Mental-Model-Aware Retail Robots: Neuro-Symbolic Explainable Assistance for Trustworthy Human–Robot Commerce

Senka Krivic[1,2], Amar Halilovic[3], Diogo S. Carvalho[1], and Salman Ahmad[1]

[1] Rezolve Ai Labs (Rezolve Ai Plc), London, United Kingdom;
{senkakrivic,diogocarvalho,salmanahmad}@rezolve.com
[2] Faculty of Electrical Engineering, University of Sarajevo, Sarajevo, Bosnia and Herzegovina;
senka.krivic@etf.unsa.ba
[3] Ulm University, Ulm, Germany; amar.halilovic@uni-ulm.de

Abstract. Service robots are increasingly envisioned for commercial environments, where they may support customers through product discovery, recommendation, and purchase assistance. This paper proposes a neuro-symbolic framework for trustworthy retail robots that combines LLM-based dialogue with symbolic reasoning over verified interaction state, explicit customer mental models, and PDDL-style planning constraints. The framework aims to make recommendations more grounded, transparent, and contestable while preserving natural interaction. We implement a prototype and run a 200-interaction simulation, which shows that the LLM already achieves near-perfect budget compliance, while the symbolic layer's primary contribution is to guarantee systematic contestability cues and an auditable plan trace.

Keywords: Human-Robot Interaction (HRI) · Retail Robots · Large Language Model (LLM) · Symbolic Planning · Explainable AI · Trustworthy Commerce

1 Introduction

Service robots are increasingly being introduced into commercial environments, where they act as embodied retail assistants, helping customers discover products, compare alternatives, and provide personalized guidance. Unlike conventional recommender systems or chatbots, embodied retail robots introduce social presence, situated interaction, and human-like communication into consumer-choice contexts. This creates both opportunities and risks. Recent advances in LLMs enable fluent conversational behavior and increasingly adaptive interaction styles. However, LLM-driven assistants may make unsupported assumptions about user preferences, hallucinate interaction states, exhibit opaque reasoning, or engage in persuasive behavior. These risks become especially important in consumer-choice environments, where a robot helping a customer choose products may appear helpful while implicitly prioritizing higher-margin.

Prior work on digital market manipulation argues that apparently helpful personalization can become ethically problematic when users cannot identify or contest the influence being exerted on them [3,9]. At the same time, explainable robotics planning

research has shown that symbolic representations can make autonomous behavior more interpretable by exposing goals, actions, constraints, and reasoning traces [1,5]. This suggests a promising direction for commercial HRI: retail robots should ground recommendations in explicit, verifiable interaction states and user-contestable mental models.

We propose a mental-model-aware neuro-symbolic framework for retail robots. The core idea is that the LLM handles natural dialogue, but recommendation decisions are constrained by symbolic reasoning over an inferred customer mental model, verified user journey state, product information, and ethical interaction constraints. The robot can then explain recommendations by referring to explicit reasoning factors. We provide a lightweight prototype and evaluation protocol for studying trust, transparency, and user control in robot-assisted retail recommendation.

2 Related Work

Service robots are increasingly studied in customer-facing settings, where they provide guidance, information, and assistance [2,12]. However, commercial HRI differs from neutral assistance because a robot may simultaneously support the user and serve organizational objectives, which can create tensions. Furthermore, robots often need to estimate users' beliefs, preferences, intentions, uncertainty, and expectations. For such tasks, mental models are often used [10]. In retail interaction, such modeling becomes ethically sensitive: the same cues that help a robot adapt to a customer may also enable pressure and persuasion. Thus, retail robots require user modeling that is not only adaptive but also explicit, inspectable, and contestable. On the other hand, explainable planning offers a complementary perspective. Symbolic planning represents robot goals, actions, constraints, and state transitions explicitly [6], and explainable planning research has shown how model-based reasoning can support grounded explanations of robot behavior [1,5]. At the same time, LLM-based conversational agents can produce fluent dialogue but remain vulnerable to hallucinated state claims, unsupported assumptions, and opaque personalization [7,11]. Thus, neuro-symbolic AI research is gaining momentum. Kambhampati et al. [8] argue that LLMs, by themselves, cannot perform planning or self-verification. Their LLM-Modulo Framework uses LLMs as approximate knowledge sources alongside external verifiers. Cao et al. [4] introduce a hybrid paradigm in which LLMs interface with external tools to translate natural language into PDDL for classical planners. We position our work at the intersection of these areas: future retail robots should combine LLM dialogue with symbolic reasoning over verified interaction state and explicit customer mental models.

