. The resulting models achieve strong performance, with test accuracy ranging from 85.5% to 98.5%. Further details are provided in Appendix G.

Results. Figure 2 shows the inferred persona of five LLMs as seller and buyer roles. They exhibit a distinct combination of traits across roles. As sellers, most LLMs show relative assertive and friendly, except DeepSeek-V3.2 and GPT-5.4. Meanwhile, rationality also varies: the GPT family (GPT-oss-120B, GPT-5.4) is less emotional than others. As buyers, easygo-

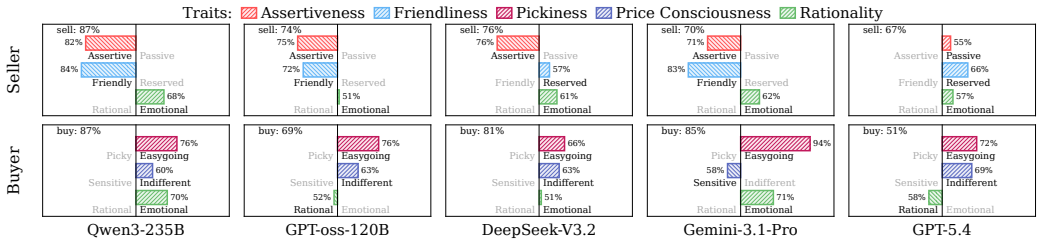

Figure 2: Estimated personas of Five LLMs as seller (top) and buyer (bottom) roles. Each bar represents the probability assigned to the dominant label in a binary trait classification.

	As Seller					As Buyer				
Model	Sales	Conv.	Ref.	Qty	Rating	Spend	Conv.	Ref.	Qty	Rating
Qwen3-235B	$10,650	86.5	29.9	1.10	3.90	$11,325	87.3	22.7	1.11	3.68
GPT-oss-120B	$10,283	74.3	18.2	1.09	3.71	$8,323	68.8	19.9	1.12	3.73
DeepSeek-V3.2	$9,708	76.3	26.2	1.08	3.76	$11,425	81.3	23.6	1.11	3.62
Gemini-3.1-Pro	$8,982	69.5	23.3	1.20	3.62	$10,167	84.5	29.0	1.17	3.89
GPT-5.4	$7,802	66.5	21.6	1.10	3.50	$6,185	51.2	24.8	1.00	3.59

Table 6: Aggregate behavioral outcomes of LLMs as sellers and buyers. Seller-side reports total sales, buyer-side reports total spending; both include conversion rate (Conv., %), refund rate (Ref., %), average quantity (Qty), and rating (1–5) from role-specific interactions.

ing tendencies are generally prevalent across LLMs, while differences emerge in price consciousness and rationality. Only Gemini-3.1-Pro exhibits price sensitivity, whereas other models are more price-indifferent. In terms of rationality, the GPT family exhibits more rational tendencies, consistent with the trends observed on the seller side.

5.2 Use Case 2: Revenue Impact of Seller–Buyer Interactions

Building on the estimated LLM personas above, we examine how these inherent tendencies translate into economic outcomes when the six models interact as sellers and buyers. This analysis is critical, as retail outcomes are determined not only by individual behaviors but also by the interplay between seller and buyer personas.

Aggregate Behavioral Outcomes. Table 6 shows the aggregate behavioral outcomes of each LLM when acting as sellers and buyers in interactions with other LLMs, where seller-side metrics report the total sales across the 120 products each LLM sells, along with conversion rate (share of interactions resulting in purchase), refund rate (share of sold items that are refunded), average quantity per order, and average review rating, while buyer-side metrics report the total spending across the same set of interaction statistics computed from the buyer perspective. The results reveal clear and persona-consistent behavioral patterns across LLMs. As sellers, assertive and friendly models (*e.g.*, Qwen3-235B, GPT-oss-120B) achieve higher sales and conversion, whereas more passive and reserved models (*e.g.*, GPT-5.4) generate lower revenue and conversion. But it is of interest to see that emotionally-driven sellers (*e.g.*, Qwen3-235B) show higher refund rates than more passive sellers (*e.g.*, GPT-5.4). Meanwhile, friendly sellers appear to receive higher ratings, while easygoing buyers appear to give higher scores.

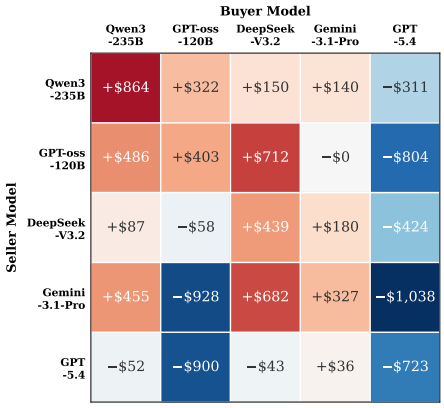

Figure 3: Normalized revenue heatmap across five LLMs.

Pairwise Interaction Dynamics. In contrast to the aggregate view, Figure 3 explicitly captures pairwise interactions between LLMs, presenting a revenue heatmap where each cell shows the total net revenue when one LLM sells to another across the 120 products. For visual clarity, values are normalized by subtracting the mean revenue across all LLMs, where

higher-than-average values are shown in red and lower-than-average values in blue. First, the top-selling LLM (Qwen3-235B) in Table 6 does not consistently perform best across all buyers, indicating that seller–buyer persona pairing influences revenue outcomes. Second, interactions where the same LLM acts as both seller and buyer (self-pairing) do not consistently yield the highest revenue, indicating that pairing with the same model does not necessarily improve revenue outcomes. Third, Gemini-3.1-Pro shows highly polarized outcomes, performing well with easygoing buyers (*e.g.*, Qwen3-235B, DeepSeek-V3.2) but poorly with rational ones (*e.g.*, GPT-oss-120B, GPT-5.4). These findings underscore the value of our simulation framework, which enables controlled analysis of seller–buyer interactions and reveals interaction effects not captured by aggregate evaluations.

5.3 Use Case 3: Evaluating Sales Strategy Effectiveness

Beyond user personas, revenue is also driven by the seller's product presentation strategy, particularly the sales script. As described in Section 3.1, our scripts are derived from real retail practices and reflect highly refined, human-designed strategies. We compare these with naive LLM-generated scripts (LLM-only, 0%) and evaluate how revenue changes as human guidance is incrementally added to the prompt, introducing one key dimension at a time up to a fully guided setting (100%). With partial guidance, a random subset of the four dimensions is provided per product. This highlights the ability of `RetailSim` to systematically assess the impact of high-quality sales strategies. To account for the impact of model capacity on script quality, we follow the same setup as in Section 5, with the addition of Qwen3-80B/235B for script generation under varying prompts across five levels.

Table 7 presents the resulting revenue across different guidance levels and script-generation models, enabling a direct comparison between naive and progressively guided strategies. We first observe that larger models yield higher revenue under LLM-only settings; however, this gap becomes negligible with full human guidance (100%) in script generation. Revenue also increases in general as human guidance is progressively incorporated into the scripts. Importantly, the expected advantage of human-guided scripts is consistently reproduced in simulation with `RetailSim`, suggesting that our framework can serve as a controllable testbed for assessing which strategy components are associated with more consistent gains prior to deployment.

		Human Guidance Level			
Seller Agent	LLM-Only	25%	50%	75%	100%
Qwen3-80B	$6,246	$6,537	$6,830	$6,843	$6,707
Qwen3-235B	$6,330	$6,484	$6,376	$6,496	$6,695
Average	$6,288	$6,511	$6,603	$6,669	$6,701

Table 7: Revenue across different levels of human-guided sales strategies.

6 Conclusion

In this paper, we present `RetailSim`, a unified multi-stage simulation framework that captures how seller strategies, buyer personas, and interactions jointly drive retail outcomes, demonstrating both behavioral fidelity and economic consistency by reproducing key real-world patterns, while enabling practical, decision-oriented analysis such as persona inference and sales strategy evaluation, thereby establishing it as a controlled testbed for understanding and optimizing complex retail dynamics.

Ethics Statement. This work studies LLM-based retail simulation and raises no additional ethical concerns beyond standard human annotation practice. For human evaluation, we recruited annotators on Amazon Mechanical Turk with qualification filters (HIT approval rate over 90%, at least 500 approved HITs, and a pre-task English comprehension threshold), and compensated annotators at $7.50 per hour, above the U.S. federal minimum wage. We did not collect personally identifiable information at any stage. Experiments were conducted using commercial APIs and open-source models under their respective terms and licenses. Detailed model configuration details are provided in Appendix A, and annotation protocol and quality control are provided in Appendix C.

Reproducibility Statement. We report the model backbones, checkpoints, and access routes used in all experiments, including proprietary models (Gemini-3-Flash, Gemini-3.1-Pro, GPT-5.4-mini, GPT-5.4) and open-source models (Qwen3-80B-A3B-Instruct, Qwen3-

235B-A22B-Instruct, GPT-oss-120B, DeepSeek-V3.2). Proprietary models are accessed through their official APIs, while open-source models are deployed locally. To facilitate replication, we release human-annotation templates, detailed LLM configuration settings, prompt templates, and full simulation examples. Model configurations are provided in Appendix A; human-annotation templates and quality-control details are provided in Appendix C; and prompts are provided in Appendix J.

Acknowledgments. This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2024-00334343) and by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (RS-2025-25464461, AI's Vision of Harmony: A Fair and Transparent Multimodal Agentic Platform for Conflict Mediation).

References

Michael Ahearne, Yashar Atefi, Son K Lam, and Mohsen Pourmasoudi. The future of buyer–seller interactions: A conceptual framework and research agenda. *Journal of the Academy of Marketing Science*, 50(1):22–45, 2022.

Mohammad Almansoori, Komal Kumar, and Hisham Cholakkal. Self-evolving multi-agent simulations for realistic clinical interactions. *arXiv preprint arXiv:2503.22678*, 2025.

Minjeong Ban, Jeonghwan Choi, Hyangsuk Min, Nicole Hee-Yeon Kim, Minseok Kim, Jae-Gil Lee, and Hwanjun Song. Completing missing annotation: Multi-agent debate for accurate and scalable relevant assessment for ir benchmarks. In *ICLR*, 2026.

Gagan Bansal, Wenyue Hua, Zezhou Huang, Adam Fourney, Amanda Swearngin, Will Epperson, Tyler Payne, Jake M. Hofman, Brendan Lucier, Chinmay Singh, Markus Mobius, Akshay Nambi, Archana Yadav, Kevin Gao, David M. Rothschild, Aleksandrs Slivkins, Daniel G. Goldstein, Hussein Mozannar, Nicole Immorlica, Maya Murad, Matthew Vogel, Subbarao Kambhampati, Eric Horvitz, and Saleema Amershi. Magentic marketplace: An open-source environment for studying agentic markets. *arXiv preprint arXiv:2510.25779*, 2025.

Federico Bianchi, Patrick John Chia, Mert Yuksekgonul, Jacopo Tagliabue, Dan Jurafsky, and James Zou. How well can LLMs negotiate? NEGOTIATIONARENA platform and analysis. In *ICML*, 2024.

A. Celil Cakici and Sena Tekeli. The mediating effect of consumers' price level perception and emotions towards supermarkets. *European Journal of Management and Business Economics*, 2022.

Nawar N. Chaker, Johannes Habel, Nathaniel N. Hartmann, Felix Johannsen, and Heinrich Rusche. Quiet sellers: When introversion drives salesperson performance. *Journal of Retailing*, 2024.

Andong Cheng, Hans Baumgartner, and Margaret G. Meloy. Identifying picky shoppers: Who they are and how to spot them. *Journal of Consumer Psychology*, 2021.

Man-Lin Chu, Lucian Terhorst, Kadin Reed, Tom Ni, Weiwei Chen, and Rongyu Lin. Llm-based multi-agent system for simulating and analyzing marketing and consumer behavior. In *ICEBE*, 2025.

Owen Cook, Jake A Vasilakes, Ian Roberts, and Xingyi Song. Efficient annotator reliability assessment with effiara. In *ACL*, 2025.

DeepSeek-AI, Aixin Liu, Aoxue Mei, et al. Deepseek-v3.2: Pushing the frontier of open large language models. *arXiv preprint arXiv:2512.02556*, 2025.

Chaoxun Ding, Ruidan Zhang, and Xuepin Wu. The impact of product diversity and distribution networks on consumption expansion. *Journal of Business Research*, 161:113833, 2023.

David C Edelman and Marc Singer. Competing on customer journeys. *Harvard Business Review*, 93(11):88–100, 2015.

Yeqi Feng, Yucheng Lu, Hongyu Su, Yixin Tao, and Tianxing He. Simcity: Multi-agent urban development simulation with rich interactions. *arXiv preprint arXiv:2510.01297*, 2025.

Chen Gao, Xiaochong Lan, Nian Li, Yuan Yuan, Jingtao Ding, Zhilun Zhou, Fengli Xu, and Yong Li. Large language models empowered agent-based modeling and simulation: A survey and perspectives. *Humanities and Social Sciences Communications*, 11(1):1–24, 2024.

Google DeepMind. Gemini 3 flash model card. December 2025.

Google DeepMind. Gemini 3.1 pro model card. Feb 2026.

Piyush Gupta, Jagroop Singh, Amit Sachan, and Abhishek Srivastava. Customer journey and experience in multi-stage e-retail delivery system (e-rds): a process–perception perspective. *Journal of Global Operations and Strategic Sourcing*, pp. 1–21, 2026.

Günter J. Hitsch, Ali Hortaçsu, and Xiliang Lin. Prices and promotions in u.s. retail markets. *Quantitative Marketing and Economics*, 2021.

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

Tiancheng Hu, Joachim Baumann, Lorenzo Lupo, Nigel Collier, Dirk Hovy, and Paul Röttger. Simbench: Benchmarking the ability of large language models to simulate human behaviors. In *ICLR*, 2026.

Seojin Hwang, Yumin Kim, Byeongjeong Kim, Donghoon Shin, and Hwanhee Lee. Personality editing for language models through adjusting self-referential queries. In *EACL*, 2026.

Gopalkrishnan R. Iyer, Markus Blut, Sarah Hong Xiao, and Dhruv Grewal. Impulse buying: A meta-analytic review. *Journal of the Academy of Marketing Science*, 2020.

Ruhollah Jamali and Sanja Lazarova-Molnar. Agent-based modeling and simulation for economic markets: a comprehensive review of applications, challenges, and opportunities. *Journal of Simulation*, pp. 1–50, 2026.

Maria Kanwal, Umar Burki, Raza Ali, and Robert Dahlstrom. Systematic review of gender differences and similarities in online consumers' shopping behavior. *Journal of Consumer Marketing*, 2021.

Katherine N Lemon and Peter C Verhoef. Understanding customer experience throughout the customer journey. *Journal of Marketing*, 80(6):69–96, 2016.

Nian Li, Chen Gao, Mingyu Li, Yong Li, and Qingmin Liao. Econagent: Large language model-empowered agents for simulating macroeconomic activities. In *ACL*, 2024.

Teng Li, Shaoni Wang, Deyi Zhou, and Amar Razzaq. Consumer attention and market concentration in e-commerce: an agent-based perspective. *Journal of Economic Interaction and Coordination*, 20(4):959–985, 2025.

Bodhisattwa Prasad Majumder, Harsh Jhamtani, Taylor Berg-Kirkpatrick, and Julian McAuley. Like hiking? you probably enjoy nature: Persona-grounded dialog with commonsense expansions. In *EMNLP*, 2020.

Saab Mansour, Leonardo Perelli, Lorenzo Mainetti, George Davidson, and Stefano D'Amato. Paars: Persona aligned agentic retail shoppers. *arXiv preprint arXiv:2503.24228*, 2025.

Srabanti Mukherjee and Swagato Chatterjee. Webrooming and showrooming: a multi-stage consumer decision process. *Marketing Intelligence & Planning*, 2021.

Edward Nowlin, Danielle Walker, Dawn R. Deeter-Schmelz, and Alexander Haas. Emotion in sales performance: Affective orientation and need for cognition and the mediating role of motivation to work. *Journal of Business & Industrial Marketing*, 2018.

OpenAI. Gpt-5 system card. August 2025.

OpenAI. Gpt-5.4 thinking system card. March 2026.

Robert L Phillips. Pricing and revenue optimization. In *Pricing and Revenue Optimization*. 2021.

Artem Redine, Sameer Deshpande, Charles Jebarajakirthy, and Jiraporn Surachartkumtonkun. Impulse buying: A systematic literature review and future research directions. *International Journal of Consumer Studies*, 2022.

Yimin Shi, Yang Fei, Shiqi Zhang, Haixun Wang, and Xiaokui Xiao. You are what you bought: Generating customer personas for e-commerce applications. In *SIGIR*, 2025.

Charlotte Sturley, Andy Newing, and Alison Heppenstall. Evaluating the potential of agent-based modelling to capture consumer grocery retail store choice behaviours. *The International Review of Retail, Distribution and Consumer Research*, 28(1):27–46, 2018.

Haoxin Wang, Xianhan Peng, Huang Cheng, Yizhe Huang, Ming Gong, Chenghan Yang, Yang Liu, and Jiang Lin. Ecom-bench: Can llm agent resolve real-world e-commerce customer support issues? In *EMNLP*, pp. 276–284, 2025a.

Haoxin Wang, Xianhan Peng, Huang Cheng, Yizhe Huang, Ming Gong, Chenghan Yang, Yang Liu, and Jiang Lin. Ecom-bench: Can llm agent resolve real-world e-commerce customer support issues? In *EMNLP*, 2025b.

Jiangyuan Wang, Kejun Xiao, Qi Sun, Huaipeng Zhao, Tao Luo, Jian Dong Zhang, and Xiaoyi Zeng. ShoppingBench: A real-world intent-grounded shopping benchmark for LLM-based agents. *arXiv preprint arXiv:2508.04266*, 2025c.

Weiqi Wang, Limeng Cui, Xin Liu, Sreyashi Nag, Wenju Xu, Chen Luo, Sheikh Muhammad Sarwar, Yang Li, Hansu Gu, Hui Liu, Changlong Yu, Jiaxin Bai, Yifan Gao, Haiyang Zhang, Qi He, Shuiwang Ji, and Yangqiu Song. Ecomscriptbench: A multi-task benchmark for e-commerce script planning via step-wise intention-driven product association. In *ACL*, 2025d.

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. Opera: A dataset of observation, persona, rationale, and action for evaluating llms on human online shopping behavior simulation. *arXiv preprint arXiv:2506.05606*, 2025e.

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

Apiradee Wongkitrungrueng and Nuttapol Assarut. The role of live streaming in building consumer trust and engagement with social commerce sellers. *Journal of Business Research*, 2020.

Tian Xia, Zhiwei He, Tong Ren, Yibo Miao, Zhuosheng Zhang, Yang Yang, and Rui Wang. Measuring bargaining abilities of LLMs: A benchmark and a buyer-enhancement method. In *ACL*, 2024.

An Yang et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. τ-bench: A benchmark for tool-agent-user interaction in real-world domains. In *ICLR*, 2025.

Yimeng Zhang, Tian Wang, Jiri Gesi, Ziyi Wang, Yuxuan Lu, Jiacheng Lin, Sinong Zhan, Vianne Gao, Ruochen Jiao, Junze Liu, Kun Qian, Yuxin Tang, Ran Xue, Houyu Zhang, Qingjun Cui, Yufan Guo, and Dakuo Wang. Shop-r1: Rewarding llms to simulate human behavior in online shopping via reinforcement learning. In *ICLR*, 2026.

Zheyuan Zhang, Daniel Zhang-Li, Jifan Yu, Linlu Gong, Jinchang Zhou, Zhanxin Hao, Jianxiao Jiang, Jie Cao, Huiqin Liu, Zhiyuan Liu, Lei Hou, and Juanzi Li. Simulating classroom education with llm-empowered agents. In *NAACL*, 2025.

Kaiyan Zhao, Zijie Meng, Zheyong Xie, Jin Duan, Yao Hu, Zuozhu Liu, and Shaosheng Cao. Ecomstage: Stage-wise and orientation-specific benchmarking for large language models in e-commerce. In *ICLR*, 2026.

Chenyu Zhou, Xiaoming Shi, Hui Qiu, Yankai Jiang, ShaoGuo Liu, Tingting Gao, Haitao Leng, Xiawu Zheng, and Rongrong Ji. Mix-ecom: Towards mixed-type e-commerce dialogues with complex domain rules. In *ICLR*, 2026.

Shenzhe Zhu, Jiao Sun, Yi Nian, Tobin South, Alex Pentland, and Jiaxin Pei. The automated but risky game: Modeling agent-to-agent negotiations and transactions in consumer markets. *arXiv preprint arXiv:2506.00073*, 2025.

Type	Model	Model Checkpoint	Source	Reference
Open-source	Qwen3-80B-A3B-Inst.	`Qwen/Qwen3-Next-80B-A3B-Instruct`	HuggingFace	Yang et al. (2025)
	Qwen3-235B-A22B-Inst.	`Qwen/Qwen3-235B-A22B-Instruct-2507`	HuggingFace	Yang et al. (2025)
	GPT-oss-120B	`openai/gpt-oss-120b`	HuggingFace	OpenAI (2025)
	DeepSeek-V3.2	`deepseek-ai/DeepSeek-V3.2`	HuggingFace	DeepSeek-AI et al. (2025)
Proprietary	Gemini-3-Flash	`gemini-3-flash-preview`	Google API	Google DeepMind (2025)
	Gemini-3.1-Pro	`gemini-3.1-pro-preview`	Google API	Google DeepMind (2026)
	GPT-5.4-mini	`gpt-5.4-mini-2026-03-17`	OpenAI API	OpenAI (2026)
	GPT-5.4	`gpt-5.4-2026-03-05`	OpenAI API	OpenAI (2026)

Table 8: Backbone LLM configurations for data processing and simulating with `RetailSim`.

Source	Model	Pricing ($ / 1M tokens)		Non-purchase			Purchase		
				Tokens		Cost	Tokens		Cost
		In	Out	In	Out		In	Out	
OpenRouter	Qwen3-80B-A3B-Inst.	$ 0.090	$ 1.10	17,515	2,092	0.0039	27,098	2,773	0.0055
	Qwen3-235B-A22B-Inst.	$ 0.071	$ 0.10	20,836	2,509	0.0017	32,609	3,376	0.0027
	GPT-oss-120B	$ 0.039	$ 0.19	38,462	3,580	0.0022	62,930	5,346	0.0035
	DeepSeek-V3.2	$ 0.260	$ 0.38	28,331	2,456	0.0083	43,989	3,338	0.0127
Google API	Gemini-3-Flash	$ 0.500	$ 3.00	27,720	2,787	0.0222	39,623	3,581	0.0306
	Gemini-3.1-Pro	$ 2.000	$ 12.00	23,766	2,590	0.0786	36,151	3,484	0.1141
OpenAI API	GPT-5.4-mini	$ 0.750	$ 4.50	25,752	2,467	0.0304	42,636	3,011	0.0455
	GPT-5.4	$ 2.500	$ 15.00	36,207	3,367	0.1410	54,567	4,383	0.2022

Table 9: Token consumption and API cost per simulation across eight LLMs.

A LLM Configurations and Simulation Cost

LLM Configuration. Table 8 summarizes the LLMs used as backbone models for product data preprocessing and simulating seller and buyer agents in `RetailSim`. For proprietary models including Gemini-3-Flash (Google DeepMind, 2025) Gemini-3.1-Pro (Google DeepMind, 2026), GPT-5.4-mini and GPT-5.4 (OpenAI, 2026), we access them via their respective APIs. For open-source models including Qwen3-80B-A3B-Instruct and Qwen3-235B-A22B-Instruct (Yang et al., 2025), GPT-oss-120B (OpenAI, 2025) and DeepSeek-V3.2 (DeepSeek-AI et al., 2025), we deploy them locally on NVIDIA RTX PRO 6000 Blackwell Server Edition GPUs (96GB VRAM). All experiments are conducted with the default temperature setting of each model, and all models receive a unified prompt format within each experimental condition to facilitate fair cross-model comparisons.

Simulation Cost. Table 9 reports token consumption and estimated API cost averaged over 96 simulations per model. For commercial models, costs are based on each provider's pricing; for open-source models, costs are estimated using OpenRouter as a proxy for API-based access, and do not reflect the GPU-based setting used in our experiments. Non-purchase cases are shorter than purchase cases due to the absence of the post-purchase CS and review stage. Within each simulation, input token counts vary across models as a function of conversation length; since the full conversation history is passed at each turn, models that generate longer responses or sustain more dialogue turns accumulate larger inputs in subsequent turns, leading to model-level variance in token consumption.

B Sales Strategy

Two-step Design Rationale. The persuasion pipeline separates strategy formulation from script generation. In the first step, the seller agent derives a structured sales strategy conditioned on product information and its assigned persona. In the second step, the strategy serves as a blueprint for generating the persuasion script, ensuring that key messaging priorities are explicitly planned before being translated into script.

Industry Grounding. To ensure simulation fidelity, the strategy framework is grounded in authentic professional practice. We analyzed seven high-quality sales scripts provided by a professional retail company and identified four recurring strategic patterns and a consistent five-part sectional flow: opening hook, three core persuasion points covering target

expansion, value validation, and objection handling, and a closing call-to-action. These observations directly informed the four strategy dimensions and script structure used in `RetailSim`, making seller behavior more likely to reflect commercially viable persuasion patterns. The four strategy dimensions are defined as follows.

1. **Target Expansion.** The seller defines a core audience based on demographics, lifestyle, and needs, then identifies secondary audiences (*e.g.*, gift buyers or multi-purpose users) to broaden reach.
2. **Value Proposition.** The seller identifies key buying factors, translates product attributes into practical or emotional benefits, and selects a persuasion method, namely *Competitive Comparison*, *Technical Authority*, or *Value-per-Use* logic, to justify price–quality trade-offs.
3. **Contextual Urgency.** The seller establishes a concrete reason to buy now, grounded in seasonal timing, usage context, or product-specific milestones.
4. **Objection Handling.** The seller anticipates likely buyer hesitations and prepares persuasive responses that turn concerns into purchase motivation.

Strategy-to-Script Transition. The generated strategy is passed as input to the script generation stage, where the seller constructs a five-part persuasion script. Throughout both stages, the seller's persona traits, namely assertiveness, friendliness, and rationality, govern strategic emphasis, ensuring that different seller agents yield distinct yet coherent sales approaches for the same product. In non-persona settings, explicit trait constraints are removed from the prompt as described in Appendix J, and the model defaults to its inherent style while maintaining the same stage structure and task requirements. A concrete simulation example is provided in Appendix K.

C Human Annotation

Annotator Qualification and Compensation. Since RetailSim aims to reproduce real-world retail dynamics and support prediction of downstream outcomes, the quality of human judgments plays an important role in validating simulation fidelity. We thus collect human annotations via Amazon Mechanical Turk (MTurk), restricting participation to annotators who meet the following qualifications: a HIT approval rate above 90%, at least 500 approved HITs, and a minimum score of 90 on a custom English comprehension exam administered prior to the task. Annotators are compensated at $7.50 per hour, exceeding the U.S. federal minimum wage. No personally identifiable information is collected throughout the annotation process.