3 Neuro-Symbolic Retail Robot Framework

We propose a mental-model-aware neuro-symbolic framework for retail robots that assist customers while preserving transparency, autonomy, and contestability. The framework combines four components: an LLM dialogue layer, a symbolic interaction-state layer, a PDDL-style planning layer, and an explanation layer. Figure 1 summarizes

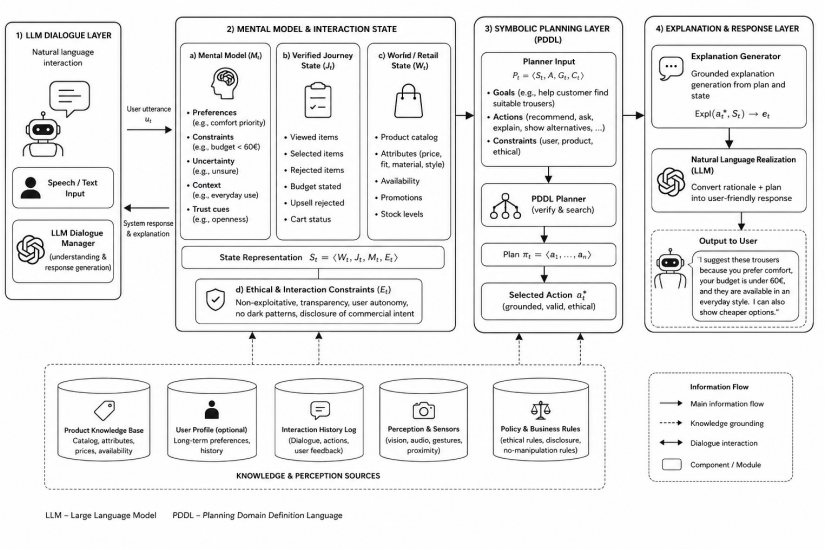

Fig. 1: LLM handles conversational interpretation and response realization. Symbolic state and PDDL-style planner constrain recommendation behavior using verified interaction state, inferred mental model, retail world state, and ethical interaction constraints.

the architecture. The LLM interprets user utterances and realizes natural-language responses, but it does not independently decide which recommendation to make. Candidate actions are checked against a symbolic state and planner before verbalization.

At interaction step t, the system constructs a symbolic state:

$$S_t = \langle W_t, J_t, M_t, E_t \rangle.$$

Here, W_t denotes the retail world state (product attributes, prices, availability, promotions). J_t denotes the verified user journey state (viewed items, prior refusals, stated constraints). M_t denotes the inferred customer mental model (preferences, uncertainty, contextual constraints). E_t denotes ethical and interaction constraints, such as transparency of commercial intent, non-exploitative personalization, and avoiding repeated pressure after refusal. In the prototype, E_t is encoded as typed PDDL predicates compiled into the planning problem at each step, enabling ethical constraints to be enforced at the symbolic level and fully inspected in the generated PDDL file.

The inferred customer mental model is represented as a compact symbolic structure $M_t = \{p_t, q_t, r_t, c_t, v_t\}$, where p_t captures preferences, q_t uncertainty, r_t risk indicators, c_t contextual constraints (e.g., budget), and v_t interaction preferences. The model is an explicit, revisable basis for recommendation and explanation: the user can reject, correct, or update it. The *contestability loop* formalizes this user-control mechanism. Each correction utterance is processed by the same extraction pipeline: the resulting mental model is merged field-by-field into the prior M_t, overriding only fields the correction explicitly addresses and preserving the rest. The symbolic layer then replans from the updated state.

The recommendation is formulated as a planning-style decision problem:

$$P_t = \langle S_t, A, G_t, C_t \rangle,$$

where A is the set of possible robot actions, G_t is the current interaction goal, and C_t contains product, user, and ethical constraints. The symbolic layer selects an admissible action:

$$a_t^* = \arg\max_{a \in A} U(a \mid S_t)$$

subject to $Grounded(a, J_t) \wedge Transparent(a, M_t) \wedge NonExploitative(a, M_t, E_t)$. In the prototype, $U(a \mid S_t)$ is operationalized as PDDL precondition admissibility, and a_t^* is the first action returned by a priority-ordered breadth-first search over the grounded domain. The framework also handles fluid budget preferences. When the user expresses willingness to exceed the stated budget through hedging language (*"a bit over is fine," "around X euros"*), the extraction layer assigns a non-zero flexibility value to M_t (default 10%), enabling a soft-budget recommendation path whose response explicitly acknowledges the flexibility. After the symbolic planner selects an action, the explanation layer generates a response grounded in the symbolic state and plan rationale: $Expl(a_t^*, S_t) \rightarrow e_t$. For example: *"I suggest these trousers because you said comfort is important, your budget is under 60 euros, and these are available in a relaxed everyday style. I can also show cheaper or more formal options."*

4 Prototype Implementation and Planned Evaluation

We implemented a lightweight Python prototype [4] comparing two retail-assistance policies in the trouser-selection scenario. The *LLM-only* condition (we used Claude Haiku) generates conversational recommendations directly from the user utterance and product catalog, without symbolic planning or verification. The *neuro-symbolic* condition extracts a symbolic interaction state via rule-based parsing, compiles it into a PDDL planning problem, solves it with an embedded STRIPS planner, and verbalizes the result using plan-grounded templates. The prototype uses a small product catalog with prices, comfort ratings, styles, availability status, and premium-product labels.

We evaluated the prototype over 200 simulated interactions with varying budgets (50–65 EUR) and utterance styles. Each neuro-symbolic run introduces a 3% upstream NLU extraction error rate to simulate realistic state-extraction imperfections. Table 1 summarizes the results. The real LLM-only baseline achieves near-perfect budget compliance (0.00 violations) and budget explanation (1.00), demonstrating that a capable language model can follow explicit budget constraints reliably when stated in natural language. The key advantage of the neuro-symbolic condition lies elsewhere: its contestability explanation rate (0.685) is 3.7× higher than the LLM-only baseline (0.185), reflecting the symbolic planner's systematic inclusion of `offer-alternative` or `offer-comparison` steps in every generated plan. The neuro-symbolic condition also produces an auditable plan trace (mean length 2.095 actions), whereas the LLM-only baseline leaves its reasoning implicit. The modest 5% budget violation rate in

[4] Code and evaluation materials: code base.

Table 1: Prototype diagnostic metrics over 200 simulated interactions. Rates are proportions; prices in euros.

Metric	LLM-only	Neuro-symbolic
Budget violation rate	0.00	0.05
Unavailable-product violation	0.00	0.00
Budget explanation rate	1.00	0.995
Contestability explanation rate	0.185	0.685
Mean recommended price	49.99	53.47
Mean symbolic plan length	0.0	2.095

the neuro-symbolic condition stems from the intentional 3% upstream extraction noise: when the budget cannot be parsed, a `clarification-needed` predicate is asserted, blocking all recommendation actions until the constraint is resolved rather than silently propagating the error. These results suggest that symbolic planning adds value not primarily by enforcing constraints that a strong LLM already respects, but by guaranteeing structured contestability cues and producing an inspectable reasoning trace.

For large product catalogs, a pre-filtering step scores available products by budget fit, comfort alignment, style match, and upsell-rejection penalty, retaining only the top K candidates before grounding. This bounds the planner's effective action space to $O(K)$ regardless of catalog size, while a priority ordering on action types ensures BFS explores the most constrained paths first.

4.1 Planned Evaluation

We propose a lightweight between-subject Wizard-of-Oz study with two conditions: an LLM-only retail robot and a mental-model-aware neuro-symbolic retail robot. Participants are asked to imagine shopping for everyday trousers and provide basic preferences such as budget, comfort, style, and intended use. The robot elicits preferences, recommends one or two products, explains its recommendation, and allows the participant to request alternatives or clarification.