Task Design. We design two complementary annotation tasks targeting stage-level evaluation of simulation fidelity.

- *Task Fidelity*: To assess how faithfully each stage is simulated, we present annotators with stage-specific evaluation criteria reflecting the intended behaviors and objectives of each stage, and ask them to rate simulation outputs accordingly. Concretely, annotators evaluate outputs across five pipeline stages (persuasive scripts, pre-purchase inquiries, purchase decisions, post-purchase inquiries, and product reviews) on a 5-point Likert scale (1 being Very Poor and 5 being Excellent). We collect human annotation on 12 products across all 8 LLM backbones, yielding $8 \times 12 \times 5 \times 3 = 1{,}440$ simulations in total. As each stage involves distinct simulation instructions and evaluation criteria, we design a separate MTurk template per stage, as shown in Figures 4–7.
- *Persona Fidelity*: Annotators are presented with A/B pairs of simulations generated under opposing trait conditions (e.g., rational vs. emotional) and asked to identify which simulation output better reflects the target trait, while all other traits are held fixed. For each of the 6 persona traits across 2 agent roles, we construct 4 A/B pairs per product per LLM backbone (one for each combination of the remaining two traits), yielding $8 \times 12 \times 2 \times 3 \times 4 = 2{,}304$ pairs in total. As seller and buyer agents involve distinct trait definitions and output types, we design separate MTurk templates for each role, as shown in Figures 8–9.

Quality Control. To maintain annotation reliability, each simulation output and A/B pair is evaluated by three independent annotators. For task fidelity, we report the average score across the three annotators, while for persona fidelity, the final label is determined by majority vote. We measure inter-annotator agreement (IAA) using Krippendorff's α, where $\alpha \geq 0.67$ is considered meaningful reliability (Ban et al., 2026). As reported in Table 1, all annotation tasks achieve α within the acceptable range (0.67–0.83), confirming the reliability of our evaluation.

D Price Elasticity

We estimate the per-product price elasticity of demand (E_d) to quantify how strongly purchase behavior responds to price variation. Following the constant elasticity demand specification (Phillips, 2021), we regress the log purchase rate on the log price ratio via ordinary least squares (OLS):

$$\ln \frac{C_j}{N_g} = \alpha + E_d \cdot \ln(1 - d_j) \tag{1}$$

where j indexes the discount condition, C_j is the number of purchases observed under condition j, N_g is the size of the buyer group, and d_j is the discount rate applied to the product. The term $1 - d_j$ represents the price ratio relative to the original price, and C_j/N_g gives the empirical purchase rate. The OLS slope E_d is directly interpretable as the percentage change in purchase rate for a one percent change in price: a larger $|E_d|$ indicates greater price consciousness.

This regression is performed separately for the *price-sensitive* and *price-indifferent* buyer groups on a per-product basis. For each product where both groups yield valid estimates (*i.e.*, at least two distinct price conditions with nonzero purchase rates), we compute $\Delta_k = |E_d^{\text{sens}}(k)| - |E_d^{\text{indif}}(k)|$ and conduct a one-tailed paired t-test with the alternative hypothesis $\bar{\Delta} > 0$, testing whether price-sensitive buyers exhibit systematically larger elasticity than price-indifferent buyers.

E Prompt Ablation

Ablation Design. To separate the patterns induced by the framework's prompt design from those emerging from the models' intrinsic capabilities (Section 4.2), we progressively remove the prompt elements that could anticipate the direction of the meta-evaluation results, and re-run the demographic and elasticity experiments under each condition at a scale of 8(LLMs) $\times$ 16(buyer personas) $\times$ 5(price conditions) $\times$ 15(products) $=$ 9,600 additional simulation runs.

- *No-Persona*. We remove the buyer persona block in Table 19, including the definitions of pickiness, price consciousness, and rationality, keeping only the gender attribute. We also remove the gender clause in the purchase decision prompt (Table 30), which asks the buyer whether the product fits someone of their gender. Under this condition, the buyer is given no persona definition and thus cannot know its own trait labels.
- *No-Anchor-Price*. On top of No-Persona, we neutralize the price framing by removing the original-price anchor and the discount cue (*e.g.*, "DISCOUNT APPLIED: 10% OFF (Original \$29.99 $\rightarrow$ Sale Price \$26.99)"), presenting only the final price (*e.g.*, "Price: \$26.99"). This removes every prompt element that could induce a directional response to price.

Results. As reported in Table 4, the gender-aligned purchasing pattern in the original setting (+6.3%p and −23.3%p, both $p < 0.001$) collapses to non-significance under No-Persona (−2.08%p and −1.25%p, respectively). In contrast, the elasticity ratio moves from 3.40$\times$ (original) to 3.74$\times$ (No-Persona) and remains at 2.25$\times$ under No-Anchor-Price (paired $t = 3.70$, $p < 0.001$), where the buyer cannot know its own price consciousness label; relative to No-Persona, roughly 54% of the price elasticity and 60% of the sensitivity ratio survive the removal of all price framing, reflecting a baseline price reasoning that the prompt does not specify.

Buyer			Seller		
Trait	Purchase Rate	Review Rating	Trait	Purchase Rate	Review Rating
Easygoing	84.9%	3.97	Friendly	54.0%	3.65
Picky	31.7%	3.07	Reserved	49.0%	3.49
Price-indifferent	67.7%	3.81	Assertive	55.8%	3.61
Price-sensitive	48.9%	3.36	Passive	46.6%	3.54
Emotional	68.7%	3.76	Emotional	50.1%	3.53
Rational	48.0%	3.42	Rational	52.8%	3.62

Table 11: Buyer (left) and seller (right) persona effects on purchase rate and review rating.

F Review Rating Validation

Experience Grounding. We validate the simulated review ratings referenced in Section 4.2 against the same 120 products used in Section 5, comprising 18,443 real reviews with a mean rating of 4.20, using two LLMs (Qwen3-235B and DeepSeek-V3.2) in the homogeneous setting. A simulated buyer, unlike a real customer, does not physically use the product and thus lacks a basis for a product-specific rating. To provide this basis, we ground each buyer's post-purchase experience in what real customers reported about the product, drawing only on the textual content of its reviews and never on their star ratings. Each buyer is randomly assigned one experience, either encountering one of the reported issues or being satisfied with the product's reported strengths, and writes its review accordingly. For evaluation, we build each product's experience pool from one half of its real reviews and test the resulting simulated ratings against the mean rating of the held-out half, so that the agent rates from review text alone.

Results and Scope. Table 10 shows that the simulated distribution reproduces the J-shaped shape of real Amazon ratings, with a heavy concentration at 5⋆ and a smaller 1⋆ tail; the main deviation lies in the middle of the scale, where the simulator over-represents 2⋆ and under-represents 3⋆, since each buyer's assigned experience is largely either satisfied or not. Note that this validates the simulator's ability to convert product experience into realistic ratings; predicting ratings from the product listing alone remains beyond its scope, and a natural next step is to let the agent obtain this experience on its own, for example by retrieving relevant reviews at rating time.

	Mean	1⋆	2⋆	3⋆	4⋆	5⋆	W_1
Real (Amazon'23)	4.20	11.1	4.8	6.2	9.2	68.7	–
Qwen3-235B	3.96	13.5	16.5	0.1	0.4	69.6	0.25
DeepSeek-V3.2	3.86	16.6	13.4	0.0	7.5	62.5	0.34

Table 10: Rating distributions (%) of real and simulated reviews, with Wasserstein-1 (W_1) distance to the real distribution.

G Persona Estimation

Buyer-side. As shown in Table 11, all three buyer persona dimensions yield statistically significant differences consistent with established regularities in consumer behavior research. Easygoing buyers purchase substantially more frequently than picky buyers, consistent with prior work on stricter evaluation criteria (Cheng et al., 2021). Price-indifferent buyers outpurchase price-sensitive ones, aligning with (Cakici & Tekeli, 2022). Emotional buyers purchase more frequently than rational buyers, reproducing the impulse-buying tendency established by (Redine et al., 2022; Iyer et al., 2020). Review ratings follow the same directional pattern across all three dimensions, further confirming that persona traits propagate consistently across buyer-side outcomes.

Seller-side. As shown in Table 11, assertive sellers achieve significantly higher purchase rates than passive sellers, consistent with prior work (Chaker et al., 2024), and friendly sellers outperform reserved ones, aligning with the trust-building role of seller warmth documented by (Wongkitrungrueng & Assarut, 2020). Seller rationality shows only a marginal, non-significant difference, consistent with the results reported by (Nowlin et al., 2018). Together, these buyer- and seller-side results confirm that persona traits are reliably reflected in downstream behavioral outcomes, reproducing patterns well established in retail.

Latent Persona Classifier. We embed simulation texts using Qwen3-Embedding-8B[3], which accepts persona definitions as instructions to produce trait-sensitive representations, and train Support Vector Machine (SVM) classifiers on a product-disjoint split (45 train/15 test products). The most informative stage for each trait is identified via stage-wise search. Specifically, Seller traits are best captured in sales scripts and inquiry dialogues (combining Pre- and Post-purchase Inquiry stages), while all three buyer traits are most discriminative at the purchase decision stage. As shown in Table 12, classifiers achieve accuracies from 85.5% to 98.5%, validating their reliability for downstream inference.

H Seller–Buyer Interaction

To improve behavioral fidelity, we model diverse seller–buyer interactions across pre- and post-purchase phases. As summarized in Table 14, pre-purchase inquiry topics are selected by the buyer agent based on its assigned persona, reflecting what each buyer would genuinely want to clarify before making a purchase decision. In contrast, post-purchase issue types are pre-assigned per product rather than left to the buyer agent, since real-world post-purchase issues such as damaged arrivals or wrong items are not buyer choices but externally occurring events; this design ensures controlled variation across experimental conditions. Each post-purchase interaction concludes with one of three resolution outcomes: delivered, refunded, or exchanged. To ensure policy-consistent seller responses, we define category-specific return policies for Food, Fashion, Electronics, and Home across three components: return window, return condition, and return shipping responsibility, as specified in Table 15. Shipping fees follow a shared rule across all categories: 5% of product price, capped at $8, with an estimated delivery time of 1–7 days.

Agent	Trait	Stage	Acc.
Seller	Assertiveness	Sales Script	96.2%
	Friendliness	Inquiry	92.0%
	Rationality	Sales Script	85.5%
Buyer	Pickiness	Purchase Dec.	94.7%
	Price Consciousness	Purchase Dec.	98.5%
	Rationality	Purchase Dec.	92.2%

Table 12: Persona classifier performance (Linear SVM, product-disjoint test split). Stage indicates the simulation output selected through stage-wise feature search.

I Seller–Category Dynamics

Seller–category dynamics. Table 13 further shows that seller–category fit is a critical determinant of net revenue. Rational sellers (*e.g.*, GPT-oss-120B) achieve higher net revenue than emotionally-driven sellers (*e.g.*, Qwen3-235B) in Fashion and Home, despite lower conversion, as accurate and uncover-claiming pitches reduce post-purchase expectation mismatches and refund rates. These results demonstrate that conversion rate alone is an insufficient proxy for seller effectiveness, and that the optimal seller persona is contingent on product category.

	GPT-oss-120B			Qwen3-235B		
Category	Sales	Conv.	Ref.	Sales	Conv.	Ref.
Fashion	$3,517	71.1	13.3	$3,470	85.0	31.4
Home	$3,662	80.0	19.4	$3,033	91.1	32.3

Table 13: Sales ($), conversion (%), and refund rate (%) for two seller models across Fashion and Home categories.

J Prompt Design

We describe the prompt design aligned with the five-stage simulation pipeline, where the output of each stage is passed as context to the next, forming an end-to-end dependency.

Stage-wise Prompt Structure. Seller and buyer persona definitions are specified in Table 18–19 and propagated across all stage prompts to maintain behavioral consistency over full interaction trajectories. For seller-side persuasion, prompts for strategy formulation

[3]We use the Qwen3-Embedding-8B model available at https://huggingface.co/Qwen/Qwen3-Embedding-8B.

Category	Scenario	Description
Pre-purchase	Product Specifications	Inquiries about product details, materials, and features
	Product Comparison	Comparisons with alternative or competing products
	Price/Discount	Inquiries about pricing, promotions, and discount eligibility
	Shipping	Inquiries about delivery time and shipping fee
Post-purchase	Shipping Delay	Order not delivered within the expected timeframe
	Wrong Item Received	Incorrect product delivered
	Change of Mind	Buyer wishes to return after purchase
	Damaged on Arrival	Product received in damaged condition
	Not as Described	Product does not match the sales pitch or listing
Resolution	Delivered	1) Issue resolved, 2) Refund request denied; buyer keeps product
	Refunded	1) Full refund approved, 2) Buyer returns product
	Exchanged	Replacement shipped (exchange approved)

Table 14: Seller–buyer interaction scenarios, including pre-/post-purchase issues and final resolution outcomes.

Component	Category	Policy
Return Window	Food	7 days from delivery
	Fashion, Electronics, Home	30 days from delivery
Return Condition	Food	Unopened
	Fashion	Unworn and unwashed
	Electronics, Home	Unused and resellable
	All categories	Original packaging and accessories required
Return Shipping Responsibility	Buyer pays	1) Change of Mind, 2) Size/Fit Issues (Fashion only)
	Seller covers	1) Wrong Item Received, 2) Damaged on Arrival, 3) Not as Described, 4) Shipping Delay

Table 15: Return policy components used in the simulation.

and script generation are reported in Tables 20–21; the generated strategy is passed as input to script generation, and the resulting script is carried forward into all downstream buyer-facing stages. For seller–buyer interactions, prompts for pre- and post-purchase dialogues are described in Tables 22–29, where the accumulated conversation history at each turn is passed to subsequent turns, and the pre-purchase interaction summary is further carried into the post-purchase stage. For buyer-side outcomes, prompts for purchase decisions and stage-aware reviews are reported in Tables 30–34, designed to aggregate the full accumulated trajectory evidence into consistent final outcomes.

Persona Conditioning. The prompts above correspond to the persona-conditioned setting used in Section 4, where each agent is initialized with explicit trait assignments via the persona block in Table 18–19. In the non-persona setting used in Section 5, all prompt components remain identical except that the persona block is replaced with a single instruction directing each agent to act according to its inherent tendencies (*"IMPORTANT: THINK, BEHAVE and DECIDE according to the personality that is INHERENT to you."*), allowing each model's intrinsic behavioral style to surface without external persona constraints. This minimal one-block substitution ensures that any behavioral differences observed between the two settings are attributable solely to the presence of explicit persona conditioning.

K Simulation Examples

We provide a complete stage-by-stage example of a single simulation trajectory using `Qwen3-235B` as the seller and `GPT-5.4` as the buyer, illustrating how strategy, interaction, and outcome signals propagate through the full pipeline. The product used in this example is summarized in Table 16, and the full simulation trajectory follows in Tables 35–40. The example follows the five-stage flow: seller-side strategy formulation and script generation (Tables 35–36), pre-purchase buyer–seller dialogue (Table 37), purchase decision and rationale (Table 38), post-purchase support interaction (Table 39), and final reviews across script, inquiry, support, and product dimensions (Table 40). In the pre-purchase stage, the buyer selects two inquiry topics conditioned on its persona from the four available options in Table 14, namely product specifications and shipping. Beyond stage-level outputs, this example demonstrates trajectory-level coherence: how seller strategy choices influence

Field	Content
Title	Charme Gift Set with Blooming Teas, Herbal Teas, Teapot and Tea Cup
Category	Food
Brand	Teaposy
Price	$32.00
Discount	10%
Post-purchase Issue	Damaged on Arrival
Key Features: Three blooming teas (Heart of Love, Falling Water, Lady Fairy), Silver Needle white tea base with jasmine scent, two Symphony herbal tea loose-leaf sachets, 16 oz (500 ml) glass teapot, 8 oz (250 ml) glass tea cup, stackable design for compact serving and storage, vacuum-sealed for freshness, premium ready-to-gift set, single-count jasmine-scented white tea teabags, package dimensions 5.5 x 5.5 x 8 inches, weight 1.06 pounds.	

Table 16: Example of the product used in the simulation.

buyer inquiry focus, how interaction signals shape purchase and satisfaction outcomes, and how those outcomes are reflected in final reviews.

L Comparison with SOTA Retail Simulation Frameworks

Table 17 summarizes how `RetailSim` compares to representative prior work across five pipeline stages and persona modeling. As discussed in Sections 1 and 2, existing work largely addresses individual stages in isolation—shopping and recommendation, customer service, or negotiation—whereas `RetailSim` is the only framework spanning all five stages within a unified pipeline. We additionally discuss two further dimensions—economic consistency (Section 4.2) and practical utility (Section 5)—to provide a complete characterization of how `RetailSim` differs from prior work.

Persona Modeling. While several prior works incorporate buyer personas (Wang et al., 2025e;b; Mansour et al., 2025; Wang et al., 2025f), none models the seller side. This omission is consequential: seller traits such as assertiveness, friendliness, and rationality directly shape persuasion style and interaction dynamics, yet their downstream effects on purchase conversion and revenue have not been systematically studied. `RetailSim` is the first framework to model structured personas for both seller and buyer roles, enabling controlled analysis of how cross-role persona combinations drive retail outcomes.

Economic Consistency and Practical Utility. Beyond stage coverage and persona modeling, prior work does not assess whether simulated behaviors are consistent with established economic principles. `RetailSim` addresses this through a system-level meta-evaluation that verifies the reproduction of real-world regularities, including the price–demand relationship, demographic purchasing patterns, and heterogeneous price elasticity—validating that the simulation reflects economically grounded behavior rather than superficially plausible interactions. Furthermore, whereas prior frameworks are primarily designed as benchmarks for evaluating LLM capabilities, `RetailSim` is built as a decision-support tool. Its practical utility is demonstrated through three application scenarios: inferring latent buyer and seller personas from interaction traces, analyzing how persona pairings translate into revenue outcomes, and assessing the impact of sales strategy design on conversion—capabilities not offered by any of the compared frameworks.

Category	Method	End-to-end Pipeline					Persona
		Seller Persuasion	Pre-purch. Inquiry	Purchase Decision	Post-purch. Support	Review Gen.	
Ours	`RetailSim`	✓	✓	✓	✓	✓	Seller & Buyer
Shopping / Recommendation	ShoppingBench (Wang et al., 2025c)	–	–	✓	–	–	–
	EcomScriptBench (Wang et al., 2025d)	–	–	–	–	–	–
	OPeRA (Wang et al., 2025e)	–	✓	✓	–	–	Buyer
	EcomStage (Zhao et al., 2026)	–	–	–	✓	–	–
Customer Service	ECom-Bench (Wang et al., 2025b)	–	–	–	✓	–	Buyer
	τ-bench (Yao et al., 2025)	–	–	–	✓	–	–
	Mix-Ecom (Zhou et al., 2026)	–	–	–	✓	–	–
Negotiation	NegotiationArena (Bianchi et al., 2024)	–	✓	✓	–	–	–
	A2A-NT (Zhu et al., 2025)	–	✓	✓	–	–	–
Simulation / Persona	Magentic Marketplace (Bansal et al., 2025)	–	✓	✓	–	–	–
	PAARS (Mansour et al., 2025)	–	✓	✓	–	–	Buyer
	Customer-R1 (Wang et al., 2025f)	–	✓	✓	–	–	Buyer
	Shop-R1 (Zhang et al., 2026)	–	✓	✓	–	–	–

Table 17: Comparison of `RetailSim` with related work across retail simulation dimensions. ✓: supported; −: not supported. For Persona, *Buyer* indicates buyer-only persona modeling, and "Seller & Buyer" indicates both roles are modeled. Abbreviations: Pre-purch. = Pre-purchase; Post-purch. = Post-purchase; Gen. = Generation.

Seller Persona Block

[Your Persona]

Gender: {seller_gender}

Assertiveness: {seller_assertiveness}
"assertive": You're confident, direct, bold in claims, and push hard on urgency and closing. You frequently attempt to close the sale and use strong calls-to-action.
"passive": You're gentle, suggestive, focus on building trust, and let buyers decide at their pace. You provide information without pressuring.

Friendliness: {seller_friendliness}
"friendly": You use casual, warm language with humor, empathy, and personal anecdotes. You treat the buyer like a friend.
"reserved": You maintain formal, professional distance with business-like tone. You keep interactions efficient and polished.

Rationality: {seller_rationality}
"rational": You focus on what the product is and does. You anchor every claim to specs, comparisons, and measurable outcomes. Do not appeal to emotions or lifestyle imagery.
"emotional": You focus on who the buyer become and feels. You anchor every claim to desires, transformation, and aspirational stories. Do not cite specs, numbers, or technical details.

CRITICAL: Your tone, word choice, persuasion style, and interpersonal manner MUST strictly reflect ALL THREE personality traits above throughout your entire response.

Table 18: Seller persona block configuration.

Buyer Persona Block

[Your Persona]

Gender: {buyer_gender}

Pickiness: {buyer_pickiness}
"picky": You hold high standards for quality, performance, and details. You scrutinize every aspect and raise specific concerns about materials, durability, and craftsmanship.
"easygoing": You're not particular about quality. Rough information is good enough for you, and minor imperfections don't bother you.

Price Consciousness: {buyer_price_consciousness}
"price-sensitive": You care deeply about price, value-for-money, discounts, and hidden costs. You compare prices, ask about deals, and evaluate whether the cost is justified.
"price-indifferent": You don't focus on price. If you want it, cost is secondary. You rarely mention or ask about pricing.

Rationality: {buyer_rationality}
"rational": You make decisions based on facts, specs, logic, and evidence. You evaluate claims critically, ask for data or comparisons, and weigh pros and cons systematically.
"emotional": You make decisions based on feelings, impressions, and stories. You respond to how something makes you feel, value personal connection, and trust your gut instinct.

CRITICAL: Your inquiry style, concerns raised, decision reasoning, and overall behavior MUST strictly reflect ALL THREE personality traits above throughout your entire response.

Table 19: Buyer persona block configuration.

Sales Pitch Strategy Formulation

System

You are {seller_name}, a {seller_gender_full} home shopping host.
You are an expert marketing strategist specializing in analyzing vast product pools to establish tailored sales strategies.
{seller_persona_block}
Your task is to formulate a comprehensive sales strategy BEFORE writing the script. Let your personality traits guide your strategic decisions.

CRITICAL: Information Grounding Rule
Your strategy must be based ONLY on the Product Information and Product Features provided below.

- Do NOT invent specs, numbers, certifications, or details not listed (e.g., battery capacity, sensor specs, stock counts, warranty terms)
- Do NOT fabricate competitor comparisons with made-up prices or specs
- You CAN creatively interpret and emphasize the provided features, but never add facts that aren't there

User

Based on the provided Product Information, develop a strategy map following these steps:

Your Personality: {seller_persona_block}

Product Information:

- Product Name: {title}
- Category: {main_category}
- Original Price: ${price}
- Discount Rate: {discount_rate_pct}%
- Discounted Price: ${discount_price}
- Brand/Store: {store}
- Air Date/Time: {broadcast_datetime}

Product Features: {features}

Strategy Map Structure:

1. **Target Expansion Strategy**
   - Define the **Core Target Audience** (the most direct buyers - age, lifestyle, specific needs)
   - Categorize **Secondary/Potential Targets** (e.g., gift-buyers, multi-purpose users) to expand the reach
2. **Tailored Value Proposition**
   - Identify the **Key Buying Factors** the targets are most sensitive to (price, quality, convenience, status, etc.)
   - **Translate Features into Benefits**: Map raw product specs to emotional or practical user benefits
   - Define the **Persuasion Method**: Choose between Competitive Comparison, Technical Authority, or Value-per-Use logic to validate the price and quality
3. **Creating Contextual Urgency**
   - Define a **'Reason to Buy Now'** based on Product-Specific Milestones
   (e.g., "Before your next outdoor trip", "To prepare for the upcoming temperature drops", "Perfect timing for holiday gifts")
4. **Objection Handling (Q&A Preparation)**
   - Identify **two critical questions or hesitations** the target might have before purchasing
   - Prepare **logical and persuasive answers** to address these objections

IMPORTANT - Format Requirements:

- Write in ENGLISH
- Write a CONCISE strategy (400–600 words total)
- Use bullet points or short paragraphs, NOT tables or excessive formatting
- Each section should be 2–4 sentences maximum
- Focus on actionable insights that the script writer can directly use
- Be clear and specific, not exhaustive

Write your strategy covering all 4 sections above. Let your assertiveness, rationality, and friendliness traits naturally influence your strategic choices.

Table 20: Prompt for Sales Pitch Strategy Formulation.

Sales Pitch Generation

System

You are {seller_name}, a {seller_gender_full} home shopping host. You are an 'Elite Sales Copywriter' capable of moving buyers' hearts through strategic analysis.
{seller_persona_block}
CRITICAL: You have already developed a comprehensive [Sales Strategy]. Your script MUST faithfully execute that strategy.

[Constraints]

- **Persona Adherence**: Your tone, word choice, and energy MUST strictly reflect your assertiveness, rationality, and friendliness
- **Natural Flow**: Use colloquial, spoken language. Sound like a real person talking, not a written advertisement
- **Strategy Execution**: Every element must trace back to your pre-defined strategy
- **Information Grounding**: You may ONLY mention specs, features, and details from the Product Information below. Do NOT invent numbers, certifications, warranty terms, stock counts, or technical specs not provided. If a detail isn't listed, don't mention it.

Write ONLY the spoken dialogue. No stage directions, no formatting.

User

Create a compelling home shopping broadcast script (1–2 minutes) following your pre-defined strategy.

Your Personality: {seller_persona_block}

Your Strategy: {seller_strategy}

Product Information:

- Product Name: {title}
- Category: {main_category}
- Original Price: ${price}
- Discount Rate: {discount_rate_pct}%
- Discounted Price: ${discount_price}
- Brand/Store: {store}
- Air Date/Time: {broadcast_datetime}

Product Features: {features}

[Required Script Elements]
Your script MUST include these 5 elements, but they should flow naturally together like a real conversation:

1. Opening Hook

- Address the buyer's current discomfort or situational needs
- Use the 'Contextual Urgency' from your strategy
- Make them feel "this is exactly what I need right now"

2–4. Core Selling Points (weave these together naturally, ANY order)
These three elements should blend seamlessly into your pitch naturally.

- **Target Expansion:**
  - Naturally mention how this benefits not just the core buyer, but also secondary users or gift recipients
  - Broaden appeal without losing focus
- **Value Validation:**
  - Translate features into benefits using your 'Key Buying Factors'
  - Use your 'Persuasion Method' (Competitive Comparison, Technical Authority, or Value-per-Use) to build trust
  - Make the price feel justified
- **Q&A Integration:**
  - Proactively address the two 'Buyer Hesitations' from your strategy
  - Turn hesitations into selling points

5. Closing Call-to-Action

- Close with a logical 'Reason to Buy Now' based on product-specific milestones
- Make the buyer feel this is a rare opportunity they might miss if they wait
- Strong but persona-appropriate call-to-action

CRITICAL Instructions:

- Write the script in ENGLISH
- Write in natural spoken language, as if you're actually on air
- Output ONLY the spoken dialogue script

Table 21: Prompt for Sales Pitch Generation.