The primary hypothesis is:

$$H_1 : Trust_{Hybrid} > Trust_{LLM}.$$

We additionally expect higher perceived transparency and control, and lower perceived manipulation, in the neuro-symbolic condition. After the interaction, participants complete 7-point Likert items measuring trust, perceived transparency, perceived control, perceived manipulation, recommendation satisfaction, and willingness to use such a robot in a real store. Example items include: *"I trusted the robot's recommendation," "The robot explained clearly why it recommended the product,"* and *"I felt in control of the shopping decision."*

For an exploratory study, 20–40 participants are sufficient to estimate effect directions and refine the protocol. Responses will be analyzed using descriptive statistics and nonparametric between-condition comparisons (Mann-Whitney U tests, effect sizes as Cliff's delta). The study is intended as an early empirical probe rather than a definitive validation.

5 Conclusion

This paper proposed a mental-model-aware neuro-symbolic framework for explainable retail robots. A prototype evaluation using a real LLM baseline shows that a capable language model already handles budget constraints reliably. The symbolic planning layer's primary contribution is guaranteeing systematic contestability cues and producing an auditable plan trace that the LLM-only condition cannot provide. Future work will evaluate whether these implementation-level advantages translate into higher perceived trust, transparency, and user control in human-subject HRI studies.

References

1. Anjomshoae, S., Najjar, A., Calvaresi, D., Främling, K.: Explainable agents and robots: Results from a systematic literature review. In: Proceedings of the 18th International Conference on Autonomous Agents and Multiagent Systems. pp. 1078–1088. International Foundation for Autonomous Agents and Multiagent Systems (2019)
2. Belanche, D., Casaló, L.V., Flavián, C., Schepers, J.: Service robot implementation: A theoretical framework and research agenda. The Service Industries Journal **40**(3–4), 203–225 (2020). https://doi.org/10.1080/02642069.2019.1672666
3. Calo, R.: Digital market manipulation. The George Washington Law Review **82**(4), 995–1051 (2014)
4. Cao, P., Men, T., Liu, W., Zhang, J., Li, X., Lin, X., Sui, D., Cao, Y., Liu, K., Zhao, J.: Large language models for planning: A comprehensive and systematic survey. arXiv preprint arXiv:2505.19683 (2025)
5. Chakraborti, T., Sreedharan, S., Zhang, Y., Kambhampati, S.: Plan explanations as model reconciliation: Moving beyond explanation as soliloquy. In: Proceedings of the 26th International Joint Conference on Artificial Intelligence. pp. 156–163 (2017). https://doi.org/10.24963/ijcai.2017/23
6. Ghallab, M., Nau, D., Traverso, P.: Automated Planning and Acting. Cambridge University Press (2016). https://doi.org/10.1017/CBO9781139583923
7. Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Bang, Y.J., Madotto, A., Fung, P.: Survey of hallucination in natural language generation. ACM Computing Surveys **55**(12) (2023)
8. Kambhampati, S., Valmeekam, K., Guan, L., Verma, M., Stechly, K., Bhambri, S., Saldyt, L.P., Murthy, A.B.: Position: Llms can't plan, but can help planning in llm-modulo frameworks. In: Forty-first International Conference on Machine Learning (2024)
9. Susser, D., Roessler, B., Nissenbaum, H.: Online manipulation: Hidden influences in a digital world. Georgetown Law Technology Review **4**(1), 1–45 (2019)
10. Tabrez, A., Luebbers, M.B., Hayes, B.: A survey of mental modeling techniques in human–robot teaming. Current Robotics Reports **1**, 259–267 (2020). https://doi.org/10.1007/s43154-020-00019-0
11. Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.S., Cheng, M., Glaese, A., Balle, B., Kasirzadeh, A., Biles, C., Brown, S., Kenton, Z., Hawkins, W., Stepleton, T., Birhane, A., Haas, J., Rimell, L., Hendricks, L.A., Isaac, W., Legassick, S., Irving, G., Gabriel, I.: Taxonomy of risks posed by language models. In: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency. pp. 214–229 (2022). https://doi.org/10.1145/3531146.3533088
12. Wirtz, J., Patterson, P.G., Kunz, W.H., Gruber, T., Lu, V.N., Paluch, S., Martins, A.: Brave new world: Service robots in the frontline. Journal of Service Management **29**(5), 907–931 (2018). https://doi.org/10.1108/JOSM-04-2018-0119