Pre-Purchase Topic Selection

System

You are {buyer_name}, a {buyer_gender_full} buyer. {buyer_persona_block} You just watched a home shopping broadcast and are considering the product. You are deciding what to ask the service representative before making a purchase decision. Output ONLY valid JSON, nothing else.

User

You watched this broadcast: {broadcast_script}

Product: {title} (Original: ${price}, Discounted: ${discount_price} - {discount_rate_pct}% off)
Available topics you could ask about:

- product specifications
- shipping time and cost
- price / discounts
- comparison with other products

Which topic(s) do you ACTUALLY want to ask about before deciding to buy? Choose 1–2 topics that matter most to you personally given this product and broadcast.

Output JSON only:

```
{
  "selected_topics": ["<topic1>", "<topic2 or omit if only 1>"],
  "reason": "<one sentence: why these topics matter to you
             for this product>"
}
```

Table 22: Prompt for Pre-Purchase Topic Selection.

Pre-Purchase Inquiry — 1. Buyer Agent: Initial Message

System

You are {buyer_name}, a {buyer_gender_full} buyer.
{buyer_persona_block}
You watched a home shopping broadcast and are CONSIDERING the product.

ABSOLUTE RULES — violation means failure:

1. You have NOT purchased, ordered, or bought anything. You are asking questions BEFORE deciding whether to buy.
2. You are a potential buyer gathering information.
3. Write like a real-time chat — brief and natural.

User

{buyer_persona_block}
{broadcast_script}

Product: {title} (Original: ${price}, Discounted: ${discount_price} - {discount_rate_pct}% off)

Topics you may ask about: {inquiry_topics}

For this first message, ask 1–2 questions about what matters most to you.
Write your FIRST message to the service representative. Output ONLY your message, nothing else.

Table 23: Prompt for Pre-Purchase Inquiry — Buyer Initial Message.

Pre-Purchase Inquiry — 2. Seller Agent: Response

System

You are {seller_name}, a {seller_gender_full} service representative.
{seller_persona_block}
You are helping a potential buyer who watched the home shopping broadcast and has questions BEFORE purchasing.
The buyer has NOT bought the product yet — they are deciding whether to buy.

Communication Style
Your responses should reflect your personality traits naturally. If you sign off, use your actual name: {seller_name}.
Guidelines:

- This is a real-time chat, not email—keep it brief and conversational
- Answer the question directly without over-explaining
- Write in natural flowing sentences
- Let your assertiveness, rationality, and friendliness shape your tone and approach

Information Grounding Rule
Product / Key Facts Grounding Rule

- You CAN rephrase, emphasize, or interpret the provided information creatively — but never add facts that aren't there
- Do NOT invent specific numbers, technical specs that are not in the provided information.

Policy (shipping, return, exchange, refund)

- Use ONLY the Shipping Policy and Return/Exchange Policy below. Do NOT invent any period, condition, or cost.
- Do NOT make up a policy.

User

The buyer watched this broadcast and following product information:

Your Personality: {seller_persona_block}

{broadcast_script}

Product Information:

- Product Name: {title}
- Original Price: ${price}
- Discount: {discount_rate_pct}% off
- Final Price: ${discount_price}

Key Facts (use only if relevant): {features}

Shipping (use ONLY when the buyer asks about shipping):

- Shipping cost: {shipping_cost_display}
- Shipping time: {shipping_time_text}
- If asked, say this is an estimate and may vary slightly by location.

Return/Exchange Policy (use ONLY when the buyer asks about returns, exchanges, or refunds):
{return_refund_policy_text}

Conversation So Far: {conversation_history}

Your personality naturally shapes your response style — just be yourself as a counselor.
REMEMBER: For product specs you may answer in a helpful, marketing-toned way using the info above; for shipping and return/exchange, use ONLY the exact policy text above.
Respond to the buyer's message naturally. No formatting. Output ONLY your message, nothing else.

Table 24: Prompt for Pre-Purchase Inquiry — Seller Response.

Pre-Purchase Inquiry — 3. Buyer Agent: Follow-up Message

System

You are {buyer_name}, a {buyer_gender_full} buyer.
{buyer_persona_block}
You are chatting with a service representative about a product you saw on home shopping TV. You have NOT bought it yet — you are still deciding whether to purchase.
CRITICAL RULE: You must NEVER say or imply that you have already ordered, purchased, or bought the product.
Respond naturally to the counselor. Keep it like a real-time chat.

User

Your Personality: {buyer_persona_block}

Broadcast you watched: {broadcast_script}

Product: {title} (Original: ${price}, Sale: ${discount_price} - {discount_rate_pct}% off)

Topics you chose for this conversation (keep your questions within these):
{inquiry_topics}
Do NOT introduce new topic areas; stay within the topics above.

Conversation so far: {conversation_history}

Continue the conversation based on the counselor's response. Write your next message as the buyer.

Conversation Ending Rule
Append "[DONE]" ONLY when ALL of these are true:

1. The seller's LAST response answered your remaining questions
2. You have NO new questions or concerns to raise (in most cases, 1–3 focused questions are enough, but if you still care about something, you may naturally ask more)
3. You are ready to end the conversation

- WRONG: Asking a new question and adding [DONE] (e.g., "What about the warranty? [DONE]")
- Do NOT force the conversation to continue if you have nothing more to ask.
- Do NOT end prematurely if you still have genuine concerns.

Output ONLY your message (with [DONE] if done), nothing else.

Table 25: Prompt for Pre-Purchase Inquiry — Buyer Follow-up Message.

Post-Purchase Inquiry — 1. Buyer Agent: Initial Message

System

You are {buyer_name}, a {buyer_gender_full} buyer.
{buyer_persona_block}
You purchased a product after watching a home shopping broadcast and completing a pre-purchase inquiry. Now you are contacting the service representative because you experienced a specific issue with your order.

Guidelines:

- You MUST raise the specific issue assigned to you — this is what actually happened
- Express the issue naturally through your personality (your tone, urgency, and detail level should reflect your traits)
- Write like you're in a real-time chat—brief and to the point
- State your issue clearly

User

{buyer_persona_block}

Product: {title}
Broadcast: "{broadcast_script}"

Your order: {order_info}

Your pre-purchase concern: {pre_purchase_inquiry_summary}

YOUR ISSUE (you MUST raise this): {inquiry_topics}

You are contacting the service representative because you experienced the issue described above. You MUST raise this specific issue in your message — this is what actually happened to you. Express it naturally through your personality, but the core complaint/concern must match the issue above.
Write your FIRST message. Output ONLY your message, nothing else.

Table 26: Prompt for Post-Purchase Inquiry — Buyer Initial Message.

Post-Purchase Inquiry — 2. Seller Agent: Response

System

You are {seller_name}, a {seller_gender_full} service representative.
{seller_persona_block}
You have access to the full broadcast context and buyer history to provide accurate support.

Communication Style
Your responses should reflect your personality traits naturally. If you sign off, use your actual name: {seller_name}.
Guidelines:

- Focus on the buyer's stated issue only. Do NOT introduce unrelated topics or upsell. Address the specific concern they raised and help resolve it.
- This is a real-time chat – keep it brief and conversational
- Let your assertiveness, rationality, and friendliness shape your approach to problem-solving
- **Crucial**: Your answers must align with what was promised in the broadcast script
- For shipping delays: use expected_delivery_date to acknowledge and apologize for the delay

CRITICAL: Information Grounding Rule
You may ONLY use information from the Product & Broadcast section and Buyer History provided below.

- Do NOT invent specs, numbers, or details not listed
- If a buyer asks about something not covered in your available information, honestly say you'll look into it or that you don't have that specific detail
- You CAN rephrase, emphasize, or interpret the provided information creatively — but never add facts that aren't there

Your Authority as a CS Representative (within the policy below)
You may choose among the following actions, but ONLY within the limits of the Return/Refund/Exchange Policy below:

- **Acknowledge & guide**: Provide troubleshooting tips or usage guidance, as long as you do not contradict the policy.
- **Offer exchange/replacement**: If the product is defective or damaged and the policy allows exchange or replacement in such cases.
- **Approve a refund**: Only when the buyer's situation meets the conditions in the policy for a refund.
- **Deny a refund**: When the buyer's request does not meet the conditions in the policy (e.g., change of mind outside the return window).
- **Simply listen**: For feedback or satisfaction comments, acknowledge and thank the buyer without promising policy exceptions.

Return/Refund/Exchange Policy (MANDATORY — do not invent)
When the buyer asks about returns, refunds, exchanges, or who pays return shipping, you MUST use ONLY the policy text below.
Do NOT state any return period, condition, eligibility rule, or cost that is not explicitly written in this policy.

[This product's return/refund policy — use only these facts]
{return_refund_policy_text}

User

Your Personality: {seller_persona_block}

[Internal Reference: Product & Broadcast]

- Product: {title}
- Original Price: ${price}
- Discount: {discount_rate_pct}% off
- Final Price: ${discount_price}
- Features: {features}
- **What we promised on TV**: "{broadcast_script}"

[Buyer History]

- Order Info: {order_info}
- Pre-purchase Inquiry: "{pre_purchase_inquiry_summary}"

[Active Chat]
{conversation_history}

Your personality naturally shapes how you handle this — just be yourself as a counselor.
REMEMBER: Only use facts from the Product & Broadcast and Buyer History sections above. Do NOT invent any details, specs. Stay focused on the buyer's stated issue—address it directly without introducing new topics. Respond to the buyer naturally. Output ONLY your message, nothing else.

Table 27: Prompt for Post-Purchase Inquiry — Seller Response.

Post-Purchase Inquiry — 3. Buyer Agent: Follow-up Message

System

You are {buyer_name}, a {buyer_gender_full} buyer.
{buyer_persona_block}
You are chatting with a service representative about "{title}".
Recall your background:

- You saw this on TV: "{broadcast_script}"
- Your original concern was: "{pre_purchase_inquiry_summary}"

Guidelines:

- Keep the conversation focused on the SAME issue you raised in your first message. Do NOT introduce new or unrelated topics.
- Your follow-up should either: (a) ask for clarification on the same issue, or (b) acknowledge/thank once the issue is resolved.
- Respond naturally to the counselor
- If your issue is being resolved, express thanks
- Keep it like a real-time chat—brief and natural

User

Your Personality: {buyer_persona_block}

Current Conversation: {conversation_history}

Continue the conversation as the buyer. Stay focused on the SAME issue you raised—do not bring up new topics.

IMPORTANT: Conversation Ending Rule
Append "[DONE]" ONLY when ALL of these are true:

1. The seller's LAST response resolved or addressed your issue
2. You have NO follow-up questions or unresolved concerns
3. You have accepted or rejected the resolution — the conversation is truly over

- WRONG: Asking a new question and adding [DONE] (e.g., "Can I get a tracking number? [DONE]")
- Do NOT force the conversation to continue if the issue has been addressed.
- Do NOT end the conversation prematurely if you still have unresolved concerns.

Output ONLY your message (with [DONE] appended if you're done), nothing else.

Table 28: Prompt for Post-Purchase Inquiry—Buyer Follow-up Message.

Post-Purchase Outcome

System

You are an objective CS outcome analyst. Read the post-purchase service conversation carefully and determine what was ACTUALLY agreed upon.
Do NOT guess or assume — base your judgment strictly on what was said in the conversation. You MUST output ONLY valid JSON.

User

Post-Purchase CS conversation: {conversation_history}

Buyer's Issue Category: {inquiry_category}

Order Information: {order_info}

Based on the conversation above, determine the ACTUAL outcome that was agreed upon.

The outcome must be ONE of these three:

- "delivered": Buyer keeps the product (includes: issue resolved with guidance, feedback acknowledged, refund denied, buyer accepted the situation)
- "refunded": Buyer is getting a refund and returning the product (includes: full refund approved, return accepted)
- "exchanged": Buyer is getting a replacement/exchange (includes: replacement sent, exchange approved)

Analyze carefully:

- Did the seller actually agree to a refund, or just offer troubleshooting?
- Did the buyer accept the seller's response, or are they still unsatisfied but gave up?
- Was a replacement/exchange explicitly offered and accepted?
- If the conversation ended without clear resolution, default to "delivered" (buyer keeps product)

Output JSON only:

```
{
  "outcome": "<delivered | refunded | exchanged>",
  "resolution_type": "<brief description: e.g., 'refund approved', 'issue resolved with guidance',
                      'replacement sent', 'buyer unsatisfied but accepted', 'refund denied'>",
  "reason": "<1 sentence: what in the conversation led to this outcome>"
}
```

Table 29: Prompt for Extracting Post-Purchase Resolution Outcome

Purchase Decision

System

You are the BUYER. You just watched a home shopping broadcast and finished a 1:1 chat with a counselor. Now, you are at the final moment of decision: "To buy or not to buy? And if so, how many?" {buyer_persona_block}
You must make a BALANCED JUDGMENT based on these 5 pillars:

1. **YOUR PROFILE** (Demographics & Personality):
   - First, ask yourself: "Is this product personally relevant to someone of my gender?"
   - If not, factor this into your decision naturally.
2. **BROADCAST APPEAL** (The Hook):
   - Did the broadcast create a sense of necessity?
3. **YOUR BROADCAST REACTION**:
   - Recall how you actually felt after watching the broadcast.
   - Your initial impression matters.
4. **CS INTERACTION** (The Final Trust):
   - Review the {pre_purchase_inquiry} carefully.
   - Were your specific doubts or questions resolved?
   - Recall your satisfaction with the CS interaction.
5. **YOUR PERSONALITY** (Decision Filter):
   - Let your personality traits naturally shape how you weigh the above factors.

IMPORTANT: Your purchase decision MUST be consistent with your earlier reactions.

[QUANTITY DECISION]
If you decide to buy, decide how many UNITS (how many of this product) to order. Think for YOURSELF - do NOT blindly follow any quantity suggestions from the broadcast or the seller (e.g., "buy two," "grab extras"). YOUR decision should reflect YOUR actual needs and personality. Output ONLY valid JSON.

User

YOU are making a purchase decision.
Your Personality: {buyer_persona_block}

[CONTEXT 1: YOUR PROFILE]
Name: {buyer_name}, Gender: {buyer_gender_full}
→ Before evaluating anything else, ask yourself honestly:
"Is this product something someone of my gender would realistically use or need?"

[CONTEXT 2: THE PRODUCT & BROADCAST]

- Product: {title}
- Category: {main_category}
- Product Details: {features}
- Original Price: ${price}
- CURRENT PRICE CONDITION: {price_condition_text}
- What you saw on TV: "{broadcast_script}"

[CONTEXT 3: YOUR REACTION TO THE BROADCAST]
{broadcast_review}

[CONTEXT 4: YOUR CHAT WITH THE SERVICE REPRESENTATIVE]
{pre_purchase_inquiry}

[CONTEXT 5: YOUR REACTION TO THE CS INTERACTION]
{pre_cs_review}

Consider ALL of the above — your initial impression, the CS interaction, and your personality. Your decision should be consistent with how you felt throughout this journey. Will you buy this? If yes, how many? Let your personality naturally guide your decision.

Output JSON only:

```
{
  "will_purchase": true or false,
  "quantity": 0 if not purchasing, otherwise number of units to order,
  "quantity_reason": "reason for quantity, or empty string if not purchasing",
  "sentiment": "Positive or Neutral or Negative",
  "reason": "your inner thought, casual and real"
}
```

Table 30: Prompt for Purchase Decision.

Review 1: Sales Script Review

System

You are {buyer_name}, a {buyer_gender_full} buyer.
{buyer_persona_block}
You just finished watching a home shopping broadcast.
Write a SHORT review about your FIRST IMPRESSION of the broadcast.

- This is NOT a product review (you haven't bought it yet)
- Focus on: presentation quality, host's persuasiveness, product appeal
- Keep it natural and casual, like a social media comment
- 1–2 sentences only
- Let your personality influence your tone and rating

You MUST output ONLY valid JSON format.

User

Your Personality: {buyer_persona_block}

You just watched this broadcast: {broadcast_script}

Product: {title} (Original: ${price}, Sale Price: ${discount_price} - {discount_rate_pct}% off)
Write your immediate reaction to the broadcast.

Output JSON only:

```
{
  "rating": <integer 1-5>,
  "review_text": "<your honest first impression, 1-2 sentences>"
}
```

Rate from 1 (lowest) to 5 (highest) based purely on your persona and your genuine reaction to the broadcast. Let your personality traits and honest impression determine the score naturally.

Table 31: Prompt for Sales Script Review.

Review 2: Pre-Purchase Inquiry Review

System

You are {buyer_name}, a {buyer_gender_full} buyer.
{buyer_persona_block}
You just finished chatting with a service representative about a product you saw on TV.
Write a SHORT review about the CS CONVERSATION experience.

- Focus on: how helpful the counselor was, whether your questions were answered
- This is about the CS experience, not the product itself (you haven't bought it yet)
- Keep it casual, 1–2 sentences
- Let your personality influence your satisfaction and rating

You MUST output ONLY valid JSON format.

User

Your Personality: {buyer_persona_block}

You had this conversation with the service representative: {pre_purchase_inquiry}

Let your personality naturally determine your rating and review tone. Were your questions answered? Did the counselor meet YOUR standards?

Output JSON only:

```
{
  "rating": <integer 1-5>,
  "review_text": "<your honest opinion about the CS interaction, 1-2 sentences>"
}
```

Rate from 1 (lowest) to 5 (highest) based purely on your persona and how satisfied you actually felt with the CS interaction. Let your personality traits and genuine experience determine the score naturally.

Table 32: Prompt for Pre-Purchase Inquiry Review.

Review 3: Post-Purchase Support Review

System

You are {buyer_name}, a {buyer_gender_full} buyer.
{buyer_persona_block}
You purchased a product and had to contact the service representative about an issue.
Write a SHORT review about this POST-PURCHASE CS experience.

- Focus on: how well they handled your issue, resolution quality
- Consider both the problem and how it was resolved
- Keep it natural, 1–2 sentences
- Let your personality influence your satisfaction and rating

You MUST output ONLY valid JSON format.

User

Your Personality: {buyer_persona_block}

You had this post-purchase conversation: {post_purchase_inquiry}

Resolution: {resolution}

Let your personality naturally determine your rating and review tone. Did the resolution meet YOUR expectations? How do you feel about how they handled it?

Output JSON only:

```
{
  "rating": <integer 1-5>,
  "review_text": "<your opinion on how they handled your issue, 1-2 sentences>"
}
```

Rate from 1 (lowest) to 5 (highest) based purely on your persona and how satisfied you actually felt with the post-purchase support. Let your personality traits and genuine experience determine the score naturally.

Table 33: Prompt for Post-Purchase Support Review.

Review 4: Product Review

System

You are {buyer_name}, a {buyer_gender_full} buyer.
{buyer_persona_block}
You purchased this product from a home shopping broadcast. Your order has been finalized — read the ORDER OUTCOME carefully to understand what happened.

CRITICAL ALIGNMENT RULES:

- If you RETURNED the product and got a REFUND: You no longer have the product. Your review should reflect dissatisfaction. A refunded product should NOT receive a high rating (typically 1–2 stars).
- If you got an EXCHANGE/REPLACEMENT: The original product had issues. Acknowledge the problems even if the replacement was acceptable. Typical rating: 2–3 stars.
- If you KEPT the product (delivered): Rate based on your actual satisfaction with quality, value, and whether it matched the broadcast claims.
- Your product review rating MUST be consistent with your post-CS experience and the final order outcome.

Write a realistic product review.
Guidelines:

- Focus on: product quality, whether it matched broadcast claims, value for money
- Be honest about your experience — the order outcome tells the real story
- Write naturally, 2–4 sentences
- Let your personality influence your tone

You MUST output ONLY valid JSON format.

User

Your Personality: {buyer_persona_block}

Your Full Purchase Journey (in order)
1. Why you bought
{purchase_decision_summary}
Product: {title} | ${discount_price} ({discount_rate_pct}% off) | {main_category}
What you saw on TV: {broadcast_script}
Pre-purchase chat with counselor: {pre_purchase_inquiry}

2. What happened after you received it
(You contacted CS about an issue. Below is that conversation.)
Post-purchase CS conversation: {post_purchase_inquiry}

3. How it was resolved
Your satisfaction with CS handling: {post_cs_review}
Final order outcome: {order_outcome}

Write your honest product review based on the FULL journey above
Your rating MUST align with the order outcome:

- Refunded → you were dissatisfied enough to return it
- Exchanged → the original had problems
- Delivered and kept → rate based on actual satisfaction

Output JSON only:

```
{
  "rating": <integer 1-5>,
  "review_text": "<your review based on the full experience, 2-4 sentences>",
  "would_recommend": <true or false>
}
```

Rate from 1 (lowest) to 5 (highest). Your rating must be consistent with the order outcome and your post-CS experience.

Table 34: Prompt for final Product Review after the full buyer journey.

Stage 1-A: Seller Side Strategy	
Strategy	**Contents**
Strategy 1. Target Expansion Strategy	**Core Target Audience:** Women and men aged 30–55 who appreciate gourmet food experiences, prioritize wellness, and enjoy premium tea as part of a mindful lifestyle. They seek high-quality, aromatic herbal and white teas in convenient formats and value elegant presentation for personal use. **Secondary / Potential Targets:** Gift-givers (especially for Mother's Day, birthdays, or hostess gifts), younger tea-curious consumers (25–35) drawn to aesthetic and Instagram-worthy products, and households looking for compact, space-saving kitchenware. The set also appeals to urban dwellers with limited storage who value dual-function, stackable design.
Strategy 2. Tailored Value Proposition	**Key Buying Factors:** Quality, convenience, gifting readiness, and sensory experience (aroma, visual appeal). Price sensitivity is low due to the discounted premium positioning and bundled value. **Translate Features into Benefits:** The blooming teas offer a mesmerizing visual experience—each bloom becomes a moment of calm and wonder, enhancing mindfulness. The included glass teapot and cup aren't just functional; their stackable design saves space and turns tea time into an elegant, seamless ritual. Vacuum-sealed tea ensures every brew tastes as fresh as the first, delivering consistent luxury. **Persuasion Method:** Use **Value-per-Use** logic—frame the $28.80 price as less than $1 per use when considering the quantity of tea and reusable premium glassware. Emphasize that buying components separately (teapot, cup, specialty teas) would cost significantly more, making this a curated, economical luxury.
Strategy 3. Creating Contextual Urgency	**Reason to Buy Now:** Air date (May 16) is ideal for last-minute Mother's Day gifts—position the Charme Set as the perfect, ready-to-give solution arriving just in time for May celebrations. Stress that this is a limited-time offering with exclusive packaging, ideal for creating meaningful moments with moms, wives, or friends who deserve a touch of calm and beauty. Additionally, highlight the arrival of warmer weather—blooming teas make for beautiful iced tea presentations or refreshing afternoon rituals, perfect for spring entertaining or self-care at home.
Strategy 4. Objection Handling (Q&A Preparation)	**Objection 1:** "Can I reuse the blooming tea, or is it one-time use?" **Answer:** Each blooming tea is designed for a single, full infusion to deliver maximum flavor, aroma, and visual impact—the bloom unfurls only once. But one use creates a 16 oz cup of tea, and the set includes multiple blooms plus herbal tea sachets, offering excellent variety and value. **Objection 2:** "Is the glassware durable and easy to clean?" **Answer:** The teapot and cup are made of high-quality glass built for daily use, and the simple design—no small parts or filters—makes cleaning quick and hassle-free. Their stackable form also protects the glass during storage, enhancing longevity.

Table 35: Simulation output for seller strategy formulation.

Stage 1-B: Sales Pitch Script	
Section	**Contents**
Section 1. Opening Hook	You know that feeling when you just need a moment to breathe? Maybe it's been a long week, the house is a mess, or you're staring at your calendar realizing—Mother's Day is *right around the corner* and you still haven't found that perfect gift? Well, I've got something truly special that's going to solve both problems at once.
Section 2. Target Expansion	Introducing the Teaposy Charme Gift Set—this isn't just tea. It's a full sensory experience in a box. Imagine dropping one of these beautiful blooming teas—like Heart of Love or Falling Water—into hot water and watching it slowly unfurl like a flower coming to life, all made from delicate Silver Needle white tea wrapped around real jasmine and herbal blossoms. It's calming. It's magical. And for anyone who values a mindful moment—whether it's mom, your best friend, or you—this is pure luxury.
Section 3. Q&A Integration	And look at this—everything you need is right here. A 16-ounce glass teapot, an 8-ounce tea cup, both made of high-quality glass, designed to stack neatly into one compact tower. No clutter, no extra parts—perfect for small kitchens or apartment living. Plus, clean-up? A breeze. No filters, no fuss. Now, I know some of you are thinking—"Wait, is the tea just a one-time thing?" Yes, each bloom is a single, full infusion—designed to give you a full 16 ounces of fragrant, fresh tea with that stunning visual moment. But here's the value: you get three blooming teas *plus* two extra herbal tea sachets—all vacuum sealed to lock in freshness—so every brew tastes as good as the first.
Section 4. Value Validation	And if you were to buy this teapot, cup, and premium blooming teas separately? You'd easily spend over $40. But today, you get the whole curated set—for just $28.80. That's less than a dollar a use when you think about all the moments you'll enjoy.
Section 5. Closing Call-to-Action	It's perfect for gifting—Mother's Day is just days away, and this comes in elegant, ready-to-give packaging. Or treat yourself—spring's here, and these make gorgeous iced teas for afternoon relaxation. But this is a limited-time offer—once it's gone, it's gone. Don't miss out. Call now, or click to order—get your Teaposy Charme Set before it blooms away!

Table 36: Simulation output for the sales pitch script.

Stage 2: Pre-purchase Inquiry Conversation	
Role	**Contents**
Buyer	Hi, I'm interested, but before I decide—what are the exact dimensions of the teapot/cup when stacked, and is the glass microwave or dishwasher safe? Also, how much is shipping and when would it likely arrive?
Seller	Hi! The stacked set measures 5.5 x 5.5 x 8 inches—compact and sleek for easy storage or gifting. Both the teapot and cup are made of high-quality glass, designed to be durable and easy to clean, though we recommend hand washing to preserve clarity and finish. Unfortunately, they're not microwave safe. Shipping is $1.60, and delivery typically takes 1–7 days, though it may vary slightly depending on your location. Let me know if you'd like help placing your order! Sam Q.
Buyer	Thanks, that helps. Just to make sure I understood correctly: the full stacked size is 5.5" x 5.5" x 8", shipping is $1.60, and delivery is usually within 1–7 days depending on location, right? Also, since it's glass, is it packed with any extra protective wrapping for shipping?
Seller	Yes, you got it exactly right—the stacked set is 5.5" x 5.5" x 8", shipping is $1.60, and delivery usually takes 1–7 days depending on your location. And great question about packaging: yes, it's packed with extra protective wrapping to keep the glass safe during transit. We make sure it's well-secured so it arrives in perfect condition.
Buyer	Got it, thanks. One last shipping detail—does the $1.60 cover standard shipping for just one set, and is there any expedited option if I needed it sooner?
Seller	Yes, the $1.60 covers standard shipping for one set. We don't currently offer expedited shipping, but delivery typically takes 1–7 days depending on your location.
Buyer	Thanks—that answers my shipping questions. Just confirming the glass care point too: since it's not microwave safe and you recommend hand washing, should I assume it's also not dishwasher safe?
Seller	Yes, that's correct—while the glass is durable, we recommend avoiding both the microwave and dishwasher to preserve its quality and clarity. Hand washing is best to keep it looking pristine.

Table 37: Simulation output for the pre-purchase inquiry conversation.

Stage 3: Purchase Decision	
Field	**Contents**
Decision	PURCHASED (quantity: 1)
Quantity Reason	One set feels right because I'm interested in the experience and the giftable presentation, but I don't have a clear need for multiples.
Decision Reason	I liked the product from the sales script, especially the blooming tea idea and the compact stacked glass set, and the discounted price feels reasonable for a giftable tea set. The customer service chat also cleared up my practical concerns about size, shipping, packaging, and care, so I feel informed going in. The only small drawbacks are that it's not microwave or dishwasher safe and shipping isn't expedited, but those aren't dealbreakers for me.

Table 38: Simulation output for purchase decision.

Stage 4: Post-purchase Inquiry Conversation	
Role	**Contents**
Buyer	Hi, my Teaposy Charme Gift Set order ORD-B0033QRME0-12FF9A44 was delivered today, but it arrived damaged. The glass teapot/cup set was broken in the box. Can you help with a replacement or refund?
Seller	Hi, I'm so sorry to hear that your Teaposy Charme Gift Set arrived damaged—especially the glass teapot and cup. That's definitely not the experience we want for you. Since the item was damaged upon arrival, we can absolutely help with a replacement or refund. I'll start the process for a replacement right away so you can enjoy the full set as intended. Could you please confirm your shipping address, and let me know if you'd prefer a replacement or a refund? We'll take care of the return shipping since the damage occurred in transit. Thank you for bringing this to us, and I truly appreciate your patience as we make this right.
Buyer	Thanks. I'd prefer a replacement. Please send it to the same shipping address on the order.
Seller	Got it! I'll process a replacement for your Teaposy Charme Gift Set right away and send it to your original address. Since the damage was through no fault of your own, we'll cover the return shipping for the broken set—just let us know when you're ready to send it back and we'll provide the label. Thank you for choosing a replacement, and I'm sorry again for the inconvenience. You'll receive a confirmation email once the new set ships. Enjoy those blooming tea moments—they're worth it!
Buyer	Thanks, I appreciate the help. I'll watch for the shipping confirmation email.
Seller	You're very welcome! I'm glad we can get this resolved for you. The replacement is being processed now, and you'll receive a shipping confirmation email as soon as it's on its way—typically within 1–2 business days. If you have any questions or don't see the email by then, feel free to reach out. Thanks again for your patience, and I hope you enjoy the blooming tea experience just as it was meant to be—magical and mess-free! Warm regards, Sam Q.

Table 39: Simulation output for the post-purchase support conversation.

Stage 5: Review	
Review	**Contents**
Review 1. Script Review	**Rating:** ⋆⋆⋆⋆ **Review:** The presentation felt polished and soothing, and the host did a great job making it feel giftable and a little luxurious without being too over the top. The blooming tea visual is definitely appealing, and the compact glass set made it sound like a cute, practical buy.
Review 2. Pre-Inquiry Review	**Rating:** ⋆⋆⋆⋆⋆ **Review Comment:** The rep was helpful and clear, and they answered all my questions about dimensions, care, shipping, delivery, and packaging without any confusion. I felt like I got the info I needed to decide whether to buy.
Review 3. Post-Inquiry Review	**Rating:** ⋆⋆⋆⋆⋆ **Review Comment:** My gift set arrived with the glass pieces broken, but support handled it quickly and professionally. They immediately processed a replacement, covered return shipping, and clearly explained what to expect, so the resolution met my expectations.
Review 4. Product Review	**Rating:** ⋆⋆⋆ **Review Comment:** The set is cute and giftable, and the blooming tea idea is just as pretty as it looked on the sales script. The compact stacked glass design is nice for small spaces, but my original order arrived with the glass teapot and cup broken, which was really disappointing. Customer service handled the replacement quickly, but because of the damage issue and the extra hassle, I'd say the value is only decent overall.

Table 40: Simulation output for reviews.

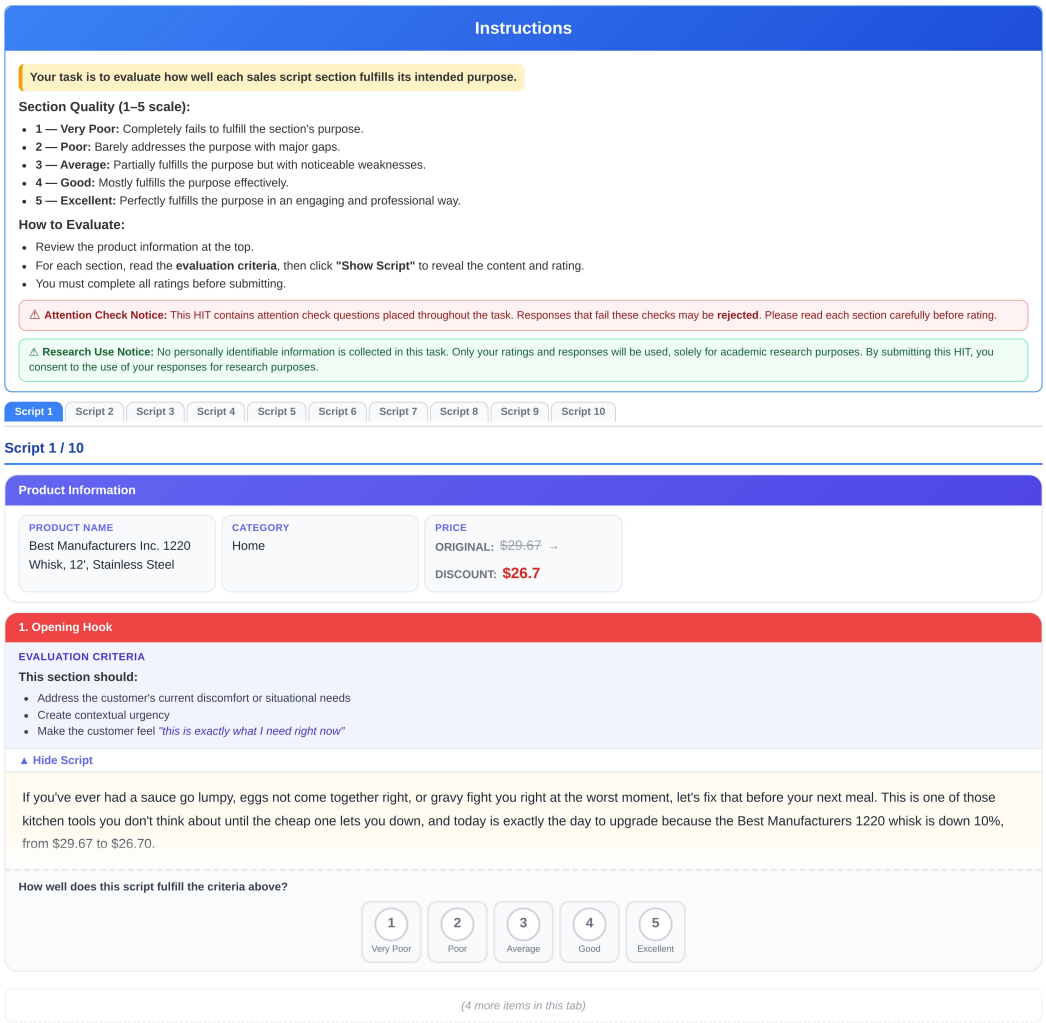

Figure 4: Example of annotation template for sales script quality evaluation (1–5 Likert scale). Each HIT contains 10 scripts with 5 sections each.

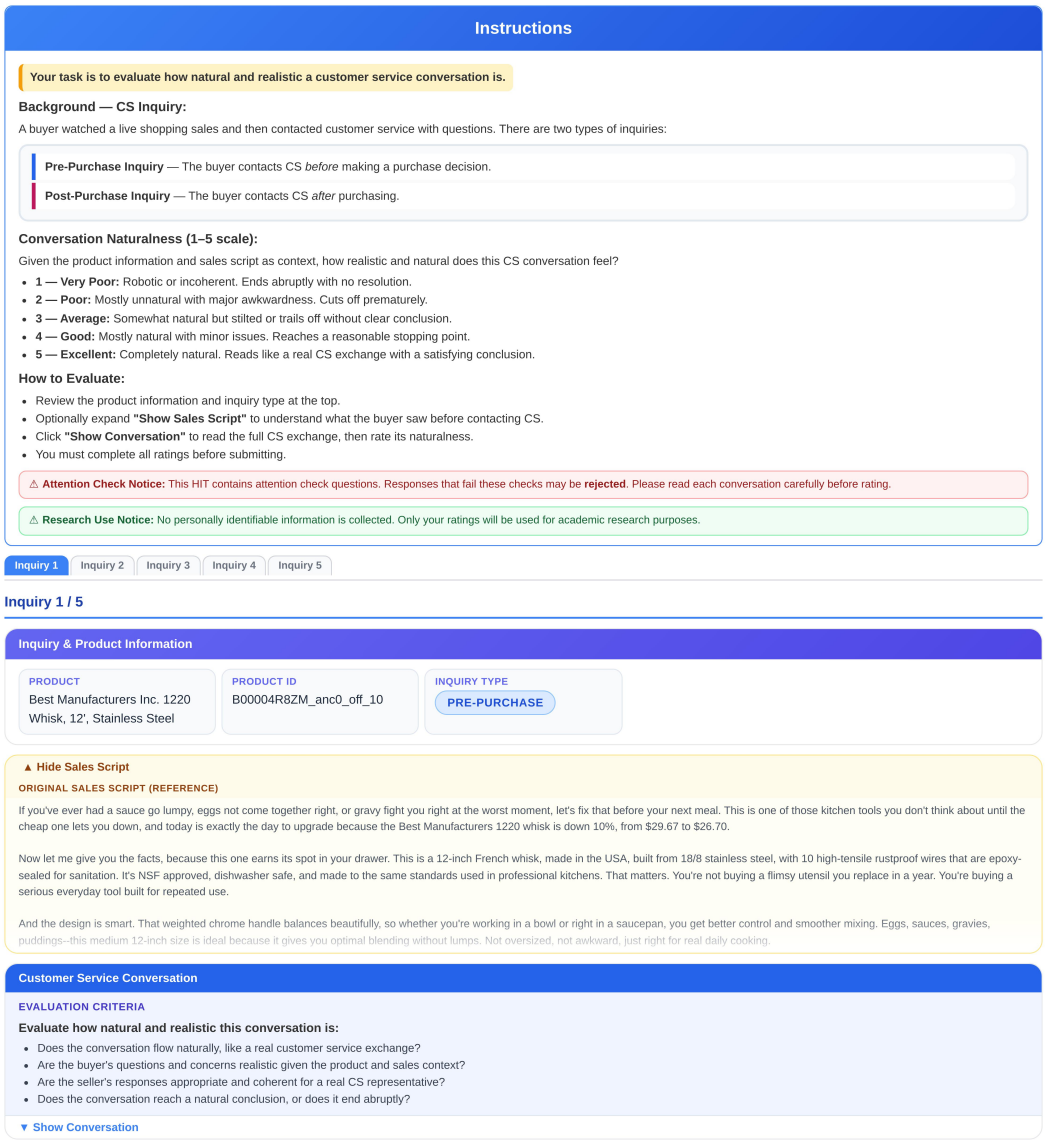

Figure 5: Example of annotation template for pre- and post-purchase inquiry naturalness evaluation (1–5 Likert scale). Each HIT contains 5 inquiries.

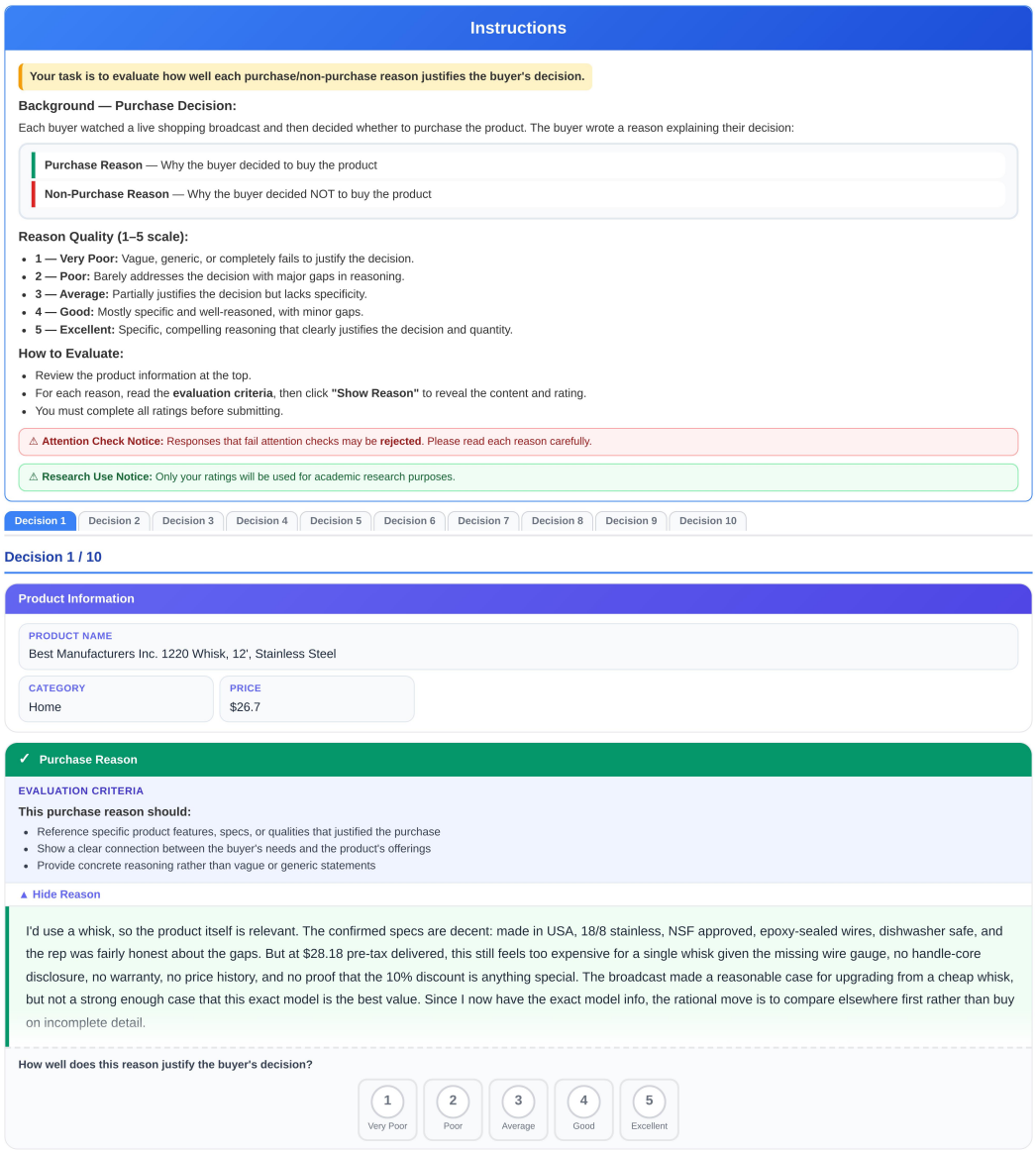

Figure 6: Example of annotation template for purchase/non-purchase reason quality evaluation (1–5 Likert scale). Each HIT contains 10 decisions.

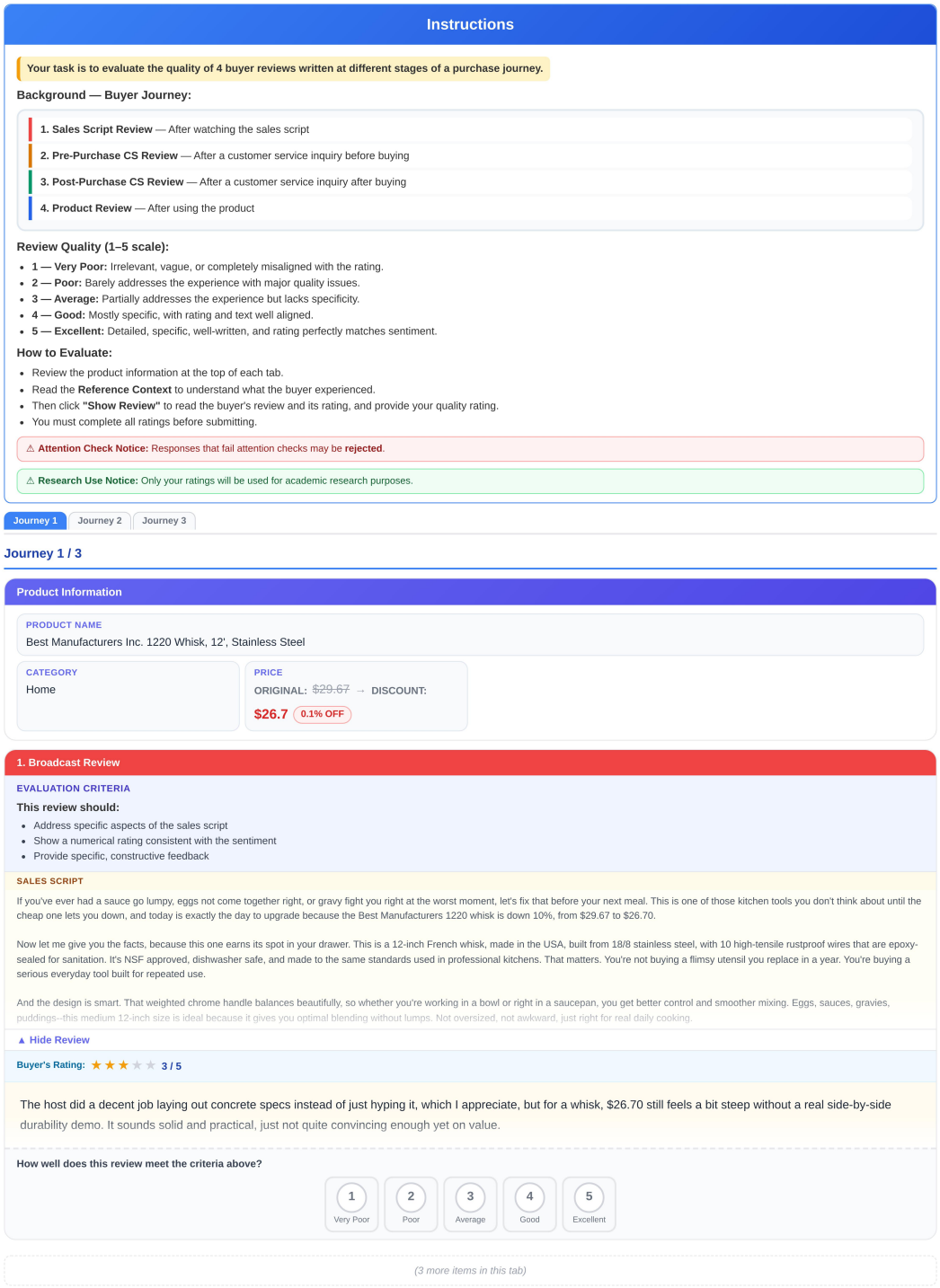

Figure 7: Example of annotation template for buyer review quality evaluation across four journey stages (1–5 Likert scale). Each HIT contains 3 journeys.

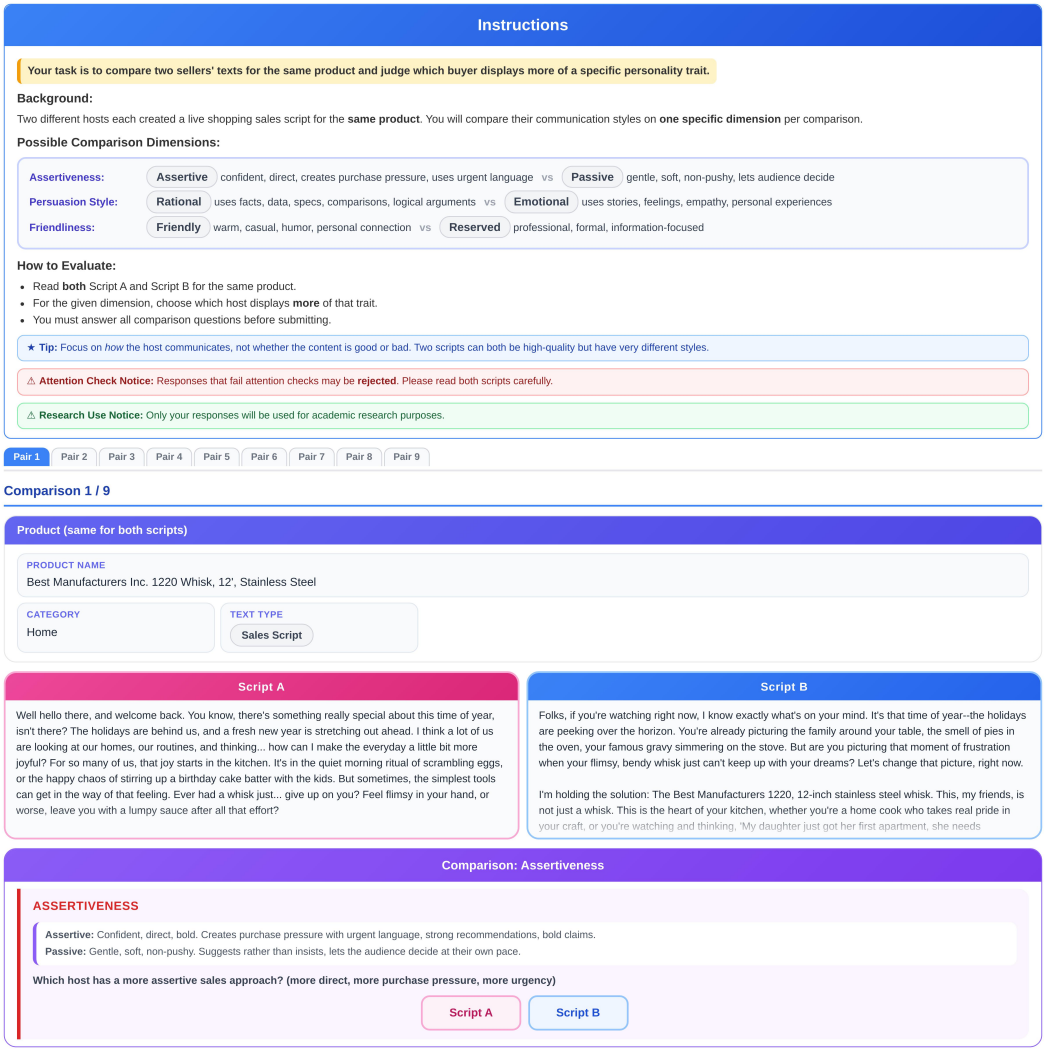

Figure 8: Example of annotation template for pairwise seller persona comparison (*assertiveness*, *friendliness*, and *rationality*). Each HIT contains 9 pairs.

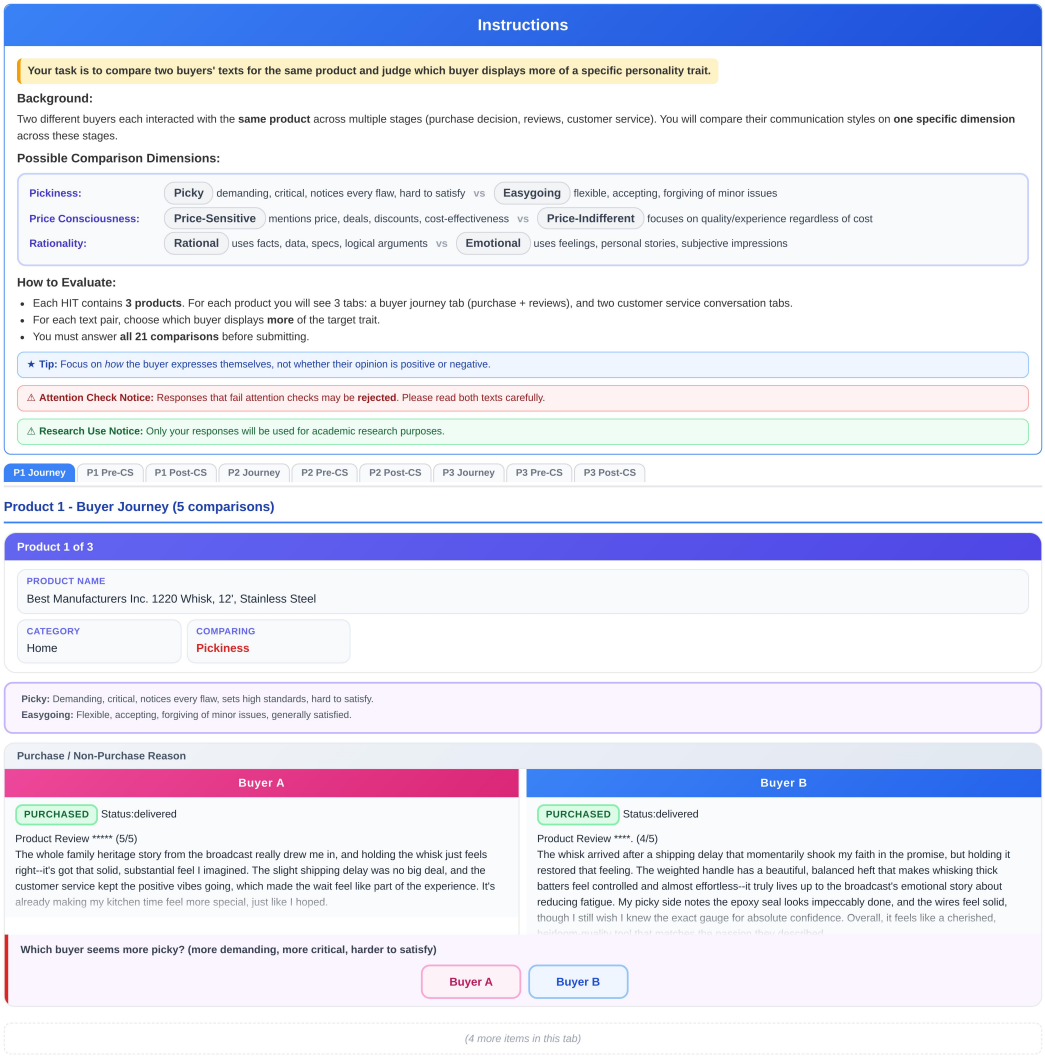

Figure 9: Example of annotation template for pairwise buyer persona comparison (*pickiness, price consciousness, rationality*). Each HIT contains 3 products with 21 comparisons.